NetMob Madrid 2023

4-6 October 2023, Universidad Carlos III de Madrid, Spain

Book of Abstracts

Steering Committee

Vincent Blondel (University of Louvain)
Leo Ferres (IDSUDD, ISI Foundation & Telefónica)
Marco Fiore (IMDEA Networks Institute & Net AI)
Vanessa Frias-Martinez (University of Maryland)
Renaud Lambiotte (University of Oxford)
Esteban Moro (MIT and Universidad Carlos III de Madrid)
Alex 'Sandy' Pentland (MIT)

Main conference

Talks

Behavior-based dependency networks shape economic resilience of cities

Takahiro Yabe[1], Bernardo Garcia Bulle Bueno[1], Morgan Frank[1,2], Alex Pentland[1], Esteban Moro[1,3]
[1]MIT, [2]University of Pittsburgh, [3]Universidad Carlos III de Madrid

Keywords: economic resilience, human mobility, urban networks

Introduction

Quantifying the economic costs of businesses caused by extreme shocks, such as the COVID-19 pandemic and natural disasters, is crucial for developing preparation, mitigation, and recovery plans. Conventionally, survey data have been the primary source of information used to measure losses inflicted on businesses by negative shocks, however, drops in foot traffic quantified using large scale human mobility data (e.g., mobile phone GPS) have recently been used as low-cost and scalable proxies for such losses, especially for businesses that rely on physical visits to stores, such as restaurants and cafes (Yabe et al., 2019). Such studies and analyses often quantify the losses in foot traffic based on individual points-of-interest (POIs), neglecting the interdependent relationships that may exist between businesses and other facilities. For example, university campus lockdowns imposed during the COVID-19 pandemic may severely impact foot traffic to student-dependent local businesses. Such dependent relationships between businesses could cause secondary and tertiary cascading impacts of shocks and policies, posing a significant threat to the economic resilience of business networks (Zhai and Yue, 2021).

Data and Methods

To identify such spillover effects across places due to behavioral changes, we build a network of dependencies between business using mobility data. We used large-scale anonymous, privacy-enhanced mobility data of more than 1 million devices from five metropolitan areas in the US, collected in 2019 to 2021 (Moro et al., 2021). The dependency scores were computed based on the foot traffic patterns before the pandemic (September 2019 to January 2020). We compute the dependence of a target POI i on a source POI j by $dep(i,j) = n_{ij}/n_i$, where v_i and v_j denote the sets of users who visit POIs i and j respectively. Because the denominator is based on the number of users who visit the target POI i, $dep(i,j) \neq dep(j,i)$. This is a simple but intuitive measure that considers the asymmetric nature of dependencies between POIs. The set of users who visit each POI in a specific period is computed using mobility data collected from mobile phone devices. Figure 1 shows an overview of the methods on how the dependency between two POIs is computed. The map shows the dependency network in the New York's Manhattan area. Commercial areas such as Times Square, Hudson Yards, as well as college and arts institutions, such as New York University and the Metropolitan Museum, can be seen as places with high degrees of in-weights. The average dependency weight between POIs i and j decreases as the Haversine distance increases, however, a simple regression model shows that the structural factors (distance between nodes, size, area, and category of nodes) only explain around 17% to 22% of the variance observed in the dependency weights, suggesting that the dependency network encodes specific and nuanced relationships between places.

Economic Impacts of Dependency Networks

Statistical analysis of the behavior-based dependency network shows that the networks encode place-specific relationships between businesses and urban amenities that cannot be fully captured by physical characteristics alone. Here we empirically investigate whether the dependency network could improve the predictability of how shocks cascade across places in cities, using the COVID-19 pandemic as an example of an extreme shock. The observed change in visits to different places is computed by $\widetilde{v_i} = (v_i^{COVID}/v_i^{pre}) - 1$, where v_i^{COVID} and v_i^{pre}denote the number of visits to place i during the pandemic (March - June 2020) and the pre-pandemic period (September to December 2019), respectively. The following model regresses the loss in visits to the ego using the loss of visits to the alters weighted by the dependency network weights, along with some fixed effects about the area and subcategory that the ego

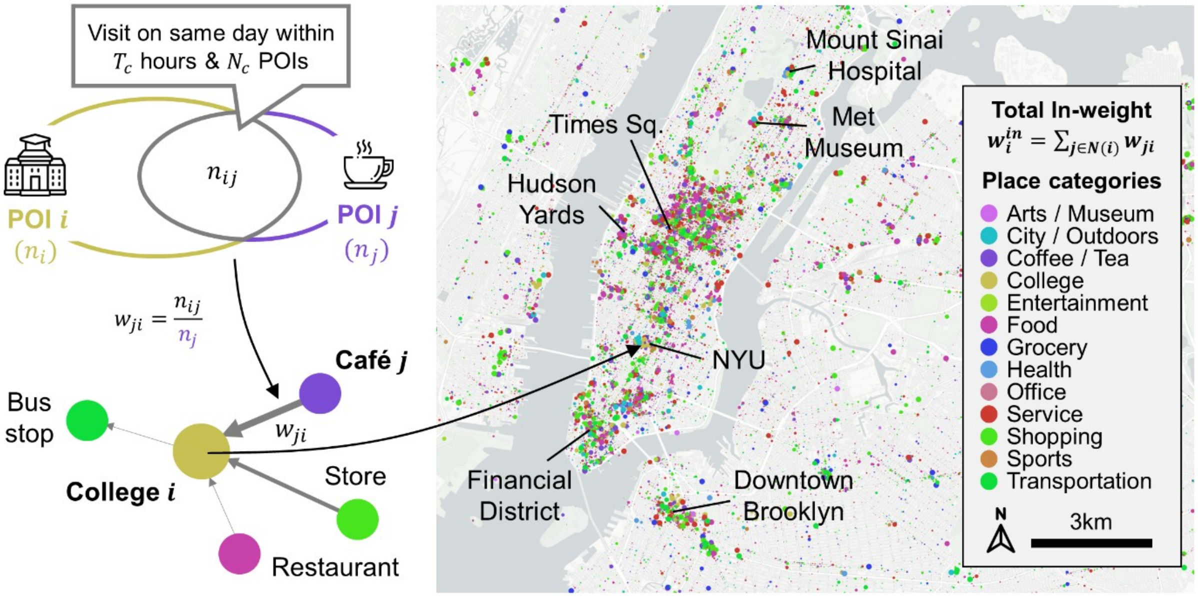


Figure 1. Behavior-based dependency network. (left) Definition of behavior-based dependency network. Dependency between POIs i and j are computed using the number of common and total visitors to each POI. (right) Dependency in-weights of POIs in the New York Manhattan area shows high dependence of places on commercial districts such as Times Square, Hudson Yards, and to institutions such as NYU, Mount Sinai Hospital, and the Metropolitan Museum.

POI is located at. As a baseline for comparison, we also prepare an alternative model that models the dependency between places based on the physical distance instead of the dependency weight w_{ij}.

$$\tilde{v_i} \sim \sum_j w_{ij}\, \tilde{v_j} + Category_i + Area_i$$

Across the five metropolitan areas, and across different periods during the pandemic, we found that using the behavior-based dependency network significantly improves the R^2 compared to using the physical distance-based network model. Comparing the regression coefficients (variables were standardized before analysis), the effects of the behavior-based dependency were 2 to 3 times in magnitude compared to the physical distance-based dependency. This methodology enables us to further investigate the effects of behavior-based dependency networks, and its impacts on the local economy with various hypothetical urban shocks, including not just pandemics but also natural disasters, public transport disruptions, and also positive interventions such as festivals and public events. Ongoing work investigates how behavior-based dependency network amplifies the spatial cascade of various urban shocks.

Acknowledgment

We thank Spectus for providing the mobility data used for the analysis through their Social Impact program.

References

Yabe, T., Zhang, Y., & Ukkusuri, S. V. (2020). Quantifying the economic impact of disasters on businesses using human mobility data: a Bayesian causal inference approach. *EPJ Data Science*, *9*(1), 36.

Zhai, W., & Yue, H. (2022). Economic resilience during COVID-19: An insight from permanent business closures. *Environment and Planning A: Economy and Space*, *54*(2), 219-221.

Moro, E., Calacci, D., Dong, X., & Pentland, A. (2021). Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications*, *12*(1), 1-10.

Foot Traffic Prediction Using Graph Neural Networks

Hasan Alp Boz[1], Mohsen Bahrami[2], Massimiliano Luca[3], Selim Balcisoy[1], Alex Pentland[2]

[1] Sabanci University, Istanbul, Turkey {bozhasan, balcisoy}@sabanciuniv.edu
[2] MIT Connection Science, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA {bahrami, pentland}@mit.edu
[3] Fondazione Bruno Kessler (FBK), Italy and Free University of Bolzano, Italy mluca@fbk.eu

Abstract

Estimating foot traffic potential for a new business with precision is of utmost importance, as it directly influences revenue generation capabilities. Nevertheless, this task is inherently challenging due to the intricate and ever-changing nature of human mobility patterns. The site selection process demands meticulous deliberation on factors like accessibility, demographics, and competition, ensuring the attraction of the intended customer base and the maximization of earning potential. Classical spatial methods, such as Huff model[1], use a business's attractiveness to customers (e.g., store area) and distance to customer location to model the probability of a customer visiting a particular business location. However, such methods fail to capture the socio-economic drivers behind human mobility. In this research, we investigate the use of Graph Neural Networks (GNNs) for foot traffic prediction to address this inherit complexity. To this end, we use large-scale longitudinal mobility data and frame the relationship between neighborhoods and businesses as a directed bipartite graph, converted into a dual graph [2] to predict the visit flux from neighborhoods to a particular business with the help of GNNs in a classification setting.

We use *SafeGraph Weekly Visit Patterns*[1] dataset, which provides weekly aggregated visits from Census Block Groups (CBG) to businesses also referred to as point-of-interests (POI), collected from mobile phones of users who opted to share their location data through various applications. To obtain a more balanced distribution, we aggregate the visits from CBG level to census tract level which covers a larger area and contain more residents. Then, we construct a bipartite graph between census tracts and POIs, in which edge weights store the yearly aggregated number of visits from a census tract to a POI. For each census tract, we compile a multitude of socio-economic features using 2014-2018 5-year American Community Survey (ACS) data[2], and assign them as node features in addition to POI count and the diversity of POI categories in the area. POIs are represented with their business category embeddings, geo-location, store area in square meters, and the socio-economic features of their home census tract. We assign edge labels based on the quartiles of aggregated number of visits in each business category.

GraphSAGE[3] is an inductive GNN framework, in which a node's local neighborhood is considered to generate its embeddings, instead of considering all the nodes in the graph. To predict the potential foot traffic, we use a GNN architecture that relies on GraphSAGE census tract and POI embeddings. We use two stacked GraphSAGE layers to extract the node embeddings in a heterogeneous graph setting. The resulting census tract and POI embeddings are concatenated to be fed to three linear layers with *leakyReLU* activation functions in between to obtain the final logit values for each class label. Figure 1 depicts the overall model architecture.

[1] https://docs.safegraph.com/docs/weekly-patterns

[2] https://data.census.gov/

We evaluate the proposed method in Monroe County, NY, USA (192 census tracts, 6K POIs and 384K edges) in the year 2018. To account for the sampling bias in the mobility data, we apply post-stratification on the number of visits based on the ratio of observed mobile devices in each time step and the actual population of a census tract. We obtain the train and test sets by randomly splitting the edges of the bipartite graph and train the model in 3000 epochs with Adam optimizer and calculate loss with Cross Entropy Loss. Moreover, we use the Huff model as our baseline, in which model outputs are again transformed to their corresponding quartiles in the same setting. The GNN model yields a 0.33 F1 score, while the Huff model returns 0.3.

Our preliminary result indicates that our model is suitable for analyzing the total customer flux and demographics of a new business's customer base. We hypothesize that the inductive learning of GraphSAGE layers enables our model to generalize embeddings of neighborhood and business node embeddings. As a future research direction, we are investigating new GNN architectures that predict not only the number of visitors a new business would receive but also its impact on visitor distribution for existing businesses.

References

[1] David L. Huff. A Probabilistic Analysis of Shopping Center Trade Areas. *Land Economics*, 39(1):81–90, 1963.

[2] Frank Harary and C St JA Nash-Williams. On eulerian and hamiltonian graphs and line graphs. *Canadian Mathematical Bulletin*, 8(6):701–709, 1965.

[3] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

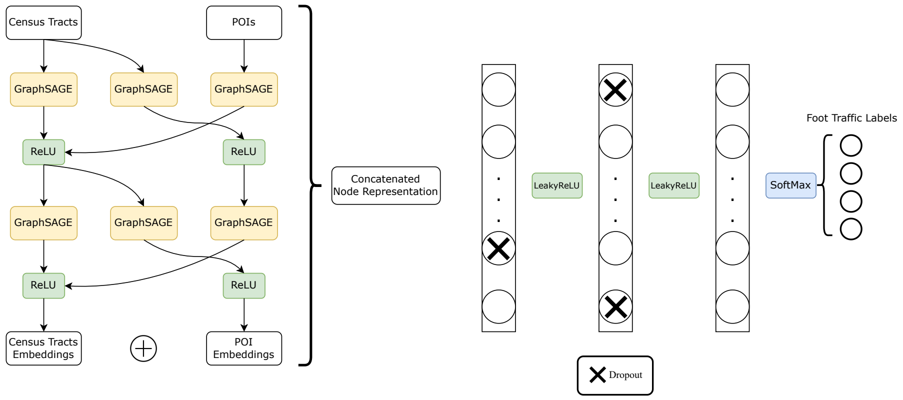


Figure 1: Overview of the graph neural network architecture used to predict the potential foot traffic. GraphSAGE layers learn the embeddings of census tract and POI nodes in the bipartite network. Resulting embeddings are concatenated and fed to stacked linear layers with dropouts in between to perform the final prediction of foot traffic labels.

Assessing the utility of mobile network operator data in geospatial models of poverty - a Ghana case study

Christopher J Brooks[†]*, James Harrison[†], Omar Seidu[×], Veronique Lefebvre[†]
[†] Flowminder Foundation, [×] Ghana Statistical Service
*corresponding author email: chris.brooks@flowminder.org

Poverty mapping is becoming an important means of informing geographic targeting of programmes by governments, NGOs and other actors involved in progressing poverty reduction across low- and middle-income countries (LMICs). Regular production of subnational estimates of poverty rates has the potential to provide decision makers with up-to-date information with which to dynamically allocate often scarce programme resources in a more optimal fashion. Traditional approaches to poverty mapping rely heavily on census data, which limits their intercensal relevance. There has been an increasing interest in combining household surveys with earth observation (EO) data, owing to its increasing quality, availability and geographic coverage, with demonstrable predictive strength across many poverty mapping studies. The additional integration of mobile network operator (MNO) data has also garnered much attention owing to the ubiquitous nature of mobile phones, with similarly positive results being reported. However, exactly how much added value proprietary MNO data has in comparison to freely available EO data on this type of modelling should be assessed, particularly given the acquisition costs for development and humanitarian programs. In recent years, there has been a strong drive by the Ghanaian Government to integrate non-traditional data sources into the production of a range of national statistics. The current study is being conducted under one such initiative; the Data for Good Partnership involving Ghana Statistical Service (GSS), Vodafone Ghana and Flowminder Foundation, which aims to leverage MNO data to support decision making across the Ghanaian Government.

We report on progress made towards integrating MNO data, in the form of Call Detail Records (CDR) and network coverage predictions, into subnational poverty maps of Ghana. Using microdata from the 2021 Ghana Population and Housing Census (PHC) and geo-referenced data from three rounds of the 2022 Ghana Annual Household Income and Expenditure Survey (AHIES), we obtain both a non-monetary Wealth Index and the Ghana Multidimensional Poverty Index (MPI). A comprehensive set of geospatial covariates were derived from publicly available EO and other spatial data sources, ranging from building footprints to vegetation indices. A candidate set of interpretable features were derived from CDR data covering the period of the census and household panel survey rounds, and together with network coverage estimates, were spatially harmonised to the geospatial covariates. Both subscriber-centric and cell tower-centric features were extracted, including assortativity measures across settlement types and urban morphologies, urban-rural mobility, commuting patterns, weekly and seasonal mobility and usage patterns, and others. The consolidated dataset was then used to address the question of whether adding CDR-derived covariates alongside existing geospatial covariates can reliably enhance estimates of non-monetary poverty at the neighbourhood level across Ghana. We report on initial results from this work-in-progress, outline the upcoming elements of the study, and provide additional reflections on future directions for the initiative.

Acknowledgements. The project was financed by the William and Flora Hewlett Foundation, and the Vodafone Foundation. This work was undertaken with cooperation from Vodafone Ghana and Ghana Statistical Service.

Mobility data reveals the trade-off between sustainability and social segregation in the 15-minute city*

Timur Abbiasov*[1], Cate Heine*[1], Sadegh Sabouri*[1], Arianna Salazar-Miranda*[1]✉,
Paolo Santi[1], Edward Glaeser*[2], and Carlo Ratti[1]

[1]Senseable City Lab, MIT
[2]Harvard University

February 2023

Abstract

Amid rising congestion and transport emissions, policymakers are embracing the "15-minute city" paradigm, which envisions neighborhoods where basic needs can be met within a short walk from home. Prior research primarily examined amenity access without examining its relationship to behavior. We introduce a measure of local trip behavior using GPS data from 40 million US mobile devices, defining *15-minute usage* as the proportion of consumption-related trips made within a 15-minute walk from home. Our findings show that the median resident makes only 14% of daily consumption trips locally. Differences in access to local amenities can explain 84% and 74% of the variation in 15-minute usage across and within urban areas respectively. Historical data from New York zoning policies suggest a causal relationship between local access and 15-minute usage. However, we find a trade-off: increased local usage correlates with higher experienced segregation for low-income residents, signaling potential socio-economic challenges in achieving local living.

Keywords: 15-minute city, mobility, sustainability, land use, walkability

*These authors contributed equally. Corresponding Author: ✉ ariana@mit.edu.

Neighborhood Detection from Mobility Data

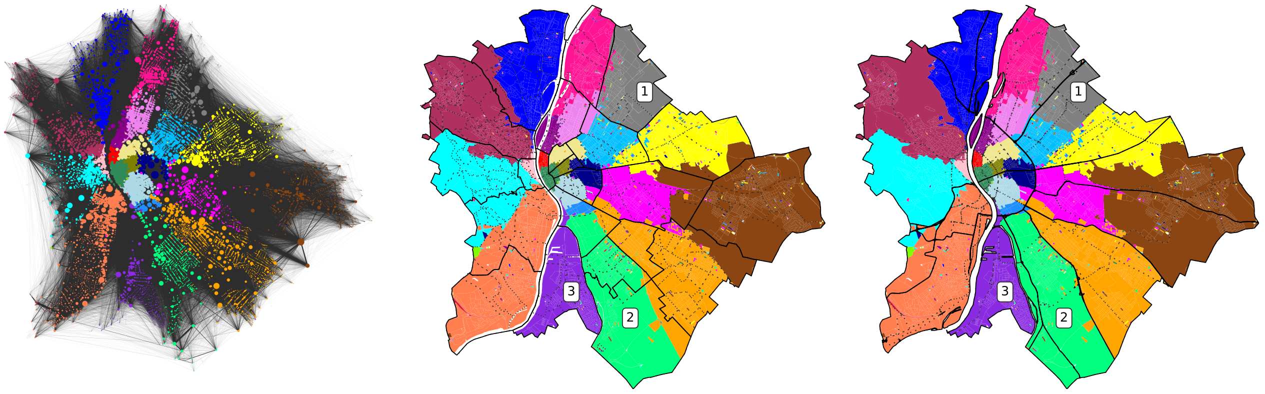
Gergő Pintér *1 and Balázs Lengyel †1,2,3

1NETI Lab, Corvinus University of Budapest, Budapest, Hungary
2ANET Lab, ELKH Center for Economic and Regional Sciences, Budapest, Hungary
3Institute for Data Analytics and Information Systems, Corvinus University of Budapest, Budapest, Hungary

Keywords: *mobile positioning data, urban mobility, community detection, barriers, polygon similarity*

A growing literature investigates the mobility of individuals to better understand social segregation in cities [1, 2, 3, 4, 5, 6, 7]. The socio-economic status of neighborhoods in urban areas are known to influence mobility patterns and separate social strata from one another [8, 9]. Yet, it is not clear how segregation through mobility is related to separation by physical and administrative barriers in cities. In this paper, we examine aggregate networks generated from individual mobility with the Louvain community detection algorithm to detect different scales of neighborhoods, analyze the role of administrative and physical barriers in shaping the neighborhoods' scales and quantify the socio-economic coherence of detected neighborhoods.

We use GPS-based mobility data between 2019 September and 2020 February in Budapest provided by a data aggregator company that collects and combines anonymous location data from smartphone applications. Using the Infostop algorithm [10], we detect stops where a user spent some time during the day. House block polygons are extracted from OpenStreetMap. Two blocks are connected by an edge if a user had consecutive stops between the given blocks within a day. Figure 1a illustrates the network for Budapest. Then, similar to recent papers that use community detection algorithms to characterize the spatial structure of mobility networks [11, 12], we apply the Louvain method to the stop-network with different resolution parameters that clusters the blocks into communities [13]. The communities were compared with administrative boundaries, the districts of Budapest, (Figure 1b), and infrastructural barriers, e.g., higher-order roads like highways (Figure 1c). The changes in the community areas are evaluated in respect of the administrative and infrastructural barriers during the change of the resolution parameter with the symmetric area difference (Figure 2c) between the clustered blocks (i.e., community) and the administrative districts.



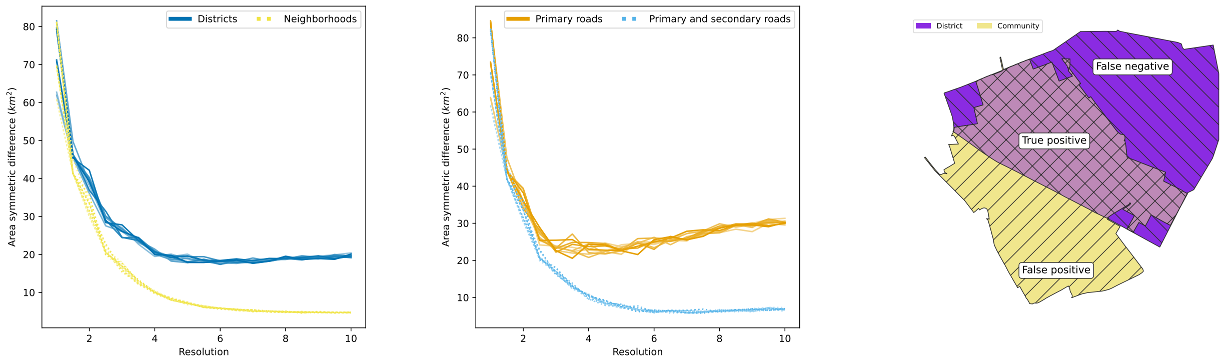
(a) (b) (c)

Figure 1: The result of the Louvain community detection, with the resolution 2.5. Figure (**a**) shows the block connectivity network, the node sizes are proportional to the degree of a node, the color represents the communities. The communities are compared to the districts of Budapest (**b**) and the areas enclosed by highways and main roads (**c**). In (**b**) and (**c**), dotted lines represent the neighborhoods and the enclosures defined by the secondary roads, respectively.

The detected communities on the mobility network tend to fit administrative and physical barriers as the resolution parameter grows (not presented). However, administrative barriers matter more in one area (Figure 1c/marker 1) but physical barriers matter more in another area (Figure 1b/marker 2). District 21 (marker 3), the district is bounded by the river Danube, which is a special case where the physical and the administrative barrier match. Note that the features of a physical barrier can affect its community-forming power as Figure 1c shows: lower order roads (displayed by dotted lines) seem to have no impact on communities at lower resolution, but higher order roads do e.g., (Figure 1c/marker 2 or 3).

*gergo.pinter@uni-corvinus.hu
†lengyel.balazs@krtk.hu

By increasing the resolution parameter of the Louvain community detection algorithm, the communities become smaller. At the same time, the community-forming power of the secondary roads and the neighborhood borders increases, shown by the significantly smaller symmetric area differences at higher resolutions (Figure 2a and 2b). Communities constructed from mobility network are bounded by administrative and infrastructural barriers as well, which seems to have a hierarchical nature in respect of the higher and lower order barriers.



(a) **(b)** **(c)**

Figure 2: The Louvain communities with different resolution parameters are compared to the administrative (**a**) and infrastructural (**b**) barriers of Budapest using symmetric area difference (**c**). In other words, the summarized area of the false negative and the false positive parts. The Louvain community detection algorithm was executed ten times with each resolution and plotted with different shades of the same color.

References

[1] Federico Botta and Charo I Del Genio. "Analysis of the communities of an urban mobile phone network". In: *PloS one* 12.3 (2017), e0174198.

[2] Elizabeth Roberto and Elizabeth Korver-Glenn. "The Spatial Structure and Local Experience of Residential Segregation". In: *Spatial Demography* 9 (2021), pp. 277–307.

[3] Esteban Moro et al. "Mobility patterns are associated with experienced income segregation in large US cities". In: *Nature communications* 12.1 (2021), p. 4633.

[4] Susan Athey et al. "Estimating experienced racial segregation in US cities using large-scale GPS data". In: *Proceedings of the National Academy of Sciences* 118.46 (2021), e2026160118.

[5] Gergő Tóth et al. "Inequality is rising where social network segregation interacts with urban topology". In: *Nature communications* 12.1 (2021), p. 1143.

[6] Zhuangyuan Fan et al. "Diversity beyond density: Experienced social mixing of urban streets". In: *PNAS nexus* 2.4 (2023), pgad077.

[7] Eszter Bokányi et al. "Universal patterns of long-distance commuting and social assortativity in cities". In: *Scientific reports* 11.1 (2021), p. 20829.

[8] Jennifer Candipan et al. "From residence to movement: The nature of racial segregation in everyday urban mobility". In: *Urban Studies* 58.15 (2021), pp. 3095–3117.

[9] Qi Wang et al. "Urban mobility and neighborhood isolation in America's 50 largest cities". In: *Proceedings of the National Academy of Sciences* 115.30 (2018), pp. 7735–7740.

[10] Ulf Aslak and Laura Alessandretti. "Infostop: scalable stop-location detection in multi-user mobility data". In: *arXiv preprint arXiv:2003.14370* (2020).

[11] Meihan Jin et al. "Identifying borders of activity spaces and quantifying border effects on intra-urban travel through spatial interaction network". In: *Computers, Environment and Urban Systems* 87 (2021), p. 101625.

[12] Mehmet Yildirimoglu and Jiwon Kim. "Identification of communities in urban mobility networks using multi-layer graphs of network traffic". In: *Transportation Research Part C: Emerging Technologies* 89 (2018), pp. 254–267.

[13] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

A Comprehensive Study of Mobile Service Demands at Indoor Cellular Networks

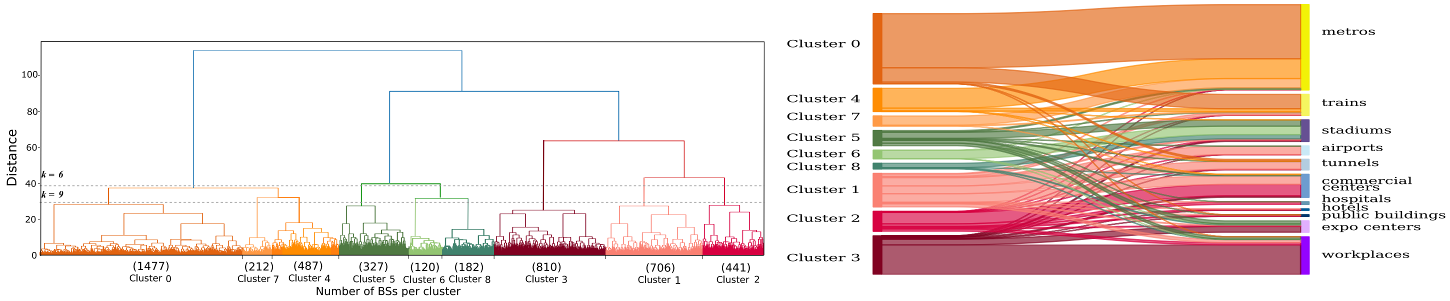
Stefanos Bakirtzis[*,5,6], André Felipe Zanella[1,3], Stefania Rubrichi[2], Cezary Ziemlicki[2], Zbigniew Smoreda[2], Ian Wassell[5], Jie Zhang[6,4], and Marco Fiore [1]

[1] IMDEA Networks, Madrid, Spain
[2] Orange Innovation - SENSE, Paris, France
[3] Universidad Carlos III de Madrid, Madrid, Spain
[4] University of Sheffield, Sheffield, United Kingdom
[5] University of Cambridge, Cambridge, United Kingdom
[6] Ranplan Wireless Network Design Ltd, Cambridge, United Kingdom

Understanding mobile network usage is an important task with manifold implications, in networking and beyond. The exploration, characterization, and modeling of mobile network traffic assumes a key role in developing and supporting an efficient wireless ecosystem, but at the same time, the traffic generated by mobile services has become an important source of insights on human activities, needs, and habits [1], and thus it has benefitted research in diverse scientific disciplines, ranging from socioeconomic to climate change [2], [3]. Interestingly, all existing studies on characterizing and exploiting mobile traffic invariably focus on measurements collected in outdoor environments. This can be ascribed to the fact that cellular networks have been primarily intended as a technology for mobile, outdoor users. Yet, this is not true anymore, and main actors in the telco ecosystem forecast that about 80% of the future cellular data traffic will be generated in indoor environments [4]. As a result, fifth-generation (5G) and beyond (B5G) systems are anticipated to align with the emerging need to serve indoor users [5]: specifically, a number of major vendors and mobile network operators (MNOs) expect 5G/B5G systems to transition from a legacy "outside-in" coverage approach, where indoor coverage is provided by antennas located outdoor, to the deployment of native indoor cellular networks (ICNs) [4]. These ICN deployments will allow improving substantially the quality of service for indoor user equipment (UE), and bring real competition to Wi-Fi technologies for the increasingly remunerative indoor market.

The emergence of pervasive ICNs renders it necessary to comprehend how such networks will be used. In this context, a considerable volume of research has focused on methods related to the modeling of radio propagation in indoor environments, the impact of the building layout on indoor network performance, and the efficient planning of ICNs. However, the dynamics and distinguishing features of the traffic data generated by in-building radio access network components have not been studied yet. Unlike outdoor base stations (BSs) that tend to observe a general-purpose use, i.e., serve concurrently numerous subscribers engaged with diverse activities during different daily life phases, ICNs are expected to target more specific use cases. For instance, ICNs are deployed in underground subway and train stations to compensate for the limited coverage of the outdoor wireless network. Likewise, corporate offices are equipped with indoor BSs to provide enhanced and reliable communications to support the work of their employees, whilst their installation is necessary in exposition centers and stadiums in order to accommodate the concentrated high-traffic demands intertwined with social events. Therefore, the ICN traffic is reasonably influenced by the context in which it is generated, which highly depends on the indoor environment type and, by extension, on the kind of activities in which users are involved in it. Eventually, the traffic dynamics of ICNs are expected to differ significantly from those of legacy communication system outdoor BSs.

[*] The work has been partially carried out while Stefanos Bakirtzis was at IMDEA Networks Institute and has been supported by equipment and resources of the institute.



(a) Dendrogram illustrating the iterative merging of antennas into clusters.

(b) Sankey diagram showing the clusters flowing into different environment types

Figure 1: Example of ICN cluster and their correlation with the indoor environment type.

To shed light on the aforementioned gaps in the existing literature, our work hinges on a countrywide ICN Internet measurement traffic data set to provide a comprehensive study of ICN mobile service demands. To this end, we define an appropriate transformation of the traffic data that enables probing the range of different Internet mobile service utilization profiles at indoor antennas. Then, employing an unsupervised learning approach, we designate that distinct service utilization clusters are inherent in indoor communication systems, as shown in Fig. 1a. This rich and diverse behavior of Internet services has not been unveiled before, and as we also demonstrate, it does not align with that of outdoor legacy communication systems. To interpret the clustering results and delve into the essence of the different clusters, we leverage techniques from the field of explainable machine learning (ML). This enables the identification of the most important features for each cluster, and consequently allows us to expound on the most important as well as on under-utilized Internet service types of each cluster. We expose that there is a strong connection between the clusters individuated by our analysis and the indoor environment type. In particular, we show that the same Internet applications manifest very heterogeneous behaviors, even for antennas in proximity, due to the determining influence of the environment type on user activities, as shown in Fig. 1b. This phenomenon is rooted in the ICN environment and has not been highlighted before. Finally, we reveal that in the various clusters the total Internet traffic data, as well as the traffic generated by the individual applications, exhibit different activity peaks and temporal patterns. Overall, our work paves the road to the characterization of service-level mobile traffic demands in indoor environments and offers a number of insights into patterns that were not observed or demonstrated quantitatively before.

References

[1] M. Ghahramani, M. Zhou, and G. Wang, “Urban sensing based on mobile phone data: Approaches, applications, and challenges,” *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 3, pp. 627–637, 2020.

[2] A. Llorente, M. Garcia-Herranz, M. Cebrian, and E. Moro, “Social media fingerprints of unemployment,” *PloS one*, vol. 10, no. 5, e0128692, 2015.

[3] S. Dujardin, D. Jacques, J. Steele, and C. Linard, “Mobile phone data for urban climate change adaptation: Reviewing applications, opportunities and key challenges,” *Sustainability*, vol. 12, no. 4, p. 1501, 2020.

[4] Cisco, “White paper: Cisco vision: 5G-thriving indoors,” Cisco, Tech. Rep., 2017. [Online]. Available: `https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/5g-ran-indoor.pdf`.

[5] M. Agiwal, A. Roy, and N. Saxena, “Next generation 5G wireless networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 2016.

A methodology for studying social contact: application to the analysis of socio-economic segregation in the Madrid region

Oriol García-Llopis - Nommon Solutions and Technologies - oriol.garcia@nommon.es
Raquel Sánchez-Cauce - Nommon Solutions and Technologies - raquel.sanchez@nommon.es
Miguel Picornell Tronch - Nommon Solutions and Technologies - miguel.picornell@nommon.es
Oliva Cantú Ros - Nommon Solutions and Technologies - oliva.garcia-cantu@nommon.es
Ricardo Herranz - Nommon Solutions and Technologies - ricardo.herranz@nommon.es

In recent years, measuring social contact has gained relevance in many fields. Relevant examples are epidemiology, where social contact is a key driver of disease contagion, urban planning (e.g., to study segregation), and politics, where it is used to study the diffusion of political opinions and trends. The widespread use of mobile devices allows a highly detailed analysis of the activities and mobility patterns of the population throughout the day, enabling a better understanding of the interactions between individuals. Moreover, the combination of these data with other data sources, such as census and income data, can be used to perform segregation analysis based on different sociodemographic characteristics of the population.

This contribution presents a methodology to compute a social contact indicator that characterises social interactions among different population groups. This methodology is based on Nommon's solution for obtaining activity patterns of individuals from mobile network data (Population Insights)[1]. This solution generates ‘activity diaries’ for the sampled mobile phone users and expands them to the total population using census data[2], providing, among others, the location of each activity and its duration. This information is used to characterise a zone based on the profile of the individuals present in that zone at different times of the day. From this characterisation of the zones, individuals are characterised based on the zones they have visited.

We use this approach to analyse socioeconomic segregation based on net income in different areas of the Region of Madrid. Social contact indicators are calculated both at zone and individual level. The data cover one week of March 2017. We analyse an average working day (average of the social contact indicators for Tuesday, Wednesday, and Thursday) and an average weekend day (average of the indicators for Saturday and Sunday).

The results of the study revealed differences in the spatial distribution of the social contact indicator at a zone level both within day and when comparing working and weekend days (see Figure 1). We can observe that in working days high values of social contact are limited to few areas located mostly in the south-central area of the region. While on weekends, the areas with high values of social contact increase and are distributed more or less homogeneously with a prominence in the northern part.

[1] https://www.nommon.es/products/population-insights/

[2] For a detailed explanation of the methodology for obtaining activity-travel diaries from the sample of mobile phone users and an expansion to the total population see Bassolas, A., Ramasco, J. J., Herranz, R., & Cantú-Ros, O. G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transportation Research Part A: Policy and Practice, 121, 56–74.

However, there are zones that maintain the levels of segregation over the whole week. For instance, some zones in the south-west of Madrid which are highly segregated, and some areas in the north-west are highly integrated over the week. In contrast, we can observe marked differences between working days and weekends in the north west area where segregation levels decrease on the weekend. This can be explained by the fact these zones are a frequent place for weekend excursionists. The opposite pattern occurs in business areas with little touristic interests. For instance, the zones in the north-centre area of the region show a slight increase of segregation during the weekend.

Like in the case of the zonal indicator, when calculated at an individual level, the social contact indicator also presents interday variations. Moreover, in the majority of zones, the average individual social contact of the residents of a given area is highly correlated with the zonal social contact indicator of it. We have also noticed that this does not hold for tourist zones which present a higher zonal social contact indicator than the average individual social contact indicator of the residents.

Finally, we have analysed the time distribution of the indicator. This seems to follow a sinusoidal sequence over time, showing a unique peak per day at midday (which may mean that the maximum level of integration is reached when people are at work or lunch time) in the working days. However, on Fridays or some weekend days sometimes it exhibits a second peak in the evening, which might be produced by people who have dinner or go out.

According to the results obtained, we can conclude that mobile network data have a great potential for the detailed analysis of segregation dynamics in the cities.

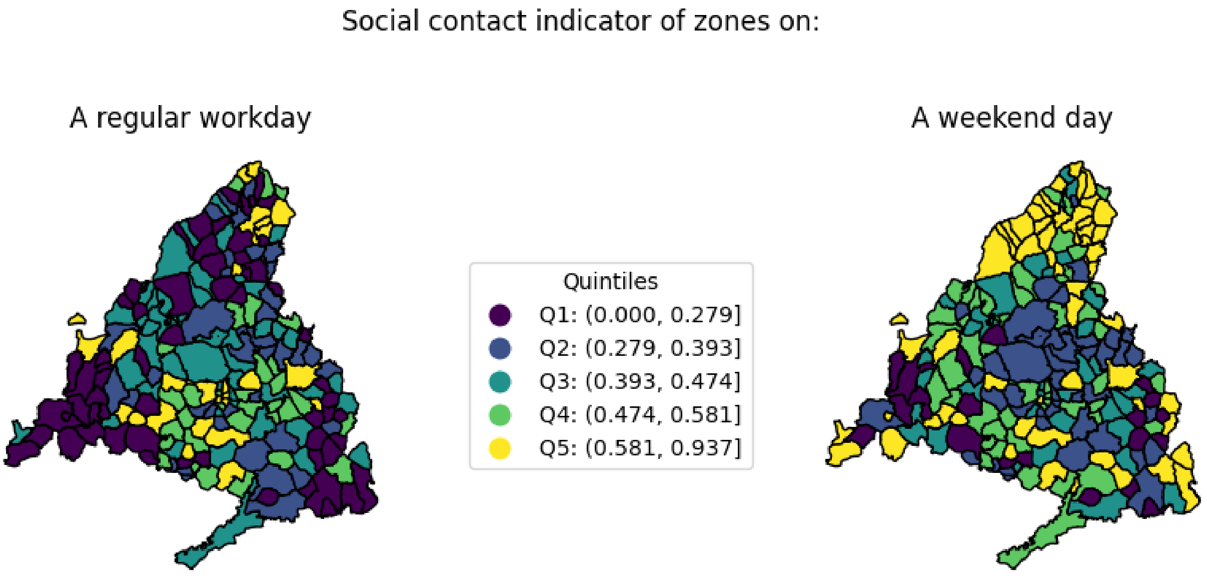


Figure 1. Zonal social contact indicator levels in the Community of Madrid on a regular working day and on a regular weekend day on the second of March 2017. We calculate the quintiles of the social contact indicator, the greater the indicator is, the less segregated the zone is. Therefore, the yellowish zones are the ones less segregated and the blueish are the ones more segregated.

Time-dinamic of income segregation at neighborhoods scale

Lavinia Rossi Mori[1,2,3], Vittorio Loreto[1,3,4,5], and Riccardo Di Clemente[6,7,*]

[1]Centro Ricerche Enrico Fermi, Via Panisperna 89/A, 00184, Rome, Italy
[2]Physics Department, Unversità di Roma Tor Vergata, 00133, Rome, Italy
[3]Sony Computer Science Laboratories Rome, Joint Initiative CREF-Sony, Centro Ricerche Enrico Fermi, Via Panisperna 89/A, 00184, Rome, Italy
[4]Sony Computer Science Laboratories Paris, 6, Rue Amyot, 75005, Paris, France
[5]Physics Department, ”La Sapienza” Unversità di Roma, Piazzale Aldo Moro 2, 00185, Rome, Italy
[6]Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.
[7]The Alan Turing Institute, London, NW12DB, United Kingdom
[*]corresponding author: riccardo.diclemente@nulondon.ac.uk

Income segregation, the uneven interaction of communities within an urban environment due to income disparities, poses a significant challenge to economic growth and social stability by exacerbating poverty and crime rates (Kawachi, Am J Public Health, 1997, DOI:10.2105/ajph.87.9.1491). Furthermore, it impacts access to opportunities and services (Tammaru, IZA Discussion Paper, 2017, DOI:docs.iza.org/dp10980), leading to disparities in commute duration and insufficient social mixing, which reduce the quality of life for lower-income individuals and limit productive interactions across income groups. Segregation, urban topology, and human mobility are interwoven, influencing daily routines, activities, and interactions within urban spaces (Tóth, Nat. Comm., 2021, DOI:10.1038/s41467-021-21465-0). People’s whereabouts depend on necessities: we commute for work, shop grocery, and dine at some restaurant according to our lifestyle (Di Clemente, Nat. Comm., 2018, DOI:10.1038/s41467-018-05690-8). Key factors that guide these movements are the Points Of Interest (POIs). The interaction with these POIs varies based on time, location, with individual needs or preferences, which consequently impacts our social activity and the level of our experienced segregation (Moro, Nat. Comm., 2021, DOI:10.1038/s41467-021-24899-8). As individuals traverse various neighborhoods as part of their daily activities, spatial segregation and social mixing dynamics change from morning to evening. Some neighborhoods provide thought the day a consistently diverse population across income groups, while others might exhibit fluctuations in social mixing, perhaps providing an inclusive atmosphere during the afternoon, only to become highly segregated at night.

The challenge is to dynamically observe these changes and understand the topological characteristics that can promote more inclusive neighborhoods. Could these be linked to factors like accessibility or the diversity and uniqueness of activities that a neighborhood offers? To answer these questions, we propose an in-depth analysis of the dynamics of social mixing in space and time in the city of Milan over an eight-month period. We introduce space-time metrics to characterize segregation. Furthermore, we present Attractivity, Liveability, and Accessibility (ALA metrics) leading to a clustering of neighborhoods that provides an indication of their service quality, with no information on who lives or visits these places yet. Accessibility is represented by the ease of getting to a given place with public transport (Biazzo, R. Soc. Open Sci., 2019, DOI:10.1098/rsos.190979). Liveability is measured through three quantities: number of supermarkets, their diversity, and the number of schools. For attractivity we define: the Fitness (Tacchella, Sci. rep., 2012, DOI:10.1038/srep00723) of a neighborhood, that represents both the diversity and uniqueness of POIs categories, the diversity of categories and prices, the median of prices, and reviews. We integrate a mobility dimension into our analysis to capture the effects of social mixing, focusing on the income percentage in a given neighborhood at a given time. This dimension leverages trajectory Location-Based Services (LBS) data from approximately 100,000 users with 24 million of pings, and a geolocated dataset of rent per square meter, serving as a proxy for user income. Our data and income assumptions are validated with census data, obtaining Pearson correlation values greater than 0.8.

Our approach allows us to observe changes in social interaction over time and in relation to ALA clusters. We find a high level of spatial segregation at night due to residential segregation but observe increased social mixing during the day when residential segregation relaxes. Neighborhoods exhibiting interaction with middle-income groups appear more inclusive, as individuals from these groups commonly attend places frequented by both low- and high-income groups. Our study indicates that segregation remains present depending on the ALA cluster, with the type of facilities offered by a neighborhood and their usage at different times by different income groups influencing segregation levels. We define the temporal mixing of a neighborhood as the temporal pattern of a function of the Gini coefficient. We identify three clusters – inclusive, mixed, and segregated – and study the similarity between these clusters and those determined by the ALA metrics. Through regression analysis, we identify key neighborhood features driving inclusivity, finding that the most influential features in making a neighborhood inclusive are velocity score, fitness, price diversity, and median price.

In conclusion, our study offers a novel perspective on time dynamics of urban income segregation by city areas. We connect the social mixing by income with clear topological metrics such as POIs diversity, public transport accessibility, and POIs prices. The identification of three distinct neighborhood clusters based on the ALA metrics, the temporal shifts in their composition, and the varying degrees of inclusivity they exhibit, can provide critical insights for urban planning and development. By acknowledging and addressing these dynamic aspects of segregation, cities can work towards creating more inclusive and diverse neighborhoods, enhancing overall urban accessibility, and providing a more equitable distribution of services and opportunities.

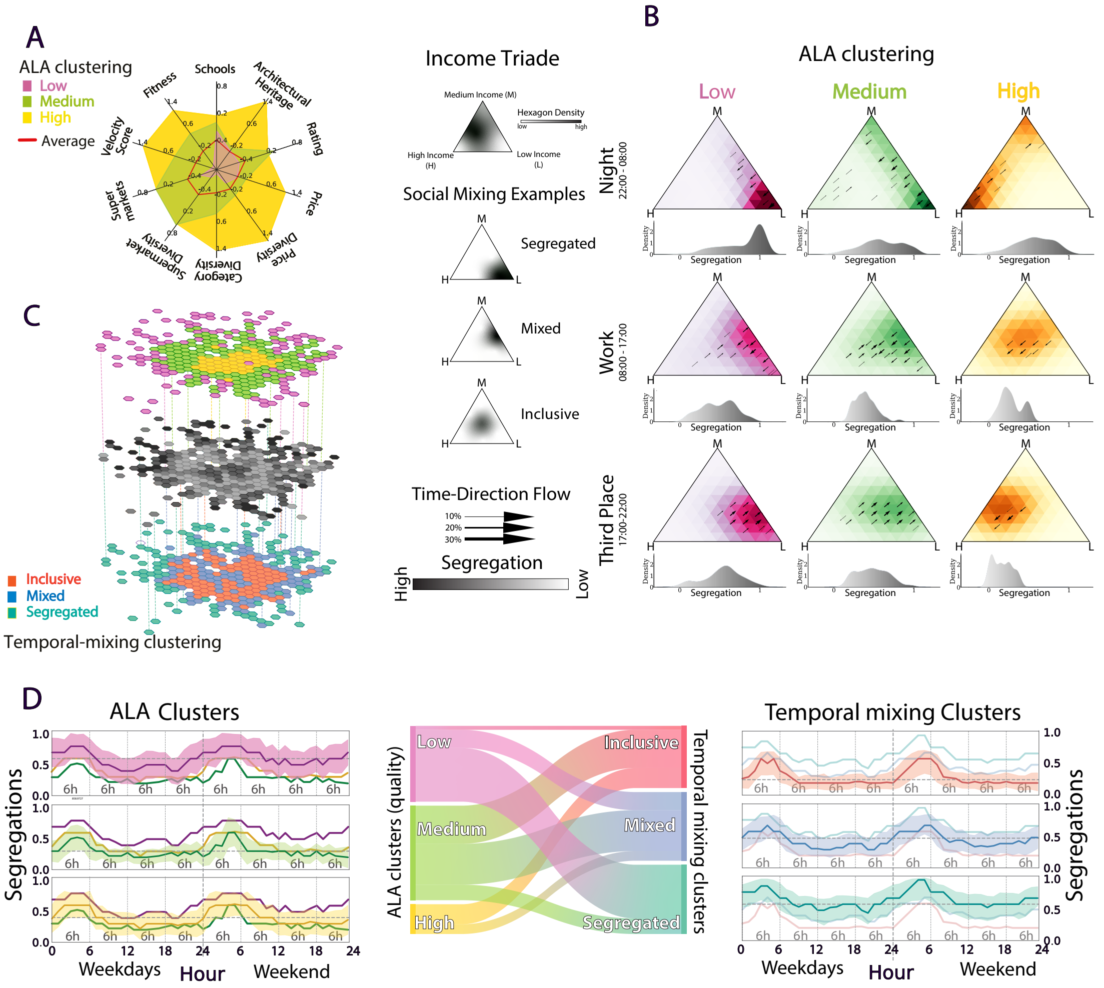


Figure 1. (**A**) Median of the ALA metrics within each cluster and all neighborhoods in red. (**B**) Movements of people in ALA clusters: at the vertices of the triangles there are incomes, the more a neighborhood is visited by an income the closer it is to one of the vertices; the more the neighborhood is well distributed among the incomes the closer it is to the center of the triangle. Each column indicates the cluster to which the neighborhood belongs, and each row indicates the time of day being observed. (**C**) First level: location of ALA clusters, second level: location of leisure segregation, third level: location of segregation profile clusters. (**D**) Segregation profile for ALA clusters, Sankey diagram for the two different clusterization and last segregation profile divided in three cluster -inclusive, mixed and segregated.

Milan's Diverse Mosaic: Urbanization and Foreign Relations Shaping Ethnic Plurality

İrem Betül Koçak*, and Oğuz Yücel*
kocaki@itu.edu.tr, yucelog@itu.edu.tr
* Management Engineering Department, Istanbul Technical University, 34469 Maçka, Istanbul, Turkey.

I. Introduction

Cities are built in areas with surplus resources, attracting people from rural areas to move and settle there. Consequently, these densely populated areas foster a culture of coexisting as strangers with different backgrounds, ethnicity and cultures. In this study, we utilize Call Detail Records data from the Telecom Italia Dataset[1], which includes the total number of calls, SMS, and Internet calls originating from square IDs within the Milano grid map. We examine the country codes present in the data to determine which countries are most frequently communicated with using these various telecommunication methods. Our aim is to examine how the historical roots, style of urbanization, and the relationship with foreigners impact the cultural and ethnic fabric of Milan, the financial capital of Italy. We also seek to understand the extent of ethnical diversity in the city.

II. Methodology

There are different metrics to measure the diversity in an area and entropy index is one of them. The entropy index is commonly referred to as the Shannon index because it is related to information theory [1]. It reaches its maximum, log N, when all subgroups have the same proportion [2] as in the equation:

$$H = -\sum_{i=1}^{N} p_i * log(p_i) \tag{1}$$

In this study, the entropy index, total call, and total SMS amounts were used to calculate and analyze the diversity within cells in detail. Three distinct locations in the city were selected to examine how entropy changes throughout the day and to assess the factors that may lead to diversity in these areas.

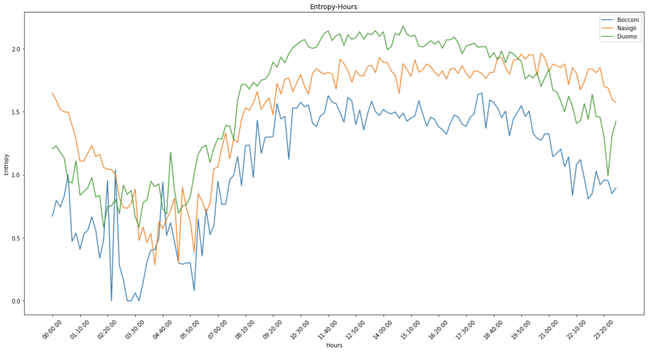


Fig. 1. Entropy changes in three different areas.

As mentioned earlier, we highlighted Milan as one of the prominent cities in Europe in terms of education, tourism, and entertainment. This study examines the entropy values of three distinct areas in Milan: Bocconi University, Navigli, and Duomo Square. Bocconi University shows lower diversity during nighttime but increases in the morning. Duomo Square consistently exhibits high entropy values due to its tourist attraction, while Navigli maintains high entropy values as a vibrant nightlife hub. Duomo Square generally has the highest daily entropy, while Bocconi University has the lowest. This analysis reveals the temporal variations and factors influencing diversity in these areas, providing insights into Milan's education, tourism, and entertainment dynamics.

III. Segregation Analysis

We create maps of dominant countries, referring to the most frequently communicated countries, in these bits of Milan grid for two different uses of the city: one for private use during the hours when people are usually sleeping or at home, and the other for public use when people are primarily working or socializing, and tourists are typically exploring the city. We examine the dominance map for these two different uses in Figure 2 and Figure 3 which represents each country with different colors.

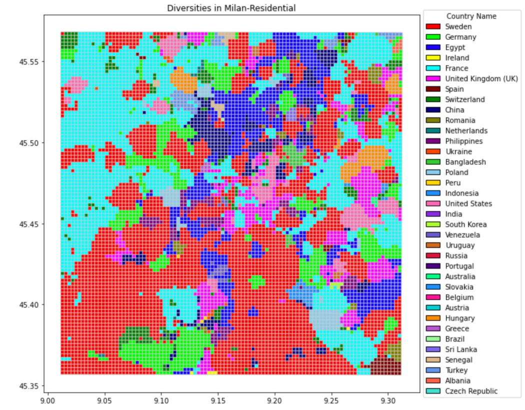


Fig. 2. Milan map with the dominant countries on residential use hours.

We assume that there are four important reasons for communication with a country can dominate a region in the city: People from these countries can emigrate to Italy and communicate with their home countries, people from Milan can migrate to these countries and communicate with their relatives from Milan, tourists can communicate with their home countries, and people communicate internationally for their work. We will consider these four main reasons and leave out the possible other reasons when we talk about dominant countries. Also, we need to consider the communication reason for work less in the residential use hours.

Before examining the dominant countries in particular regions and their relationship with Italy, it is important to discuss the city structure and urbanization of Milan. According to Karaulan [3], the urbanized portion of Milan is located in the middle of the grid, while the peripheral areas are primarily used for industry and

[1] Details can be found at http://dx.doi.org/10.7910/DVN/UTLAHU

agriculture. Consequently, the peripheral areas cannot be classified as highly urbanized. At first glance, Figure 2 and Figure 3 suggests that the central areas of Milan are more diverse compared to the peripheral area of the city, and this observation supports the acclaimed relationship with urbanization and diversity.

We can also see that there are more countries that could become dominant in their regions in the times of use of residential in certain areas. We can speculate that people tend to group in their housing preferences, so people related to certain countries tend to cluster together and segregate with others more in their residence choice as Schelling claimed [4].

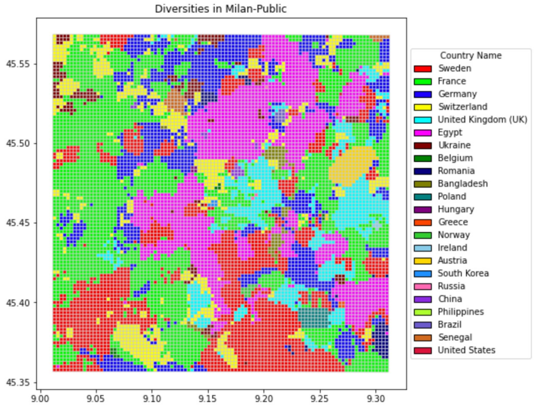


Fig. 3. Milan map with the dominant countries on public use hours.

We observe from both figures that the countries of the European Union (EU) occupy the top five positions. Italy stands out as one of the countries with the highest number of tourist arrivals worldwide, ranking as the third largest among European countries in 2013, according to the UNWTO. Milan, renowned for its historical and tourist attractions, emerges as one of the most alluring cities in Italy. So, there exist numerous factors that attract European visitors to Milan. Notably, one plausible explanation is the significant Italian diaspora in these dominant countries, as reported by the EUROSTAT [5]. Additionally, the facilitated regulations and extensive inter-connectivity between EU nations and Italy further support the presence of these countries in the top five rankings.

In Figures 2 and 3, Egyptian dominance in the urban area is evident. This is a result of Egyptians migrating to European countries like Italy and France in search of better opportunities due to economic challenges and limited job prospects in Egypt. Milan has become a significant destination for Egyptian migrants, who have secured employment in diverse establishments.

Another important aspect is to identify which countries tend to have close coexistence. Both figures demonstrate that regions with similar communication patterns are generally located in close proximity to each other. Areas that engage in communication with EU member states also exhibit spatial proximity. We also see that France often share borders with other countries where French is also spoken. This suggests that proximity practices are established not only through national identity and economic agreements but also through shared language.

In general, these interpretations based on dominance can be further expanded and detailed through second or third majority/minority mapping, aiding our understanding of Milan's ethnic and cultural life and how this situation of urban dominance translates into daily life and urban discrimination.

IV. Attraction Point Analysis

In the second phase of the study, we will examine whether the diversity in Milan contributes to the attractiveness of certain locations within the city. The Tripadvisor dataset from November 2013 was utilized to determine the attractiveness of cells. User opinions were collected to label certain points as attractive. Four different supervised learning methods were employed to predict whether regions are attractive points or not, by using variables from the CDR dataset and entropy. The data was analyzed to generate predictions for the respective regions.

Method	Accuracy	Precision	Recall	F1-Score
Decision Tree (DT)	0.891	0.564	0.373	0.449
Artificial Neural network (ANN)	0.909	0.674	0.492	0.569
Logistic Regression (LR)	0.911	0.852	0.365	0.511
Linear Discriminant Analysis (LDA)	0.911	0.744	0.460	0.569

TABLE I
Classification metrics

The table summarizes the performance metrics of four supervised learning methods: Decision Tree (DT) with 3 maximum-depth, Artificial Neural Network (ANN), Logistic Regression (LR), and Linear Discriminant Analysis (LDA). The accuracy values for all methods are relatively high, ranging from 0.891 to 0.911, indicating overall successful predictions. Precision scores vary across the methods, with LR and ANN demonstrating higher precision values compared to DT and LDA. The recall and F1-scores also exhibit variations among the methods, indicating differences in their abilities to correctly identify positive instances. Overall, ANN and LR show more balanced performance in terms of precision, recall, and F1-score, suggesting their effectiveness in predicting the attractiveness of regions using the given variables and entropy.

V. Conclusion

This study utilizes mobile phone data to gain insights into the nationalities of migrants, tourists, and diaspora in Milan. It highlights the influence of both historical and cultural factors in explaining this phenomenon. Unlike previous studies focusing on diversity density, this research explores the significance of diversity and segregation created by specific countries in Milan. By employing four different supervised learning methods, the study determines the presence of attractive points in different locations. The findings provide valuable patterns and dynamics within Milan's diverse population, offering a broader understanding of diversity in the city.

References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] J. Iceland, "The multigroup entropy index (also known as theil's h or the information theory index)," *US Census Bureau. Retrieved July*, vol. 31, no. 2006, p. 60, 2004.

[3] D. S. Karaulan and A. S. Kubat, "Analyzing fringe belt phenomenon in the historico–geographical structure of milan, italy," *ICONARP International Journal of Architecture and Planning*, 2018.

[4] T. C. Schelling, "Dynamic models of segregation," *Journal of mathematical sociology*, vol. 1, no. 2, pp. 143–186, 1971.

[5] S. Marino, V. Corbanese, and G. Rosas, "Inclusion of migrants in the labour market a comparative analysis of italy and other eu countries."

Spatiotemporal gender differences in urban vibrancy

Thomas Ryan Collins*,[1], Riccardo Di Clemente[1,2,3], Mario Gutiérrez-Roig[4], and Federico Botta[1,2]
[1] *Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United Kingdom.*
[2] *The Alan Turing Institute, London, NW1 2DB, United Kingdom.*
[3] *Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.*
[4] *Department of Mathematical Sciences, University of Essex, Colchester, CO4 3SQ, United Kingdom.*
*trc207@exeter.ac.uk

Keywords (1) urban vibrancy, (2) urban gender segregation, (3) mobile phone data (4) spatial data science.

Extended Abstract

Urban vibrancy is the dynamic activity of humans in urban locations. Urban vibrancy can vary with urban features and the opportunities for human interactions [1], but might also differ according to the social conditions. Different demographic groups exhibit heterogeneity in preferences, accessibility, and large-scale mobility behaviours. These variations can contribute to the spatial separation of genders, as individuals with different preferences or mobility patterns may choose to reside in different spatial locations, leading to the potential for gender segregation [2]. Because of a lack of awareness, segregation and inequalities have been hidden components of cities historically [3]. Traditional studies have failed to capture these characteristics; the link between urban vibrancy and urban features, or how this might differ for different genders, is not fully understood. We asked how urban features might contribute to a vibrant environment and how they might vary across social groups, especially concerning gender.

Our results show that (1) there are differences between males and females in terms of urban vibrancy activity, (2) the differences relate to ‘Points of Interest’ both generally and with a social focus, and (3) that there are both positive and negative ‘spatial spillovers’ existing across each city.

For this, we use Call Detail Record (CDR) data in a quantitative approach–taking advantage of the near-ubiquitous use of mobile phones [4, 5]–to gain high-frequency observations of spatial behaviours across the seven most prominent cities of Italy (Milan only shown here for clarity; Figure 1, A). Taken from the largest telecommunications services provider operating in Italy [6], we use two months’ CDR data, divided by gender, that describes a value proportional to the network user activity. This value, therefore, indicates the presence of male and female network users across the cities. Before analysis, we validate our CDR data by assessing the correlation between the nighttime CDR data and census tracts interpolated to the CDR grid. We use Moran’s I analyses to detect spatial clustering in male-female activity differences [7] and construct and compare spatial lag and spatial error models [8]. Our spatial lag models return with a relatively high pseudo-R-squared used for goodness of fit (pseudo-R2 = 0.72). We use models a spatial model comparison approach of the direct and spillover effects from urban features on male-female activity differences.

Our areal interpolation highlights a match between nighttime CDR and census patterns where, across all cities, results were positive and significant at the 5% level, simultaneously confirming the utilization of CDR data methodology and providing supporting evidence that demonstrates how CDR data reveals the locations of individuals throughout the day. Our Moran’s I analyses highlight spatial clustering of male-female activity differences (Moran’s $I = 0.38$; $p < 0.0001$; $n = 794$ grid cells).

We suggest these differences might relate to (1) the difference between men and women in their socioeconomic backgrounds; (2) that social norms related to public space use where male and female behaviours are divergent in some way; or (3) the divergent gender roles that men and women have relating to work and opportunities.

This research suggests that variations in preferences and mobility patterns among different demographic groups, including genders, may contribute to spatial separation: gender segregation can occur within cities due to choices to reside in different areas based on their preferences and behaviours. We find evidence to suggest that urban features and social conditions influence urban vibrancy. Understanding and promoting urban vibrancy can contribute to creating more lively and engaging urban environments. Positive and negative spatial spillovers existed across each city; this implies that certain urban features or activities may have effects that extend beyond their immediate locations. Finally, segregation and inequalities have historically been hidden components of cities. Addressing and acknowledging these hidden components is important to create more inclusive and equitable urban environments.

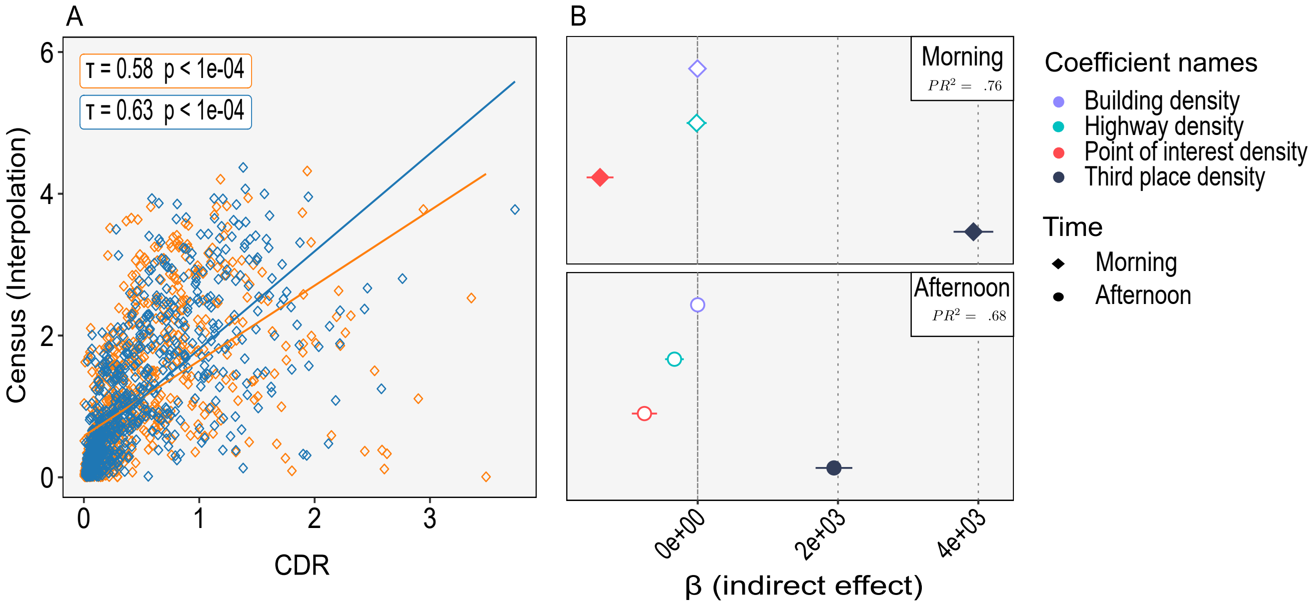


Figure 1: Milano male-female differences in urban vibrancy. (A) Correlation between Call Detail Record (CDR; 20:00-0800) data and census data. Orange markers are females whereas blue markers are males. (B) The relationship between density in features and male-female differences. Filled shapes indicates a significant result at the 5% level.

[1] Federico Botta and Mario Gutiérrez-Roig. "Modelling Urban Vibrancy with Mobile Phone and OpenStreetMap Data". In: *PLoS ONE* 16 (6 June 2021), pp. 1–19. pmid: 34077441.

[2] André De Palma and Yorgos Y. Papageorgiou. "Heterogeneity in States and Urban Structure". In: *Regional Science and Urban Economics* 18.1 (Feb. 1988), pp. 37–56.

[3] Robert M. Blackburn, Jennifer Jarman, and Bradley Brooks. "The Puzzle of Gender Segregation and Inequality: A Cross-National Analysis". In: *European Sociological Review* 16.2 (June 1, 2000), pp. 119–135.

[4] David Lazer et al. "Life in the Network: The Coming Age of Computational Social Science". In: *Science (New York, N.Y.)* 323.5915 (Feb. 6, 2009), pp. 721–723. pmid: 19197046.

[5] Tobias Preis et al. "Quantifying the Advantage of Looking Forward". In: *Scientific Reports* 2.1 (2012), pp. 1–2.

[6] GruppoTIM. *Dataset Source: TIM Big Data Challenge, https://www.gruppotim.it/en.html.* 2015. URL: http://localhost:4000/cases/telecom-italias-big-data-challenge.html (visited on 03/14/2023).

[7] Arthur Getis and J. K. Ord. "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24.3 (1992), pp. 189–206.

[8] James P. Lesage and Manfred M. Fischer. "Spatial Growth Regressions: Model Specification, Estimation and Interpretation". In: *Spatial Economic Analysis* 3.3 (Nov. 1, 2008), pp. 275–304.

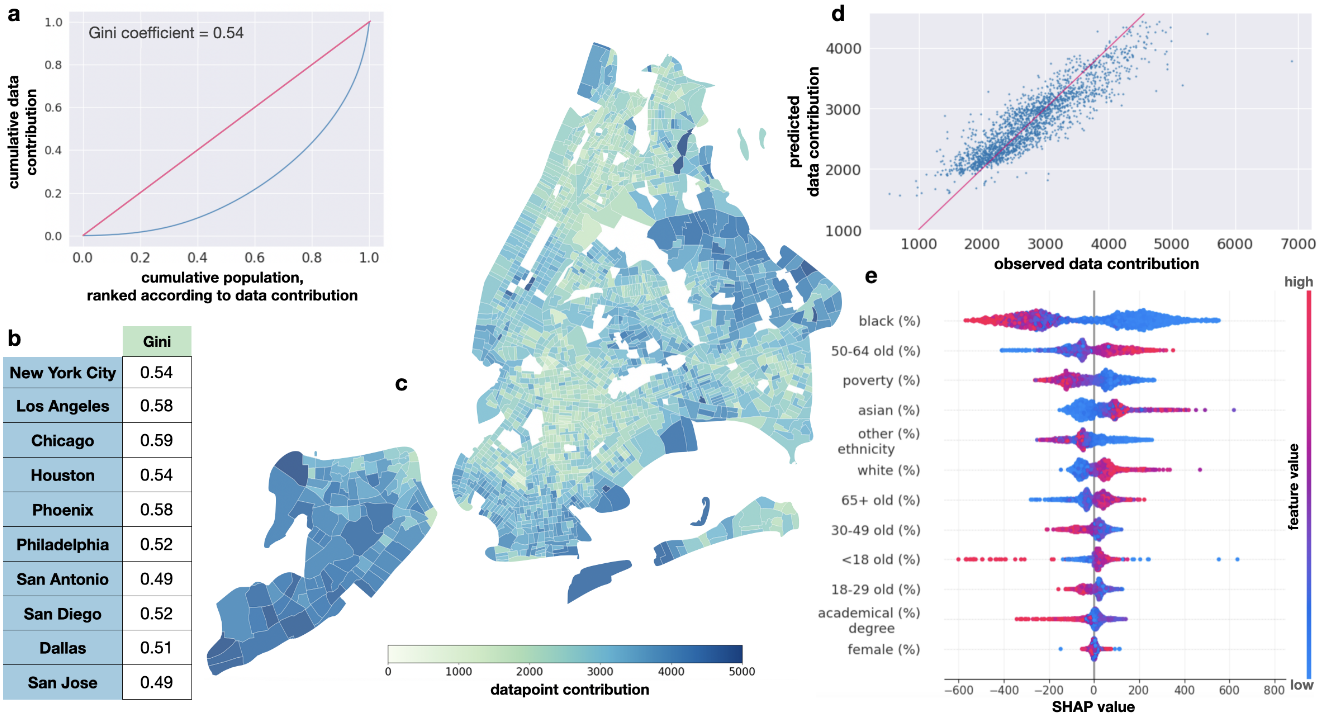
Representativeness Explained: Understanding Data Production Biases in Human Mobility Data

Katinka den Nijs[1], Elisa Omodei[2], and Vedran Sekara[3]
[1] IT University of Copenhagen, Denmark - katinkadennijs@gmail.com
[2] Department of Network and Data Science, Central European University, Austria - omodeie@ceu.edu
[3] IT University of Copenhagen, Denmark - vsek@itu.dk

Our capabilities to collect, store and analyze vast amounts of human mobility data have greatly increased in the past decades [1]. Today these datasets play a critical role in a majority of algorithmic systems [2], business processes [3], and policy decisions [4]. While lots of progress has been made in developing new models to analyze the data, there has been much less focus on understanding the fundamental shortcomings of these big datasets [5]. Here we focus on understanding the representativeness of high-resolution human mobility datasets with the aim of developing methodologies to debias data.

Large-scale human mobility is often measured using digital tools such as smartphones. However, it remains an open question how truthfully these digital proxies represent the actual travel behavior of the general population. The literature shows that smartphones ownership is unequally distributed across society. For instance, in the US only 81 out of 100 people own a smartphone [6], with younger and wealthier groups being more likely to own one [7]. As such, it is a well-established fact that not everybody is properly represented in digital datasets. However, we show this type of bias is not the only one to be mindful of. It is equally important to understand *how* people are represented in data, i.e. is data of similar utility, amount, and quality.

To understand data representativeness we focus on the amount of data each individual generates. We analyze mobility traces (on GPS resolution) collected using smartphones. The mobility data is provided by Spectus, a location intelligence platform. Data is collected from anonymized users who have opted-in to provide access to their location data anonymously, through a CCPA and GDPR-compliant framework.



	Gini
New York City	0.54
Los Angeles	0.58
Chicago	0.59
Houston	0.54
Phoenix	0.58
Philadelphia	0.52
San Antonio	0.49
San Diego	0.52
Dallas	0.51
San Jose	0.49

Figure 1: **Datapoint representation April 2019 (pre-COVID). a,** Lorenz curve for New York City (NYC) with a Gini coefficient of 0.54. **b,** Gini coefficients for the ten largest cities in the US. **c,** Observed median dataproduction levels for each census tract in NYC. The median is calculated over all individuals living in a census tract. **d,** Predicted datapoint contribution per census tract using a random forest model ($R^2 = 0.78$). **e,** SHAP values for the random forest model.

Through its Social Impact program, Spectus provides mobility insights for academic research and humanitarian initiatives. Our focus is on data generated by individuals living in the ten largest cities in the United States. In each city, people are linked to census tracts based on their inferred home location (home locations are identified on the block group level as the most frequent location during night).

It is well-documented that wealth, income, and consumption are all unequally distributed in society [8]. Here we show that data production, i.e. the number of datapoints mobility datasets contain for each individual, are also unequally distributed. Fig. 1a shows the Lorenz curve for data production for New York City (NYC). It has a Gini coefficient of 0.54, which is comparable to how wealth is distributed in the city (Gini = 0.55 [9]). NYC is not an outlier, other large cities in the US have similar Gini coefficient (Fig. 1b). This inequality in datapoints will undeniably affect how well mobility traces can be reconstructed, trusted, and applied for different purposes.

To understand what causes data inequality, and ultimately also gain an understanding of how to make mobility datasets more representative, we build a model to estimate data production (here measured as the median number of produced datapoints) on census tract level. Fig 1c shows the observed median datapoint contribution for census tracts in NYC for one month, April 2019. We focus on census-tracts as there exists rich demographic information on poverty, sex, age, education, and ethnicity, from the United States Census Bureau. Using a random forest model, because it is able to model non-linear relationships, we find that data production rates can be predicted with a high confidence (Fig 1d, R^2=0.78).

To explain model predictions we use a SHAP plot [10] (Fig 1e). The demographic feature which has the highest influence on data-production rates is the percentage of people who self-identify as black living in a census tract. Higher fractions drastically reduce the number of datapoints for a census tract (up to 600 datapoints less for the specific month), while tracts with low rates contribute positively (with up to 550 datapoints more during this month). Put differently, tracts with high percentages of black individuals are worse represented in mobility datasets. Poverty has a similar effect. Poorer tracts produce less datapoints, with up to 300 datapoints less, compared with wealthier tracts. Interestingly, tracts with high percentages of older demographics, 50-64 and 65+ year-olds, produce more datapoints. While tracts with high percentages of 18-29 year-olds contribute to fewer datapoints being collected.

Our model reveals a complex interplay between demographics factors and whose data is represented in human mobility datasets. Some of these effects can be caused by certain demographics having less access to the technological means of dataproduction (in our case smartphones). However, it is hard to believe that is what caused 18-29 year-olds to produce less data. This can potentially be the result of algorithmic confounders [11]. Our mobility dataset is created by pooling data from multiple smartphone apps. Each of these may be using different social engineering mechanisms to collect data, which will invisibly nudge their users towards specific behaviors—future work will involve untangling these effects.

Biased data leads to biased algorithms. Our analysis shows how certain demographic factors can severely impact data representation. This a work in progress, our next steps include uncovering the effects of data-production on mobility networks. Will worse data representativeness lead to incomplete travel networks? This work constitutes a first step towards understanding how to develop techniques that correct for biases in widely used human mobility datasets.

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[2] C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad, "Measuring algorithmically infused societies," *Nature*, vol. 595, no. 7866, pp. 197–204, 2021.

[3] C. Demunter, "Tourism statistics: Early adopters of big data," *Publications Office of the European Union*, 2017.

[4] J. E. Blumenstock, "Estimating economic characteristics with phone data," in *AEA Papers and Proceedings*, vol. 108, pp. 72–76, 2018.

[5] K. Crawford and R. Calo, "There is a blind spot in AI research," *Nature*, vol. 538, no. 7625, pp. 311–313, 2016.

[6] Pew Research Center, 2019. Smartphone ownership is growing rapidly around the world, but not always equally.

[7] Pew Research Center, "Mobile fact sheet," 2021. [Last accessed 2023-02-22] https://www.pewresearch.org/internet/fact-sheet/mobile/.

[8] L. Chancel, T. Piketty, E. Saez, and G. Zucman, *World inequality report 2022*. Harvard University Press, 2022.

[9] T. Bach, "The 10 us cities with the largest income inequality gaps," *US news and world report*, 2020.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[11] M. J. Salganik, *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

The impact of apparent ‘teleports’ on the estimation of mobility from Call Detail Records (CDRs)

Robert Eyre (robert.eyre@flowminder.org), Galina Veres, Véronique Lefebvre
Flowminder Foundation

Population mobility and call routing. The assumption underpinning the extraction of mobility information from Call Detail Records (CDRs) or other mobile operator data is that, if a subscriber is recorded from cells that are in different locations, then the subscriber has moved. This assumption is not always true however, as a subscriber’s consecutive calls can be routed by distant cells even if the subscriber has not moved. Call rerouting (or rerouting of SMS and data sessions) therefore creates spurious movements if a change in subscriber recorded location is interpreted as movement, which can make it appear as if subscribers travel further, and more often than they actually do.

Spurious movements can be detected when a change in recorded location occurs too quickly to be a plausible movement (e.g. 2 calls made with a 5 sec interval and routed by cells distant by 20km, which would be a speed of 14,400 km/h), which we call ‘teleports’. However, rerouting can also create spurious movements that are plausible (e.g. 2 calls made with a 5 hour interval and routed by cells distant by 20km) and not distinguishable from actual movements. Therefore we can only measure the tip of the iceberg (‘teleports’) of the larger ‘rerouting-without-movement’ problem.

Measuring and weighting ‘teleports’. To investigate the rerouting-without-movement problem we start by measuring its visible part (the teleports) and experiment with reducing their numbers through 1) spatial aggregation of recorded locations, and 2) weighing down the number of connections (recorded changes in subscriber location) observed for pairs of locations where teleports are common.

We used 7 months of CDRs from a random sample of 40,000 subscribers of Digicel Haiti and counted the number of subscribers’ consecutive connections for each pair of cell towers, and for each pair of level 3 administrative units (admin3). We then classified as ‘teleports’ all connections made at a speed exceeding 120km/h and computed the proportion of teleports over all connections, for each pair of locations.

Figure 1 shows the ratio of teleports to all connections, between each pair of cell towers (left) and each pair of admin3 units (right). For some pairs of towers none of the observed connections are real movements (all connections are ‘teleports’) and many pairs have high ratios of ‘teleports’. In comparison, aggregating towers by admin3 (which we commonly do in our mobility indicators) leads to lower proportions of ‘teleports’ as rerouting is more likely to occur over shorter distances and so more within admin3 units than across units. Grouping of towers near to administrative boundaries and taking cell coverage and direction into account would further reduce teleports (and all spurious movements).

We also observe that teleports are more likely in some locations, such as densely covered areas and areas where cells have longer coverage (e.g. coastal areas, flat terrain). Connections classified as teleports can be excluded from counts or ‘1 - proportion of teleports’ can be used to weight down aggregated numbers of connections for each pair of locations. This would need to be re-computed over time or averaged for specific time periods, as ‘teleports’ are more likely during the day or during events (e.g. new year’s eve) when the network is busy. Weighing down the number of connections using the proportion of teleports would improve comparisons of mobility across different locations, and over time, e.g. avoiding that mobility estimates are dominated pairs of locations with large proportions of ‘teleports’. However, ‘teleports’ are only the visible spurious movements and the number and proportion of all spurious movements are not accounted for.

Identifying subscribers’ ‘meaningful locations’ to attenuate the rerouting-without-movement influence. In addition to adjusting aggregated statistics we can also attempt to correct individual CDR trajectories by clustering nearby cells that are frequently recorded for each subscriber, to obtain a set of ‘meaningful locations’ for each subscriber. The assumption is that most rerouting-without-movement would occur within the meaningful locations, and that a change in meaningful location has a higher chance of being a real movement. We used the Hartigan’s Leading algorithm[1] to explore this method on our sample. **Figure 2** shows the cell towers used by a synthetic subscriber (left), how these towers would be clustered if only taking their location into account (middle), and the clusters obtained for this particular synthetic subscriber, i.e. their ‘meaningful locations’ (right). The proportion of teleports between meaningful locations is smaller than the proportion of teleports between fixed cell clusters (based on cell locations only), however when aggregating clusters by admin3 units the proportion of teleports is the same whether clusters were derived from individual CDR traces or from cell locations only, so

[1] Hartigan, John A. 1975. Clustering Algorithms. Wiley.

the identification of individual meaningful locations would only be worth computing for problems requiring sub-admin3 resolution.

Population mobility and call routing during crises. When the network is busy, increased rerouting is necessary, and to more distant locations, to maximise the chance for all subscribers to access the network. Increased phone usage and rerouting-without-movement may occur in crises (e.g. following disasters or during conflicts), which could lead to overestimates in crisis-driven displacements and erroneous ranking of most affected locations. This is particularly concerning as CDR-derived estimates could mislead response and recovery efforts by focusing on the wrong locations. Methods to correct for spurious movements (e.g. weighing by teleport proportions and identification of meaningful locations) would need updating immediately at the onset of a crisis and regularly throughout, to cope with fast changes in both network activity and mobility, and keeping with constraints of limited compute power and the requirement to provide updates in a timely manner.

Immediately upcoming work will investigate the rerouting-without-movement problem on the Digicel Haiti subscriber base during and after the 2021 earthquake in Haiti, and other recent displacement crises, to compare with routine patterns and and further test the envisaged correcting methods. Methods we implement operationally also need to allow for frequent (ideally daily) production of mobility estimates on a limited infrastructure at mobile network operators, to provide critical information to guide humanitarian and developmental efforts in the low and middle-income countries we work in.

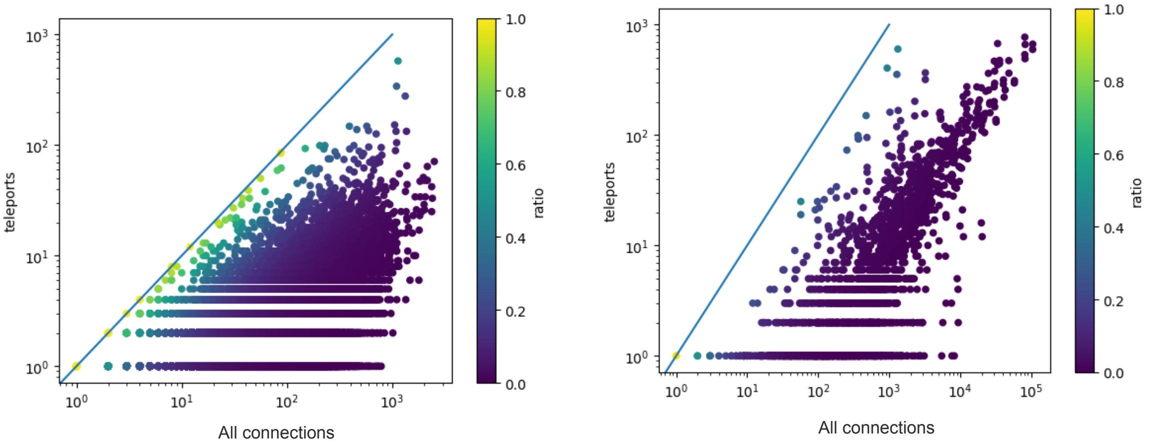


Figure 1: **Total teleports over the study period vs all connections for each pair of cell towers (left) and admin3 units (right)**. Large ratios indicate that the number of teleports is close to the number of connections (i.e. none or few of the recorded changes in location correspond to movements). The proportions of teleports between admin3 units are lower than between cell towers as rerouting is a local process.

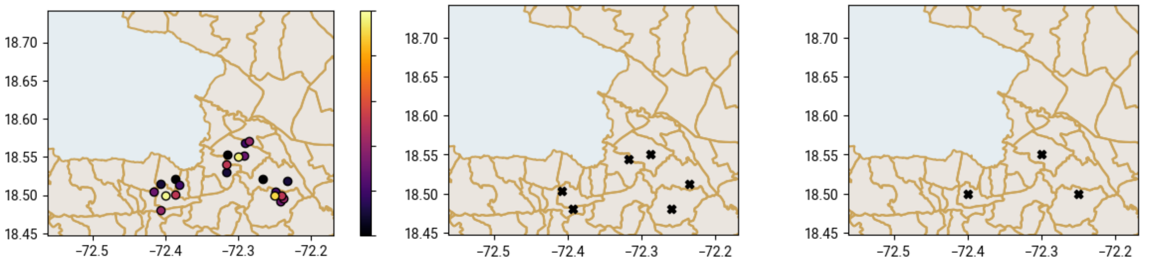


Figure 2: **Synthetic data: Identifying a subscriber's 'meaningful locations' to attenuate the rerouting influence. Left**: Synthetic example of the (randomly generated) cell towers serving a single (randomly generated) subscriber over a period of time, coloured by frequency of records. **Middle**: Example of clustering of cell towers based on their location - same clustering for all subscribers (crosses correspond to cluster centroids). With such clustering there is a risk to separate nearby locations that are commonly visited by a subscriber (or nearby cell towers that reroute their calls when they don't move), potentially resulting in an inflated number of trips between clusters. **Right**: Clustering of cell towers derived from the synthetic records of the (randomly generated) subscriber, identifying their 'meaningful locations' which aggregate both short movements and rerouting into a single spatial unit (which needs to be updated over time).

Auditing the Fairness of Place-Based Crime Prediction Models Implemented with Deep Learning Approaches

Saad Mohammad Abrar[1], Naman Awasthi[1], Jiahui Wu[1], Enrique Frías-Martínez [2], and Vanessa Frías-Martínez[1]

[1]University of Maryland, {sabrar, nawasthi, jeffwu, vfrias}@umd.edu
[2]Universidad Camilo José Cela, enrique.frias@ucjc.edu

Environmental criminology provides a theoretical foundation to study crimes from the perspective of places *i.e.,* places with different urban functions are viewed as crime attractors or crime generators [3]. Through the lens of place-based crime prediction, researchers have studied the relationship between future crimes, historical crimes, social interactions and the built environment across different regions using a diverse set of approaches [6]. For example, kernel density estimation - which was very common in the early efforts of crime prediction - uses the estimated density of historical crimes as a measure of risk for future crime areas [2]; while epidemiological models [10] propose an epidemic-type aftershock sequence model to utilize the near repetition patterns of historical crimes whereby, as prior work has shown, the spatio-temporal patterns of crimes in one location increase the probability of other incidents occurring at nearby locations. More recently, however, place-based crime prediction has been dominated by deep learning models. Deep-learning models for place-based crime prediction claim to outperform prior models by accurately predicting the non-linear spatio-temporal patterns of crime and its relationship with the built environment via temporal and spatio-temporal neural architectures [7]. Despite their increased predictive accuracy, researchers have failed to discuss the fairness of place-based crime prediction models implemented with deep learning.

Prior work has shown that non-deep learning crime predictive models predicted non-white crime locations at roughly 1.5 times the rate of white crime locations, and that these predictions were in part a reinforcement or an amplification of bias present in the crime data used to trained the models ([8]). For example, White and wealthy crime victims or female-headed household victims in the US are less likely to report to the police [13]; and the police are less likely to record to their databases minor crimes in majority minority-race and immigrant neighborhoods in the US due to the *unworthy victim* perspective [14]. We propose to carry out the first systematic analysis of the fairness of place-based crime prediction models implemented with deep learning (with a focus on race and ethnicity), and an evaluation of how that fairness - or lack thereof - might be related to bias in the crime data used to train the models.

Crime opportunity theories attempt to explain the occurrence of crimes in terms of human behaviors by looking into how variations in people's routine activities - measured at the intersection between place and human mobility - might affect the *opportunities* for crime [4]. Prior work in place-based crime prediction using non-deep learning approaches has shown that incorporating human mobility variables extracted from cell phone data can improve the accuracy of the predictions including the number of people present at a given place (footfall) [1], the pass-through flows [15] or the urban spatial structure [16]. Place-based crime prediction models implemented with deep learning approaches have also incorporated crime opportunity theories by modeling routine activities using human mobility data; and have also shown that incorporating mobility data can improve the accuracy of the predictions [12]. Nevertheless, the human mobility data used in these studies - generally collected from cell phones and cell phone apps - might suffer from sampling bias due to the fact that cell phone ownership is not equal across social groups [11]. In fact, recent work by Coston *et al.* has shown that certain protected groups in North Carolina, such as Black population and elder residents, are more under-represented in the smartphone mobility data collected by the location intelligence company SafeGraph than White and younger groups [5].

In this paper we aim to fill in the existing research gaps by (1) quantifying the fairness of place-based crime prediction models implemented with deep learning and trained with either crime or both crime and mobility data; (2) measuring changes in the fairness of place-based crime prediction models implemented with deep learning when human mobility data collected from cell phones is incorporated to the model *i.e.,* does adding mobility data decrease fairness, despite its apparent improvement on prediction performance?; and (3) analyzing the root causes behind the changes in fairness metrics, looking into crime data bias, mobility data bias and algorithmic bias. Quantifying fairness in this context is complex due to the diverse set of parameters to be considered: different deep learning models, different geographical areas (*e.g.,* cities) and different types of crimes. At the same time, a large amount of work has shown that there often exists a trade-off between performance and fairness in predictive models [9, 17]. Thus, to narrow down our analysis, we first implement a large variety of state-of-the-art place-based deep learning crime prediction models using crime only or both crime and mobility data across four American cities (Baltimore, Minneapolis, Austin and Chicago) and eight types of crimes, and we then select the one with the best performance to carry out an in-depth audit of the fairness of that model across cities and types of crimes, with a focus on race and ethnicity. We posit that by analyzing fairness for the best performing model, our results will reveal the most extreme cases of unfairness.

Ultimately, the focus of this paper is to bring to light the fairness issues embedded in current crime prediction approaches that focus on deep learning methods and on the use of human mobility data; as well as to encourage researchers and decision makers to think more critically about the development and deployment of place-based crime prediction software based on deep learning approaches. The main findings of our paper are:

- Our results show that the best performing place-based crime prediction deep learning models, trained with crime or both crime and mobility data, generally output unfair predictions for groups that have historically suffered from racial and ethnic discrimination.
- Our results show that although mobility features can enhance the prediction performance over deep learning models that only use crime data, it sometimes comes at the cost of a decrease in racial/ethnic fairness.

References

[1] A. Bogomolov, B. Lepri, J. Staiano, E. Letouzé, N. Oliver, F. Pianesi, and A. Pentland. Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data*, 3(3):148–158, Sept. 2015.

[2] K. J. Bowers, S. D. Johnson, and K. Pease. Prospective Hot-Spotting: The future of crime mapping? *Br. J. Criminol.*, 44(5):641–658, Sept. 2004.

[3] P. Brantingham and P. Brantingham. Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research*, 3(3):5–26, Sept. 1995.

[4] C. R. Browning, N. P. Pinchak, and C. A. Calder. Human mobility and crime: Theoretical approaches and novel data collection strategies. *Annu. Rev. Criminol.*, Jan. 2021.

[5] A. Coston, N. Guha, D. Ouyang, L. Lu, A. Chouldechova, and D. E. Ho. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 173–184, New York, NY, USA, Mar. 2021. Association for Computing Machinery.

[6] R. N. Davidson. *Crime and Environment.* St. Martin's Press, New York, 1981.

[7] L. Duan, T. Hu, E. Cheng, J. Zhu, and C. Gao. Deep convolutional neural networks for spatiotemporal crime prediction. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, pages 61–67. csce.ucmss.com, 2017.

[8] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, Oct. 2016.

[9] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[10] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *J. Am. Stat. Assoc.*, 110(512):1399–1411, Oct. 2015.

[11] Pew. Pew research center: Mobile fact sheet. `https://www.pewresearch.org/internet/fact-sheet/mobile/`, 2021. [Online; accessed 18-May-2022].

[12] A. Stec and D. Klabjan. Forecasting crime with deep learning. *arXiv preprint arXiv:1806.01486*, 2018.

[13] TheMarkup. Crime prediction software promised to be free of biases. new data shows it perpetuates them. `https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them`, 2021. [Online; accessed 2-May-2022].

[14] S. P. Varano, J. A. Schafer, J. M. Cancino, and M. L. Swatt. Constructing crime: Neighborhood characteristics and police recording behavior. *Journal of Criminal Justice*, 37(6):553–563, Nov. 2009.

[15] J. Wu, E. Frias-Martinez, and V. Frias-Martinez. Addressing Under-Reporting to enhance fairness and accuracy in mobility-based crime prediction. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '20, pages 325–336, New York, NY, USA, Nov. 2020. Association for Computing Machinery.

[16] J. Wu, E. Frias-Martinez, and V. Frias-Martinez. Spatial sensitivity analysis for urban hotspots using cell phone traces. *Environment and Planning B: Urban Analytics and City Science*, 2021.

[17] H. Zhao and G. Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685, 2019.

Title: Tackling uncertainty in the spatial density estimation from Mobile Network Operator data

Author: Marco Ramljak

Email: marco.ramljak@ramboll.com

Affiliation: Senior Geospatial Data Analyst, Ramboll

Associated repository: https://github.com/R-ramljak/MNO_uncertainty

Document: Two-page abstract for NetMob 2023

Introduction: Mobile network operator (MNO) data represent a rich potential source for estimating the spatial distribution of mobile phones at a given time. Moving from proof-of-concept towards the implementation of MNO data into official statistical production processes, [1] emphasize the importance of a reference methodological framework and offer a first blueprint. The proposed processing pipeline from raw MNO data to the final spatial density map requires modeling the (approximate) spatial footprint of cells – a task called *cell geolocation*. Two approaches are identified in the literature on how to operationalize cell geolocation: (i) *tessellations*, which divide the reference area into spatially disjoint coverage areas associated with different (groups of) cells. And (ii) *overlapping cells*, which are methods that allow cell coverage areas to overlap. Most of the past applications utilize tessellations, assuming that every mobile device connects to its closest antenna, which is the inherent assumption of the popular Voronoi estimator [2], [3]. Essentially, any tessellation method reduces to a simple area-proportional computation and is, therefore, considered a deterministic estimation strategy. While its simplicity seems quite powerful in the context of scalability, the assumption of non-overlapping coverage areas does not match reality. MNOs purposefully architect their network in a way that their coverage areas overlap. Cell overlap reduces coverage holes, assuring well-enough reception quality across an operating area (in theory).

Cell overlapping geolocation methods, proposed by [4], [5], use radio propagation simulation techniques to actually model individual cell footprints. This allows the implementation of many cell-specific characteristics, if known, and can explicitly handle overlapping coverage areas by introducing a probabilistic framework. Recent simulation studies have shown that, with appropriate estimation methods based on probabilistic models, the availability of more detailed coverage area information allows to improve the spatial accuracy of the final estimate considerably, compared with the simpler traditional methods relying on Voronoi geolocations [1], [6]. However, such results were obtained (i) under the assumption of perfect knowledge on the coverage areas, i.e., omniscient network topology data, and (ii) limited to a single simulated network scenario characterized by a dense, multi-layer coverage pattern with a high degree of cell overlapping. Under no realistic circumstances is it possible to model coverage areas perfectly accurately because they depend on many parameters, some of them being very volatile or immeasurable, such as weather conditions. It is questionable whether we can expect such accurate spatial density estimations when using network topology data of a realistic certainty level, i.e., imperfect, and if they are robust across different network scenarios.

Research objective: This methodological sensitivity analysis draws directly on the work by [1], [6] continuing the methodological research and focusing on the aspects of geolocation uncertainty within the spatial density estimation. We investigate through simulations the robustness of probabilistic estimators to uncertainties and inaccuracies in the model input parameters, namely (i) the matrix of emission probabilities and (ii) prior information.

Methodology: We design and conduct a simulation study, based on semi-synthetic data with all state-of-the-art estimation strategies containing deterministic and probabilistic cell geolocation methods. We develop parametric techniques that purposefully introduce inaccuracies into the geolocation with tunable magnitude. One mismatch technique introduces spatially sensitive random noise, while the other technique quantizes the variability level of the available signal strength information. Also, we consider distinct prior information vectors with varying levels of informativeness. To substantiate the robustness of our findings, we assess the impact of different network scenario parameters (e.g., cell density, multiple cell layers) on the relative performances of the various estimators. Owing to the spatial nature of the estimation problem, we use the Kantorovich-Wasserstein distance to measure the (dis)similarity between the estimated density and the true population distribution.

To conduct our experiments and mimic MNO-like data, we use the MNO-simulator workflow, which is constructed within the open-source programming language R and supports with development of the necessary non-trivial software.[1] The MNO-simulator enables the user to work with any kind of data availability – experiments can be conducted with complete synthetic scenarios (no real-world data available), semi-synthetic scenarios (some real-world data available, e.g., census data or (partial) cell coverage data), and real scenarios if actual data are available.

Results: Our results indicate that probabilistic estimators are robust towards inaccuracies in the emission probabilities. We find that probabilistic estimators deliver more accurate results than the Voronoi methods in all scenarios, even when confronted with extremely mismatched estimation models. Given that these observations are robust across different network scenarios, we advise investing further academic efforts in developing more performant estimation strategies based on probabilistic geolocation methods. For iterative estimators, we observe divergence, which occurs in some special cases under severe mismatching conditions, pointing to the need to improve further the numerical methods adopted by probabilistic estimators. We expect our results to encourage further research on the probabilistic framework and novel estimation strategies.

References:

[1] Ricciato, F., Lanzieri, G., Wirthmann, A., & Seynaeve, G. (2020). Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, *68*, 101263.

[2] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779-782.

[3] Sakarovitch, B., Bellefon, M. P. D., Givord, P., & Vanhoof, M. (2018). Estimating the residential population from mobile phone data, an initial exploration.*Economie et Statistique*,*505*(1), 109-132.

[4] Ricciato, F., Widhalm, P., Craglia, M., & Pantisano, F. (2015). *Estimating population density distribution from network-based mobile phone data*. Luxembourg: Publications Office of the European Union.

[5] Tennekes, M., & Gootzen, Y. (2022). Bayesian location estimation of mobile devices using a signal strength model. *Journal of Spatial Information Science*, (25), 29-66.

[6] Ricciato, F., & Coluccia, A. (2021). On the estimation of spatial density from mobile network operator data. *IEEE Transactions on Mobile Computing*.

[1] We created the MNO-simulator workflow within the development of this project, however, given its flexibility and potential usage for different projects, we have generalized it. An introductory presentation linking also to other resources can be found here: https://github.com/R-ramljak/MNO_simulator .

Correcting measurement biases in the detection of long and short stay locations in sparse Call Detail Records (CDRs)

Galina Veres*, Jono Gray, James Harrison, Véronique Lefebvre
Flowminder Foundation
*corresponding author email: galina.veres@flowminder.org

Introduction. With mobile phone technologies continuing to spread around the world, Call Detail Records (CDRs) represent an attractive additional data source for inferring internal migration in addition to traditional data sources such as surveys and administrative records, especially in low and middle-income countries (LMICs), where traditional statistics can be either unavailable or difficult to collect. Stay locations are the basis of most mobility statistics derived from location data and CDRs. Detecting stay location is crucial to migration statistics (Where do people live? When and to where do they change their residence?), to disaster statistics (Where have people lost their homes? Where are they displaced to? Did they then return home and when?), and to a multitude of other applications from informing disease spread to tourism statistics. Though a number of methods were proposed in the literature to detect stay locations from CDR data, some challenges remain especially in LMICs - where CDRs tend to be sparser - due to irregularity and low frequency of phone use, network instability, so called 'ping-pong' effect as artefact of mobile communications, and resulting conflation between changes in phone usage and changes in mobility. We present our solution to detect stay locations (long and short stays) and relocations from CDRs, which addresses the above issues related to particularly sparse data in LMICs and can be run on mobile operator infrastructure (constrained in memory and compute power) to ensure data privacy.

Problem formulation. In this paper, we address two problems of using CDR data for migration and disaster statistics: 1) robust detection of stay locations and relocations in individual CDR traces and 2) developing aggregated mobility indicators corrected for biases stemming from changes in phone usage. Stay location is often assumed to be a location of the last call of the day for an overnight stay location detection, or the modal location of the last call of the day for longer periods (e.g. detecting 'home location'). However, such methods of relocation detection between two stay locations lead to ~ 79% of false discoveries at the daily level, from experiments we conducted on CDRs traces using a subset of 781 Digicel subscribers in Haiti for whom we manually labelled stay locations (at administrative level 3) and relocations. This indicates the need for better performing methods to detect relocations and stay locations from individual CDR traces. Another source of error arises from summing the detected stay locations for each region each day or each month to estimate the number of 'residents' of each region (or subscribers who 'stay' in each region). However, numbers of residents and their temporal variations computed as 'stay location counts' contain both variations in phone usage and variations in internal mobility. We quantified the proportion of temporal variation in 'stay location counts' due either to phone usage or to mobility, using CDRs from Digicel Haiti, and found that only 23% of the monthly 'stay location counts' variations (on average across regions) in a 24-month study period are attributable to mobility in this case. This indicates the need to derive resident numbers directly from observed mobility (relocations).

Method development and validation. We propose a fast and elegant solution to fix such measurement biases, while ensuring it is operationally feasible: in near real time (updating every day), and on infrastructure constrained in compute power and memory. To ensure data privacy constraints we impose for all computations on individual data to be done on a server located at the mobile operator premises, behind their firewall.

We improved on the common methodology of capturing the modal location of the last call of the day by using a system of two moving windows: a short window to estimate a daily overnight location and a longer window (length depending on the type of stay to be detected) within which we check for a dominant location. For the short window, we tested several methods and concluded the last call of a day method is a trade-off between performance and required execution time for operational purposes. We compared the modal last call of a day, the modal distinct day and the anchor methods for detecting stay locations within 7-day windows at the fine spatial level of the group of cell towers. The last call of a day and anchor methods performed similarly for stay location detection, based on our labelled subset. Recall was better by 5% for the last call of a day, precision and false positive rate was better for anchor method by 19% and 24% respectively for relocation detection taking location of relocation into consideration. However, execution time increased by 2.5 times for the anchor method in comparison to the last call of a day, which was retained as the method to use as 'first pass' on the data in the short window. We use a 7 day rolling window (short window) and assign a daily stay location as a modal location of the last call of a day as a 'first pass' over individual trajectory. In a 'second pass' over the time series, we use a longer window (e.g. 28 days) to search for a dominant location within the locations returned by the modal last call of day location method. This is particularly relevant when searching for home location, when the subscriber could be absent from their home. If there is a dominant location, then a subscriber is assigned this location as their residence location, otherwise the subscriber is 'unlocatable', i.e. avoiding to assign a residence or stay location when a subscriber has been mainly on the move. This effectively reduces the number of false relocations compared to the simpler common method, and creates a sample of subscribers who are stable enough and active enough so that their stay location can be detected. This two windows method also permits an approximation of stay duration that is robust to missing data and noise. Then a relocation is simply detected as a change in stay location for each subscriber. We tested the proposed method for detecting residence location and relocation on a manually labelled subset of subscribers and found that false discoveries were reduced by 33% for daily relocation detection, from ~79% for a simple modal last call of a day location method to 53% (our novel method). We are working towards further reduction of false discoveries in our work of suppressing the ping-pong (or re-routing) artefact of CDR data by weighting the number of relocations between regions.

Secondly, while this approach creates a more robust detection of stay location at the individual level, the number of residents (or stays) in a location cannot simply be computed by summing the number of stay locations detected, as this may mainly be driven by the number subscribers becoming (un)locatable. To ensure that we only retain variations in the number of residents derived from mobility (changes in stay locations), we propose to estimate the number of residents directly from the number of relocations (e.g. change in stay locations between two months). The difference in the number of residents between two months is equal to the total number of subscribers relocating into the location minus the number who relocated out of the location. The number of estimated residents is then the cumulative sum of these differences added to a baseline or starting time as described by

$$\hat{residents}(i,t) = \hat{residents}\,(i,0) + \sum_{j=1}^{t}(reloc\_to(i,i-1,j) - reloc\_from(i,j-1,j)),$$

where $i\,(i = 1,..., N)$ is a location i from N locations in the area of interest, $\hat{residents}(i,0)$ is an estimated number of residents in the baseline period or start point. Using this method, large fluctuations in subscriptions (often observed in urban areas) are suppressed. The method ensures that the variations of our resident indicator are now derived from observed mobility, at least for short time periods (changes in the subscriber base will still create a drift over years, which needs to be addressed by further weighting factors and auxiliary survey data - a problem we are also working on).

Operational use. We have been able to use our stay detection method and resident indicator design on an operational level, computing resident estimates and internal migration monthly in 3 countries (Haiti, Ghana and the DRC), and adjusting them for representation biases using survey data. Such data can be used for service provision planning and for refining other statistics taking population mobility into account such as disease incidence and prevalence. We also use the method with shorter time windows to detect disaster-driven displacements and returns and inform disaster management.

Using survey data to correct for representation biases in mobility indicators derived from mobile operator data to produce high-frequency estimates of population and internal migration

Roland Hosner (roland.hosner@flowminder.org), Zachary Strain-Fajth, Véronique Lefebvre
Flowminder Foundation

Background

While mobile operator data constitute an important source of evidence on population mobility, particularly in data-poor settings such as Low- and Middle-Income Countries (LMICs), they remain partial in terms of population coverage, and prone to representation biases which are difficult to measure, control, and correct for in the absence of independent auxiliary data.

The use of mobile operator data (including Call Detail Records, CDRs) to estimate changes of sub-regional population counts over time and population mobility often relies on a series of assumptions, including that movements observable for mobile phone users are similar to the movements of the general population. But mobile phone users have been shown to be different from the general population in many characteristics (gender, age, socio-economic status (SES) including education, degree of urbanisation of place of residence), and such characteristics also lead to differences in their mobility compared to that of the non-phone-users. As a result, CDRs cannot be used as a source of population statistics unless representation biases are corrected for.

Although methodological development and research on the application of CDR data has progressed in recent years, few solutions have been offered so far for the adjustment of such representation biases in CDR-based indicators. The inherent biases of these data require correction through joint modelling with traditional data sources such as surveys. Unadjusted indicators may severely misrepresent population movements and deducted population distributions, particularly in regions where phone use is low and subscriber numbers are small.

We see from the available survey data that a range of important parameters differs between relocations (per combination of origin and destination locations): the average number of SIM cards used, the number of travellers per SIM card, the share of phone users and the MNO market shares do differ between relocation flows. This also means that we observe mobility differences between phone users and non-users.

These problems of selection bias and representativity in general are common to all big data analyses, where often the erroneous assumption is made that the quantity of data renders representation bias negligeable. We propose here a method that corrects for such biases in CDR-derived indicators. What is more, the method could be generalised to other types of estimates and big data datasets.

Data

Flowminder has ongoing access to CDR data in Haiti, Ghana and the DRC, and has commissioned or joined primary survey data collection in these countries. Our method relies on CDR aggregates, survey data and existing population estimates from National Statistical Offices, the United Nations, WorldPop and other sources.

For the DRC, we use CDR aggregates from Vodacom with survey data from a microcensus and a telephone survey in 2021 to estimate relocations and residents by health zone for all months since February 2020.

For Haiti, we use CDR aggregates from Digicel with survey data from a general population survey in 2022 to estimate relocations and residents by communal section for all months since January 2020.

For Ghana, we will combine CDR aggregates from Vodafone with census data from 2021, survey data from a telephone survey and a general population survey in 2022 to estimate relocations and residents by district for all months since December 2019.

Method

As a first step, we assess all available national, regional and sub-regional population estimates from National Statistical Offices and other sources. Sub-regional population estimates for the baseline month - the first month for which CDR aggregates are available - are then derived from existing estimates or projections. Importantly, we are not using CDR-derived counts of home locations for scaling to estimate residents, because our analyses show these counts strongly depend on phone coverage and phone use, and are not suitable for direct scaling.

In the second step, we apply adjustment and scaling factors only to CDR-derived relocations, i.e. detected changes of home locations or stay locations over time. These CDR-derived inflows and outflows of (frequent and locatable) phone users are adjusted and scaled based on survey-derived parameters. The approach is to scale the bilateral relocations between sub-regions (flows) from one month to the next, aggregate total scaled inflows and outflows per sub-region, and add scaled net flows in a cumulative manner over time to baseline estimates to arrive at time-series estimates of residents. As a result, the difference between total inflows to and outflows from a subregion corresponds to the change in estimated residents from one month to the next.

As a final step, we apply factors to adjust for overall population growth or decline, derived from estimates provided by National Statistics Offices or the United Nations.

Discussion

Our method corrects for differences in mobility between phone users and non-users and further biases observed in the survey data. However, there are few data sources to validate such estimates. Additionally, to enable the ongoing production of population and mobility statistics from mobile operator data, longitudinal survey data are needed to capture potential changes in these mobility differentials of the phone using and non-phone using populations.

In summary, we have been able to develop and apply a bias-correcting methodology to produce estimates of internal migration and sub-regional population change in two LMIC countries (Haiti and the DRC) and have started work on a third country (Ghana). This data can be used for a wide range of use cases, from the health sector to humanitarian work, disaster preparedness, and to official statistics, including reporting on development indicators. For example, these estimates can highlight sub-regions with monotonic population increases (above natural population growth), or sub-regions with fluctuating populations such as those affected by large population displacements in the DRC and Haiti.

Monitoring the impact of the 2020-2021 COVID-19 mobility restrictions: Flowminder CDR analysis in seven low- and middle-income countries

James Harrison[1a]*, Veronique Lefebvre[1a], Tracey Li[2a], Xavier Vollenweider[1a], Chris Brooks[1a], Jonathan Gray[1a], Sophie Delaporte[1a], Galina Veres[1a], Robert Eyre[1], Thomas Smallwood[1], Michael Harper[2a], Caterina Irdi[1], Wole Ademola Adewole[2], Omar Seidu[3], Apphia Yuma[1], Caroline Reeves[1], Cathy Riley[1], Daniel Power[1], Linus Bengtsson[1]

[1] Flowminder Foundation, [2] Formerly Flowminder Foundation, [3] Ghana Statistical Service

[a] Equal significant contribution to the set up of the solution

* Corresponding author. Email: james.harrison@flowminder.org

Abstract

Introduction

During the COVID-19 pandemic, many countries imposed mobility restrictions to reduce transmission and control the spread of the disease. Anonymised and aggregated mobile operator data, such as call detail records (CDRs), can provide near-real time insights into population mobility with high spatial and temporal resolution, across a whole country. Such data are useful for monitoring the impact of these restrictions on mobility and therefore help assess their potential impact on the spread of COVID-19. Flowminder supported the global response to the COVID-19 pandemic by working with mobile network operators, governments, and development actors, throughout 2020-2021 to rapidly generate CDR-derived insights related to population distribution and mobility changes caused by the pandemic and the associated mobility restrictions in seven countries (Curaçao, the Democratic Republic of the Congo (DRC), Ghana, Haiti, Namibia, Papua New Guinea, and Sierra Leone). Here, we detail the key analyses conducted to assess the impact of the response to the pandemic and mobility restrictions on mobility and the key learnings on rapid analysis of fast varying mobility indicators in multiple countries.

Data and Methods

We used anonymised CDR aggregates from 7 countries to produce indicators determining how mobility patterns changed during the COVID-19 pandemic and, in particular, in response to the announcement, implementation and lifting of mobility restrictions by governments.

We introduced a systematic classification at the time of short- and longer-term mobility indicators. Short-term indicators included: presence, the number of unique subscribers observed in an area (i.e. made a call routed by a cell site in that area) during a given period of time (e.g. hour, day); trips, the number of unique subscribers observed in an area having been previously observed in another area during a given time period (e.g. day); and the average number of areas visited by a subscribers each day. Longer-term indicators included residents, the number of subscribers whose home location (i.e. the area containing the cell site that most commonly routed a subscribers last call of the day) is assigned to an area; and relocations, the number of subscribers whose home location changed to an area from another area in a given period of time (e.g. week, month).

Using these indicators, Flowminder was able to generate insights into how mobility and population distributions changed, relative to a pre-pandemic baseline, following different government interventions. These included: how much has travel decreased; to what extent are people staying home more; how the distribution of population between different areas (e.g. urban vs rural) has changed; and how much has population mixing reduced (or increased) as a result of the measures.

Results

The mobility indicators derived from CDR data show a substantial, sharp reduction in mobility immediately following government restrictions in all seven countries (e.g. Fig.2), which may have helped control the spread of the disease as intended. However, the lifting of restrictions resulted in slow recovery of mobility towards the pre-pandemic baseline, suggesting a longer term impact of mobility restrictions on the economy. For example, in Namibia, the population in the core economic areas around Windhoek similarly remained below the pre-pandemic baseline beyond September 2020.

Mobility restrictions may also have unintended impacts which facilitated the spread of the disease. In both Haiti and Ghana we observed people relocating from urban to rural areas particularly in between the announcement and implementation of restrictions. We also note the redistribution in Haiti (Fig. 3) was similar to that usually observed around the Christmas period, suggesting that people returned to family homes outside of the cities. In Namibia, there was also a sharp increase in mobility following the announcement of the restrictions, prior to their implementation.

Discussion

Our multi-country COVID-19 response was significant in the development of Flowminder work on population mobility. It allowed us to test and implement a new way of analysing mobile operator data (through assisting mobile operators with running our code) and highlighted the need for a systematic categorisation of mobility indicators to support analysts and decision-makers to access relevant and impactful insights and compare them across countries and contexts. Also the large amount of data we analysed enabled us to explore limitations in our methodologies, particularly the impact of changes in phone use behaviour on the mobility information extracted from CDRs. For example, changes to tariffs in Namibia led to a reduction in mobility indicators, in fact mainly driven by a reduction in phone usage. Disentangling phone usage effects from mobility is of on-going development but is already more robust as a result of this work. This limitation is also more problematic when assessing routine mobility rather than large scale unusual changes such as those triggered by the COVID-19 mobility restrictions.

In conclusion, anonymised CDR aggregates can provide useful insights into both the short- and long-term impacts of mobility restrictions implemented by governments in LMICs. These can support decision-makers to monitor and evaluate the effectiveness of interventions to limit mobility, and therefore the spread of infectious disease, and to assess recovery of mobility after restrictions are lifted and the associated economic impact on economic activity.

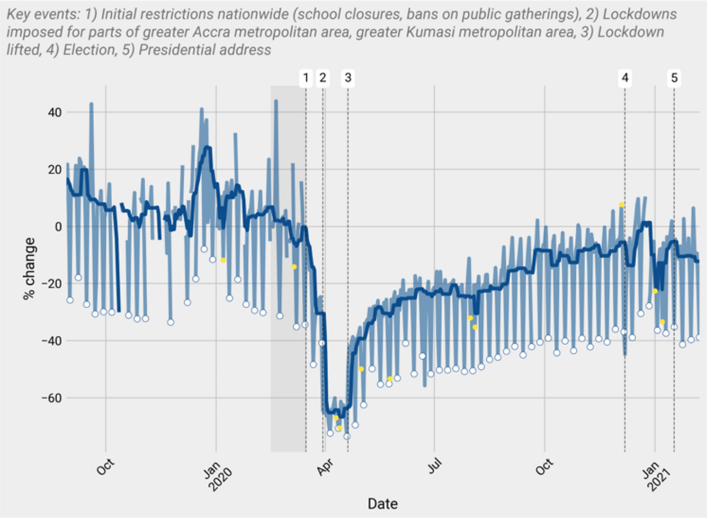


Figure 1. **Ghana. Percentage change in the number of trips between any two districts in Greater Accra**, each day, relative to the baseline value, overlaid with a seven day rolling average. Yellow and white dots denote public holidays and weekends, respectively, and the baseline period is indicated by the shaded region. The restrictions led to an immediate drop in mobility, followed by a slow recovery continuing long after the restrictions were lifted.

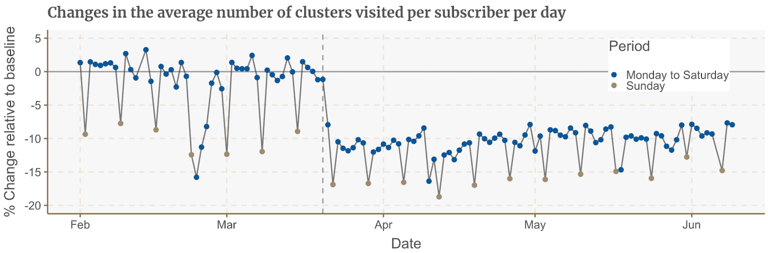


Figure 2. **Haiti. Percentage change in the mean number of locations (clusters of cell sites) visited per subscriber, each day,** relative to a pre-pandemic baseline period. The dotted line represents the introduction of mobility restrictions in Haiti. Covid-19 restrictions lead to an immediate drop in mobility, comparable to that of a normal Sunday.

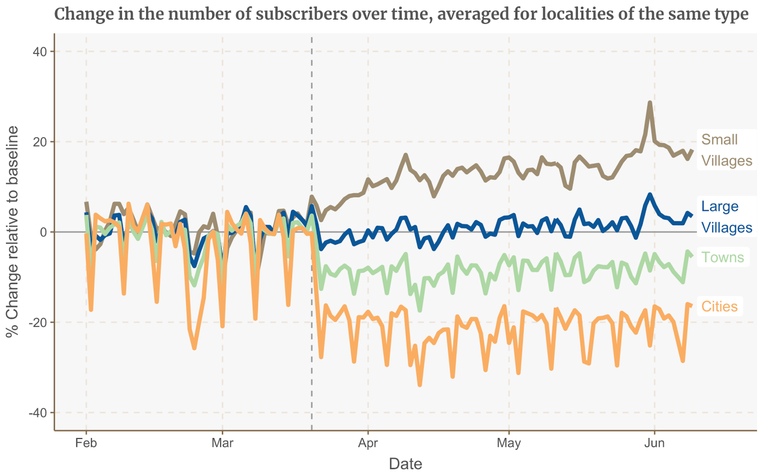


Figure 3. **Haiti. Average percentage change in the number of subscribers in localities of different levels of urbanisation** (cities, towns, large villages, small villages), each week. The dotted line represents the introduction of mobility restrictions in Haiti. Restrictions lead to a redistribution of the population, from the cities to the villages, similar to what is observed in Haiti during the end of year period.

COVID-19 is linked to changes in the time-space dimension of human mobility

[1,2]Clodomir Santana, [1,3]Federico Botta, [1]Hugo Barbosa, [4]Filippo Privitera, [1,3,5]Ronaldo Menezes, [3,6]Riccardo Di Clemente*

(1) University of Exeter, Computer Science Department, Exeter, United Kingdom, (2) Institute of History, Polish Academy of Sciences, Warsaw, poland, (3) The Alan Turing Institute, London, United Kingdom, (4) Spectus, New York, United States, (5) Federal University of Ceará, Fortaleza, Brazil, (6) Complex Connections Lab, Network Science Institute, Northeastern University London, London, United Kingdom

Extended Abstract

Society produces digital records of, for example, the places we visit, the products we buy, and the people we call. These digital records proved valuable in studying different aspects of human behaviour [1]. Here, we leverage Location Base Service (LBS) data from mobile phone users to study how citizen mobility patterns have been affected during the COVID-19 pandemic in the UK. Assessing the effects of the mobility restriction policies on daily routines relies on investigating the relationship between space and time-based population mobility patterns. We employ the radius of gyration [2] to gauge the span of the urban movement (spatial dimension). We also define mobility synchronisation as a time metric that quantifies the co-temporal occurrence of the daily mobility motifs [4] – i.e. leave/return home from work happens periodically at the same time (temporal dimension). Combining these space and time metrics, we can estimate the effect of the mobility restrictions on the population. Using the radius of gyration (Fig. 1 A orange), we could identify that the effect of the first lockdown was more significant than the others in changing the spatial characteristics of citizens' movement. Among the reasons that could lead to this result, we can mention more strict policies adopted in the first lockdown and the lockdown duration [5, 6]. However, further investigation is needed to obtain more evidence to support these hypotheses. In contrast to the spatial dimension of mobility, the results indicate that the temporal (Fig. 1 A green) one was more impacted during the second lockdown when more flexible mobility restriction policies were enforced. After the first lockdown, we argue that people who could not work from home were allowed to leave home and work in the office as long as they respected social distancing rules [5, 3]. The results also indicate that the two lockdowns affected the synchronisation of people's movement differently (Fig. 1 B). The mobility synchronisation displays a recovery latency compared to the gyration radius. Furthermore, how we respond to mobility restriction measures is interwoven with the characteristics of the geographical space, such as income groups, economic activities, and population density. In particular, we focus on disentangling how the population density in rural-urban areas and the different socio-economic groups have adjusted their routine to comply with the mobility restrictions imposed. We observed that the radius reduction was slightly more significant in rural areas than the urban ones. In contrast, the decrease in synchronisation levels was more notable in urban areas than rural ones. We noticed that high-income groups displayed a more considerable reduction in the radius and synchronisation than the low-income groups. We also studied the differences concerning the duration and the type of trips. Fig. 1 B illustrates that high-income groups have the most reduction in the duration of their work-related trips. While Fig. 1 C depicts that rural areas presented the most significant increase in park trips compared to the baseline year of 2019. In summary, the analysis of the spatial dimension of human mobility coupled with the insights from the study of the time dimension allows us to characterise the impact of *stay-at-home* policies on the population of different areas/socioeconomics. These differences suggest that each group experiences, in a particular way, the emergence of asynchronous mobility patterns primarily due to mobility restriction policies and the ascension of new habits (e.g. home office and home education).

*corresponding author: r.di-clemente@exeter.ac.uk

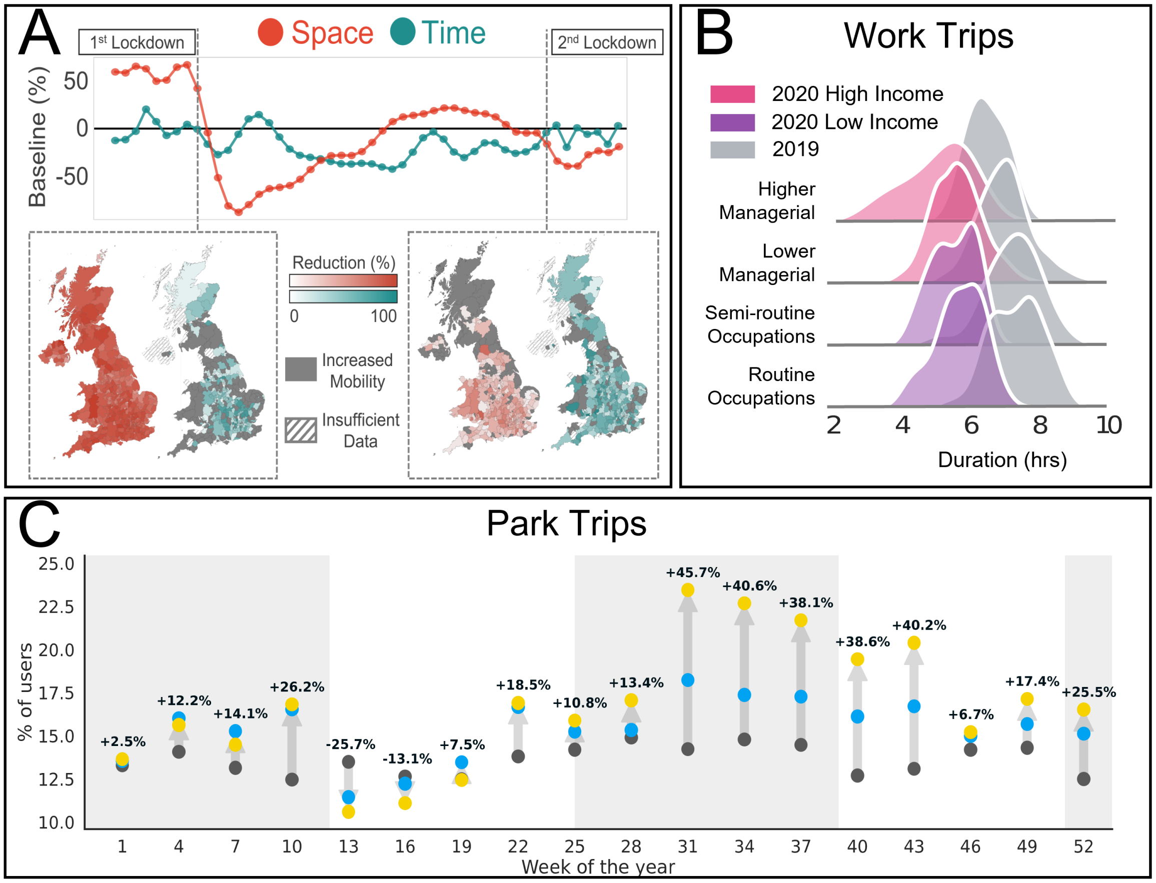


Figure 1: Changes in mobility patterns during the COVID-19 pandemic. Panel A illustrates the mobility difference over time and for the first two English national lockdowns measured with mobility synchronisation (time) and the radius of gyration metrics (space). Panel B looks at the mobility related to the duration of work-related trips before (2019) and during the pandemic (2020). Using the NS-SEC classification, we show the different behaviour of income and socioeconomic groups. Lastly, Panel C depicts the differences in the number of trips to green spaces such as parks compared to the baseline year of 2019. Note that in this case, we group the local authorities according to their level of urbanisation.

References

[1] Caroline O Buckee et al. "Aggregated mobility data could help fight COVID-19". In: *Science* 368.6487 (2020), pp. 145–146.

[2] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns". In: *nature* 453.7196 (2008), pp. 779–782.

[3] World Health Organization et al. *Global surveillance for COVID-19 caused by human infection with COVID-19 virus: interim guidance, 20 March 2020*. Tech. rep. World Health Organization, 2020.

[4] Clodomir Santana et al. "Analysis of human mobility in the UK during the COVID-19 pandemic". In: *preprint@ github* (2020).

[5] Gabriel Scally, Bobbie Jacobson, and Kamran Abbasi. *The UK's public health response to covid-19*. 2020.

[6] Legislation. gov. uk. "The Health Protection (Coronavirus, Restrictions)(England) Regulations 2020". In: *Queen's Printer of Acts of Parliament* (2020). URL: `https://www.legislation.gov.uk/uksi/2020/350/contents/made`.

Visitation patterns of COVID-19: POIs interactions within urban structure

Robert Eyre[1], **Lavinia Rossi Mori**[2,3,+], **Antonio Desiderio**[2,3,+], **Filippo Simini**[4], **and Riccardo Di Clemente**[5,6,*]

[1]University of Bristol, Department of Engineering Mathematics, Bristol BS8 1UB, United Kingdom
[2]Physics Department and INFN, Tor Vergata University of Rome, 00133 Rome, Italy.
[3]Centro Ricerche Enrico Fermi, 00184 Rome, Italy.
[4]Argonne Leadership Computing Facility Argonne National Laboratory Lemont, IL 60439 United States.
[5]Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, GBR.
[6]The Alan Turing Institute, London, NW12DB, GBR.
[*]riccardo.diclemente@nulondon.ac.uk

The COVID-19 pandemic has been an unprecedented global crisis, leading to the disruption of human mobility and the rhythm of urban life. The behavioral changes we undergo reshape our interactions with the urban environment (Yabe, Nat. Comm., 2023, DOI:10.1038/s41467-023-37913-y), and raising new challenges about resilience and recovery of the urban interactions and economy. A city thrives when its Points of Interest (POIs) - restaurants, museums, cafes, etc., - synchronize to create a vibrant and captivating urban landscape (Glaeser, New York: Penguin, 2012). Human mobility acts as the lifeblood of a city, fueling the interactions and pulse among its POIs. The pandemic-induced lockdown has profoundly impacted our everyday behavior, evident in less time spent in amenities, simpler routines, and increased predictability (Lucchini, Sci. rep., 2021, DOI:10.1038/s41598-021-04139-1). These behavioural changes are not just individual responses; they have macro-level implications for our urban landscapes (Santana, ArXiv, 2023,DOI:10.48550/arXiv.2201.06527 especially for the POIs that heavily rely on human visiting patterns. What are the topological factors shaping the visitation patterns of POIs during post-COVID recovery? Can we distinguish resilience or vulnerability in terms of POI diversity and utilization within urban areas?

We aim to uncover the common drivers that shape the recovery visitation patterns of POIs in different urban landscapes. We strive to understand how in a particular area the diversity of available POIs, the strength of community engagement, and the presence of infrastructure contribute to the reshaping of activities during the recovery phase after the COVID-19 restrictions have been lifted. By examining the trajectories of POIs and analyzing shifts in citizen visitation patterns, we aim to unravel the complex interplay between POI location and their interactions, all at an aggregated level. We leverage the information from GPS data provided from the location services company Spectus.ai from September 2019 to September 2020. To answer our questions we first need to deduce various visitation patterns to POIs across Italy, and then to predict changes in visitor counts, specifically during two periods—immediately after the initial lockdown in March 2020 and the following summer. We consider the GPS traces generated from 181k unique user over the whole of Italy during the 12 months, from which we extracted 52 million stops data. To preserve citizen privacy, the users pings were then aggregated into hourly visitor count data at the stop level. By applying the DBSCAN algorithm to all unique stopping locations, we identified popular location hotspots across Italy, which were then enriched with OpenStreetMap data (POIs). This integration allowed us to assign each hotspot to a specific POI, such as restaurants, museums, or stations.The resulting data set (60k POIs) provides an anonymized but detailed view of visitation patterns at a city-wide scale. We define the hourly visitation profile of a POI to be the normalised time series of weekly visitor counts, along with normalized counts of visitors during average weekdays and weekends. We cluster these time series of visitors to each POI in Italy such that we can identify motifs of 'visitation profiles' (whether they are seasonal, non-seasonal, mainly visited on weekends, mainly visited on weekdays etc) and then groups of POIs between different regions. We employ the unsupervised machine learning technique known as Self-Organizing Map (SOM) to cluster the time series data Fig. 1 (**A**). The main objective of a SOM is to transform unstructured input data into a simplified representation, by mapping them onto a lower-dimensional grid of neurons. By training a grid of neurons to fire in localized areas for different types of input, self-organizing maps enable the identification of clusters of neurons that exhibit similar behavior, gradually adjusting the weight vectors to converge towards the underlying data distribution. In order to cluster the grid space, for each cell we follow a path to the nearest neighbour that is closer on average to its own neighbours, and keep following until we find a grid cell is closer on average to its neighbours than its neighbours, allowing for the emergence of clusters without needed to specific the number of clusters. The neighborhood size affects the cluster formation in the SOM: when it is larger, neighboring neurons have a greater influence on each other, leading to more cohesive and broader clusters. In the pre-lockdown time series, we discover 6 different visitation patterns, Fig. 1 (**A**). These POIs'clusters capture unique modes of activity (feature from

OpenStreetMap) and are general across different cities. With the same analysis in post-lockdown phase only three visitation patterns emerge (Bonaccorsi, Sci. rep., 2021, DOI:10.1038/s41598-021-99548-7): struggling businesses unable to reopen (L-shaped), businesses that recover following the lockdown (U-shaped), and those that recover until the summer season when they experience a dip in visitor numbers (W shaped), Fig. 1 (**A**).

We initially apply Gradient Boosting to predict shifts in POI recovery profiles, classifying them as either L-shaped or U-shaped. This model uses pre-lockdown features and context-specific variables that describe each POI's environment, such as density and variability. This approach yields a high prediction accuracy of 87By computing SHapely Additive ExPlanations (SHAP), we explore why the model predicts these outcomes, Fig. 1 (**B**-left) to capture the reasons why certain businesses decrease their visiting activity after the lockdown. We show that changes to visitor counts and visitations patterns can be explained by features unique to the POI itself (pre-lockdown visitation profile), but also by features about the surrounding of the POI. We also consider a reduced model, excluding the pre-lockdown features, and it still yields a high prediction accuracy of 84%. The reduced model underlies that the visitation profile of a POI can be inferred by its surrounding context (Fig. 1 (**B**)), highlighting the dynamic nature of cities.

As urban infrastructure guides how people navigate between POIs and the visitation profile of POI is influenced by its location, then the topology of the street network encompasses the valuable information about the interactions among POIs (Su, EPJ Data Sci., 2022, DOI:10.1140/epjds/s13688-022-00355-5) . We start by extracting a weighted network of POIs from the street network. The weight between two POIs is determined based on the shortest path length between the streets where the POIs are situated. The weight is calculated as the sum of the lengths of the streets that are crossed. Next, we utilize the DBSCAN algorithm along with the weighted distance matrix to cluster the POIs, Fig. 1 (**C**). This clustering is based on the density distribution of the POIs at different levels of reachability. By increasing the radius that defines neighboring points, we can identify a specific scale where isolated points no longer exist, where we cover the entire city. To evaluate the homogeneity of the clusters, we compute the average post-lockdown profile of the cluster; which can be categorized as either L-shaped or U-shaped. Remarkably, at the identified scale, we observe that the clusters predominantly fall into one of these two categories (with average homogeneity 0.5).

We have disclosed that both individual features of POIs and the broader urban structure significantly influence visitation patterns during the recovery phase after a disaster, not using individual trajectory data. By leveraging this framework, we can identify the extent of the area damaged by the natural disaster and establish their link with the city landscape. By analyzing the interactions among points of interest and combining them with economic data, we can derive new metrics that will inform resilient urban planning strategies for cities, enabling cities to thrive and recover effectively in the aftermath of future major disruptions and economic shift (Fan, PNAS Nexus, 2023, DOI:10.1093/pnasnexus/pgad077).

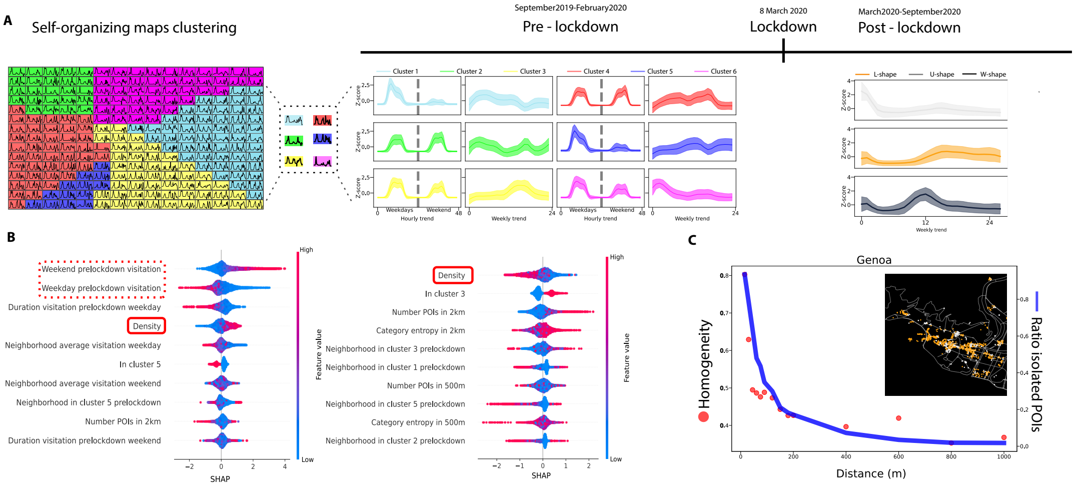


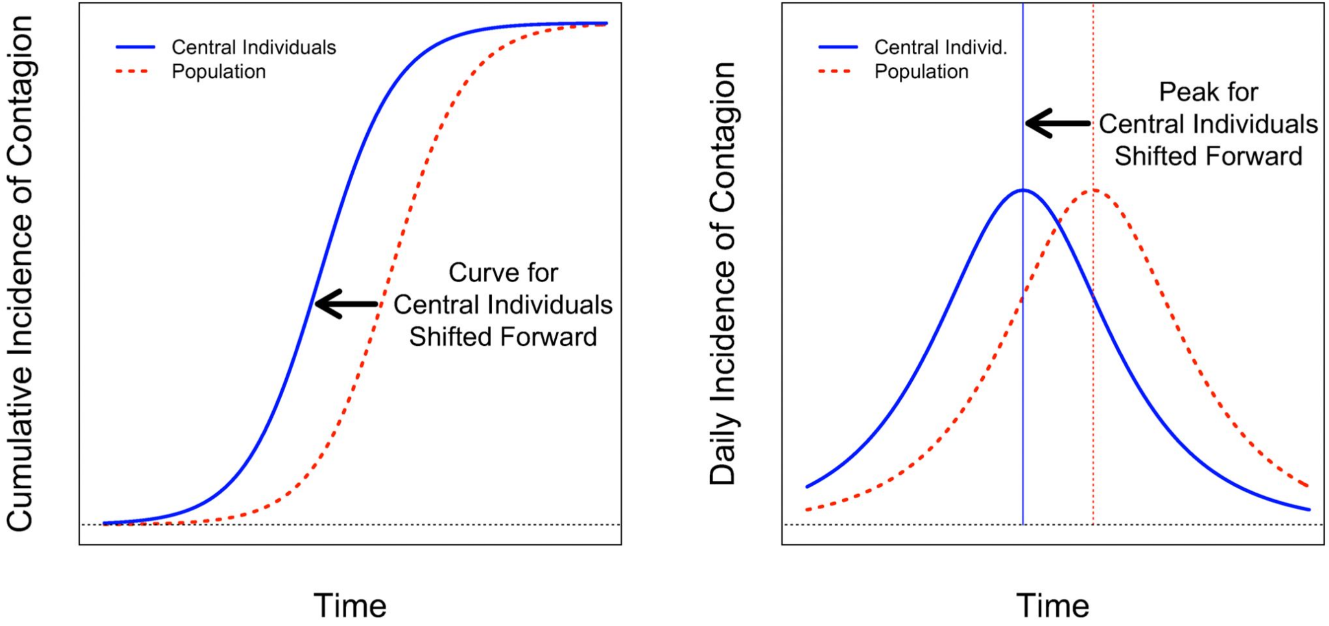
Figure 1. Visitation Patterns and Human Interactions within the Urban Network. A) Self-Organizing Maps and Time Series Clustering. On the left, the self-organizing maps for the pre-COVID clusters. In the center the average profile of each pre-COVID cluster, while on the right the average profile of the post-COVID phase. **B)** XGBoost and SHAP Explainability. On the left the key features for the full-model, ranked by the SHAP values, while on the right the key features for the restricted model. **C)** Urban Area Survival and DBSCAN in Genoa: The plot displays the homogeneity (red dots) and the percentage of of isolated POIs (blue line) of the emergent clusters found by DBSCAN at the given scale. In the inset the city map of Genoa, where the resilient POIs colored in orange and the other in grey.

Social Media Sensors to Detect Early Warnings of Influenza at Scale

David Martín-Corral, Manuel G. Herranz,
Manuel Cebrián, Esteban Moro

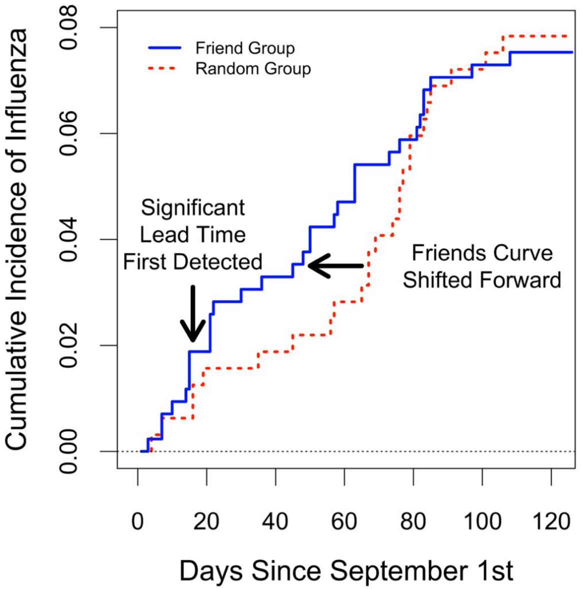
uc3m | Universidad Carlos III de Madrid

The spreading of information on networks tell us that central nodes are the first to be aware of a gossip or a virus.

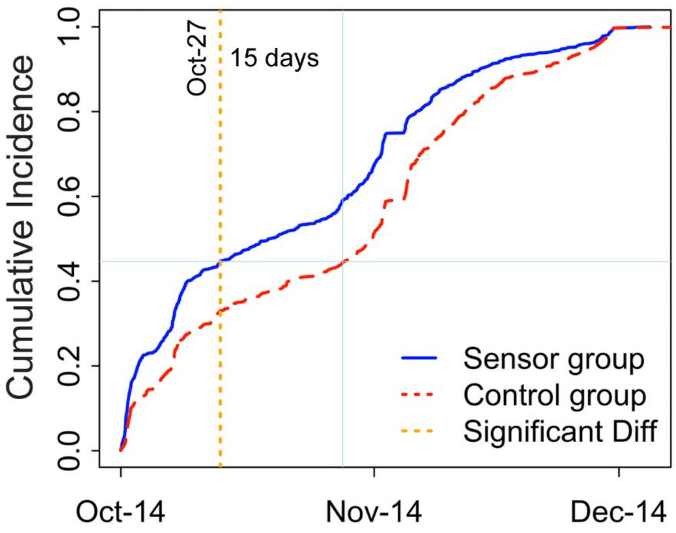


Source: Christakis, Nicholas A., and James H. Fowler. "Social network sensors for early detection of contagious outbreaks." PloS one 5.9 (2010): e12948.

Central nodes (sensors) have been used in biological and informational outbreaks but never used together.



Christakis, Nicholas A., and James H. Fowler. "Social network sensors for early detection of contagious outbreaks." PloS one 5.9 (2010): e12948.

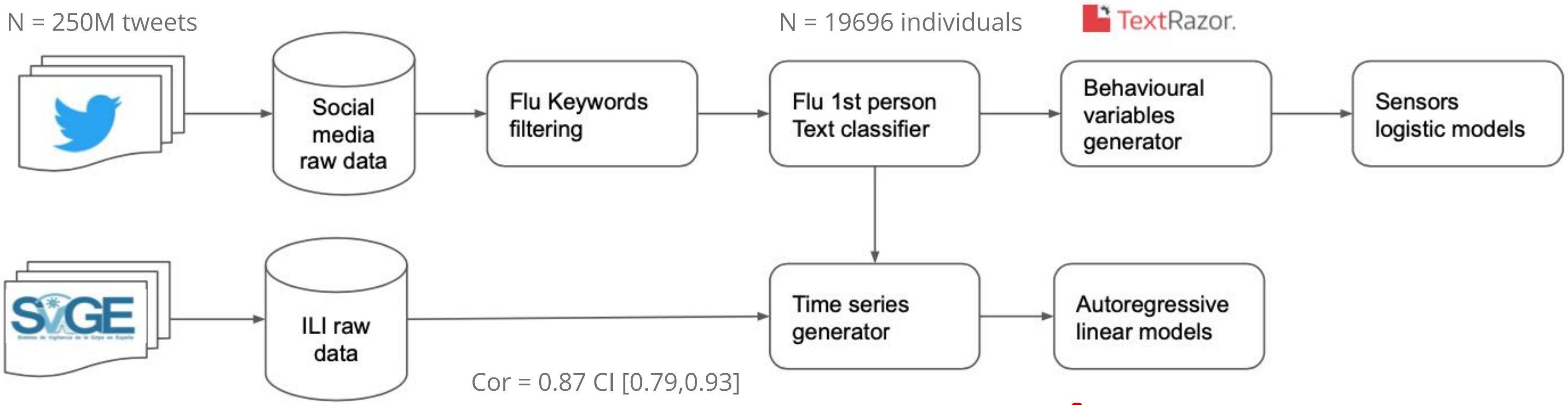


Garcia-Herranz, Manuel, et al. "Using friends as sensors to detect global-scale contagious outbreaks." PloS one 9.4 (2014): e92413.

Research questions

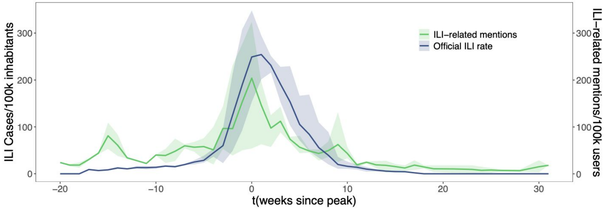
- Do **highly-connected users in social media** (social sensors) **mirror highly-connect individuals** (biological sensors)?
  - Online networks mimic offline contacts' connections, similarity, and spatial organization.
- How well do **social sensors from digital platforms capture biological viral outbreaks**?
  - If social sensors = biological sensors, then we can use them as early warning signals.
- Can we **characterize social sensors** based on their behaviours?
  - Beyond network centrality, are sensors defined by other type of features?

We used social media traces for indirect observation of an ILI epidemic, extract centrality metrics and define sensors.



Time period = Dec 2012 to April 2015
Geography = Spain mainland

Cor = 0.87 CI [0.79,0.93]



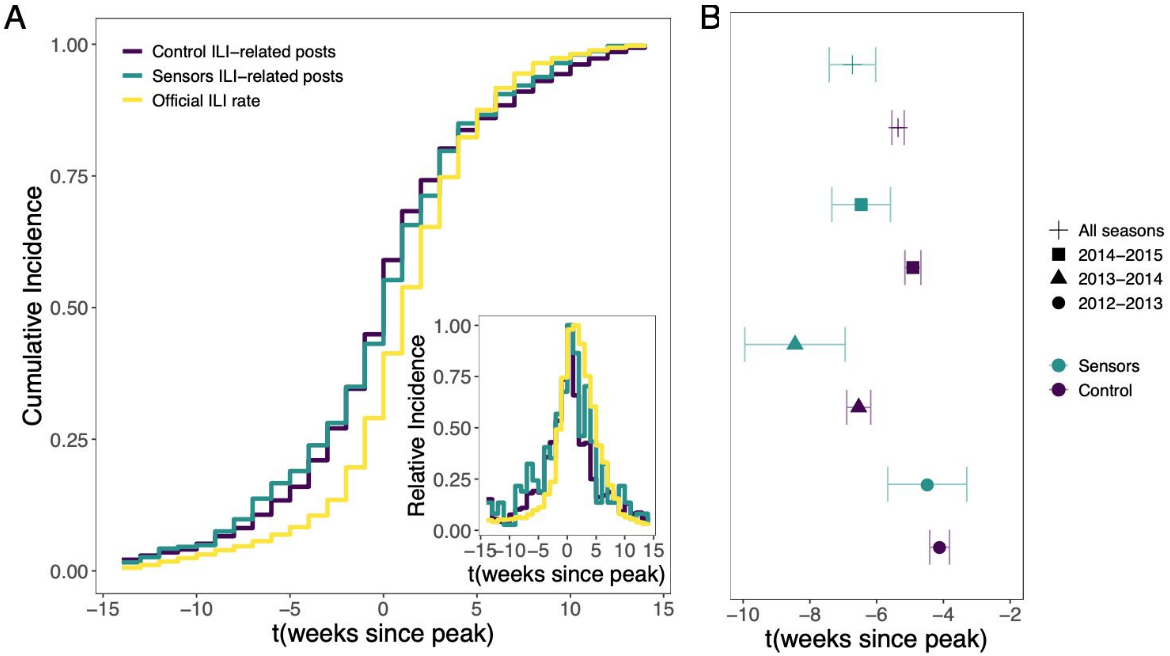
Sensors

We define a **sensors** as individuals with an **out-degree higher than 1000** followees.

$$d_{out,i} > 1,000.$$

Source: Martin-Corral, D., García-Herranz, M., Cebrian, M., & Moro, E. (2022). Social Media Sensors to Detect Early Warnings of Influenza at Scale. medRxiv, 2022-11.

Sensors post 6.72 weeks before the peak and 1.37 weeks earlier than control group.



Statistically significance

$$\Delta t_C = \langle t_i^{post} - t^{peak} \rangle_{i \in C} = \text{-}5.35 \text{ (CI } [-5.54, -5.17])$$

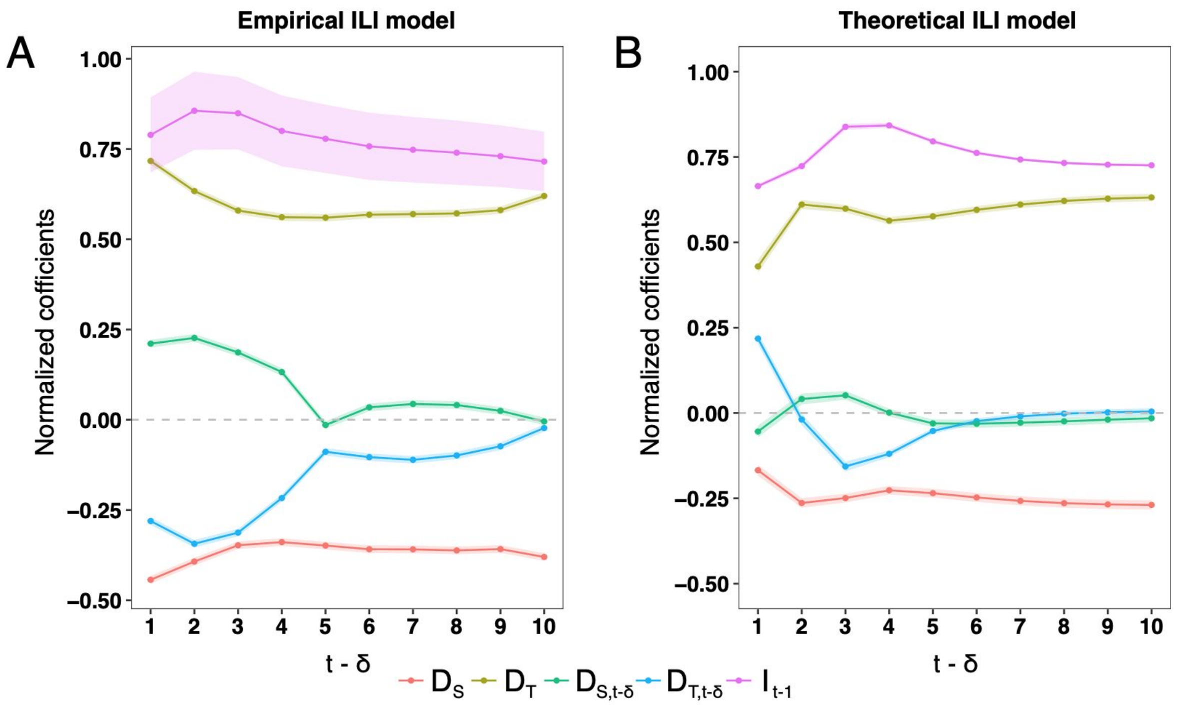
$$\Delta t_S = \langle t_i^{post} - t^{peak} \rangle_{i \in S} = \text{-}6.72 \text{ (CI } [-7.42, -6.02])$$

$$\Delta t_S - \Delta t_C = -1.37 \text{ (CI } [-2.08, -0.64]$$

Mean difference between control and sensors are statistically significant for all seasons, except 2012-2013.

Source: Martin-Corral, D., García-Herranz, M., Cebrian, M., & Moro, E. (2022). Social Media Sensors to Detect Early Warnings of Influenza at Scale. medRxiv, 2022-11.

We used social out-degrees to predict empirical ILI cases and we validated our results against a theoretical network.



Centrality metrics

$$D_{T,t} = \sum_{i\in\Omega_t} d_{out,i,t} \qquad D_{S,t} = \sum_{i\in\Omega_t^*} d_{out,i,t}$$

Autoregressive model

$$I_t = \beta_0 + \beta_1 I_{t-1} + \sum_{\delta\geq 0}(\alpha_\delta D_{T,t-\delta} + \gamma_\delta D_{S,t-\delta}) + \epsilon_t.$$

R2 Explainability

$$R^2_{adj,I} = 0.8552$$

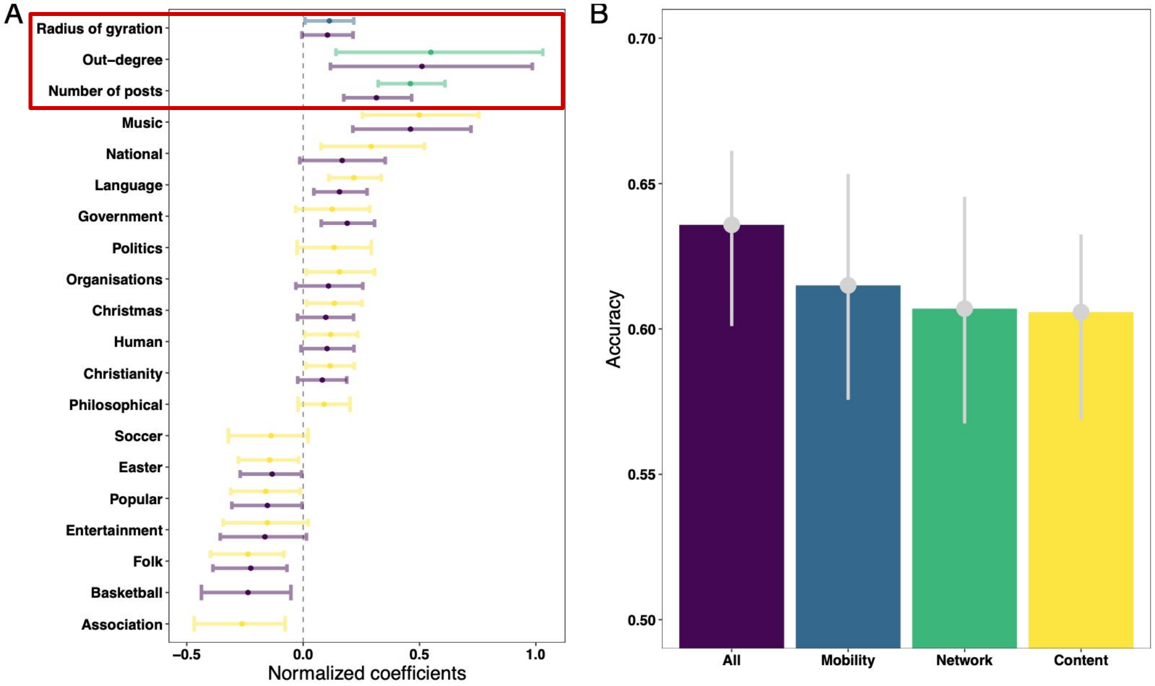
$$R^2_{adj,I,T,S} = 0.924$$

Theoretical ILI model

We used an ABM to simulate a SIR epidemic spread of ILI upon a theoretical network generated by the Barabasi-Albert model.

Source: Martin-Corral, D., García-Herranz, M., Cebrian, M., & Moro, E. (2022). Social Media Sensors to Detect Early Warnings of Influenza at Scale. medRxiv, 2022-11.

Beyond network features, can we use other behavioural features to predict who talks about ILI earlier (better sensors)?



Behavioural features

M = Mobility

N = Network

C = Content

Logistic regression model

$$\Pr(i \in \Omega^*) = \text{logit}^{-1}[\beta_0 + \sum_l \alpha_l M_l^i + \sum_n \beta_n N_n^i + \sum_m \gamma_m C_m^i]$$

We predict the **probability of an individual (sensor) to talk about ILI before it does.**

Source: Martin-Corral, D., García-Herranz, M., Cebrian, M., & Moro, E. (2022). Social Media Sensors to Detect Early Warnings of Influenza at Scale. medRxiv, 2022-11.

Social sensors mimic biological sensors, can be used as early warning signals and are defined beyond network centrality.

- We have developed a **methodology to process novel data streams** from social media data that embedded human health-related behaviors.

- Our **method exploits the user heterogeneity underneath social media sites** to detect more efficiently earlier outbreaks from disease-informational epidemics.

- We have proved that **epidemic social sensors from social media data can be built** for more efficient early warning epidemiological systems.

Our research have two clear paths that could expand knowledge further about epidemic sensors.

- Generalized to other **regions, epidemics and platforms**.
- Further research on **super-sensors personality and behavioural traits**.

Does mobility data help in regional COVID-19 case predictions?

Saad Mohammad Abrar, Naman Awasthi, Daniel Smolyak and Vanessa Frías-Martínez

University of Maryland
{sabrar, nawasthi, dsmolyak, vfrias}@umd.edu

The COVID-19 pandemic has mainstreamed human mobility data into the public domain, and beyond academic networks. During the early stages of the pandemic, the importance of limiting mobility to contain the epidemic became evident, with cities, states and countries taking various non-pharmacological interventions (NPIs) focused on mobility such as national lockdowns or work-from-home approaches [11, 4]. To evaluate the effect of these interventions, public health experts, the CDC, city departments and journalists explored the use of mobility data which, at the time, was made open and freely available. Companies like Apple, Google, SafeGraph or Descartes shared different types of aggregated mobility datasets to characterize behaviors such as the volume of visits to specific places (e.g., schools, workplaces or restaurants), the volume of trips between regions (e.g., trips between two counties), or the volume of trips by type of transportation (e.g., driving vs. public transit).

Beyond understanding the impact of mobility reduction policies, the increased access to mobility data sources has also supported research on regional COVID-19 case prediction models, with the assumption that how people moved within a region in the past could potentially provide additional information about how people get infected in the future. COVID-19 case prediction models focus on providing regional-level estimates for future number of cases in the short- and long-term via lookahead analysis performance *i.e.,* measuring region-level prediction performance for various temporal windows such as daily, weekly or monthly [7]. For example, researchers have shown that SafeGraph data can help predict weekly COVID-19 cases at the county level in the US, providing higher accuracy when compared to non-mobility baselines [10]. There exist a wide variety of models to predict regional COVID-19 cases including epidemiological [6, 1], machine learning [8, 5] and statistical models [7, 5]. In this paper, we focus on statistical models (linear regression and ARIMA) because we are interested in the deployment of models that are interpretable by decision makers, rather than implementing black-box predictive approaches that are harder to explain [12].

Nevertheless, there are several gaps in the current state of the art in regional COVID-19 case prediction using mobility data. First, performance results - measured as RMSE or correlation between actual and predicted regional COVID-19 cases - are reported as averages across regions, masking individual region-level performance, which is critical to inform local interventions and policies [7]. For example, past research has shown that mobility data enhances COVID-19 case predictions, on average, across counties in the US; however, that average might be masking counties for which it did not work [8, 3]. Second, performance results are often times not compared against non-mobility baselines, making it hard to measure the effectiveness of adding mobility data to the prediction model [8, 3, 14]. Third, prior work has shown that mobility data might suffer from sampling bias whereby certain demographic groups *e.g.,* Black, elder and low-income individuals can be under-represented in the data due to lower smartphone and cell phone ownership rates [13, 2]. Nevertheless, prior work focused on building COVID-19 case prediction models tends to ignore the bias present in the mobility data, which in turn, might affect the performance of regional COVID-19 case prediction models depending on the population of that region [7, 8, 9]. Fourth, current approaches tend to provide narrow evaluations, focused on a few models, or on one or a few mobility datasets, with little research broadly looking into the impact of different prediction models, mobility datasets, and training approaches that use more or less data, on model performance. Given (1) the high cost of acquiring human mobility data for COVID-19 prediction purposes, now that it is no longer freely accessible, and (2) that COVID-19 case predictions are going to be used to assess NPIs such as mobility reduction, or vaccine distribution at the local level, we posit that it is critical to understand the conditions under which mobility data helps (or not) at the individual regional level so that it can adequately inform local decision-making.

In this paper, we aim to analyze the conditions under which human mobility data provides an enhancement over individual regional COVID-19 case prediction models that do not use mobility as a source of information. Our main objective is to inform regional decision makers about the potential of region-level COVID-19 case prediction models that use mobility data, which we posit should be well understood given the high cost of human mobility data. The main contributions of this paper are:

- Focusing on US counties, we evaluate the number of counties that benefit from adding mobility data, and quantify the improvements. Our analyses show that, at most, 60% of counties improve their performance over non-mobility baselines; and that those improvements are modest, happening mostly for longer-term predictions. Looking into the counties that benefit from adding mobility data, 50% of those counties show modest correlation improvements of at most 0.1 and 25% show correlation improvements of at most 0.3.

- We present and discuss an approach to assess whether mobility data bias - characterized by demographic and socio-economic characteristics of each county - might explain the differences in the performance of COVID-19 prediction models across counties. We show that correlation improvements were lower for counties with higher Black, Hispanic, and other non-White populations as well as low-income and rural populations, pointing to potential bias in the mobility data negatively impacting performance.

- We analyze whether the differences in the performance of mobility-based models over non-mobility baselines vary depending on the mobility datasets; the predictive model; or the training approach. Our results reveal that the improvements brought about by mobility data are similar across mobility datasets, albeit with slightly better values for Apple and SafeGraph; that linear regressions are associated with larger improvements; and that the training approach might also affect the scale of the improvements.

References

[1] S. Chang, M. L. Wilson, B. Lewis, Z. Mehrab, K. K. Dudakiya, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, et al. Supporting covid-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2632–2642, 2021.

[2] A. Coston, N. Guha, D. Ouyang, L. Lu, A. Chouldechova, and D. E. Ho. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for covid-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, 2021.

[3] T. T. da Silva, R. Francisquini, and M. C. Nascimento. Meteorological and human mobility data on predicting covid-19 cases by a novel hybrid decomposition method with anomaly detection analysis: A case study in the capitals of brazil. *Expert Systems with Applications*, 182:115190, 2021.

[4] L. Feng, T. Zhang, Q. Wang, Y. Xie, Z. Peng, J. Zheng, Y. Qin, M. Zhang, S. Lai, D. Wang, et al. Impact of covid-19 outbreaks and interventions on influenza in china and the united states. *Nature communications*, 12(1):1–8, 2021.

[5] S. García-Cremades, J. Morales-García, R. Hernández-Sanjaime, R. Martínez-España, A. Bueno-Crespo, E. Hernández-Orallo, J. J. López-Espín, and J. M. Cecilia. Improving prediction of covid-19 evolution by fusing epidemiological and mobility data. *Scientific Reports*, 11(1):1–16, 2021.

[6] X. Hou, S. Gao, Q. Li, Y. Kang, N. Chen, K. Chen, J. Rao, J. S. Ellenberg, and J. A. Patz. Intracounty modeling of covid-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences*, 118(24):e2020524118, 2021.

[7] C. Ilin, S. Annan-Phan, X. H. Tai, S. Mehra, S. Hsiang, and J. E. Blumenstock. Public mobility data enables covid-19 forecasting and management at local and global scales. *Scientific reports*, 11(1):1–11, 2021.

[8] C.-P. Kuo and J. S. Fu. Evaluating the impact of mobility on covid-19 pandemic with machine learning hybrid predictions. *Science of The Total Environment*, 758:144151, 2021.

[9] Z. Mehrab, A. Adiga, M. V. Marathe, S. Venkatramanan, and S. Swarup. Evaluating the utility of high-resolution proximity metrics in predicting the spread of covid-19. *ACM Transactions on Spatial Systems and Algorithms*, 2021.

[10] B. Nikparvar, M. Rahman, F. Hatami, J.-C. Thill, et al. Spatio-temporal prediction of the covid-19 pandemic in us counties: modeling with a deep lstm neural network. *Scientific reports*, 11(1):1–12, 2021.

[11] N. Perra. Non-pharmaceutical interventions during the covid-19 pandemic: A review. *Physics Reports*, 913:1–52, 2021.

[12] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[13] F. Schlosser, V. Sekara, D. Brockmann, and M. Garcia-Herranz. Biases in human mobility data impact epidemic modeling. *arXiv preprint arXiv:2112.12521*, 2021.

[14] L. Wang, X. Ben, A. Adiga, A. Sadilek, A. Tendulkar, S. Venkatramanan, A. Vullikanti, G. Aggarwal, A. Talekar, J. Chen, et al. Using mobility data to understand and forecast covid19 dynamics. *medRxiv*, 2020.

Epidemic spreading dynamics under exploration and preferential return mobility

Alfonso de Miguel Arribas[1,2], Alberto Aleta[1,2], Yamir Moreno[1,2,3] and Esteban Moro[4,5]
[1] *Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018, Zaragoza, Spain.*
[2] *Department of Theoretical Physics, University of Zaragoza, 50018, Zaragoza, Spain.*
[3] *Centai Institute, Turin, Italy*
[4] *Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. and*
[5] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, Leganés, Spain.*

Within the epidemic modeling literature, the metapopulation approach is a powerful and versatile framework to explore and analyze spatial epidemics both theoretically and in a data-driven way. Metapopulation models present three key ingredients: 1) spatial structure, 2) mobility model, and 3) spreading dynamics. Regarding mobility, our main concern here, it has been typically assumed some kind of Markovian random walk-like model with little ground on the observed real human behavior.

There have been interesting and insightful discoveries in the last decade, though, that open new venues for the integration of more realistic mobility assumptions in the context of metapopulations. In Song et al. 2010 [1], based on an exhaustive analysis of empirical data captured by mobile-phone traces, two principles governing human trajectories are introduced: exploration and preferential return. Here, the EPR microscopic model for human mobility emerges. Later, in Pappalardo et al. 2015 [2], through both mobile phone and GPS data, the authors effectively unveil the distinction between two main mobility profiles: explorers and returners, and propose a more refined micro-mobility model where the exploration mechanism is based on the gravity law of mobility (d-EPR model). This feature of human mobility has been remarked as very relevant to different aspects of human dynamics, including the impact of epidemic spreading phenomena among others. However, to the best of our knowledge, little attention has been paid to this specific question in the literature. In this work, we contribute to filling this gap by doing an exhaustive exploration of the impact at the urban scale of an epidemic spreading under an exploration and preferential return mobility dynamics. In particular, we work with the standard compartmental SIR model, appropriate for influenza-like illnesses, and the d-EPR model.

This model of microscopic mobility (i.e. agent-based) conceives that at a given time step, every agent either explores a new location with probability $P_{\text{exp}} = \rho S^{-\gamma}$, or returns to an already visited location with $1-P_{\text{exp}}$. Here, $\rho \in [0,1]$ is a parameter that characterizes the mobility profile of the agent, being $\rho \to 0$ for an agent with a very recurrent movement, and $\rho \to 1$ on the other extreme, characterizing agents with a very high rate of location discovery. Then, S is the number of different visited locations at a certain time step, and γ is a parameter that characterizes the exploration decay with S. Typically, ρ is assumed identical for every agent, but here we introduce individual variability in the population and explore the dynamics with a homogeneous and a heterogeneous (beta) distribution. The exploration stage follows the gravity model as $p_{ll'} \propto A_{l'}/r_{ll'}^2$, where A is the so-called attractiveness of the location, and $r_{ll'}$ is the Euclidean distance between origin and destination. In the case of returning to an already explored location, the destination l' follows a preferential attachment selection rule based on the visit frequency $f_{l'}$ of location l' at that point in time. We assume homogeneous and unitary sojourn times for every agent in locations. Regarding the epidemic model, we assume that contagion events occur at locations under a well-mixing approach, where the generation of new cases follows a binomial process with a probability of contagion $p_{S\to I,l} = 1-(1-\beta/N_l)^{I_l}$. Here N_l is the population of location l at the time of the contagion event, I_l is the infected population in l at that time, and β is the transmission rate, related to the disease's basic reproduction number by $R_0 = \beta T_I$, being T_I the infectious period. We work mainly under a low transmissibility scenario and set $R_0 = 1.2$. As for the spatial structure, we worked with both synthetic and real urban systems. Here, we focus on the last case and build the attractiveness field from the aggregation of high-resolution real human mobility trajectories from mobile phone data, taking place in the Greater Boston Area.

We conducted extensive Monte Carlo simulations for this coupled d-EPR+SIR dynamics in a low-density scenario. To compare the results from the d-EPR model we propose a series of baseline scenarios: i) memoryless, where the non-Markovian component of the return is removed, ii) uniform, where the spatial variability of locations' attractiveness is removed, and iii) plain, with no memory and uniform attractiveness.

First, we look at the invasion process, which we define here as a secondary outbreak for the first time in any location other than the epicenter. In figure 1 we plot the invasion share and the invasion times per ρ group for the d-EPR and the proposed baselines. We also compare all this model against the null assumption where ρ plays no special role in the invasion and thus these are random events. Under the Gaussian distribution, the deviation from the null assumption is very weak, whereas clearly, in a more heterogeneous setting, explorers drive the invasion process. Interestingly, in terms of the invasion share, both the d-EPR model and the uniform baseline behave very similarly. Invasion times for the beta distribution also clearly reflect that explorers invade significantly faster.

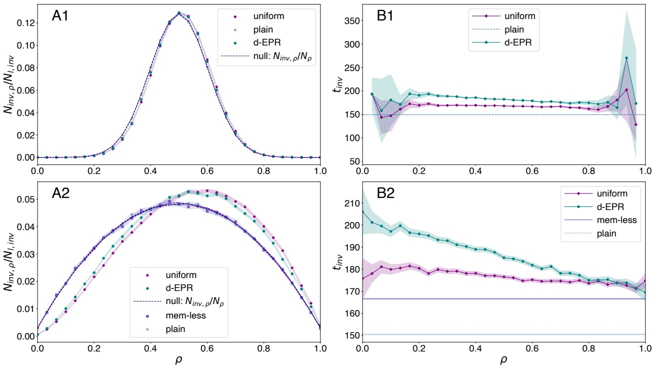


FIG. 1: **Invasion events**. Figures A1 and A2 show the share of invasions per mobility group ρ for a Gaussian (A1) and a beta distribution (A2). We compare the d-EPR model against the memory-less, the uniform, and the plain baselines, and also against a null reference where the invasion would occur randomly independently of ρ. Figures B1 and B2 depict the invasion times per mobility group, respectively for the Gaussian and the beta ρ distributions. Comparisons of the d-EPR results are also done against the previous baselines. Dots and dashed lines represent average values, while shaded areas cover the 95% confidence interval.

Then, we look at the overall infections in the system and similarly, we focus on infection share per mobility group ρ (disease prevalence) and average infection times. We find that under the Gaussian ρ distribution, all the global prevalence outcomes from the different scenarios collapse nearly under the same value and there is no differential impact across ρ groups. However, under the beta distribution, the different scenarios bring prevalence outcomes. Interestingly, only under the d-EPR scenario, there is a clear differential impact across mobility groups.

To conclude this abstract, we characterize where agents are infected, the frequency with which they visited the location where they got the infection, and its attractiveness 3. From $\rho \approx 0.33$, agents tend to spend more time outside their home location, and thus they tend to be infected there (60% of all infections). These places actually are rarely visited, but they are locations with higher attractiveness, where contagion events happen more frequently.

In conclusion, we see that explorers take the lead in bringing the disease to new locations, that is, help to "connect" the system, and also present a higher risk of contracting the infection. Admittedly, though, heterogeneity must be present in the population to appreciate some effects. Inspection of contagion event features and attractiveness of locations (not shown here), shows that the most attractive locations sustain outbreaks much more frequently than the least attractive locations. The next step in this project will include comparing the d-EPR model against the spreading under real mobility trajectories. Another interesting direction is to run these dynamics in structurally different urban environments.

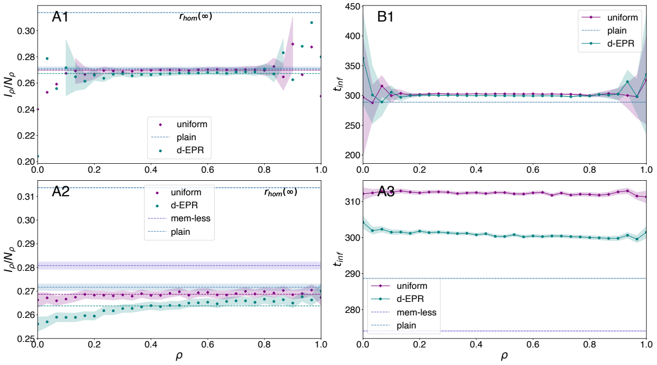


FIG. 2: **Overall infections**. Figures A1 and A2 represent the prevalence per mobility group for a Gaussian (A1) and a beta distribution (A2). We compare the d-EPR against the proposed baselines and also against the classical single-population SIR solution $r_{hom}(\infty)$. Figures B1 and B2 depict the infection times per mobility group ρ. Dots and dashed lines represent average values and the shaded areas represent the 95% confidence interval.

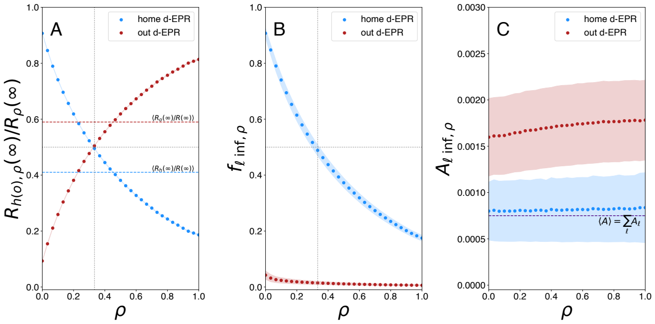


FIG. 3: **Infections at home and outside home locations**. A) Fraction of agents infected at home (outside). B) Frequency of stays at the infection location. C) Attractiveness of those locations. Dots represent average values.

[1] Song, C., Wang, P., Barabási, AL. Modelling the scaling properties of human mobility. Nature Physics, vol. 6, no. 10, pp. 818823, 2010.

[2] Pappalardo, L. et al. Returners and explorers dichotomy in human mobility. Nat Commun. 2015 Sep 8;6:8166.

You are where you eat: Effect of mobile food environments on fast food visits

Bernardo Garcia-Bulle[a], Abigail L. Horn[b], Brooke M. Bell[b,c], Mohsen Bahrami[a], Burcin Bozkaya[d], Alex Pentland[a], Kayla de la Haye[b], and Esteban Moro[a,e]

[a]MIT, [b] University of Southern California, [c] Department of Chronic Disease Epidemiology, Yale School of Public Health, Yale University, [d] Sabanci University, [e]Universidad Carlos III de Madrid

Poor diets are a leading cause of morbidity and mortality. Exposure to low-quality food environments, such as 'food swamps' saturated with fast food outlets (FFO), is hypothesized to negatively impact diet and related diseases. However, research linking such exposure to diet and health outcomes has generated mixed findings and led to unsuccessful policy interventions. To date, research into the relationship between food swamps or deserts and food choice has predominantly focused on predefined local and static food environments such as the environment around a home or workplace. That limited focus may explain the mixed results, given that a growing proportion of food acquisition and consumption occurs miles from our homes. For instance, Cooksey et al. [1] presented findings that food swamps predict higher rates of obesity at the neighborhood level. However, their results are weaker in neighborhoods where residents are more mobile (i.e., more residents who travel to work by car or public transport).

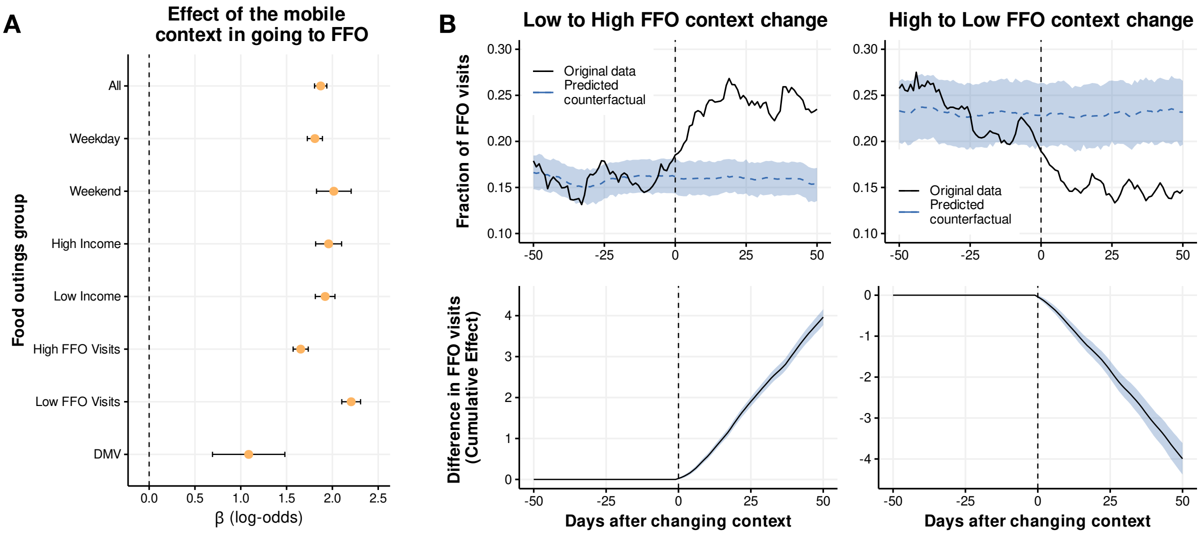
Here, we use a large, privacy-preserving, population-scale mobility dataset spanning a 6-month period during 2016-2017 and 11 metropolitan areas in the US to examine peoples' visits to food outlets (FO) and FFO in and beyond their home neighborhood. We investigate how these FFO visits are linked to features of the mobile food environments they are exposed to throughout their daily routines. Our analyses focus on visits to FFOs as the key outcome because greater intake of fast food is a well-established risk factor for diet related diseases.

Results

Our study focuses on the dynamic environments to which users are exposed throughout their days. To study the impact of these environments on our decisions, we did a logistic regression model trying to explain people's visits to FFO during lunchtime. As regressors, we used individual fixed effects along with a variable describing the food environment, which was ratio of number of FFO to all FO around a user within a 1km radius (ϕ). Because our users moved through the days, for food environment we selected the last place where the user was seen before 11h30, as that is generally before lunchtime.

Our results, illustrated in the Figure below, show a strong association between features of the mobile food environment and visiting a FFO. The model produces a log-odds of $\beta = 1.87 \pm 0.033$ for all FO visits at lunchtime: when the context includes 10% more FFO, there is an increase in the odds to visit an FFO of $(e^{\beta \times 0.1} - 1) \times 100 \simeq 20\%$. This influence of the food environment one is exposed to before going to lunch was similar during weekdays and weekends. In the results, we also include an analysis done using only the context when people went to a DMV (the location where one can get a license or ID), which we consider to be a close-to-random location giving us a natural experiment to verify the robustness of our results.

Despite finding an effect of mobile food environments on visits to FFO, it could be that the lack of non-fast food options predominantly affects individuals when they are in a new place. To address this question, we propose a semi-causal framework using a natural experiment to investigate the relationship between habitual FFO context and FO decisions. In this experiment, we observe people who changed their quotidian context during the study. Those users were split into four groups depending on whether they changed to a context with similar or different low ($\phi < 0.13$) or high ($\phi > 0.13$) exposure to fast food. We compare the FFO visits of group that changed their FFO contexts (Low → High and High → Low) with the counterfactual of those that, despite changing their context, were exposed to similar FFO food environments. Results are presented in the figure above, which shows that the group that changed from Low → High



A: Effect of the mobile food environment on visiting a FFO at different times, locations, or for different income or FFO visitation groups. Values show the coefficient of the ratio of FFO to all FO in a logistic regression for the food visits (outings) corresponding to the different groups. Bars indicate the standard error of the coefficient. B: Evolution of the fraction of FFO visits (top) and cumulative difference in FFO visits (bottom) for groups of users that change their contexts from Low to High FFO environments (left) and High to Low FFO environments. The dashed horizontal line is the predicted counterfactual for groups of users that changed their context.

FFO exposure increased their fraction of FFO visits from $\sim 16\%$ to $\sim 25\%$. The counterfactual of users that changed contexts but remained exposed to food environments with similar FFO ratios maintained a similar fraction of FFO visits.

The observed relationship between food environments with high ratios of FFOs and increased visits to FFO, specifically for mobile food environments, implies that more targeted interventions to reduce visits to FFO can be designed. Our findings highlight that FFO visits often take place well beyond the home neighborhood, and suggest that strategies that solely focus on geography and spatial access to food outlets in the home neighborhood are likely to lead to sub-optimal intervention effects. We then used the results of our observational study to identify the optimal locations to intervene in food environments to have the greatest impact on decreasing FFO visits. Specifically, these will be contexts demonstrating the highest ratios of FFO to FO, the highest frequencies of user exposure and FFO visits, and the largest observed impact of food environment features on a population's FFO decisions. We investigate the likely effects of intervention strategies that change the ratio of FFO to FO in these optimal impact locations vs. interventions targeting locations such as neighborhood food deserts and food swamps, the traditional choice locations for intervention. Overall, an intervention considering the mobile food environments would be 2x to 4x times more efficient in decreasing FFO visits than interventions that used only the FFO context where decisions are made or around the home neighborhood. We find that the groups of POIs related to "Malls", "Industry / Factory", "Airport" or "Office" are more likely to appear in our targeted areas than in the rest of the areas in the city and the rest of interventions.

References

[1] Kristen Cooksey-Stowers, Marlene B Schwartz, and Kelly D Brownell. Food swamps predict obesity rates better than food deserts in the united states. *International journal of environmental research and public health*, 14(11):1366, 2017.

A detour for snacks and beverages? A cross-sectional assessment of selective daily mobility bias in food outlet exposure along the commuting route and dietary intakes

Lai Wei [a*], Joreintje D. Mackenbach [b,c], Maartje P. Poelman [d], Roel Vermeulen [e,f], Marco Helbich [a]

[a] *Department of Human Geography and Spatial Planning, Utrecht University, Utrecht, the Netherlands*
[b] *Amsterdam UMC Location Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam, the Netherlands*
[c] *Upstream Team, Amsterdam UMC, the Netherlands*
[d] *Chair Group Consumption and Healthy Lifestyles, Wageningen University & Research, the Netherlands*
[e] *Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands*
[f] *Julius Centre for Health Sciences and Primary Care, University Medical Centre, Utrecht University, Utrecht, the Netherlands*

* Correspondence: l.wei@uu.nl

Abstract

Background: There are concerns about whether using global positioning system (GPS) data in food studies introduces selective daily mobility bias (SDMB) (1), possibly distorting associations between the food environment and diet-related outcomes. Evidence of the existence of SDMB is limited. Therefore, this study aimed to examine whether SDMB influenced the associations between exposure to food outlets regarding snacks and soft drinks along commuting routes and dietary intake among young Dutch adults.

Methods: We used 7-day smartphone-based GPS tracking data for 67 adults aged 25 to 45 living in urban areas in the Netherlands. In addition to the GPS-tracked commuting routes, we computed each participant's shortest-path commuting route. An absolute and a relative measure of exposure to food outlets where people could purchase snacks and soft drinks was assessed on both routes using 100 m and 250 m route buffers. Participants' self-reported average daily consumption of soft drinks, small snacks, and large snacks were collected through the baseline food frequency questionnaire. We used paired Wilcoxon test to compare the food outlet exposure along both routes and fitted Tobit regressions to examine their associations with daily dietary intake outcomes.

Results: Our results showed that the absolute food outlet exposure based on GPS-tracked routes was significantly lower than those along the shortest paths route in 100 m (median: 13

vs. 18 outlets) and 250 m (median: 30 vs. 43 outlets) buffers (Figure 1). No statistical differences were observed between the two types of routes for relative food outlet exposure. Furthermore, only soft drink intake was negatively associated with relative food outlet exposure on the GPS-tracked route in the 100 m (β: -0.03, 95%CI: -0.05, -0.01) and 250 m (β: -0.04, 95%CI: -0.07, -0.01) buffers. Associations between all dietary intake outcomes and food outlet exposures on the shortest path route were null.

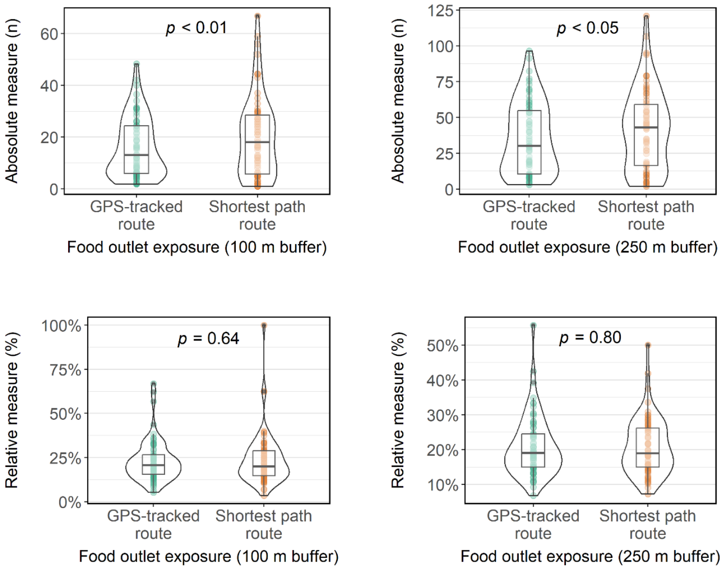


Figure 1. Differences in individual-level food outlet exposures along GPS-tracked routes and shortest path routes across 100 m and 250 m buffers in the FoodTrack study (*N* = 69). Mean differences were statistically tested by means of paired Wilcoxon signed rank tests.

Conclusions: Our study showed no evidence of SDMB in the relationship between exposure to food retailers regarding snacks and soft drinks along commuting routes and snacks and beverage consumption in this sample. The relative exposure to food outlets on the shortest path routes was sufficiently similar to that on GPS-tracked routes to allow for shortest path route data to act as a potentially surrogate for GPS-tracked route data. It should be noted that our findings might be limitedly generalizable to other food environment and were based on a small but nuanced dataset.

References

(1) Chaix B, Meline J, Duncan S, Merrien C, Karusisi N, Perchoux C, et al. GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? Health Place 2013;21:46-51.

A methodology for monitoring travel demand in Latin American cities combining mobile network data and other demand data

Javier Burrieza-Galán - Nommon Solutions and Technologies - javier.burrieza@nommon.es
Miguel Picornell Tronch - Nommon Solutions and Technologies - miguel.picornell@nommon.es
Ricardo Herranz - Nommon Solutions and Technologies - ricardo.herranz@nommon.es
Ellin Ivarsson - World Bank Group - eivarsson@worldbank.org
Aiga Stokenberga - World Bank Group - astokenberga@worldbank.org

The COVID-19 crisis had a severe impact on public transport demand around the world. Latin America has not been an exception: public transport ridership decreased by 60-90% in most cities during 2020, hitting fare revenues and compromising services' financial viability. In this context, the World Bank Group (WBG) identified the increasing availability of geolocation data as an opportunity for the region's transport sector to monitor the evolution of travel demand patterns and better understand mobility trends. To take advantage of this opportunity, the WBG commissioned Nommon, a Spanish mobility data analytics company, to develop a methodology for calculating travel demand indicators that could be applied to any Latin American city, based on the fusion of mobile network data, public transport smart card data and mobility surveys. The WBG engaged the cities of Buenos Aires (Argentina) and Bogota (Colombia) as case studies to demonstrate the new methodology, which has already been replicated in Medellín (Colombia). The presentation in the NetMob conference will outline the developed methodology and present case study results.

The methodology is based on Nommon's solution for obtaining travel demand information from anonymised mobile network data (Mobility Insights[1]), which generates 'activity-travel diaries' for the sampled mobile phone users and expands these diaries to the total population using census data[2]. The methodology addresses one of the key challenges when analysing urban mobility patterns: the identification of transport mode. While mobile network data offers an unprecedented detail of trip generation and distribution patterns, it is often not enough to characterise mode choice: a single sequence of mobile events is often compatible with many mode options. The framework provides three alternative methods to overcome this problem: (i) analysing public transport smart card data to obtain origin-destination matrices of the public transport system, which then serve as a reference for assigning trips to public transport services; (ii) training machine learning with available mobility surveys, which then allows to classify the trips observed with mobile network data; (iii) a hybrid approach that combines both methods. Additionally, the project developed algorithms for identification of trips associated with professional drivers (which therefore are not potential demand of public transport services) based on the observed non-home based trip patterns. The output consists of hourly origin-destination matrices segmented by traveller profile (home location and other sociodemographic attributes such as age and gender, depending on the availability of information about anonymous mobile phone users included in the sample) and trip characteristics inferred from the analysis of mobile phone records (trip purpose, trip distance, trip mode and passenger/professional trip).

[1] https://www.nommon.es/products/mobility-insights/

[2] For a detailed explanation of the methodology for obtaining activity-travel diaries from the sample of mobile phone users and an expansion to the total population see Bassolas, A., Ramasco, J. J., Herranz, R., & Cantú-Ros, O. G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transportation Research Part A: Policy and Practice, 121, 56–74.

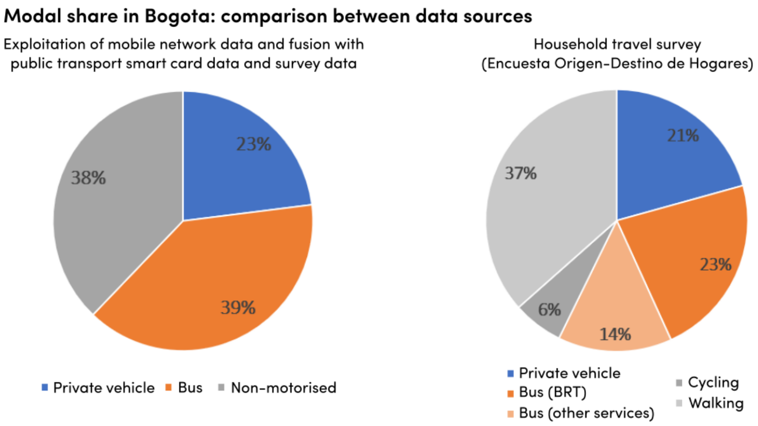


Figure 1. Comparison of modal share in Bogotá obtained with the methodology developed in the project (left) and the results of the latest household travel survey (right).

The case studies covered several working days, Saturdays and Sundays from 2019, 2020 and 2021. The results helped Bogota and Buenos Aires transport authorities understand how citizens modified their travel behaviour during the pandemic. The analysis of mobile network data revealed that COVID-19 had a long-lasting impact on trip generation rates (i.e., average number of daily trips per person): in 2021, after the end of all social distancing measures, trip rates were well still below the levels registered in 2019 (-7.5% in Bogota and -17.5% in Buenos Aires). The decline was more noticeable among trips performed by those with higher socioeconomic levels and among home-work trips, which may reveal differences in the adoption of telework across the population.

In 2021, public transport ridership was still well below 2019 levels, showing a 20% drop in Bogotá and a 40% drop in Buenos Aires. The results suggest that this was the result of two factors: (i) the sharp decrease in the overall demand for medium and long-distance trips within both metropolitan areas (trips larger than 5 km); and (ii) the shifts from public transport to private vehicles, particularly in Buenos Aires, where private vehicle modal share increased from 38% to 47% in 2020 and remained higher during 2021 (43%).

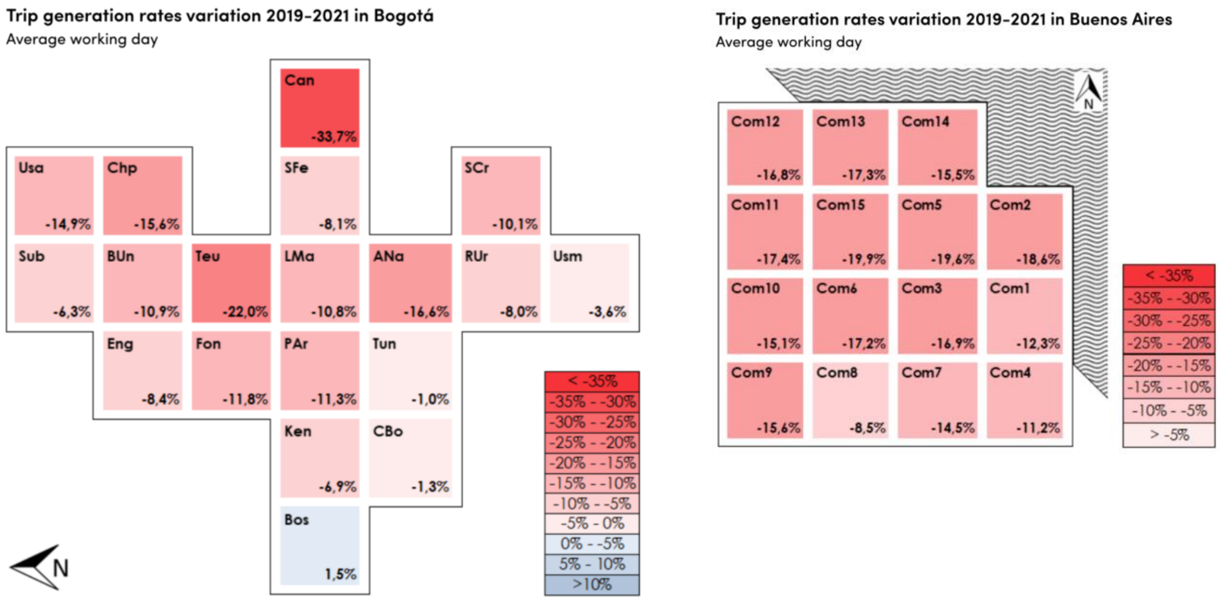


Figure 2. Trip generation rates variation from 2019 to 2021 in an average working day in Bogotá (top) and Buenos Aires (bottom)

Paths Reconstruction From Cell Phone Data Using ANN in Havana

Andy Rodríguez Daniel A. Amaro Ramos[2] Orlando Martínez[1]
Milton García-Borroto[2] Alejandro Lage-Castellanos[2]

[1] IMDEA Networks, [2]Centro de Sistemas Complejos, Facultad de Física, Universidad de La Habana, (**Mail to:** ale.lage@gmail.com)

In this short paper we attempt to reconstruct intermediate missing points in trips identified from mobile phone-tower data. We will study a simplified version of the path reconstruction problem[2]:

[Intermediate point path reconstruction, IPPR]: Predict the chances to locate a traveler at a given Transport Area when it has already traveled a fraction of the distance between a given origin and destination.

We use an artificial neural network to produce probable intermediate values in trips. Before applying it to the case of Havana, in Cuba, we test the quality of the approach in an artificial city model.

Path reconstruction

We assume that users locations are close to the towers handling their mobile signals. We use the Location Area registries (a type of network-driven record) previously anonymized and granted access in the servers of Etecsa, the telecom company in Cuba. [1]

Registries are in the form in the form (IMSI, CELL, DATETIME). IMSI field corresponds to an anonymized hash of the users IMSI. The CELL field is designing a group of nearby 2G and 3G antennas, whose approximate position was also granted by the company. DATETIME is a timing indicator of the moment the mobile interacted with the tower.

For each user we identify trips, *i.e.* sequences of cells that span more than 2 km altogether, without having stops of more than 2 hours in any given cell. The parameters and the methodology for such process is described in [1], where it is shown that they correlate well with known information about population dynamics in Havana. We adopt a description at the level of Transport Area (TA). The city transport system is organized in 134 TAs that cover the city surface. We map the GPS coordinates of every CELL to a corresponding Transport Area, as done in [1]. Trips, corresponding to sequences of TAs, look like:

```
[95, 46, 48, 31, 28, 23, 14, 42, 128, 116, 117]
[14, 61, 68]
[95, 128, 112, 117]
```

This type of data is quite heterogeneous and the actual number of intermediate TAs can vary a lot, which is the reason why it is interesting to attempt a reconstruction.

1 Path reconstruction with ANN

We want to train a network that can estimate the probabilities of finding a traveler at any given Transport Area, when it has covered a fraction of trip between a given origin and destination. The input layer of the ANN will receive 5 numerical values:

- I_1 a latitude for the origin point;
- I_2 a longitude for the origin point;
- I_3 a latitude for the destination point;
- I_4 a longitude for the destination point;
- I_5 real number in $[0, 1]$ for the fraction of the trip distance;

When $I_5 = 0$ ($I_5 = 1$) we are at the beginning (end) of a trip, while in between values are used to map positions of travelers from one point to the other.

The last layer of the network is a *softmax* activation function, that naturally produces probability distributions. We have a 5 input, 134 outputs through a *softmax*,

[1] Even after anonymization, no inter-user metadata is stored, and researchers signed a confidentiality agreement.

while other architecture details of the network are set by experience, intuition, or simple trial and error.

We implemented the ANN in Keras using 3 hidden ReLu layers. We used categorical cross-entropy as the loss function to minimize during training.

Preprocessing is needed to turn sequences of Transport Areas into triplets,

```
(To, Tm, tm, Td)
```

where T_o and T_d are the origin and destination TAs of the paths, and $t_m \in [0, 1]$ is a fraction of the whole trip time at which user is detected in TA T_m. For a given trip, each intermediate point will generate a quadruplet. With this input data we can train the network by feeding it the with lots of

```
( Lat(To) , Lon(To), Lat(Td) , Lon(Td), tm )
```

and minimizing the output cross-entropy with the real T_d.

Artificial city model In order to have a benchmarking for our ANN, We developed an artificial 2D city model where we can compute exactly the statistical properties of the trips. In this L × L square lattice city, each link (i, j) is randomly weighted ω_{ij}.

We only consider paths $\mathcal{P} = \{h_0, h_1, \ldots, h_n, h_{n+1}\}$ that are efficient, meaning that every step moves the traveler closer to their destination. Given an origin h_0 and a destination h_{n+1} the total weight (cost, or energy) of a path as $\mathcal{H}(\mathcal{P}) = \sum_{(i,j)\in\mathcal{P}} \omega_{i,j}$, and the probability to take a certain path $\mathcal{P}$ between nodes h and m can be expressed as $P(\mathcal{P}) \sim \exp(-\beta\mathcal{H}(\mathcal{P}))$ Real valued β is a scale parameter giving more or less relevance to the weights in the selection of paths, with $\beta = 0$ producing equiprobable paths.

Testing ANN on the artificial city

We generated an artificial square city to test our ANN, with L = 12 and weights ω_{ij} randomly sampled from a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. Selecting random points as origin and destination of trips, we sampled many paths according to the distribution of their weights to generate an file with a similar structure to the data obtained from trips identification from Location Update records.

From the city model we generated a set with 10^6 inputs, where 15% is used as validation and the rest as training of the ANN. We found a high correlation ($C = 0.9523$, for $\beta = 1.0$) between the ANN predicted probabilities and the exact probabilities to find a traveler at any intermediate point between any two origin and destinations in the city.

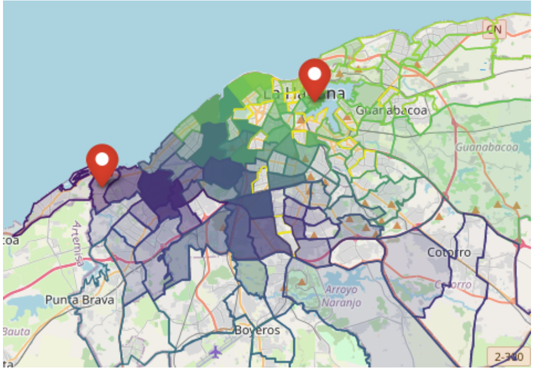


Figure 1: Havana heat map of travels between transport zones 2 and 33. Color code is proportional to time, and transparency to the probability of visiting each zone.

Real data: Havana, Cuba

Location Area updates were collected from the 2G-3G network for 15 days in March 2020, just before COVID affected the population dynamics in Havana, in order to have a normality snapshot of Havana City.

An ANN was trained using de data collected. Using the trained network we were able to produce likely paths from any pair of origin an destination Transport Areas. In figure 1 we show an example, starting in Playa and ending in Old Havana, the result is consistent with what is expected from our knowledge of the city transport system.

Conclusions. We trained a multilayered ANN to produce the probabilities of intermediate missing points in sequences of transport areas for travelers, identified from mobile-tower data. Although we lack a ground truth to validate our predictions on real data, testing on the artificial city and knowledge of the mobility patterns in Havana both suggest that the reconstruction is sound.

References

[1] Intra-day population fluxes from mobile phone data in havana, cuba. in preparation.

[2] Mingxiao Li, Song Gao, Feng Lu, and Hengcai Zhang. Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. *Computers, Environment and Urban Systems*, 77:101346, 2019.

Inter-city Trajectory Clustering Using Mobile Phone Records

Laura Gutiérrez-Lastra [1] Claudia González-Alfonso[1] Yairobis Salazar-Matos[1]
Ernesto Rosales[2] Alejandro Lage-Castellanos[3] Milton García-Borroto[3]

[1] Unión de Ferrocarriles de Cuba, [2] AlaSoluciones , [3]Centro de Sistemas Complejos, Facultad de Física, Universidad de La Habana, (**Mail to:** milton.garcia@gmail.com)

We present an end-to-end solution for clustering the trajectories of Cuban travelers on long-distance trips, using anonymized Location Area mobile phone records. We encode users data into a feature vector that is then automatically clustered using a custom-defined distance measure. Results are validated in two ways, comparing to expert knowledge of Cuba's transport system and also by comparing large clusters with train timetables.

Phone records. Location Area Update (LAU) registries of a one month period (March 8th-April 8th, 2023) were accessed at the facilities of ETECSA, the single Cuban Telecom. The logs were fully anonymized, ensuring privacy protection for users. Furthermore, registries did not contain user-to-user metadata, and researchers accessing the data signed a confidentially agreement.

Location Area registries are network driven, and contain, among others, a user-id (anonymized), a date-time stamp and a cell-tower-id. Telecom company provided approximate GPS locations for its towers, together with their corresponding tower-id. Throughout this paper, we use the tower position as a surrogate of user's actual position when registries are generated.

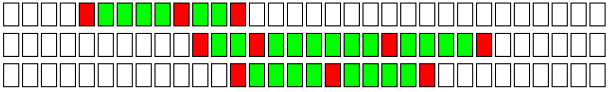
Among the actions that trigger the logs, the change in Location Area is particularly useful for mobility studies. The volume of the data was around 300 million records per day, efficiently stored in a Hadoop cluster.

GPS data. For validation purposes, GPS logs were obtained from inter-provincial transportation busses covering some routes between distant locations in Cuba. Each GPS record contained a continuous record from the first day of the study to the last one, consisting of tuples (daytime, latitude, longitude). Preprocessing of this data allowed for a clear extraction of many instances of inter-provincial bus journeys.

Trajectory feature extraction. We restrict our analysis to users-id's that are identified in at least two different provinces in times spans of at most 2 days. We isolate the trajectory fragment between the starting and ending provinces, ignoring registries that are not in this interval.

We created features vectors of spatial (lat, long) coordinates with $172.8s$ time intervals, dividing any 2 days period in 1000 fragments. Every coordinate in the vector is filled with the average of all cellphone tower positions in the user registries during the time interval. While this vectors are intentionally left empty before/after the trip begins/ends, they also abound in empty entries during the trips. These gaps were filled with a linear interpolation between GPS coordinates of consecutive logs.

Diagram represent feature vectors for three different users. In white are the empty spaces before and after the identified trip. Red positions are filled with actual data, while green are interpolations.



Distance and Clustering Algorithm Distances among feature vectors are defined in a $[0, 1]$ interval. A pair of users whose non empty time slots do not overlap enough, are directly set to have the maximal distance 1. For the rest, geodesic distance is computed for every pair of non empty slots. The distance between the vectors is given by the fraction of these slots that has a geodesic distance above 5000m.

We used a hierarchical clustering method, that results in a hierarchical structure of agglomerative clusters. We consider two parameters for the clustering method. The first parameter is the linkage criterion to determine the distance between clusters. These criteria define how the distance between two clusters is calculated based on the distances of their constituent data points. We considered seven values: **single, complete, average, ward, centroid, median, and weighted** linkage, each with its strengths and weaknesses.

The second parameter is the cutoff point, *i.e.* the maximum allowed distance in the same cluster. Lower threshold values lead to more fine-grained clusters, where data points within each cluster are more similar or closer together.

Clustering tuning. We evaluated seven linkage criteria (”single,” ”complete,” ”average,” ”weighted,” ”centroid,” ”median,” and ”ward”) and (2, 4, 6, 8, 10, 12, 14) cutoff points.

Effectiveness of our clustering will be measured by comparing to “ground truth” clusters. Using GPS from public transport between provinces, we find feature vectors that are consistent with the bus trajectory, and group them as C^0_{GPS} ground truth clusters. Therefore, a high-quality clustering outcome would entail the successful grouping of all the GPS-related trajectories within a single cluster. Comparing the intersections of C^0_{GPS} with the most overlapping cluster C^{max}_c resulting from our analysis, we compute standard precision and recall metrics and the resulting F-measure ($f1 = 2 * (Prec * Rec)/(Prec + Rec)$). This is done for every linkage and cut-off parameters.

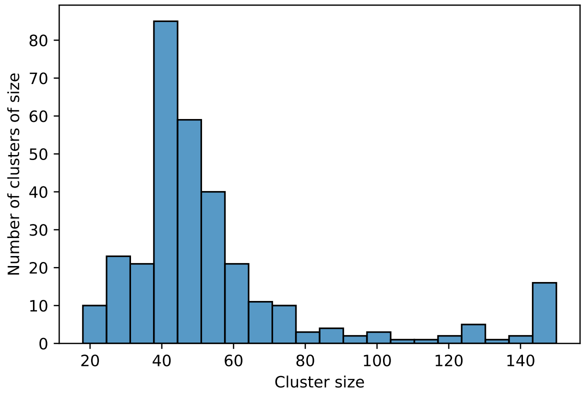


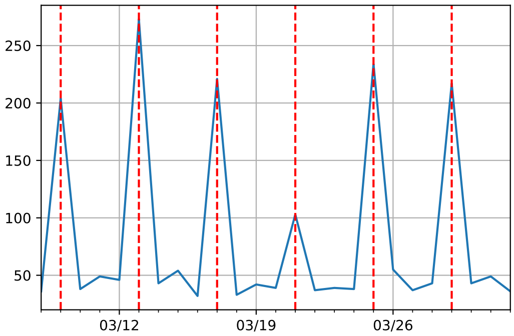
Figure 1: Five clusters with the largest number of elements in the Guantánamo - La Habana trajectory on date 2023/03/18

The F-measure values for each GPS trajectory were ranked, and the resulting rankings were then averaged. Based on the high average rank achieved by **ward** linkage and cut-off 10 and its favorable standard deviation in the measure, we have chosen it as the best candidate for performing the clustering task.

Clusters validation. In this paper, we employed two indirect measures to validate the introduced clustering method. We first examine the size of the clusters, and whether they capture meaningful patterns and segregates the data into cohesive groups. The second measure evaluates the correspondence in time and size of large clusters with inter-provincial trains.

In Figure 1, we present the histogram of transportation cluster sizes. To simplify the visualization, we combined all cluster sizes above 150 into a single category labeled as 150. The peak around 40 persons is consistent with the most common bus/truck transportation. Values above 60 -although uncommon- cannot be fully explained by the existing transportation methods.

In Figure 2, the size of the largest identified cluster is depicted for each available route and day. Additionally, the days when the national train connecting these provinces operated are represented with dashed lines.



(a) Guantánamo - La Habana

Figure 2: Size of the largest cluster found on each day of the week for Guantanamo-Habana route. Dashed lines indicate the days when the train operated on that route

As can be seen in the graphs, the trajectories have a certain baseline value that remains stable over time, with peaks occurring on the specific days where trains connected La Habana and Guantánamo. Detailed analysis shows that these clusters coincide in time profile with the known train trajectories.

Conclusion. By leveraging the hierarchical clustering approach and evaluating different linkage criteria and cut-off thresholds, we were able to uncover meaningful clusters representing various travel routes.

The clustering algorithm successfully captures the inherent structure of the trajectory data, revealing distinct clusters corresponding to specific travel routes between provinces. The identified clusters align well with the known national train schedule, further validating the accuracy and reliability of the clustering approach. The evaluation of cluster sizes and the comparison with the national train schedule highlight the potential utility of the proposed method in understanding and analyzing transportation patterns.

Is it a work or leisure travel? Applying text classification to identify work-related travel on social networks

Lucas Félix
Univesidade Federal de Minas Gerais
Brasil
lucas.felix@dcc.ufmg.br

Washington Cunha
Univesidade Federal de Minas Gerais
Brasil
washingtoncunha@dcc.ufmg.br

Jussara Almeida
Univesidade Federal de Minas Gerais
Brasil
jussara@dcc.ufmg.br

1 INTRODUCTION

Through the internet, users gather in forums, communities, travel blogs, and *Social Networks* (**SNs**). Indeed, these platforms are useful tools considering that they made it possible to seek *Points of Interest* (**POI**), provide and receive information about places, and share experiences from a single point of view. Studies show that 80% of American travelers use *SNs* while traveling, and more than half of this percentage share their journey information with their contacts [12]. This study also showed 'how' and 'why' *SNs* are the tool most adopted by usual travelers.

Between the platforms available on the web Location-Based SN (**LBSN**) and Travel SN (**TSN**) are more focused on tourism and have been extensively used in the literature in tourism works [5]. The LBSNs' main purpose (e.g., *Yelp* and *Foursquare*) focus on the user's current location sharing their footsteps. On the other hand, the *TSNs* (e.g., *TripAdvisor*) are responsible for joining in the same platform several pieces of information, such as locations, accommodations, transport, food, attractions, and services [12], enabling their users to describe better the full trip experience. Furthermore, *TSNs* allows users to plan trips, look at different perspectives on a place, and get recommendations about users with the same taste [6], while comparing prices.

Although TSN and LBSN platforms concentrate most of the information needed to plan a full trip experience to a specific place (or set of places), users still encounter difficulty with the huge amount of available data, precluding to distinguish which option (or set of options) is the best [9]. Therefore, *Recommender Systems* (**RS**) solutions are usually applied to tackle the information overload problem, focusing on suggesting POIs based on the user history. RS approaches reduce the number of options for the users and enable them to choose between a smaller (and possibly better) option set.

Many widely studied aspects, such as physical constraints, social and temporal influences, make the POI recommendation scenario harder than the classic one. Consequently, the literature has naturally adopted contextual information to leverage recommendations quality. Indeed, additional data allows for more accurate recommendations than traditional methods (e.g., Collaborative Filtering) [10].

One additional piece of information that can be used for tourism recommendations is the user **travel purpose**. In the literature, this has not been properly exploited due to the fact that this is not usually available in datasets and, when available, is not usually filled by the users. On the other hand, this can be used to properly characterize tourists' actual behavior while traveling [3], and providing, in this way, a more suited recommendation for these users.

Accordingly, in this work, we propose a model to predict whether a trip is leisure or work-related. We aim to enable authors to characterize better travels made by users and perform better recommendations in specific scenarios. To accomplish so, we compare different state-of-the-art **Automatic Text Classification (ATC)** models, namely BERT [2], RoBERTa [8], and BART [7].

2 PROPOSED STRATEGY

2.1 Data Collection and Pre-Processing

From the datasets available in the literature, only a few present essential features for real-world evaluation, such as the POI working hours, availability, and cost. In this scenario, data from TSN, like TripAdvisor, are more suited for such analyzes. Besides containing more information, TripAdvisor is currently the most popular travel website [5], with about 390 million monthly unique visitors. Therefore, we adopted TripAdvisor review datasets to train an ATC model capable of distinguishing leisure and work-related travels.

To accomplish so, we extracted TripAdvisors' data through a web crawler responsible for automatically browsing the website and collecting all users' available content and POIs. The collected data was firstly in an unstructured format (e.g., HTML). Then, it was parsed to retrieve the content within each page, pre-processed, and stored in semi-structured format (e.g., CSV) data.

Our collection was initially focused on users from five different touristic cities worldwide and their complete visit history (including other cities): Tiradentes and Ouro Preto (Brazil), San Gimignano, Cannes, and Ibiza [1]. Since our dataset possessed users' complete history, containing the visits to numerous different cities, in this work, we focused on the English reviews available [2] that are associated with the trip label (e.g. family, romantic, friends, work-related, alone) [3]. We focused the analysis on English reviews due to the fact that the text classification algorithms employed in our methodology demonstrate better performance with English rather than in other languages [11]. To identify the English reviews on the datasets used in this work, we adopted the pre-trained model proposed in [4].

To augment our dataset, we labeled visits occurring within the same city and time period (month and year) that lacked a label. To assign a label, we utilized the available classification from another POI visited by the same user in the same city and time period.

Considering our aim is to define leisure and work-related travels to characterize travelers further initially, we modeled our problem as a binary classification problem by aggregating non-work classes as leisure, leaving only two classes. Following the aggregation

[1] TripAdvisor Complete Dataset
[2] TripAdvisor English
[3] TripAdvisor English w/ classes

process, it was observed that 87.67% of the instances consisted of leisure-related reviews, whereas 12.33% were categorized as work-related reviews. This indicates an imbalanced dataset context, where most instances pertain to leisure-related reviews.

Table 1 presents a summary of TripAdvisor data used in this work.

Dataset	# Instances
TripAdvisor Complete	11, 443, 663
TripAdvisor English	2, 434, 252
TripAdvisor English w/ classes	639, 997

Table 1: TripAdvisors' data summary

2.2 Text Classification

In the ATC block of our modeling, we aim to classify the users reviews, identifying if their travels were leisure or work-related. To accomplish so, we compare three different neural approaches that are the state-of-the-art on text classification task, achieving the best results in the benchmarks used by researchers [1], namely we use Bert [2], RoBERTa [8] and BART [7].

To compare the effectiveness of each approach, we follow the procedure defined in [1]. We evaluate the effectiveness of the proposals using Macro and Micro Averaged F1. The experiments were executed using a 5 fold cross-validation procedure. To compare the cross-validation results, we evaluate the statistical significance using a paired t-test with 95% confidence with Bonferroni correction to account for the multiple tests.

As mentioned earlier, our dataset exhibits a significant skew, making it challenging to obtain accurate predictions for both classes. To address this issue, we implemented a class balancing procedure during the training phase for each fold, in which we randomly chose N_{mino} instances from the majority class, ensuring equilibrium with the N_{mino} instances derived from the minority class. During the testing phase, on the other hand, we maintained the overall distribution of the datasets to ensure a realistic representation of the actual data.

3 EXPERIMENTAL EVALUATION

In this section, we present the results achieved by our proposal. First, we present and discuss some instances used to train the models. Then, we present the metrics for each model.

3.1 Dataset

In Table 2, we present some examples of each class present in our dataset. In these examples, it is easy to distinguish each class, especially in the second case, due to the fact that the author states that it is work-related travel. Nevertheless, not all work-related instances can be easily distinguished, as in this example. Illustrating that, we have the following example, which is work-related travel: '"I don't really understand much of the exhibits, but it can be an eye-opener. The free exhibits took us slightly less than 2hrs to complete. The navigation is easy. Premise is clean and spacious.". However, in this particular example, there are no indications suggesting that this trip is work-related. Therefore, as part of our future work, we aim to incorporate additional features that can help us differentiate between leisure and work-related trips. Our intention is to leverage features that are commonly found across various datasets, thereby enabling the application of our proposed model in different scenarios.

Class	Review
Leisure	"What a fantastic example of the Brazilian Barroc art..."
Work	"..This hotel is one of my favorite I've stayed in for work travel..."

Table 2: Instances used to train in each class

3.2 Classification Results

Table 3 results show the effectiveness of each model evaluated in our methodology, presenting Macro-F1, Micro-F1, and the Confidence Interval (CI) of each algorithm. Even though the RoBerta approach presents the best average results, all the algorithms are statistically equivalent.

Model	Macro-F1	Micro-F1
Bart	69.1(1.52)	80.86(1.89)
Bert	67.36(1.45)	79.78(1.36)
RoBerta	70.16(1.57)	82.15(2.11)

Table 3: Models metrics and CI

Based on the results, it can be inferred that the algorithms demonstrate strong predictive capabilities for both classes. The evaluation of Macro-F1 metrics supports the notion that the algorithms achieve favorable classification outcomes for both classes. However, as mentioned earlier, we acknowledge that in certain instances, relying solely on reviews may not provide sufficient discernment to accurately differentiate trip labels. Consequently, this limitation has an impact on the Macro-F1 results.

ACKNOWLEDGMENTS

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the CAPES-STIC-AMSUD 22-STIC-07 LINT project.

REFERENCES

[1] Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos Gonçalves. 2023. A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM CSUR* (2023).

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Linus W Dietz, Rinita Roy, and Wolfgang Wörndl. 2019. Characterisation of traveller types using check-in data from location-based social networks. In *Information and Communication Technologies in Tourism 2019: Proceedings of the International Conference in Nicosia, Cyprus, January 30–February 1, 2019.* Springer.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[5] Amir Khatibi, Fabiano Belem, Ana P Silva, Dennis Shasha, Marcos A Goncalves, et al. 2018. Improving tourism prediction models using climate and social media data. In *Twelfth International AAAI Conference on Web and Social Media.*

[6] Jinyoung Kim, Hyungjin Kim, and Jung-hee Ryu. 2009. TripTip: a trip planning service with tag-based recommendation. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems.*

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint 1907.11692* (2019).

[9] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* (2008).

[10] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. 2013. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL.*

[11] Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840* (2019).

[12] Kyung-Hyan Yoo, Marianna Sigala, and Ulrike Gretzel. 2016. Exploring TripAdvisor. In *Open tourism.* Springer.

Mixing Individual and Collective Behaviours in Mobility Models

Sebastiano Bontorin sbontorin@fbk.eu,[1, 2] Riccardo Gallotti rgallotti@fbk.eu,[1] Luca Pappalardo luca.pappalrdo@isti.cnr.it,[3] Simone Centellegher centellegher@fbk.eu,[1] Manlio De Domenico manlio.dedomenico@unipd.it,[4, 5, 6] Bruno Lepri lepri@fbk.eu,[1] and Massimiliano Luca mluca@fbk.eu[1, 7]

[1]*Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy*
[2]*Department of Physics, University of Trento, Via Sommarive 14, 38123 Povo (TN), Italy*
[3]*ISTI - National Research Council, Via Giuseppe Moruzzi 1, 56127 Pisa (PI), Italy*
[4]*University of Padua, Via Francesco Marzolo 8, 35131, Padua, Italy*
[5]*Padua Center for Network Medicine, University of Padua*
[6]*Istituto Nazionale di Fisica Nucleare, Sez. Padova, Italy*
[7]*Free University of Bolzano, Piazza Università 1, 39100 Bolzano (BZ), Italy*

Understanding human mobility and performing next-location prediction at the individual level is a challenging task [1, 2], which is critical for numerous research and practical applications, such as urban planning, public health and sociology. Knowledge of individual mobility patterns enables the development of effective policies that meet the needs of individuals and communities.

Recent studies suggest that some intrinsic properties of mobility datasets limit the predictability of individual-level human mobility. For example, the visitation frequency to places (e.g., points of interest) follows a long-tail distribution. Thus, if we aim to predict the next destination an individual will visit, a model will often see a limited number of frequently visited locations during the training phase while many other potential destinations may not be represented. This leads to a core problem in the prediction of locations not seen in training, defined as novel mobility prediction, a challenge where even SOTA deep recurrent models fail.

Inspired by the literature on the interplay between individual and collective decisions of intelligent systems, we design a model that leverages collective mobility decisions dynamically to generalize mobility prediction for out-of-routine mobility traces. In our work, we propose to support next-location predictors via collective origin-destination matrices [3]. We study minimal Markov models and exploit the physics-grounded definition of entropy as an effective definition of an individual's unpredictability seen in training.

Specifically, given a user u and its current trajectory point i, we retrieve its individual (IND) probability distribution T_{ij}^{IND-u} over all the possible target locations $\{L\}$. We then quantify the amount of uncertainty by measuring its Shannon's entropy:

$$\alpha_i^{IND-u} = \frac{1}{H_{max}} \sum_{j \in \{L\}} T_{ij}^{IND-u} \cdot log(T_{ij}^{IND-u}). \qquad (1)$$

Where the normalizing factor H_{max}, defined as the maximum entropy over the set of possible destinations, allows α_i^{IND-u} to be understood as a compression rate. This allows us to bias the individual Markov model with the collective transition probability in an effectively parameter-free model that only requires the knowledge of inherent physical property of human trajectories (see Fig. 1). α_i^{IND-u} encodes how much collective (COL) behaviors should impact the prediction of the next location: this complex interplay is then encoded in the Markov-Chain (MC) transition matrix

$$MC_i^{IC-u} = MC_i^{IND-u} \cdot (1 - \alpha_i^u) + MC_i^{COL} \cdot \alpha_i^u, \qquad (2)$$

where IC stays for "Individual-Collective". We demonstrate the effectiveness of our approach by using a large-scale dataset that contains trajectories of more than 2 million users collected over nine months in 2020 in New York City, Seattle and Boston, provided by Cuebiq. We build collective and individual transition matrices from this dataset and use them to perform next-location prediction via our entropy-based model.

Our results show that our approach improves the accuracy of next-location prediction compared to existing state-of-the-art methods (see Fig. 2). The model biases the prediction probability when the unpredictability of movement is high or the model lacks information about individual mobility because it didn't see it during the training. Moreover, we also exploit this minimal entropy-based model to shed light on features of human individual mobility linked to the urban environment: specific urban areas are characterized by larger movement entropy and benefit more from the collective dynamics information. Furthermore, we investigate the changes in human mobility behavior and routines during the COVID-19 pandemic by analyzing the entropy and predictability of individual mobility trajectories.

Our approach not only improves the accuracy of next-location prediction, but also generalizes to the concepts of exploration of urban scenarios and sheds light on inner properties of the mobility trajectories of the urban environments in areas characterized by high densities of POIs. Furthermore, the simplicity of our model allows it to be integrated into more complex mobility models by means of additional layers, to integrate collective mobility information.

The combination of collective origin-destination matrices and entropy-based models provides a powerful tool for predicting human mobility and exploring the dynamics of complex systems. We believe that this approach has the potential to open up new avenues of research in the field of computational social science and complex systems, and provide valuable insights into the behavior of individuals and communities in urban environments.

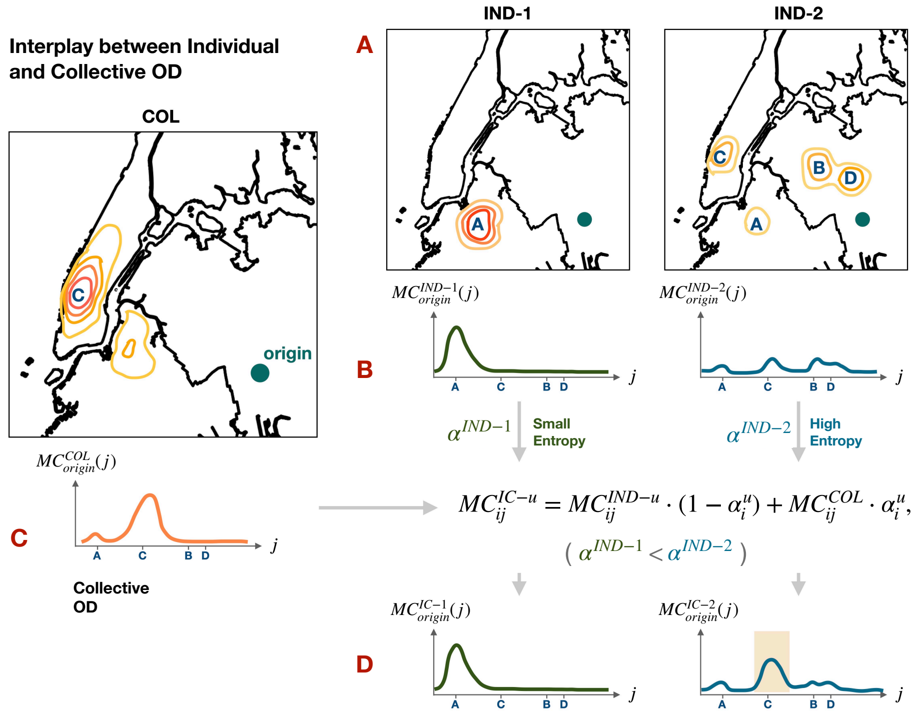


FIG. 1. **Graphical depiction of the interplay between Individual and Collective Mobility: (A)** individual transition probabilities MC^{IND-u} are shown in the urban space: color intensity maps the probability of moving to one of four possible target locations form a given sample origin (green). Two exemplary individuals (1 and 2) are characterized by different distributions, **(B)** Individual 1's transition probability peaks in location A, resulting in a small entropy. While user 2's transition probabilities are more uniform across possible targets (IND-2), thus having larger entropy and larger unpredictability. Collective flows MC^{COL} information **(C)** is integrated in the prediction via α^u_i, resulting in the MC^{IC} probabilities **(D)** where the collective biasing for IND-2 is larger due to its higher origin entropy, and a bias towards destination C appears in the distribution.

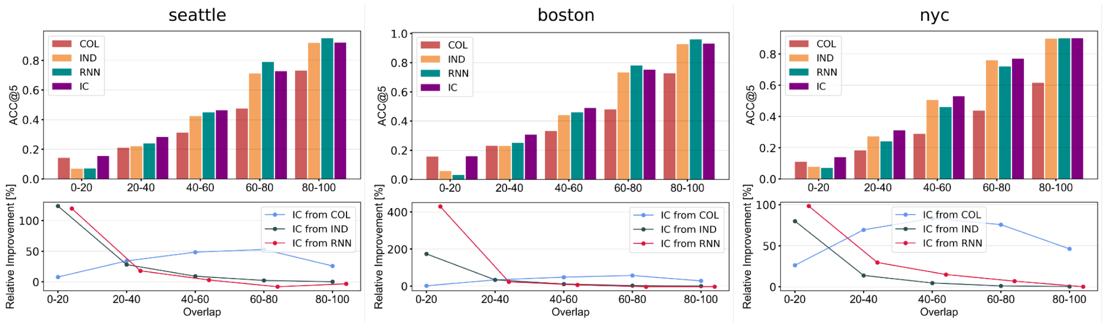


FIG. 2. **Accuracies in having the correct destination in the top 5 predicted locations with highest confidence (ACC@5):** Individual, Collective and IC Markov-based models are compared to predictions from R-NN deep models for NYC, Boston and Seattle in the pre-covid period. Models are tested in different scenarios for novel mobility seen during the prediction phase (quantified via the longest common subsequence (LCSS) overlaps between training and test sets). The Individual-Collective model shows a relative improvement in accuracy, with respect to the individual Markov model, up to 150 % in the case of small overlaps (test trajectories characterized by mostly novel transitions not seen in training), where the collective behavior's information aids the predictive capabilities.

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, Human mobility: Models and applications, Physics Reports **734**, 1 (2018).

[2] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, A survey on deep learning for human mobility, ACM Computing Surveys (CSUR) **55**, 1 (2021).

[3] F. Calabrese, G. Di Lorenzo, and C. Ratti, Human mobility prediction based on individual and collective geographical preferences, in *13th international IEEE conference on intelligent transportation systems* (IEEE, 2010) pp. 312–317.

Explaining Mobility Flows in Different Temporal Settings: City-Scale Deep Gravity

Fátima Velásquez-Rojas fatima@ifisc.uib-csic.es,[1] Massimiliano Luca
mluca@fbk.eu,[2, 3] and José Javier Ramasco jramasco@ifisc.uib-csic.es[1]

[1]*Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC), Palma de Mallorca, Spain*
[2]*Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy*
[3]*Free University of Bolzano, Piazza Università 1, 39100 Bolzano (BZ), Italy*

Cities occupy 2% of the surface of our planet and 50% of the world's population lives there. Also, cities are responsible for 75% of energy consumption and produce 80% of CO2 emissions [1]. Thus, understanding people's behavior at the scale of cities is paramount for many applications including, but not limited to, disease spreading, crime, and reduction of inequalities [2]. Nowadays, the recent advances in computer science and the adoption of ubiquitous technologies allow researchers to employ deep learning models on big mobility data to gain insights about the population's whereabouts [3]. While there are multiple models based on different techniques to predict the flow of people within cities[3], it remains unclear which are the dynamics that drive mobility flow in different places in different temporal settings. To answer this question, we leverage Deep Neural Networks (DNNs) and explainable AI tools (i.e., SHAP [4]), which are the spatial and geographical factors driving human mobility flows in different temporal settings (i.e., predominant factors in the weekdays and in the weekends). We conducted the study in 20 different cities worldwide collected through a Location-based Social Network in 2014-2021 (excluding 2020 due to the COVID-19 pandemic). Each city we investigate is tasseled into 1×1 km squares, and an Origin-Destination (OD) matrix is estimated. Given a city divided into n regions, the related OD is a matrix $OD = \mathbb{N}^n \times \mathbb{N}^n$ and $OD_{i,j}$ is the number of people going from i to j. Note that self-loops are not included in the dataset. Thus, we have diag$(OD) = 0$ for each city. For each area, we also download some geographic features from OpenStreetMap to characterize why people may decide to commute to a certain place. We model the flow using a Deep Gravity-like algorithm. In particular, we adapted the Deep Gravity model to operate on a city scale by removing the concept of tile presented in the original paper. Also, to avoid overfitting, the new model can dynamically adjust the number of hidden layers needed to predict/generate realistic mobility flows. We named this model City-Scale Deep Gravity (CSDG). The model is evaluated in terms of the Common part of Commuters [3] or, in short, CPC. CPC measures the similarity between real flows, y^r, and generated flows, y^g:

$$CPC = \frac{2\sum_{i,j} min(y^g(l_i, l_j), y^r(l_i, l_j))}{\sum_{i,j} y^g(l_i, l_j) + \sum_{i,j} y^r(l_i, l_j)} \tag{1}$$

It is always positive and contained in the closed interval $(0, 1)$, with 1 indicating a perfect match between the generated flows and the ground truth and 0 highlighting the bad performance with no overlap. Note that when the generated total outflow is equal to the real total outflow, as for all the models we consider in this paper, CPC is equivalent to the accuracy, i.e., the fraction of trips' destinations correctly predicted by the model. In Table I, we show the performances of the proposed model (CSDG) and a Gravity model (G). Regardless of the temporal setting (e.g., weekends - WE, weekdays - WD), our model outperforms the baseline. As suggested in the literature, this behavior is not unexpected, as deep learning techniques can capture non-linear patterns in the data and automatize the feature engineering process.

By leveraging XAI tools like SHAP [4], we can capture the role of geographical features on people's commuting decisions. The importance of the features, as captured with Shapely values, can be used to spot differences in the attractiveness of different geographical features in different temporal settings. In Figure 1, we can see an example of some preliminary results of SHAP values for the city of Madrid. In the Figure, we can see the feature importance of the geographical features collected using OpenStreetMap, population, and distances. The model's target is to generate flows that match the ones we collected from Twitter over 2014-2021 (2020 excluded).

In this abstract, we only have the opportunity to show some results on the city of Madrid, the city hosting NetMob '23, as an example of the results we obtained for all the cities. In Figure 1, we show the average importance of the features in the case of weekends (Figure 1.A) and weekdays (Figure 1.B). Also, we only report the top 8 features for the two temporal settings. The order in which the features appear from top to bottom is related to their importance (e.g., the first one is the most important while the last to appear is the 8^{th} most important). The arrows associated with the features can be blue or red. Red arrows are features that contribute positively to the generation of the flows (i.e., as bigger the feature associated with a red arrow, as big the generated flow will be). Similarly, features associated with blue arrows contributed toward a reduction in the generated flow. As bigger the number in the arrow is, as larger the impact of the associated feature. Note that such numbers are not comparable across cities. What emerge from SHAP is that in general, populations and distances tend to play an important role when modeling mobility flows at the city scale. However, in most cases, the most important features are geographical-related features (e.g., the number of POIs of a certain kind or types of land use).

As you can see, during the weekdays, in Madrid, the most important POIs that increase the flow of people between areas are the presence of schools (all levels, university included) and the presence of transportation systems in origin. Also, people do not travel to green spaces during the week. In other terms, people in Madrid tend to go to

	CSDG (WE)	CSDG (WD)	G (WE)	G (WD)
Barcelona	68.5	68.8	47.6	48.9
Bogota	64.8	62.4	49.2	49.1
Cairo	69.4	63.7	52.7	53.1
Chicago	63.2	63.8	51.1	54.3
Copenhagen	49.8	49.9	41.8	41.5
Detroit	62.8	64.8	53.9	51.7
Istanbul	57.4	53.3	47.2	45.8
Jakarta	71.8	68.9	61.3	60.6
Los Angeles	66.4	68.1	55.2	55.6
London	62.7	62.7	49.8	49.4
Manila	69.9	62.3	50.7	50.2
C. de Mexico	54.6	55.8	41.8	44.3
Moscow	59.8	60.0	42.8	42.9
New York City	63.2	57.9	47.4	48.3
Osaka	58.6	61.1	52.7	53.4
Paris	55.3	55.9	47.6	48.0
Sao Paulo	62.7	63.2	50.9	51.2
Seattle	58.2	59.8	44.9	44.9
Stockholm	59.6	59.5	51.9	51.9
Tokyo	57.8	60.2	49.8	49.9

TABLE I. Accuracies of Gravity model (G) and our model (CSDG) during weekends (WE) and weekdays (WD).

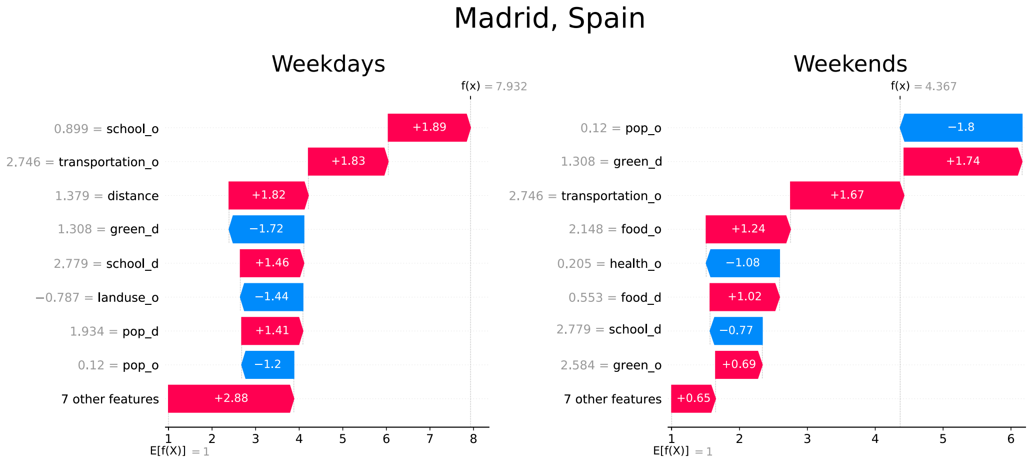


FIG. 1. Distribution of Shapely values for all features used for Madrid on A) Weekdays and B) Weekends.

schools and it appears that people prefer to commute public transportation in the city (maybe due to the excellent metro network). On the other hand, in Madrid, people prefer outdoor activities and areas with food-related POIs (e.g., bars or restaurants). For instance, it is possible to see that green spaces in the destination contribute negatively during the week while they become the most important positive contribution during the weekends. Also, food-related POIs appear in the top 8 most important features. Interestingly, other features change their role in these two different temporal settings. An example is schools that, during the weekends, contribute to a decrease in the flow. It is not surprising is information that can be used to validate the Shapely values we are obtaining. Similarly, some features remain important and do not change their role. Examples are the populations in origin and destinations and the role of public transportation. While analyzing a single city is interesting but not informative, clustering the cities according to their shapely values may shed some light on potentially similar functional roles that different cities may have. At the same time, if we focus on single commuters between specific origins and specific destinations, we may be able to cluster cities according to the behaviors of people living there. Developing such clustering techniques may provide unique insights to policymakers and urban planners, and it is an ongoing work we will present in the paper associated with this study. Finally, outcomes of our study suggest that fine-tuning generative models for flow generation without stratifying mobility profiles (e.g., temporal settings, socio-demographic settings, others) may lead to an unprecise generation of mobility flows as features' weights strongly depend on the mobility profile we want to model. Future works may validate the obtained shapely values by directly looking at city structures.

[1] W. Carlo Ratti, These four numbers define the importance of our cities: 2, 50, 75 and 80 (2016).

[2] M. Luca, G. M. Campedelli, S. Centellegher, M. Tizzoni, and B. Lepri, Crime, inequality and public health: a survey of emerging trends in urban data science, Frontiers in Big Data **6**, 1124526 (2023).

[3] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, A survey on deep learning for human mobility, ACM Computing Surveys (CSUR) **55**, 1 (2021).

[4] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 4765–4774.

A methodology for studying residential migrations: application to the analysis of migrations from the Madrid region

Raquel Sánchez-Cauce - Nommon Solutions and Technologies - raquel.sanchez@nommon.es
Jorge Mallo Fuentes-Lojo - Nommon Solutions and Technologies - jorge.mallo@nommon.es
Javier Burrieza-Galán - Nommon Solutions and Technologies - javier.burrieza@nommon.es
Oliva Cantú Ros - Nommon Solutions and Technologies - oliva.garcia-cantu@nommon.es
Ricardo Herranz - Nommon Solutions and Technologies - ricardo.herranz@nommon.es
Miguel Picornell Tronch - Nommon Solutions and Technologies - miguel.picornell@nommon.es
Juan Carlos García Palomares - Universidad Complutense de Madrid - jcgarcia@ghis.ucm.es
Gustavo Romanillos Arroyo - Universidad Complutense de Madrid - gustavro@ucm.es
Enrique Santiago Iglesias - Universidad Complutense de Madrid - ensantia@ucm.es

The COVID-19 crisis has caused an irruption of teleworking that, among other effects, has modified the residence patterns of the population. By leveraging the rich longitudinal information provided by passively collected data from mobile networks, we have analysed residential migrations in the Madrid Region since the start of the COVID-19 pandemic. This study is part of the DARUMA[1] research project, developed in the frame of the EIG Concert-Japan cooperation initiative, whose general goal is to study and predict the impact of disruptive events on social behaviour.

The methodology to identify residential migrations is based on Nommon's solution for obtaining activity patterns information from anonymised mobile network data (Population Insights[2]). The solution generates 'activity diaries' for the sampled mobile phone users and expands them to the total population using census data[3], providing a detailed activity characterisation, including the activity type (home, work, education, etc.), length of stay, etc. This information is used to compare the home location of the common sample of users in two time periods. Based on this comparison, residential migrations are identified.

The use of mobile network data makes it possible to capture movements that are not recorded in official statistics and to observe the influence of factors such as income level on residential mobility patterns, constituting a powerful tool to monitor and analyse the behaviour of the housing market.

The analysis of migratory flows in the Madrid Region considers the period 2020-2022. In particular, a week of February 2020 and the same week of February 2022 were chosen as reference to perform the comparison. The analysis shows that 19% of the people living in the Madrid Region in February 2020 changed their residence location by February 2022. From those, 25% move to other Spanish regions.

The results reveal that the zones that attracted more population were two bordering provinces with Madrid: Toledo (with 15.8% of the external migrations) and Guadalajara (with 6.1% of the external

[1] https://concert-japan-daruma.github.io/
[2] https://www.nommon.es/products/population-insights/
[3] For a detailed explanation of the methodology for obtaining activity-travel diaries from the sample of mobile phone users and an expansion to the total population see Bassolas, A., Ramasco, J. J., Herranz, R., & Cantú-Ros, O. G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transportation Research Part A: Policy and Practice, 121, 56–74.

migrations). This group of migrants has an average income belonging to the second income quintile of the Madrid Region. These migrations are probably motivated by the combination of proximity to the capital and much cheaper housing offer. It is also worth noting the migratory flows towards the Mediterranean coast, which account for 18.4% of the total number of external migrations, probably due to second residences of those people who can work from home. In this case, this group of migrants has an average income belonging to the fourth income quintile of Madrid.

Regarding the migrations within the Madrid Region (Figure 1), the centre of the City of Madrid is the area that experienced the greatest reduction in population. Other big cities in the surroundings of Madrid, such as Móstoles, Fuenlabrada, Majadahonda and Pozuelo de Alarcón, also reduced their population due to these migration flows. On the contrary, areas near the Madrid Mountain Chain, such as El Molar and Collado Mediano, received new inhabitants. This is probably because, with the option of teleworking, many people preferred to move to cheaper and more liveable areas. We also identified other regions near the City of Madrid, such as Navalcarnero or San Fernando de Henares, as large recipients of migrants, which shows the attractiveness of new urban developments in these areas.

This study is still ongoing. As future steps, we will analyse the impact of sociodemographic characteristics such as age, gender, and income on migratory flows.

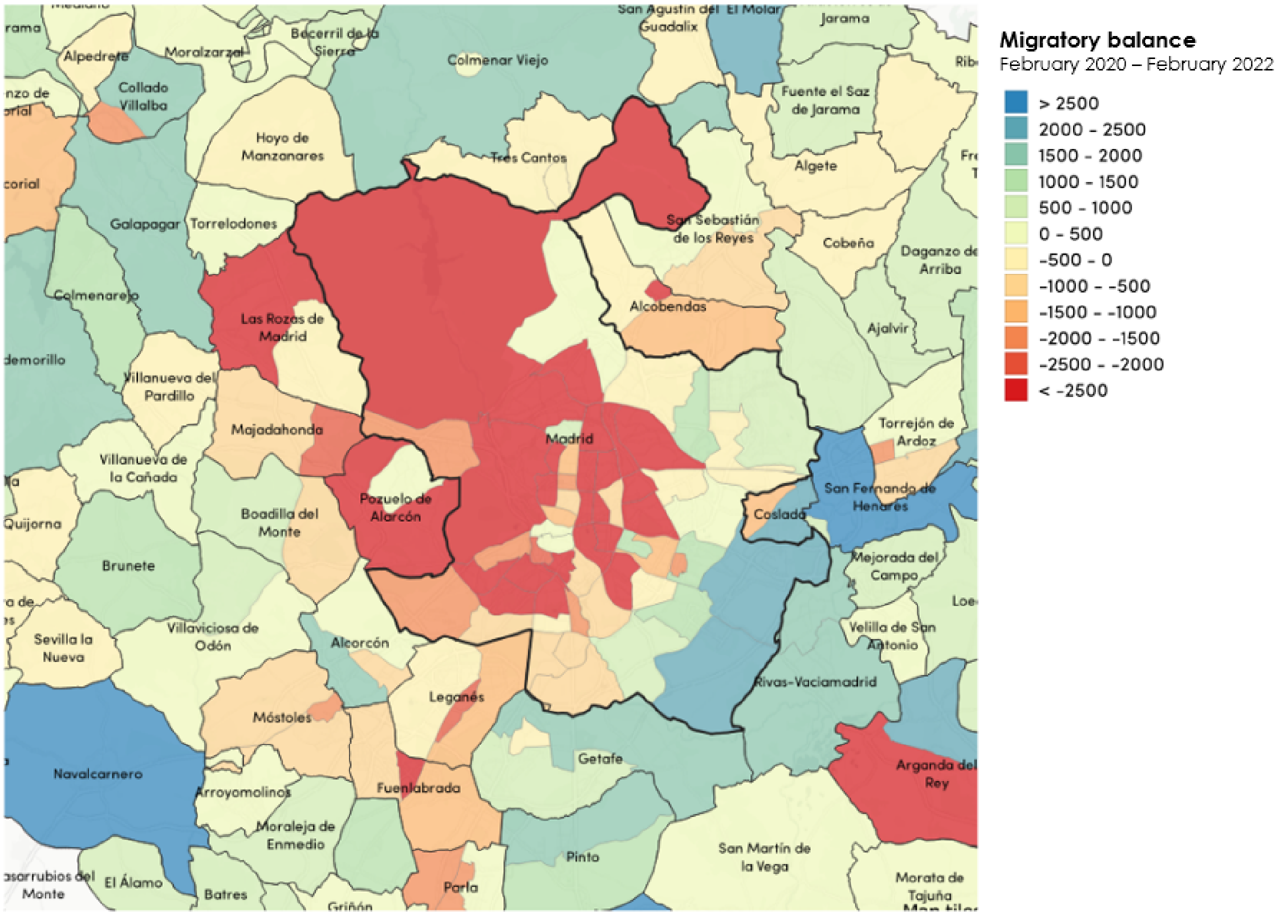


Figure 1. Migratory balance in the Madrid region between February 2020 and February 2022.

A new, easy-to-interpret and fast prediction algorithm inspired by the next-location prediction problem

Kamil Smolak[1]

[1] *Institute of Geodesy and Geoinformatics, Wrocław University of Environmental and Life Sciences, Wrocław, Poland ,*
email: kamil.smolak@upwr.edu.pl

1 Introduction

Next-location prediction is a task of forecasting future whereabouts of an individual. It is relevant for multiple applications such as traffic forecasting, urban planning, public health and well-being [1].

The next-location prediction is based on the mobility sequence of an individual, where this sequence represents movement of a person. It is a series of symbols, where each symbol represents a visited location. The goal of the next-location prediction is to predict the next location which a person will visit, based on historical data. In practice, the goal is to predict what will be the next symbol in the sequence. Therefore, this task can be treated as a multi-class classification problem.

Many studies were devoted to finding a prediction approach which will eventually lead to increased accuracy of the next-location predictions. This has resulted in a plethora of works which applied various algorithms to this problem, spanning from Markov Chains, through machine learning models, up to complex deep learning approaches. Because of the complexity of human mobility, increasing advancement of applied prediction algorithms led to increased accuracy of predictions. Complex prediction algorithms are able to extract and capture hidden mobility patterns, which are useful in the next-location prediction task. Simpler models, such as Markov Chains, are unable to capture long-distance dependencies (LDDs) exhibited in human mobility [2]. These are the patterns observed on significant distances appearing in mobility sequences, which carry important information. When disregarded, this may result in the under-performance of models. More complex models, such as neural networks, can model LDDs, which is the main reason for their good performance.

On the other hand, complex models suffer from low interpretability, which impedes analyses of factors driving the movement of humans. Another problem is the training time of these models, which is usually long. To address these problems, I propose a new sequence prediction algorithm, which is easy to interpret, has fast training times, and most importantly, in many cases outperforms complex solutions. This algorithm addresses the inability of Markov Chains to capture LDDs and is their extension to a more general case. The core of the algorithm is based on the method used to calculate *equally sparse repeatability* (*ESR*) metric [3].

This conference paper presents only part of the algorithm, which is already developed. However, it misses important optimisations which will further improve its accuracy and computation time.

2 Data

The tests were conducted on an actual human mobility dataset as well as on generated sequences. The human mobility dataset was collected from mobile devices of people living in Rio de Janeiro, Brazil. The used subsample of the dataset consists of 130 randomly chosen trajectories, collected over a period ranging from 17 to 22 days. Trajectories were processed into mobility sequences, with an average length of 80 symbols.

Generated sequences follow the deterministic sequence of symbols where $x_1 \to x_2 \to ... \to x_n \to x_1 \to$ ... with probability p. Their length varied in range from 100 to 400 symbols, with 1 to 30 unique symbols, and p from 0.1 to 0.9.

3 Method

A mobility sequence T_u of a person u consists of symbols p_t representing visited location at time t. The algorithm, which is called a Sparse Chain (SC), utilises common subsequence detection to search for patterns in the user's mobility sequence. The algorithm iterates through the T_u, at each iteration splitting it into two subsets, T_{u1} and T_{u2}, at index $i \in \{1, t-1\}$, where $t \in \{0, N\}$ and N is the sequence length. The goal is to find occurrences of common subsequences in T_{u2} and T_{u1}, with a condition that symbols in matched subsequences appear in the same order and are at the same relative positions. The algorithm searches for all matching subsequences of any length, starting at any index in T_{u2}. For each match, a symbol following the matched subsequence in T_{u1} is assigned to the match as a candidate for prediction. The process is also repeated to find all subsequences of T_{u1} appearing in T_{u2}. All detected subsequences, with matching symbols, their relative positions, and symbols following these matches are recorded within the model and used for the prediction.

Example 1 *Given*
$T_{u1} = [A, B, C, D, A, B, C, C, D]$ *and*
$T_{u2} = [A, B, A, D, A, B, A, D]$ *one of the subsequences meeting the requirements would be* $[A, B, \_, D, A, B, \_, \_]$ *starting at index* 1 *in* T_{u2}*. "_" are empty symbols that were not matching. The symbol following the match in* T_{u1} *is* D*, at index* 9*.*

At the prediction, the algorithm is given a context in the form of a past mobility sequence T_u with the last position recorded at time t. The algorithm predicts the next symbol at time $t+1$. Context is matched with the learned patterns. Patterns overlapping with the context are assigned a score, which is the total number of overlapping symbols in the pattern. Scores are then grouped by each candidate symbol. The symbol with the highest score is predicted. Additionally, scores are weighted by the length of the matched pattern and the recency of the pattern. The former is motivated by the importance of longer patterns over short matches. The latter is related to the existence of LDDs, which decay over time, hence patterns matched with the most recent symbol in the context should be given higher scores.

Example 2 *We are given a context* $[B, A, B, A, C]$ *and a model with a pattern* $[A, B, \_, D, A, B, \_, \_]$ *associated with a* D *symbol which follows it. In this case, only symbols at indices* 2, 3 *of context and* 5, 6 *of the pattern are matching. Therefore, a* D *symbol would be given a score of two.*

4 Results

Sequences were split into training and test sets in a such way that training set stated for 80% of the mobility sequence. First, 20% of the training set was used for validation in order to select the best weights. Then, the whole training set was used to fit the algorithm. Presented results were calculated on the test set. We compare the results of SC algorithm to other widely used models, namely Higher-order Markov Chains (MC), Random Forest (RF), Gated Recurrent Unit with Embedding (GRU), Bidirectional Long short-term memory with Embedding (Bi-LSTM), and Bidirectional Gated Recurrent Unit with Embedding (Bi-GRU). Results are presented in Table 1 and Table 2.

Table 1: Next-location prediction results for compared algorithms calculated using the generated dataset. The best score for each metric is marked in bold.

	ACC@1 [%]	ACC@3 [%]	F1-score [%]
MC	39.6	49.0	39.3
RF	43.9	56.1	43.9
GRU	39.7	54.1	39.7
Bi-LSTM	34.4	51.3	34.4
Bi-GRU	42.8	55.5	42.8
SC	**51.1**	**61.4**	**51.1**

Table 2: Next-location prediction results for compared algorithms calculated using the human mobility dataset. The best score for each metric is marked in bold.

	ACC@1 [%]	ACC@3 [%]	F1-score [%]
MC	33.0	46.3	32.9
RF	26.0	43.4	26.0
GRU	32.2	**48.3**	32.2
Bi-LSTM	27.4	42.2	27.4
Bi-GRU	27.4	40.3	27.4
SC	**34.0**	47.1	**34.0**

The SC algorithm in most cases outperforms other applied algorithms. The generated dataset was designed to demonstrate the ability of the SC algorithm to capture patterns in sequences in the presence of noise introduced by the p parameter. Specifically, when the value of the p parameter was high, algorithms performed with similar accuracy. However, when p decreased, hence noise was introduced, SC performed better than other algorithms. The human mobility dataset consisted of relatively short sequences, which especially impacted the performance of neural networks and RF. MC and SC algorithms are better suited to learn patterns from short sequences, and therefore, they perform better on this dataset. Another important note is the training times necessary to fit all the models, which were much shorter for the SC algorithm (from 20 to 30 seconds) than those required by neural networks (from 15 to 70 minutes) and RF algorithm (13 to 18 minutes).

In the future works, SC will be validated on a wider range of datasets, preferably benchmarks, and compared to other state-of-the-art algorithms. Moreover, the interpretability of the SC algorithm will be fully explored. Each detected pattern and the total score assigned to each pattern at each prediction step can be analysed. This can be used to extract the most impactful patterns in the mobility sequence and provide insights into the driving factors of human mobility. It is important to note, that the applicability of the SC algorithm is far beyond the human mobility prediction area, as it can be utilised on any sequential dataset.

References

[1] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–44, 2021.

[2] V. Kulkarni, A. Mahalunkar, B. Garbinato, and J. D. Kelleher, "Examining the limits of predictability of human mobility," *Entropy*, vol. 21, no. 4, p. 432, 2019.

[3] K. Smolak, W. Rohm, and K. Sila-Nowicka, "Explaining human mobility predictions through a pattern matching algorithm," *EPJ Data Science*, vol. 11, no. 1, p. 45, 2022.

A PROCESSING PIPELINE FOR EUROPEAN OFFICIAL STATISTICS:

TOWARDS STANDARDISATION OF MOBILE NETWORK OPERATOR DATA PROCESSING

Authors:

Tiziana Tuoto (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), tuoto@istat.it
Roberta Radini (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), radini@istat.it
Paolo Mattera (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), paolo.mattera@istat.it
Erika Cerasti (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), erika.cerasti@istat.it
Cristina Faricelli (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), cristina.faricelli@istat.it
Giorgia Simeoni (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), simeoni@istat.it
Gabriele Ascari (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), gabascari@istat.it
Luca Valentino (Istituto Nazionale di Statistica / Italian Statistical Institute - ISTAT), luvalent@istat.it
Matthias Offermans (Statistics Netherlands - CBS), mpw.offermans@cbs.nl
Edwin de Jonge (Statistics Netherlands - CBS), e.dejonge@cbs.nl
Ricardo Herranz (Nommon Solutions and Technologies), ricardo.herranz@nommon.es
Margus Tiru (Positium), margus.tiru@positium.com
Florabela Carausu (GOPA Worldwide Consultants), florabela.carausu@gopa.de
Miguel Picornell (Nommon Solutions and Technologies), miguel.picornell@nommon.es
Villem Tonnison (Positium), villem.tonisson@positium.com
Cristina Escribano (Nommon Solutions and Technologies), cristina.escribano@nommon.es

- The views in this abstract are those of the authors and do not necessarily reflect the position of the European Commission (EC) or national statistical institutes-

Abstract:

The European Statistical System (ESS) – the partnership between the EU statistical authority (Eurostat) and national statistical institutes (NSI) and other statistical authorities in the European member states – considers Mobile Network Operator (MNO) data as one of the most promising new data sources for future statistical production. The production of official statistics based on MNO data has the potential to provide considerable societal value. In this context, the ESS emphasises the need for standardised reference methods adhering to the principles of statistical production, such as quality, privacy protection, and transparency.

In line with the ESS Innovation Agenda, following an open call for tenders[1], in December 2022, Eurostat awarded the service contract "*Development, implementation and demonstration of a reference processing pipeline for the future production of official statistics based on Multiple Mobile Network Operator data (TSS multi-MNO)*". The project is a significant milestone towards the future reuse of MNO data for the production of official statistics at EU level. The goal of the project is to develop a complete, open end-to-end processing pipeline that should serve as starting point towards the regular production of future official statistics based on MNO data Europe-wide. This "processing pipeline" encompasses a combination of a fully documented open methodological and quality framework, plus the implementation of a reference open-source software pipeline compliant with the said framework. The processing pipeline will be demonstrated across data from multiple MNOs. If successful, the reference pipeline developed by the project will be proposed for adoption by the ESS as a methodological standard.

[1] Reference ESTAT/2022/OP/001. Technical specifications available at: eTendering - Data (europa.eu)

The project is being implemented by a consortium providing extensive experience from both the business and the official statistics domains. The consortium is composed by GOPA Worldwide Consultants GmbH (DE) - lead, Nommon Solutions and Technologies SL (ES), Positium OÜ (EE), Statistics Netherlands (NL) and the Italian Statistical Institute (IT). Additionally, five European MNOs from four distinct countries will be involved in the testing of the pipeline.

This collaborative endeavor aligns with the European Data Strategy's goal of providing comparable and reliable statistics across European countries. The project addresses the challenge of providing open and standardised methodologies for official statistics, without hampering the development of future private initiatives, nor the continuation of the range of analytic products based on MNO data that have been developed and commercialised by mobile operators or other third-party entities for purposes other than European official statistics.

While the project is financed by Eurostat (the EU statistical office), its ultimate success will depend on the potential endorsement of the project result by the larger ESS community (integrating all EU statistical offices and other national authorities). It is expected that this will have positive implications for future activities and may serve as a model that can be replicated in other domains, along with seeking a closer collaboration with industry or business partners, more in general, in the context of initiating or strengthening co-development undertakings for the production of official statistics.

This contribution will focus on the presentation of the overall pipeline architecture, and the description of an initial version of the processing pipeline. The architecture design will adhere to highest technical requirements and methodological soundness. The proposed pipeline considers the division between data processing at the MNO environments and additional processing steps at the NSI or other parties.

The software will be divided into modules for (1) the processing of disaggregated data exclusively at each MNO' secured environment, and (2) the post-processing of aggregated and anonymous data at national statistical offices. The latter is particularly relevant since the post-processing will be performed on aggregated data after the application of statistical procedures, such as Statistical Disclosure Control (SDC), that ensure that individual data cannot be referenced back. Comprehensive documentation, including functionality, implementation details, and usage instructions, will accompany the software. Reference test data, consisting of synthetic or semi-synthetic samples, will be created for each software module to ensure reproducibility and ease the development of alternative but fully compliant software implementations by independent entities.

The entire open-source pipeline, including the codes and related documentation, as well as the methodological framework will be openly published. The software codes will be published under EUPL license, promoting transparency and accessibility, facilitating the replication and adoption of the developed software solutions, encouraging collaboration and further advancements in the field of statistical production.
The reference implementation of the pipeline will be public and results will be communicated to interested audience through public official channels.

Flowminder standards in producing mobility and population estimates from call details records in low- and middle-income countries

Véronique Lefebvre*[a], James Harrison[a], Jonathan Gray[a], Roland Hosner[a], Galina Veres[a], Chris Brooks[a], Robert Eyre[a], Thomas Smallwood[a], Romain Goldenberg[a], Zachary Strain-Fajth, Joachim Jellinek, John Roberts, Xavier Vollenweider, Sophie Delaporte, Cathy Riley, Daniel Power, Linus Bengtsson

Flowminder Foundation

* Corresponding author. email: veronique.lefebvre@flowminder.org

[a] Equal significant contribution

Processing and analysing CDRs in LMICs: Extracting the population mobility information contained in Call Detail Records (CDRs) is of critical importance in data poor contexts such as in low- and middle-income countries (LMICs), where it can support humanitarian and human development efforts. Such contexts however present additional challenges compared to high-income countries (HICs) for mobility analysis from mobile operator data: often only CDRs are available and they are sparser over time and space, mobile networks are more unstable, particular in crises, which are often more frequent, and the geographic coordinates of cells are sometimes missing and erroneous. Further, the proportion of the general population using mobile phones is significantly lower in LMICs (e.g. 47% of households on average in 7 provinces of the DRC, down to 35% in the more rural provinces) and therefore differences in the mobility of phone users and non users have a larger impact on the representativity and applicability of CDR-derived statistics. At Flowminder we have specialised in addressing such challenges and we present here an overview of our live systems, from ingestion and automated quality assurance (QA) checks of pseudonymised CDR data and cell data, to the extraction of mobility information from CDRs and bias correction using survey data, resulting in the semi-automated production of a set of standard indicators, ready to be disseminated to decision makers in LMICs through dashboards, standard reports or as data sheets.

Flowminder system: At Flowminder we made the choice to conduct all CDR data processing within the firewall of the Mobile Network Operator (MNO). While this comes with constraints on compute power and memory, it is essential to protect subscribers' data privacy. We also ensure we do not have access to subscriber phone numbers and instruct MNOs on how to pseudonymise the records. We built the ‘FlowKit software’ to handle all data processing, from ingesting the pseudonymised CDRs to outputting mobility and population estimates. FlowKit is an open-source CDR data processing toolkit, consisting of a PostgreSQL database along with tools for automating data ingestion and QA, implementations of our methods for extracting mobility information from CDRs, scaling, combining and formatting the mobility estimates for end usage. We describe below the different steps of the pipeline as per our overview diagram in **Figure 1**.

Pseudonymised CDR ingestion and checks: Pseudonymised CDR data are received from the MNOs daily, and are automatically ingested into the FlowKit database, ready for processing. QA checks are run on the CDR data during ingestion to ensure data quality, and exclude incomplete or unreliable data from analysis. Cell locations are received from the MNO at regular intervals (ideally monthly), and QA checks are performed to identify mislocated cells. We also set up a monitoring system that tracks the status of our at-MNO servers, FlowKit installations and automated QA checks, and sends alerts so that any detected problem can be addressed quickly. This is important particularly for crisis preparedness so that CDRs are available and specific processing can begin promptly if a crisis occurs or is forecast.

Mobility extraction and bias correction: We use FlowKit to process the pseudonymised CDR data within the MNO's firewall to extract mobility information. Recent method development led to a more robust detection of subscriber’s home (or ‘stay’) location from sparse CDRs, and estimates of relocations. Further we collect field and phone survey data to correct for representation biases in the number of relocations between each subregion summed over subscribers. We then derive monthly population estimates (residents) from the adjusted relocations and from baseline population estimates, which scales the relocations and attenuates the effect of phone usage changes and subscriber churn on changes of resident numbers. Similar work to disambiguate between phone usage and mobility for population presence and travel and to scale these indicators is ongoing.

Reporting and dissemination: We have standardised the way we provide analytical reports and data (via datasheets or dashboards) to inform a range of applications, from disaster management, to official migration statistics and immunisation planning, and are working towards further automating the production of these end products for each application. Currently we produce mobility and population estimates for Haiti, Ghana and the DRC.

Flowminder standards: While specific to CDRs and to data issues encountered in LMICs, we propose that our infrastructure characteristics, QA checks, automation framework, methods for mobility information extraction and for correcting representation biases, as well as our standard set of mobility indicators may be of use to anyone attempting to produce mobility statistics from CDRs or related data types in any country.

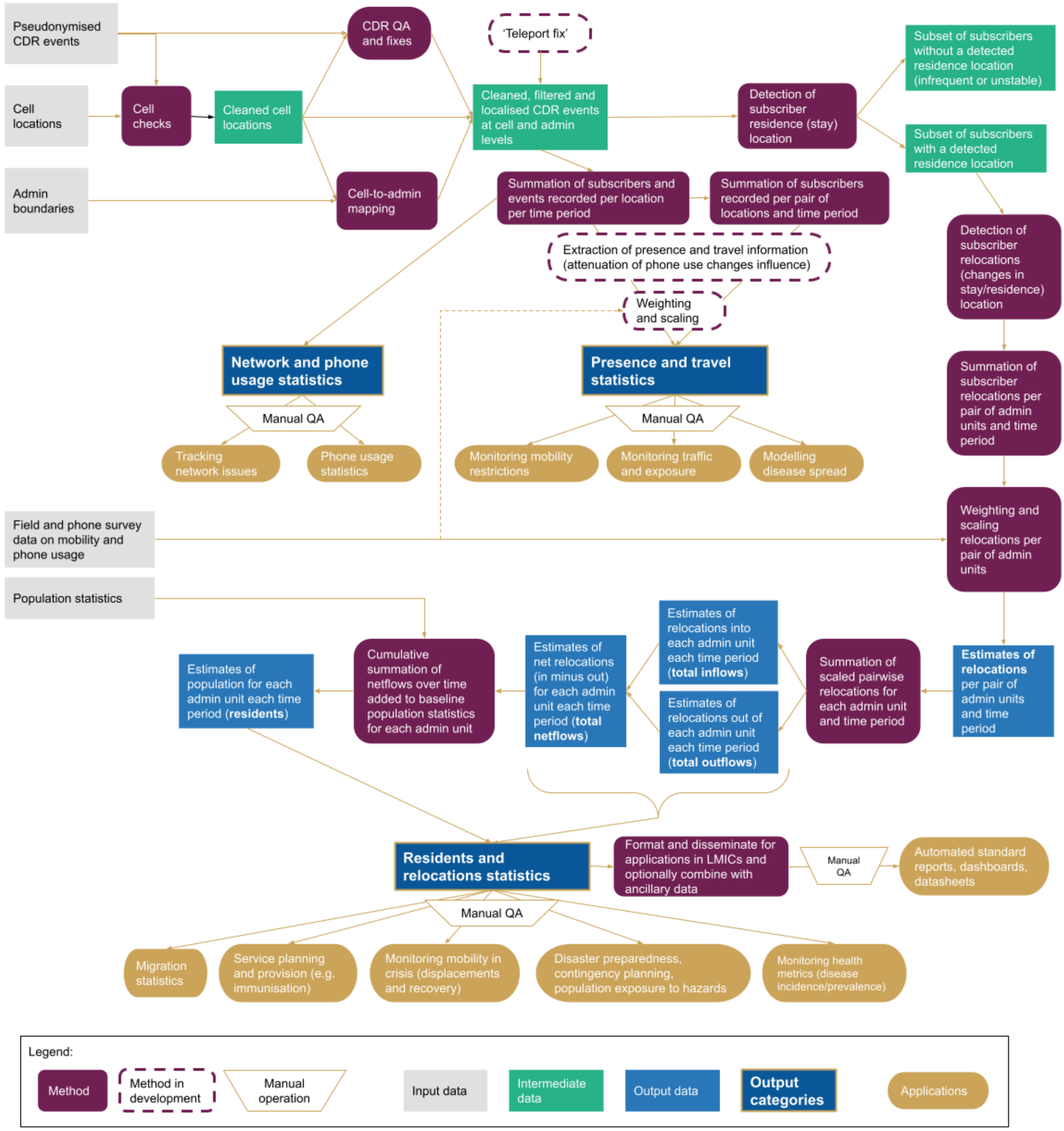

Pseudonymised CDR events
Cell locations
Admin boundaries
Cell checks
Cleaned cell locations
CDR QA and fixes
Cell-to-admin mapping
'Teleport fix'
Cleaned, filtered and localised CDR events at cell and admin levels
Detection of subscriber residence (stay) location
Subset of subscribers without a detected residence location (infrequent or unstable)
Subset of subscribers with a detected residence location
Summation of subscribers and events recorded per location per time period
Summation of subscribers recorded per pair of locations and time period
Extraction of presence and travel information (attenuation of phone use changes influence)
Weighting and scaling
Detection of subscriber relocations (changes in stay/residence) location
Network and phone usage statistics
Presence and travel statistics
Manual QA
Manual QA
Tracking network issues
Phone usage statistics
Monitoring mobility restrictions
Monitoring traffic and exposure
Modelling disease spread
Summation of subscriber relocations per pair of admin units and time period
Field and phone survey data on mobility and phone usage
Weighting and scaling relocations per pair of admin units
Population statistics
Estimates of population for each admin unit each time period (residents)
Cumulative summation of netflows over time added to baseline population statistics for each admin unit
Estimates of net relocations (in minus out) for each admin unit each time period (total netflows)
Estimates of relocations into each admin unit each time period (total inflows)
Estimates of relocations out of each admin unit each time period (total outflows)
Summation of scaled pairwise relocations for each admin unit and time period
Estimates of relocations per pair of admin units and time period
Residents and relocations statistics
Format and disseminate for applications in LMICs and optionally combine with ancillary data
Manual QA
Automated standard reports, dashboards, datasheets
Manual QA
Migration statistics
Service planning and provision (e.g. immunisation)
Monitoring mobility in crisis (displacements and recovery)
Disaster preparedness, contingency planning, population exposure to hazards
Monitoring health metrics (disease incidence/prevalence)
Legend:
Method
Method in development
Manual operation
Input data
Intermediate data
Output data
Output categories
Applications

Figure 1: Flowminder pipeline for processing and analysing call detail records to inform humanitarian and development applications in low- and middle-income countries.

Enabling the Ghanaian National Statistics Office to produce official statistics derived from mobile operator data with applications across the public sector

Romain Goldenberg[1]*, Thomas Smallwood[1], Cathy Riley[1], James Harrison[1], Richard Attandoh[1], Omar Seidu[2] and Veronique Lefebvre[1]

[1] Flowminder Foundation, [2] Ghana Statistical Service

* Corresponding author. email: romain.goldenberg@flowminder.org

Abstract

Human mobility has important implications for decision-makers working across a broad range of sectors, including those working in government. Anonymised and aggregated mobile operator data provide near-real time insights into human mobility at high spatial and temporal resolution across a whole country and, after correcting for representation biases using survey data, can support decision-makers across a broad range of applications. However, there are a number of barriers to the use of mobile operator data by national statistics offices (NSOs) and other government agencies, including gaining access to data and the technical capacity to effectively process and analyse the data while preserving the individual privacy of subscribers.

To integrate mobile operator data into official statistics, and support branches of the government in using novel data sources for improved decision-making, Ghana Statistical Service (GSS), Vodafone Ghana and Flowminder entered a unique and long-standing partnership - the Data for Good partnership - for the wellbeing of all in Ghana. Within this partnership, Flowminder has been strengthening the capacity of GSS to process and analyse mobile phone data to produce routine mobility statistics and support other parts of the Government, independently. Flowminder is using its tools and expertise in the secure and privacy-preserving processing and analysis of call detail record (CDR) data to to set up a data and analytical pipeline for GSS to eventually run independently, receiving daily pseudonymised CDRs and extracting and bias-correcting mobility indicators of migration and travel. These indicators will then be integrated into decision-making in ministries, departments and agencies (MDAs) across the Ghanaian government.

In order to support GSS and other MDAs, the Data for Good partnership has focussed on delivering CDR-derived indicators for several applications including official mobility statistics, dynamic public health indicators, and dynamic disaster risk mapping which we describe below.

(1) As a result of this joint effort, a standardised mobility product has been created, primarily aimed at comprehending overall mobility patterns in the country and for key administrative areas of interest. It is intended to be generated at regular intervals (6 month periods) and in an automated manner, to ensure consistent and timely updates on a collection of recent mobility trends, such as migrations or pendular movements (Fig. 1).

(2) Flowminder and GSS are also collaborating with Ghana Health Service (GHS) and the Ministry of Health to identify public health metrics which may be improved by incorporating dynamic estimates of population density. Many public health metrics, including metrics of infectious disease such as disease incidence and prevalence, are calculated using the population of an area as a denominator (i.e. are calculated per capita). However, the traditional data sources most often used to estimate population size, such as censuses or surveys, only provide a static snapshot of the population, which ignores seasonal variations and migration trends, and therefore may be inaccurate and outdated. Flowminder is currently producing four experimental health metrics, covering infectious and non-infectious disease incidence, vaccination coverage and resource allocation, with monthly district-level population estimates.

(3) The National Disaster Management Organisation (NADMO) has requested support from Flowminder and GSS to help improve disaster preparedness by incorporating dynamic population estimates into risk analyses. Exposure, or the number of people in an area who may be exposed to a given disaster, is an important dimension of risk. However, exposure can vary as people move in or out of an area over both the short- and long-term (e.g. commuting and migration, respectively). Flowminder has produced preliminary dynamic flooding risk maps capturing the change in district-level flooding risk with hourly resolution for different days of the week, covering the whole of Ghana. The analyses demonstrate that CDRs can effectively be used to estimate the large variations in the number of people who may be exposed to flooding, and therefore the risk, in some districts depending on the time of day and day of the week due notably to commuting movements (Fig. 2). This variation in risk can have important implications for disaster preparedness. Future work will aim to ensure CDR-derived hourly metrics are corrected for biases related to phone usage variations and for representation biases so that the resulting risk maps can be used operationally.

The Data for Good partnership is a ground-breaking example of a sustainable system for the use of CDR data by a National Statistics Office and other MDAs to support decision-makers across multiple sectors in a middle-income country. Furthermore, partners are continuing to develop the applications described here and Flowminder is also testing additional applications, such as combining CDR data and other geospatial data for predictive modelling of poverty indicators.

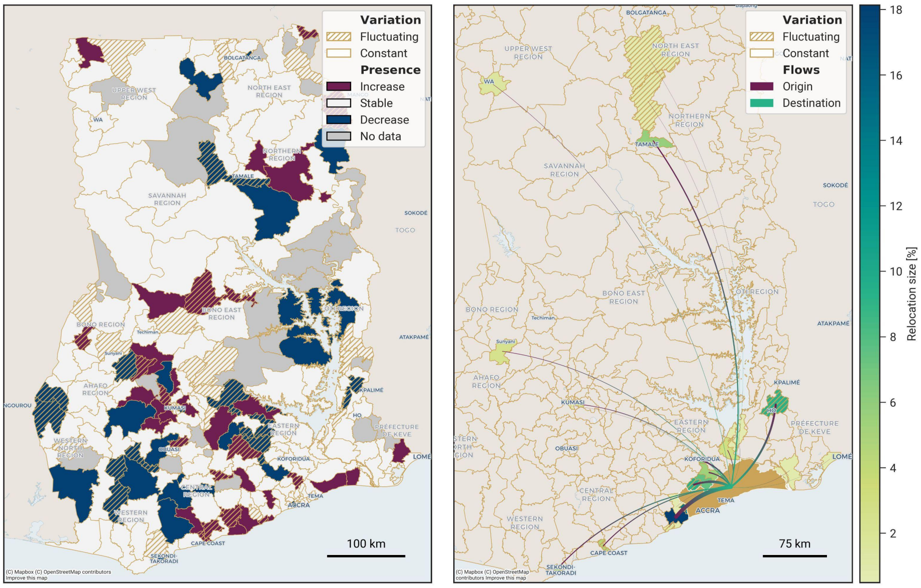


Figure 1. *Trends in residents counts for each district in Ghana (resulting from internal migration in the country) are shown on the left-hand side. On the right-hand side is an example of such relocations, showing migrations to the Greater Accra region. Both figures are averages over the period 2020 to 2022.*

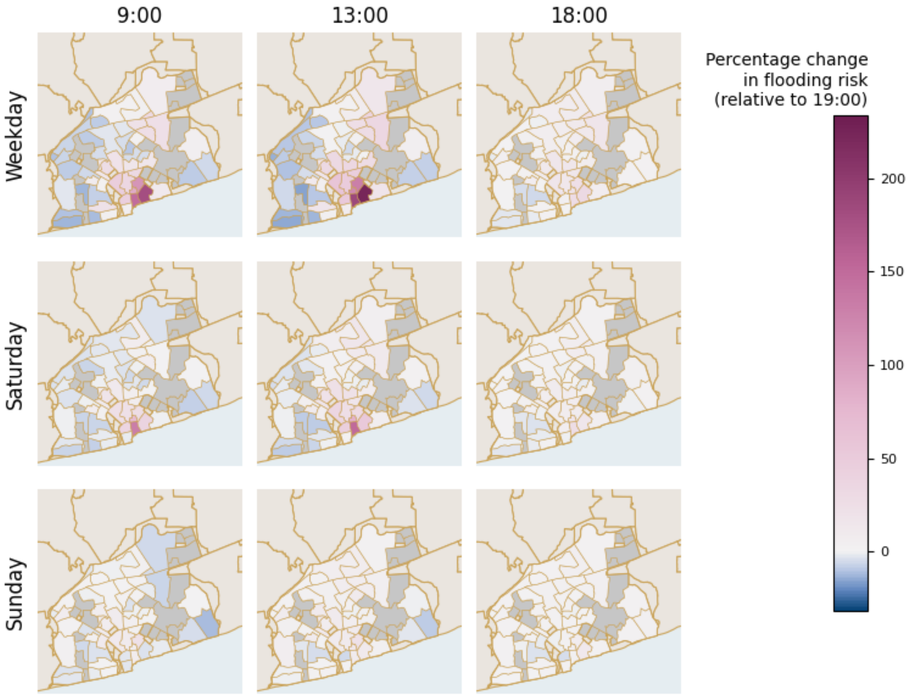


Figure 2. *Percentage change in flooding risk in central Accra at different times of day and different days of the week, relative to the risk at 19:00.*

SIMBox fraud: How well can *they* mimic *your* communication behavior?

Anne Josiane Kouam*, Aline Carneiro Viana*, Alain Tchana†
* INRIA, France. † Grenoble INP, France.
E-mail: {anne-josiane.kouam-djuigne, aline.viana}@inria.fr*, alain.tchana@grenoble-inp.fr†

***Abstract*—Being one of the most prevalent scams in cellular networks, *SIMBox* fraud causes a significant financial loss, national security threats, and phone conversation privacy breaches. Yet, *SIMBox* fraud is still an open issue being little addressed and hardly detected by operators due to: (c1) the scarcity of ground-truth fraudulent datasets and (c2) the constant evolution of fraudulent strategies aiming to disguise by mimicking legitimate communication behaviors. This paper introduces the *FraudZen* framework to tackle (c1) by generating mobile communication datasets (i.e., Charging Data Records/ CDRs) with realistic fraudulent ground truth. Such CDRs are associated with explicit knowledge of a *fraud model*, thus filling the gap for tackling challenge (c2). Through *FraudZen*, we show fraudsters can mimic legitimate communication behaviors from literature almost perfectly, raising the need to advance current detection.**

I. Problem statement

SIMBox fraud consists of diverting the international voice traffic from the regulated routes through VoIP established links [1]. The diverted traffic is received at the level of a *SIMBox* (VoIP to GSM gateway) in the destination country and re-originated as a national mobile call to its recipient. Hence, destination mobile operators perceive national termination fees instead of international ones, which are much higher. The impact is enormous, affecting states' and operators' revenues with a loss estimated to USD 3.11 Billion in 2021 [2]. More critically, *SIMBox* fraud allows attackers to act as national subscribers, which international terrorists could use for covert operations. *SIMBox* appliances also allow eavesdropping on international phone conversations [3], impeding user privacy and giving the possibility of international espionage, *striking impact attesting SIMBox fraud deserves much more attention.*

Yet, the *SIMBox* fraud mitigation remains little tackled by researchers: only 15 literature approaches since 2011. This paper tackles the two main challenges inherent to fraud mitigation.

Scarcity of fraudulent ground-truth. *SIMBox* fraud investigation relies on network-related datasets (i.e., CDRs) to distinguish between legitimate users and *SIMBox* fraudulent communication behavior. Unfortunately, CDRs suffer a deficiency of ground-truth on known fraudulent traffic, resulting from limited operators' detection capability: operators are generally aware of no or a low percentage of fraudulent users compared to the total amount of users in CDRs. The remaining large portion of users is considered legitimate. Hence, *detection built from such partial fraudulent knowledge likely cause many false negatives*.

Constant fraud evolution. Aiming to be indistinguishable from legitimate users, *SIMBox* fraudsters constantly create and refine their appliance functionalities to mimic human traffic and mobility behavior. Unfortunately, such fraud evolution is not followed by detection: *Current CDR-based detection methods are mostly validated in very particular contexts with no guarantee of generalized efficiency for evolved fraud behaviors*.

This paper introduces *FraudZen* as the first-of-the-literature framework to ease research on the fraud detection. As explained next, *FraudZen* rationale is to break through a *lagging-behind detection*, by providing *updated information* on the fraud behavior that can only *come from the fraudsters*.

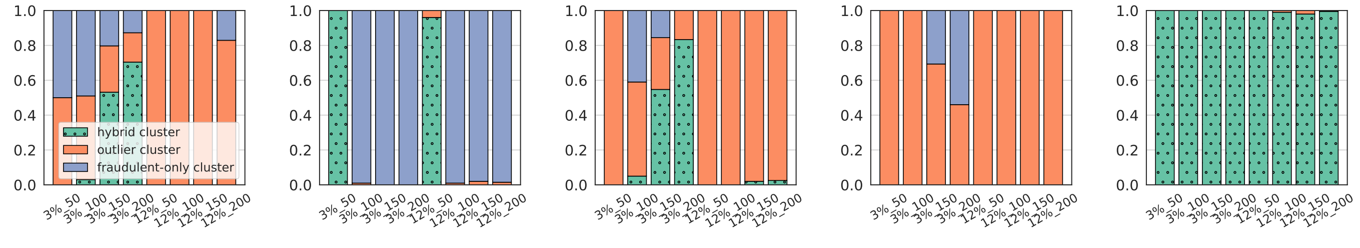
II. How we tackle the problem

FraudZen unleashes the current barriers to fraud investigation in *(c1) providing CDRs with realistic fraudulent ground truth while (c2) associating them with an explicit and measured knowledge of fraud behavior, thus filling the gap for detection leveraging*. Accordingly, researchers and mobile operators can use this tool to (i) inject fraudulent traffic (generated from existing fraud methods or prospected ones) in their CDRs or (ii) investigate the validity of current and future detection methods, in a exhaustive and controlled way. This result is achieved through the following thorough methodology.

a) ***SIMBox market study****:* First, we identify that *SIMBox*'s fraudulent traffic is highly linked to the *SIMBox* appliances generating the fraud. Thus, the *SIMBox fraud's imprint in CDRs results from handling functionalities offered by SIMBox manufacturers with known intent.* With that in mind, we perform a comprehensive *SIMBox* market study to assess the current fraud capabilities [1]. We review the functionalities of all 94 *SIMBox* appliances from the major *SIMBox* manufacturers in the international market. Our study encompasses appliances used by over 2000 fraudsters in more than 31 countries for ten years [4]. It uncovers and categorizes existing *SIMBox* functionalities from their purpose in detection evasion.

b) ***SIMBox fraud modeling****:* Second, we propose the first-of-the-literature *SIMBox* fraud modeling. From the above survey of in-market *SIMBox* functionalities alternatives, we extract information to *"think like a fraudster"*. Our modeling thus provides a framework to define meaningful *SIMBox* frauds. Precisely, through the analysis of 1212 guiding articles and 66 video tutorials [5] from *SIMBox* fraud assisting services (i.e., GoAntiFraud and Antrax), we extract intuitive and easy-to-interpret traits covering the fraud action areas, i.e., traffic, mobility, and social communication behaviors.

c) ***FraudZen framework****:* We then embed our *SIMBox* fraud modeling in the design of *FraudZen*: *an environment for the scalable simulation of SIMBox frauds*. *FraudZen* implements a realistic cellular network architecture involving multiple operators' topology, legitimate users, and *SIMBox* architectures



(a) *fd_naive* (b) *fd_traffic* (c) *fd_mobility* (d) *fd_social* (e) *fd_all*

Fig. 1: In-crowd-blending metric per fraud model: distribution of fraudulent users in OC/HC/FC clusters.

of varying sizes and configurable locations. Such simulation scenarios are calibrated by the configuration of 122 real-world parameters. Hence, from input legitimate (i.e., traffic and mobility models) and fraudulent (*SIMBox* fraud model) users' parameters, *FraudZen* generates CDRs traces of both legitimate and fraudulent users.

d) ***In-crowd-blending capability****:* At last, we define a metric capturing *a fraud model's efficiency* from *FraudZen* generated traces. The *in-crowd-blending capability* of a *SIMBox* fraud model refers to its ability to make fraudulent users blend into the crowd of legitimate ones. It comes from the intuitive idea that the more a fraud model yields users' behaviors close to human ones, the harder it is to detect such fraudulent users.

To infer such capability, we consider the *FraudZen* CDRs generated from a given fraud model fm. From these traces, we get for each user a vector of features reflecting its communication behavior *from CDR-based literature works* (e.g., number of calls per day; number of unique cell Ids).

We then apply a multi-variate unsupervised clustering (e.g., DBSCAN) to the gotten users' feature vectors to group users with similar cellular communication behavior. The users populating the same behavioral group define a particular cluster.

Hence, we distinguish three categories of fraudulent users: (i) isolated users (named *outlier cluster, i.e., OC*), (ii) users in the same clusters as legitimate users (named *hybrid cluster*, i.e., *HC*), and (iii) users in a cluster of only fraudulent users (named *fraudulent-only cluster*, i.e., FC), described hereafter. The distribution of users into the three aforementioned categories reveals how efficient each *SIMBox* fraud model fm is in blending into the legitimate crowd. We compute such *in-crowd-blending capability* as: $ICB(fm) = \frac{|HC|}{|HC|+|FC|+|OC|}$.

III. Validation and Key observations

This section validates *FraudZen* ability for efficient *SIMBox* fraud generation, under realistic cellular network scenarios, by characterizing the ICB of five generated fraud models (fm).

We perform experimental setup with 21K legitimate users from a fully anonymized real-world traffic CDRs. This dataset is enriched with our realistically-emulated trajectories using the *Enhanced-Working Day Mobility Model* (En-WDM) [6].

Regarding fraudulent users' behavior, we design an advanced *SIMBox* fraud model (i.e., *fd_all*) following insights from a GSM termination tutorial [7]. [7] indicates the *fd_all*'s *SIMBox* configurations in terms of traffic, mobility, and social (i.e., calling interactions) behavior to provide as input to the *FraudZen* framework. Then we distinguish four fraud models obtained by reproducing such advanced configurations for no (i.e., *fd_naive*) or a unique behavioral feature (i.e., *fd_traffic*, *fd_mobility*, or *fd_social*) allowing to assess their significance at the fraud efficiency. Multiple simulation scenarios are obtained by combining values of (i) percentages of incoming international traffic (i.e., 3% and 12%) and (ii) numbers of SIM cards in the fraudulent architecture (i.e., 50, 100, 150 and 200).

Fig. 1 reports the *in-crowd-blending capability* per fraud model and under the different considered scenarios: The *fd_all* fraud model (Fig. 1e) yields all fraudulent users in hybrid clusters regardless of the scenario. *This attests FraudZen capability to leverage the current in-market SIMBox functionalities at the generation of frauds very close to human behavior, thus highly efficient.* This provides evidence of the need to enhance current CDR-based *SIMBox* fraud detection of the literature.

Further analysis indicates (i) *fd_traffic* fraud model is counter-intuitively worse than *fd_naive* due to its naive mobility behavior shared by all fraudulent users. (ii) *fd_mobility* fraud model follows the same trend as *fd_naive* while being more efficient. This suggests improving mobility rather than traffic has a better impact on the effectiveness of fraud strategies. (iii) At last, *fd_social* fraud model's results show the social component does not greatly impact the fraud effectiveness.

IV. What we plan next

FraudZen will be released to ease and promote research on SIMBox fraud mitigation. We believe such a tool is indispensable for research in this field where data is intrinsically private. In future works, we plan to implement literature *SIMBox* fraud detection approaches and assess their performance and limitations against the above fraud models through a comprehensive evaluation given by multiple parameters. We will then have hints to build a more resilient detection approach.

References

[1] A. J. Kouam, A. C. Viana, and A. Tchana, "Simbox bypass frauds in cellular networks: Strategies, evolution, detection, and future directions", *IEEE Communications Surveys & Tutorials*, 2021.

[2] C. F. C. Association, "Fraud loss survey", tech. rep., 2021.

[3] GoAntiFraud, "Call recording". https://goantifraud.com/en/ejointech-skyline-gsm-termination-solution#call-recording, March 3 accessed 2023.

[4] GoAntiFraud, "Top 5 popular gsm gateway manufacturers".

[5] GoAntiFraud, "Goantifraud blog". https://goantifraud.com/en/blog, n.d.

[6] A. J. Kouam, A. Carneiro Viana, A. Garivier, and A. Tchana, "Génération de traces cellulaires réalistes", in *CORES*, 2022.

[7] A. FlamesGroup, "Answers of voip communications expert olga saiko". https://www.youtube.com/watch?v=YodIXMqw6Ek, August 2011.

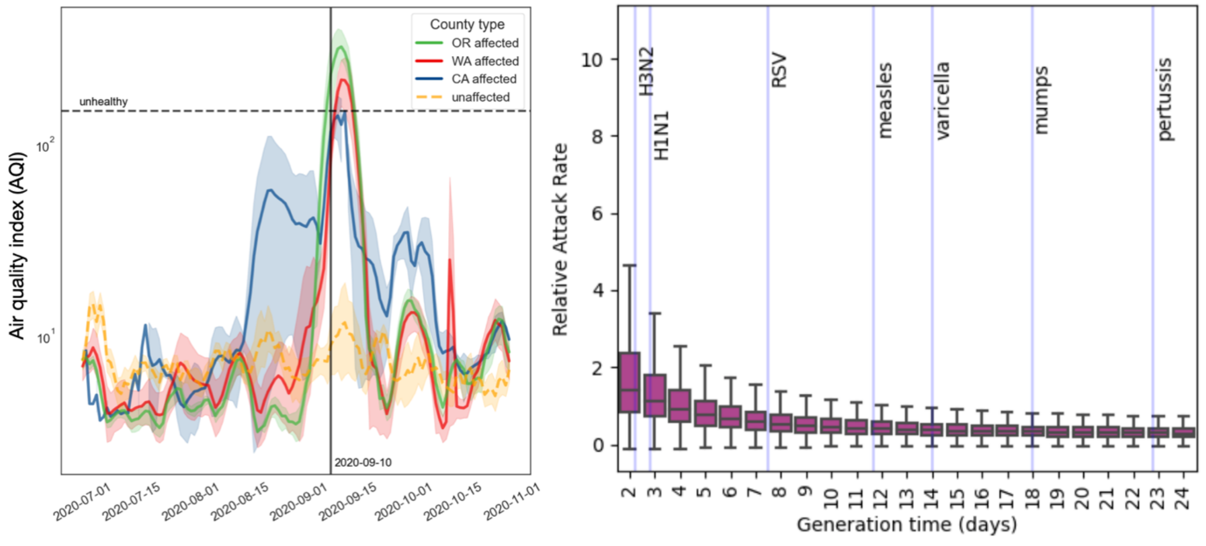
Disruption of outdoor activities caused by wildfire smoke shapes circulation of respiratory pathogens

Claudio Ascione[1,*], Beatriz Arregui García[2], Arianna Pera[3], Davide Stocco[4], Boxuan Wang[1,5], Eugenio Valdano[1], Giulia Pullano[6,†]

1 Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F75012, Paris, France
2 Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain.
3 Department of Computer Science, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark
4 Department of Mathematics, Politecnico di Milano, Via Bonardi 9, I-20133 Milan, Italy (IT)
5 EHESP French School of Public Health, F-35000 Rennes, France
6 Department of Biology, Regents Hall, Georgetown University, 37th and O Streets NW, Washington, DC, 20057-1229, USA
* Presenting author, † Corresponding author

Climate change is causing larger, longer and more devastating wildfires everywhere, and this is especially visible in the United States of America. Adopting this perspective, the New York Times defines wildfires ravaging the West Coast as climate fires [1]. The 2020 wildfire season in the Western United States occurred during the warmest period since global climate records began in 1880, and it was one of the seasons with the highest number of wildfires in the last two decades. Specifically, in August 2020, thunderstorms triggered multiple wildfires across the states of Oregon, Washington, and California which were then followed by additional fire outbreaks along the West Coast in early September 2020.

Wildfires cause harm beyond the direct destruction they cause: atmospheric circulation carries the smoke to areas unaffected by the wildfires, compromising air quality over large swathes of land. Smoke exacerbates pre-existing respiratory chronic diseases, increasing their morbidity and mortality [2]. For this reason, the Centers for Disease Control (CDC) provides guidelines to limit outdoor physical activity when and where wildfire smoke makes the air unsafe. But wildfires may threaten human health in other ways, by changing behavioral patterns. Directly perceived air quality deterioration and guidelines induce changes in the ability and willingness of individuals to engage in outdoor activities, altering recurrent human behavior habits, and these behavioral changes may affect the spread of communicable disease [3]. This is especially true in the case of airborne respiratory pathogens: influenza, SARS-CoV-2, and Respiratory Syncytial Virus (RSV). Their



spread is greatly driven by the social behavior of individuals, and indoor activities facilitate transmission and play a significant role in shaping the variability of disease dynamics [4]. While seasonal human behavior changes and their impact on hampering or mitigating epidemic activity have been largely studied [5], the impact of Wildfire smokes on human behaviors and the cascade effect for respiratory disease transmission is still unclear.

Here we study the effects of smoke generated by severe wildfires in the U.S. states of California, Oregon, Washington in September 2020. We assess the impact on human behavior and the cascade effect on epidemic spread using data on indoor and outdoor human habits extracted from SafeGraph mobile phone data in [4], and the potential consequences for the emergence of respiratory diseases. Our findings reveal a significant shift towards indoor activities in counties within Oregon and Washington during wildfires. However, a discernible change in mobility patterns is not evident in California. This discrepancy may arise from the familiarity of Californian residents with wildfires and air quality index alerts, which have become integrated into their daily routines. Consequently, their mobility patterns may be less affected during such incidents compared to individuals in other regions. We then use a deterministic compartmental model of epidemic spread to quantify the impact of the describe behavioral changes on epidemic circulation. Our findings show that counties with disrupted air exhibit higher cumulated and peak incidence of cases compared to unaffected counties, with the exception of California. Additionally, flu-like epidemics, featuring low reproduction ratio and short generation time, are most affected by the behavioral changes under study. Our findings may help improve public health response in a context of larger, more frequent wildfires triggered by climate change.

References:

[1] Stuart A. Thompson and Yaryna Serkez. “Opinion — Every Place Has Its Own Climate Risk. What Is It Where You Live?” en-US. In: The New York Times (Sept. 2020). issn: 0362-4331

[2] Marina Romanello et al. “The 2022 report of the Lancet Countdown on health and climate change: health at the mercy of fossil fuels”. English. In: The Lancet 400.10363 (Nov. 2022). Publisher: Elsevier, pp. 1619–1654. issn: 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(22)01540-9.

[3] Giulia Pullano et al. “Underdetection of cases of COVID-19 in France threatens epidemic control”.en. In: Nature 590.7844 (Feb. 2021). Number: 7844 Publisher: Nature Publishing Group, pp. 134–139. issn: 1476-4687. doi: 10.1038/s41586-020-03095-6.

[4] Zachary Susswein, Eva C Rest, and Shweta Bansal. “Disentangling the rhythms of human activity in the built environment for airborne transmission risk: An analysis of large-scale mobility data”. In: eLife 12 (Apr. 2023). Ed. by Niel Hens, Diane M Harper, and Guillaume Béraud. Publisher: eLife Sciences Publications, Ltd, e80466. issn: 2050-084X. doi: 10.7554/eLife.80466

[5] Amy Maxmen. “Can tracking people through phone-call data improve lives?” en. In: Nature 569.7758 (May 2019). Bandiera abtest: a Cg type: News Feature Number: 7758 Publisher: Nature Publishing Group Subject term: Epidemiology, Malaria, Research data, Ebola virus, pp. 614–617. doi: 10.1038/d41586-019-01679-5

Re-organisation of socioeconomic networks in Sierra Leone due to external shocks

Ludovico Napoli[1], Vedran Sekara[2], Manuel García-Herranz[3] and Márton Karsai[1,4*]

[1] Department of Network and Data Science, Central European University, Vienna, Austria.
[2] IT University of Copenhagen, Copenhagen, Denmark.
[3] UNICEF, Stockholm, Sweden.
[4] Rényi Institute of Mathematics, Budapest, Hungary.
*Corresponding author: karsaim@ceu.edu

Individual socioeconomic status is a crucial driver of macroscopic phenomena in social networks. Status homophily is one of the key elements that influence network evolution and potentially lead to observable social segregation patterns [1]. Moreover, homophilic mechanisms are adaptive to the external environment, leading to changes in the social structure. However, the observation of these phenomena in large-scale systems is still problematic due the poor availability of proper digital data, especially in developing countries.

Here we combine large-scale mobile phone communication data from a major telecommunication provider in Sierra Leone with a fine-grained socioeconomic map to build a large socioeconomic network. Our goal is to understand how network segregation patterns change due to external shocks in two simultaneously observed behavioural aspects of the same population: their spatial mobility and their social communication networks. We follow these changes on the short run and at different scales in response to the first COVID-19 restriction policies in the country.

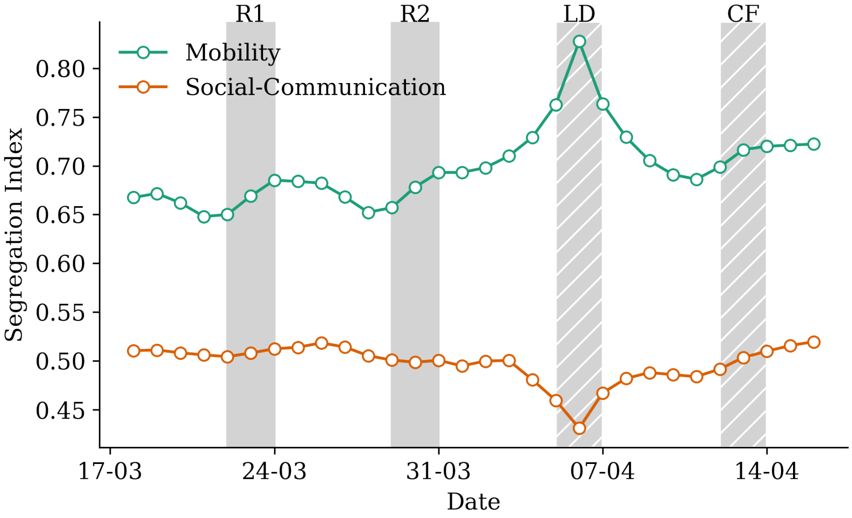


Figure 1. Global segregation dynamics in Sierra Leone. The segregation index is measured as the socioeconomic network assortativity coefficient in the mobility (green curve) mobile communication (orange curve) networks day by day, with a three-day rolling time window, during a four-week period of observation. Shaded periods indicate two reference periods (R1 and R2), the three days lock-down period (LD) and the corresponding three days during the subsequent curfew period (CF).

In general, as seen in Figure 1, during the reference periods (R1 and R2) significant level of segregation is present in both networks, as compared to the segregation reproduced by simple reference models (not shown) accounting for different confounding factors like physical distance, income distribution, or network degree distribution. Interestingly, by following the segregation index of the socioeconomic assortativity, we find opposite dynamics in the two networks. While in the mobility network we observe an expected increased segregation (as observed by others too [2, 3]). induced by the mobility restrictions installed during this period, disallowing people from different socioeconomic classes to mix as usual. At the same time, strikingly, segregation is decreasing globally in the social communication networks suggesting that people from different socioeconomic classes interact more among classes, this way decreasing assortativity during this period. We found that this way of re-organisation of the social-communication network is induced by a relative increase of communication between the richest area around the capital city Freetown and the rest of the country.

Looking at the individual point of view (i.e., measuring the individual level of assortativity in the network [4]), we can draw a more sophisticated picture. We observe that the network positions in the new network configurations are strongly dependent on socioeconomic status, leading to a new equilibrium between classes. Indeed, if at each time we standardise individual values over the current global level of segregation, we find that during lockdown rich people occupy more segregated positions in both network than before, while poor people become the most segregated in the mobility network, while they occupy more integrated position in the social-communication network.

Our study highlights that while emergency policies can have an overall strong effect on the socioeconomic structure of social networks, they do not impact the same way people from different socioeconomic classes. Studying inequalities in intervention effects can lead to rich phenomenology that may motivate future research both on the observational and on the modelling side of re-organisation of socioeconomic networks induced by external shocks.

References

[1] Y. Leo et al., Journal of The Royal Society Interface, 13 125 (2016)

[2] A. Glodeanu, et al., Health & Place 70 102580 (2021)

[3] L. Peel et al., PNAS, 115 16 (2018)

[4] G. Bonaccorsi, G., et al., PNAS 117 27 (2020)

Climate variability and temporary migration

Evidence from three years of mobile phone data in Senegal

Paul Blanchard[1], Virginie Comblon[2], Flore Gubert[3], Erwan Le Quentrec[2], Anne-Sophie Robilliard[3], Stefania Rubrichi[2]
[1] Trinity College Dublin, [2] Orange Innovation - SENSE (Sociology and Economics of Networks and SErvices),
[3] IRD, UMR LEDa-DIAL, PSL, Université Paris-Dauphine, CNRS, Paris, France.

Households in Sahelian countries mostly rely on subsistence agriculture and livelihood means are generally dependent on the quality of a single rainy season (June to October). Income streams are thus often marked with a high degree of seasonality – especially in the rural sector – and are exposed to shocks with the occasional occurrence of poor rainy season conditions. Such variations in the level of productivity across space and over time could induce some degree of short-term labor mobility whereby individuals temporarily relocate to places where they are more productive. For instance, annual circular migration movements have been observed historically, indicating that individuals move across space in response to seasonal fluctuations in the availability of natural capital (e.g. water and fodder) and the demand for labor, especially in the rural sector. However, little is known about the scale and precise nature of temporary migration decisions as a viable strategy in the face of climate stresses.

In Senegal, about 87% of agricultural households practice rainfed agriculture, and more than half of the population resides in rural areas so that economic activities are highly dependent on precipitations during the rainy season, between June and October. In the medium term, we observe important inter-annual variations in the amount and spatio-temporal distribution of rainfall, with the occasional occurrence of very good years (i.e. above-average rainfall) and droughts (i.e. significant rainfall deficits). Over the past two decades, the country has experienced four major drought episodes, three of which after 2010 (2011, 2014, 2018). Taken together, these three events are reported to have affected over 1.8 million people[1].

Research on the relationship between environmental and climate change and human mobility has increased rapidly during the past decades. Recent studies underline how the climate-migration relation is complex and contingent upon several factors at macro-, meso- and micro-scales. There is a broad consensus - as highlighted in the 6th IPCC Assessment Report Working Group 2 report - that environment and climate conditions are important drivers of different forms of mobility, but that specific (im)mobility outcomes are context-specific and strongly influenced by economic, social, political and demographic processes (IPCC 2022). Targeted studies are therefore required to improve our understanding of migration responses and eventually to better assess their consequences.

Although the literature has mostly focused on permanent migration movements and urbanization processes, a relatively marginal body of research has nonetheless highlighted the importance of short-term migrations. But little is known about the scale and precise nature of temporary migration decisions as a viable strategy in the face of climate stresses. This can be partly attributed to the difficulty of measuring subtler human movements with traditional survey methods and the resulting lack of detailed data on short-term movements.

In this study, we exploit a multi-year mobile phone dataset in Senegal to gain new insights into the role of rainy season conditions in shaping temporary migration decisions. We develop algorithmic methods that allow us to extract the temporary migration history of millions of users, that we argue represent a very large section of the population. We construct a time-disaggregated migration matrix that describes temporary migration patterns over three years and across more than 900 locations covering the entire country. We combine it with satellite-based measures of the quality of rainy seasons and we estimate gravity-type models to identify the effect of rainfall conditions at both origin and destination on temporary migration decisions.

As a preliminary step, we carefully examine the characteristics of our sample in order to address usual concerns about the representativeness of such non-traditional data sources. We rely on secondary survey data such as the Demographic and Health Surveys (DHS) to assess whether and to what extent users in our sample differ from the at-large population. We argue that, although non-random, our set of users likely represents a very large fraction of both rural and urban sub-populations in Senegal. We define a “high-quality ” subset of users that are frequently observed for a relatively long period of time in order to allow for the detection of temporary migration events, while ensuring this filtering procedure does not exacerbate selection. The resulting dataset consists of about 2 million users moving over a network of 916 locations.

With this dataset in hand, we develop a mixed method that employs both frequency-based approaches and segment-based migration detection procedures inspired from Chi et al. (2020), to extract the temporary migration patterns. Our clustering method identifies periods of continuous presence of a user at a particular location, allowing for idiosyncratic deviations, i.e., short-term trips. We apply this procedure to our dataset and obtain the migration history of each user that we can flexibly aggregate over time and across locations. Each segment is associated with a destination, a start date, an end date and a duration.

Our findings reveal that temporary migration is ubiquitous in Senegal, with an estimated 11.3 million temporary migration events of at least 20 days over the period 2013-2015. The granularity of our data allows us to clearly show the degree of connections across locations induced by short-term movements, as illustrated in Figure 1. We find a high degree of interconnectedness between rural and urban locations and, perhaps surprisingly, relatively large rural-to-rural and urban-to-urban movements. Unsurprisingly, urban locations appear to be net

[1] EM-DAT, CRED / UCLouvain, Brussels, Belgium – www.emdat.be

receivers of temporary migrants while rural locations are net senders. In particular, Dakar seems to be attracting a relatively large fractions of the temporary migration flows from all categories.

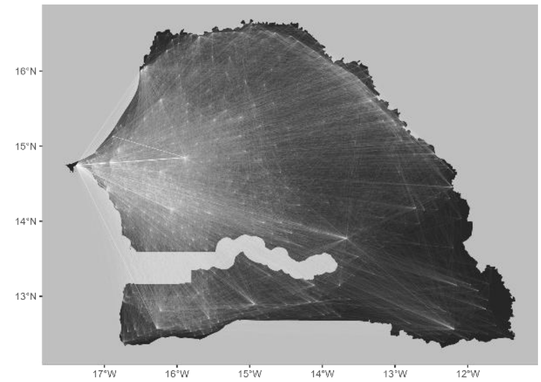


Figure 1: Migration flows across locations in Senegal, 2013.

Over the three years investigated, we consistently observe marked seasonal patterns with migration peaks towards the end of rainy seasons (August-September), as illustrated by Figure 2. The magnitude of the observed systematic increase from June to September is striking: we find a more than two-fold difference in 2013, i.e. an additional 500,000 migrants. we cannot unequivocally identify the motives for such movements, but the results are suggestive of the existence of migration-based income diversification strategies during the agricultural season.

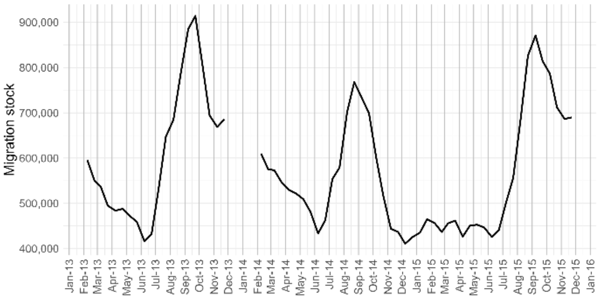


Figure 2 Evolution of the stock of temporary migrants over time.

Armed with a finer understanding of the temporary migration profile of Senegal, we then get to the core of this study and investigate to what extent rainy season conditions affect temporary migration decisions. To do so, we combine a granular pseudo-panel of temporary migration estimates with satellite-based measures of the quality of rainy seasons for the study period. We quantify the quality of rainy seasons with precipitation anomalies based on the Standardized Precipitation Index (SPI), which measures the normalized distance of local precipitation estimates to long-term means (McKee et al., 1993).

We estimate a gravity model with a Poisson Pseudo-Maximum Likelihood, to quantify the effect of rainy season conditions at origin and destination on the propensity to migrate, $P_{o,d,h,\ y}$.

$$P_{o,d,s,y} = \exp\Big(\beta_0 + \sum_{k=1}^{24} \Big(\big(\beta_{1,k} \log(x_{o,s,y})\, \alpha_s^k + \beta_{2,k} \log(x_{d,s,y})\, \alpha_s^k\big) + \gamma_{o,d,h} + \mu_{o,d,s,y}\Big)$$

$P_{o,d,s,y}$ is the fraction of residents in *o* that are in migration at destination d for half-month s of year *y*. $\gamma_{o,d,h}$ is an origin/destination/half-month fixed effect that controls for origin/destination/half-month time-invariant characteristics. Compared to usual approaches that use survey data with poor information on migration destinations, this allows us to estimate a gravity model that essentially controls for the supply side of migration by accounting for the effect of rainy season conditions at destination on its level of attractiveness.

Results are showed in Figure 3 where light dots represent the effect of conditions at origin on the propensity to out-migrate, and darker dots the effect of conditions at destination on its relative attractiveness to potential temporary migrants.

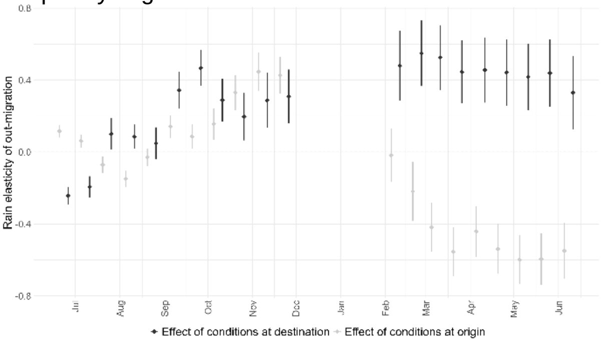


Figure 3 Effect of the SPI at origin and destination on out-migration over the agricultural year.

Interestingly, effects of conditions at origin and destination for the harvest season (October-November) are positive and comparable in size. This important result supports the idea that, at least for the harvest period, both local conditions at origin and destination participate in shaping temporary migration decisions, with poor rainfall conditions increasing liquidity constraints at origin and decreasing the level of attractiveness at destination. On the other hand, between February and June, on average, rainy season conditions at origin and destination create opposite forces in the decision to temporarily migrate: poorer conditions at origin act as a push factor while similar conditions at destination decrease the value of migrating to that destination.

Our result clearly highlights the value phone-derived highly granular migration estimates to identify the effect of interest. We have proposed a first interpretation of the results based on existing evidence in the migration literature, but additional empirical field research is needed to qualitatively confirm the patterns inferred from mobile phone data analysis.

References

IPCC (2022): Climate Change 2022: Impacts, Adaptation, and Vulnerability. Pörtner, H.-O. et al. (Eds.): Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, USA.

Chi, G., F. Lin, G. Chi, and J. Blumenstock (2020): A general approach to detecting migration events in digital trace data, PLoS ONE, 15.

McKee, T. B., J. D. Nolan, and J. Kleist (1993): The relationship of drought frequency and duration with time scales in Eight conference on applied climatology, American Meteorological Society, 179–184.

High resolution urban air quality sensing at scale

Anastasios Noulas
Yasin Acikmese
Firefly
New York, USA
{tassos,yasin}@fireflyon.com

Renaud Lamitotte
University of Oxford
Oxford, UK
renaud.lambiotte@maths.ox.ac.uk

Marta Gonzalez
UC Berkeley
California, USA
martag@berkeley.edu

The proposition of exploiting taxi fleets as a platform to sense urban environments has come into prominence in recent years due to the many advantages it offers both in terms of the scale and the resolution that monitoring can take place, but also due to economic factors that can determine the feasibility and longevity of a sensing project [1]. Taxi units as well as ride-sharing vehicles are omnipresent across time and space within the territory of a city and collecting various forms of signals in their surroundings at high spatio-temporal granularity is possible. Moreover, unlike purpose-built monitoring vehicles that could be deployed by city authorities or private organizations interested in environmental monitoring for instance, taxis do not require explicit commissioning to move about a city, a process that is very expensive to execute presenting a major unit economics challenge. In the meantime, one of the most significant issues that has concerned citizens, urban authorities and other governing bodies over the past decades is the ability to consistently monitor urban air quality. Growing research evidence points that poor environmental and atmospheric conditions in urban environments are one of the lead causes of premature death and a number of age long conditions such as asthma or other respiratory diseases including lung cancer [2], negatively affecting billions of people worldwide that reside at urban and urban-proximate areas. One of the most resonating ideas to perform air quality monitoring at scale has been the initiation of citizen led projects [3] which have been inspired by numerous crowd-sourcing projects that have emerged in the last twenty years and which have enabled the successful collection of data on mapping, social activity and mobility amongst others. Those projects come however with their own challenges such as the need for wide citizen participation rates in addition to deployment related and operational obstacles. Ubiquitous sensing technologies on the other hand can be deployed in existing taxi fleets at scale in an economically viable manner. Over the recent years Firefly [1] has developed and deployed a nation-wide and internationally expanding technology platform the core components of which are a digital display deployed on top of taxis and ride sharing vehicles, and a cloud orchestrated edge-capable software system that communicates information across the display network in real time. The display installation has offered a unique opportunity to deploy a number of sensors that can collect contextual signals of a vehicle's environment as it navigates the city. A key sensor in the display panel is a dust sensor measuring the number of units of suspended particulate matter PM2.5 in the air volume (particle concentration with diameters that are generally 2.5 micrometers and smaller in $\mu g/m^3$). Each sensor measures PM2.5 particles approximately every 60 seconds. As Firefly taxis navigate the city covering over time a large part of the street network, these data is coupled with GPS sourced mobility information effectively offering a highly detailed view of pollution levels in the city. In Figure 1, we demonstrate the spatial resolution of the data by visualizing the maximum PM2.5 concentration recorded at the street segment level. Immediately two key observations can be made. Firstly, the large heterogeneity of pollution levels emerging in the various locations of the city, highlighting how novel analytical insights can be obtained when diving beyond the aggregate view of a large geographic area. Secondly, higher pollution levels are recorded around the center, which hosts the denser parts of the build environment, as well as the main street network arteries which form the traffic backbone of the city. In Figure 2 we plot the probability distributions of PM2.5 readings in histogram form across the eight cities we study in the dataset (value range 0-30). While all cities follow a common pattern with the probability dropping significantly as PM2.5 values rise, there are notable variations across cities with some having their probability mass shifted more towards higher values. Important questions in this setting revolve around the understanding of what urban structural as well as environmental and population activity characteristics raise the probability of higher pollution concentrations in a city. We also compare the data collected by Firefly with air quality monitors by citizen led projects as well as government agencies. In Table 1 we report basic statistical properties of the data collected by Firefly taxis, ranking cities according to mean pollution level. We note how dense cities known

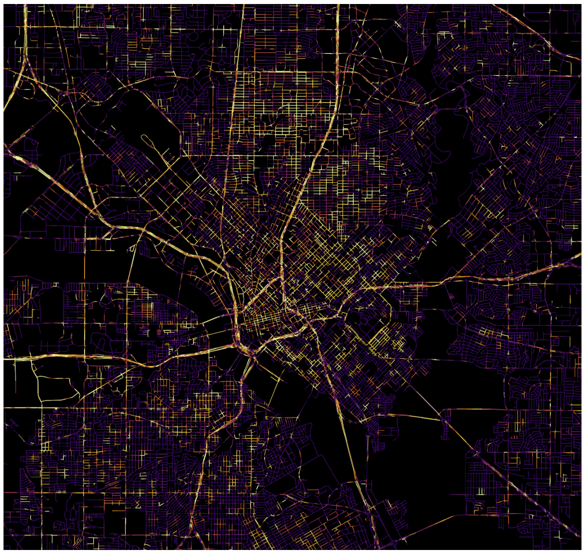


Figure 1: Dallas Street Network Colored according to max PM2.5 levels recorded (brighter color signifies higher levels).

[1] www.fireflyon.com

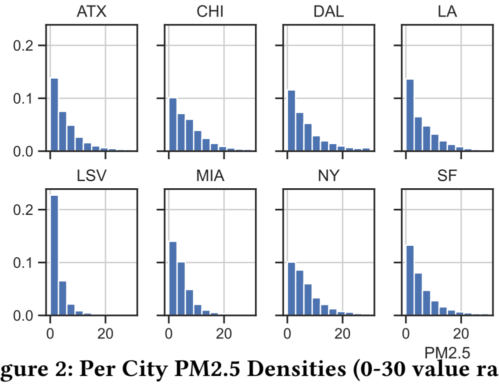


Figure 2: Per City PM2.5 Densities (0-30 value range).

City	Data points	Mean	SD	Rank
Dallas	3,337,678.00	8.11	12.81	1.00
New York	42,272,355.00	7.58	11.82	2.00
Chicago	16,557,911.00	7.45	8.73	3.00
Los Angeles	15,761,095.00	6.38	8.71	4.00
San Francisco	13,418,809.00	6.22	8.07	5.00
Austin	4,246,054.00	6.11	8.69	6.00
Miami	12,067,138.00	4.73	6.20	7.00
Las Vegas	37,189,840.00	2.64	6.83	8.00

Table 1: City level PM2.5 basic statistics, Firefly Taxis

City	PM2.5
Los Angeles-Long Beach-Anaheim, CA	12.1
Las Vegas-Henderson-Paradise, NV	10.5
Chicago-Naperville-Elgin, IL-IN-WI	10
Dallas-Fort Worth-Arlington, TX	9.8
San Francisco-Oakland-Hayward, CA	9.9
Austin-Round Rock, TX	9.5
Miami-Fort Lauderdale-West Palm Beach, FL	9.4
New York-Newark-Jersey City, NY-NJ-PA	8.7

Table 2: Annual mean values reported by the US EPA (https://www.epa.gov/) in similar regions (measured in μg/m3).

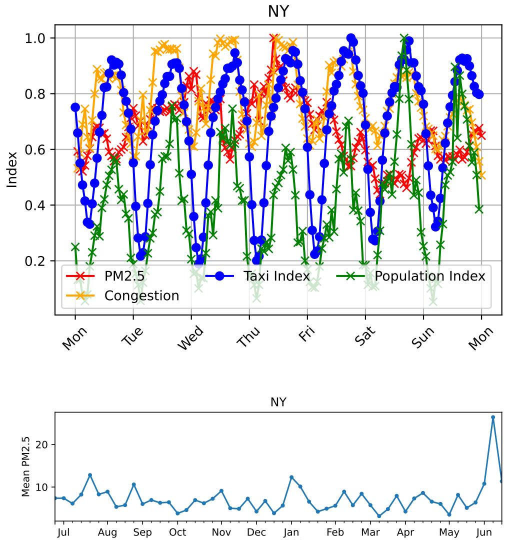


Figure 3: Weekly patterns of activity signals (top) and seasonal mean PM2.5 view (bottom) in the New York area.

for their car-centric infrastructure and intense traffic conditions feature lower air quality standards. Similarly, in Table 2 we provide a ranking according to the annual mean values reported by the U.S. Environmental Protection Agency for similar regions. While there is a general agreement between the two sources of air quality measurement, disagreements are also apparent. We discuss potential sources of this discord including differences in measurement infrastructure as well as the regions across which measurement is carried out. As an example, in New York Firefly taxis are primarily active in the highly populous boroughs of Manhattan, Brooklyn and Queens whereas the US EPA reports measurement across a very wide region around New York City, which includes New Jersey. We discuss key sources of bias in the data, where taxi-based measurement is naturally skewed towards areas that taxis go to. They come however with the advantage of a very high spatial resolution view of the city when stations typically installed by government agencies are more sparsely dispersed across geographies. We argue that different sources of air quality monitoring are not only complementary, but essential, considering the environmental emergency our planet is going through and the epidemiological importance of air quality for urban populations. Another insightful perspective on air quality monitoring is the study of the temporal variability of pollution levels over time. In Figure 3 (top) we report mean pollution levels recorded by Firefly taxis in New York City for each hour of the week, where each data point corresponds to the mean pollution level observed at that hour. We normalize each data point with respect to the max observation during the 168-hour time window of a week and compare this signal with population fluctuations, taxi activity as well as traffic congestion signals. Despite the fact that pollutant particle diffusion and concentration patterns in the atmospheric realm of a city vary also due to weather conditions (e.g. wind patterns, humidity and temperature), here we concentrate our investigation on the link between pollution and human activity dynamics since the latter is the primary source of urban pollutants and a defining factor of their concentration dynamics. We also discuss seasonal variations in air quality patterns. In this context, we present evidence which suggests a connection between spikes in air pollution levels and significant climatic or social events. In Figure 3 (bottom) we present weekly mean PM2.5 levels in the New York region across the course of the year featuring an unusual increase in air pollution levels in early June 2023, a phenomenon known to have been induced by forest fires in Canada.

REFERENCES

[1] K. O'Keeffe, A. Anjomshoaa, S Strogatz, P. Santi, and C. Ratti. Quantifying the sensing power of vehicle fleets. *PNAS*, 2019.

[2] R. Li, R. Zhou, and J. Zhang. Function of PM2.5 in the pathogenesis of lung cancer and chronic airway inflammatory diseases. *Oncology letters*, 2018.

[3] N. Castell, FR Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international*, 2017.

A methodology for studying the exposure of the population to air pollution: application to the analysis of the Madrid region

María Merino Asenjo - Nommon Solutions and Technologies - maria.merino@nommon.es
Raquel Sánchez-Cauce - Nommon Solutions and Technologies - raquel.sanchez@nommon.es
Miguel Picornell Tronch - Nommon Solutions and Technologies - miguel.picornell@nommon.es
Oliva Cantú Ros - Nommon Solutions and Technologies - oliva.garcia-cantu@nommon.es
Ricardo Herranz - Nommon Solutions and Technologies - ricardo.herranz@nommon.es
Rafael Borge - Universidad Politécnica de Madrid - rafael.borge@upm.es
Javier Pérez - Universidad Politécnica de Madrid - javier.perezr@upm.es
David de la Paz - Universidad Politécnica de Madrid - david.delapaz@upm.es

Air pollution is one of the biggest challenges cities are facing today. The expected growth of urban population up to 70% of overall world population turns cities into one of the greatest pollutant emitting sources, compromising citizens' health and calling for specific actions in order to assess and reduce pollution. To efficiently determine the most appropriate policies, it is essential to carry out assessments of the population's exposure to pollutants. Traditionally, one of the main limitations associated with these studies has been the lack of information on the distribution of the population in the city throughout the day (population dynamics). The widespread use of mobile devices makes it possible to overcome these limitations, allowing a highly detailed analysis of the activities and mobility patterns of the population throughout the day.

In this contribution we present a data processing pipeline that allows us to calculate the pollution exposure diary of the people living in a certain region. The pipeline is based on a dynamic exposure approach that merges population presence information obtained from mobile network data with pollutant concentrations from a high-precision advection-diffusion model. The pipeline has been conceived to provide the necessary scalability to deal with large datasets, a mandatory feature when it comes to the study of large metropolitan areas. The proposed approach is demonstrated for the city of Madrid, in the context of a project aimed at assessing air quality improvement policies. The calculation has been performed for a week of March 2018, considering population presence on an hourly basis and hourly concentrations of O3, NO2, and PM2.5 measured in micrograms per m3. This abstract outlines the developed methodology and presents some of the results from its application to the Madrid case study.

The methodology benefits from Nommon's solution for obtaining presence information from anonymised mobile network data (Population Insights), which generates activity diaries for the sampled mobile phone users and expands these diaries to the total population using census for residents and official tourism statistics for non-resident population. Then, activity diaries are merged with a netCDF file containing pollutant concentration data. The joint is not only spatial, as it might traditionally happen in urban spatial analysis, but also time is considered. Exposure is calculated for each user, at every time step and for a given zone, as the product of population presence and pollutants' concentration level at the zone, weighted by the time of exposure within the defined time interval.

The results are aggregated to calculate two types of indicators: the average exposure of the residents of each district, and the average exposure per inhabitant experienced at each visited district. Figure 1 shows the average daily exposure to NO2 of the residents of each municipality of the Madrid region and each district for the second week of March 2018. As can be seen in the figure, the residents of the Madrid city experience the highest levels of exposure, specially the ones of the central area, while the zones further away from the capital have the lowest levels of exposure.

These indicators provide information such as which is the district whose inhabitants are more exposed to pollution during the day or where the population affected by the pollution of a zone lives. This information can inform the design of policies that effectively reduce population exposure to pollution.

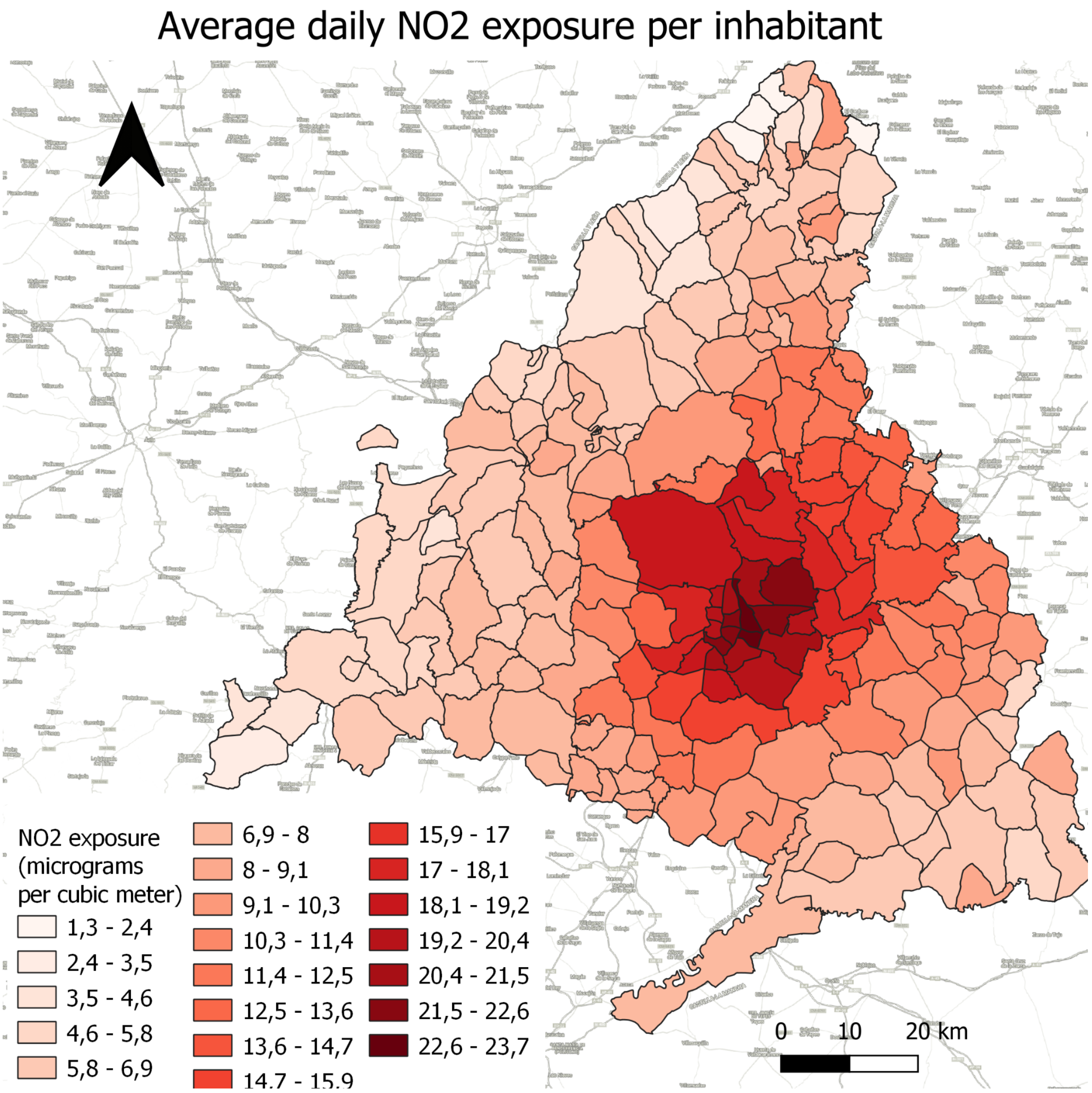


Figure 1. Average daily exposure to NO2 experienced by the residents of each municipality of the Madrid region and each district of the city of Madrid for the week from 5 to 11 March, 2018.

Coordinated Traffic Assignment for Sustainable Urban Transportation

Giuliano Cornacchia[1,2], Mirco Nanni[2], and Luca Pappalardo[2]

[1] Department of Computer Science, University of Pisa, Pisa, Italy (giuliano.cornacchia@phd.unipi.it)
[2] ISTI, CNR, Pisa, Italy (luca.pappalardo@isti.cnr.it)

Traffic Assignment (TA), the process of assigning routes to a collection of trips, has become a critical issue in modern times due to the rapid growth of urbanization and the escalating problem of traffic congestion [4, 7]. Effective TA is pivotal in achieving several United Nations' Sustainable Development Goals (SDGs) by promoting efficient traffic management and reducing greenhouse gas emissions. The need for an efficient TA is even more crucial now given the proliferation of various GPS navigation services (e.g., Google Maps and TomTom), whose aggregated outcome may potentially lead to unintended negative consequences such as the increase of CO2 emissions [1, 6, 5].

Several works have explored the efficient distribution of vehicles across the road network using alternative routing (AR) methods [3]. AR algorithms focus on offering alternative routes from an origin to a destination to individual users, aiming to strike a balance between proximity to the fastest route and route diversity [3, 2]. However, the individualistic nature of AR algorithms tends to overlook vehicle interactions, resulting in sub-optimal outcomes at a collective level. Consequently, they often contribute to heightened congestion and an increased environmental impact.

To address these limitations, we propose Metis, a novel approach that enhances TA by incorporating alternative routing, edge penalization, and informed route scoring. Unlike conventional AR algorithms, Metis is a centralized entity that suggests routes to drivers based on real-time traffic estimates. When Metis recommends a route for a trip from origin o to destination d, it follows these steps. First, Metis employs Forward-Looking Edge Penalization (FLEP), which estimates the current positions of in-transit vehicles on the road network and applies a penalty factor p to the weights of edges these vehicles are projected to traverse. This increase in edge weights reflects the dynamic changes in travel time due to traffic volume, encouraging alternative route choices and a more balanced traffic distribution. Second, utilizing the penalized road networks, Metis applies a state-of-the-art AR algorithm KMD [2] to generate k dissimilar alternative routes between the origin o and destination d while adhering to a user-defined cost threshold ϵ. Finally, Metis evaluates and ranks the set of alternative routes produced by KMD. To identify the best route among the k options, it assigns a score that discourages selecting popular routes and favours routes with higher capacity, capable of accommodating increased traffic flow. By avoiding popular routes, likely to be chosen by other vehicles, Metis can distribute vehicles on the road network more efficiently.

We evaluate Metis against several one-shot TA solutions through experiments conducted in Milan, Florence, and Rome. To assess the effectiveness of Metis and the baselines, we consider three measures: total CO2 emissions, road coverage (percentage of visited edges), and time redundancy (average edge utilization within a temporal window). Figure 1a-c demonstrate METIS's significant contributions, showing impressive reductions of CO2 emissions (18% in Milan, 28% in Florence, and 46% in Rome) compared to the best baseline. Furthermore, Metis achieves the highest road coverage in Florence (79.66%) and Milan (86.68%), and the second-highest in Rome (48.51%) (Figure 1d-f). Moreover, Metis demonstrates low time redundancy, indicating an efficient route allocation within a 5-minute temporal window (Figure 1g-i). Figure 2 visually displays the spatial distribution of sample routes generated by Metis and KMD (the second-best model) in Milan: it is evident that Metis produces more evenly distributed routes, leading to higher road coverage and lower time redundancy than KMD.

In conclusion, Metis offers a promising solution to TA: by introducing a coordinated and collective approach, Metis leverages the power of route diversification and a central unit to optimize traffic distribution.

References

[1] G. Cornacchia, M. Böhm, G. Mauro, M. Nanni, D. Pedreschi, and L. Pappalardo. How routing strategies impact urban emissions. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '22, New York, NY, USA, 2022. Association for Computing Machinery.

[2] C. Häcker, P. Bouros, T. Chondrogiannis, and E. Althaus. Most diverse near-shortest paths. In *ACM SIGSPATIAL GIS*, page 229–239, 2021.

[3] L. Li, M. Cheema, H. Lu, M. Ali, and A. N. Toosi. Comparing alternative route planning techniques: A comparative user study on melbourne, dhaka and copenhagen road networks. *IEEE Trans. Knowl. Data Eng.*, 34(11):5552–5557, 2022.

[4] L. Pappalardo, E. Manley, V. Sekara, and L. Alessandretti. Future directions in human mobility science. *Nature Computational Science*, pages 1–13, 2023.

[5] D. Pedreschi, F. Dignum, V. Morini, V. Pansanella, and G. Cornacchia. *Towards a Social Artificial Intelligence*, pages 415–428. 04 2023.

[6] D. Pedreschi, L. Pappalardo, R. Baeza-Yates, A.-L. Barabasi, F. Dignum, V. Dignum, T. Eliassi-Rad, F. Giannotti, J. Kertesz, A. Knott, Y. Ioannidis, P. Lukowicz, A. Passarella, A. Pentland, J. Shawe-Taylor, and A. Vespignani. Social ai and the challenges of the human-ai ecosystem, 06 2023.

[7] Y. Wang, W. Y. Szeto, K. Han, and T. L. Friesz. Dynamic traffic assignment: A review of the methodological advances for environmentally sustainable road transportation applications. *Transport. Res. B-Meth*, 111:370–394, 2018.

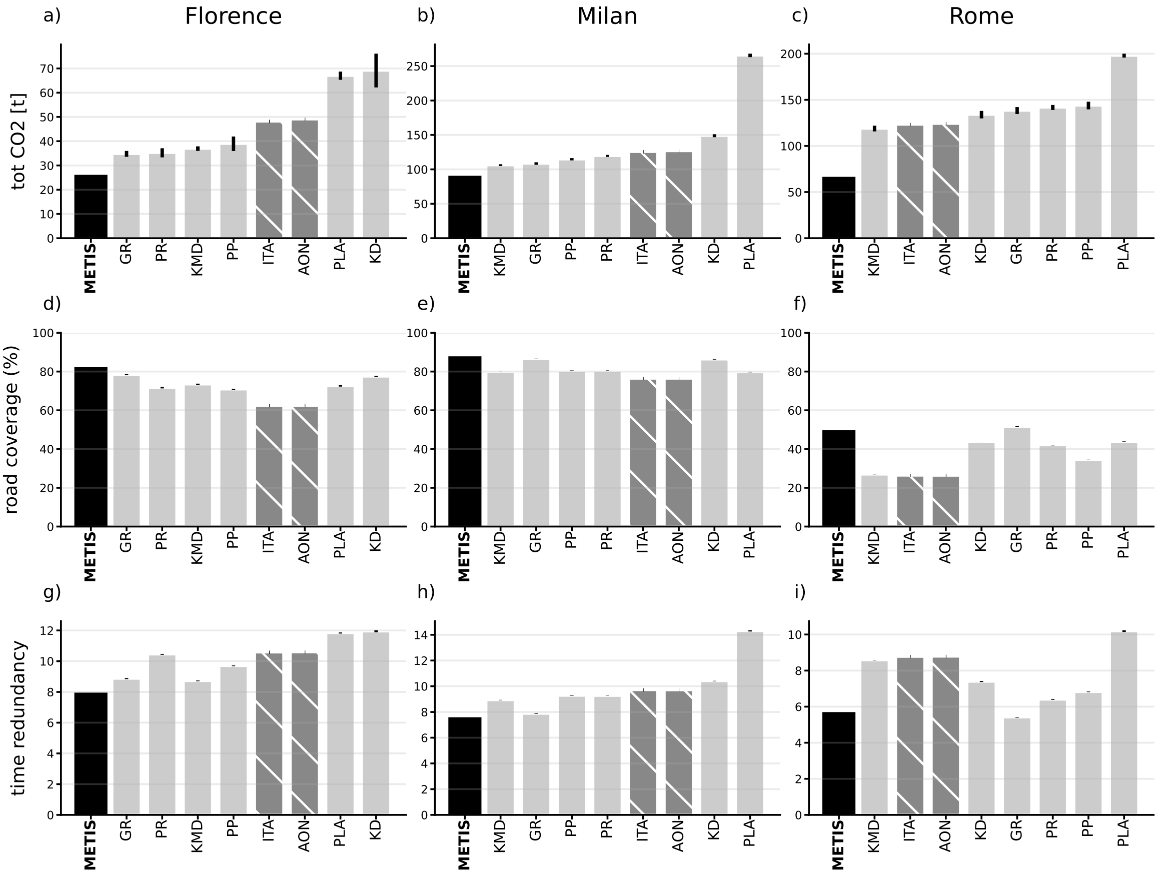
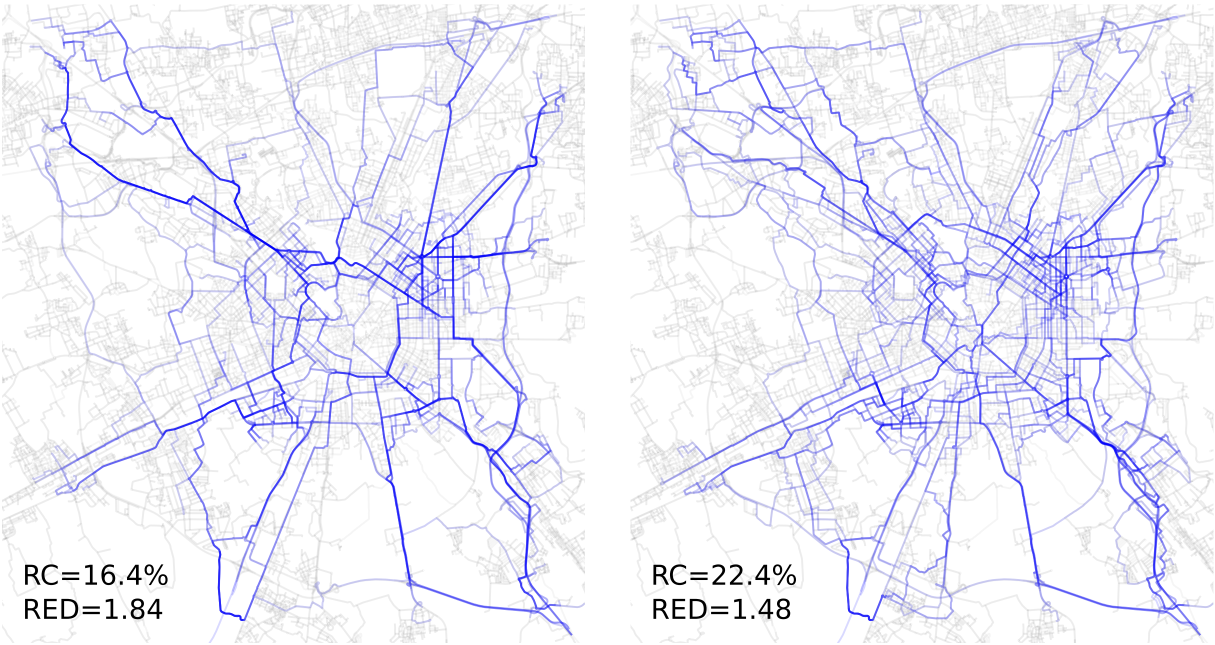


Figure 1: Comparison of METIS (black bar) with the baselines in Florence, Milan, and Rome on CO2 emissions (in tons), road coverage (in %), and time redundancy.



(a) KMD

(b) METIS

Figure 2: Routes generated by KMD (a) and METIS (b) in Milan for 150 trips. RC is the road coverage, and RED is the time redundancy (5-minute window).

Reconstructing temporal social contexts as multilayer networks using contact data

Javier Ureña-Carrion*, Sara Heydari, Talayeh Aledavood, Jari Saramäki, Mikko Kivelä

Aalto University, Finland

*javier.urenacarrion@aalto.fi

Introduction

Social scientists have long recognized the importance of tie multiplexity in social networks: the idea that people interact in social contexts that encompass different types of relationships –combinations of friends, family, work colleagues, and a myriad of acquaintances–. Despite their importance, analyzing real-world social contexts and tie multiplexity can be challenging, as it may require either prohibitively expensive surveys or additional data sources. Simultaneously, massive auto-recorded contact datasets, such as phone call logs, have been crucial for the development of computational social science and our current understanding of social networks. For such large-scale data, tie multiplexity has been out of reach.

We bridge this gap by proposing a method for inferring the social context of a tie from contact data, which has been typically previously discarded in an aggregation process, and leveraging population-level behaviour as a basis for social contexts. Our core assumption is that, on the aggregate, categories of ties are associated with particular timings during the week – e.g., work ties will be largely active during work hours–, so that information on tie multiplexity is encoded in population-wide temporal patterns. Our work follows the framework of Feld's focus theory [1], which proposes that (i) relationships and multiplexity largely arise within social foci – workplaces, gatherings, or any social structure that facilitates interactions–, and that (ii) such focused organization leads to prominent patterns in social networks such as clustering, bridginess, and small-world phenomena. Under this assumption, our target is not to infer particular social foci, but the temporal expression of collections of foci constrained by social timings. Using a large dataset of Call Detail Records from a single country, we reconstruct networks that show non-random structures that confirm many topological hypotheses of focus theory.

Results

We use a large sample of ties to construct a matrix of dyadic activity during each hour of the week and obtain *population-level signals* of via Non-Negative Matrix Factorization (NNMF) with orthogonality constraints [2]. Our case study, in Figure 1a, exhibits signals that can be interpreted as societal temporal rhythms: working times, working day evenings, and weekend day times and evenings. We then propose two different methods based on NNMF and maximum likelihood estimation for inferring each tie's weights for each signal –the tie's *socio-temporal context* –. These weights allow for the construction of a multilayer network, where we add an error layer that captures estimation errors from uncommon call patterns (e.g., late-night contacts). We validate our results by shuffling each tie's contacts according to the aggregated population call signal, finding that this process does not preserve the patterns consistent with focal theory.

Our approach reveals several features of social systems including (a) a heterogeneous distribution of weights across layers, (b) a link between such distribution to tie bridginess, and (c) layer-specific clustering dynamics that might arise from social foci. In more detail, (a) we find that ties with different number of calls have vastly different layer-allocation profiles (Figure 1b). Such profiles reveal a new form of Granovetter's strength of weak ties hypothesis [3]: (b) we find that ties that are highly associated with only a few layers tend to serve as bridges, while ties that allocate their weights evenly across layers are more likely to be embedded in communities (Figure 1c). This relationship holds *even in low-call ties*, providing a richer characterization of ties traditionally regarded as weak [4]. A uniform layer-allocation pattern might suggest that two people interact in multiple contexts. According to focus theory, this focused interaction increases the possibility that they have common friends, even if they such interactions occur sparsely. Last, (c) we analyze transitivity, the idea that if two people are connected to a third, they are also connected among themselves. On Figure 1d, we observe that a triad is more likely to be closed if the two ego-alter pairs allocate their communication in similar proportions across layers, as measured by cosine distance. Note that the latter metric does not account for the magnitude of the layer weights; remarkably, the similarity of weight magnitudes within a layer explains transitivity *only on weekend layers.* Intuitively, this might be understood as people calling close friends and family during the weekends –people who are likely to also know each other–, while calling diverse acquaintances during the working week. This strongly suggests that weekend layers are more informative of related social foci in the sense that the three members of the triad participate in joint activities; whereas weektime layers capture collections of foci that are socially constrained, although not necessarily related.

Discussion

Our results strongly suggest that population-level signals can serve as a basis for inferring social contexts and tie multiplexity from contact data. Multilayer networks reconstructed from such signals display patterns consistent with focus theory and shed new light on known properties of communication systems, such as the topological bridginess of ties and clustering phenomena. The main limitations of our work currently stem from inferring population-level signals using NNMF, a method that requires a prior specification of the number of signals and might get stuck on vastly different local optimums. We hypothesise such limitations might not affect results about layer allocation patterns, but will inevitably capture different layer-specific dynamics. In this sense, we cannot claim that layers are unique, but our results show strong evidence that embedding weekly patterns onto layers reveals social contexts that display structural effects. We believe our framework has ramifications in several areas of social network analysis, such as spreading phenomena and egocentric networks.

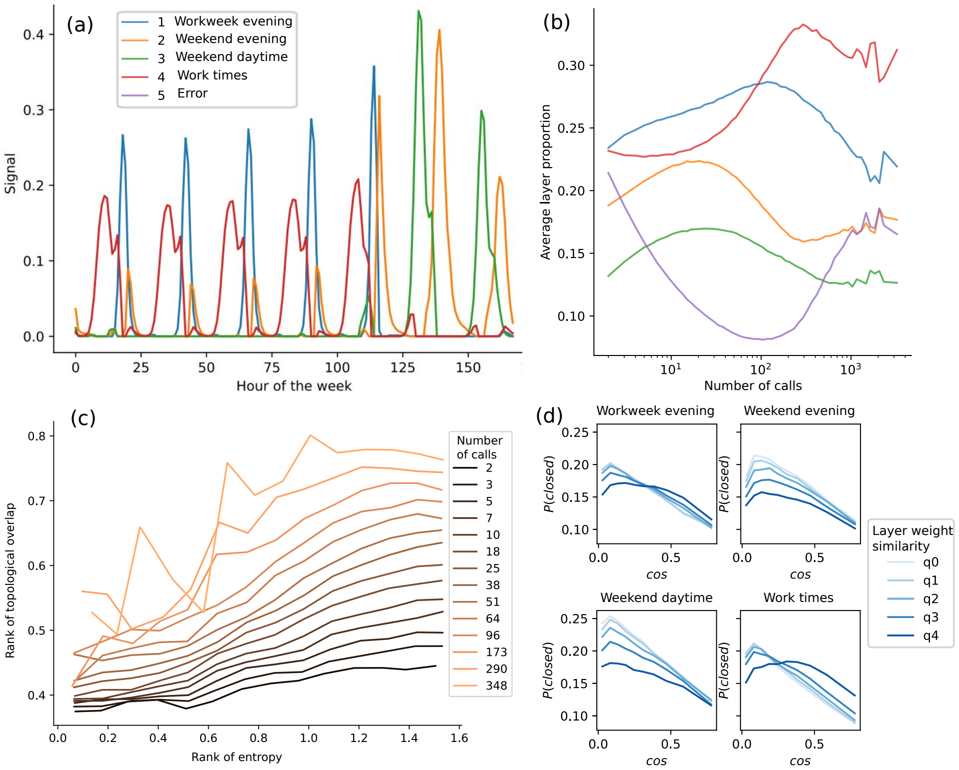


Figure 1: *(a)*. Population-level signals from dyadic interactions obtained via NNMF. We classify weekly behaviour by different signals, which may overlap at different hours. *(b)* Distribution of layer proportions given the number of contacts. *(c)* Effect of weight allocation on bridginess of ties. The x-axis depicts the entropy of layer proportions, where higher values imply a more uniform weight allocation across layers; the y-axis depicts ranked overlap – a proxy for embeddedness defined as the proportion of common neighbors over all neighbors–, and colors represents the number of contacts. Across all numbers of contacts, a more uniform weight allocation (entropy) is associated with lower bridginess (high overlap). *(d)* Probability that a triad is closed based on the activity of two ego-alter pairs for the same ego. The x-axis represents the cosine distance between layer proportions of two ties, the y-axis the probability that the alters are connected in that layer. Colors represent the quantiles of scale-independent distance between the layer weights: for w_{l1} and w_{l2} the weights in layer l, the scale-independent distance $\frac{|w_{l1}-w_{l2}|}{w_{l1}+w_{l2}}$ is not skewed by extremely large values. For weekend layers more similar magnitudes predict more closed triangles, suggesting that similar allocation patterns exist within the same focus; for the weektime layers only the largest differences in magnitude reveal clustering disparities.

[1] Feld, S. L., *The Focused Organization of Social Ties*, In Americn Journal of Sociology (Vol. 86, Issue 5, pp. 1015-1035). (1981)

[2] Kimura, K. Tanaka, Y., & Kudo, M., *A Fast Hierarchical Alternating Least Squares Algorithm for Orthogonal Nonnegative Matrix Factorization.* In Proceeding of the Sixth Asian Conference on Machine Learning (Vol. 39, pp. 129-141). (2015)

[3] Granovetter, M. S. *The Strength of Weak Ties.* In American Journal of Sociology (Vol. 78, pp. 1360–1380). (1973).

[4] Ureña-Carrion J., Saramäki, J., & Kivelä, M.m *Estimating tie strength in social networks using temporal communication data.* In EPJ Data Science (Vol. 9, Issue 1). Springer Scinece and Business Media LLC, (2020).

Quantifying Rhetoric Alignment using Node Embeddings on Temporal Graphs

Keeley Erhardt and Alex Pentland

MIT Media Lab, Cambridge, Massachusetts, USA
{keeley,pentland}@mit.edu

Introduction

In recent years, there has been a growing interest in leveraging node embeddings within graph structures to gain insights into various domains. While text embeddings have been extensively explored for natural language processing tasks, researchers have only recently begun to apply node embeddings in graphs to analyze their underlying structure. This paper utilizes vector representations of a network of Russian and Chinese diplomats' online activity to map the degree of similarity between the diplomats' discourse. This approach extends the application of temporal graph learning, which extracts knowledge from evolving networks, to the domain of political communication and international relations. The analysis provides a deeper understanding of diplomatic interactions within the evolving landscape of digital communication.

Node Embeddings

Node embeddings are low-dimensional representations or vectors that capture the semantic and structural characteristics of nodes within a graph. These embeddings encode information about the relationships, connectivity, and context of nodes. *Node2vec*[1], the most widely used method for node embedding, aims to learn low-dimensional vector representations, denoted as $x_v \in \mathbb{R}^d$, for each node $v \in V$. The algorithm balances the exploration-exploitation trade-off by defining parameters p and q. The parameter p, also known as the return parameter, controls the likelihood of returning to the previous node in a random walk. A higher value of p increases the probability of returning, leading to more localized exploration of the graph. On the other hand, the parameter q, known as the inout parameter, determines the likelihood of visiting nodes that are farther away from the current node in the random walk. A higher value of q favors visiting such nodes, resulting in more global exploration of the graph.

Node2vec maximizes the average log-probability of observing a context node u given a target node v, using the Skip-gram model. The objective function is formulated as maximizing $\sum \log P(u|v)$, where $P(u|v)$ models the probability of observing context node u given target node v. By optimizing this objective function, *node2vec* effectively preserves important graph properties, such as node proximity and structural similarity, in the vector space. In the context of this paper, *node2vec* serves as a powerful tool for extracting meaningful representations of Twitter accounts belonging to Russian and Chinese diplomats, facilitating the analysis of convergence and divergence in their online rhetoric.

Methods

This study investigates the Twitter content shared by 308 Russian diplomats and 222 Chinese diplomats[1] from January 1, 2022, to March 31, 2022. Collectively, these accounts generated 182,693 tweets, which we collected using the Twitter v2 Search API. To examine the temporal evolution of the alignment in online discourse between diplomats from both countries, we construct a graph $G(V,E)$ for each month. This graph captures the dynamic connectivity patterns between the diplomats' online interactions, allowing us to analyze changes in alignment over time. In this graph, V represents the set of Twitter accounts, while E represents the edges connecting pairs of nodes. The edges are determined based on a shared topic criterion, meaning that an edge exists between two nodes if they discuss the same topic. For any two nodes u and v, an edge $(u,v) \in E$ exists if and only if there is a shared topic T that both u and v discuss. To determine these topics, we utilize Twitter's analysis of the tweet's content, returned by the Twitter v2 Search API as context annotations[2].

For each graph G, we learn an embedding over all nodes using `PecanPy`[2], an optimized Python implementation of *node2vec*. We use the `FirstOrderUnweighted` mode with p and q both set to 1. By considering the first-order transition probabilities, the algorithm captures the local neighborhood structure and guides the random walk towards exploring similar or related nodes in the graph. All parameters are set to their default – the dimension of the final embedding is 128, ten random walks are generated from each node, and the length of each random walk is 80.

[1] https://github.com/schliebs/disinfo

[2] https://github.com/twitterdev/twitter-context-annotations

Results

Between January and March 2022, online discussions of Russian and Chinese diplomats reveal a converging pattern. This convergence is supported by a decrease in the Euclidean distance between their respective vector representations. A Euclidean distance of zero would signify complete convergence in the topics addressed by the diplomats. Figure 1 displays the average Euclidean distances, revealing a reduction of over threefold in the average distance between the node embeddings of Russian and Chinese accounts. This intriguing trend suggests evolving alignment in diplomatic rhetoric between the two countries.

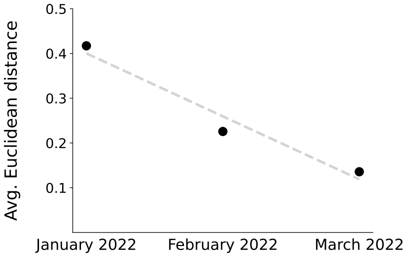


Figure 1. Average Euclidean distance between node embeddings associated with different countries

To provide a visual representation of the node embeddings, we employ t-SNE, a dimensionality reduction algorithm. Figure 2 showcases the visualization of these embeddings in a two-dimensional mapping space. By applying t-SNE, the algorithm learns a mapping from the original high-dimensional vectors to a lower-dimensional space. The resulting mapping ensures that if two high-dimensional vectors, denoted as u and v, are close in proximity, their corresponding mapped points, $map(u)$ and $map(v)$, are also closely positioned within the 2-d mapping space.

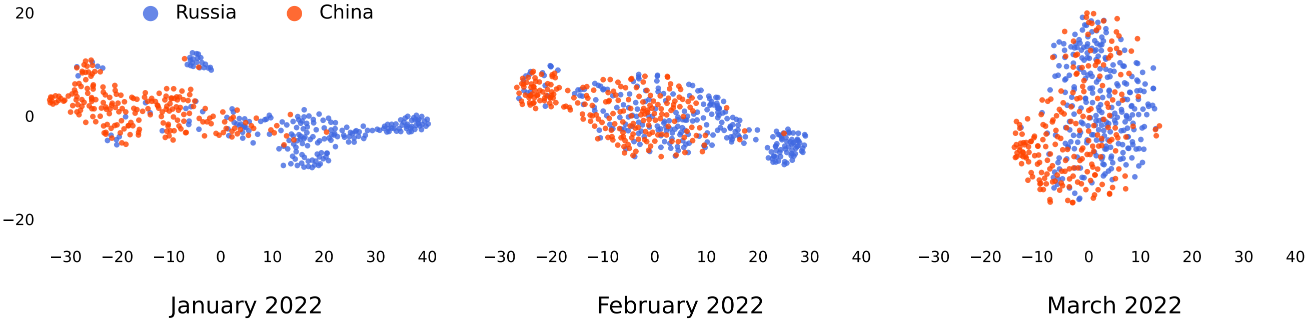


Figure 2. Visualization of node embeddings using t-SNE

In future work, we plan to delve deeper into analysis of the alignment observed between Russian and Chinese diplomats' online discourse. Our focus will be on identifying and examining the specific topics that demonstrate the most notable convergence. This research contributes to a more nuanced understanding of the evolving diplomatic dynamics and signaling strategies employed by both countries, providing valuable insights into their communication patterns and diplomacy efforts.

References

1. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864 (2016).

2. Liu, R. & Krishnan, A. Pecanpy: a fast, efficient and parallelized python implementation of node2vec. *Bioinformatics* **37**, 3377–3379 (2021).

Acknowledgements

Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Evidences of a growing second-level digital in France

Sachit Mishra*†, Zbigniew Smoreda‡ and Marco Fiore*
*IMDEA Networks Institute, Spain, †Universidad Carlos III de Madrid, Spain, ‡SENSE / Orange Innovation, France
{sachit.mishra, marco.fiore}@imdea.org, zbigniew.smoreda@orange.com

***Abstract*—We study the interaction between the consumption of digital services via mobile devices and urbanization levels, using measurement data collected in an operational network serving the whole territory of France. We unveil that such an interaction follows a power law, or, in other words, there exists an emergent behavior that prompts subscribers living in increasingly extended and populated urban areas to exhibit a surging individual consumption of mobile traffic. The result holds for global traffic where we unveil the phenomenon of how the imbalance in per-capita mobile traffic shifts has grown steadily and substantially from 2014–2019 time frame for bigger cities in France. Our study raises interesting arguments on the presence of second-level digital divides in developed countries and on the necessity of better understanding them.**

I. Introduction

The relationship between the population of a specific geographical area and the volume of mobile traffic generated therein is known to be governed by power laws. Hence, it holds that $p \propto t^{\beta}$, where p is the population density (*e.g.*, individuals/km^2) and t is a measure of the mobile traffic (different metrics have been used to date, including the number of voice calls or text messages, bytes of data traffic, or user activity metadata) per spatial unit [1], [2]. Interestingly, previous investigations have proven that $\beta < 1$, implying an emergent behavior according to which larger traffic volumes are in fact generated by a proportionally diminishing population.

Contribution –In this paper, we investigate the presence of a second-level digital divide (*which is a difference in the use of the internet and online services between two groups despite having equal access [3]*) at a national scale in a prominent European country, *i.e.*, France. Specifically, we explore correlations between the number of inhabitants and the volume of consumed mobile data traffic in thousands of individual cities and towns in the country. Our analysis is based on substantial measurement data collected in the production network of a major operator between 2014 and 2019, and leads to the following key observations about the observed second-level digital divide.

(i) The relationship between mobile traffic usage and number of inhabitants of a urban settlement is well described ($R > 0.80$) by a power law with exponent higher than 1. In other words, there exists an *emergent behavior* according to which the larger is the city a mobile subscriber lives in, the higher the mean volume of mobile data he or she consumes.

(ii) The aforementioned phenomenon has *amplified* over the 2014–2019 time frame, as inhabitants of larger cities have increased their consumption of mobile data traffic at a faster rate than inhabitants of smaller towns.

(iii) The growth of inequality in mobile service consumption over time is not explained by the geographical coverage of the network infrastructure, age or income of the local populations, or the commuting of workers to larger cities. The robustness of the behavior to such potential confounding factors reinforces the hypothesis that the living environment (*i.e.*, a large conurbation opposed to a small village, with the sociological implications that the difference involves) is the root cause of the observed divide.

		2014	2016	2017	2019
Commune	R	0.84	0.80	0.84	0.85
Urban unit	R	0.89	0.86	0.88	0.90

TABLE I: Accuracy (R) for different years to substantiate power law hypothesis as R > 0.80 for all cases.

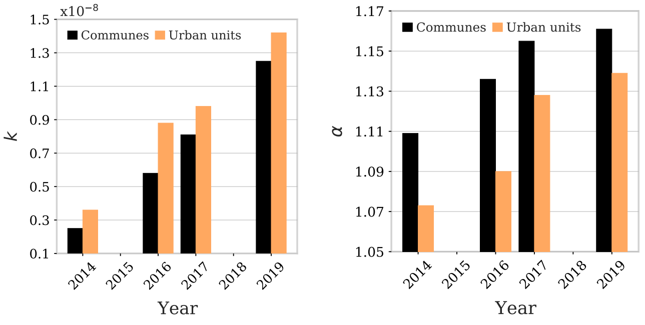
II. Data and methods

We employ anonymized information about the daily mobile data traffic consumption aggregated at the level of individual base stations over the whole of France, and collected in similar time periods over four different years, *i.e.*,, 2014, 2016, 2017, and 2019. We use the geographical locations of the base stations to draw a Voronoi tessellation of the French territory, and uniformly distribute the traffic associated to each base station over its Voronoi cell. A weighted spatial interpolation is performed to map the mobile traffic recorded in Voronoi cells to that generated within each *commune* (an administrative unit analogous to civil township in the United States) and *Urban unit* definition of a city base on built-up area of France. We restrict the analysis to communes and urban units of non-negligible size that have a population count of at least $2,000$ inhabitants, which results in a dataset of daily traffic for $4,837$ communes and $2,296$ urban units. As we are interested in a general model that captures standard correlations, the typical mobile data traffic per day in every city is computed as the average of that recorded in all available days, separately for each year. The per-commune population information is obtained by linearly interpolating in time census data provided by the French National Institute of Statistics and Economic Studies (INSEE) for years 2011 and 2016.

We perform a joint linear fitting and outlier detection on the logarithmic transformations of per-city mobile data traffic and population, via a suitably calibrated Random sample consensus (RANSAC). The rationale is that we observe outlying behaviors for a subset of the smaller towns, due to the unavoidable approximation of the spatial interpolation between Voronoi cells and commune surfaces when their ratio is large (which happens at locations where the user population is sparse). According to RANSAC, 142 communes and 58 urban units are affected by this behavior, and are thus removed from the dataset before drawing results.

III. Results

We confirm that a power law relates mobile data traffic consumption and population size across cities and towns of a developed country like France. Table I shows the R for the linear fitting in log-log scale, corresponding to a power law $t = k \cdot p^{\alpha}$ for linear scales. Here t is the daily mobile network traffic demand measured in a commune (*e.g.*, in bytes, although all results are normalized by the maximum daily load observed in the data, so as not to disclose the actual volume of traffic of the operator), and p is the local population (in inhabitants). The multiplier k and the exponent α are the fitted model parameters. Figure 1(a) shows the value of k which is a per-capita traffic behavior on a per-city basis. The value of a k is increasing which shows the increase in usage of mobile service during these years while by comparing the exponent α in the 2014–



(a) Multiplier

(b) Exponent

Fig. 1: Evolution of the parameters $\boldsymbol{k}$ and $\boldsymbol{\alpha}$ of the power law model of the total traffic and population across years.

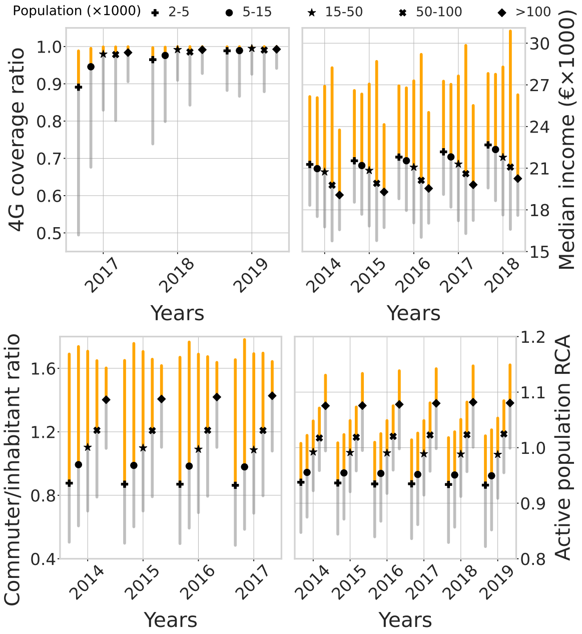


Fig. 2: Mean (marker) and standard deviation (errorbars) of four potential confounding factors, for five population-based classes of communes (denoted by different markers) and in different years.

2019 time frame, we observe a monotonic increase in its value for both communes and Urban units. The result is illustrated in the bar chart in Figure 1(b), and reveals that the imbalance above has been in fact growing in the observed four years. The outcome is somewhat surprising, as one could reasonably expect that the digital divide is actually decreasing in a developed country like France, where, *e.g.*, broadband mobile network technologies are pervasive. Therefore we are led down a path of questioning regarding what the causes for such an anomalous behavior might be.

A. Confounding factor analysis

A legitimate question is whether the disparity in per-capita mobile traffic consumption is biased by confounding factors. So we conduct an analysis of the following social factors:

Broadband coverage. The first obvious aspect that may affect the results is the availability of broadband 4G coverage in the set of studied communes. Using the network coverage information from open source, available for years from 2017 to 2019, we compute for every commune its 4G coverage ratio, *i.e.*, the portion of its territory covered by 4G connectivity, separately in each year. We then group communes into five classes, based on their population: $3,000$–$5,000$, $5,000$–$15,000$, $15,000$–$30,000$, $50,000$–$100,000$, and over $100,000$ inhabitants. For each class, we compute the mean and the standard deviation of the 4G coverage ratio of the associated communes. The top-left plot in Figure 2 shows the result: as expected, 4G coverage has improved significantly between 2017 and 2019, reaching values close to 100% in almost all considered communes, with minimal differences among the five classes. As no 5G service was yet deployed in 2019 in France [**?**], the result highlights a *reduction* of the digital divide in terms of mobile broadband accessibility. This is in stark contrast with the inequality in usage and confirms that the imbalance we unveil configures as a form of the second-level digital divide.

Income. Previous studies on this phenomenon have also pointed at income as a source of inequality in mobile service adoption. By using income data which is open source by INSEE and adopting a similar approach of clustering communes as done for 4G coverage, we find no clear relationship between economic wealth and traffic usage at a commune level. The top-right plot in Figure 2 shows minimum changes in the mean value of the income across communes or years, which cannot therefore explain the observed divide.

Commuting. Larger cities tend to attract commuters during working hours, which inflates their actual mobile subscriber population with respect to the inhabitants recorded on the census. The bottom-left plot in Figure 2 is computed using the work mobility survey data and highlights how the ratio of commuters per inhabitant actually grows with the number of residents in the commune. However, the effect is constant across years, hence cannot explain the increasing imbalance over years in per-capita mobile traffic consumption.

Active population. Mobile services are utilized in diverse ways and quantity by people of different age. In France, individuals in the 14–59 age range are the most active mobile. We thus investigate whether an uneven presence of inhabitants of that age may represent a confounding factor for our study. Specifically, we employ the population age data to compute the Revealed Comparative Advantage (RCA) of residents who are 14 to 59 years old in each commune; the RCA measures the higher or lower incidence of people within that age range with respect to the average over all communes. The bottom-right plot in Figure 2 shows that a difference exists across cities, and larger municipalities indeed present a higher fraction of more active mobile network users. Yet again, this diversity does not vary across years, hence cannot be the cause of the growing divide.

Finally, we can speculate that it may reflect a stronger natural inclination (which has been reinforced over recent years) of inhabitants of larger cities to rely on and benefit from mobile services as none of the social factors explain the cause of the second-level digital divide.

Acknowledgements

S. Mishra's work has been supported by NetSense, grant no. 2019-T1/TIC- 16037 funded by Comunidad de Madrid. This work has also been funded by the research project CoCo5G (Traffic Collection), grant no. ANR-22-CE25-0016, funded by the French National Research Agency (ANR).

References

[1] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15888–15893, 2014.

[2] R. W. Douglass, D. A. Meyer, M. Ram, D. Rideout, and D. Song, "High resolution population estimates from telecommunications data," *EPJ Data Science*, vol. 4, pp. 1–13, 2015.

[3] P. DiMaggio, E. Hargittai, C. Celeste, and S. Shafer, "From unequal access to differentiated use: A literature review and agenda for research on digital inequality," in *Social inequality* (K. Neckerman, ed.), pp. 355–400, 2001.

Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters

Takahiro Yabe[1], Bernardo Garcia Bulle Bueno[1], Xiaowen Dong[1,2], Alex Pentland[1], Esteban Moro[1,3]
[1]MIT, [2]University of Oxford, [3]Universidad Carlos III de Madrid

Keywords: urban segregation, human mobility, behavior modeling

Introduction

Studies have shown that the diversity of social networks is a significant factor of economic growth, social capital, and community resilience. However, in addition to the already rising inequality and segregation, the COVID-19 pandemic and the consequential countermeasures including mobility restrictions have posed significant challenges for maintaining both the quantity and quality of such physical encounters in cities. Large-scale location data have been used to understand the nature of physical encounters of people in cities (Eagle et al., 2009). A recent study using mobile phone data in 10 American cities revealed that peoples' mobility behavior, as opposed to their residential locations, account for 55% of urban segregation (Moro et al., 2021). Such studies based on mobility data have provided a more comprehensive understanding of income segregation in urban environments. The aftermath of the pandemic has brought also significant changes in behavior in our cities, including less use of public transportation, more hours working from home, and higher usage of online food and goods delivery services. However, little is understood about how much longitudinal effects the pandemic has had on the quantity and quality of our encounters in urban environments. Measuring the dynamics and potential causes of changes in the diversity of urban encounters across different periods of the pandemic could be valuable in understanding the long-term impacts of the pandemic on cities, and for developing resilient policies to better prepare for future outbreaks.

Data and Methods

Using a large and longitudinal dataset of GPS location records in four major metropolitan areas in the US across more than three years, we analyze how experienced income diversity of urban encounters has changed during different periods of the COVID-19 pandemic. Specifically, we analyze the dynamics of income diversity of encounters at the level of individual places and individual users in cities. Mobility data was provided by Spectus, who supplied anonymized, privacy-enhanced, and high-resolution mobile location pings for more than 1 million devices across four U.S. census core-based statistical areas (CBSAs). All devices within the study opted-in to anonymized data collection for research purposes under a GDPR and CCPA compliant framework. Our second data source is a collection of 433K verified places across four CBSAs, obtained via the Foursquare API. Each anonymized individual user in the dataset was assigned a socioeconomic status (SES) proxy, estimated from their home census block group (CBG) using the 2016-2020 5-year American Community Survey (ACS), and were then categorized into four equally sized SES quantiles. Given the estimated SES quantiles of individual users and the visited POIs, we measured the experienced income diversity at each place α (denoted as D_α) and experienced by each individual i (denoted as D_i). The diversity measures were computed for each 2-month moving window to ensure a sufficient number of visits to POIs, and were deseasonalized using monthly trends observed in 2019.

Diversity has decreased even though mobility has recovered

Fig. 1a shows how experienced income diversity at places around the Boston and Cambridge area substantially decreased during the first wave of the pandemic, and has not fully even after more than 1 and a half years from the lockdown, in October 2021. Given the recovery of aggregate mobility metrics shown in Figure 1b, one could expect the income diversity of encounters to also return back to pre-pandemic levels by late 2021. However, as shown in Fig. 1c, the income diversity experienced at places and by individuals is consistently lower than the pre-pandemic levels for all four cities even after 2 years into the pandemic. Cities experienced the most decrease in diversity in April 2020, 30% lower than pre-pandemic levels during the lockdown. A second peak in the loss of diversity is observed in late 2020, which corresponds to the

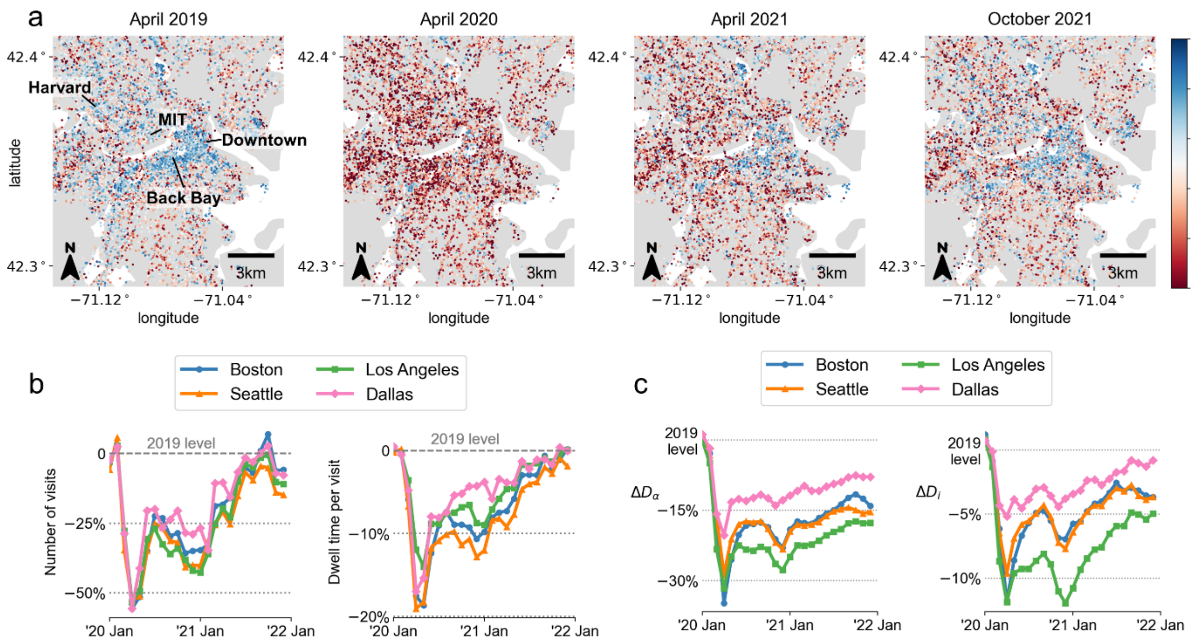


Figure 1. Decreased diversity of urban encounters. a) Map shows that the income diversity of encounters in places in the Boston and Cambridge area decreased during the pandemic. Diversity gradually recovers with reopening, albeit not fully compared to pre-pandemic levels, even in October 2021. b) Aggregate mobility metrics, such as the daily number of visits per individual and daily amount of time spent at places have returned to pre-pandemic levels by late 2021. c) Despite the recovery in mobility statistics, the diversity of encounters experienced at places and by individuals has decreased by 10% to 20% at places, and have not recovered back to pre-pandemic levels.

increase in cases due to the first SARS-CoV-2 variant. Despite the recovery of individual mobility metrics, income diversity of encounters is still around 10% less than pre-pandemic levels even by late 2021.

Why has diversity decreased? – Reduction in social exploration

To investigate the behavioral factors that led to the consistent decrease in income diversity experienced at places and by individuals, we consider three possible hierarchical levels of changes in the behavior of individuals due to the pandemic: (i) reduction in the total amount of time spent at places outside homes and workplaces, (ii) changes in travel distances for each income quantile, and (iii) microscopic changes in place preferences, including changes in exploration behavior and visitation patterns across place subcategories. Using counterfactual mobility simulations and the social exploration and preferential return model, we revealed that reduction in activity and travel distances explain around 55% of the decreased diversity during the first wave of the pandemic, however, the remaining 45% is due to more microscopic behavioral changes. In particular, people's willingness to socially explore substantially decreased by 5% to 10% compared to the 2019 levels in all four cities, leading to less experienced diversity.

Acknowledgment

We thank Spectus for providing the mobility data used for the analysis through their Social Impact program.

References

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 106(36), 15274-15278.

Moro, E., Calacci, D., Dong, X., & Pentland, A. (2021). Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications*, *12*(1), 1-10.

Modeling and understanding the impact of COVID-19 containment policies on mobile traffic consumption for cities in France

André Felipe Zanella*†, Stefania Rubrichi‡, Zbigniew Smoreda‡ and Marco Fiore*
*IMDEA Networks Institute, Spain, †Universidad Carlos III de Madrid, Spain, ‡SENSE / Orange Innovation, France

Introduction: The COVID-19 pandemic impacted people's lives in a way never seen since on the digital era. Governments worldwide adopted measures to restrict the spread of the virus, resulting in major changes of routine for local populations. Those changes not only affected individuals, but resulted in new patterns of utilization of digital services. As new waves of contagion happened, different restriction measures were put in place according to the severity of COVID-19 cases in each region. Understanding the impact due to those measures on telecommunication services can help operators comprehend how their products are affected by large scale changes in populations and help authorities comprehend the effectiveness of their measures. Previous studies explored how mobile phones data could help identify changes in mobility and the relation with socioeconomic indicators on multiple countries [1]–[5] and changes in international mobility [6]; explored how to detect COVID-19 hospitalization and derive epidemic risk maps [7] and the patterns in usage across time and space [8], [9]. Those focus on country-level dynamics on earlier stages of the pandemic, in special the first wave. We'll be exploring the impacts on smartphone usage due to later stages mobility restrictions of the COVID-19 pandemic inside the biggest cities in France, aiming to find a set of socioeconomic indicators that can explain those changes.

Data: We studied the three transitional periods $[T_L, T_{D1}, T_{D2}]$, each representing the difference of mobile traffic consumption on the 14 days before/after the transition of restriction measures: T_L represents changes when the country entered its 3rd nationwide lockdown on 03/04/21, resulting in closed schools, non-essential shops and prohibition of non essential travels; T_{D1} represents the 1st phase of lifting restrictions after the lockdown ended on 05/05/21, removing the restriction of trips and reopening schools; T_{D2} represents the 2nd phase of lift-off on 19/05/21, with the reopening of most non essential shops, cinemas, theatres, museums and places to eat. Data was collected on the production network of Orange. We restrict our studied period to 7pm-1am of workdays (when people leave their work, go home or to leisure locations); those can help correlate changes of traffic with urban units (IRIS) level data from the Census (related to the place of residency) and leisure locations. We chose the 10 cities with highest population in France: Paris, Marseille, Lyon, Toulouse, Nice, Nantes, Montpellier, Strasbourg, Bordeaux and Lille; aiming to understand the complex changes of mobile traffic consumption in big urban centers in France due to different mobility restrictions imposed by the government. We also standardized the traffic for every city and period, to make comparisons fair.

Results: Figure 1A shows the per differences for Paris on studied periods, with red/blue indicating a decrease/increase in mobile traffic consumption in relation to the mean traffic difference on each IRIS. We note a spatial dependency on the changes across IRIS, which can be matched for entire neighborhoods of the cities. Those changes can be related to residential and commercial zones of cities: on T_L, commercial and leisure areas saw a significant reduction of mobile traffic consumption, while residential areas saw an increase; this pattern was promptly reverted during T_{D1} and with changes in T_{D2} being even more focused across leisure areas. We also note that the differences seen when the lockdown started were reverted after restrictions were loosened, matching what was seen for mobile traffic at nationwide level [9]. To further understand the traffic difference patterns, we propose a spatial lag model (SLM) using socioeconomic indicators as regressors, with the goal to predict mobile traffic variations with a singular model for the 10 biggest cities in France. We choose as features: population density, median income and leisure density (composed of restaurants and non essential stores). The geographical distribution of features of Paris is seen on Figure 1C. We achieve a satisfactory Pearson correlation coefficient between real and predicted values of $[T_L : 0.703; T_{D1} : 0.85; T_{D2} : 0.867]$; values per city are always above 0.6, meaning all cities fit well to the general model. We further observe the satisfactory prediction power when comparing the maps predicted by the model on Fig. 1B with the real traffic difference of Fig. 1A, where spatial patterns were well predicted and the only differences being in minor shifts of the absolute values.

We also analyze the coefficients for each $[T_L, T_{D1} T_{D2}]$, seen on Fig. 2, to understand the relations of changes in mobile traffic consumption with socioeconomic characteristics of the cities. The first trend is the shift between positive/negative values across periods, related to the reversing effects previously seen on the traffic difference maps. When the nationwide lockdown T_L started, we note densely populated areas with an overall growth of mobile traffic usagem which can be due to people being unable to leave far from their residence. Similarly, areas with a dense presence of leisure locations had a decrease of traffic usage, which relates to most establishments being forced to close. Median income presents another interesting trend, where regions with a higher income were related with less traffic consumption, which could indicate that this population was willing to leave their homes during lockdown, such as move out of the city for secondary residencies. On the opposite end, areas with a lower median income saw an increase of traffic, meaning people were unable to leave their residence for the lockdown and utilized more their mobile data at home (in a similar trend to the growth of traffic in areas with denser population). T_{D1} saw an initial reversal of trends: since the government started to slowly lift restrictions, the wealthier share of the population came back into their urban residences, resulting in a growth of mobile data consumption in regions with higher median income. Even though officially they were not allowed to open, areas with a denser presence of leisure locations started to see activity again, indicated by a growth traffic. We conclude that T_{D1} represented a slow change of dynamics due to the small values of coefficients. Finally, the second phase T_{D2} further accentuated the reversal slowly initiated in T_{D1}, with coefficient values for population and leisure density in opposition to T_L. This can be interpreted as a return to normality: people were allowed to stay out of their homes in the early evening (which wasn't officially possible since late 2020), represented by a negative coefficient for the population density feature. This transitioned also occurred during spring: due to improvements in weather, later hours for sun setting and looser restriction measures, people are willing more to out of their homes (if allowed). It also resulted in a positive coefficient in areas with a dense presence of leisure places: since being officially open, those spaces had a surge of mobile traffic consumption due to visitors being welcomed again. Finally, we see that the trend of people being less in their homes and more at stores and restaurants is uniform across different income levels, due to the median income coefficient having a negligible value.

Conclusions: The study sheds a light in the vast diversity of smartphone usage during the pandemic. The patterns are highly dependent on in real life events, completely shifting well known behaviors. Our study helps understand population changes across residential and leisure areas of the cities due to mobility restrictions. This can help governments understand the effectiveness of their actions across different indirect data sources and guide companies comprehend the regional impacts over their networks during major events, allowing better scalability of systems for disruption due to major anomalies.

Acknowledgements: A. F. Zanella's work was supported by BANYAN project, which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 860239.

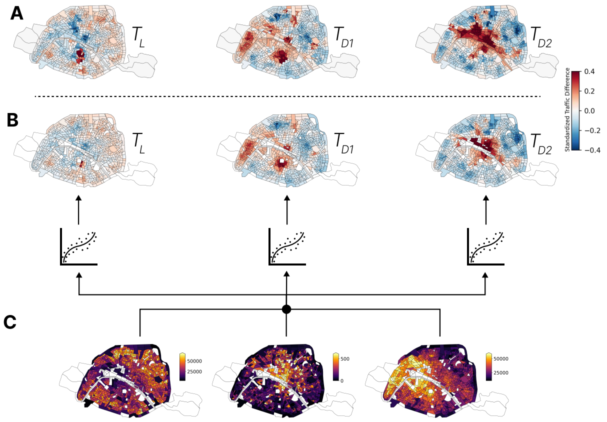


Fig. 1. (a) Standardized traffic difference in Paris for the three studied periods; (b) predicted values by the proposed spatial model; (c) Spatial distribution for the selected model features.

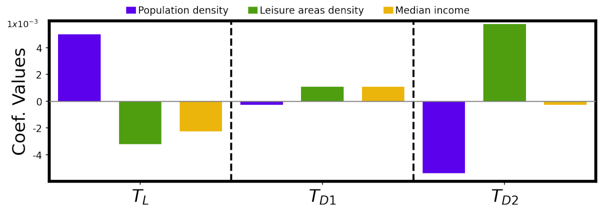


Fig. 2. Fitted coefficients for the spatial model across all cities, for each transitional period.

References

[1] G. Pullano, E. Valdano, N. Scarpa, S. Rubrichi, and V. Colizza, "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the covid-19 epidemic in france under lockdown: a population-based study," *The Lancet Digital Health*, vol. 2, no. 12, pp. e638–e649, 2020.

[2] E. Valdano, J. Lee, S. Bansal, S. Rubrichi, and V. Colizza, "Highlighting socio-economic constraints on mobility reductions during COVID-19 restrictions in France can inform effective and equitable pandemic response," *Journal of Travel Medicine*, vol. 28, 04 2021. taab045.

[3] L. Gauvin, P. Bajardi, E. Pepe, B. Lake, F. Privitera, and M. Tizzoni, "Socio-economic determinants of mobility responses during the first wave of COVID-19 in italy: from provinces to neighbourhoods," *J. R. Soc. Interface*, vol. 18, p. 20210092, Aug. 2021.

[4] J. Kim and M.-P. Kwan, "The impact of the covid-19 pandemic on people's mobility: A longitudinal study of the u.s. from march to september of 2020," *Journal of Transport Geography*, vol. 93, p. 103039, 2021.

[5] A. Glodeanu, P. Gullón, and U. Bilal, "Social inequalities in mobility during and following the covid-19 associated lockdown of the madrid metropolitan area in spain," *Health & Place*, vol. 70, p. 102580, 2021.

[6] Luca, Massimiliano, Lepri, Bruno, Frias-Martinez, Enrique, and Lutu, Andra, "Modeling international mobility using roaming cell phone traces during covid-19 pandemic," *EPJ Data Sci.*, vol. 11, no. 1, p. 22, 2022.

[7] E. Cabana, A. Lutu, E. Frias-Martinez, and N. Laoutaris, "Using mobile network data to color epidemic risk maps," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology*, SpatialEpi '22, (New York, NY, USA), p. 35–44, Association for Computing Machinery, 2022.

[8] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis, "A year in lockdown: How the waves of covid-19 impact internet traffic," *Commun. ACM*, vol. 64, p. 101–108, jun 2021.

[9] A. F. Zanella, O. E. Martínez-Durive, S. Mishra, Z. Smoreda, and M. Fiore, "Impact of later-stages covid-19 response measures on spatiotemporal mobile service usage," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pp. 970–979, 2022.

Main conference

Posters

Mobility data assimilation to forecast electricity demand in a sobriety context in France

Nathan Doumèche[1,2] **Yann Allioux**[2] **Yannig Goude**[2] **Stefania Rubrichi**[3]

[1] Sorbonne University, CNRS,
Laboratoire de Probabilités, Statistique et Modélisation, LPSM,
F-75005 Paris, France
nathan.doumeche@sorbonne-universite.fr
[2] Électricité de France R&D
{nathan.doumeche, yann.allioux, yannig.goude}@edf.fr
[3] Orange Innovation, SENSE lab
stefania.rubrichi@orange.fr

Energy is at the very core of modern economies and politics. In addition, its impact on climate change is forcing our society to change its consumption patterns. As a result, there is a growing interest in energy savings and in the transition to sustainable energy sources. In France, electricity is the main source in the energy mix. The French ecological transition plan is based on on a massive electrification of the uses powered by nuclear energy. The recent context of high energy prices due to the post-covid industrial renewal and to the war in Ukraine has had an important impact on European economies and has been accompanied by energy savings in Europe. In this article, we document electricity savings in France during the winter of 2022-2023, which various media have described as a sobriety period (The New York Times, 2022).

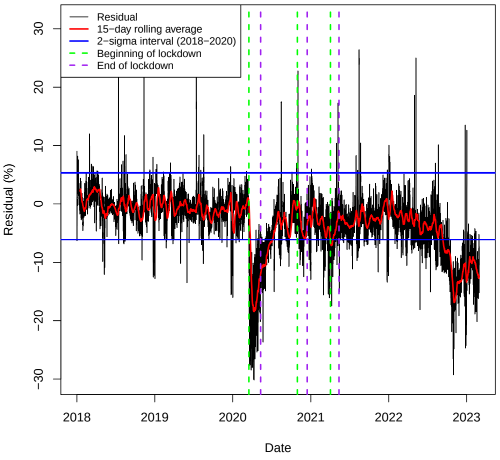


Figure 1: Gap between the effective French electricity demand and the expected demand

However, there is no well-established modeling of the impact of price increases and government incentives on electricity consumption patterns. In addition, traditional statistical models based on meteorological data struggle to adapt to such brutal ruptures. Recently, machine learning techniques have been applied to electricity load forecasting to ensure the balance of the electricity grid and to reduce electric waste. Indeed, because electricity storage capacity is limited and expensive, the supply of electricity must match demand at all times. As a consequence, electricity load forecasting at different forecasting horizons has attracted a growing interest in recent years. This work focuses on 24-hour ahead load forecasting, which is particularly relevant for operational applications in industry and the electricity market (Nti et al., 2020). Recent research in electricity load forecasting has demonstrated the advantages of adaptive methods that adapt to changes in regimes. In particular, state-space models and online aggregation of experts have proven successful in capturing changes in patterns caused by the COVID-19 crisis (Obst et al., 2021). Most state-of-the-art models rely on historical data of past electricity loads, calendar data such as holidays or the position of the day in the week, and meteorological data such as temperature and humidity. However, such data cannot accurately describe the complex human behaviors that affect the variability of energy demand. As a result, traditional models have struggled to account for brutal societal events such as COVID-19 lockdowns. A deeper knowledge of human behaviors is thus necessary to better model the electricity demand. Over the past decades, complementary datasets generated from cell phone networks, location-based services (LBS), and remote sensors in general have emerged to provide a more fine-grained description of human behavior (Blondel et al., 2015). Regarding day-ahead electricity load forecasting, mobility data from SafeGraph, Google, and Apple mobility reports are strongly correlated with electricity load drops in the U.S. during the COVID-19 outbreaks (Ruan et al., 2020). Though these datasets are very informative about traffic in key urban areas, they were not intended to precisely account for the presence or displacement of the global population. Indeed, there is an intrinsic bias in the data collection, corresponding to the bias of using a specific application for example, which causes the need to adjust the indicators.

Thus, the originality of this work relies on the use of the adjusted high-quality data on human presence provided by the Orange business service Flux Vision to model electricity demand during the sobriety period in France in 2022-2023. This dataset is based on measurements of mobile antenna traffic, while most datasets are based on users' consent to share their geolocation with specific applications. The resulting signal is very stable compared to other datasets where the user base may vary as users unsubscribe from the application. We start by characterizing electricity savings during the sobriety period in France. We then show that mobility data from Orange are correlated with socio-economic indicators. Finally, we show that models using mobility data outperform the state-of-the-art in electricity demand forecasting. Our results show how sequential learning methods are able to capture the changes in mobility behaviors and their impact on the electricity demand.

References

V.D. Blondel, A. Decuyper, and G. Krings. Understanding vehicular routing behavior with location-based service data. EPJ Data Science, 4(10), 2015.

I.K. Nti, M. Teimeh, O. Nyarko-Boateng, and A.F. Adekoya. Electricity load forecasting: a systematic review. Journal of Electrical Systems and Information Technology, 7(13):2314–7172, 2020.

D. Obst, J. de Vilmarest, and Y. Goude. Adaptive methods for short-term electricity load forecasting during covid-19 lockdown in france. IEEE Transactions on Power Systems, 36(5):4754–4763, 2021.

G. Ruan, D. Wu, X. Zheng, H. Zhong, C. Kang, M.A. Dahleh, S. Sivaranjani, and L. Xie. A cross-domain approach to analyzing the short-run impact of covid-19 on the U.S. electricity sector. Joule, 4(11):2322–2337, 2020.

The New York Times. As Russia chokes Europe's gas, France enters era of energy sobriety. Available: https://www.nytimes.com/2022/09/05/business/russia-gas-europe-france.html, 2022. [Accessed: May 11, 2023].

Understanding urban mobility during special events by public transport mobile phone application

Zhiren Huang[1] (zhiren.huang@aalto.fi), Charalampos Sipetas[1], Alonso Espinosa Mireles de Villafranca[1],
Tri Quach[2] and Jari Saramäki[1]
[1]Aalto University, 02150 Espoo, Finland
[2]Helsinki Regional Transport Authority, 00520 Helsinki, Finland

I. Introduction

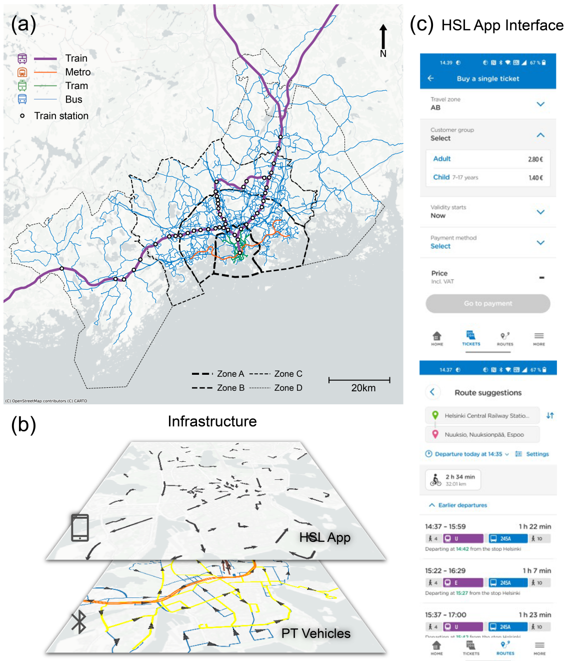
Large-scale special events are crucial for a city's vibrancy and economic growth, but they also present significant challenges to transportation systems due to the complex mobility patterns they generate. For example, people's daily routines may change, leading to changes in travel mode choice and the overall travel demand. Without efficient crowd management and public traffic coordination, such changes can lead to severe congestion or even fatal stampedes, such as the Seoul Halloween crowd crush in 2022. Hence, understanding people's travel demands and behaviours during crowded events is of major importance.

For understanding crowd behaviour, video surveillance has been widely applied to study pedestrian movement patterns within event venues. However, those studies focus on site-specific features, such as the exits or corridors. With the rapid development of ICT, sensing large-scale crowd movements has become feasible. Mobile phone data starts to play an important role in sensing large-scale crowd-gathering processes [1]. Nevertheless, the common limitation of mobile phone data is the lack of comprehensive travel mode information during such events. Other data sources include traffic data collected from intelligent transportation systems, such as smart card data [2]. Those approaches are limited to specific traffic modes. Therefore, how people shifted their travel mode choice during the respective crowded event is unknown. For example, how many people use public transport (PT) to arrive at critical locations related to a special event could be used for PT operation planning. Hence, crowd management and evacuation design could benefit from travel mode information concerning crowded events and special celebrations.

To tackle the obscure travel mode problem, this study presents the spatio-temporal analysis of multi-modal mobility patterns during special events by leveraging an emerging data source. More specifically, ”TravelSense” is a pilot project of the Helsinki Regional Transport Authority (HSL) that uses Bluetooth beacons installed in PT vehicles combined with a mobile phone ticket app to collect door-to-door trajectories from anonymous PT users [3].

II. Materials

This study utilizes data collected from the 1st to the 31st of May 2022, obtained from HSL's TravelSense pilot project. HSL provides PT services for around 1.2 million residents in Helsinki metropolitan area. The study area and the PT network are illustrated in Fig.1a. To facilitate PT trips, HSL offers a mobile phone ticketing application that enables passengers to easily buy tickets, search the best routes, and receive timely updates about PT operations (Fig.1c). The data collection infrastructure (Fig. 1b) incorporates two components, the first employs the HSL mobile ticketing application to sense non-PT trips, while the other utilizes Bluetooth beacons (Fig.1b) installed in PT vehicles (including subway, train, tram, bus, and ferry) to capture PT trips. By integrating data from these infrastructures, we acquire door-to-door trajectories of anonymised passengers. To ensure high data privacy standards, devices are anonymised daily with random IDs, and location coordinates are only accurate to grid cells of 250m x 250m. Additionally, location timestamps outside the PT network are rounded to the nearest quarter-hour.



Buy a single ticket

Travel zone
AB

Customer group
Select

Adult 2.80 €
Child 7-17 years 1.40 €

Validity starts
Now

Payment method
Select

Price
Incl. VAT
-

Go to payment

Route suggestions

Helsinki Central Railway Statio...
Nuuksio, Nuuksionpää, Espoo

Departure today at 14:35 Settings

2 h 34 min
32.01 km

Earlier departures

14:37 - 15:59 1 h 22 min
4 U 245A 10
Departing at 14:42 from the stop Helsinki

15:22 - 16:29 1 h 7 min
4 E 245A 10
Departing at 15:27 from the stop Helsinki

15:37 - 17:00 1 h 23 min
4 U 245A 10

Fig. 1. Data and study area overview. (a) PT network of Helsinki metropolitan area; (b) Infrastructure of TravelSense; (c) Screenshots of HSL mobile ticket application.

The data prepossessing steps, as described in [3], involved

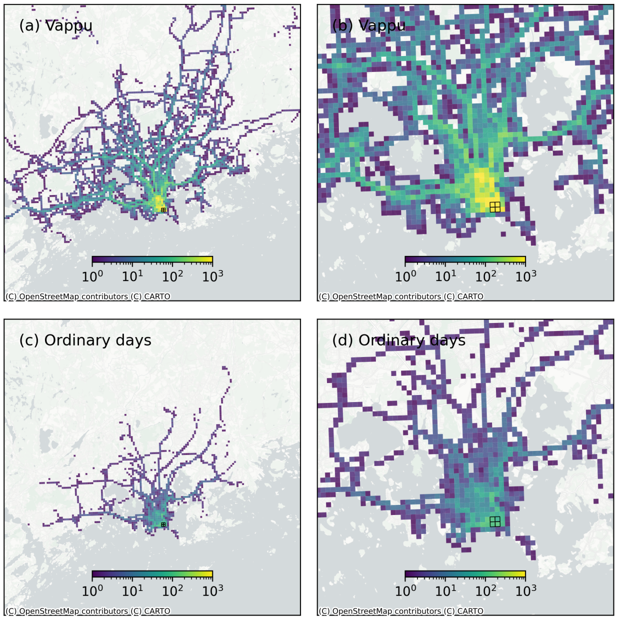


Fig. 2. Heatmap of trajectories visiting the event's main location (a, b) and during other non-working days (c, d). Grids in black lines represent the primary location for Vappu.

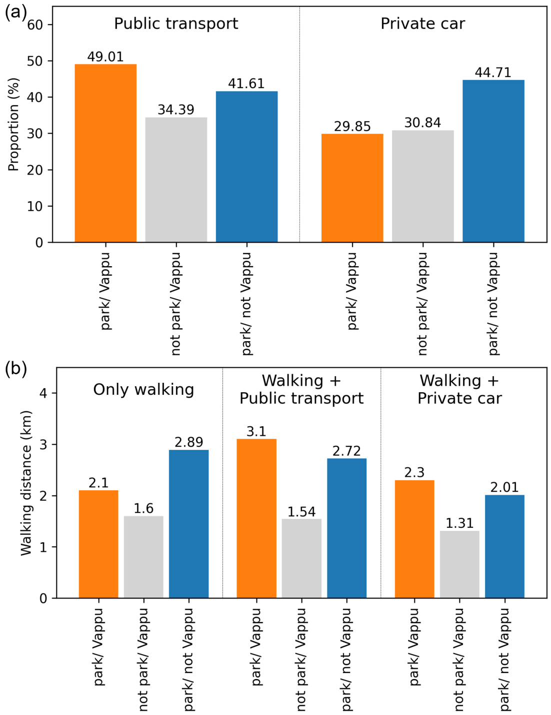


Fig. 3. (a) Proportion of trajectories that contain at least one PT trip or at least one private car trip. (b) Walking distance for trajectories that include only walking, walking and PT, walking and private car.

cleaning the raw data and aggregating the data at different levels. The information is structured based on *legs* (i.e., trips) and *trip chains* (i.e., journeys or trajectories). A trip chain is characterised by the legs that make up each segment and the travel modes used in each of them.

III. Case Study

This case study refers to the Vappu (1st of May) celebration of 2022 which happened to be a Sunday and includes TravelSense data from the entire metropolitan area of Helsinki. For comparison, data from other non-working days are also utilized. By focusing on the travelers who accessed the main event location (i.e., Kaivopuisto park), Fig.2 presents a heatmap resulting from the respective trajectories that visited Kaivopuisto during Vappu and other non-working days. It is apparent that the park was visited by many more travelers during Vappu compared with other days, but also that the trajectories which visited it are more scattered during the special day. It is also implied by visual inspection that the distances traveled to access the park are longer than usual.

Fig. 3a focuses on the proportion of trajectories that visited Kaivopuisto during Vappu ("park/ Vappu"), did not visit Kaivopuisto during Vappu ("not park/ Vappu") and visited Kaivopuisto during ordinary non-working days ("park/ not Vappu"). The analysis refers to PT and private cars users. Travelers that visited the event's main location during Vappu used PT more than private cars, in contrast to what happens during ordinary non-working days. The usage of PT among users that visited this location is also greater compared to PT users who didn't visit this location during Vappu.

Fig. 3b shows the average walking distance for three types of trajectories, including only walking, walking and PT, walking and private cars. People are willing to walk longer distances during Vappu to join the event when the trajectory combines walking with PT or a private car. When people use walking and PT as primary modes for joining the event, they are willing to walk about 3.1 km on average during the whole journey, which is the highest of the three types of trajectories. Analysing the proportion of walking distance in the total distance traveled in a trajectory shows that the highest proportion of walking occurs in the trajectories that combine walking and PT. Overall, it is observed that PT users are willing to walk more during Vappu day to access the event's main location.

Through this case study, we observe that people tend to favor public transport over private cars and are prepared to walk long distances to participate in the event. The study underscores the value of using comprehensive multi-modal data to better understand and manage transportation during large-scale events.

References

[1] J. Candia et al., "Uncovering individual and collective human dynamics from mobile phone records," Journal of Physics A: Mathematical and Theoretical, vol. 41, no. 22, p. 224015, 2008.

[2] Z. Huang, P. Wang, F. Zhang, J. Gao, and M. Schich, "A mobility network approach to identify and anticipate large crowd gatherings," Transportation Research Part B: Methodological, vol. 114, pp. 147–170, 2018.

[3] Z. Huang, A. Espinosa Mireles de Villafranca, and C. Sipetas, "Sensing Multi-modal Mobility Patterns: A Case Study of Helsinki using Bluetooth Beacons and a Mobile Application," in Proceedings of the 2022 IEEE International Conference on Big Data (Big Data). Osaka, Japan: IEEE, pp. 2007–2016. 2022.

Evaluation of Home Detection Algorithms on Mobile Phone Data using Individual-Level Ground Truth

Luca Pappalardo[1], Leo Ferres[2,3], Manuel Sacasa[2]
Ciro Cattuto[3] and Loreto Bravo[2]

[1] ISTI-CNR, Pisa, Italy. email: luca.pappalardo@isti.cnr.it
[2] Universidad del Desarrollo, Santiago de Chile
[3] ISI Foundation, Turin Italy

1 Introduction

Governments and statistical offices want to incorporate new digital data into official statistics to improve efficiency and obtain faster data-driven solutions to sophisticated societal problems. This includes using diverse digital sources like social media, GPS traces, and mobile phone records to estimate well-being indicators [1].

The many possible applications of data analytics to official statistics, such as estimating population density, commuting and migration flows and developing realistic epidemic models, depend critically on our ability to identify *where someone lives*, i.e., detecting an individual's home location. The knowledge of the home location of individuals forms the crucial link between digital data and census data, making it a key enabler for the integration of these two sources of information.

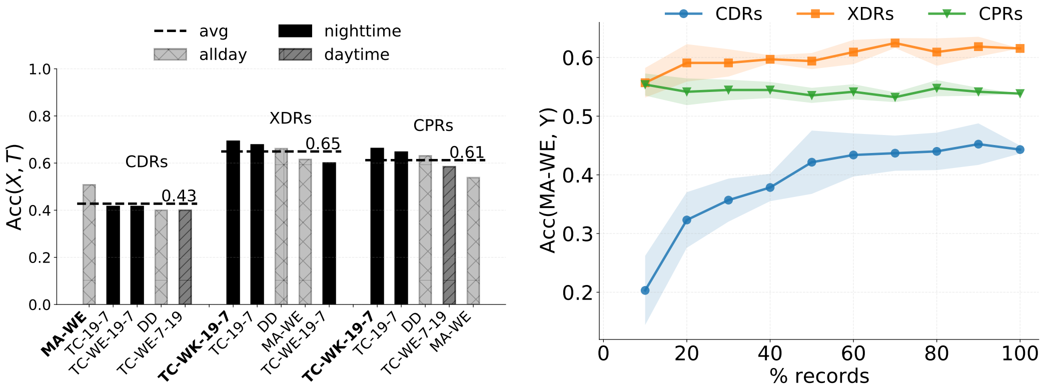
Most of the home detection algorithms (HDAs) proposed in the literature [2–4] process mobile phone records according to *ad-hoc* heuristics rather than principled approaches. They rely on simple decision rules based on how much, and when, an individual calls in each location during the period of observation. The simplest HDA identifies an individual's home location as the one in which they made the highest number of calls during nighttime (e.g., between 7pm and 7am). Other HDAs use a combination of criteria or slight variations of the one mentioned above [4]. Although these algorithms have been used in many works and tools, a thorough validation of their accuracy is still missing.

One reason for this is that, with few exceptions [2, 3], ground truth data at the individual level are not provided by mobile providers for privacy protection reasons, making it difficult to obtain a large enough sample of users for which complete information about positions and residence are available at the same time. For example, Vanhoof et al. [4] provide a high-level validation of the most popular HDAs by comparing each mobile phone tower's population as estimated by official censuses with the number of users whose home is detected to be in that tower. They conclude that there is an urgent need for validation of HDAs at the individual level, i.e., evaluating the performance in detecting home location for individuals for which the actual home location is known [4].

Moreover, validation of HDAs primarily relies on sparse Call Detail Records (CDRs), which only capture user positions during calls [5]. However, due to irregular inter-event times between calls, CDRs offer an incomplete picture of an individual's positions over time. It remains unclear whether eXtended Detail Records (XDRs), generated by individuals and phone devices, or Control Plane Records (CPRs), triggered solely by the mobile phone network, can address CDR limitations and provide more accurate estimates of an individual's home location.

We propose a study that aims at a fine-grained validation of HDAs on individual-level ground truth data and three streams of mobile phone records – CDRs, XDRs, and CPRs. Specifically, 65 users working for Telefónica Chile gave their written consent to provide us access to their phone records for two weeks, as well as their actual address of residence. This information allowed us to correctly assess the accuracy of HDAs, i.e., their capacity to detect a user's actual home correctly, on a ground truth dataset. Our validation reveals, for each data stream, the most accurate state-of-the-art HDA, and that XDRs and CPRs improve the accuracy of HDAs considerably with respect to CDRs (Figure 1a). Moreover, we set up a data minimization experiment to study how the accuracy of detecting home locations changes by the stream used and the number of records for each user. We find that, depending on the stream, just a small fraction of the records is enough to achieve reasonably accurate estimations of an individual's home location (Figure 1b), hence providing a tool to manage the uncertainty and utility trade-off in geo-privacy.

Our individual-level validation paves the road towards the definition of more accurate HDAs and a standardized method for home detection that could make studies more comparable.



(a)

(b)

Fig. 1. (a) Home detection accuracy of the top five HDAs for each stream. Dark bars are algorithms based on nighttime records, dark grey bars are those based on daytime records, and light grey bars are those using all-day records. The dashed line indicates the average accuracy across the top five HDAs. (b) Results of the data minimization experiment. Number of records randomly selected versus the avg and std of home detection accuracy for algorithm MA-WE.

References

1. V. Voukelatou, L. Gabrielli, I. Miliou, S. Cresci, R. Sharma, M. Tesconi, and L. Pappalardo, "Measuring objective and subjective well-being: dimensions and data sources," *International Journal of Data Science and Analytics*, 2020.
2. R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010.
3. V. Frias-Martinez, J. Virseda, A. Rubio, and E. Frias-Martinez, "Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data," in *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, ICTD '10, (New York, NY, USA), pp. 11:1–11:10, ACM, 2010.
4. M. Vanhoof, C. Lee, and Z. Smoreda, *Performance and Sensitivities of Home Detection on Mobile Phone Data*, ch. 8, pp. 245–271. John Wiley Sons, Ltd, 2020.
5. V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, vol. 4, no. 1, p. 10, 2015.

Generating Mobility networks with Generative AI

Giovanni Mauro [1,2], Massimiliano Luca[3,5], Antonio Longa[4,5], Bruno Lepri[5], Luca Pappalardo[1]

[1] ISTI-CNR, Pisa, Italy; [2] University of Pisa and IMT Lucca, Italy; [3] Free University of Bolzano, Italy; [4]University of Trento; [5]Fondazione Bruno Kessler

A mobility network, also known as an origin-destination matrix, is a representation of human flows between geographic locations within a given region. Typically constructed from GPS traces or mobile phone records, mobility networks play a crucial role in mobility analytics and simulation. However, there are instances where these conventional data sources are either unavailable or insufficient for network extraction. This paper addresses the challenge of generating synthetic mobility networks using Artificial Intelligence.

In particular, we propose MoGAN (Mobility Generative Adversarial Network), a deep learning architecture based on Deep Convolutional Generative Adversarial Networks (DCGANs), a particular type of Generative Adversarial Networks (GANs) [1]. MoGAN consists of a generator G, which learns how to produce new synthetic mobility networks, and a discriminator D, which has the task of distinguishing between real and fake (artificial) mobility networks. G and D are trained in an adversarial manner: D maximizes the probability of correctly classifying real and fake mobility networks; G maximizes the probability of fooling D, i.e., to produce fake mobility networks classified by D as real. Both D and G are Convolutional Neural Networks (CNNs), which are proven effective in capturing spatial patterns in the data.

We compare MoGAN with two classical approaches for mobility flows' generation: the Gravity and the Radiation models [2] and a Random Weighted model (RW) that creates a mobility network where the weight of each edge is randomly chosen from the distribution of weights for that edge in the training set. We use four public datasets describing trips with taxis and bikes in New York City and Chicago during 2018 and 2019 (730 daily networks). Two datasets contain daily information regarding bike-sharing services: the City Bike Dataset for New York City and the Divvy Bike Dataset for Chicago. Each record describes the coordinates of each ride's starting and ending stations and the starting and ending times. We define locations (nodes) based on a squared spatial tessellation and compute the flows (edges) as the number of people moving between pairs of these tiles.

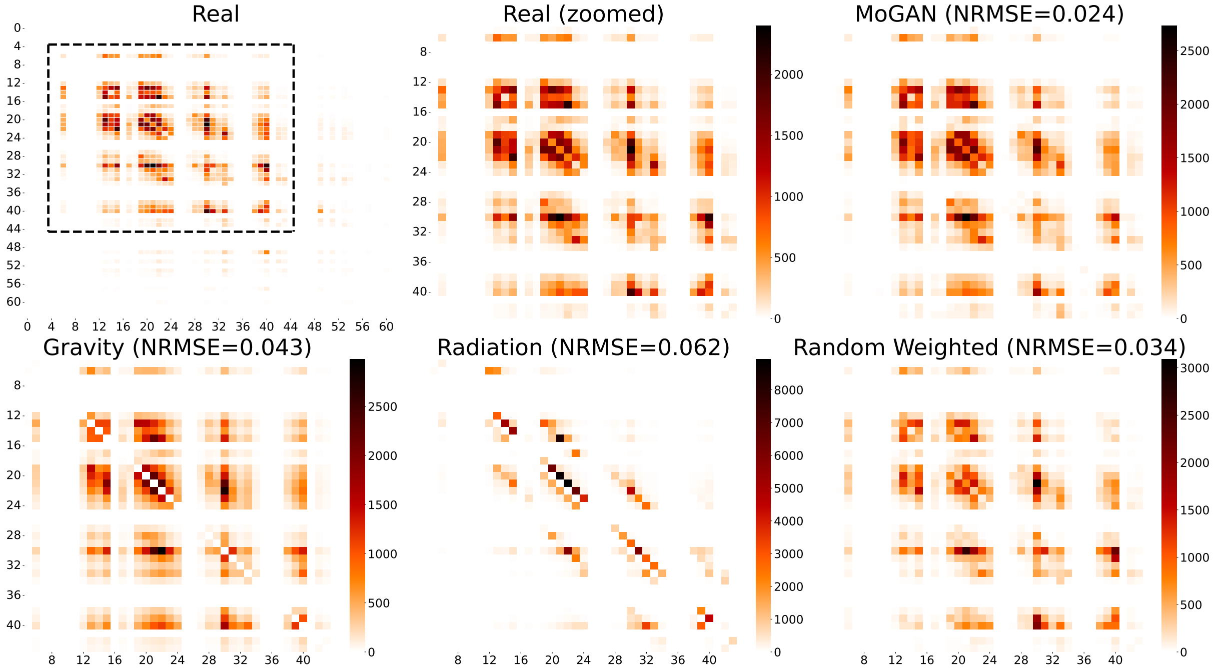
We developed a tailored approach to evaluate the realism of the generated mobility networks. We construct a mobility network for each dataset for each day, obtaining 730 real mobility networks. We split the 730 networks into a training set (584 networks) and a test set (146 networks). We train MoGAN on the training set and generate 146 synthetic mobility networks (synthetic set). We then compute the difference between each network in the synthetic set and each network in the test set, so obtaining $146 \times 146 = 21,316$ values. If the generated mobility networks are realistic, they should differ from the real networks to the same extent real networks differ between themselves. To stress this aspect, we create a set of 146 mobility networks (mixed set), in which half of them are chosen uniformly at random from the test set, and the other half is chosen uniformly at random from the synthetic set. Finally, we compute the pairwise difference between any possible pair of mobility networks in the mixed set.

In our experiments, we calculate several error metrics: such as Normalized Root Mean Square Error (NRMSE), Common Part of Commuters (CPC), and Cut Distance (CD), but also metrics based on the topology of the networks like the Jensen–Shannon (JS) divergence among the weight distributions and several others. MoGAN's CPC distributions overlap almost entirely in all four datasets, meaning that MoGAN generates mobility networks that are indistinguishable from real ones and way more realistic than those generated by the baselines. CPC results are consistent with the results for the other measures (NRMSE, CD, and JS) [3].

We provide another visualization sample of our model's performances in Figure 1. MoGAN is way better than the Gravity model, the best baseline model, at predicting flows between close tiles. The two models reach a similar performance for flows regarding tiles that are very distant to each other.

An important aspect to investigate as future work is also to what extent MoGAN is geographically transferable [4], i.e., it can be trained on a specific city and then used to generate mobility networks in a different city effectively. Another promising future direction is developing a GAN to generate a realistic mobility network for a specific condition (e.g., a rainy day or a day with some public events in the city). In the meantime, our study demonstrates the great potential of artificial intelligence to improve solutions to crucial problems in human mobility, such as the generation of realistic mobility networks. MoGAN can synthesize aggregated movements within a city into a realistic generator, which can be used for data augmentation, simulations, and what-if analysis. Given the flexibility of the training phase, our model can be easily extended to synthesize specific types of mobility, such as aggregated movements during workdays, weekends, specific periods of the year, or in the presence of pandemic-driven mobility restrictions, events, and natural disasters.

The code to train/test MoGAN and reproduce our analyses, mainly conduct and the links to the datasets used in our experiments, can be found at `https://github.com/jonpappalord/GAN-flow`.


Real

Real (zoomed)

MoGAN (NRMSE=0.024)

Gravity (NRMSE=0.043)

Radiation (NRMSE=0.062)

Random Weighted (NRMSE=0.034)

Figure 1: **Visual comparison of the adjacency matrices of the Mobility Networks.** Visualization of the more dense part of the mobility networks of NYC Bikes having the maximum sum of flows observed in the Test Set (Real Zoomed) and of the Mobility Networks having the maximum sum of flows observed in the fake sets produced by all of the other models. Per each generated matrix, we reported the RMSE with respect to the Real matrix. In the top left panel, we show the full 64×64 mobility network and highlight the most dense zones, on which we focus in the other plots of the figure.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[2] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[3] Giovanni Mauro, Massimiliano Luca, Antonio Longa, Bruno Lepri, and Luca Pappalardo. Generating mobility networks with generative adversarial networks. *EPJ data science*, 11(1):58, 2022.

[4] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1):1–44, 2021.

Individual level drivers of seasonal agricultural mobility in Turkey

Bilgeçağ Aydoğdu[1], David Natarajan[1], Subhi Güneş[2], Albert Ali Salah[1,3]
[1] Utrecht University, The Netherlands
[2] Turkcell Technology, Türkiye
[3] Boğaziçi University, Türkiye
Emails: {b.aydogdu,a.a.salah}@uu.nl, d.j.natarajan@students.uu.nl, subhi.gunes@turkcell.com.tr

I. Introduction

Seasonal agricultural migration in Turkey is a complex phenomenon, largely influenced by social, demographic and economic factors. Over the past decade, the landscape of seasonal agricultural employment in Turkey has undergone a substantial shift, primarily driven by the massive influx of refugees from Syria. With the onset of the Syrian refugee crisis in 2013, the wages got lowered and working conditions got worsened in the informal agricultural industry, which pushed some natives out of seasonal agricultural employment [1].

In this work, we aim at developing a model for predicting seasonal mobility from Istanbul, the most populous city in Turkey, to the northeastern hazelnut-producing cities of Ordu, Giresun, and Trabzon between July and August 2020. We also assess the drivers of seasonal mobility using extended detail records (xDR) from the largest telecom operator in Turkey. Recent research done with a survey collected in 2018 [2] demonstrated that the agriculture sector is one of the primary employment areas, especially for refugee women. Surveys and census studies give insights on the characteristics of seasonal workers, but are costly to apply.

The contribution of this study is to (1) develop xDR-based indicators of seasonal mobility and (2) show the usefulness of xDR-based mobility and census-based features in generating insights on the characteristics of seasonal agricultural workers. Previous studies have analyzed the usefulness of call detail records (CDR) [3] [4] for measuring seasonal migration, yet xDR is an underused data source in migration analysis. Temporal frequency of signals in xDR is higher than CDR, which helps for creating more complete trajectories and more accurate mobility and migration indicators, thus closing some data gaps in seasonal agricultural mobility.

II. Data

The xDR datase we use is prepared within the scope of the HummingBird EU H2020 Project [5], following the fine-grained mobility approach (and accompanying ethical framework for data anonymization and aggregation) suggested in Data for Development (D4D) [6] and Data for Refugees (D4R) challenges [7]. It includes user id, timestamp, and the id of the cell tower used by the user. We enriched the xDR with (noisy) nationality and sex flags using the telco indicators, and the noise in these flags serves further anonymization. Up to thirty percent of users flagged as ”male” may be females. The datasets are shared with our research group under strict data use agreements and ethical approvals.

We have approximately 100,000 users subsampled with replacement from a pool of 4 million users every two weeks throughout 2020. The short sampling period is used to prevent profiling of users. We combine these data with various other data sources, such as the night light satellite data created from the Earth Observation Group (EOG) sources, and neighborhood level indicators collected in Istanbul as a part of the Mahallem Istanbul project[1]. We used spatial scaling methods to recalculate neighborhood-level indicators around the cell towers. These indicators give insights into the demographic characteristics of the Turkish population living around the cell tower such as education, age, and maritial status.

III. Methodology

We processed the anonymized fine grained mobility data sets to identify the users whose home location (based on activity during 18:00-06:00) is in Istanbul, and who have been to the Northeastern cities in the two week sampling period. The number of travellers originating from Istanbul to the harvesting cities peaks in the first week of August. In order to analyze the drivers of this mobility behavior, we focus on the trips that took place between 15/07/2020 and 31/08/2020, the high season for harvesting hazelnuts[2]. Although we cannot know whether trips are indeed associated with agricultural work with high certainty, this approach optimizes the chances that is the case. The gender and nationality breakdown of the groups are given in Table I.

We used the scikit-mobility library [8] to calculate six different mobility-related indicators, namely, the radius of gyration, maximum distance from home, number of visits, number of locations, maximum distance travelled, and total distance travelled calculated in straight linesseperately for night (between 18:00 and 06:00) and day for all days that the user spent in Istanbul.

This work was funded by the European Union Horizon 2020 Research and Innovation funding Program, project number 870661.

[1] https://www.kalkinmakutuphanesi.gov.tr/dokuman/mahallem-istanbul/554

[2] https://arastirma.tarimorman.gov.tr/findik/Sayfalar/Detay.aspx?SayfaId=32

	Likely Agricultural Migrant	Likely Non-Migrant
Turkish Male	532	25614
Turkish Female	185	11310
Syrian Male	188	16432
Syrian Female	76	6747
Total	981	60103

Table I: Number of users grouped by category.

	F1 (mean)	F1 (std)	AUC-ROC (mean)	AUC-ROC (std)
Turkish Male	0.68	0.03	0.75	0.04
Turkish Female	0.63	0.05	0.70	0.05
Syrian Male	0.72	0.05	0.80	0.04
Syrian Female	0.67	0.09	0.76	0.08

Table II: Performance metrics across subsets.

Following [9], we first develop a model to predict seasonal mobility from urban mobility related metrics (which indicates employment) and census-driven indicators, and then we analyze feature importances to understand the drivers. We develop four different prediction models for each demographic group. For Syrians, we do not use census features, as they were collected for the Turkish population. Including census features improves the prediction model for Turkish group only slightly, whereas it lowers the performance for Syrians.

In our data set, mobility related metrics were highly correlated. We used a principal component analysis (PCA) to day and night mobility related variables seperately to reduce the dimensionality to four by keeping 80 percent of the variance. Furthermore, we dropped various demographic features that were in high correlation with other features (correlation estimate higher than 0.8) based on a collinearity test.

The challenge of our data is the great imbalance between the number of users who have been and who have not been to the harvest cities. To address this, we randomly downsampled the non-migrant group to the same numbers with the migrant group [9], for a hundred times. We applied 5-fold cross validation in each such sample. We used logistic regression, as our aim is to have simple and interpretable model. The feature importances are computed for each fold, in each sample.

IV. Results and Discussion

We share the model performances in Table II, measured by using F1 and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) scores that are averaged across all experimental samples. The model we developed has the highest performance for predicting the seasonal mobility of Syrian males, and the AUC-ROC scores are comparable to similar studies in the literature [9].

To understand feature importances we report the odds ratio (OR) for all features averaged across all samples in Table III. In logistic regression, OR measures the impact of features on the individual decision to go to the harvesting regions during the high season (value above 1 indicates the chances of going). In Table III, we report the top three most positively (blue) and negatively (red) influential features for each group.

	Turkish M	Turkish F	Syrian M	Syrian F
Day mobility PC 1	1.15	1.04	1.43	1.24
Day mobility PC 2	0.63	0.59	0.55	0.62
Night mobility PC 1	1.86	1.96	2.27	1.95
Night mobility PC 2	0.74	0.91	0.77	0.62
Nightlight brightness	1.08	1.14	1.24	1.21
Percentage nocturnal	0.96	0.96	0.91	0.84
Average Age	0.84	0.83	-	-
Age Dependency ratio	1.47	0.70	-	-
Baby boomer ratio	0.58	1.10	-	-
Illiterate ratio	1.31	1.36	-	-
Literate but no education ratio	0.42	0.94	-	-
Average education dur. (men)	1.10	1.16	-	-
Married ratio	1.29	1.46	-	-
Total female population	1.02	1.22	-	-
Population density	1.08	0.86	-	-

Table III: Mean Odds Ratios per subset.

The results show that urban mobility related principal components (PC) are highly influential in decision making for seasonal mobility for all groups. While the ”maximum distance” has a positive association, ”radius of gyration” and ”number of locations” have a negative association with seasonal mobility (i.e. seasonal migrants tend to be less mobile, but travel greater distances). For users with less mobility, the unemployment probability is higher. The illiterate population ratio around the home location cell tower is positively associated with the target variable for the Turkish group. For women, we also found that ”married ratio” is associated positively, but ”average age” is associated negatively with seasonal mobility.

In this study, we show that the xDR data can help to improve our understanding of the drivers of seasonal mobility. Both for Syrians and Turkish, the biggest driver of seasonal migration is the unemployment. For Turkish population, we also found that the illiteracy can be one of the biggest drivers. We argue that these results can complement the information collected on the seasonal agricultural workers through surveys.

References

[1] Z. Bayramoglu and M. Bozdemir, “Dış göçlerin mevsimlik tarım işçiliği üzerine etkilerinin değerlendirilmesi,” *Journal of the Institute of Science and Technology*, vol. 9, no. 2, pp. 1164–1176, 2019.

[2] M. Demirci and M. G. Kırdar, “The labor market integration of syrian refugees in turkey,” *World Development*, vol. 162, p. 106138, 2023.

[3] P. J. Zufiria *et al.*, “Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. application in food security,” *PloS one*, vol. 13, no. 4, p. e0195714, 2018.

[4] S. Turper Alışık *et al.*, “Seasonal labor migration among Syrian refugees and urban deep map for integration in Turkey,” in *Guide to Mobile Data Analytics in Refugee Scenarios*. Springer, 2019, pp. 305–328.

[5] B. Aydogdu, A. A. Salah, O. Ones, and B. Gurbuz, “Description of the mobile CDR database,” *HumMingBird Project Deliverable*, vol. 6, no. 1, 2021. [Online]. Available: https://hummingbird-h2020.eu/publications

[6] V. D. Blondel *et al.*, “Data for development: the D4D challenge on mobile phone data,” *arXiv preprint arXiv:1210.0137*, 2012.

[7] A. Salah, A. Pentland, B. Lepri, and E. Letouzé, *Guide to Mobile Data Analytics in Refugee Scenarios*. Springer, 2019.

[8] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini, “scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data,” *arXiv preprint arXiv:1907.07062*, 2019.

[9] V. Dias, L. Fernando, Y. Lin, V. Frias-Martinez, and L. Raschid, “Framework to study migration decisions using call detail record (CDR) data,” *IEEE Transactions on Computational Social Systems*, 2022.

Anomaly Detection in Mobile Networks

Veena B. Mendiratta
Northwestern University, USA
veena.mendiratta@northwestern.edu

I. Introduction

Mobile networks are designed to be highly reliable, and it is important to monitor and detect anomalous behaviors in a timely manner since the occurrence of network failures can be costly. The network complexity exacerbates the difficulty of failure detection as the the symptoms may be small degradation in performance, making it difficult to detect with typical Key Performance Indicators and threshold based methods. Nevertheless, large amounts of log data are generated by these networks where a mobile network server can generate hundreds of thousands of records per minute, each with hundreds of fields.

The focus of this work is the development of an anomaly detection framework using network log data. We develop a novel scheme for data aggregation to obtain multivariate summary data that represents the *system* state for a given time period. Next, we propose a multivariate unsupervised anomaly detection model that can be used over time regardless of the network evolution. To find the subspace where the variation of the normal data does not change over time, we decompose the normal data using Principal Component Analysis (PCA) and choose the principal components (PC) with low variation. With these PCs, we build a detection model that discriminates the anomalies from the normal data. The impact of our work is the proactive detection of network anomalies, thereby improving network reliability and availability. The approach is not specific to the wireless domain, and can be applied to other types of network anomaly detection with large feature sets using unsupervised learning methods.

The algorithms are developed and tested with Per Call Measurement Data (PCMD) records from a 4G-LTE network. PCMD records are generated by the Mobility Management Entity (MME) of the LTE network and provides access to network data for voice, text and data calls or sessions activated on the network, and are available in near real-time. The data are collected on a per-procedure basis with over 70 different procedures defined in PCMD. Each record contains 250 fields related to metrics such as error codes, signaling and handover performance, data throughput, etc.; and are generated when a procedure ends. The PCMD reporting period is tunable, and we analyze the data in 10-second intervals.

In the following sections we present the proposed methodology and experimental results.

II. Methodology

When the number of anomalies is limited, an unsupervised approach is used, which characterizes the variation of the normal instances based on which a detection model is developed. Before developing a detection model, we apply PCA on the normal instances to capture the true behavior of the normal instances, and to reduce the number of dimensions and derive a new set of uncorrelated features.

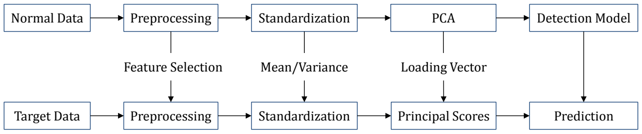


Fig. 1. Anomaly detection procedure

Among d principal components we consider some of the middle principal components which have non-negligible variation that is time invariant, and characterize the true behavior of the normal instances. With the middle principal components, we derive the following statistic which follows a Chi-square distribution, where the degrees of freedom is the number of middle principal components, if the original features follow a normal distribution due to the linearity of Y with respect to A and the orthogonality of X:

$$\sum_{l<\lambda_j<u} \frac{Y_{ij}^2}{\lambda_j} \tag{1}$$

where, Y_{ij} is the j^{th} principal component of the i^{th} instance, λ_j represents the sample variance of the j^{th} principal component, and l and u represent the lower and upper limits for selecting the middle principal components.

This statistic geometrically measures the squared Mahalanobis distance of the i^{th} instance to the origin in the reduced principal component space, consisting of the middle principal components. Hence, an instance having a large distance can be considered an anomaly since it deviates by a large amount from the average behavior of the normal instances. Therefore, we use the statistic in (1) to detect anomalous behaviors by comparing the values of the statistic to the reference value $\chi^2_\alpha(m)$, where α represents the significance level and m is the number of middle principal components.

Figure 1 shows the anomaly detection methodology.

III. Experimental Results

Normal data from 2014, 2015 and 2017 (total of 130 minutes) and outage data from 2014 and 2015 was used for the experiments. As described above, we build a detection model using the normal data and test its performance on the outage cases. After data aggregation, the features that are not

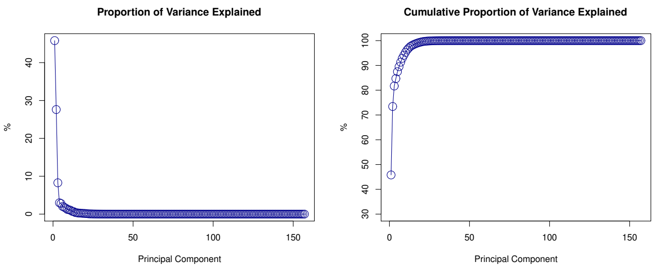


Fig. 2. Proportion of variance explained, by principal component

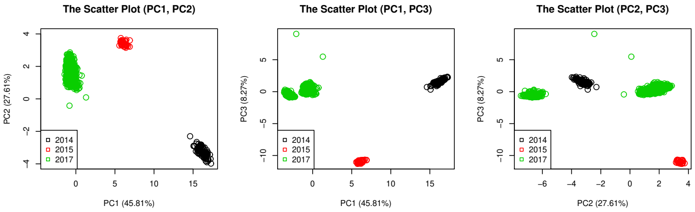


Fig. 3. PCA results (PC1, PC2, PC3)

relevant or have no variation are dropped, reducing the number of features from 230 to 140.

PCA is applied to the pre-processed data to decompose the variation of the normal data. The left plot in Fig. 2 shows variation explained by each principal component, and the right plot shows the cumulative distribution of the explained variation. The first three PCs explain about 81.68% of the total variation as shown in Fig. 3, The normal instances are clustered by year in the scatter plots of any two PCs (Fig. 3). The scatter plot of PC1 and PC2 shows that the normal instances from 2014 and 2015 are located farther from the origin and have large values of PC1 while those from 2017 are located near the origin, indicating that the first three PCs capture the variation related to the evolution of the system over time.

A review of the PCA loading plots (not shown due to space limitations) show that the features contributing to the first three PCs are related to network throughput, which is expected to improve, as the technology evolves over time. This observation explains the occurrence of the grouping patterns in Fig. 3, and, since our goal is to find PCs that capture time-invariant behavior of the normal data, these PCs are excluded. On the other hand, if we plot PC4 and PC5 which account for 2.99% and 2.79% of the variation in the normal data, respectively, the grouping patterns seen in Fig. 3 disappear, as shown in Fig. 4. In Fig. 4, all the instances are distributed close to the origin, well characterizing the boundary of normal data.

As discussed above, while PC4 and PC5 capture the system-independent characteristics of the normal data, PC1, PC2, and PC3 capture the system-dependent characteristics. Our goal is to develop a detection model for use across the system, so we use the middle principal components to characterize the behavior of normal data. To exclude the first three PCs, we set l and u as follows: $l = 0.02 \times \sum_{1\leq i\leq d} \lambda_i$, $u = 0.05 \times \sum_{1\leq i\leq d} \lambda_i$. This can be interpreted as: include the middle PCs which explain more than 2% but less than 5% of the total variation. Utilizing the Chi-square statistic in (1), we derive the following detection rule:

$$\text{Classify as an anomaly if } \sum_{l<\lambda_j<u} Y_{ij}^2/\lambda_j > \chi^2_\alpha(m)$$

where m is the number of PCs included in the model, and α is the significance level. Among 157 PCs, only 3 PCs explain more than 2% but less than 5% of the total variation resulting in $m = 3$.

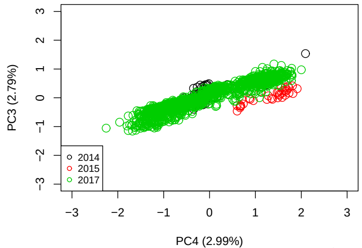


Fig. 4. PCA results (PC4, PC5)

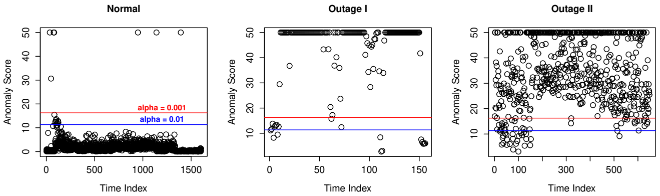


Fig. 5. Anomaly detection results

The anomaly detection results for the normal data and the two outage cases are shown in Figure 5, varying the α level from 0.01 to 0.001. The $\max(\textit{anomaly\_score}, 50)$ is plotted and the thresholds are represented as horizontal lines. Detection results are obtained by classifying the points above the lines as anomalies. It can be observed from Figure 5 that the detection results of the normal time periods are mostly normal with a few exceptions. For the two outage cases, the majority of instances have large anomaly scores being classified as anomalies, demonstrating that the two outage cases are successfully detected using the proposed anomaly detection model. Moreover, the detection model precisely identifies the timing of the anomaly in the first outage if $\alpha = 0.001$ is used. As shown in the middle plot of Figure 5, anomaly scores stay below the red line until the anomaly occurs and explode immediately after it occurs. For the second outage case, the entire data is considered to be an anomaly as observed in the third plot of Figure 5. This case is well captured by both alpha levels since the time intervals classified as anomalies are in almost all time periods. Though there are some fluctuations in the early periods, we can conclude that an anomaly is occurring since there are more anomaly time intervals than normal ones.

Acknowledgements. For a longer version of this work, including a section on root cause analysis using finite-state machine methods, the reader is referred to:
C. Kim, V. B. Mendiratta, and M. Thottan. *Unsupervised anomaly detection and root cause analysis in mobile networks.* Proceedings **IEEE COMSNETS 2020**.

Extended Abstract : Is more better? Testing feature extraction for poverty estimates from telecom data in Côte d'Ivoire *

Sveta Milusheva[a], Oscar Barriga-Cabanillas[a], Oumaima Makhlouk[b], and Ruiwen Zhang[a]

1 Introduction

Targeting the poor is an integral part of social program design in low-income countries, with widely studied benefits on improving the program's cost-effectiveness (2; 4). Geographical targeting gives priority to areas with high concentrations of poverty. However, traditional data sources, such as household surveys, lack the spatial resolution to estimate poverty at a highly disaggregated level and are costly to collect on a regular basis.

We leverage the proliferation of big data obtained from mobile devices and satellites to produce a poverty map at the commune-level for Côte d'Ivoire and provide two main contributions. First, we benchmark our results by replicating methods used across the literature, building evidence on the extent that similar prediction models and information differ in predictive power across countries and periods. Second, previous applications rely on computationally intensive methods to extract information from raw cellphone transaction data. We show how similar levels of prediction accuracy can be achieved by using key performance indicators (KPIs) that Mobile Network Operators produce regularly as part of their operations. This lowers the financial and data access barriers to estimate and update prediction models.

2 Data

Poverty Data We use a representative household survey, the 2018/19 Enquête Harmonisée sur les Conditions de Vie des Ménages (EHCVM), which contains a sample of $12,992$ households across 433 communes (out of 518 communes). We use this data to calculate four welfare measures at the commune level, which are used as the ground-truth measures in the prediction models: the Poverty Rate, a composite wealth index, Multidimensional Poverty Index (MPI), and the Intensity of Poverty (MPA).

CDR Data We use anonymized Call Detail Records (CDR) data for a representative sample of 800,000 users from April 2021 to May 2022. Individual level transaction data was processed directly by the MNO using open source libraries (5; 1) which produce a set of features on phone and data usage, mobile money transactions and top-ups at the individual level aggregated per month (referred to as 'CIDER'). Additionally, a set of indicators at the commune level were generated to capture mobility between geographic areas following (7) (referred to as 'mobility'). Finally, individual-level KPIs (referred to as 'KPIs') follow transactions on a monthly basis. There were 462 CIDER features, 90 KPI features and 11 mobility features.

Environment Data We compile satellite and geospatial indicators from Nighttime lights data, soil types, land type classification, built environment indicators, population density statistics, and zonal statistics on access to key social goods and services.

3 Methods

We aggregated individual-level indicators into commune level statistics (min, median, mean, q1, q3, max and std across users) at the monthly level and average across months. All features are normalized between [-1,1]. We test the models used in (10) and (9), which include multivariate linear regression (which has been commonly used in the literature, e.g. (6)), lasso, elastic net, ridge, and Bayesian hierarchical models. We test out using all features from the KPIs, cider and mobility library, as well as adding in the environmental variables and testing out each set of indicators independently.

4 Results

The Bayesian Hierarchical Model using all features provides the best performance with a Pearson correlation of 0.78 and R^2 of 0.58 (Table 1). Relying only on KPI and environmental indicators reduces model performance marginally with only small re-ranking of the communes (Figure 1). This reduction of performance must be weighted against the computational and access costs that more involved featurization processes involve. In line with previous research, the model performs best predicting the wealth index and is better when focused on urban areas (10).

*We thank Ezra Rwakazooba for his research assistance. Financing from the G2Px trust fund is gratefully acknowledged. a) World Bank b) Intel; Contact: smilusheva@worldbank.org; obarriga@worldbank.org

Table 1: Prediction Model Results

Features	Model	Pearson corr	RMSE	R2
All	Lasso	**0.59**	35.7	0.51
All	EN	**0.69**	33.8	0.55
KPIs, CIDER Mobility	EN	**0.64**	33.2	0.54
KPIs, Env	EN	**0.58**	39.2	0.38
KPIs	EN	**0.53**	39.4	0.34
All	BHM	**0.78**	0.32	0.58
KPIs, Env	BHM	**0.74**	0.32	0.50
KPIs	BHM	**0.72**	0.34	0.48

Notes: – All features includes KPIs, Bandicoot, Mobility and Environmental features; EN=Elastic Net; BHM = Bayesian Hierarchical Model. Results show average for the test data across 30 iterations using an 80/20 split.

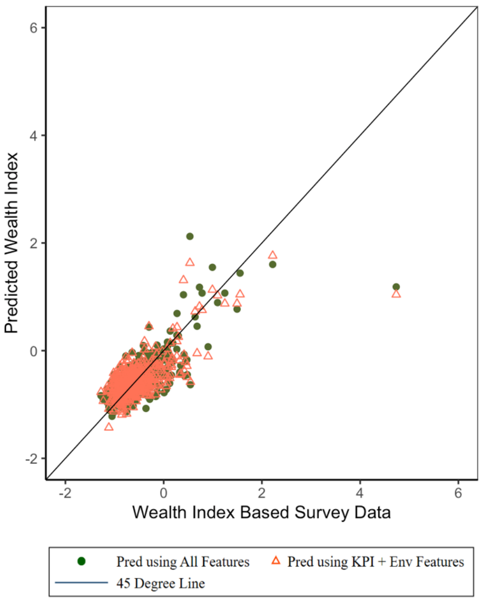


Figure 1: Comparison of Wealth Index Estimated Using Survey Data vs Predicted Wealth Index

5 Discussion

Our results have practical implications on the feasibility of producing, and updating, poverty maps on a regular basis. Previous results demonstrate that satellite and mobile phone records are strongly correlated with household welfare to the extent that they can predict area-level results from standard poverty map implementations(3; 10; 8). However, even if cellphone records have the advantage of being passively produced, the featurization of raw data into meaningful indicators requires processing large amounts of data, is computationally expensive, and involves sorting several layers of data privacy regulations. In contrast, KPI data are regularly processed at the individual level by mobile network operators. This not only lowers the barrier in terms of computational resources required, but also addresses several challenges in terms of data access by providing pre-processed indicators that are more easily anonymized and are usually stored for longer periods.

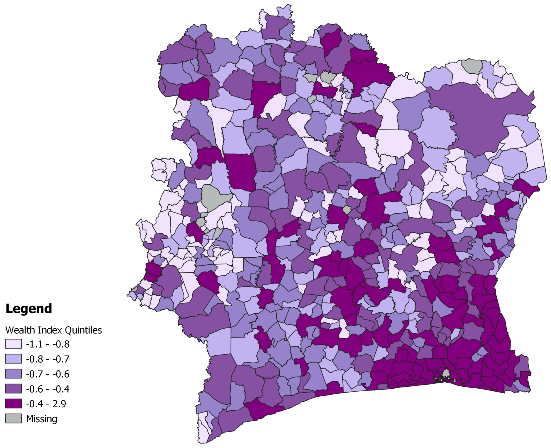


Figure 2: Map of Predicted Wealth Index at Commune Level

References

[1] Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*, 7903 (2022).

[2] Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. Targeting the poor: evidence from a field experiment in indonesia. *American Economic Review 102*, 4 (2012).

[3] Blumenstock, J., Cadamuro, G., and On, R. Predicting poverty and wealth from mobile phone metadata. *Science 350*, 6264 (2015), 1073–1076.

[4] Coady, D., Grosh, M., and Hoddinott, J. Targeting outcomes redux. *The World Bank Research Observer 19*, 1 (2004).

[5] De Montjoye, Y.-A., Rocher, L., and Pentland, A. S. bandicoot: A python toolbox for mobile phone metadata. *The Journal of Machine Learning Research 17*, 1 (2016), 6100–6104.

[6] Frias-Martinez, V., and Virseda, J. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development.*

[7] Milusheva, S., Lewin, A., Gomez, T. B., Matekenya, D., and Reid, K. Challenges and opportunities in accessing mobile phone data for covid-19 response in developing countries. *Data & Policy 3* (2021), e20.

[8] Njuguna, C., and McSharry, P. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research 70* (2017), 318–327.

[9] Pokhriyal, N., and Jacques, D. C. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences 114*, 46 (2017), E9783–E9792.

[10] Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., De Montjoye, Y.-A., Iqbal, A. M., et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface 14*, 127 (2017), 20160690.

Title: Assessing Bias in Mobile App Data Used for Daily Mobility Analysis: A Case Study of Auckland's Ethnic Groups

Olena Holubowska, KU Leuven, Belgium, olena.holubowska@kuleuven.be

Ate Poorthuis, KU Leuven, Belgium, ate.poorthuis@kuleuven.be

Mobile app data has become increasingly valuable for modeling daily mobility patterns on a city scale. This approach enables researchers to gain insights into the inequalities in daily mobility experienced by residents from different demographic groups. However, a key challenge arises from the inherent bias present in mobile app data, as certain apps might be used predominantly by specific groups of people. Consequently, relying on data from a single app may introduce a bias in the analysis and limit the generalizability of the results (Wesolowski *et al.*, 2013).

Partially to mitigate this bias, researchers have begun utilizing datasets that incorporate data from multiple sources (Huang *et al.*, 2021). Combining data from various apps might overcome bias associated with individual apps, but it is essential to recognize that different apps not only exhibit variations in user representation but also in the methods and frequency of data collection. Consequently, even when using datasets containing data from multiple apps, bias may still be present, affecting the analysis and subsequent findings.

Many existing methods aimed at addressing this bias involve comparing the number of home locations detected in the dataset to the number of residents recorded in census data (Li *et al.*, 2023). Although this approach facilitates the consideration of bias stemming from over or underrepresentation of residents within a particular neighborhood, it does not differentiate other types of bias induced by various mobile apps. As a result, additional analysis is needed to understand the impact of different mobile app biases on daily mobility modeling.

In this research, we conducted an analysis of human mobility patterns among residents of Auckland, with a focus on three ethnic groups: European, Asian, and Pacific Peoples. We used a dataset spanning six months and consisting of data from 168,213 users, who have collectively used 2156 applications that recorded their location, resulting in 70,122,575 data points. The mobile apps within the dataset were categorized into four groups: apps with a user representation proportional to the proportion of ethnic groups in Auckland as a whole, apps with an overrepresentation of residents from the Asian ethnic group, apps with an overrepresentation of residents from the European ethnic group, and apps with an overrepresentation of residents from the Pacific Peoples ethnic group. To make potential bias of apps within these groups tangible, we compare specific metrics for each group of apps against the perspective given by the dataset as whole. We investigate the following metrics: time difference between consecutive data points; distance between consecutive data points; average radius of gyration; most frequented locations; and the degree of segregation indicated by the proportion of residents from the same ethnic group in the visited areas.

Substantial dissimilarities were observed among the characteristics of different mobile applications. For instance, applications exhibiting an overrepresentation of Europeans (96 apps with a total of 319 users) exhibited an average data point frequency of 1 km and 22 minutes, while apps with an overrepresentation of Asians (68 apps with 181 users) recorded data-stamps at intervals of 41 meters and 4 minutes. This indicates the prominence of a different type of app in the latter group that captures more mobile behavior (e.g. a running app). Moreover, the most frequently visited locations varied significantly between the two groups. The average radius of gyration also demonstrated variability, with apps overrepresenting Asians exhibiting an average of 3.5 km, apps overrepresenting Europeans displaying 4.9 km, apps with balanced representation across all ethnic groups showing 4.2 km, and apps overrepresenting Pacific populations indicating 8.1 km.

Notably, the captured segregation patterns exhibited significant variations in our study. Apps with balanced representation across ethnic groups revealed that Europeans tend to spend their activity time in areas characterized by 51.4% Europeans, 22.8% Asians, and 12.8% Pacific residents. When considering the overall proportions of these ethnic groups in the entire area (49.7% Europeans, 22.7% Asians, and 14.2% Pacific residents), this trend suggests a slight preference for locations with a higher proportion of Europeans and a smaller proportion of Pacific residents, with an ever starker difference for data from apps with an overrepresentation of Europeans. Conversely, when examining apps with an overrepresentation of Pacific ethnicities, it was observed that the average locations visited by Europeans were characterized by 33.7% Europeans, 23.8% Asians, and 26.1% Pacific residents. Notably, this distribution exhibited a decrease of 15.6% in the proportion of Europeans compared to the entire area, while there was an increase of 11.9% in the proportion of Pacific residents. These findings demonstrate that conclusions regarding segregation in activity spaces can vary depending on the specific groups of apps utilized for analysis.

These findings emphasize the significance of considering the selection and composition of app data in mobility modeling studies, as it directly impacts the variability of results. In the context of mobility studies, where location data plays a pivotal role, understanding the data collection methodology becomes crucial. It should be noted that biases can be intricate, as certain apps are utilized not only during specific activities but also by specific demographic groups. While data from apps with representation of the global population can be expected to be more general and popular among a wider audience, the use of apps targeting specific groups or ethnicities can reveal distinct movement characteristics for those groups. Thus, researchers must reflect on the group composition in selecting app datasets, taking into account their research objectives and considering potential biases to enhance the interpretation and generalization of findings in mobility modeling studies.

Huang, X. *et al.* (2021) ‘The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the U.S. during the COVID-19 pandemic’, *International Journal of Digital Earth*, 14(4), pp. 424–442. Available at: https://doi.org/10.1080/17538947.2021.1886358.

Li, Z. *et al.* (2023) ‘Understanding the Bias of Mobile Location Data Across Spatial Scales and Over Time: A Comprehensive Analysis of SafeGraph Data in the United States’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.4383333.

Wesolowski, A. *et al.* (2013) ‘The impact of biases in mobile phone ownership on estimates of human mobility’, *Journal of the Royal Society, Interface*, 10(81), p. 20120986. Available at: https://doi.org/10.1098/rsif.2012.0986.

AD FINGERPRINTING: CONCEPT, VIABILITY AND SOLUTIONS

Miguel A Bermejo-Agueda[*], Universidad Carlos III Madrid
Patricia Callejo[†], Universidad Carlos III de Madrid, uc3m-Santander Big Data Institute
Rubén Cuevas[‡], Universidad Carlos III de Madrid, uc3m-Santander Big Data Institute
Ángel Cuevas[§], Universidad Carlos III de Madrid, uc3m-Santander Big Data Institute

Abstract

This paper introduces *ad fingerprinting*, a new form of fingerprinting where devices are fingerprinted from code inserted in ads. We have built *adF*, our own system to perform ad fingerprinting and run it in several ad campaigns that delivered 2,13M ad impressions. The collected data allow us to assess the vulnerability of current mobile devices to ad fingerprinting.

Web browsers have developed a set of functions and APIs that allow accessing different information about the browser, the operating system, or the device. The original goal of these functions is to optimize the functionality of browsers and the user experience. Web fingerprinting is implemented through a script embedded in the HTML code of the web page. This script leverages the different functions and APIs offered by the browser to collect different attributes from the browser configuration (e.g., browser add-ons or configured languages), the OS (installed fonts), and the device (sound device or graphic card information). In conjunction, the values of all these attributes create a fingerprint of the device. If a sufficiently large number of attributes are collected, the resulting fingerprint of the device might be unique even among a large pool of devices. Ad fingerprinting is complementary to web fingerprinting. It is a specialized form of web fingerprinting wherein a script is embedded in an ad (rather than a web page) and gathers the attributes. Ad fingerprinting tracks the interaction of individual users with a specific advertiser's ads. This is very useful for building attribution models; being an attribution model an extensively used and very valuable tool for advertisers to understand which ad campaigns, web pages, and ad channels perform better. Instead, web fingerprinting, for instance, is used to track a user's browsing history so that individual users can be profiled based on their interests.

To achieve our goals, we aim to evaluate the vulnerability of different device configurations against ad fingerprinting. In the context of this paper, a device configuration is defined by the combination of three elements: device type (mobile), operating system (e.g., Android or iOS), and browser (e.g., Chrome, Safari, or MiuiBrowser). In addition, we also consider the specific case of apps on mobile devices (Android and iOS). To this end, we have executed our system in real ad campaigns, delivering a total of 1.24M and 890k in webpages and mobile apps, respectively, across 15 countries in Europe, Africa, and America.

Based on our results and putting them into context, adding all the delivered configurations covers over 90% of the devices, representing the most common configurations in the current Web. We estimate that 47% of mobile devices can be uniquely fingerprinted with our ad fingerprinting system. However, the resilience to ad fingerprinting varies significantly across browsers and mobile apps, with Android as the most vulnerable configuration for both. Regarding browsers, (at least) 36% of mobile devices can be fingerprinted with ads delivered to browsers. As such, different device configurations offer a significantly different vulnerability to ad fingerprinting. The main takeaways from our findings suggest that users are more exposed to ad fingerprinting in mobile apps than in browsers. Moreover, iOS offers significantly better protection for ad fingerprinting than Android.

Furthermore, *adF* also serves as an auditing tool to assess the performance of the proposed anti-fingerprinting solutions by browsers and mobile app developers. The browser industry is developing different techniques to counter fingerprinting. These techniques, in essence, try to block the reporting of specific info related to browser settings, OS or hardware.

Following this principle, we propose *ShieldF* as a countermeasure to ad fingerprinting. *ShieldF* is a simple solution which blocks the reporting by browsers of those attributes that we found in the analysis of our dataset that present the most significant discrimination power. Our experiments reveal that *ShieldF* outperforms all anti-fingerprinting solutions proposed by major browsers (Chrome and Safari), offering an increase in the resilience offered to ad fingerprinting up to 55% for some device configurations. *ShieldF* is available as an add-on for any Chromium-based browser. Moreover, it is readily adoptable by browser and mobile app developers. Its widespread use would lead to a significant improvement in the protection offered by browsers and mobile apps to ad fingerprinting but also to other forms of fingerprinting.

[*] mibermej@pa.uc3m.es
[†] pcallejo@it.uc3m.es
[‡] rcuevas@it.uc3m.es
[§] acrumin@it.uc3m.es

Challenges in predicting individual poverty status from mobile operator customer segmentation metrics and phone surveys: a Papua New Guinea case study

Galina Veres*, Veronique Lefebvre, Savita Ragoonanan***, Caterina Irdi, Xavier Vollenweider, Shohei Nakamura**

Flowminder Foundation, ** The World Bank,*** Digicel Pacific

*corresponding author email: galina.veres@flowminder.org

Introduction. Papua New Guinea (PNG) faces challenges in estimating the geospatial distribution of population poverty due to data availability and quality. The last census was conducted in 2011, and the next census has been delayed until 2024. Conducting field surveys in Papua New Guinea can be challenging due to various factors such as the country's rugged terrain limiting accessibility, low population density in many areas, the high cost of transportation and limited infrastructure, and security issues in some areas. Papua New Guinea is a culturally diverse country with over 800 languages spoken, which can make it difficult to gather accurate information. Thus, data collected by Mobile Network Operators (MNOs) represent an attractive non-traditional data source to estimate poverty in PNG. In this paper, we investigate statistical dependencies between Digicel customer segmentation metrics based on mobile phone usage - which are routinely collected for each subscriber for commercial purposes, and the poverty status of subscribers as estimated by the World Bank high-frequency phone survey.

Data. Customer segmentation metrics collected by Digicel for each subscriber are handset type, monthly average revenue per user (ARPU), amount and number of top-ups per day, daily usage of mobile phone measured by number of call minutes, number of SMS and megabytes of data, centroids of commercial clusters per subscriber used as a proxy for home location, urbanicity of home location and available technology (2G, 3G, 4G). The World Bank shared the high-frequency phone survey data for five rounds conducted between December 2020 and June 2022 with approximately 5 months between rounds. For round 1, a stratified random sample was drawn from Digicel subscriber base using sample sizes proportional to 22 provinces' populations. In the next rounds, respondents were selected in two stages: 1)Contacting all respondents of previous rounds (panel cases); 2)Purposive sampling of subscribers ("replacement" cases) to reach respondents from all socio-economic groups: only respondents who did not send/received sms messages, received only incoming calls and had the majority of credit transferred were selected. Attrition rates are very high from round to round: from 52% between rounds 4 and 5 to 86% between rounds 2 and 3. Thus the resulting dataset is not a random sample of Digicel subscribers, and does not contain the full spectrum of each segmentation feature. The Word Bank also provided national estimates of wealth deciles for each respondent in each round, which defined the poverty status of respondents as poor (0.7 decile and below) and non-poor (0.8 decile and above). Linkage of the two data sources for each respondent was conducted by Digicel and only pseudonymised records on customer metrics (not location data) were shared with Flowminder.

Methodology. The framework for predicting the poverty status of respondents consisted of the following steps: feature engineering based on customer segmentation metrics; statistical analysis of segmentation features; feature selection; training, testing and comparing machine learning algorithms; analysing the results. Segmentation feature engineering was done by calculating statistics for relevant segmentation metrics and each round, such as min, max, mean, median and std. Statistical analysis of segmentation features showed complex classification problems with strongly overlapped classes, with not a single segmentation feature having a strong correlation with poverty status on its own. This could be partly due to the purposive sampling with features not representing the features full distributions. Then we removed segmentation features with strong pairwise Pearson correlation and similar means, and identified the most relevant segmentation features for predicting poverty status using ANOVA one-way test, Lasso with Cross-Validation, and Mutual information. Two scenarios were investigated: 1)To assess the variations in segmentation features for each class, and correlation between segmentation features and poverty status on a round basis, we trained a model on each round and tested predictions of respondents' poverty status on the same round, and 2) to check whether dynamic prediction of poverty was possible in PNG, we trained a model on earlier rounds and tested predictions on the last round. Several machine learning classification algorithms were trained and tested such as Logistic Regression, Random Forest, Decision Trees, Gaussian

Process Classifier, AdaBoost and Neural Networks. The results below are shown for the Logistic regression (LR) model due to similar performance to other algorithms and easy implementation on the MNO's server for operational use.

Results. Figure 1 shows the classification performance measures for individual rounds. Training and testing sets were created by proportional stratified sampling to preserve proportion of poor and non-poor in both training (75% of respondents) and testing (25% of respondents) sets. The results show the mean performance measures for 1,000 repeats and average comparison to chance. The LR model achieved precision and accuracy above 60% in all rounds except round 2. Recall (TP) is slightly above 50% for rounds 2 and 3 and above 75% in rounds 4 and 5. Comparison with chance shows improvements by between 19% and 40% for poor, and between 10% and 34% for non-poor. However, the later rounds are the least representative of the Digicel subscriber base due to the purposive sampling of replacement respondents using features also used for prediction, which may inflate accuracy and precision statistics. The classification power of the features is probably closer to round 2 with less targeted sample, though only approximately half of round 2 respondents were selected randomly.

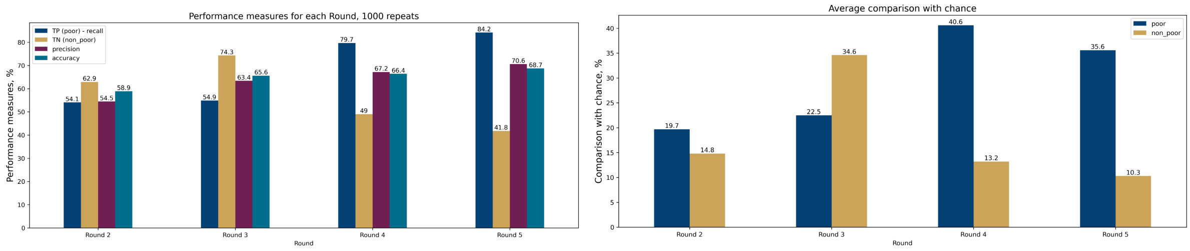


Figure 1. Performance measure for each Round (left) and average comparison to chance (right).

Figure 2 shows classification results on the second scenario: training on the previous rounds 2, 3 and 4, and testing on round 5. Precision (~74%) and accuracy (~67%) is very similar when training and testing on round 5 only. Recall (~70%) is lower, however the trade-off is an improvement in correctly classifying non-poor (~61%). Comparison with chance shows improvements by ~61% for non-poor and by ~13.5% for poor. These results show potential for predicting poverty status based on the past data for the respondents participating in the World Bank high-frequency survey (bearing the above caveat on sampling).

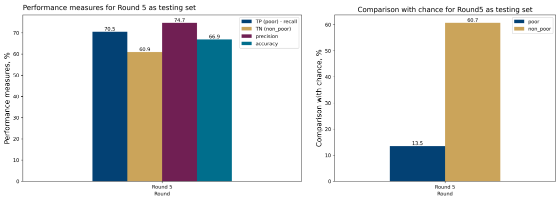


Figure 2. Performance measure for the second scenario (left) and average comparison to chance (right).

Challenges and recommendations.Though customer segmentation metrics or other MNO data appear related to poverty, we have identified the following challenges and recommendations to develop predictive models from these data: 1)The high-frequency phone survey was conducted using purposive sampling for some rounds, thus no inference is possible neither to Digicel subscriber base nor to a general population, i.e. random sampling needed. 2)The purposive sampling was based on some of the segmentation features the project wanted to evaluate, which prevents the possibility of verifying assumptions on statistical relationships between these features (SMS, credit) and poverty, i.e. random sampling needed. 3)Inference to the general population requires data on poverty of phone users and non-users, and subscribers of different networks, i.e. field survey data required. 4)Customer segmentation metrics are more challenging for poverty prediction compared to other MNO data such as mobility, social networks and mobile banking, i.e. integrating the promising features from MNO would improve models, so would ancillary data such earth observation data.

Acknowledgements. The project was financed by the Australian Government through the World Bank. This work was done with a cooperation from Digicel Pacific PNG.

Dynamic Road Network Criticality Computation using Call Detail Records for Enhancing Healthcare Accessibility

M. Nunez-del-Prado[1], V. Gauthier[2], G. Roujanski[2],
H. Alatrista-Salas[3], M. Tariverdi[1]

[1] The World Bank {mnunezdelpradoco,mtariverdi}@worldbank.org
[2] SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris
{vincent.gauthier, gatien.roujansk}@telecom-sudparis.eu
[3] Pontificia Universidad Católica del Perú {halatrista}@pucp.pe

Nowadays, the states play a crucial role by providing basic public services such as education, justice, and healthcare, among others, to the population. This is particularly true for health systems, where states must promote infrastructure development to ensure equitable access to health services for the entire population.

The present work focuses on accessibility to healthcare system services without losing generality. Accordingly, access to the healthcare system relies on the road network infrastructure, which enables the population to reach health services. Consequently, different works in the literature studied the criticality of the road network to rank the most critical links in the road network [1, 2, 3, 4]. The idea behind the criticality metric is to capture the importance of a road segment, which inhabitants use to reach the health system. Thus, criticality depends on the number of inhabitants in a given region. Nonetheless, the before mentioned works are based on static population data, such as census information.

In the current effort, we extend the work of Tariverdi et al. [5] to introduce a dynamic count of inhabitants to compute criticality metrics varying over time. Thus, the idea is to capture the importance of road segments as an effect of the change in the number of people in a given region as an effect of the change in time. For instance, dense residential populated zones decrease the number of inhabitants during workdays because people go to other regions for work. Therefore some road segments near residential areas have less criticality during weekday work hours. On the contrary, road segments close to business areas increase their criticality simultaneously. The dynamic criticality metric allows a more comprehensive view of the crucial and vulnerable segments [6] to reach the healthcare system at different hours of the day. This metric will enable policymakers to prioritize multi-sectorial investments better to maintain critical road segments for enhancing healthcare service accessibility. This metric is important since people living far from healthcare facilities sometimes prefer not to go to the health system when the accessibility is not easy, which could have fatal effects on people's health [7, 8, 9].

The dynamic estimation of the population through the Call Detail Records (CDR) has been studied in the literature [10]. In this effort, we enhance the model [10] by using a VoronoiBoost [11] to model an antenna's more realistic coverage area. In addition, we derive the population presents a

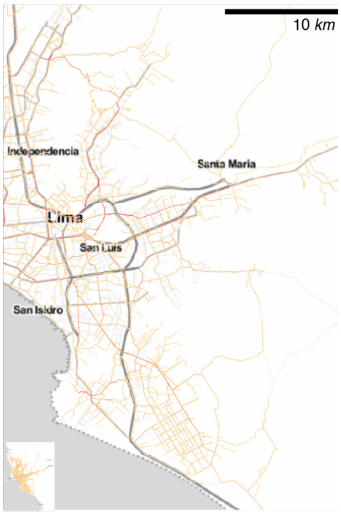


Figure 1: Criticality distribution for Lima city.

given time into the hexagonal hierarchical geospatial indexing system (H3)[1] at a resolution of nine to represent smaller areas inside the antennas. The idea is to represent the inhabitants' presence more accurately since antenna coverage could be huge, especially in rural areas. Finally, the number of connected cellphones to a given antenna is distributed uniformly to each H3 hexagon.

Using the generated population count based on CDR activity, the graph $G(V, E)$ representing the road network from Openstreetmaps (OSM), and healthcare facilities location, we can estimate the probability of visiting a healthcare facility from a given node i using the Huff model as shown in Eq. 1.

$$P_{i,j} = \frac{\frac{A_j}{t_{i,j}}}{\sum_i^V \frac{A_j}{t_{i,j}}} \tag{1}$$

$$A_j = 2 * \exp(-min(distance, 7km) * \log(2)/7)) - 1 \tag{2}$$

Where A is the attractivity representing the complexity of the healthcare facility. Therefore, using these probabilities, we compute the road network criticality.

The criticality of the road segments regarding a single service is built based on population density and preference-weighted edge betweenness centrality. To compute criticality, we consider the population at node origin w_i weighted by the probability $P_{i,j}$ to visit an amenity multiplied by an important factor α as shown in Eq. 3.

$$C_{v_i} = \alpha \cdot w_s \cdot P_{S_i,j} \tag{3}$$

The important factor is a list of public services' relevance provided by policymakers. Once the criticality for the node C_i is computed, it is imputed to all edges E_{ij} composing the trip from origin i to j.

For visualizing the important segments in the road network, we build a heatmap ranging from yellow (least critical) to red (most critical), as illustrated in Fig. 1. In this map, the weights represent the number of times individuals passed through a road segment to reach a healthcare facility.

Preliminary results show the most critical axis to reach healthcare facilities is the *República de Panama Ave.*, the main highway of Lima city, connecting the city from north to south.

[1] Hexagonal hierarchical geospatial indexing system: `https://h3geo.org/`

For future steps, we will compare the evolution of the road criticality based on dynamic population count from CDR information over the different hours of workdays in Lima.

References

[1] G Petri, P Expert, HJ Jensen, and JW Polak. "Entangled communities and spatial synchronization lead to criticality in urban traffic". In: *Scientific reports* 3.1 (2013), pp. 1–8.

[2] H Hamedmoghadam, M Jalili, HL Vu, and L Stone. "Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks". In: *Nature communications* 12.1 (2021), pp. 1–10.

[3] W Wang, S Yang, HE Stanley, and J Gao. "Local floods induce large-scale abrupt failures of road networks". In: *Nature communications* 10.1 (2019), pp. 1–11.

[4] S Loreti, E Ser-Giacomi, A Zischg, et al. "Local impacts on road networks and access to critical locations during extreme floods". In: *Scientific Reports* 12.1 (2022), pp. 1–15.

[5] M Tariverdi, M Nunez-Del-Prado, N Leonova, and J Rentschler. "Measuring accessibility to public services and infrastructure criticality for disasters risk management". In: *Scientific reports* 13.1 (2023), p. 1569.

[6] E Koks, J Rozenberg, M Tariverdi, et al. "A global assessment of national road network vulnerability". In: *Environmental Research: Infrastructure and Sustainability* 3.2 (2023), p. 025008.

[7] KE Battle, D Bisanzio, HS Gibson, et al. "Treatment-seeking rates in malaria endemic countries". In: *Malaria journal* 15 (2016), pp. 1–11.

[8] VA Alegana, J Wright, C Pezzulo, et al. "Treatment-seeking behaviour in low-and middle-income countries estimated using a Bayesian model". In: *BMC medical research methodology* 17.1 (2017), pp. 1–12.

[9] R Manongi, F Mtei, G Mtove, et al. "Inpatient child mortality by travel time to hospital in a rural area of T anzania". In: *Tropical medicine & international health* 19.5 (2014), pp. 555–562.

[10] G Khodabandelou, V Gauthier, M Fiore, and MA El-Yacoubi. "Estimation of static and dynamic urban populations with mobile network metadata". In: *IEEE Transactions on Mobile Computing* 18.9 (2018), pp. 2034–2047.

[11] OE Martínez-Durive, T Couturieux, C Ziemlicki, and M Fiore. "VoronoiBoost: Data-driven Probabilistic Spatial Mapping of Mobile Network Metadata". In: *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. 2022, pp. 100–108. DOI: 10.1109/SECON55815.2022.9918610.

Mind the Gap: Studying the Impact of Covid-19 on Cellular Users Presence in Inland Areas

Andrea Pimpinella[1], Carmelo Ignaccolo[2], Cristina Boniotti[3]
[1] *Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano*
[2] *Department of Urban Studies and Planning (DUSP), Massachusetts Institute of Technology*
[3] *Department of Architecture, Built Environment and Construction Engineering (DABC), Politecnico di Milano*
email: {name.surname}@polimi.it[1,3], carmeloi@mit.edu[2]

I. Introduction

It is widely known that there is a close relationship between the habits of people and the characteristics of the environment where such people live. Also, due to the ubiquity of mobile radio access (mobile subscriptions are expected to overcome 9 billions in 2028 [1]), the study of cellular users communication activity can reveal important insights about social, topological, and technological phenomena at large scales [2].

The goal of grouping mobile radio access sites (e.g., eNodeBs) according to the spatial and temporal characteristics of their network activity is typically pursued through the design of *clustering* algorithms. A common choice is to group network sites according to the dynamics of the served traffic [3]: this option offers several advantages to service providers, who put effort to discover regularities of traffic loads and target pro-activity in network resources management.

Recently, the huge potential of mobile data has been exploited to examine the impact of Covid-19 pandemic on people's mobility and social life[1]. In this context, many studies in literature observe a shift in the presence of people from dense urban cities to inner areas, apparently inverting the trend of depopulation of rural areas that has been observed since decades in many countries worldwide [4], [5].

Using a real-world cellular network dataset, this study focuses on the effects of Covid-19 pandemic on cellular users presence in the Italian region of Valtellina, which consists of small, rural villages progressively shrinking due to a long-term depopulation phenomenon. By clustering network sites based on the changes (i.e., gaps) in the relative share of connected users after the pandemic, we observe that: i) users' behavior in Valtellina has changed and ii) such change is spatially heterogeneous, indicating a modification in the attractiveness of urban settlements according to post-pandemic needs.

II. Data and Methodology

This work leverages a dataset coming from the LTE network of a popular European mobile operator, containing radio access network measurements collected at 61 eNodeBs in the region of Valtellina, in the form of hourly sampled time series. We consider here the number of users that are Radio Resource Control (RRC) connected to each cell site as a proxy of people presence in the valley. We focus on two 1-month periods, T_1={20/01/2020, 16/02/2020} (i.e., before pandemic breakout in Italy[2]) and T_2={24/01/2022, 20/02/2022} (i.e., when national government removed most of Covid-19 restrictions).

Operatively, we rely on k-means clustering to group together eNodeBs based on the Euclidean distance among their *gap* signals. In detail, the gap signal $\mathbf{g}_i$ of the i-th eNodeB is generated as it follows:

1) Compute the Median Weekly Signature (MWS) of the number of connected users in T_1 and T_2 (computational details can be found here [3]), namely $\mathbf{m}_{(1,i)}$ and $\mathbf{m}_{(2,i)}$.
2) Normalize each hour of both signatures to the hourly sum of connected users, referring to them as $\mathbf{m}^n_{(1,i)}$ and $\mathbf{m}^n_{(2,i)}$.
3) Compute $\mathbf{g}_i$ as the difference between $\mathbf{m}^n_{(1,i)}$ and $\mathbf{m}^n_{(2,i)}$.

Repeating this process for each eNodeB in the network allows to perform k-means clustering with the gap signals $\mathbf{g}_i$ as input, setting $k = 4$ through a data-driven approach [3]. We underline that our approach takes into account both the spatial and temporal dimensions of the problem. On the one hand, we consider the share of the total number of cellular users connected in Valtellina taken every hour by each site through the normalization process as described above (step 2), i.e., we embed spatial domain information in the clustering objects. On the other hand, the algorithm clusters $\mathbf{g}_i$ according to its temporal dynamics, thus using the time domain information of the signals. Note also that each input is normalised to mean and variance in the time domain before being used, as our interest is to group eNodeBs regardless of the amplitude of the users' presence variation.

III. Experimental Results

We plot in Figure 1 the centroids of the recognised clusters, each one representing the gap signals of all eNodebs in the corresponding cluster. For each centroid, positive (negative) gaps represent the case where the variation of the number of users served from T_1 to T_2 at that hour is greater (smaller) than the centroid's average weekly gap, that is represented by the black horizontal dashed line.

Moving downwards in Figure 1 from top to bottom plot, we interpret the centroid profiles as it follows:

1) The red cluster groups 20 eNodeBs (33% of the total), and is the largest of our configuration. Its centroid has

[1] Examples can be found at these web sites: i) https://senseable.mit.edu/stockholm-19/, ii) https://wfh-inequalities.netlify.app

[2] First recognized Covid-19 case in Italy dates to 20/02/2020.

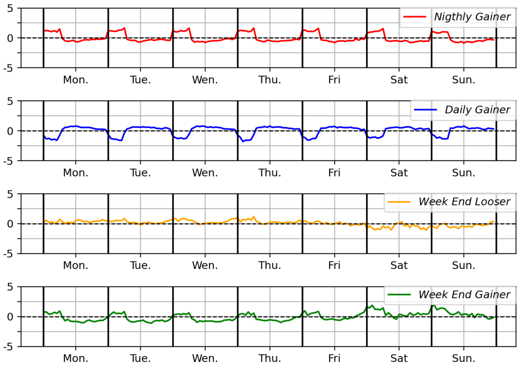


Fig. 1. Weekly profiles of centroids representative of the recognised clusters.

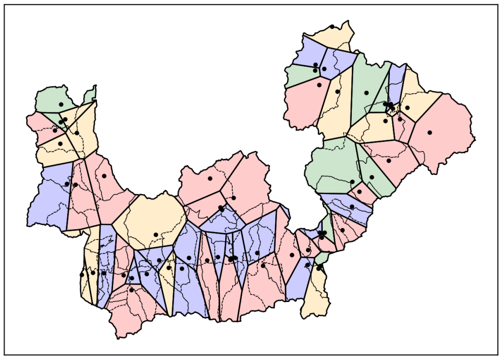


Fig. 2. Geographical map of Valtellina. Voronoi polygons are color-coded according to the cluster the corresponding eNodeB belongs to.

high gaps during the night (from 0:00 a.m. to 7:00 a.m.), while they are slightly below the average during the day, for both week days and week ends. We therefore label the eNodeBs of this cluster as *Nightly Gainers*.

2) The blue centroid repesents 19 eNodeBs (31%), which we name as *Daily Gainers*. In fact, we observe gaps greater than the weekly average from 9:00 a.m. to 11:00 p.m in week days (up to 8:00 p.m. in the week end), while they lay below the average during the night.
3) The orange centroid, representing 13 out of 61 eNodeBs (21%), has a profile that is stable around the average during week days. Differently, gaps turn negative during the week end: this means that the eNodeBs of this cluster have more likely lost users during the week end in T_2 as compared to T_1 than what observed during the week days. So, we label such eNodebs as *Week End Loosers*.
4) The green cluster contains only 9 out of 61 eNodeBs (14%), and is the smallest of the four clusters. As one can see, its centroid has a week day profile similar to the one of the red cluster but with a positive upward trend. In fact, gaps are positive during the week end: we thus name the eNodeBs in this cluster as *Week End Gainers*.

TABLE I
PER-CLUSTER DISTRIBUTION OF ENODEBS AND CUMULATIVE SHARE OF CONNECTED USERS DURING T_2.

Cluster Label	Color	Cluster Members (%)	Served Users (T_2) (%)
Nightly Gainer	Red	32.79	30.73
Daily Gainers	Blue	31.15	42.92
Week-End Gainer	Orange	21.31	14.24
Week-End Looser	Green	14.75	12.12

We summarise in Table III the percentage of eNodeBs of the network grouped in each cluster and the corresponding fraction of the overall number of connected users that were served during T_2. To understand the spatial distribution of the clusters, we represent in Figure 2 a colour-coded map of Valtellina, where each color corresponds to the mentioned clusters. In the map, the black markers pinpoint the eNodeBs location, while the thicker black boundaries outline the radio coverage area of each eNodeB, which for simplicity is here assumed to coincide with the corresponding Voronoi polygon[3]. Also, thinner black lines outline the administrative boundaries of the municipalities in the valley. As one can see, most of the Daily Gainers (blue polygons) are located in the middle Valtellina. In fact, municipalities in this area have recently upgraded their digital infrastructures thanks to recent public investments[4]: we believe communications technology upgrades greatly benefit users enjoyement, especially due to the enabling of remote working options. Additionally, we recall that blueish eNodeBs gain users also during the week end: this can be the results of the many efforts local public entities devoted to foster tourism in these areas. Besides, lateral valleys mostly host Nightly Gainers (red polygons), probably due to the scarcity of digital infrastructures and services as well as job opportunities, which induce citizens to move outwards during daily hours and return home during the night. Finally, we mention that the municipality of Grosio (whose administrative boundaries mostly overlap the green Voronoi polygon located eastwards close to Swiss border) is a Week End Gainer: it became in 2020 a neutral delivery point for optical cables distribution to surrounding districts[5]. Future works will regard a more specific analysis on the relationship between these findings and the characteristics of the built environment in the valley.

REFERENCES

[1] "Ericsson mobility report," https://www.ericsson.com/en/reports-and-papers/mobility-report, 2022.

[2] R. Singh *et al.*, "Urban vibes and rural charms: Analysis of geographic diversity in mobile service usage at national scale," in *The W.W.W. Conf.*, 2019, pp. 1724–1734.

[3] A. Pimpinella *et al.*, "Forecasting busy-hour downlink traffic in cellular networks," in *ICC 2022 - IEEE Int. Conf. on Communications*, 2022, pp. 4336–4341.

[4] G. Lanza *et al.*, "Impacts of the covid-19 pandemic in inner areas. remote work and near-home tourism through mobile phone data in piacenza apennine," *TEMA*, vol. 2, pp. 73–89, 2022.

[5] E. Willberg *et al.*, "Escaping from cities during the covid-19 crisis: Using mobile phone data to trace mobility in finland," *ISPRS Int. Journal of Geo-Inf.*, vol. 10, no. 2, p. 103, 2021.

[6] O. E. Martínez-Durive *et al.*, "Voronoiboost: Data-driven probabilistic spatial mapping of mobile network metadata," in *2022 19th Annual IEEE Int. Conf. on Sensing, Communication, and Networking (SECON)*. IEEE, 2022, pp. 100–108.

[3] How to represent the coverage area of a radio access point without the help of ad-hoc coverage maps is still a topic of literature debate [6].

[4] Ultra-Broadband National Plan (2015) and National Recovery and Resiliency Plan (2021).

[5] https://bandaultralarga.italia.it/en/map/

Impact of Mobility Patterns on Federated Learning applied to Human Mobility Prediction

João Paulo Esper*, Aline Carneiro Viana†, Jussara M. Almeida*
*Universidade Federal de Minas Gerais, Brazil. † INRIA, France.
E-mail: {joaopauloesper, jussara}@dcc.ufmg.br*, aline.viana@inria.fr†

Abstract—**The Federated Learning (FL) framework has been applied in multiple domains, offering solutions that provide both accuracy and data privacy protection. Yet, specifically for human mobility prediction, prior solutions have been analyzed on mobility datasets that are spatially and temporally sparse, AND neglected the impact of the *heterogeneity* of users' mobility patterns. Heterogeneity on the (fine-grained) spatial and temporal mobility patterns directly impact prediction, hardening the FL performance analysis. As such, prior evaluations of FL on mobility prediction are limited and may overestimate the robustness of the proposed solutions. We here aim to fill this gap by analyzing the impact that different mobility patterns (e.g., repetitive and/or exploratory patterns) have on the performance of FL-based human mobility prediction models, in terms of both model effectiveness and efficiency.**

Index Terms—**Mobility prediction, federated learning, privacy.**

I. Introduction

Human mobility prediction drives the solution to a diverse range of complex problems, including resource allocation, traffic management, and epidemic prevention, to name a few [1]. Human mobility is typically represented as a time series describing an ordered temporal sequence of visited locations by a user on a daily basis, i.e., a circadian trajectory. The *prediction task considered in this paper is the inference of the next location to be visited by a user at the next timestamp*.

Recently proposed human mobility prediction models often exploit machine learning techniques, notably deep learning models [2], [3], to deliver state-of-the-art prediction accuracy. Yet, such approaches present concerning privacy issues that are rooted in their *centralized* architecture, which requires the uploading of sensitive data (e.g., individuals' visited locations) to a server where the prediction model training occurs.

Towards meeting users' increasing demands for privacy guarantees in various domains (notably mobility prediction), the decentralized Federated Learning (FL) framework was proposed [4], [5]. In FL, a shared model is first *decentralized* trained on multiple devices, using only local data. That is, private location data is only stored and analyzed locally on the corresponding device and used to train a local prediction model. The local prediction models are then uploaded to a central server. Then, a global model is generated by the server, that first aggregates the received local parameters, then chooses candidate devices and finally distributes the updated global model. The previous steps are repeated until convergence, and an optimal model for mobility prediction is generated [4]. Note that only the locally trained models, i.e., model parameters (e.g., weights), naturally less sensitive than users' raw location data, are transferred to the server. As such, FL offers the possibility of producing (accurate) predictions while still restricting access to users' sensitive data.

FL has been applied to various problems, from image classification to next word prediction [6], [7]. For mobility prediction specifically, some prior efforts adapted solutions from other domains to mobility problems. For example, in [6], an image classification model was used for transportation mode prediction by converting coordinates into pixels. Yet, directly adapting solutions from other domains can be both challenging and inefficient due to the spatial and temporal related specificities of mobility prediction. Indeed, visits to certain point of interests are directly correlated to the users' routines and preferences. These are hard to embed in models for *next word* prediction or image classification, which are more concerned with grammatical structures of a language and recognition of patterns on a static low-dimensional space.

There are only a few FL solutions that were designed for mobility prediction [4], [5]. Yet, prior analyses of them neglected the impact that the naturally heterogeneous human patterns may have on FL effectiveness. Also, the usage of social network datasets, that are both sparse in space and coarse in time, challenges routine and mobility patterns characterization.

In this work, we aim to fill this gap by analyzing the performance of alternative FL-based mobility predictions in scenarios with users with varying mobility patterns, identified in real and less sparse human mobility data. We aim to answer the following question: *How do existing FL-based mobility prediction models perform for users with very different mobility patterns, such as very repetitive behavior (e.g., routines) or more exploratory visiting patterns (e.g., tourists)?* Our analyses comprise both model effectiveness (accuracy) and efficiency (resource usage and execution time), and offer insights into possible improvements to current FL solutions.

II. Evaluation

A. Methodology: Models and dataset

After searching for FL models designed for mobility prediction in the literature, we settled with two different models introduced in [5], namely *GRU-Spatial* and *Flashback*, as these were the only ones with publicly accessible code which we were able to run with no execution errors. *GRU-Spatial* uses Gated Recurrent Units (GRUs) which, given a sequence, learn which data is relevant keeping or not. GRU was proposed as an evolution from Recurrent Neural Networks (RNNs), which had their learning process impacted by very large/small gradient values, an issue tackled by the gating mechanism in GRU. *Flashback*, in turn, was first proposed as a centralized model [3] and later adapted to a federated setup [5]. It is an RNN-based prediction model proposed for sparse mobility data,

Table I: Federated results considering 5 clients and samples with 3,000 users. Resource measured in vRAM consumption.

	Acc@1 Flashback	**Acc@1 GRU-Spt.**	**Acc@5 Flashback**	**Acc@5 GRU-Spt.**	**Resource Flashback**	**Resource GRU-Spt.**	**Time Flashback**	**Time GRU-Spt.**
Scouters	0.111 ± 0.02	0.301 ± 0.02	0.270 ± 0.02	0.538 ± 0.02	1,662 MB	1,746 MB	26 min 17 sec	14 min 41 sec
Regulars	0.160 ± 0.01	0.395 ± 0.02	0.322 ± 0.03	0.618 ± 0.01	1,604 MB	1,686 MB	21 min 41 sec	11 min 45 sec
Routiners	0.213 ± 0.02	0.500 ± 0.02	0.384 ± 0.02	0.721 ± 0.01	1,510 MB	1,538 MB	19 min 10 sec	10 min 18 sec
Mixed	0.169 ± 0.01	0.397 ± 0.05	0.349 ± 0.02	0.635 ± 0.04	1,622 MB	1,706 MB	21 min 31 sec	11 min 33 sec

Table II: User profiles and metrics of their trajectories.

	Trajectory length	**Stationarity**	**Diversity**
Scouters	420.66 ± 4.11	0.391 ± 0.01	0.840 ± 0.00
Regulars	345.21 ± 1.72	0.566 ± 0.01	0.598 ± 0.00
Routiners	320.53 ± 2.27	0.668 ± 0.01	0.404 ± 0.01

which explores spatial-temporal contexts, searching historical hidden states with equivalent context to improve its prediction.

Our study uses a fully anonymized *Call Detail Record* (CDR) dataset gathered by a major telecom operator in a metropolitan area in China. It contains records of 58,502 users spanning over a two week period. Unlike real raw CDRs, the recorded location is not the cell tower's coordinates the user was attached to, but rather the centroid of a 200 m^2 cell in a grid representation nearest to the tower the user stayed mostly attached for each hour. No user identity or the corresponding operator's cell tower infrastructure is provided. To deal with missing records and homogenize the number of records per user, the data was handled as in [1], resulting in 41,460 users.

To characterize the users in our dataset, we followed the clustering mechanism proposed in [1] to group them into three mobility profiles based on their trajectories: (i) *Scouters* are users more prone to explore and discover new areas; (ii) *Regulars* are users who constantly alternate between explorations and revisits; and (iii) *Routiners* are users who rarely explore and prefer to stick to their known places. We also used two metrics, *Stationarity* and *Diversity* [8], to analyze each user trajectory. *Stationarity* measures the number of records for which the user stays continuously in the same place. *Diversity* quantifies the number of distinct segments of trajectories given a time-ordered sequence of locations visited by a user.

B. Results

We started by clustering the 41,460 users in our dataset into the three aforementioned mobility profiles and characterizing their trajectories. We found 5,743 (14%) users characterized as *Scouters*, 25,032 (60%) users as *Regulars*, and 10,685 (26%) users as *Routiners*. Table II presents a broad characterization of the trajectories of users in each cluster, including average trajectory length, *Stationarity* and *Diversity*. *Scouters* tend to have much longer and diverse trajectories, while more data may favor model training, greater *Diversity* also makes pattern learning more challenging, motivating the study of model effectiveness for different profiles.

We have carried out a set of experiments evaluating effectiveness and efficiency of both *Flashback* and *GRU-Spatial* for different subsets of users in various scenarios. For illustration purposes, Table I presents some of our results for a sample of 3,000 users from each profile as well as a sample of 3,000 users with mixed profiles, keeping the fractions of users of each profile the same as in the original dataset. The table shows results for both model effectiveness (accuracy) and efficiency (resource usage and total execution time). The users were federated into 5 clients, each client containing data from 600 users, which is a setup comparable to the one used in [5].

III. Discussions and future work

Considering the proposed scenarios, we noticed the considerable impact that different mobility patterns can have on both the effectiveness and efficiency of the FL models. Easier to predict users (i.e., *Regulars* and *Routiners*) experience great improvements on the accuracy of both models, accelerated the learning process and reduced resource consumption. *Scouters*, in turn, really challenged both models on every aspect, specially accuracy. Indeed, even a small fraction of *Scouters* (14%) greatly impacted both models in the *Mixed* scenario, evidencing the impact that heterogeneity has on the solutions.

These initial results offer many future directions of exploration. For example, we aim to study strategies to explicitly incorporate into model training the different properties of the mobility profiles, as well as investigate the performance trade-offs of favoring one particular profile over the others. Studying how the FL solutions can adapt to fluctuations in mobility patterns (and even profiles) over time is also worth pursuing. Finally, prior work has discussed the reconstruction of trajectories in an FL environment by inferring patterns given the model weights [4]. Analyzing how such security issues may correlate with the mobility patterns and how they impact model effectiveness is also an interesting avenue to pursue.

Acknowledgements

This work was supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and the CAPES-STIC-AMSUD 22-STIC-07 LINT project.

References

[1] L. Amichi *et al.*, "Understanding individuals' proclivity for novelty seeking," in *Proc. of the 28th International Conference on Advances in Geographic Information Systems*, 2020.

[2] J. Feng *et al.*, "Deepmove: Predicting human mobility with attentional recurrent networks," in *Proc. of the world wide web conference*, 2018.

[3] D. Yang *et al.*, "Location prediction over sparse user mobility traces using RNNs," in *Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[4] J. Feng *et al.*, "PMF: A privacy-preserving human mobility prediction framework via federated learning," *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, 2020.

[5] C. E. J. Ezequiel *et al.*, "Federated Learning for Privacy-Aware Human Mobility Modeling," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[6] F. Yu *et al.*, "Privacy-preserving federated learning for transportation mode prediction based on personal mobility data," *High-Confidence Computing*, vol. 2, no. 4, 2022.

[7] T. Yang *et al.*, "Applied federated learning: Improving google keyboard query suggestions," *arXiv preprint arXiv:1812.02903*, 2018.

[8] D. do Couto Teixeira *et al.*, "On estimating the predictability of human mobility: the role of routine," *EPJ Data Science*, vol. 10, no. 1, 2021.

Characterising Intersectional Mobility Patterns Using Multiple Correspondence Analysis and Network Entropy

Arthur Vandervoort[1,*], Karyn Morrissey[2], Riccardo Di Clemente[3,4], and Sabina Leonelli[5]

[1]UKRI CDT in Environmental Intelligence, University of Exeter, Exeter, GBR.
[2]Department of Technology, Management and Economics, Technical University of Denmark, Kongens Lyngby, DEN.
[3]Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, GBR.
[4]The Alan Turing Institute, London, NW12DB, GBR.
[5]Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter, Exeter, GBR.
[*]a.vandervoort@exeter.ac.uk

A person's mobility – or lack thereof – is a key constitutive factor in their ability and propensity to access amenities, employment, and social networks among nearly every other facet of social, modern life[1]. The intersection of mobility and gender has been studied for decades, leading to the identification of specifically gendered patterns with regard to trip chaining, transport mode choice, and commute length, among others[2–4]. Importantly, these differences reflect the trade-offs that women have to make in terms of their mobility compared to men. For instance, feminist scholars of mobility point to the patriarchal norms around care responsibilities as an important constraint to women's propensity to move; or their *motility*[5]. Gender as a category operates on an intersectional basis; whereby social categories intersect to produce different outcomes for different population segments, and it is therefore important to be careful not to treat the effects of social categories like gender as exclusive and separable[6].

It is within this context that we produce an account of inter- and intra-household mobility differences that incorporates an intersectional understanding of how mobility and gender operate in São Paulo. Feminist modellers that aim to operationalise intersectionality in quantitative research have identified Multiple Correspondence Analysis (MCA) as one potential technique that addresses these aforementioned issues of inexclusivity and inseparability of social dynamics[7,8]. This analysis is based on the commuting patterns of 5332 households (aged 26-65) within the city of São Paulo, as well as socio-demographic information drawn from the 2017 São Paulo O/D survey. The results are spread over a time of 1 year and 9 months, and capture the weekly mobility patterns of households. The average household reports six trips, spread over three locations.

We test the extent to which network metrics – in this case the Shannon entropy of a given person's adjacency matrix – can help characterise inter- and intra-household differences in mobility patterns by comparing two MCAs: one typical MCA modelled after Manderscheid's implementation of the method, and another which includes household entropy. To do this we take origin-destination data for all reported mobility within the São Paulo metropolitan area (fig. 1A), subset by household, and calculate the average Shannon entropy per household member, by gender (fig. 1B). We then compare factor maps for a typical MCA (fig. 1C), and for an MCA which includes entropy (fig. 1D). We define a weighted, directed mobility network by aggregating all trips taken by an individual, where the link weights are defined by the amount of times a trip is undertaken by any member of a given individual. Nodes are defined as "mobility zones", which are defined by the survey and of which there are 342 in São Paulo. Based on these parameters, we generate an adjacency matrix for each household, with which we calculate the Shannon entropy of the edges and edge weights.

Our findings indicate that the inclusion of entropy helps characterise the trade-offs women have to make with regards to their mobility, by showing the interrelation between the presence of children in a household (*famSit* on plots 1C and 1D), women's commuting times, socio-economic status, and entropy, with the inclusion of entropy accounting for a great deal of variance in terms of household structure.

References

1. Lenormand, M. *et al.* Influence of sociodemographic characteristics on human mobility. *Sci. Reports* **5**, 10075, DOI: 10.1038/srep10075 (2015).

2. Gauvin, L. *et al.* Gender gaps in urban (2020).

3. Gordon, P., Kumar, A. & Richardson, H. W. Gender differences in metropolitan travel behaviour. *Reg. Stud.* **23**, 499–510, DOI: 10.1080/00343408912331345672 (1989).

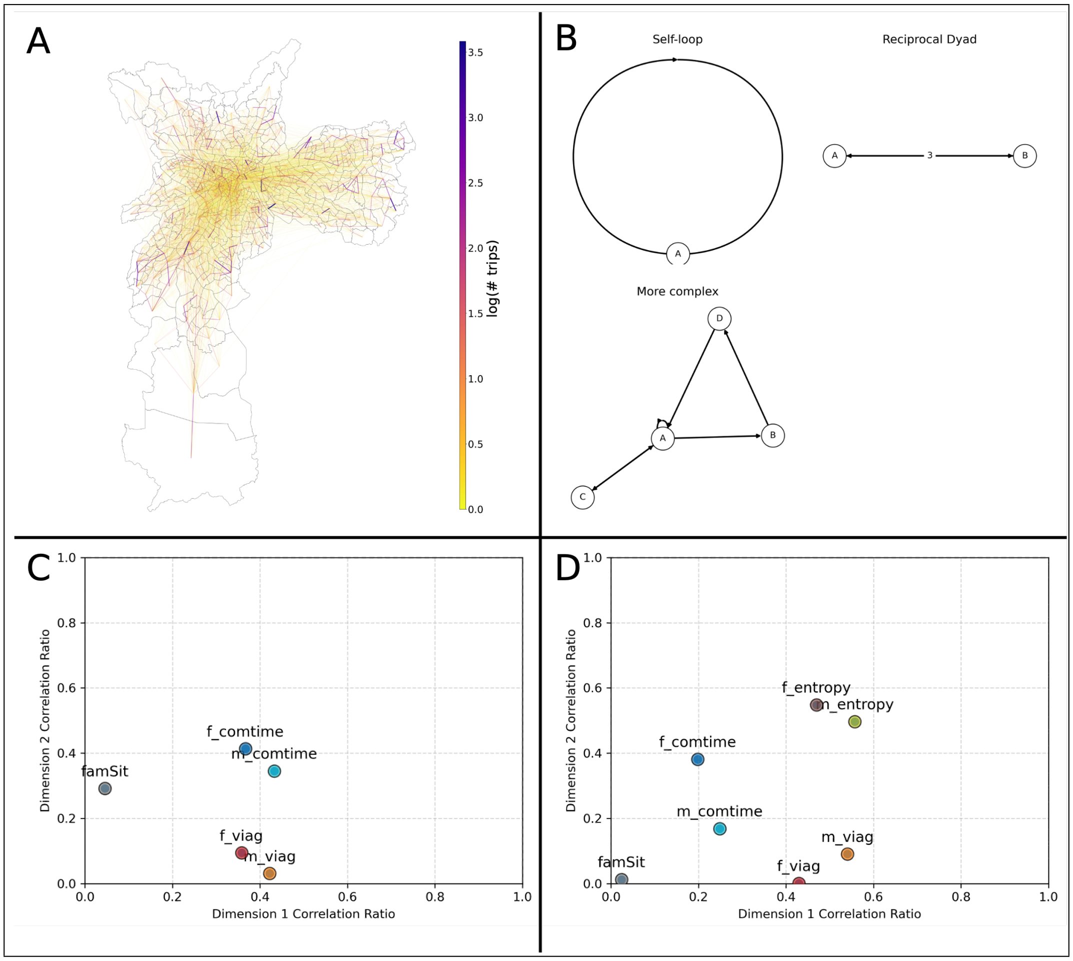


Figure 1. São Paulo household O/D networks and their MCA factor maps. A) Origin-destination network based on SP O/D 2017. Edge colour computed using $\ln(\#trips)$. Mobility zone boundaries pictured here form the boundaries for nodes in (fig. 1B). **B)** Example individual networks. Node labels are based on mobility zones defined in the survey. Edges are based on trips made between two mobility zones, with weights indicating multiple trips. Plots **C** (MCA result without entropy) and **D** (MCA result *with* entropy) showcase the effects of including entropy in an MCA. Differences in household structure (*famSit*) are nearly reduced to 0 in terms of correlation ratio, while gender differences in commute time and number of trips are emphasised.

4. Macedo, M., Lotero, L., Cardillo, A., Menezes, R. & Barbosa, H. Differences in the spatial landscape of urban mobility: Gender and socioeconomic perspectives. *PLOS ONE* **17**, e0260874, DOI: 10.1371/journal.pone.0260874 (2022).

5. Kaufmann, V., Dubois, Y. & Ravalet, E. Measuring and typifying mobility using motility. *Appl. Mobilities* **3**, 198–213, DOI: 10.1080/23800127.2017.1364540 (2018).

6. Crenshaw, K. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanf. Law Rev.* **43**, 1241–1300 (1990).

7. Sigle-Rushton, W. Essentially quantified? towards a more feminist modeling strategy. In Evans, M., Johnstone, H., Henry, M. & Hemmings, C. (eds.) *The SAGE Handbook of Feminist Theory* (SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP, 2014). DOI: 10.4135/9781473909502.

8. Manderscheid, K. Criticising the solitary mobile subject: Researching relational mobilities and reflecting on mobile methods. *Mobilities* **9**, 188–219, DOI: 10.1080/17450101.2013.830406 (2014).

Data challenge

Talks

Unfolding urban fragmentation: The interconnection between app usage and city landscape

Antonio Desiderio[1, 2], **Zsófia Zádor**[3], **Riccardo Di Clemente**[4,5,*,‡], **and Laura Alessandretti**[6,*,†]

[1]Physics Department and INFN, Tor Vergata University of Rome, 00133 Rome, Italy.
[2]Centro Ricerche Enrico Fermi, 00184 Rome, Italy.
[3]Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.
[4]Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.
[5]The Alan Turing Institute, London, NW12DB, United Kingdom.
[6]Technical University of Denmark, DK-2800 Kgs., Lyngby, Denmark.
[*]equal contributions
[‡]riccardo.diclemente@nulondon.ac.uk
[†]lauale@dtu.dk

In today's digital world, smartphones are the primary means through which we communicate and stay connected with one another. We spend a considerable amount of time on our phones, with a daily average of 2.5 hours smartphone usage with 101 app switching[1]. While smartphones undeniably offer a multitude of benefits, we are also understanding the negative consequences associated with excessive screen time, such as adverse mental health outcomes, increased depressive symptoms, and sleep problems[2]. In particular, smartphone usage can have more severe negative consequences for specific socio-demographic groups, such as youngsters or vulnerable individuals[2]. Therefore, understanding the factors impacting phone usage across different socio-economic groups is crucial. One hypothesis is that the physical location of individuals strongly impacts their smartphone usage patterns. In particular, it is suggested that individuals are less likely to engage in prolonged screen time when they were in settings that facilitated face-to-face interactions or leisure activities[3]. Such urban environments are often described as having urban vibrancy that foster positive, social interactions.

Small-scale empirical studies suggest indeed that young individuals are more likely to spend time engaging with screens when there are few services, convenience goods or public open spaces in the physical spaces around them[3], while the presence of high quality parks reduces children's screen time[4]. Understanding the interaction between screen time and outdoor activity is relevant not only because of the lack of social interactions associated with higher smartphone usage but also due to the negative impact of reduced physical activity[5]. These findings highlight the importance of understanding how smartphone usage patterns changes across different urban locations[6] and to what extent urban features affect the formation and maintenance of social interactions[7]. However, our understanding of how smartphone usage varies in different urban settings remains incomplete, primarily because of insufficient data that encompasses both individuals' smartphone behavior and their presence within specific environments. In this research, we explore to what extent individuals' exposure to different kinds of urban forms - such as parks, residential areas, vibrant city centers and suburbs - explain their usage of different smartphone apps - from social media, to games, and news. Our work will provide a comprehensive quantitative description of the interplay between environmental and smartphone usage, exploring these relations for individuals with different socio-economic backgrounds.

Cities blend buildings, infrastructure and populations into vibrant landscapes. To capture the interplay between smatphone usage and the cityscape, we integrate the NetMob dataset of the app activity[8] with Point of Interest data extracted from OpenStreetMap and census data from INSEE. Our analysis focuses on the study of the quarter-hourly activity $z_i^{\alpha}(t)$, defined as the z-score of the raw activities for a given app aggregated across all weekdays[9]. We have, then, computed the average activity $\langle z_i^{\alpha} \rangle_t$ to seize the app-usage over the day. Our preliminary analyses focused on understanding the role played by park proximity on app usage. Fig. 1 (**A**) reveals that app-usage is not influenced by the park proximity. Fig. 1 (**B**) shows the relative Facebook usage in a rich neighbourhood of Lyon near park. Therefore, these preliminary results indicate that there might be further underlying features apart from the vicinity of parks that affect mobile phone usage. To better understand the role played by the physical environment on app-usage we have quantified the diversity of app-usage. In particular, the sequence of the apps used during a day can reveal patterns in how people use their devices throughout the day, providing insights into how individuals allocate their time on different types of apps and in conjunction with each other. From the activity $z_i^{\alpha}(t)$ we have extracted a sequence $s_i(t)$ of apps that are used the most at the given time t in the tile i ($s_i(t) = \max_{\alpha} z_i^{\alpha}(t)$). By scanning the sequence $s_i(t)$ chronologically, we have computed the predictability of each tile as the number of unique patterns of apps encountered. App usage patterns are highly heterogeneous within the city of Lyon, with some areas characterized by many highly used apps, and others by only a few, as depicted in Fig. 1 (**C**) with a highly fragmented map.

Therefore, we aim to further analyse which urban features and city landscapes are associated with a reduction of social media usage and the diversity of apps, and how temporal variations such as time of day, and weekdays versus weekends might influence social media activity. Additionally, we plan to extend the analysis by comparing multiple cities to identify common patterns and unique characteristics that contribute to differences in social media usage across urban environments. By providing a more nuanced understanding of the interconnections between virtual and physical spaces, our study aims to shed light on urban fragmentation and its implications for app and social media usage. The insights gained from this research can inform urban planning, and the design of strategies to promote digital well-being.

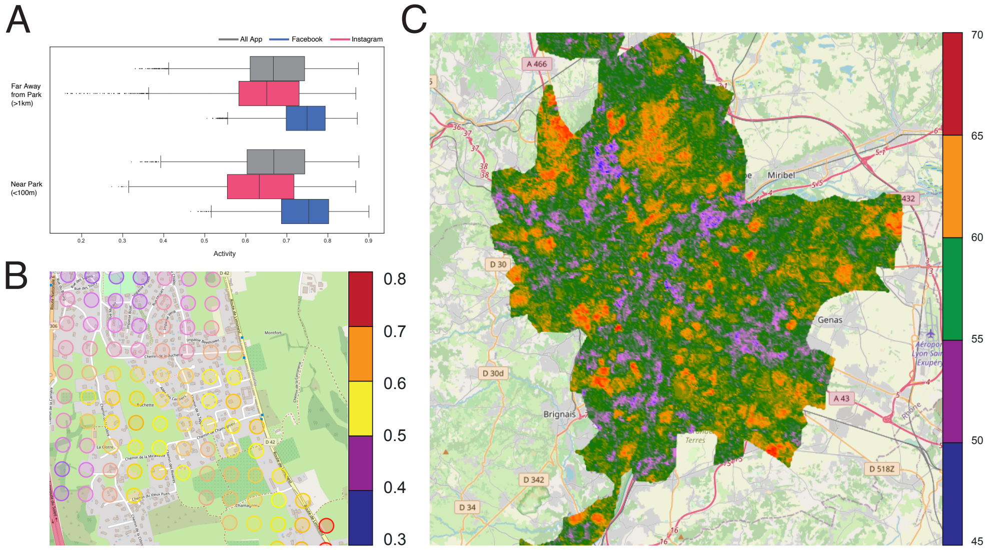


Figure 1. App Usage Patterns and Distance to Park Lyon. A) Comparison of Activity for Different Apps (All Apps, Facebook, Instagram) in Proximity to (distance below 100m) and Far from Parks (distance above 1km). **B)** Map of Facebook Activity: Visualizing Variation in Activity Levels across Different Tiles. **C)** Heat Map of App Diversity Usage in Lyon.

References

1. Deng, T. *et al.* Measuring smartphone usage and task switching with log tracking and self-reports. *Mob. Media & Commun.* **7**, 3–23, DOI: 10.1177/2050157918761491 (2019).
2. Thomée, S. Mobile Phone Use and Mental Health. A Review of the Research That Takes a Psychological Perspective on Exposure. *Int. J. Environ. Res. Public Heal.* **15**, 2692, DOI: 10.3390/ijerph15122692 (2018).
3. Christian, H. *et al.* Nowhere to Go and Nothing to Do but Sit? Youth Screen Time and the Association With Access to Neighborhood Destinations. *Environ. Behav.* **49**, 84–108, DOI: 10.1177/0013916515606189 (2017).
4. Veitch, J. *et al.* Is the Neighbourhood Environment Associated with Sedentary Behaviour Outside of School Hours Among Children? *Annals Behav. Medicine* **41**, 333–341, DOI: 10.1007/s12160-011-9260-6 (2011).
5. Rey-López, J. P., Vicente-Rodríguez, G., Biosca, M. & Moreno, L. A. Sedentary behaviour and obesity development in children and adolescents. *Nutr. metabolism cardiovascular diseases* **18**, 242–251 (2008).
6. De Nadai, M. *et al.* The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. In *Proceedings of the 25th International Conference on World Wide Web*, 413–423, DOI: 10.1145/2872427.2883084 (International World Wide Web Conferences Steering Committee, Montréal Québec Canada, 2016).
7. Botta, F. & Gutiérrez-Roig, M. Modelling urban vibrancy with mobile phone and OpenStreetMap data. *PLOS ONE* **16**, e0252015, DOI: 10.1371/journal.pone.0252015 (2021).
8. Martínez-Durive, O. E. *et al.* The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography (2023). ArXiv:2305.06933 [cs].
9. Toole, J. L., Ulm, M., González, M. C. & Bauer, D. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, 1–8 (2012).

Unmasking Socioeconomic Disparities: A Study of Urban Segregation through the Lens of Mobile App Usage Patterns

Yuya Shibuya[1*], Santiago Garcia Gabilondo[2], Sun Chenchen[2], Yu Peiran[3], Ma Jue[2], and Yoshihide Sekimoto[1]

[1]Center for Spatial Information Science, The University of Tokyo
[2]Department of Civil Engineering, The University of Tokyo
[3]Department of Advanced Interdisciplinary Studies, The University of Tokyo
[*]yuya-shibuya@csis.u-tokyo.ac.jp

Human mobility dynamics are interconnected with diversified lifestyles and human behavioral contexts in modern cities. By understanding how digital and real-world lifestyles intersect, we can gain insight into issues such as work-life balance, the divide between physical and digital experiences, and economic disparities in urban living. In this study, we examined mobile app usage data in France to explore how digital lifestyles are connected to urban segregation. Our findings show that app usage relates not only to land use patterns and time and week of days but also to income levels. For example, lower-income places use games and message apps relatively highly compared with higher-income places by 0.6 and 0.3 points, respectively. On the other hand, higher-income places use video call and email apps relatively highly compared with higher-income places by 0.4 points and 0.2 points, respectively. Additionally, our study sheds light on the importance of synchronicity in digital lifestyles in relation to the physical world. We found clusters with a strong synchrony of digital lifestyle. Further investigation on the impacts of socioeconomic and urban function on synchrony is needed.

Keywords Digital lifestyles · Segregation · Behavior · App usage · Digital devide

In the digital age, investigating the urban digital lifestyle, which captures people's digital practices and socioeconomic and socio-spatial characteristics, is critical to understanding people's modern lives and socioeconomic disparities [1]. In this light, mobile phone apps have been widely studied. Previous research revealed several factors contributing to people's app usage, such as people's location [2, 3, 3], time of day and day of week [4], political, social, and sports event [5]. While existing studies have shed light on user-app interactions, little attention has been given to the relationship between app usage and key socioeconomic contextual factors [1, 6]. In this study, we analyze netmob23

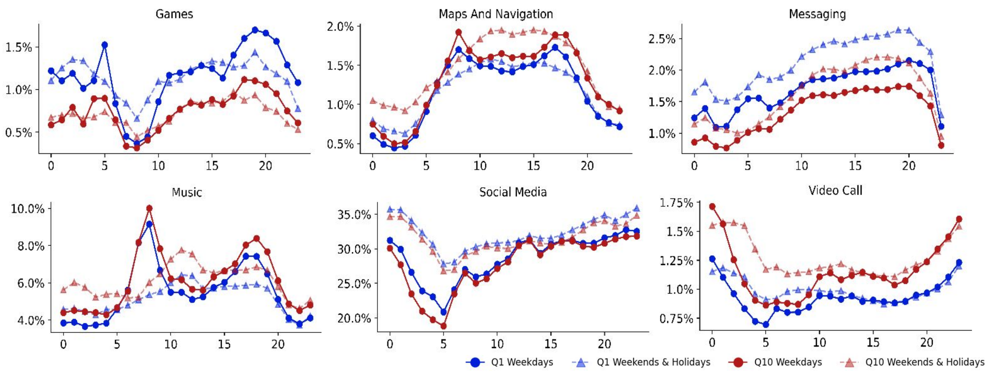


Figure 1: Hourly average app usage (%) on working and non-working days in the highest and lowest income quantiles of major French 20 cities. The x-axis represents the hours of a day.

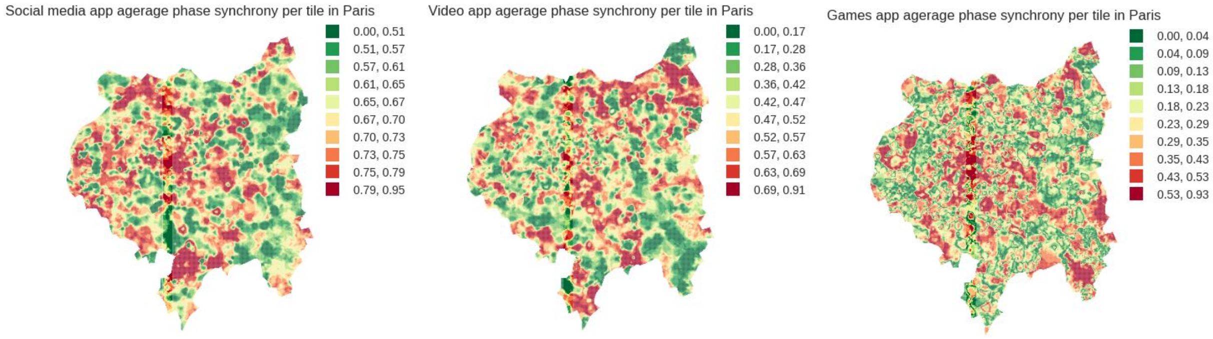


Figure 2: Synchrony of app usage for Social Media, Video, and Game apps in Paris. The color-coded scale represents synchrony values on a scale from 0 to 1, where higher values denote higher synchrony.

datasets [7] to capture interrelationships between app usage and urban socioeconomic context by focusing on through the lens of pattern differences of app usage and synchronicities of app usage.

First, we compare the highest and lowest income quantile areas' app usage patterns across working and non-working days (Saturdays, Sundays, and French holidays). In doing so, we summed hourly application traffic volume, uplink, and downlink. Then, we calculated the percentage of application use types at the tile levels every hour for working and non-working days. Fig. 1 shows the gaps in app usage between higher and lower-income places on working and non-working days. Lower-income areas have higher game and messaging app use ratios than higher-income areas (average 0.45 and 0.30 points, respectively, from 10 a.m. to 8 p.m. In contrast, higher-income areas have higher email and video call use ratios (average 0.42 points and 0.19 points, respectively, from 10 a.m. to 8 p.m.). Furthermore, music, maps, and navigation apps are used more during commuting times only on working days. These income levels' relations with app usage are also observed even within the same land use areas. For example, Game apps are highly used among the lowest income level areas within both industrial and residential places, respectively, while Maps and Navigation apps were more used among the highest income level areas within both residential and industrial places, respectively.

Next, we analyze the synchronicity of digital lifestyles to examine variations in app usage volume, i.e., rhythms, across different temporal and spatial dimensions. To establish a baseline, we used the first two weeks of our dataset to identify the most prominent frequency components in the regularity of app usage by employing the Lomb-Scargle periodogram on our dataset. Our findings revealed that app usage patterns exhibit a strong correlation with 2-hour time intervals. We quantified the regularity of these 2-hour intervals by calculating a rolling window correlation (24-hour window) to compare the expected regularity of app usage with the actual data for the entire dataset. Fig. 2 illustrates the average synchrony per app type at the tile level within Paris. A synchrony value approaching 1 indicates that the app usage rhythms are highly regular over time. As the figure shows, there are several clusters with higher synchrony for each app type. Additionally, in the case of Paris, a weak positive correlation (correlation coefficient: 0.22) was observed. Further analysis on investigating the factors affecting the digital lifestyle synchrony is needed.

References

[1] Tali Hatuka, Hadas Zur, and Jose Antonio Mendoza. The urban digital lifestyle: An analytical framework for placing digital practices in a spatial context and for developing applicable policy. *Cities*, 111:102978, April 2021.

[2] Donghan Yu, Yong Li, Fengli Xu, Pengyu Zhang, and Vassilis Kostakos. Smartphone App Usage Prediction Using Points of Interest. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–21, 2018.

[3] Abhinav Mehrotra, Sandrine R. Müller, Gabriella M. Harari, Samuel D. Gosling, Cecilia Mascolo, Mirco Musolesi, and Peter J. Rentfrow. Understanding the Role of Places and Activities on Mobile Phone Interaction and Usage Patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–22, 2017.

[4] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56. Association for Computing Machinery, 2011.

[5] Steven Van Canneyt, Marc Bron, Andy Haines, and Mounia Lalmas. Describing Patterns and Disruptions in Large Scale Mobile App Usage Data. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 1579–1584. ACM Press, 2017.

[6] Yi Ren, Tong Xia, Yong Li, and Xiang Chen. Predicting socio-economic levels of urban regions via offline and online indicators. *PLOS ONE*, 14(7):e0219058, 2019.

[7] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

Assessing the space-use efficiency of French cities by coupling 3D city models with mobile data traffic

Zhexuan Tan, Yuehan Yang, and Markus Schläpfer*
Department of Civil Engineering and Engineering Mechanics
Columbia University, New York, USA

In light of climate change, resource scarcity, and population growth, it becomes increasingly important to use the existing built-up space of our cities as efficiently as possible. However, despite this pressing societal challenge, the degree to which the available three-dimensional (3D) urban space is currently being utilized by human activities has never been studied systematically. Here, we couple mobile data traffic with 3D city models for 20 major cities in France to explore their space-use efficiency in terms of the built-up volumes that cover the time-varying presence of people. We find a clear, 'donut-like' organization of cities, where the space-use efficiency is surprisingly low in the city center, becomes high in the immediate surroundings, and then low again in the suburbs. This hitherto hidden regularity reveals a large potential to increase the overall resource efficiency of cities by increasing the utilization of under-used spaces in the city centers.

I. INTRODUCTION

The ongoing worldwide population growth and rapid urbanization, together with growing concerns over climate change and resource scarcity, make it important to utilize the existing urban built-up spaces as efficiently as possible [1, 2]. This is particularly relevant for the existing building stock of cities. Through their vertical expansion, buildings increase the available floor space and provide venues for interactions and socioeconomic activities [3], but they also consume substantial amounts of material and energy [1]. Meanwhile, accelerated by the recent pandemic, the shift towards remote work and online shopping has been leading to an under-usage of many buildings [4]. Therefore, using buildings - and the urban built-up space in general - efficiently and in a more agile way promises to build a bridge between demand and supply, eventually facilitating more resource-efficient and climate-friendly cities [1].

Understanding how citizens are using the available built-up space is of great value for designing policies aiming at efficient space utilization. Indeed, several methods have been developed to investigate dynamic building occupancy, such as agent-based modeling [5] or analysis of mobile phone records [6]. However, these approaches focus on occupancy-driven building energy use and do not assess the actual utilization of the available built-up space.

On these premises, the main objectives of our study are: 1) To assess the feasibility of using mobile data traffic to estimate the time-varying population distribution in cities; 2) To assess the potential of the resulting dynamic population maps for the systematic identification of urban areas with underused built-up space; 3) To explore the potential of such a data-driven approach for the identification of recommendations towards more efficient and agile uses of the existing urban built-up spaces.

II. METHODOLOGY AND DATASETS

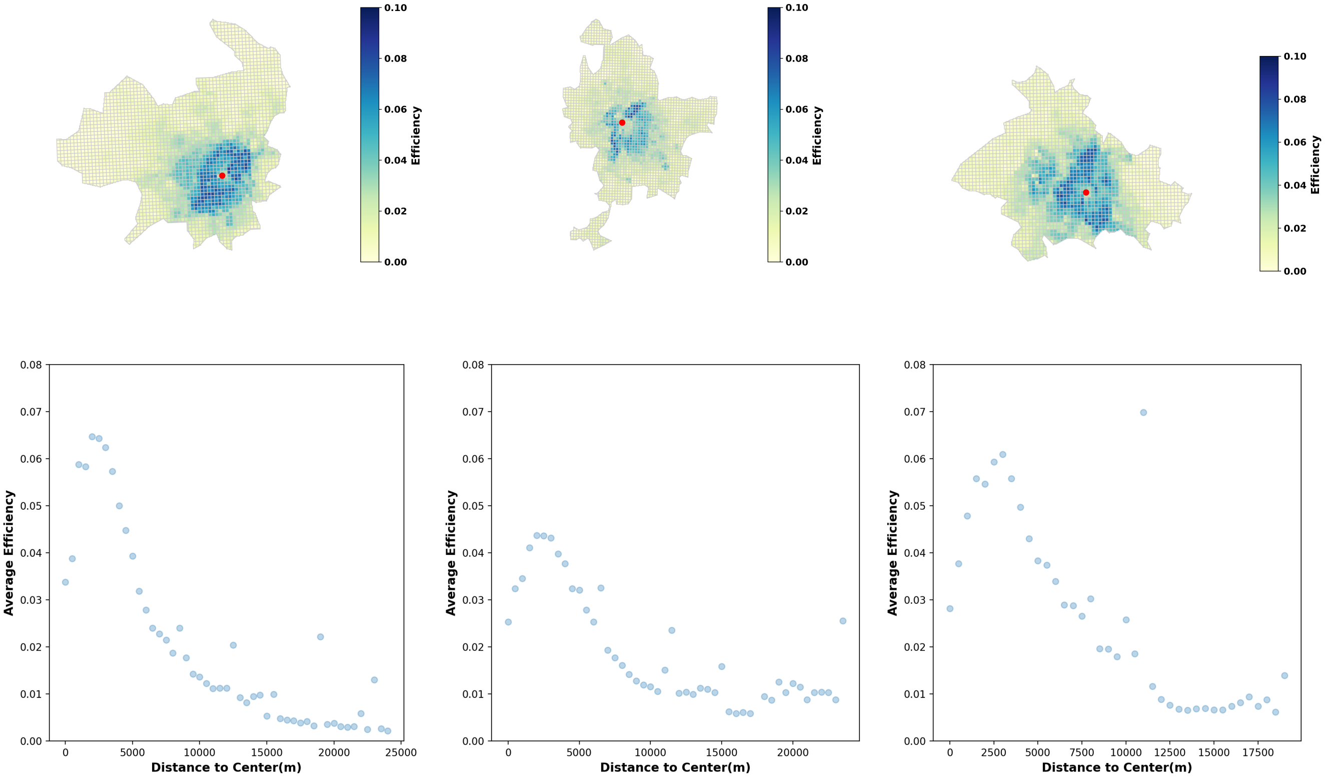
We develop a framework that: 1) Applies machine learning methods to estimate the dynamic population distribution in a city from aggregate mobile data traffic; 2) Combines the resulting population maps with detailed 3D city models and quantifies the built-up space use efficiency at a fine-grained spatial resolution. The datasets involved in our study include: 3D city models [7], mobile traffic data [8], census population [9], existing dynamic population data [10], and land use information [11].

III. RESULTS AND DISCUSSION

The comparison among our models and cross-validation with the previous dynamic population dataset show that a random forest model performs well in predicting the dynamic population.

Based on our predicted population, we calculate the space-use efficiency within and across cities. An interesting phenomenon is discovered whereas in almost all cities there is a clear, 'donut-like' organization of the built-up space utilization: it is low in the city center, becomes high in the immediate surroundings, and then low again in the suburbs. Figure 1 depicts this pattern for three exemplary cities. A possible explanation is that the buildings in the city centers are mostly commercial and office buildings which have lower occupancy-over-capacity rates than residential buildings and are largely empty during night hours and weekends [12]. This points to a general under-utilization of the built-up areas in the centers of the studied cities (with Paris being an exception).

This "hollowing out" of cities has potentially been further amplified by the recent global pandemic, with people moving out of the city centers due to the emergence of working from home and online shopping [13]. As such, our results indicate a large potential to increase the efficiency of cities by increasing the utilization of city centers through diversification of attraction points or through an improvement of the existing urban infrastructures.



(a) Bordeaux

(b) Lyon

(c) Toulouse

FIG. 1: Urban space-use efficiency patterns for the examples of Bordeaux, Lyon, and Toulouse. The red dots in the maps indicate the location of the city center (city hall). The bottom row shows the behavior of the average efficiency values with increasing distance from the city centers.

[1] D. Dodman et al., Cities, settlements and key infrastructure, in *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2022).

[2] United Nations Environment Programme, & International Resource Panel, The weight of cities: Resource requirements of future urbanization (2018).

[3] M. Schläpfer, J. Lee, and L. Bettencourt, Urban skylines: Building heights and shapes as measures of city size (2015), arXiv preprint arXiv:1512.00946.

[4] E. L. Glaeser and C. Ratti, 26 Empire State Buildings Could Fit Into New York's Empty Office Space. That's a Sign, The New York Times (2023).

[5] Y. Chen, T. Hong, and X. Luo, An agent-based stochastic occupancy simulator, Building Simulation **11**, 37 (2018).

[6] E. Barbour et al., Planning for sustainable cities by estimating building occupancy with mobile phones, Nat. Commun. **10**, 3736 (2019).

[7] `https://land.copernicus.eu/local/urban-atlas/building-height-2012`.

[8] O. E. Martínez-Durive et al., The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography (2023), arXiv:2305.06933.

[9] National Institute of Geographic and Forest Information (IGN), 'The reference database for the infra-municipal dissemination of the results of the population census by iris, of decametric precision', `https://geoservices.ign.fr/contoursiris`, accessed: 2023-06-15.

[10] F. Batista e Silva, S. Freire, M. Schiavina, K. Rosina, M. A. Marín-Herrera, L. Ziemba, M. Craglia, E. Koomen, and C. Lavalle, Uncovering temporal changes in Europe's population density patterns using a data fusion approach, Nature Communications **11**, 4631 (2020).

[11] `https://opendata.apur.org`.

[12] E. Barbour, C. C. Davila, S. Gupta, C. Reinhart, J. Kaur, and M. C. González, Planning for sustainable cities by estimating building occupancy with mobile phones, Nature Communications **10**, 3736 (2019).

[13] M. Batty, The post-pandemic world: Are big cities hollowing out?, Environment and Planning B: Urban Analytics and City Science **50**, 1409 (2023).

* m.schlaepfer@columbia.edu

NetMob 2023 Data Analysis with ASCA

José Camacho, CITIC, University of Granada, Granada, Spain, josecamacho@ugr.es

***Abstract*—We provide insights into the Netmob 2023 Data Challenge using ANOVA Simultaneous Component Analysis (ASCA), employed mainly in the clinical sciences. ASCA is a combination of Analysis of Variance (ANOVA) and Principal Component Analysis (PCA). We also present the first Big Data extension of ASCA. This work was supported by the Agencia Estatal de Investigación in Spain, MCIN/AEI/ 10.13039/501100011033, grant No PID2020-113462RB-I00.**

I. Summary

We can generally organize the traffic statistics of the Netmob 2023 Data Challenge in 3-way tensors of ‘space’ x ‘time’ x ‘service’. This data arrangement is useful to unveil spatio-temporal trends in the service consumption. We can use this approach at different scales (resolutions) of time and space giving way to hierarchical/multi-scale models that can be used to inspect the data at different levels of detail.

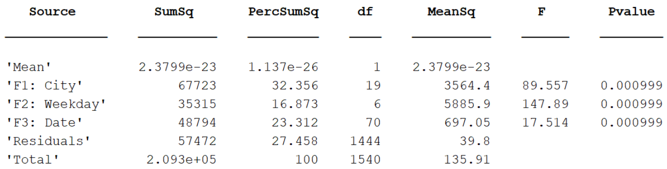
Resulting tensors can be analyzed in dozens of ways depending on the interest of the analyst. In particular, multivariate models are useful to understand complex interactions among the services. In this paper we focus on ASCA in order to understand spatio-temporal and sociological patterns in the data. ASCA i) partitions the data according to a set of factors, like time or location, allowing us to untangle temporal and spacial patterns, ii) performs statistical inference and iii) allows visualizations of the patterns with fully interpretable PCA plots. Furthermore, we normalize the data using the Centered-Log-Ratio (CLR), because we are more interested on the relative composition of the traffic in terms of the usage of services, than in actual traffic values, which trivially depend on the amount of population. We also apply auto-scaling, in order to give the same relative importance to all services in the visualizations.

Our research group has been a main code developer of ASCA in the last years. We integrate ASCA code into the MEDA Toolbox, available for free in Github, and all the plots and software used in this paper will be made available as soon as the organizers allow us to.

In the following we provide three analyses: i) high-level spatio-temporal ASCA model of ‘cities’ x ‘days’ x ‘services’, ii) high-level spatio-social ASCA+PCA model of ‘cities’ x ‘services | population statistics | profession statistics’ and iii) city-level models based on Big Data ASCA.

A. High-level spatio-temporal ASCA model

We propose an ANOVA factorization in three factors: the ‘City’, the ‘Weekday’ and the ‘Date’. Following ANOVA theory[1], factors ‘City’ and ‘Weekday’ are fixed and crossed, and factor ‘Date’ is random and nested in ‘Weekday’. After variance factorization and statistical inference, we get the table in Figure 1. This table can be interpreted like any ANOVA table. It shows that all factors are statistically significant (their influence in the traffic is important) and we can look at the ‘MeanSq’ to see their relative importance: in order of relevance we get ‘Weekday’, ‘City’ and ‘Date’.



```
     Source          SumSq     PercSumSq   df      MeanSq        F       Pvalue
 ______________   ___________  __________  ____  ____________  _______  _________

 'Mean'           2.3799e-23   1.137e-26      1   2.3799e-23
 'F1: City'            67723      32.356     19       3564.4   89.557   0.000999
 'F2: Weekday'         35315      16.873      6       5885.9   147.89   0.000999
 'F3: Date'            48794      23.312     70       697.05   17.514   0.000999
 'Residuals'           57472      27.458   1444         39.8
 'Total'           2.093e+05         100   1540       135.91
```

Fig. 1. ANOVA Table for the high-level spatio-temporal ASCA model.

PCA[2] biplots (scatter plots of observations and services) allow to relate spatio-temporal patterns with the services. Basically, observations showing high (relative) usage of a service will be located in the plot far from the center of coordinates towards the same direction of that service. Biplots for data corresponding to Factors ‘Weekday’ and ‘City’ are shown in Figure 2[3]. Figure 2(a) shows a clear separation between working days and weekends in the first PC (horizontal axis). Working days have a higher relative usage of services like ‘Web-Finance’, ‘Microsoft-Mail’ or ‘Google-Mail’, while in weekends the relative usage of recreational services like ‘WhatsApp’, ‘Apple-Video’ or ‘Youtube’ is higher. We can also distinguish the average usage patterns of Fridays (‘Google-Play-Store’ and ‘Spotify’), Saturdays (‘Deezer’ and ‘Apple-Music’) and Sundays (‘Netflix’ and ‘Web-Transp’). Figure 2(b) shows a clear separation among cities. To give some selected examples, the relative usage of ‘Uber’ is higher in Paris, while ‘Apple-iCloud’ is most popular (in relative terms) in Nice. Finally, the biplot for factor ‘Date’ (not shown) unveils trends in the 70 days: ‘Spotify’ significantly reduces its traffic during the time interval under study while ‘Apple-iTunes’ significantly increases it. ‘Apple-App-Store’ and ‘Snapchat’ also lose traffic by the end of the period.

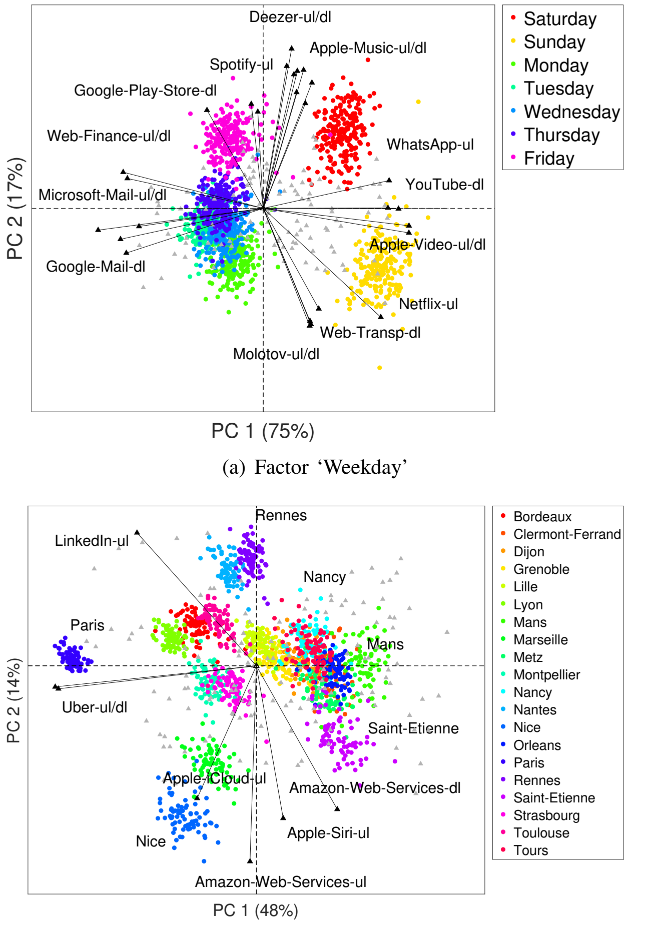
Previous insights (and other more subtle ones we did not discuss for the sake of brevity) improve our understanding of the traffic in terms of the factors under study. ASCA untangles the effect of each factor on the traffic, simplifying

[1] In the summary, we assume reasonable knowledge of ANOVA. See for instance Montgomery, Douglas C. Design and analysis of experiments. John wiley & sons, 2017.

[2] In the summary, we assume reasonable knowledge of PCA interpretation. See for instance Jolliffe, Ian. "Principal component analysis." Encyclopedia of statistics in behavioral science (2005).

[3] For the sake of simplicity, only selected services are labelled in the plots.

978-1-6654-0601-7/22/$31.00 © 2022 European Union



(a) Factor ‘Weekday’

(b) Factor ‘City’

Fig. 2. High-level spatio-temporal ASCA model of days x cities x services: biplot of (a) factor City and (b) factor Weekday.

interpretation. For instance, from the ASCA partition for factor ‘City’, where temporal patterns are filtered out, we can compute the relative distribution of selected services over the map of France, which is simpler to interpret than a time-series (a video) of maps. Also, confirmatory figures are a perfect complement for PCA visualizations. All patterns found in ASCA where validated with confirmatory plots.

B. High-level spatio-social ASCA model

In this section we combine the previous factorization of ASCA for factor ‘City’ with population and professional statistics. The result is a matrix of 20 cities x 246 variables that we explore with PCA in Figure 3. Care should be taken when interpreting this Figure, as we do not have statistical inference, as we do in ASCA. Besides, the data is highly dimensional, and clear patterns are difficult to observe. Yet, the visualization can be useful to unveil potential correlations that should be confirmed with independent data or increasing the spatial or temporal resolution. The first PC (horizontal axis) shows that there is a higher relative usage of recreational services in, e.g. Saint-Etienne, than what can be found in Paris. This pattern seems to be positively correlated with the proportion of men without schooling for more than 15y and with a CAP/BEP degree, and negatively correlated with the proportion of women without schooling for more than 15y and with a superior degree. Nice seems to have higher proportion of old population, which can influence the specific usage of services (e.g., usage of more expensive Apple devices). As future work, we would like to apply sparse and variable selection approaches to improve the interpretability of the analysis. However, a more effective direction would be to obtain population and professional statistics at equivalent resolution than the traffic data.

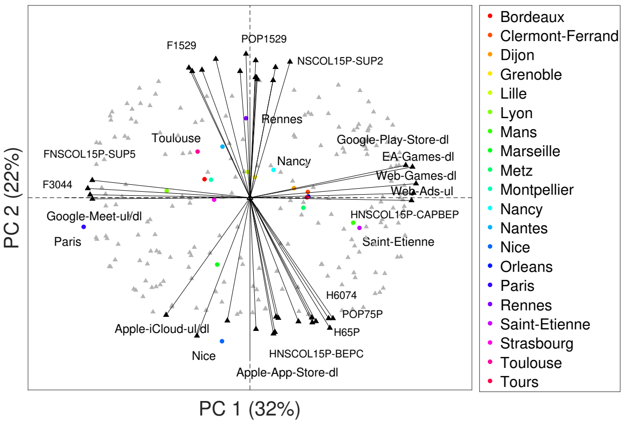
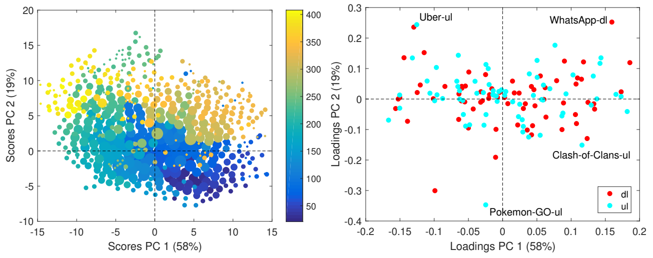


Fig. 3. Biplot of the high-level spatio-social PCA model of ‘cities’ x ‘services | population statistics | profession statistics’.



(a) Scores

(b) Loadings

Fig. 4. Paris Big Data ASCA factorization for Factor ‘Row’. Selected labels have the format: row-column-date.

C. City-level spatio-temporal ASCA models

We have specifically developed for this challenge the first Big Data (out-of-the core) implementation of ASCA. We use clustering to allow visualizations of unlimited numbers of observations, inspired in our previous Big Data implementations of other multivariate tools. We have computed Big Data ASCA models for each of the 20 cities, from their corresponding tensor of days x tiles x services. For the sake of space, we can only provide here a summary of results and one example of visualization in Figure 4. City-level results allow to conclude that the main factor affecting the traffic is always the ‘Weekday’, and that there is systematically differential usage of services in time and location within all cities. For instance, in Figure 4(a), weekend patterns are highlighted towards the right part, and working day patters towards the left part. The scores also show some separability during the weekend along the vertical axis. In Figure 4(b), we see that ’Instagram-dl’ (among others not labelled) is connected to weekend patterns. All patterns found in ASCA where confirmed with service-level city maps, but are not shown here for the sake of brevity.

Maintaining App Services in Disrupted Cities: A Crisis and Resilience Evaluation Tool

Leon Würsching
Secure Mobile Networking Lab
Technical University of Darmstadt
Darmstadt, Germany
lwuersching@seemoo.tu-darmstadt.de

Matthias Hollick
Secure Mobile Networking Lab
Technical University of Darmstadt
Darmstadt, Germany
mhollick@seemoo.tu-darmstadt.de

ABSTRACT

Disaster scenarios can disconnect entire cities from the core network (CN), isolating base stations (BSs) and disrupting the Internet connection of app services for many users. Such a disruption is particularly disastrous when it affects *critical app services* such as communication, information, and navigation. Deploying local app servers at the network edge can solve this issue but leaves mobile network operators (MNOs) faced with design decisions regarding the criticality of traffic flows, the BS topology, and the app server deployment. **We present the Crisis and Resilience Evaluation Tool (CARET) for crisis-mode radio access networks (RANs)**, enabling MNOs to make informed decisions about a city's RAN configuration based on real-world data of the NetMob23 dataset.

1 SCENARIO AND GOAL

We assume a disaster scenario where a city is disconnected from the CN, but a set of BSs is still available. **Our goal is to maintain app services for the users of a disrupted city by deploying local app services on the network edge**. In this context, we focus on the user plane, so our scenario assumes that the MNO has already attended to the availability of the control plane and the necessary CN functions. Any approaches to reconnect the city to the CN or the Internet are outside the scope of this work as we provide a solution within the disrupted city.

Starting from today's cellular networks, many decisions have to be made so that the app services can be maintained in disrupted cities: Our vision (cf. Figure 1) is that the users remain connected to the closest BS, app services are deployed on local edge servers, and BSs route traffic between users and app services via wireless links.

2 CONTRIBUTION

For the RAN to support such a crisis scenario, the MNO has to make many decisions, e.g., about the available infrastructure and supported traffic. **Our contribution is the evaluation tool CARET for crisis-relevant decisions with real-life mobile traffic data from the NetMob23 data set**. CARET (cf. Figure 2) expects mobile traffic data and the following crisis decisions as input:

- **Apps**. Which apps should supported in crisis mode?
- **Base Stations**. Which BSs are available in the city
- **Edge Servers**. On which BSs can edge servers be provided?
- **App Servers**. Where are the app services deployed?
- **Links**. Which links are available to connect BSs?
- **Routing**. How is traffic routed through the city?

It then evaluates how well the given decisions perform for the provided traffic data and returns the following evaluation metrics:

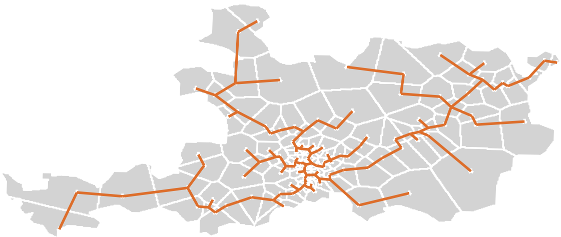


Figure 1: The RAN of Saint-Etienne in crisis mode showcasing each BS's coverage (gray tiles) and the minimum set of inter-BS links connecting all BSs (orange lines).

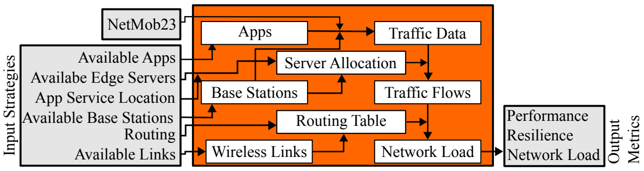


Figure 2: System model of CARET, including input parameters, the data flow, and the resulting output metrics.

- **Resilience**. Which fraction of the traffic can be served?
- **Performance**. How much traffic is routed wirelessly?
- **Load**. What is the overall traffic load in the city?

3 DATA PREPARATION

Each input parameter to CARET [3] is either a concrete *configuration*, e.g., the set of available BSs, or a *strategy*, such as `HIGH TRAFFIC 80`, which uses the provided traffic data to identify the BSs with the highest traffic volume and make 80% of them available.

CARET's required traffic data format is very similar to that of the NetMob23 data set [2], with the following differences: On the spatial dimension, we consider BS-wise instead of tile-wise traffic utilizing the BS information available at Cartoradio [1]. On the temporal dimension, we deviate from the NetMob23 directory and file structure to obtain a file for each time slot, facilitating parallel evaluation. **We provide a conversion tool for the NetMob23 data set to be compatible with CARET.**

4 EVALUATION

As depicted in Figure 2, we filter the input traffic data by available apps and BSs. For each app service, we select one edge-capable BS

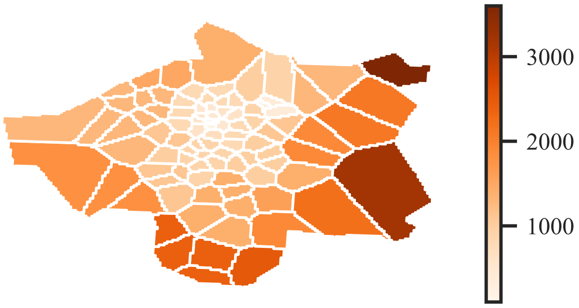


Figure 3: Minimum wireless link range [m] required for the BSs in the city of Nancy to connect to the inter-BS network.

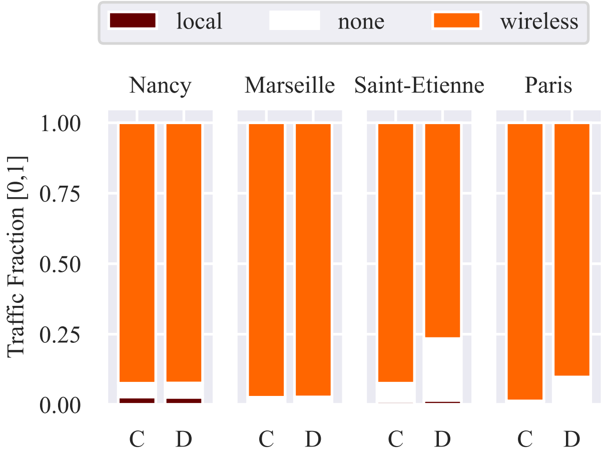


Figure 4: RAN performance for a connectivity of 85%. Depicted is the traffic fraction that is handled locally, routed wirelessly, and cannot be served for the app service strategies `CENTRAL` (C) and `DECENTRAL` (D).

to host the local app service. This enables us to generate traffic flows while considering two traffic profiles: The uplink traffic of each BS is partitioned per app and routed to the corresponding app service, and the downlink traffic is routed on the reverse path. Each traffic flow is routed through the city based on the routing table and the available links. Then, we compute the network load by accumulating the individual link loads.

5 RESULTS

CARET is applicable to the entire NetMob23 data set [2]. However, we limit our reporting to the following cities to showcase different combinations of size and population density: Nancy (small and sparse), Marseille (small and dense), Saint-Etienne (large and sparse), and Paris (large and dense). For this section, we assume the role of an MNO configuring a city's RAN in response to a crisis.

5.1 Link Range vs. Energy Consumption

Wireless link establishment in ad-hoc networks is a well-researched problem, and an optimal configuration highly depends on individual infrastructure characteristics. Therefore, we skip concrete configurations and instead showcase CARET's functionality with a simple model where each BS connects to all neighboring BSs in a radius of r meters. The MNO has to trade off service coverage and energy consumption: A larger radius leads to higher inter-BS connectivity but causes higher energy consumption. Figure 3 shows the minimum link range required for each BS in Nancy to connect to the inter-BS network. The required radius for a connectivity of 85% is 1200 m in Paris, 1400 m in Marseille, 1500 m in Nancy, and 4300 m in Saint-Etienne. The MNO can use CARET to evaluate different link establishment algorithms and decide which link range is required to achieve the desired connectivity.

5.2 Central vs. Decentral Service Deployment

Deploying app services on the local network edge is a non-trivial allocation problem, so we showcase two simple deployment strategies: Deploying all app services on one edge server with the most traffic (`CENTRAL`) and deploying each service at the edge server where there is the most app-specific traffic (`DECENTRAL`). For Figure 4, we evaluated the traffic data of May 31, 2019, with the following strategies: All apps and all BSs are available, all BSs are edge-enabled, and each city's link range is set to achieve 85% connectivity with minimum distance routing. The fraction of locally handled traffic n dense cities is negligible, and non-serviceable traffic is lower than in sparse cities. While both strategies perform similarly in small cities, the `CENTRAL` strategy dominates in large cities. The high fraction of non-serviceable traffic for the `DECENTRAL` strategy in Saint-Etienne is caused by app services deployed at one of the 15% disconnected BSs. CARET enables MNOs to evaluate the performance of deployment strategies for local app services based on recorded data, e.g., the NetMob23 data set [2].

6 CONCLUSION AND FUTURE WORK

The goal of this work is to maintain app services in disrupted cities by deploying app services on the local network edge. To this end, we introduce the Crisis and Resilience Evaluation Tool (CARET) [3], supporting MNOs to make informed decisions for the configuration of crisis-mode RANs. Furthermore, we provide a conversion tool that enables MNOs to use CARET with the NetMob23 data set.

For future work, we envision a cellular network that can transition into crisis mode, as described in Section 1. We will continue the development of CARET to support MNOs in this endeavor.

REFERENCES

[1] Cartoradio. 2023. The map of radio sites and wave measurements. https://cartoradio.fr/#/. [Available online. Last accessed 2023-09-16].

[2] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. 2023. The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv:2305.06933 [cs.NI]

[3] Leon Würsching and Matthias Hollick. 2023. *CARET: Crisis and Resilience Evaluation Tool.* Secure Mobile Networking Lab. https://github.com/seemoo-lab/caret

Traffic and Resource Optimization in 5G Multi-Layer Edge Networks

Marcello Pietri, Natalia Selini Hadjidimitriou, Marco Mamei, Marco Picone, Enrico Rossini
University of Modena and Reggio Emilia, Italy
{marcello.pietri,selini.hadjidimitriou,marco.mamei,marco.picone,enrico.rossini}@unimore.it

The next generation of mobile networks - 5G and beyond - will need to support services and applications with a wide range of different requirements in terms of Quality of Service (QoS). The key technology to meet such requirements is the combination of: (i) data-driven solutions and machine learning models to forecast mobile demand and throughput (ii) the virtualization of networks functions and servers to provide operators with unprecedented flexibility on how to allocate resources, re-route traffic and slice the network dynamically

In this research, we propose automating network management through data-driven intelligence, with a particular focus on anomalies and network traffic during specific events or periods (e.g., gatherings, sporting events, concerts, etc.). We analyze the NetMob23 dataset [1] with the goal of forecasting mobile demand for different classes of services, and we provide algorithms to optimize the resources allocated to network slices and optimize traffic distribution (load balancing) within the operator's network to match the demand.

Building upon existing work by [2], [3], [4], the contribution of this paper is to provide a "full-stack" solution dealing with:

- Modeling the mobile network. We create a network model based on a vertical partitioning base on network slices, and an horizontal partitioning based on multiple layers of core and edge computing nodes.
- Demand forecasting. We trained a deep-learning model - based on LSTM - to forecast network demand for different classes of services. We focus on predicting spikes that can saturate network resources and capacity.
- Network optimization algorithms. We developed high-level optimization algorithms to improve network performance by sharing/re-balancing resource among slices, and sharing/re-balancing traffic among network nodes.

In our model the mobile network consists of a set of network nodes / data-centers comprising VNFs and application servers (see Figure 1). Nodes are organized in hierarchical layers. At the bottom there are the *edge* layers. Such network nodes are associated to a subset of BSs and are in charge of managing traffic to/from such BSs. Top layers represent the *core* network, such network nodes are in charge of processing data from lower layers and interacting with cloud services and resources outside the mobile operator network. Network traffic runs from BSs to edge nodes, to core nodes and vice versa.

We assume that the mobile traffic generated by BSs is divided into different classes of services. The resources in each network node are divided between multiple network slices. There is a network slice for each class of service, and each slice is in charge of processing the traffic of that class of service (for example, the slice associated to video-services will handle all the traffic generated by videos). This architecture is general enough to accommodate a wide range of modern network deployments. Finally, in order to link the traffic demand on a node with a measure of performance – following [3], [4] – we introduce, for each node n and slice s, a capacity $C_{n,s}$ representing the amount of traffic that the resources associated to that node in the slice can handle. This is a strong simplification as $C_{n,s}$ subsumes different resources in terms of: networking, memory, storage and computation. The capacity should effectively handle normal traffic but it might be saturated during peak events. Referring to the traffic in a node n in slice s at time t as $T_{n,s,t}$, we assume that the network performance degrades if $T_{n,s,t} > C_{n,s}$. In order to optimize the network operations we will try to minimize the times in which $T_{n,s,t} > C_{n,s}$ and the amount of the excess $T_{n,s,t} - C_{n,s}$.

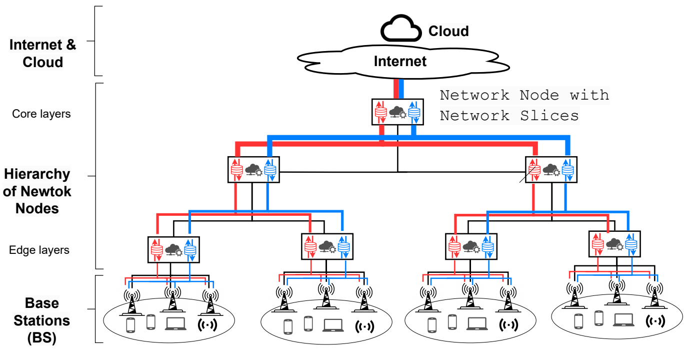


Fig. 1: A schematic representation of the hierarchical and sliced architecture associated with the 5G edge networks.

Given the above scenario, we devised a methodology to optimize network operations on the basis of forecast traffic on the different slices.

Our base hypothesis is that the network allocates resources fairly statically according to the modeled capacities $C_{n,s}$. If traffic is manageable $T_{n,s,t} < C_{n,s}$ the network does not modify its resources. Vice versa, if the traffic is predicted to grow beyond capacity, the network can temporarily adjust resources to meet the surge in demand. Once traffic is predicted to diminish, the network will return to its base allocation $C_{n,s}$.

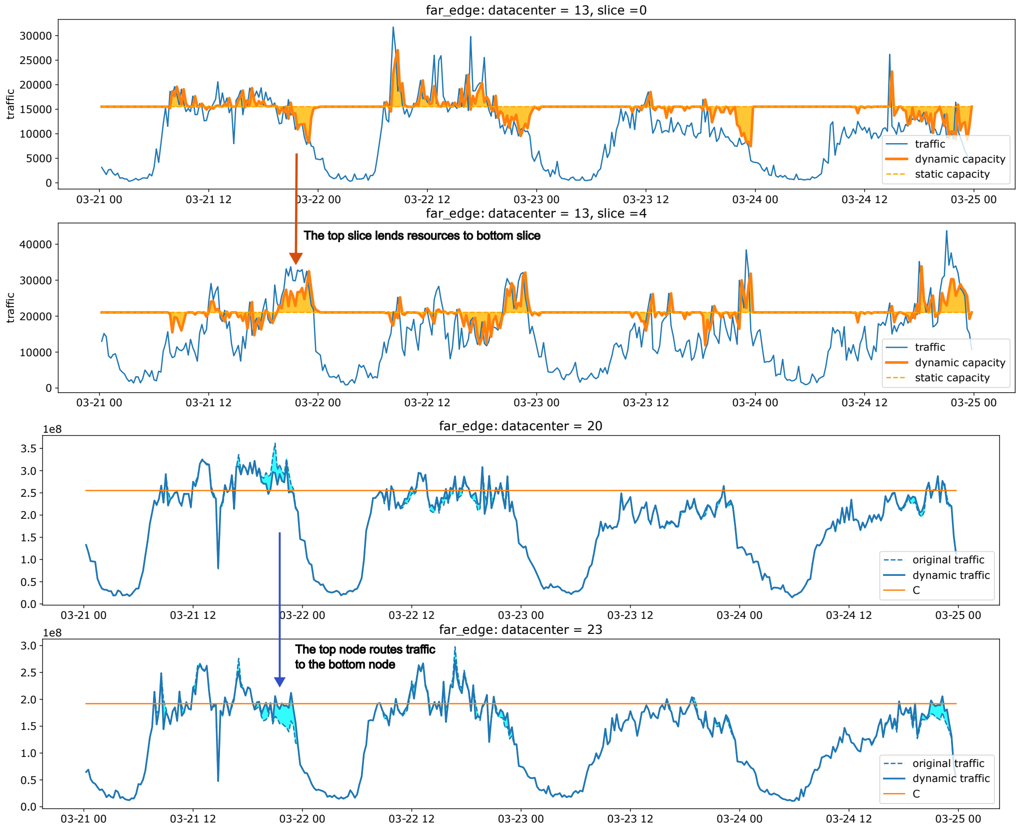


Fig. 2: Dynamic capacity optimization and Dynamic traffic optimization

Our approach is based on two complementary mechanisms: **Resource sharing/re-balancing among slices.** Our network model is based on a set of hierarchical nodes / data-centers processing network traffic. The resources in each node C are divided to handle multiple slices. C_s are the resources allocated to handle slice s ($C = \sum C_s$). If the traffic associated to a slice $s1$ at time t is greater than the capacity $T_{s1,t} > C_{s1}$ the network under-performs. If there is another slice $s2$ in the same data-center with spare capacity $T_{s2,t} < C_{s2}$, some resources could be reallocated from $s2$ to $s1$. The idea of this mechanism is to apply this pattern on the basis of forecast traffic: if the network forecasts that at future time t, $T_{s1,t} > C_{s1}$ and $T_{s2,t} < C_{s2}$, then it re-balances resources in advance, leaving the total amount C unchanged - see Fig. 2 - top. **Traffic sharing/re-balancing among network nodes/data-centers.** The dual approach is to re-balance traffic instead of resources: if the above resource-balancing-within-a-node can not accommodate all the demand ($\sum_s T_{s,t} > C$), a node can re-route part of the traffic to other nodes. This re-route can take place: *(i)* among the nodes of the same network layer, e.g. part of the traffic to a given edge node is routed to a nearby node at the same edge layer (i.e., horizontal offloading). *(ii)* among the nodes at different network layers, e.g. part of the traffic to a given edge node is routed to and processed by a node at an upper layer of the hierarchy - a core node (i.e., vertical offloading) - see Fig. 2 - bottom.

Experimental results show that dynamic reallocation of resources among slices and traffic between nodes improves performance by more than 11% on average.

Acknowledgment. We would like to thank Andrea Buldorini, Edoardo Maria Sanna, Daniela Di Rienzo and Jovanka Adzic from Telecom Italia (TIM) for useful discussion and suggestions in setting up the research and the manuscript.

This work is supported by Ecomobility project co-funded by the European Commission JU-KDT grant agreement No. 101112306 (Research and Innovation Action).

References

[1] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," 2023.

[2] A. Ceselli, M. Fiore, A. Furno, M. Premoli, S. Secci, and R. Stanica, "Prescriptive analytics for mec orchestration," in *2018 IFIP Networking Conference (IFIP Networking) and Workshops*, 2018.

[3] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18, New Delhi, India, 2018. [Online]. Available: https://doi.org/10.1145/3241539.3241567

[4] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez, "Resource sharing efficiency in network slicing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019.

Network Sharing for sustainability and resilience in the era of 5G and beyond

Marco Ajmone Marsan*, Michela Meo*, Maoquan Ni*, Daniela Renga*

Abstract—**The transition towards the era of 5G and beyond is characterized by the widespread penetration of extremely demanding communication services entailing the need for exchanging increasingly larger traffic volumes with stringent throughput and latency requirements. Whereas the extensive densification of radio access networks (RANs) aims at coping with this challenging scenario, sustainability concerns arise due to the consequent huge increase of energy consumption and to the costly deployment of new infrastructure that, being dimensioned for future peak demands, may result underutilized for long periods of time. In addition, new potential vulnerabilities emerge that may impair the provisioning of resilient communication services. In this context, sharing network resources among different mobile operators (MOs) may play a key role to improve energy efficiency and to enhance resilience of future mobile networks. We aim at investigating the potential benefits derived from the sharing of network infrastructure (primarily Base Stations with their portion of spectrum) among different network operators in a urban environment. Based on the NetMob real traffic data, we will design data-driven strategies to dynamically offload traffic among Base Stations owned by different MOs, allowing the switch off of unneeded resources. We will then evaluate the performance of these strategies in reducing network energy consumption and cost, and in enhancing the network resilience to power outages, to enable a feasible deployment of 5G scenarios and a sustainable evolution towards 6G.**

I. Concept

The deployment of 5G networks is expected to provide excellent quality of service (QoS) to extremely high numbers of devices, to enable flawless services during user mobility, and to enhance energy efficiency with respect to previous communication technologies [1], [2]. Currently, the 5G technology is in its initial commercialization stage [2]. However, we are still witnessing relevant challenges in the actual realization of the 5G era, that is characterized by the widespread penetration of extremely demanding communication services, and requiring edge caching and computing to support smart mobility [3]. Whereas the extensive densification of radio access networks (RANs) aims at coping with this challenging scenario, sustainability concerns arise due to the consequent huge increase of network energy consumption and to the installation of new infrastructure that, being dimensioned for future peak demands, may result underutilized for long periods of time. Furthermore, mobile operators (MOs) face increasingly higher expenses, in terms of both operational cost (OPEX), due to the growing energy demand, and capital expenditures (CAPEX), due to the need for integrating new expensive network components based on 5G technology in their current infrastructure. Finally, new potential vulnerabilities emerge that may impair the provisioning of resilient communication services, entailed by the fast expansion of the network and by possible overload on the energy grid. In this context, sharing network resources among different MOs may play a key role to improve energy efficiency and to enhance the resilience of future mobile networks [3], [4].

Our research aims at investigating the potential of network sharing to achieve a number of objectives: (i) to improve the energy efficiency of RANs; (ii) to reduce the OPEX due to the network energy demand; (iii) to limit the CAPEX faced by MOs to install new proprietary network nodes based on 5G technology; (iv) to enhance network resilience in case of power outages due to electric grid overload, cyber attacks, natural disasters, or emergency situations. We hence consider a portion of the mobile access network, focusing on a densely populated urban environment. We assume that the capacity provided by Base Stations (BSs) owned by different MOs can be shared, based on predefined agreements, as long as the coverage of the shared BSs is overlapping.

The objective of this study consists in evaluating the performance of these strategies in reducing the network energy consumption, decreasing operational cost, limiting CAPEX, and enhancing the network resilience to power outages, to enable a feasible deployment of 5G scenarios and a sustainable evolution towards 6G. Clearly, an effective analysis should be performed based on real and up-to-date mobile traffic data to derive reliable outcomes about network sharing that result actually representative of a rapidly evolving scenario. To this extent, the NetMob23 Dataset presented in [5] results extremely suitable to perform our research.

II. Methodology

We consider a simple case study in a urban area, in which two MOs provide mobile access service through a number of LTE BSs. The network sharing is enabled by implementing a traffic offloading strategy that operates on pairs of BSs owned by different operators, denoted by Op_1 and Op_2, that are located in the same site. At every time step a check is performed to identify the BS that is characterized by the highest residual capacity. Let us assume this BS is owned by Op_1. If its current residual capacity is sufficient to host the traffic volume handled by the other BS, owned by Op_2, this traffic volume is moved to the Op_1 BS, provided that the Op_1 BS capacity is not saturated above a threshold that we denote C_{th}. The Op_2 BS can hence be deactivated to save energy. To estimate the BS energy consumption we adopt the power models detailed in [6] for LTE technology. Real traces of normalized mobile traffic from the NetMob dataset, provided by a French MO, are used to model the BS traffic demand [5]. The traffic patterns cover a period of 77 days, with samples collected every 15 minutes with a spatial resolution of 100×100 m^2, and represent more than 60 different mobile services. We consider real data about the geographical distribution of sites hosting mobile BSs from different MOs [7] and refer to real electricity market prices [8]. We focus on a sample area in the city of Lyon, France, derived from the aggregation of $n \times n$ contiguous tiles, for which

*Politecnico di Torino. Email: marco.ajmone@formerfaculty.polito.it, michela.meo@polito.it, maoquan.ni@studenti.polito.it, daniela.renga@polito.it

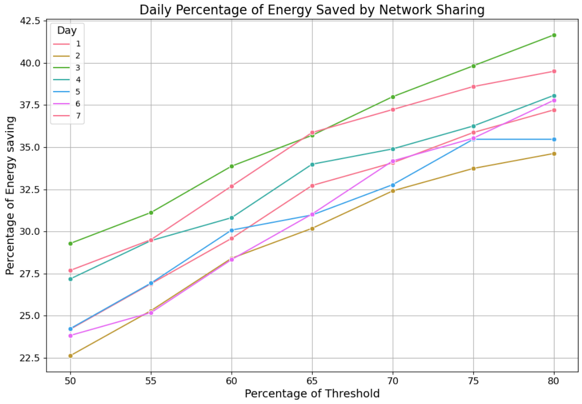


Fig. 1: Energy saving under different C_{th} settings of the threshold C_{th} per each day of a sample week.

traffic traces from Op_1 are available based on NetMob dataset. We then consider the actual location of BSs installed within the defined area that are owned by operators Op_1 and Op_2. Only downlink traffic generated by video and web services is considered, since they account for the largest fraction of mobile traffic demand. To derive realistic traces of the traffic volumes handled by the Op_1 BSs located in the considered areas, we map the traces of traffic volumes distributed over the $n \times n$ tiles to each of the BSs from Op_1 included in the considered area, so that each traffic time series is associated to the closest BS from Op_1. Finally, the traffic trace representing the actual traffic volumes handled by each BS is derived aggregating the traffic data series associated to that BS and further scaling the traffic proportionally to the actual BS bandwidth capacity. In addition, each BS owned by Op_2, for which actual traffic profiles are not available, is assigned the traffic volume profile corresponding to the closest Op_1 BS, proportionally scaled to its capacity.

III. Network sharing analysis

We first focus on characterizing the available traffic data, also investigating how traffic volumes are distributed in space and time. We hence perform a preliminary evaluation of the network sharing potential, based on the aforementioned traffic characterization and on the analysis of how the BSs from different MOs are distributed in the city of Lyon. Finally, to investigate the effectiveness of network sharing in achieving sustainability and resilience goals, we evaluate the performance via simulation over a period of one week, with time step duration of 15 minutes. The identified urban area, whose size is 20×20 tiles (4 km^2), includes two pairs of co-located BSs from different operators, denoted S_1 and S_2.

Due to the limited space, we present only a sample of our results, that demonstrate the remarkable benefits yielded by the application of the network sharing paradigm. In particular, our performance analysis shows that the implementation of network sharing strategies allows to achieve relevant energy saving, of up to more than 40% with respect to the case in which no network sharing is applied, depending on the offloading strategy configuration settings. Fig. 1 depicts the energy saving that can be obtained for the BS pair S_1 in the considered urban area when network sharing is applied under different settings of C_{th}, for each day of the week. The savings raise as the threshold C_{th} is set to higher values, with relevant variability observed over different days, confirming that network sharing effectiveness is influenced by traffic load and user behavior. Interestingly, more than 20% of energy can be saved even under the most conservative threshold settings, highlighting the potential of network sharing to trade off sustainability goals and QoS requirements. Considering the surface of the urban area under evaluation, in which two pairs of BSs are suitable for network sharing application, the traffic offloading strategy allows to save up to 37 kWh/km^2 per week. Assuming an average electricity price of 0.174 €/kWh, more than 132 €/km^2 can be saved per week. Although these figures may seem limited, scaling up these savings to the entire area of mobile network deployment over the whole country owned by the considered operators and over an entire year results in a remarkable reduction of the electricity bill.

IV. Conclusion

Our study highlights the potential of network sharing to enable a feasible deployment of 5G scenarios and a sustainable evolution towards 6G. In particular, our preliminary results show that sharing the network infrastructure among different MOs is effective in reducing the network energy consumption by up to more than 40%, entailing further benefits in terms of reduction of the electricity bill. As future work, we aim to expand our research on network sharing exploring wider portions of the radio access network, considering more complex multi-operator scenarios, and envisioning the integration of local renewable energy supply to power BSs. The availability of an extensive set of recent data on mobile network operations opens many possibilities for further relevant research directions, in relation to the contribution that network sharing may give to resilience in 5G and beyond networks and to the definition of a path toward net-zero networks.

Acknowledgement

This paper was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”, Focused Project R4R).

References

[1] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, “A Survey on Resource Allocation for 5G Heterogeneous Networks: Current Research, Future Trends, and Challenges,” *IEEE Communications Surveys Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.

[2] V. K. Quy, A. Chehri, N. M. Quy, N. D. Han, and N. T. Ban, “Innovative Trends in the 6G Era: A Comprehensive Survey of Architecture, Applications, Technologies, and Challenges,” *IEEE Access*, vol. 11, pp. 39 824–39 844, 2023.

[3] N. Slamnik-Kriještorac, H. Kremo, M. Ruffini, and J. M. Marquez-Barja, “Sharing Distributed and Heterogeneous Resources toward End-to-End 5G Networks: A Comprehensive Survey and a Taxonomy,” *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1592–1628, 2020.

[4] A. Gomes, J. Kibiłda, A. Farhang, R. Farrell, and L. A. DaSilva, “Multi-Operator Connectivity Sharing for Reliable Networks: A Data-Driven Risk Analysis,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 2800–2811, 2021.

[5] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, “The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography,” arXiv:2305.06933 [cs.NI], 2023.

[6] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermark *et al.*, “D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown,” *Earth*, vol. 20, no. 10, 2010.

[7] “Agence Nationale des Fréquences (ANFR).” [Online]. Available: https://data.anfr.fr, https://cartoradio.fr [Accessed on 28 June 2023]

[8] “Réseau de Transport d’Électricité,” https://www.rte-france.com/, [Online: accessed 28 June 2023].

Inferring urban spatial dynamics through mobile phone traffic activity

Ulysse Marquis[1,4], Sebastiano Bontorin[1,2], Massimiliano Luca[1,3], Andrea Guizzo[1], Simone Centellegher[1], Bruno Lepri[1], Riccardo Gallotti[1]

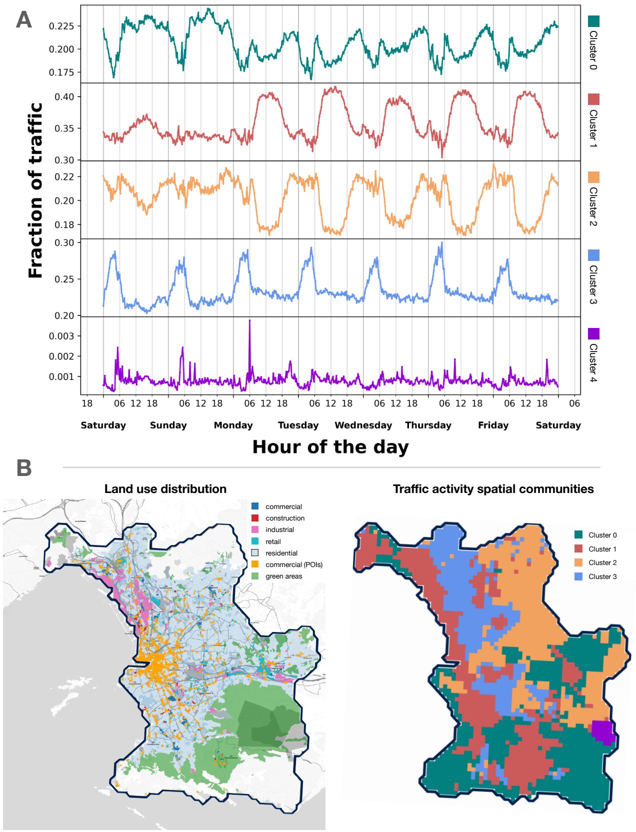
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy
[2]Department of Physics, University of Trento, Via Sommarive 14, 38123 Povo (TN), Italy
[3]Free University of Bolzano, piazza Università 1, 9100, Bozen-Bolzano, Italy
[4]Department of Mathematics, University of Trento, Via Sommarive 14, 38123 Povo (TN), Italy

The surge in cell phone data consumption, combined with advancements in communication infrastructure, has opened new insights to the fields of urban analysis and human activity modeling [2] [7] [9] [10].

The aim of this work is to unravel macroscopic patterns from data traffic extracted from NetMob23 dataset [5]. To accomplish this, we employ a network-based procedure, making use of Louvain's community detection method [3], proposed in previous works [1], [4] and employed with cellphone records to examine the complex interplay between the spatial and temporal dimensions of human activity in the urban space [4]. Additionally, we extend this community detection approach on the traffic dynamics decomposed across the different applications. This approach allows us to effectively identify a small number of communities of applications with similar temporal dynamics and usage. Thus shedding light on the inherent functional organization of traffic usage and its relation with urban space and land use.

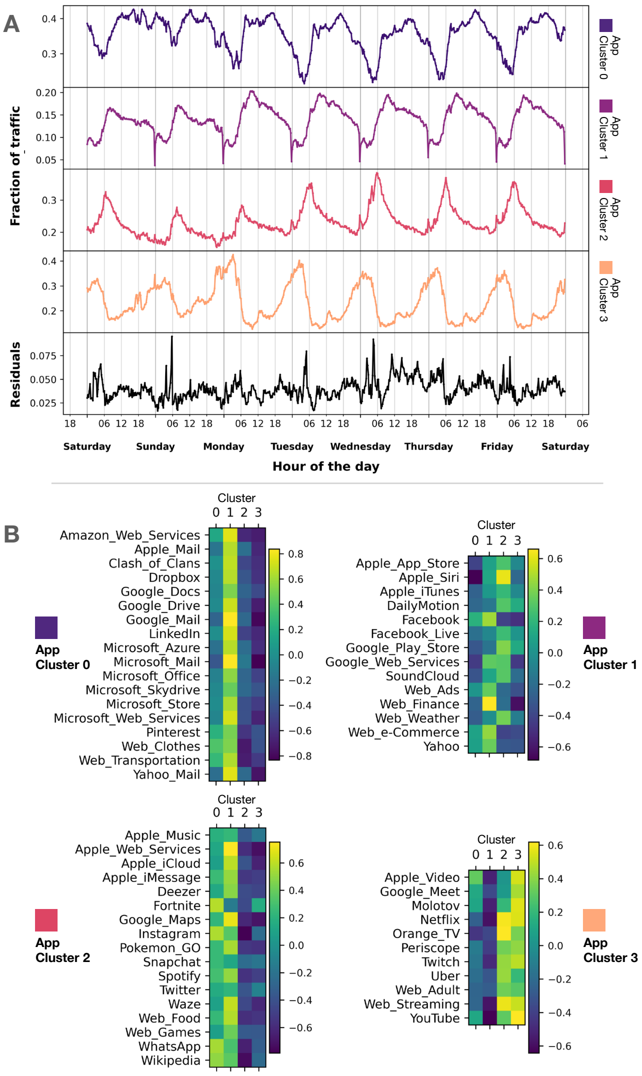
We first propose a comprehensive case study that focuses on the urban area of Marseille. This in-depth examination explores the interplay between the aggregate traffic temporal dynamics and the underlying urban space and land use, obtained from Open Street Map [6]. We assess the robustness of the proposed method by comparing the properties of the communities found through this method with those previously identified using phone call records in earlier studies. The case study on the urban area of Marseille reveals several interesting aspects of the discussed method. Firstly, the proposed method is able to identify, temporal patterns already noticed in previous works [4]. Moreover, we can map these clusters to the underlying land use. Additionally, the method is sensitive to significant groups of tiles with out-of-distribution behavior, as the detection of the military basis of ”Camp de Carpiagne“ reveals. This example demonstrates the precision and flexibility of the method at capturing macroscopic patterns with meaningful characteristics and size. Figure 1 displays the temporal and spatial structure of the identified communities.

Furthermore, the community detection method applied on the set of applications is able to identify four clusters, exhibiting characteristic temporal patterns. These communities gather together applications linked to professional activities (such as mail), or grouping streaming applications. Computing relationship between these functional app clusters and the spatial communities of aggregated traffic dynamics unravels additional dependencies in human activity and mobile phone usage. Figure 2 displays the temporal dynamics of cumulative densities for the cluster of applications and their correlation with the spatial clusters identified in Marseille.



Cumulative traffic density patterns of identified communities in Marseille (A) an their spatial layout compared with land use distribution (B). We remark the emergence of patterns corresponding to work and study (Cluster 1 in red), logistics (Cluster 2) and residential (Cluster 3) activites from [4]. Cluster 0 appears to be related to leisure and non-work daily activities, being largely laid out along the beach and the ”Calanques“ area. Interestingly, the purple community corresponds geographically to a military basis (”Camp de Carpiagne“), highlighting the resolution capability of this technique in identifying patterns which deviate from more standard land uses/activities. Moreover, clusters show coarsening in space as well complex organization patterns. It is also worth noticing interesting coincidence patterns between communities and land uses. For instance, the industrial parts of the coast belong to the red cluster, associated with work, while the leisure portion of the coast coincides exclusively with the ”leisure“ cluster. The land use distribution map is generated using OpenStreetMap data [6].

Finally, we extend these analyses to a subset of 9 other cities composed of Nancy, Nice, Tours, Metz, Le Mans, Clermont-Ferrand, Dijon, Marseille and Grenoble. Across the diverse cities examined in our study, we discern similarities in both in the temporal and spatial coarsening of the temporal dynamics. More precisely, we were able to consistently detect three typical patterns across every urban area,



Cumulative traffic dynamics of application clusters (A) and correlation between traffic of application groups and spatial clusters in Marseille (B). In panel (A) we notice peculiar behaviours in the clusters of applications. Cluster 0 is characterized by work-related apps such as emails, App Clusters 1 and 2 show predominant daily activity with most of social media and music streaming. And finally App Cluster 3 with night activity (composed of video and streaming applications). The Residuals (black curve) corresponds to the density of non-clustered applications. In the figure (B), we remark high correlation between work-related applications and the "work" community in the temporal cluter of Marseille (see Fig. 1) , and high correlation between streaming applications and the "logistics" and "residential" communities. Additionally, noticeable levels of anti-correlation are shown between applications between streaming applications and "work" cluster.

corresponding to "*work*", "*residential*" and "*industrial*" activity patterns, as defined in [4].

Moreover, we compare the relationship between identified communities and land use over the set of cities and unravel the interplay between human activity and land use organization, exploiting metrics such as Block entropy, introduced in [4], and Ripley's index [8].

In conclusion, we analyze the data provided by the NetMob23 dataset [5] through the lens of a network-based community detection method. We observe results replicating previous works' observations [4] in the spatial and temporal organization of detected communities. We propose also a new perspective of the temporal and spatial dynamics through an investigation exploiting the dataset decomposition of groups of akin applications. We propose an application clustering and unravel a matching between spatial clusters and these functional application clusters. Our exploration of mobile data traffic unveils compelling insights on the interplay between land use organization and human activity. Moreover, the addition of a functional dimension to the analysis, allowed by the highly detailed and rich NetMob23 dataset [5], permits a novel characterization of human activity in urban space through traffic data.

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[2] V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ data science*, 4(1):10, 2015.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[4] M. Lenormand, M. Picornell, O. G. Cantú-Ros, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, M. San Miguel, and J. Ramasco. Comparing and modelling land use organization in cities. *Royal Society Open Science*, 2(12):150449, 2015.

[5] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

[6] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org , 2017.

[7] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, and L. Bravo. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ data science*, 10(1):29, 2021.

[8] B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977.

[9] V. Soto and E. Frías-Martínez. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM International Workshop on MobiArch*, page 17–22, New York, NY, USA, 2011. Association for Computing Machinery.

[10] J. L. Toole, M. Ulm, D. Bauer, and M. C. Gonzalez. Inferring land use from mobile phone activity, 2012.

Characterising Temporal Patterns in Visits to Locations for Human Mobility Modelling

Prathyush Sambaturu[1], Bernardo Gutierrez[1,2], Moritz U.G. Kraemer[1,3]
1. University of Oxford, UK; 2. Universidad San Francisco de Quito USFQ, Ecuador;
3. Pandemic Sciences Institute, University of Oxford, UK
{prathyush.sambaturu, bernardo.gutierrez, moritz.kraemer}@biology.ox.ac.uk

Introduction

Human mobility data is key to understanding the transmission of infectious diseases and controlling their spread effectively [1]. Empirical human mobility data can be obtained through multiple sources; for example, mobile phone data is typically available in the form of Call Detail Records (CDR), eXtended Details Records (XDR), and GPS traces [2]. However, accessing this type of data at spatial and temporal resolutions that are epidemiologically meaningful is challenging, as it pertains to individual privacy rights. Therefore, there is a growing need for generative human mobility models that can produce synthetic trajectories that are as realistic as possible.

The seminal work by Song et al. [3] proposed the Exploration and Preferential Model (EPR), which is capable of generating individual synthetic trajectories. Recently, many extensions of this model were proposed, of which some of the main themes were to include preferential exploration (both the Density-EPR or d-EPR [4,5] and Preferential Exploration and Preferential Return, PEPR) [4,5], or social and temporal dimensions (as in STS-EPR) [6]. The d-EPR [4,5] model assumes that an individual explores new locations that are *i)* geographically close to their current location, and *ii)* preferentially choose locations that are densely populated; together, these two conditions determine the 'gravity' of the newly explored location. In this sense, the d-EPR model conceptually shares characteristics with the gravity model, which describes movements between locations based on their population sizes and geographical distance between them. On the other hand, the PEPR model[4,5] assumes that individuals prefer to explore new locations depending on their "attractiveness" and distance.

The PEPR models assume that the gravity or attractiveness of a location remains constant throughout the period for which trajectories are generated. However, the gravity or attractiveness of a location are not static and change over time: the attractiveness of a location can be perturbed by mass events such as sports matches or concerts happening within its vicinity, which result in the attraction of a larger number of visitors than normal for a brief period of time we hereby refer to as 'crowding spikes'. Further, in some places, there is also a periodicity to some of these events. Capturing such patterns or events that result in crowding is useful to better understand the transmission dynamics of infectious diseases in urban areas [7]. The existing models are not capable of generating these short-lived, sometimes periodic spikes in visits to a location in a way that is congruent with empirical observations. Therefore, in this project, our aim is to better understand the temporal nature of the attractiveness of a location in order to develop a new mobility model that incorporates such mechanisms.

Methods and Results

Data preprocessing and the maximum visit time series generation. We consider Facebook and Whatsapp service data for the six largest cities of France: Paris, Marseille, Lyon, Toulouse, Nice, and Nantes for a period of 77 days between 16/03/2019 and 31/05/2019 from the data provided in the NetMob 2023 Challenge[8]. For a given location, we take the maximum traffic in any 15-minute time period of a day as a measure of the busiest period for that location (a 100 m^2 tile). Facebook and WhatsApp data tend to show peak values at different times; we therefore select the maximum value from whichever is highest between both services over the same day. By taking the busiest time of a location on a given day, we intend to capture these 'crowding spikes', which show a sharp increase in magnitude of active users in a location. This selection produces a time series for each location, where each point in the series represents the maximum number of visitors on a particular day. (**Figure 1a)** shows the time series for randomly sampled (25% of total) locations in Paris. We refer to the shape of this time series over time as the 'crowding behaviour' of that location.

Clustering of locations using self organizing map neural networks. We use Self-Organizing Map (SOM) neural networks [9], an unsupervised learning algorithm, to cluster together locations that show similarities in the crowding behaviour over time. The implementation is done using the Minisom[10] and Tslearn[11] packages in Python. Our choice of parameters results in 9 clusters per run (which we use for comparability between locations), while the optimal number of clusters can be found using the silhouette score. (**Figure 1b)** visualises the clusters resulting from SOM, where locations within the same cluster are filled with the same colour. A vast majority of locations (shown in **Figure 1c**) outside the center of Paris belong to the same cluster and exhibit low crowding spikes. On the other hand, the city center has many clusters and exhibits a variety of crowding behaviours and magnitudes. For instance, the locations in cluster 3 (**Figure 1d**) has only 14 locations **(Figure 1c)** all contiguously located in the eastern portion of the city, a little away from the city center (**Figure 1b**). The cluster is surrounded by relatively less busy areas as compared to the center of the city. These locations typically pull less crowds on most days; however, there is a lone crowding spike (**Figure 1d**) in the middle of the week starting on April 6th, 2019 (06/04/2019 - 13/04/2019) of magnitude at least 3 times of any other crowding spike. This suggests that there was an event that pulled a large crowd to the locations in this cluster during that day. This pattern is an anomaly in the crowding behaviour of the locations in this cluster. Cluster 5, with 150 locations (**Figure 1f**), also exhibits interesting crowding behaviour with a weekly pattern (periodicity is a week) (**Figure 1f**), where there is less crowd during the weekends and relatively larger crowds during weekdays with occasional spikes.

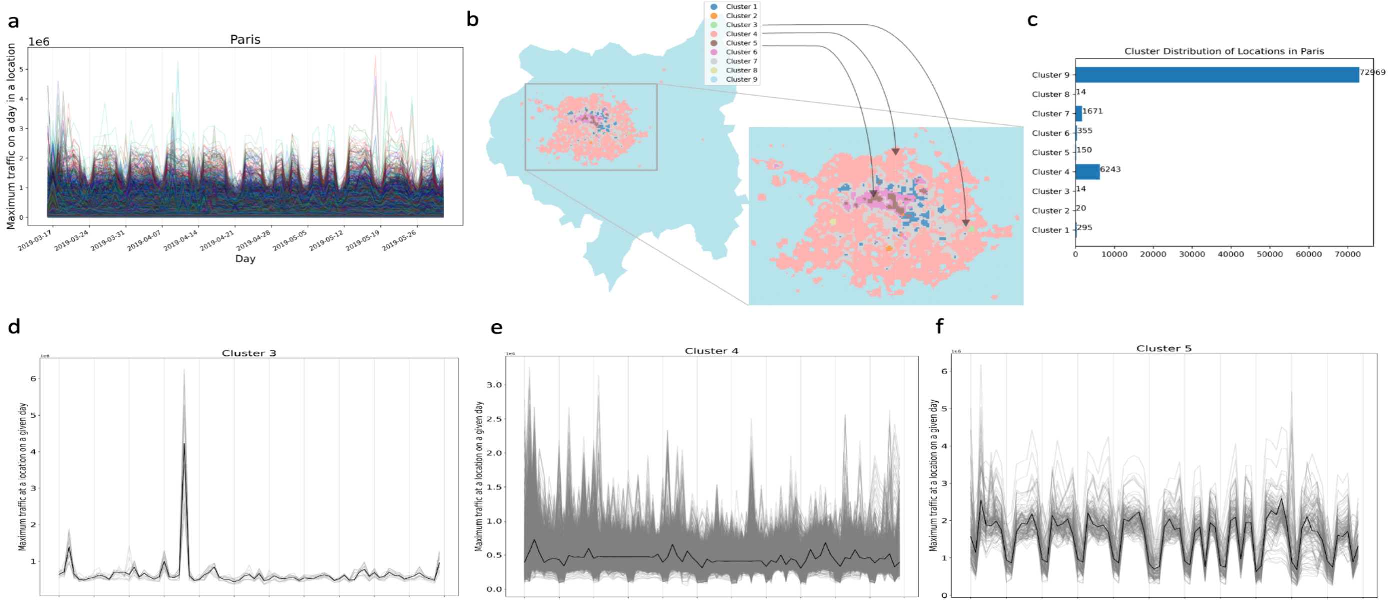


Figure 1. *Temporal variations and trends in maximum visits per day to locations in Paris from data obtained by combining Facebook and WhatsApp records.* ***a)*** *Maximum traffic on a day in a sampled set of locations (25% of total) in Paris between 16/03/2019 and 31/05/2019 (77 days). The x-axis represents the days with markings for Sundays, and the y-axis represents the maximum number of visits in any 15–minute time period of a day. Each trace corresponds to a single location (100 m² tile) in the city.* ***b)*** *Map of Paris showing the 9 different groups of locations clustered by Self-Organizing Maps (SOM) based on the similarity in time series corresponding to peak-time visitors over the 77 days. The region in the box is the center of the city; arrows show locations belonging to clusters 3, 4, and 5.* ***c)*** *Distribution of locations inferred to belong within each cluster, showing a vast majority of locations (outside the city center) being grouped together, whereas more diversity in crowding patterns is observed within the centre of Paris.* ***d, e, and f)*** *Each of these plots shows the time series corresponding to all locations in each cluster. The central dark line corresponds to the barycenter of the cluster showing the crowding behaviour.*

Conclusions and Future Work

The experiments in our work show that examining the maximum number of users in a location at any time of a day is useful in identifying crowding behaviour in cities of France. The clusters obtained using our methods provide key insights into various types of crowding behaviours, some of which may not be captured by existing human mobility models. An essential next step is to experimentally verify whether the synthetic mobility trajectories generated by PEPR models, with just the population distribution of a city as input, are capable of generating these crowding behaviours. A future direction is to use further analysis to incorporate the mechanisms behind these crowding patterns, observed in various clusters of a city, into the PEPR model so as to be able to generate such patterns in synthetic data.

References

1. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
2. Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C. & Bravo, L. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Sci* **10**, 29 (2021).
3. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823 (2010).
4. Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. *Procedia Comput. Sci.* **83**, 934–939 (2016).
5. Schläpfer, M. *et al.* The universal visitation law of human mobility. *Nature* **593**, 522–527 (2021).
6. STS-EPR: Modelling individual mobility considering the spatial, temporal, and social dimensions together. *Procedia Comput. Sci.* **184**, 258–265 (2021).
7. Rutten, P., Lees, M. H., Klous, S., Heesterbeek, H. & Sloot, P. M. A. Modelling the dynamic relationship between spread of infection and observed crowd movement patterns at large scale events. *Sci. Rep.* **12**, 1–16 (2022).
8. Martínez-Durive, O. E. *et al.* The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography. (2023).
9. Javed, A., Rizzo, D. M., Lee, B. S. & Gramling, R. SOMTimeS: Self Organizing Maps for Time Series Clustering and its Application to Serious Illness Conversations. (2021).
10. GitHub - JustGlowing/minisom: :red_circle: MiniSom is a minimalistic implementation of the Self Organizing Maps. *GitHub* https://github.com/JustGlowing/minisom.
11. Tavenard, R. *et al.* Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **21**, 1–6 (2020).

Semantic Maps to Generate Mobile Phone Traffic Data

Chiara Pugliese
ISTI-CNR, University of Pisa, Italy
chiara.pugliese@isti.cnr.it

Francesco Lettich
ISTI-CNR, Pisa, Italy
francesco.lettich@isti.cnr.it

Giulio Loddi
IMT Lucca, Italy
giulio.loddi@imtlucca.it

Fabio Pinelli
IMT Lucca, Italy
fabio.pinelli@imtlucca.it

Chiara Renso
ISTI-CNR, Pisa, Italy
chiara.renso@isti.cnr.it

1 INTRODUCTION

Generative AI, a subfield of artificial intelligence, has witnessed remarkable advancements in recent years, enabling machines to create, mimic, and generate content that closely resembles human creations. This field encompasses a variety of techniques and models designed to generate original and realistic outputs based on learned patterns and data. These typically utilize deep learning architectures to generate diverse content such as images, music, text, and even interactive experiences. By identifying underlying patterns and statistical properties of training data, generative AI models can produce novel outputs with a wide range of applications in various domains, including the generation of high-quality synthetic datasets[1].

In the realm of urban mobility data, numerous deep learning models have been developed in recent years to address various challenges. For example, models utilizing the encoder-decoder paradigm have shown promise in learning trajectory embeddings [1–3], which are compact vector representations that encapsulate spatio-temporal attributes of trajectories and can be also used for generating synthetic trajectories, and network-level traffic flow generation [4]. [5] proposes a two-stage GAN method (TSG) that uses a map-based approach to generate synthetic trajectories that adhere to the constraints and characteristics of the map. Similarly, considering various types of information attached to each cell in the NetMob dataset – such as population density and Points of Interest (POIs) – could aid in producing more realistic data. For instance, the approach proposed in [6] uses semantic knowledge (e.g., travel mode, trip purpose, POIs) as a guiding signal to simulate human mobility.

Overall, the generation of synthetic mobility datasets can offer numerous benefits in terms of privacy, confidentiality, and proprietary concerns. However, it is crucial to balance two key factors: the utility of the generated data for analytical purposes and the need for the data to be sufficiently different from the original to maintain its integrity and avoid privacy issues. In [7] the authors proposed a hybrid CNN/LSTM architecture to generate synthetic and dynamic high-resolution population density data from static low-resolution data. In the experiments, the authors show that the performance improves significantly when increasing the use of POIs in the model.

Building upon these insights, our work for Net Mob Challenge 2023 proposes a method for **generating reliable datasets of synthetic mobile data traffic** that leverages the diversity of city areas. The idea could be to transform the city into an image with dimensions of $N \times N$ pixels, where each cell in the image corresponds to a specific area and contains multiple channels representing the associated semantic information.

By leveraging the specificities of individual cells within the city image representation, we can generate cellular traffic data that reflect the expected traffic patterns. A model can be trained on existing data from a specific city and thereafter applied to synthesize traffic patterns for other cities with different traffic patterns. We argue that the proposed method holds potential for various applications, such as urban planning, infrastructure optimization, and simulating the impact of policy changes on mobile data usage within urban environments.

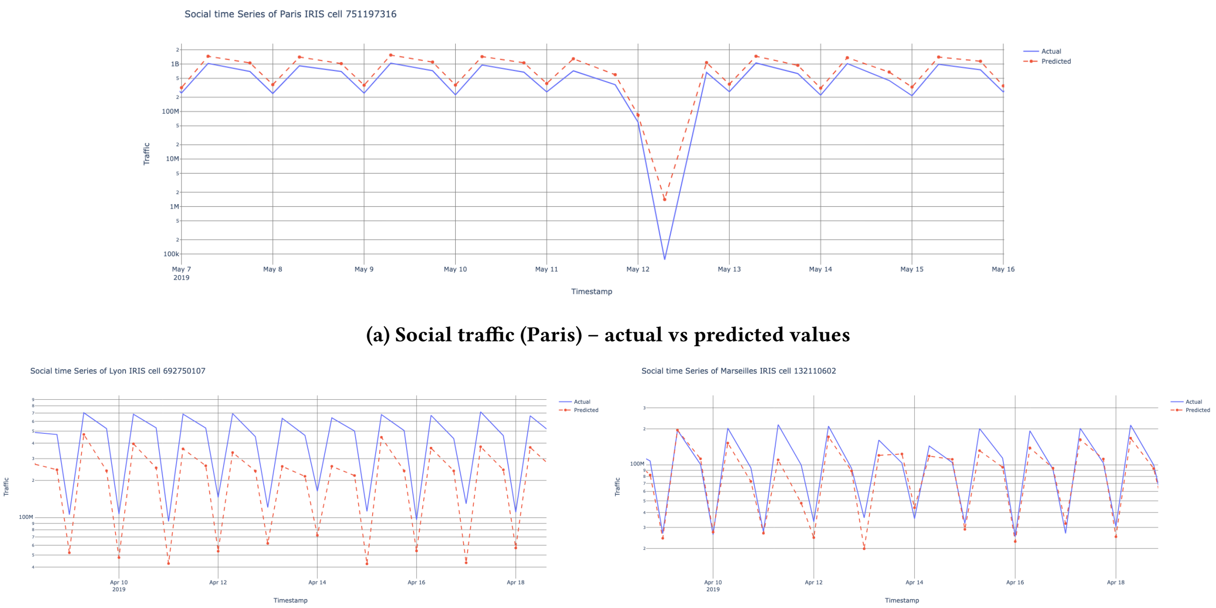
2 THE PROPOSED METHOD

In the following, we first formalize the problem we want to consider and then sketch the method we intend to address it. Let us consider a city partitioned into a set of tiles, where each tile represents a specific geographical area and is associated with a set of features that describe its characteristics, e.g., population density, road connectivity, land use, POI, satellite images, and more. Moreover, suppose that, for each tile, several historical time series are available, providing traffic volume data for N distinct groups of similar mobile services (e.g., work, leisure, gaming, and so on). The goal is to implement a set of generative models – one for each group of mobile services – that take in input the set of features describing each tile and generate the relative time series of traffic volume for each group at the tile level. More formally, let:

- C be a city divided into K tiles, denoted by $C = \{Tile_1, Tile_2, ..., Tile_K\}$.
- F be the set of features describing each tile, denoted by $F = \{Feature_1, Feature_2, ..., Feature_M\}$.
- For each group g of mobile services, let T^g be the historical time series data of traffic volume, represented as $T^g = \{T_1^g, T_2^g, ..., T_L^g\}$, where L is the number of time points in the historical data.

The problem is then to find a set of generative models $\{G^1, G^2, \ldots, G^N\}$, where $G^g : F \rightarrow T^g$ is the generative model for group g that maps the set of features of each tile to the corresponding time series of traffic volume for each group of mobile services. Each generative model G^g aims to learn the underlying patterns and relationships between the features of the tiles and the traffic volume for the g-th group of mobile services to output realistic and accurate synthetic time series data – in other words,

[1] See also the Nature article *Synthetic data could be better than real data*, available at: https://www.nature.com/articles/d41586-023-01445-8



(a) Social traffic (Paris) – actual vs predicted values

(b) Social traffic (Lyon) – actual vs predicted values

(c) Social traffic (Marseille) – actual vs predicted values

Figure 1: Comparison between actual and predicted traffic using an NN model for an IRIS cell (one for each city) based on the temporally aggregated social traffic datasets in the *3 time-slots* configuration.

the output should exhibit the temporal dynamics and fluctuations of the traffic volume observed in real data.

The success of the model G^g is measured by evaluating the generated time series for each group against the respective historical data using appropriate metrics, i.e., mean squared error (MSE), and mean absolute error (MAE).

Given the semantic information and the traffic volume historical time series associated with each city tile, to solve the problem, we propose a method that can generate synthetic mobile traffic information.

3 EXPERIMENTAL RESULTS

We considered as external additional data sources the INSEE dataset[2] and OpenStreetMap[3]. From INSEE, we selected 204 demographic indicators, standards of living, housing, and electricity supply. We downloaded from OSM map feature data of categories: amenity, aeroway, healthcare, historic, landuse, office, public_transport, railway, shop, tourism, leisure, place, and highway concerned Paris, Lyon, and Marseille. We selected the Social and Gaming activity traffic.

We employed deep learning and machine learning methods, specifically DNN and XGBoost. For the two methods, we aggregated data along the temporal dimension: hourly, three slots in a day (morning, afternoon, night), average weekday, and average weekday with three temporal slots. An illustration of these models' application to traffic data is depicted in Figure 1. Specifically, Figure 1 displays the temporal comparison between actual traffic and predicted traffic for the social category in three cities: Paris (a), Lyon (b), and Marseille (c). In this instance, traffic data is aggregated in 8-hour slots per day, resulting in three data points representing each day, and only a 10-day window is presented.

From these experiments, it is evident that the models can capture certain patterns and generalize to some extent. They are relatively close to real values and generate more temporally detailed data, although they may miss anomalous peaks. These findings are consistent across all our conducted experiments, and while they are preliminary, they show promise for further expansion of this research direction.

REFERENCES

[1] Z. Fang, Y. Du, X. Zhu, D. Hu, L. Chen, Y. Gao, and C. S. Jensen. Spatio-temporal trajectory similarity learning in road networks. In *28th ACM SIGKDD*, 2022.

[2] T.-Y. Fu and W.-C. Lee. Trembr: Exploring road networks for trajectory representation learning. *ACM (TIST)*, 11(1):1–25, 2020.

[3] X. Li, K. Zhao, G. Cong, C. S. Jensen, and W. Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th ICDE*, pages 617–628. IEEE, 2018.

[4] G. Mauro, M. Luca, A. Longa, B. Lepri, and L. Pappalardo. Generating mobility networks with generative adversarial networks. *EPJ Data Sci.*, 11(1):58, 2022.

[5] X. Wang, X. Liu, Z. Lu, and H. Yang. Large scale GPS trajectory generation using map based on two stage gan. *Journal of Data Science*, 19(1):126–141, 2021.

[6] G. Xiong, Z. Li, M. Zhao, Y. Zhang, Q. Miao, Y. Lv, and F.-Y. Wang. Trajsgan: A semantic-guiding adversarial network for urban trajectory generation. *IEEE Transactions on Computational Social Systems*, 2023.

[7] Z. Zong, J. Feng, K. Liu, H. Shi, and Y. Li. Deepdpm: Dynamic population mapping via deep neural network. In *Proc. of the 31st AAAI conf. on Artificial Intelligence*, volume 33, pages 1294–1301, 2019.

[2] https://www.data.gouv.fr/fr/datasets/1526-variables-regroupees-en-19-classes-surles-49461-iris-de-france/

[3] http://www.openstreetmap.org

Data challenge

Posters

Spatio-Temporal Analysis of Mobile Service Consumption for Social Signature Clustering

Z. Hu[1], C. Li[1,2], V. Gauthier[1], M. Nunez-del-Prado[3], H. Moungla[1,2]
[1] SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris
[2] LIPADE, University of Paris-Cité
[3] The World Bank
{zhaobo.hu, chuan.li, vincent.gauthier, hassine.moungla}@telecom-sudparis.eu
mnunezdelpradoco@worldbank.org

Mobile phone metadata is now heavily used to large scale extract socio-economic activity metrics for cities or regions. Properly using this data can provide unique insights into downstream tasks and business value. We propose a novel scalable method for clustering urban areas based on the spatio-temporal characteristics of mobile traffic data. The development of Deep Learning techniques makes time series deep clustering feasible. Our approach utilizes deep learning and meta-learning methods, including RNN-based variational auto-encoder and Graph Neural Networks (GNN), the former captures temporal information, and the latter offers a spatial correlation between neighboring regions, respectively. Moreover, meta-learning uses the additional geographical position embedding to generate neuron parameters for the Neural Networks described above to avoid the shared RNN over different regions. In this way, the resulting latent embeddings are clustered eventually. Furthermore, the model's generalization must be guaranteed, and the model can be easily transferred to other cities.

Introduction

The rapid expansion of urbanization worldwide poses unique challenges (mobility, climate change, risk assessment, inequality) that require new technical approaches to study urban dynamics. To address these challenges, new data sources, such as mobile phone metadata, have catalyzed a new, cost-effective approach to creating novel socio-economic activity metrics. In this context, Ucar *et al.* [1] correlates the consumption of mobile services with income, educational attainment, and inequality. Furno *et al.* [2] use mobile phone traffic to create a social signature of different urban areas. Khodabandelou *et al.* [3] develop a new approach for estimating population density in urban regions using mobile network metadata. Bachir *et al.* [4] to infer dynamic origin-destination flows by transport mode. Finally, Vilella *et al.* [5] extract spatial socio-economic characteristics of cities based on online news consumption.

In this paper, we propose a novel method for clustering urban areas based on the spatio-temporal characteristics of mobile traffic in a given region. Our approach takes advantage of novel deep learning methods to deal with the high-dimensional (both in time and space) traffic dataset provided in the NetMob23 challenge [6]. Since the data provided for the challenge achieves remarkable spatial accuracy, mapping to over 870,000 high-resolution regular grids, and knowing that current methods have difficulty capturing the complexity of mixing spatial and temporal relationships. We developed a novel approach that is scalable and allows us to capture the diversity of traffic in time, space, and across different applications.

First, our framework uses representation learning based on the variational auto-encoder [7] model, coupled with recurrent neural network architecture, to create a representative embedding of each time series of each application over the city grid. Secondly, we add spatial information to the time series embeddings to characterize the position of each grid cell relative to others through the following methods:

method 1. Use geographical positional encoding to consider the spatial relationship between different grid regions [8]. The encoding provides rich geographical information to guide the learning architecture to generate dynamic weight and bias parameters for neural networks (*i.e.*, meta-learning architecture). In other words, each grid should have a set of unique parameters for neural networks based on grids attribute(*i.e.*, GPS Location) and grids feature(*i.e.*, apps consumption) [9].

method 2. Use a Graph Convolutional Network layer [10] to account for the spatial relationship between different grid regions. Beyond geodesic distance, encoding the spatial dimension in a graph structure could allow us to define the relationships between spaces more broadly than just geodesic distance (*e.g.*, OD matrix, land cover similarity, activity similarity).

Finally, time series embedding of each application proposed in the dataset contextualized with spatial information (with either method 1 or 2) are clustered using a Gaussian Mixture Model (GMM).

We believe this approach could pave the way for a broader set of applications in these areas, such as a baseline for anomaly detection in traffic patterns, behavior change detection in specific geographic regions, land use classification, and metrics for comparing cities. In addition, our model is trained to learn the traffic patterns and spatial dependencies, so our model trained in one city could be easily transferred to other cities.

References

[1] I Ucar, M Gramaglia, M Fiore, et al. "News or social media? Socio-economic divide of mobile service consumption". In: *Journal of The Royal Society Interface* 18.185 (2021), p. 20210350. DOI: 10.1098/rsif.2021.0350.

[2] A Furno, M Fiore, R Stanica, et al. "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas". In: *IEEE Transactions on Mobile Computing* 16.10 (2016), pp. 2682–2696.

[3] G Khodabandelou, V Gauthier, M Fiore, and MA El-Yacoubi. "Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata". In: *IEEE Transactions on Mobile Computing* 18.9 (Sept. 2019), pp. 2034–2047. DOI: 10.1109/tmc.2018.2871156.

[4] D Bachir, G Khodabandelou, V Gauthier, et al. "Inferring dynamic origin-destination flows by transport mode using mobile phone data". In: *Transportation Research Part C: Emerging Technologies* 101 (Apr. 2019), pp. 254–275. DOI: 10.1016/j.trc.2019.02.013.

[5] S Vilella, D Paolotti, G Ruffo, and L Ferres. "News and the city: understanding online press consumption patterns through mobile data". In: *EPJ Data Science* 9 (Apr. 2020). DOI: 10.1140/epjds/s13688-020-00228-9.

[6] OE Martínez-Durive, S Mishra, C Ziemlicki, et al. *The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography.* 2023. arXiv: 2305.06933.

[7] DP Kingma and M Welling. "Auto-Encoding Variational Bayes". In: *Proceedings of 2nd The International Conference on Learning Representations.* ICLR 2014. Banff, AB, Canada, Apr. 2014. arXiv: 1312.6114.

[8] G Mai, K Janowicz, B Yan, et al. "Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells". In: *Proceedings of 8th The International Conference on Learning Representations.* ICLR 2020. Addis Ababa, Ethiopia, Apr. 2020. arXiv: 2003.00824.

[9] Z Pan, W Zhang, Y Liang, et al. "Spatio-Temporal Meta Learning for Urban Traffic Prediction". In: *IEEE Transactions on Knowledge and Data Engineering* 34.3 (2022), pp. 1462–1476. DOI: 10.1109/TKDE.2020.2995855.

[10] TN Kipf and M Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *Proceedings of 5th The International Conference on Learning Representations.* ICLR 2017. Toulon, France, Apr. 2017. arXiv: 1609.02907.

Detection of anomalous spatio-temporal patterns of app traffic in response to catastrophic events

Sofia Medina[*], Shazia'Ayn Babul[*], Rohit Sahasrabuddhe[*], John Pogué-Biyong[*], Timothy LaRock[*], Nicola Pedreschi[*], and Renaud Lambiotte[*]

[*] Mathematical Institute, University of Oxford, UK

September 22, 2023

Understanding how information propagates during and after catastrophic events is an active field of investigation [1, 6, 3, 5]. Social media and online resources have been used to track the length and intensity of responses to breaking news stories [5, 2], or to categorize types of responses from the population [3]. In this work, we analyze mobile phone app data to understand how the temporal and spatial usage pattern of different applications is perturbed in the aftermath of an unprecedented event. To this end, we use the NetMob2023 Data Challenge dataset [4], which provides mobile phone applications traffic volume data for several cities in France at a spatial resolution of $100m^2$ and a time resolution of 15 minutes for a time period ranging from March to May 2019. We analyze the spread of information before, during, and after the catastrophic Notre-Dame fire using volume of data uploaded and downloaded to different mobile applications as a proxy of information transfer dynamics. The methods we develop can be extended to other contexts to characterize mobile phone user response to unplanned catastrophic events, giving insight into how information spreads during a catastrophe in both time and space.

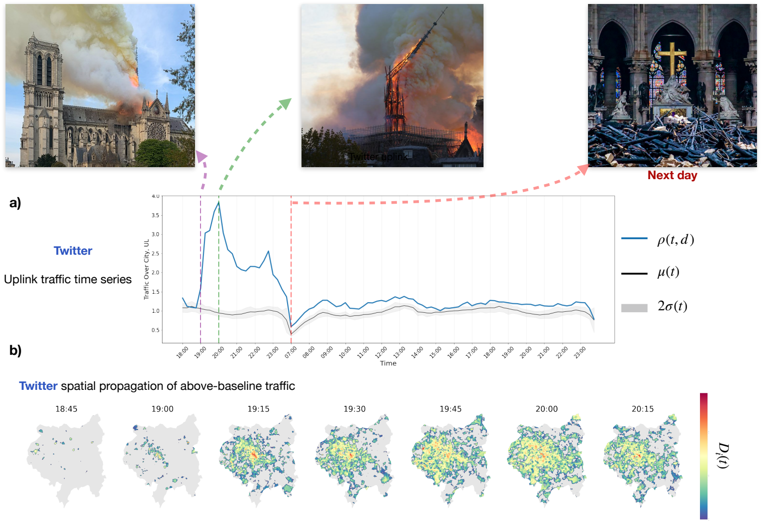


Figure 1: **Graphical introduction** A) The time series evolution of the application Twitter over a time period ranging from hour 18 to end of day on the day of the fire and hour 7 to end of day the day following the fire. The traffic on the day of is given in blue, with the mean traffic of previous weeks and two standard deviations from this mean also shown. B) The spatial evolution of application traffic of Twitter is shown in time on the outline of the city of Paris.

We present two kinds of analysis, limiting to data from Paris, Marseilles, Lyon, Montpellier, Rennes, and Strasbourg. First, we spatially aggregate data within each city and study the timeseries of app traffic per city for a subset of the available apps. We then use a simple anomaly detection method to determine whether traffic volume for each app spiked after the Notre-Dame Fire, when those spikes occurred, and for how long, exploring the differences between uplink and downlink. We use information about these spikes in traffic volume to cluster apps based on how they were used during the fire. Finally, we analyze how spikes were distributed spatially throughout the city of Paris over time as news of the fire spread and mobile phone users sought more information through social media, apps, and streaming services. We find that applications regarding social media, messaging, and video streaming generally experience spikes in traffic during and after the fire of Notre-Dame. These behaviors can be categorized by 5 clusters. Within Paris, these clusters represent apps with large spikes and sustained spiking behavior, mild spikes and interest on the day of the event, high spikes the first day with late spikes the next day, brief spikes of large amplitude, and small perturbations of spikes that are not sustained. Interestingly, the behavior of these applications does not divide cleanly into the general function of the app, such as social media or video streaming. We also investigate the evolution in time, as the fire is happening, of the spatial distribution of anomalous traffic of Twitter in the city of Paris. Our analyses show how the first areas to spike are the immediate surroundings of the cathedral; the above-baseline activity then spreads throughout the rest of the city in a manner that is surprisingly correlated with geographical distance from the catastrophe, given digital media is shared online and is available regardless of distance.

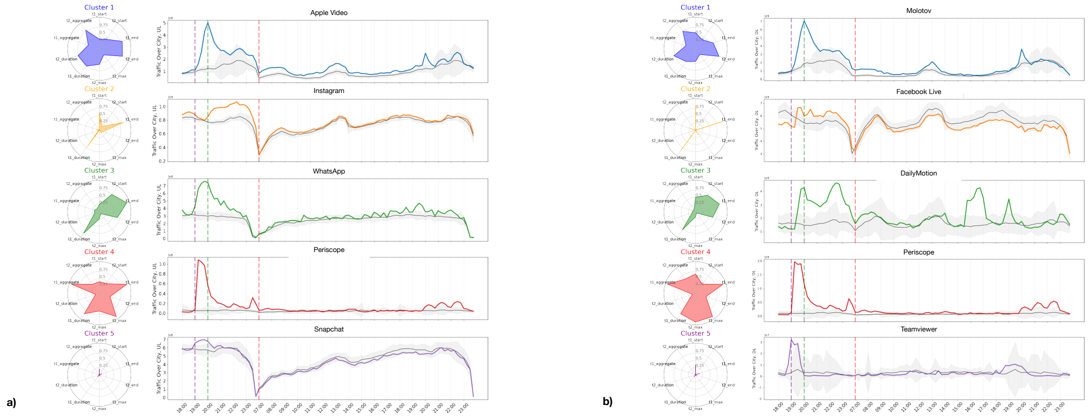


Figure 2: **Clusters and Representative apps in Paris, uplink and downlink** Uplink application activity clusters, as well as representative time series for each cluster are shown in a). Downlink clusters and representative time series are shown in b). Radar plots of each cluster show the normalised value of different features for the applications assigned to each cluster. These features include start and end times of traffic spikes on the day of the fire and day after the fire , as well as the values of a defined distance metric on the spikes.

References

[1] Ruth Garcia-Gavilanes, Milena Tsvetkova, and Taha Yasseri. Dynamics and biases of online attention: the case of aircraft crashes. *ROYAL SOCIETY OPEN SCIENCE*, 3(10), OCT 2016.

[2] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. Extracting event-related information from article updates in wikipedia. In *European Conference on Information Retrieval*, 2013.

[3] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 251–260, New York, NY, USA, 2012. Association for Computing Machinery.

[4] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

[5] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using twitter and wikipedia. In *Sigir 2012 workshop on time-aware information access*, pages 16–76. Citeseer, 2012.

[6] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

Modelling daily mobility using mobile data traffic at fine spatiotemporal scale

Panayotis Christidis[1], María Vega- Gonzalo, Miklos Radics

European Commission, Joint Research Centre, Seville Spain

Abstract

We applied a data-driven approach that explores the usability of the NetMob 2023 dataset (Martínez-Durive et al., 2023) in modelling mobility patterns within an urban context. We applied methods for the analysis used in the past on similar datasets provided- among others- by the same mobile phone oprerator (Christidis et al., 2022) in order to test how this new dataset can combined with demographic data in order to derive trip generation rates, in the line of what has been proposed by (Bwambale et al., 2019).

We identified a highly suitable external dataset that can be used as the basis for the analysis, namely the ENACT dataset (Batista e Silva et al., 2020), a 1 km x 1km grid that provides estimates of the day and night population across Europe. The night population is based on official statistics, so can be considered the ground truth. The day population corresponds to the distribution of the population across the grid and was estimated based on land use, land cover and Point-of-Interest data. Both population estimates were developed independently of the information included in the NetMob 2023 dataset and can therefore be considered as – in principle- suitable for modelling purposes. Figures 1 and 2 visualize the distribution of the night and day population in the Greater Paris area that is covered in the NetMob dataset. While a certain mobility in the population is already visible, filtering for the areas that have a significant increase or decrease between night and day suggests that there is a strong trend for mobility between the outskirts of Paris to its centre and main activity areas (Figures 3 and 4). In addition, it is worth mentioning that the total night population in the area covered is 7 million, which increases to 8 million during the day, an observation that indicates that there is a significant movement of population from/to zones outside the covered area. The population grids are available for all 20 cities in NetMob and in practically all cases reveal a high level of mobility. Our approach consisted of developing 3 sets of models that predict population per 100m x 100m cell using the data for each of the 68 services in the dataset: night, day, day (using night population as input too). The NetMob data we used is the average per weekday and time slot, a level of aggregation that leads to satisfactory precision levels, while still allowing a manageable dataset size.

Fig. 1: Distribution of population in Paris during the night

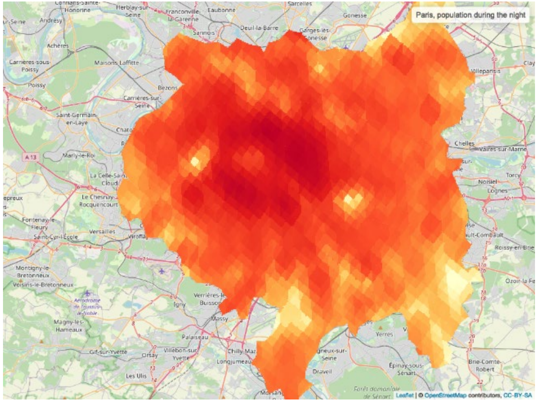
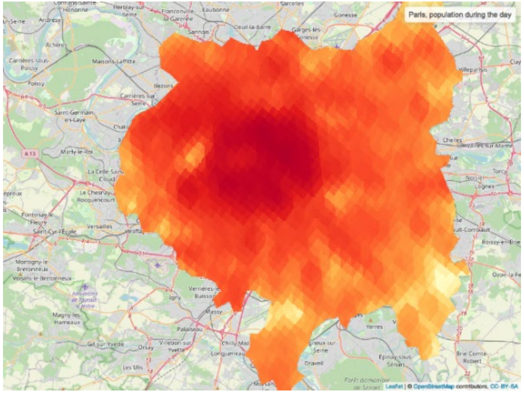


Fig. 2: Distribution of population in Paris during the day



source: ENACT dataset, converted by the authors to 100m x 100m grid

[1] Corresponding author, e-mail: Panayotis.Christidis@ec.europa.eu

Fig. 3: Areas with significant daily population increase

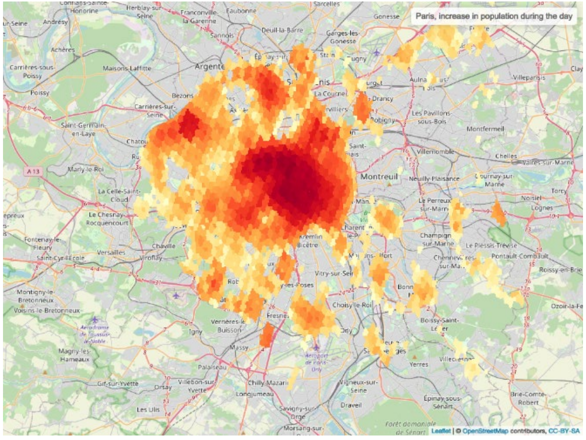
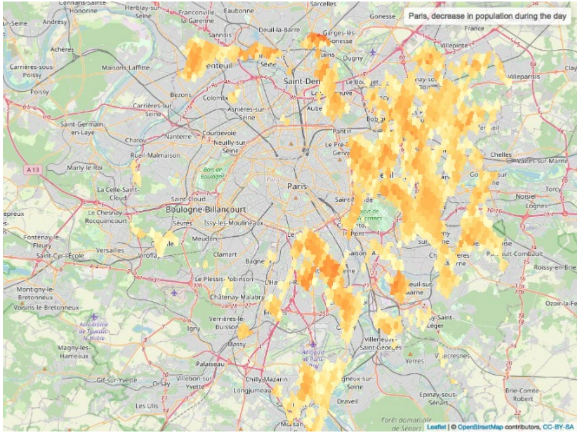

Paris, increase in population during the day
Argenteuil
Montreuil
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Fig. 4: Areas with significant daily population decrease


Paris, decrease in population during the day
Paris
Boulogne-Billancourt
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Calculating the mean 1-hour time slot total for each weekday along the 11 weeks of data available already allows several observations concerning the spatiotemporal patterns of each application use (Figures 5 and 6). The XGBoost model that we applied allows the identification of the main variables that can be used for the prediction of population in a cell (Figure 7). The results so far are highly promising.

Fig. 5: Concentration of Uber use, Mondays 9:00-10:00

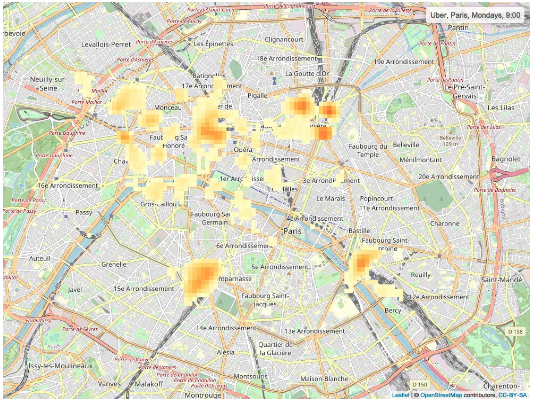

Uber, Paris, Mondays, 9:00
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Fig. 6: Concentration of Uber use, Mondays 19:00-20:00

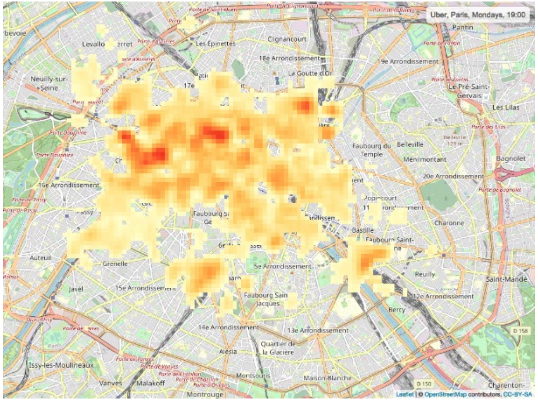
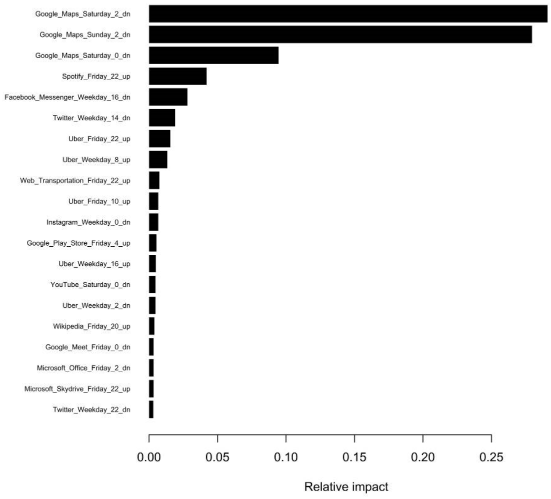

Uber, Paris, Mondays, 19:00
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Fig. 7: Example of model feature importance



Main references

Batista e Silva et al., 2020. Uncovering temporal changes in Europe's population density patterns using a data fusion approach. Nature Communications 11, 4631. https://doi.org/10.1038/s41467-020-18344-5

Bwambale et al., 2019. Modelling trip generation using mobile phone data: A latent demographics approach. J. Transp. Geogr. 76, 276–286. https://doi.org/10.1016/j.jtrangeo.2017.08.020

Christidis et al., 2022. Regional mobility during the Covid-19 pandemic: Analysis of trends and repercussions using mobile phones data across the EU. Case Studies on Transport Policy 10, 257–268. https://doi.org/10.1016/j.cstp.2021.12.007

Martínez-Durive et al., 2023. The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography.

SNN-based Federated Learning for Network Traffic Classification

Giovanni Perin*, Alessandro Buratto*, Riccardo Mazzieri*, Jacopo Pegoraro*, Francesca Meneghello*,
Leonardo Badia*, and Michele Rossi*‡
* Department of Information Engineering, University of Padova, Italy
‡ Department of Mathematics “Tullio Levi-Civita”, University of Padova, Italy
e-mail: {giovanni.perin.1, jacopo.pegoraro, francesca.meneghello.1, leonardo.badia, michele.rossi} @unipd.it,
{alessandro.buratto.1, riccardo.mazzieri} @phd.unipd.it

Abstract—**The recent growth of mobile traffic and data availability makes network practitioners face a new important challenge: how to find a new optimized balance between quality of service and energy efficiency of the network infrastructure. Artificial intelligence-aided traffic analysis is a useful tool to optimize the management of the network. Edge computing is a candidate method to solve the scalability issue in a decentralized way, and the most common form of distributed machine learning is federated learning. In this abstract, we present preliminary studies on federated mobile traffic classification using spiking neural networks. They are chosen as the most popular form of neuromorphic computing and belong to a new paradigm of brain-inspired neural networks that have shown to be highly performing in terms of energy efficiency concerning traditional models. Our results show that the model capacity of spiking neural networks is still inferior to state-of-the-art models, both in centralized and federated settings.**

Index Terms—**Federated learning, edge computing, neuromorphic computing, spiking neural networks, network traffic classification**

I. Introduction

In recent years we faced an increasing trend in the amount of data traffic of mobile applications, insofar as it has doubled in the last two years. Video streaming services generate the majority of the traffic volume, with about 80% of the total traffic, but other applications such as social networking and file downloads still use a significant share [1]. Motivated by this increase in the traffic volume, practitioners need to make the network infrastructure resource- and energy-efficient. One way to achieve this goal is to provide multi-access edge computing (MEC) facilities with pervasive and distributed artificial intelligence (AI) and machine learning (ML) to enhance the traditional platform management algorithms. Federated learning (FL) [2] is becoming a popular approach to training ML models in a distributed way via periodic averaging of locally learned parameters. Specifically, the traditional setting is composed of a set of N clients (agents) each owning their local and private training dataset, not to be shared with other agents. Clients train a private local version of a common model and periodically send the local version to the central server, which averages the parameters obtained from all the clients and returns an updated global version of the model to the clients. This procedure is typically performed on a star topology but can also be carried out on a generic mesh in a fully decentralized way, where nodes communicate their models in a peer-to-peer way with their 1-hop neighbors. However, energy efficiency problems arise when considering such a way of training since CPUs and GPUs are energy-hungry. A possible solution to this is the use of neuromorphic computing [3], [4], which is a recent neural computing paradigm more closely inspired by how the human brain works.

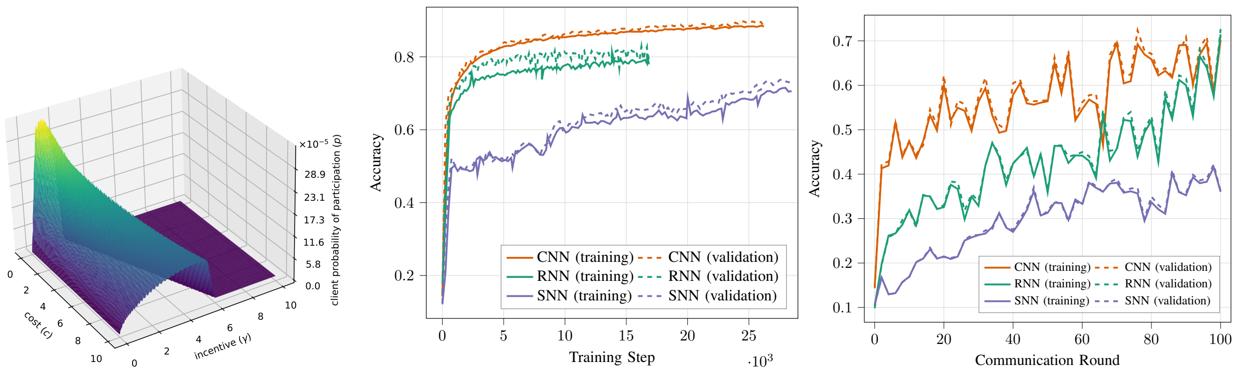
In this work, we explore the task of traffic classification of 8 popular mobile applications using the `NetMob23` dataset [5] and referring to the city of Paris. We first assess the feasibility of the task with a centralized training of a subset of the dataset, and then use a realistic FL approach where the mobile traffic is assigned to the closest long-term evolution (LTE) base station (BS) using a minimum distance criterion. The BSs locations are taken from the repository `OpenCelliD` [6]. Spiking neural networks (SNNs) are considered as a popular candidate of neuromorphic computing, since they have the advantage of performing extremely sparse computations, resulting in a much higher energy efficiency. This interesting feature makes the neuromorphic paradigm attractive for FL [7]–[10] and has also caught the interest of deep learning researchers as a whole.

II. Data preprocessing

We use the `NetMob23` traffic traces for the city of Paris, restricted to 8 popular services: Amazon Web Services, Instagram, LinkedIn, Netflix, Spotify, Twitter, Wikipedia, and Whatsapp. To simulate a real network setting, we use the OpenCelliD dataset [6] to aggregate the traffic based on the locations of the LTE BSs, using a minimum-distance criterion. Then, the traffic traces are split based on the collection date: we select the first two months for training and the rest is equally divided into validation and test sets. The uplink and downlink traces of each day are cropped into sequences of 4 hours duration, with 15 minutes granularity, and standardized. Finally, we add the time of the day information along with uplink and downlink traffic values as a third input feature for the classifier.

III. Spiking neural network classifier

The SNN architecture consists of a feed-forward network of Leaky Integrate and Fire (LIF) neurons [3] with 6 layers. Each neuron acts as a leaky integrator of an input signal and fires an output *spike*, i.e., a discrete impulse, whenever its state exceeds a pre-defined threshold. The input to the first layer



(a) Client sampling probabilities at NE

(b) Centralized accuracy

(c) Federated learning accuracy

Fig. 1. Main preliminary results of this abstract: NEs of the sampling policy and training accuracies reached by the SNN model and the benchmarks in centralized and federated settings.

is the 4 hours long sequence of uplink and downlink traffic, and the time of the day information. The final classification is obtained through an output layer with 8 neurons: the predicted class corresponds to the index of the neuron that fires the most spikes while processing the input window.

IV. Client sampling policy

We analytically derived the policy for sampling clients through a solution to the Nash Equilibrium (NE) of a static game of complete information. Each client chooses its own participation probability p_i in order to maximize its payoff function

$$u_i = -\mathbb{E}[\mathcal{D}] - cp_i - \gamma\mathbb{E}[\delta_i], \tag{1}$$

which consists of the expected duration of the process $\mathbb{E}[\mathcal{D}]$, a cost factor c and an incentive based on Age of Information $\mathbb{E}[\delta_i]$ weighted by γ. Due to the symmetries of the FL scenario, all the clients at the NE will have the same participation probability p. Fig. 1a shows the NE solution for a range of values of c and γ. The incentive is necessary to increase the participation of clients, but there is a threshold dependent on the cost factor, over which the clients will fall back to a non-collaborative strategy.

V. Results

The SNN is compared in both a centralized and federated fashion with two benchmark models representing the state-of-the-art for time-series-related ML tasks, i.e., a CNN and an RNN. Specifically, the CNN is composed of an inception-inspired block while the RNN of three gated recurrent units.

Fig. 1b shows the accuracy of the models in the centralized setting, where 150 BSs were used and trained for more than 25^3 gradient descent steps (batch size 256). The plot shows how the benchmark models learn in a smooth way reaching an accuracy of 88% and 81% for the CNN and RNN, respectively. Training the SNN is a slower process and the model accuracy after the training time given is 72%.

In Fig. 1c the performance of the SNN and the benchmark models is shown as a function of the communication rounds for the FL setting. The number of communication rounds is limited to 100 due to the lack of time and computational capacity of the hardware available for simulations. It can be seen that all the models would need more time to converge. However, from this preliminary result, we observe that the CNN is the fastest model to reach an acceptable accuracy as it reaches 60% at round 20, while the RNN takes 88 rounds to reach the same accuracy. It is confirmed that the proposed SNN model is the slowest to find a minimizer of the loss function and the model capacity seems to be overall inferior, as the reached accuracy is about 40%.

References

[1] Ericsson, “Ericsson mobility report,” June 2023. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[3] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, “Training spiking neural networks using lessons from deep learning,” *Proceedings of the IEEE*, 2023.

[4] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, “Physics for neuromorphic computing,” *Nature Reviews Physics*, vol. 2, no. 9, pp. 499–510, 2020.

[5] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, “The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography,” 2023.

[6] Unwiredlabs. (2023) Opencellid. [Online]. Available: https://opencellid.org/

[7] N. Skatchkovsky, H. Jang, and O. Simeone, “Federated Neuromorphic Learning of Spiking Neural Networks for Low-Power Edge Intelligence,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8524–8528.

[8] ——, “Spiking Neural Networks—Part III: Neuromorphic Communications,” *IEEE Communications Letters*, vol. 25, no. 6, pp. 1746–1750, 2021.

[9] K. Xie, Z. Zhang, B. Li, J. Kang, D. Niyato, S. Xie, and Y. Wu, “Efficient Federated Learning With Spike Neural Networks for Traffic Sign Recognition,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9980–9992, 2022.

[10] Y. Venkatesha, Y. Kim, L. Tassiulas, and P. Panda, “Federated Learning With Spiking Neural Networks,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 6183–6194, 2021.

Identifying socio-demographic traits associated to different lifestyles: an application to extended workload during out-of-office hours

Lorenzo Dall'Amico[1], Michele Tizzani[1], Nicolò Gozzi[1], Mattia Mazzoli[1,*], and Anna Sapienza[1,2,*]

[1]ISI Foundation, Turin, Italy
[2]Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark

Mobile apps have permeated nearly every aspect of our lives, including our health, social relationships, and mobility. Particularly noteworthy is their impact on our daily work routines. While they can enhance and streamline our work-related activities, they can also have a negative impact on our work-life balance, by increasing work-induced tiredness, burn-out rates, and more in general by reducing the quality of life outside the workplace [1, 2, 3]. Due to limited access to detailed mobile app use data, previous studies are mostly survey-based and involve a small number of participants and working contexts. Hence, the association between work-life balance and socio-demographic traits remains largely unexplored.

In this study, we narrow down this gap by leveraging high-resolution spatiotemporal mobile traffic data collected for 77 consecutive days in 20 metropolitan areas in France in 2019 [4]. The dataset includes total traffic generated by a set of mobile phone apps within 100 × 100 meters tiles every 15 minutes. By exploring the use of work-related apps throughout the week, we study the association between neighborhood socio-demographic traits and work activity outside of traditional working hours, which potentially represents struggles with work-life balance. We only consider traffic coming from apps related to working behaviors, such as cloud services, online editors, videoconferencing, and e-mailing services, ending up with 13 apps. We spatially aggregate the app traffic at the IRIS level (defined by INSEE), including ~ 2000 individuals each and their aggregated socio-demographic features. We then represent the spatiotemporal users' activity at IRIS scale as a matrix $X \in \mathbb{R}^{N \times T}$, where N represents the number of IRIS units and $T = 168$ is the number of hours in a week. Each element $X_{i,j}$ of this matrix is the median (over the 11 weeks of data) of total traffic generated by work-related apps in IRIS unit i during hour j of the week.

	Metz	Tours	Nantes	Marseille	Toulouse	Lille	Lyon	Bordeaux	Paris	Saint-Etienne
RF	0.45	**0.43**	**0.43**	0.41	**0.38**	**0.34**	**0.31**	**0.31**	**0.27**	**0.25**
OLS	**0.87**	0.00	0.39	**0.44**	0.35	0.33	0.13	0.08	0.08	0.01
RR	0.46	0.12	0.20	0.25	0.17	0.15	0.01	0.10	0.08	0.02

	Mans	Cler-Fer	Dijon	Orleans	Nice	Grenoble	Rennes	Nancy	Strasbourg	Montpellier
RF	0.22	0.20	**0.19**	**0.16**	0.12	0.08	0.08	**0.06**	**0.02**	0.00
OLS	**0.67**	**0.33**	0.06	0.11	0.11	0.21	0.09	0.01	0.01	**0.17**
RR	0.06	0.22	0.05	0.13	0.12	**0.32**	**0.19**	0.04	0.01	0.08

Table 1: **Prediction accuracy.** Out of sample R-squared for the three models in the 20 cities under study.

We perform Non-Negative Matrix Factorization (NMF) on X [5], choosing two components, which yield 82% accuracy. We find two "typical" temporal profiles (α and β), that can be associated with work (peaking during office hours) and off-work activity (peaking in early morning, late afternoon, and weekends). The NMF approximates each IRIS unit n as $X_n \approx C_{\alpha,n}\alpha + C_{\beta,n}\beta$, where $C_{\alpha,n}$, and $C_{\beta,n}$ weigh the relative contribution of the two typical temporal profiles to each IRIS traffic time series. Thus, the coefficient $C_{\beta,n}$ can be used as a proxy for the level of working activity outside of office hours for IRIS unit n.

Regions with high levels of off-work activity are not uniformly distributed over the city. This observation stimulates the investigation of the role of spatial, social, demographic, and economic features in explaining this heterogeneity. We use three regression models: an Ordinary Least Squares (OLS), a Ridge, and a Random Forest (RF) regression. The target variable is the C_β coefficient of IRIS areas and socio-economic features are the covariates. We train the models on the 80% of each city available area independently and report the out-of-sample R-squared in Table 1. The RF generally performs best at predicting off-work activity in our data, with values reaching $R^2 \sim 0.4$. These results suggest that socioeconomic factors could play a role in shaping work-related digital activities outside of typical working hours, but activity in some parts of the city is not generalizable from

$*$ Contributed equally as last author
Correspondence should be addressed to anna.sapienza@isi.it

the rest of the urban area. This may be due to the city's size but also to the local transportation network, structure, and shape. In Figure 1, we show the RF covariates shap value, on the left for two cities with high prediction accuracy, Nantes and Marseille, and two cities on the right with bad prediction accuracy, Strasbourg and Montpellier. Features' importance varies among cities. Income has a high positive impact on off-work activity in Nantes and Marseille (high accuracy), while it has an opposite and mixed impact in Strasbourg and Montpellier (low accuracy). Inequality is associated with high off-work behavior in Nantes, while the opposite happens in Strasbourg and no clear effect is present for Marseille and Montpellier. Internet speed spatial heterogeneity may bias traffic data, as this is the case for Montpellier and Marseille. Places dedicated to teaching services relate to low off-working behavior; tourism offices play little to no role in feature importance, while the rest of public services show no clear impact. Living alone and age composition impact differently across cities, highlighting that off-work activity is not directly associated with places inhabited by any specific age classes or household composition. Intuitively, the primary sector is the least important feature in all cities. While the tertiary sector encodes the most suitable jobs for off-work activity, this feature has controversial impacts, indicating that not all jobs within the tertiary sector relate to high off-work activity and that tertiary sector jobs may highly differ across cities.

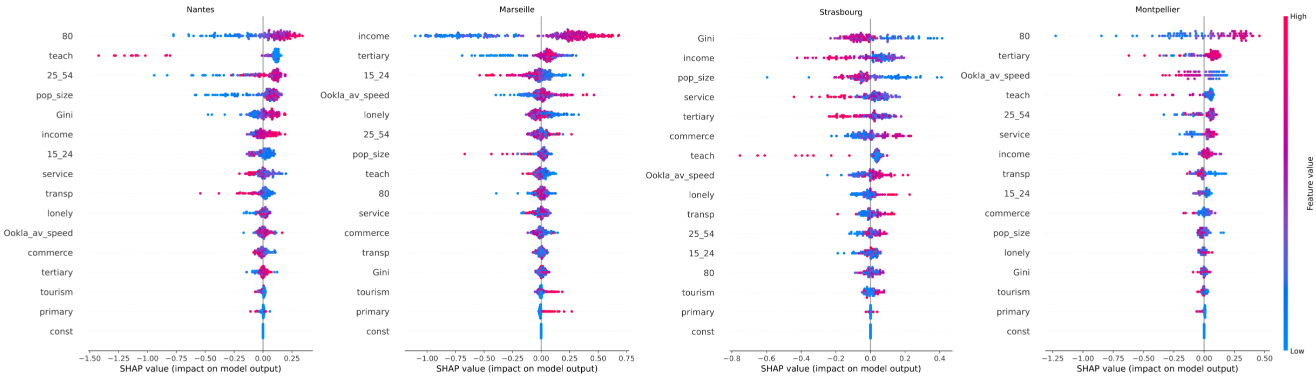


Figure 1: **Shap values in four cities.** Shap values for the 15 covariates employed in the RF regression in four key cities. From left to right, Nantes, Marseille, Strasbourg, Montpellier, in order of prediction accuracy. Features are ordered top-bottom in order of importance.

In conclusion, we find that local off-work activity is the result of a complex interplay of features that goes beyond age classes, household size, work sector, and economic disparity. While land use mixing allows workplaces in urban areas to be spatially correlated to population density, their location with respect to socio-demographic and economic traits may vary across cities of different sizes. As a result, there is no one-fits-all model to predict spatial off-work activity in cities from local socio-demographics. Some features like income, inequality, and tertiary sector exhibit opposite impacts across cities despite having high importance. This switchover effect highlights how the off-work activity is not directly related to areas inhabited by specific groups, since the spatial allocation of workplaces responsible for off-work behavior may vary, highlighting the importance of land-use mixing. Note that these results draw no conclusions on the socio-demographic and economic traits of the users who exhibit off-work behavior, app traffic in our dataset is registered on the spot and disregards of the amount of users generating the signal. Users showing off-work activities may be operating either from home or from their workplaces, no general criteria can be drawn to define their location, and no information on users' location of residence was accessed. Our method highlights the importance of this data to identify urban areas exposed to extended workloads and hence at risk of developing higher burn-out rates and work-related stress.

[1] Ruben Cambier, Daantje Derks, and Peter Vlerick. Detachment from work: A diary study on telepressure, smartphone use and empathy. *Psychologica Belgica*, 59(1):227, 2019.

[2] Yue Lok Cheung, Miu Chi Lun, and Hai-Jiang Wang. Smartphone use after work mediates the link between organizational norm of connectivity and emotional exhaustion: Will workaholism make a difference? *Stress and Health*, 38(1):130–139, 2022.

[3] Daantje Derks and Arnold B Bakker. Smartphone use, work–home interference, and burnout: A diary study on the role of recovery. *Applied Psychology*, 63(3):411–440, 2014.

[4] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography. *arXiv preprint arXiv:2305.06933*, 2023.

[5] Talayeh Aledavood, Ilkka Kivimäki, Sune Lehmann, and Jari Saramäki. Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data. *Scientific reports*, 12(1):5544, 2022.

Are mobile phone applications data a good proxy for predicting the spatio-temporal evolution of peoples activities in urban areas ?

Etienne Côme, Paul de Nailly, Angelo Furno, Benoit Matet

September 22, 2023

Aggregate data on the use of certain mobile phone applications have already been used to predict and/or infer the socio-economic characteristics of the **resident population**, see for e.g., [1]. However, since the use of such and such applications depends on both the characteristics of the individuals and the activities they perform, it seems conceivable to use them to characterise the **population present** at a given time and place, both from a socio-demographic point of view and from the point of view of the activities they perform. However, this is not a trivial task, as there are a number of interdependent factors at play. In such a case, mobile phone application data could provide an in-depth understanding of the activity profiles of a city's territories, enabling a better planning of urban spaces, the identification of disparities, and even the comparison of the structure of cities. In this study, we want to determine whether it is possible to predict the share of certain activity profiles present in statistical areas (districts) covering some of France's major cities. The originality of our study is that we use information on the use of different mobile applications provided by the NetMob challenge to try to predict the share of different activities for each statistical zone and **different time bands**. As a first step, and to investigate the possibility of such an approach, we compare the obtained results with those of a baseline models that include some information about the characteristics of built environment and Points of Interest (PoI). In the following, we present the data used as well as the preprocessing steps that we conducted on the NetMob data together with the 2 others datasets that we considered. We then present the model used and the adopted experimental setup. The reported results and their discussion provide the first insights answering the mentioned research question.

Data and pre-processings The main activity profiles by districts in 18 French cities are derived from Mobiliscope [2], a geo-visualization tool that allows exploring the population present and the social mix in cities over the course of a 24-hour typical weekday. The data come from large-scale public surveys used to quantify and qualify the populations present at different times of the day and to study the evolution their transport practices. In France, these surveys are traditionally commissioned by local authorities every 10 years. This tool provides access to the types of activities present in districts during the 24 hours of a typical weekday (Monday-Friday). Five types of activities constitute our target data: *working, home, leisure, study* and *shopping.* We collected these survey results for eighteen cities among the twenty cities available in the NetMob challenge.

To compare the interest of using the NetMob data for answering our research question, we considered, as a baseline, OpenStreetMap data describing urban facilities for each district and city of the analyzed French territory. Ultimately, for each district, we consider: the number of middle and high schools, the number of universities, the number of leisure places (including restaurants, cafes, theatres, cinemas and others), the number of railway, tram and bus stations and the surface area of parks and retails zones. OSM data are normalized by the total surface area of districts.

Our study seeks to use information on the use of the various mobile applications offered by the NetMob challenge to ascertain their impact on the correct prediction, or otherwise, of typical attendance by district. As a result, two aggregations were required here:

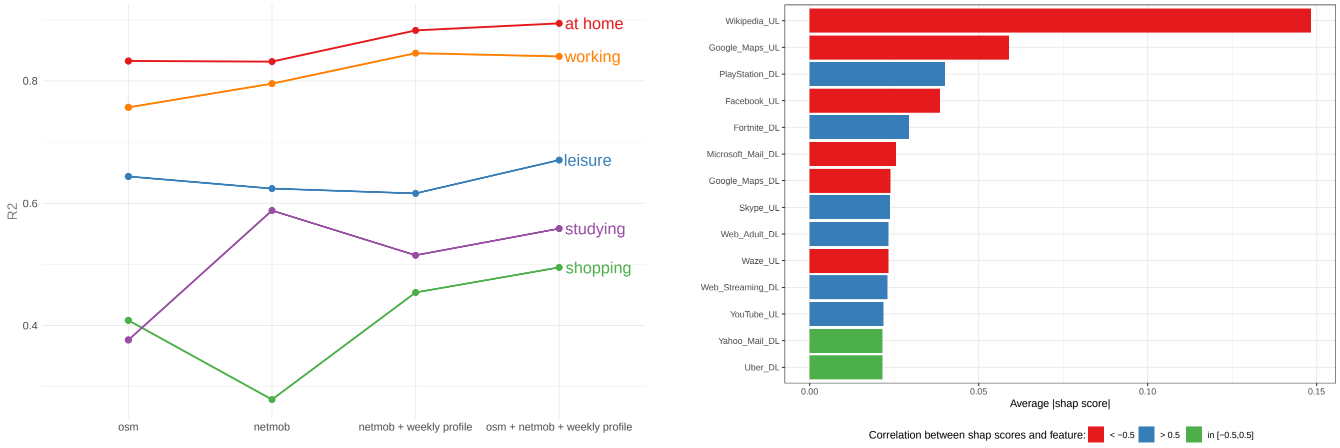
- A temporal aggregation for which we consider the median weekly profile of upload and download for each application at each tile of the NetMob dataset. We first filtered on non-holidays and non-bank holidays (major data anomalies listed in [3] were also removed) before computing median profiles.
- A spatial aggregation where we match the district areas from mobiliscope data and tiles from NetMob data. Here we link each district from the mobiliscope data with the area that covers it the most from the NetMob data (must be over 50% coverage). The tiles profiles were then aggregated at the district levels by summing them.

Together, these two preprocessing steps provide a typical weekly profile for each district, and for each application download and upload volumes.

Method Considering a set of input data built from OSM and / or NetMob data weekly profiles, we aim at predicting the shares of the activity types within the districts at each hour. To do so, we rely on Extreme gradient boosting (XGBoost) model. The learning was performed by minimizing the cross entropy between the predicted activities share and the true ones. For each experiment, we performed a grid search over the number of trees, the minimal node size, the tree depth and the learning rate, for a total of 25 distinct models. The best model was selected by a 5 folds block cross-validation by cities where each city is entirely contained in a fold. The final performances of the chosen model were eventually assessed on a test set of five cities (Bordeaux, Dijon, Nantes, Paris and Rennes).

We ran and compared multiple models that differ in terms of input data. First, we chose which data to include as input: OSM, NetMob (we kept only the Tuesday profile, as typical weekday, and linked it by hour and zone to the target dataset), or NetMob weekly usage information (percentage of use on each day of the week per application for each zone, this information is constant for each time bands but it may helps in characterizing the type of district). For each possible dataset or dataset combination, the time band values were also provided as a numeric feature. Then, for each possible input data choice, we compared different ways to handle the NetMob data: categorization of applications (game, social network, etc.) or not, normalizing by z-score or total use, at the city scale or at the city and hour scale. These pre-processings were identified as important to mitigate regional effects on the application usage data and enable generalization to new cities.

Results The R^2 metric computed on the test set, for each experiment are displayed in Figure 1 (a). In order to better visualise the results, we only show it for all cities without distinction, with z-score normalization at city/hour scale and no categorization of the applications, as it showed better results.



(a) (b)

Figure 1: (a) R^2 computed on the test sets for models with different input tuning and different targets, (b) Features importance scores for at-home activity with the NetMob dataset alone (the importance of the time band feature is not shown (shap score of 1.38) to better display the importance of the NetMob features).

Overall, taking into account the use of mobile applications improve the model's results, for the different types of activity. "Home" and "working" activities are more easily predicted (with a good accuracy), than the others, as they are common activities on a large number of districts, with a strong hourly pattern. "Studying" prediction seems to benefit greatly from mobile application data. Overall, the NetMob data with weekly profiles enable a better or a similar prediction than the baseline for all the targets and the best results are obtained by combining the NetMob data with contextual information from OSM. We focused mainly on predicting certain types of activity in different city districts. The results showed us that mobile application usage data was a valuable input for prediction, enabling prediction with relatively good accuracy on unseen cities. Mobile application usage data seems therefore an interesting proxy to study evolution of the activities performed. Far from limiting ourselves to predicting types of activity, we extended the study to the prediction of the dynamic of demographic and social indicators (level of education, socio-professional category, age distribution), available with Mobiliscope data. The predictions are not as good as the ones for activities, but some characteristics still benefited from the use of mobile applications usages.

This research was partially funded by MobiTic ANR Project, a collaboration between INSEE, mobile operator Orange and the Université Gustave Eiffel to measure people's mobility and presence in a given area.

[1] Inaki Ucar, Marco Gramaglia, Marco Fiore, Zbigniew Smoreda, and Esteban Moro. *News or social media? socio-economic divide of mobile service consumption.* Journal of the Royal Society Interface, 18(185):20210350, 2021.

[2] Vallée J, Douet A, Le Roux G, Commenges H, Lecomte C, Villard E *Mobiliscope, a geovisualization platform to explore cities around the clock (v4.2). [Data set].* Zenodo. doi: 10.5281/zenodo.7822016, 2023.

[3] Orlando E. Martínez-Durive and Sachit Mishra and Cezary Ziemlicki and Stefania Rubrichi and Zbigniew Smoreda and Marco Fiore, *The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography*, arxiv:2305.06933, 2023.

29 June 2023

Applications of state space models to mobile app data

Orest Bucicovschi[1], *David A. Meyer*[1,2], *David P. Rideout*[1], *Asif Shakeel*[1], *Jiajie Shi*[1]

[1]Department of Mathematics, University of California, San Diego
[2]Theoretical Sciences Visiting Program
Okinawa Institute of Science and Technology Graduate University
orest@gmx.net, dmeyer@ucsd.edu, drideout@ucsd.edu
asif.shakeel@gmail.com, jis254@ucsd.edu

State space models for time series consist of an unobserved state which stochastically determines the observed data, and a stochastic dynamics for the unobserved state. As an example, consider the time series formed by summing Netflix downloads over all the tiles in Bordeaux. Figure 1 shows the part of this time series from 2019 March 16 00:00 to 2019 March 31 02:00, when Daylight Saving Time started: $\big(y_t \mid t \in \{1, 2, \ldots, m_1 \times 15 + 4 \times 2\}\big)$, where $m_1 = 4 \times 24$. It appears to be approximately periodic with a period of m_1, so a simple linear state space model is

$$\begin{aligned} y_t &= \mu_t + \delta_t + \epsilon_t & \epsilon_t &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ \mu_{t+1} &= \mu_t + \xi_t & \xi_t &\sim \mathcal{N}(0, \sigma_\xi^2) \\ \delta_{t+1} &= -(\delta_t + \cdots + \delta_{t-(m_1-2)}) + \eta_t & \eta_t &\sim \mathcal{N}(0, \sigma_\eta^2), \end{aligned}$$

where μ_t is the *local level* and δ_t is the daily periodic component. We estimate this model using standard algorithms—the Kalman filter and smoothing—obtaining the blue curve in Figure 1, which is thickened to give the 2 standard deviation ($\approx 95\%$) confidence interval.

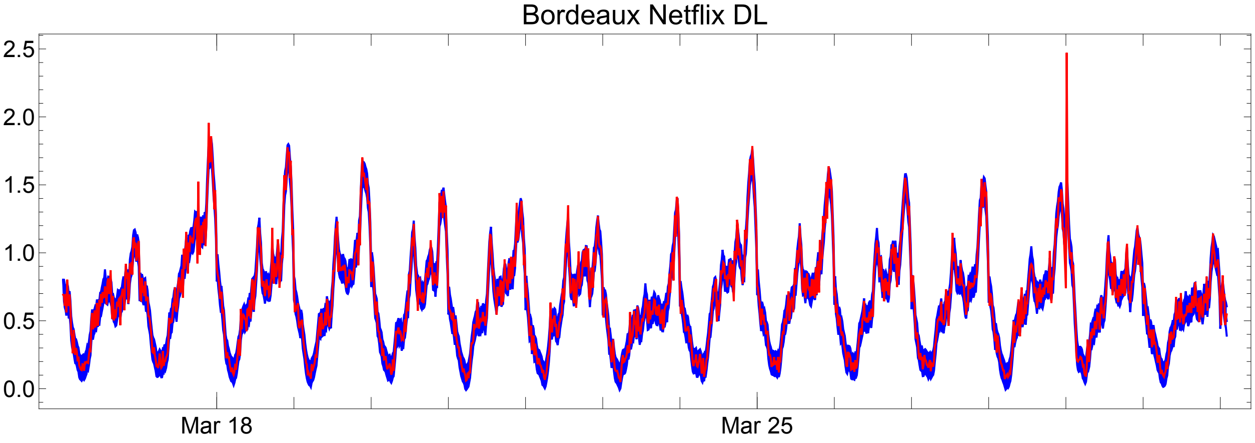


Fig. 1. Netflix download (DL) data aggregated for Bordeaux over 15 minute intervals is shown in red. The estimated state space model with its 95% confidence intervals is shown in blue. Vertical unit is 10^9.

The first, immediate, application of such a state space model is the systematic identification of anomalies, *i.e.*, data points far outside the confidence interval. Figure 2 illustrates this with the data for March 28 and March 29: the y_t are in red, the estimated $\hat{\mu}_t + \hat{\delta}_t$ curve

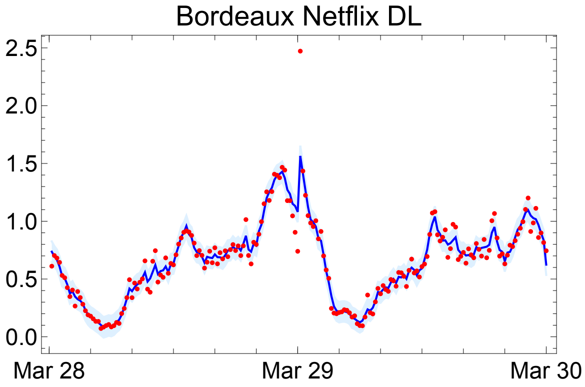


Fig. 2. A midnight anomaly.

is plotted in blue, and the confidence interval around it is shown in light blue. Notice that the confidence interval seems correct; of the $2m_1 = 192$ data points, about 7 of them are clearly outside the light blue region. One of the data points, March 29 00:15, however, is way outside the confidence interval. This constitutes an anomaly, and calls for a social explanation. On 2019 March 29, Netflix released its third French language series, *Osmosis*. Plausibly, many subscribers in Bordeaux connected just after midnight to stream it. But then why is the traffic in only one 15 minute interval anomalously high when the episode runtime is about an hour? Perhaps because Netflix releases new series at midnight *California time*, so at midnight in Bordeaux, only the trailer would have yet been available!

Having estimated a state space model from data (y_t), it is straightforward to generate *simulated data*, conditional on the real data. Thus a second application of this formalism will be to create simulated mobile app use data. We emphasize, however, that it is hard to imagine any simulation procedure that would preserve explainable anomalies like the one we just analyzed. (And, to be clear, we should confirm the presence of this anomaly in the Netflix DL time series for the other cities in the dataset before taking the *Osmosis* explanation too seriously.)

The state space formalism is extremely flexible. Because there is an approximate daily periodicity to the data, the official clock time seems likely to be important. The change to DST at 2019 March 31 02:00 jumped the time to 03:00. Using clock time to index the time series in these data leaves an hour gap without data. But state space models handle this easily, simply fitting to the data points that exist. A third application then, is to look for deviations from the model in the time period immediately after the change to DST, observing its social and economic impact.

The data also display an approximate weekly periodicity. The state space model formulated above can be generalized to include a second periodic component, and also to include a trend, although that seems perhaps less useful for the 77 day length of these data. Finally, the time series need not be a sequence of scalars; it can be a sequence of vectors, *e.g.*, comprising the data for a single app listed for all the tiles in one city, or the data for multiple apps. That is, the formalism extends to multivariate time series.

Multi-geographical-variate state space models, once estimated, can be compared across geographical regions, and analyzed in terms of disaggregated demographic and economic information. This is a fourth, important, application of these models.

Mapping mobile service usage diversities in cities

Maxime Lenormand[1, *]

[1] *TETIS, Univ Montpellier, AgroParisTech, Cirad, CNRS, INRAE, Montpellier, France*

INTRODUCTION

As mobile data traffic continues to grow globally, to gain a better overview of the mobile service usages and acquire a better understanding of their distribution in space and time is becoming relevant in many research area.

In this paper, we propose to build upon the work of [1, 2] in an attempt to map mobile service usage diversities in cities at different scales based on the data made available as part of the *NetMob 2023 Data Challenge* [3]. The dataset provides information about the traffic generated by 68 mobile services, at a high spatial resolution of 100 x 100 m^2 over 20 metropolitan areas in France during 77 consecutive days in 2019 [3]. The unusual richness of this dataset gives rise to interesting questions related to the diversity of mobile service usage in cities.

More specifically, the aim of this paper is to rely on an entropy-based metric to assess the mobile service usage diversity in cities by focusing on the hourly traffic volume information of six social network services. We focus in particular on the difference between cities according to the type of day (weekday, Saturday, Sunday, holiday).

MATERIALS AND METHODS

Data

The dataset used in this study is part of the NetMob 2023 Data Challenge [3]. It contains download and upload traffic volume information on 68 mobile services during 77 days (from March 16 to May 31 2019) accross 20 urban areas in France at a 100 x 100 m^2 spatial resolution with a 15 minute temporal resolution.

Data cleaning process

To illustrate our approach we selected six social network services among the 68 mobile services available in the dataset: *Facebook*, *Instagram*, *LinkedIn*, *Pinterest*, *Snapchat* and *Twitter*.

In order to compare the traffic volume information across cities we selected one regular week (April 1-7) without any school holidays, national holidays or anomalies from Monday to Sunday for every city. We also selected one spring school holiday week without anomalies for each city according to the school holiday zone (Figure 1). We selected the week April 22-28 for the zone A, April 15-21 for the zone B and from April 29 to May 5 for the zone C.

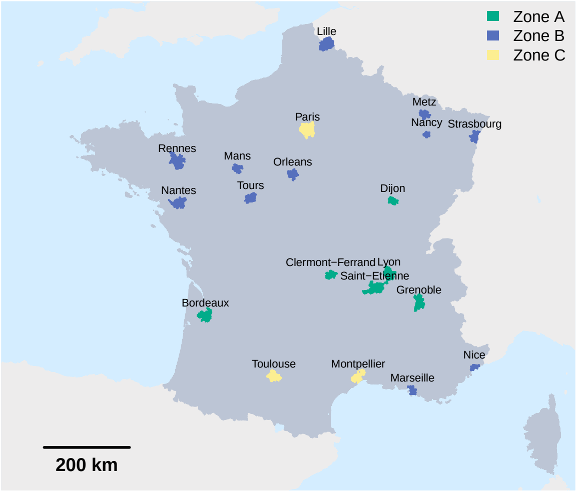


Figure 1. **Map of the studied areas.** The dataset is composed of 20 French cities. The colors represent the three school holiday zones (green for A, blue for B and yellow for C).

Then, the original dataset has been aggregated in space using 500 x 500 m^2 grid cells. We also aggregated the dataset in time using 1 hour-time slots instead of the original 15 minutes slots. We only considered four days of the week: Thursday representing a normal working day, Friday, Saturday and Sunday.

For each city c and week type (regular and holiday), the distribution of traffic volume (addition of download and upload traffic volume) for the six selected services has been stored in a three-dimensional matrix $T^c = (T^c_{s,g,h})$ representing the traffic volume associated with mobile service s in grid cell g during a given hour h. To fairly compare the traffic volume across mobile services and cities we first normalized the three-dimensional matrix by the traffic volume of mobile service s in each city c (Equation 1).

$$\hat{T^c}_{s,g,h} = \frac{T^c_{s,g,h}}{\sum_k T^c_{s,k,h}} \tag{1}$$

This quantity has then been normalized to obtain a total traffic volume summing to 1 for each grid cell (Equation 2).

$$\tilde{T}^c_{s,g,h} = \frac{\hat{T^c}_{s,g,h}}{\sum_k \hat{T^c}_{k,g,h}} \tag{2}$$

[*] Corresponding authors: maxime.lenormand@inrae.fr

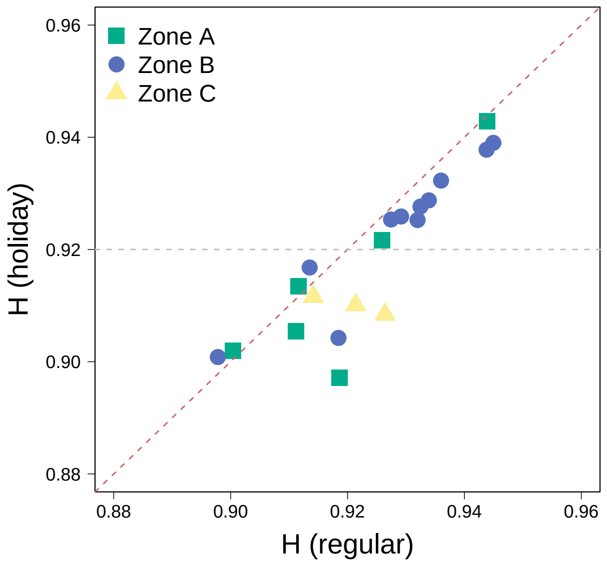


Figure 2. **Shannon diversity index as a function of the time according to the week type (regular and holiday).** The results are based on the average Shannon diversity index for a given hour and a given city. These values are then aggregated over the 20 cities. The plain line represents the average and the dotted lines the minimum and maximum values.

Diversity metric

We focused in this study on a well known diversity metric to quantify the diversity of services in a cell. The normalized Shannon diversity index [4] computed as follow for a given city c and a given cell g during a given hour h:

$$H^c_{g,h} = -\frac{1}{ln(S)} \sum_{s=1}^{S} \tilde{T}^c_{s,g,h} ln(\tilde{T}^c_{s,g,h}) \qquad (3)$$

where $S = 6$ is the number of selected mobile services.

PRELIMINARY RESULTS AND FUTURE WORK

After cleaning and formatting the data, we computed the diversity metric for every hour, city and type of week (regular and holiday). We first focused on the average Shannon diversity index per cell and per hour according to the city and the type of week (regular and holiday).

Figure 2 shows that for a large majority of cities, the diversity is lower during the holiday week than during the regular week. A diversity gradient according to the city locations can also be observed. Indeed, as shown in Figure 2 the diversity value 0.92 for the average diversity during the holiday week can be used as threshold to roughly divide the cities into two categories. The cities with a “low” diversity from one side (namely Bordeaux, Le Mans, Clermont-Ferrand, Toulouse, Montpellier, Paris, Nice, Saint-Etienne, Lyon and Marseille) and the cities with a “high” average Shannon diversity index from the other side (namely Grenoble, Tours, Nancy, Lille, Metz, Rennes, Orleans, Strasbourg, Nantes and Dijon). With the exception of Le Mans and Paris, cities belonging to the “low” diversity group are located in the southern half of France (Figure 1). It is also important to note that the four cities (Nice, Saint-Etienne, Lyon and Marseille) showing the highest differences in diversity between holiday and regular weeks belongs to this group. Inversely, cities belonging to the “high” diversity group (excepted Grenoble) are located in the northern half of France (Figure 1).

The next step will be to conduct a thorough analysis of the spatio-temporal evolution of mobile usage diversity within cities, the similarity/dissimilarity of mobile usage within and between cities and the identification of spatial cluster homogeneous in terms of mobile usage composition and the relationships between them.

[1] R. Singh, M. Fiore, M. Marina, A. Tarable, and A. Nordio. Urban Vibes and Rural Charms: Analysis of Geographic Diversity in Mobile Service Usage at National Scale. In *The World Wide Web Conference*, WWW ’19, pages 1724–1734, New York, NY, USA, 2019. Association for Computing Machinery.

[2] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda. Not All Apps Are Created Equal: Analysis of Spatiotemporal Heterogeneity in Nationwide Mobile Service Usage. In *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, CoNEXT ’17, pages 180–186, New York, NY, USA, 2017. Association for Computing Machinery.

[3] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore. The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography, 2023.

[4] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

Traffic Models and MEC Nodes Planning Using the NetMob 2023 DataSet

Shima Afshar Barji
Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN) University of Genoa
Genoa, Italy
shima.afshar.barji@edu.unige.it

Cristina Emilia Costa
Smart and Secure Networks National Laboratory (S2N Lab) National Inter-University Consortium for Telecommunications (CNIT)
Genoa, Italy
https://orcid.org/00000002-2198-2571

Fabrizio Granelli
Dept. of Information Engineering and Computer Science (DISI) University of Trento & CNIT
Trento, Italy
https://orcid.org/0000-0002-2439-277X

Raffaele Bolla
Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN) University of Genoa & CNIT
Genoa, Italy
https://orcid.org/00000003-2861-1586

I. Introduction and scope of the paper

Measurements-based approaches represent an important aspect in designing and analyzing current and future communication networks. Indeed, most of the works available in the scientific literature validate design schemes and achieved results by means of mathematical models or simulation. However, while such methods are surely useful in situations and scenarios where technologies are not yet deployed on the field or not available in the market, they limit the accuracy of results in cases where real data traces are available. NetMob 2023 DataSet [1] represents a relevant chance to use real data in the framework of the research activities on mobile networks.

This document describes how the research team of CNIT Research Units in Trento and Genoa (Italy) approach to the use the dataset provided by the NetMob 2023 Challenge. The research team is familiar with measurement-based approaches, as in several scientific works the researchers use results from testbed or real measurements to derive methods for providing strong and effective validation of their achieved results.

Mobile network data is indeed extremely variable and volatile, due to the freedom it provides to the users in terms of terminals to use and movement in a large area, while still presenting clearly defined patterns, already emerging from a brief analysis of the NetMob 2023 dataset.

We believe that the availability of such high-resolution detailed information about real usage of the mobile network can be effectively utilized to enable a more accurate and unprecedented study about network design principles and to better understand how traffic demands varies in time and space. In this framework, two relevant scenarios where a measurement-based approach based on the NetMob 2023 DataSet might provide the highest impact: (a) traffic and service modeling and prediction, and (b) MEC design and dimensioning.

Consequently, the objectives of the proposed work are the following:

1. Analyzing and processing the dataset in order to build time series of traffic, aggregate and by service type, and then using traffic modelers to derive accurate models of the traffic traces and define different models for the different service types;
2. Studying how it is possible to provide short- and long-term estimates of traffic in space and time;

The expected results are: (i) to define accurate models for the traffic and service types offered to the mobile network, both in space and time. This will be used to better understand and dimension the slices of 5G/6G networks; (ii) to build accurate models for traffic prediction based on actual data traces. Such results will be useful to accurately plan the potential deployment of MEC nodes to support the described traffic patterns, and to define the related costs in terms of deployment and operation (including power consumption). However, these last aspects will be subject to future work on the subject.

II. Methodology

IIA. Pre-processing

The Data Preprocessing phase plays a pivotal role in transforming the raw data from the NetMob23 dataset into a suitable format for subsequent analysis, specifically focusing on time series clustering. The dataset used in this research is characterized by intricate spatiotemporal dimensions, making the preprocessing step a crucial prerequisite for meaningful clustering and forecasting.

In preparation for time series clustering and subsequent mobile data traffic forecasting, the following preprocessing steps are undertaken:

Data Extraction:

- Extraction of the mobile data traffic time series data from the NetMob23 dataset, focusing on the selected subset corresponding to the City of Lyon.

- Organization of data by tile, service type (Downlink), and the chosen 15-day temporal window.

Data Normalization:

- Normalization of the time series data to ensure uniform scales for each tile's traffic observations. This step facilitates meaningful comparisons between tiles.

Feature Engineering:

- Derivation of relevant features from the time series data, which may include statistical moments, frequency domain features, and other pertinent metrics that capture the time series characteristics.

Spatial Aggregation:

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- Grouping of tiles based on their geographical proximity to form spatial clusters. These clusters may be used as a basis for initial time series clustering.

Temporal Aggregation:

- Aggregation of the 15-minute interval data into coarser temporal resolutions (e.g., hourly) to reduce the dimensionality of the time series and potentially uncover long-term patterns.

IIB. TIME SERIES CLUSTERING

Time series clustering is a pivotal component of this research, and we intend to employ Self-Organizing Maps (SOM) and Dynamic Time Warping Self-Organizing Maps (DTW-SOM) to cluster tiles based on their mobile data traffic time series patterns. These algorithms are well-suited for this task due to their ability to capture complex patterns and relationships in time series data.

Self-Organizing Maps (SOM)

Self-organizing maps are a type of artificial neural network that maps high-dimensional data onto a lower-dimensional grid while preserving the topological relationships between data points.

Dynamic Time Warping Self-Organizing Maps (DTW-SOM)

Dynamic Time Warping (DTW) is a technique for measuring the similarity between time series data, accounting for variations in time and amplitude. DTW-SOM combines the benefits of SOM with DTW to cluster time series data in a more time-aware manner.

Hyperparameter Tuning

Hyperparameter tuning is a crucial step to optimize the performance of both SOM and DTW-SOM for time series clustering. To determine the optimal values of these hyperparameters, we will employ mathematical optimization techniques such as grid search, random search, or Bayesian optimization. These methods enable to systematically explore the hyperparameter space and select the configurations that result in the best clustering outcomes.

Error Minimization

The key objective in time series clustering using SOM and DTW-SOM is to minimize the Quantization Error (QE). By minimizing QE, we ensure that the weight vectors associated with neurons on the SOM grid accurately represent the underlying patterns in the time series data. This leads to more meaningful and cohesive clusters.

IIC. SPATIAL CLUSTERING

Processed data is then analyzed in the spatial dimension in order to define spatial clusters, or traffic hotspots, for the different types of application traffic.

The usage of spatial clustering enables to outline relatively large areas characterized by similar temporal behavior, thus facilitating the analysis and prediction of traffic patterns.

It was observed that spatial clusters tend to be located in areas with specific social / life meanings, such as train stations, large plazas, residential areas, highways.

Figure 1 provides an example of spatial clustering, geo-referenced on the map of Lyon.

IID. FORECASTING MODEL DEVELOPMENT

The final stage is related to implement forecasting model. This stage is characterized by the following steps:

- Develop time series forecasting models customized for each cluster of tiles.

- Experiment with various forecasting algorithms, including ARIMA, LSTM, and hybrid models.

III. Results

Results presented at the workshop will include:

- Identification of clusters of tiles with similar mobile data traffic time series patterns.
- Development of cluster-specific forecasting models.
- Improved accuracy in mobile data traffic forecasting within small areas.

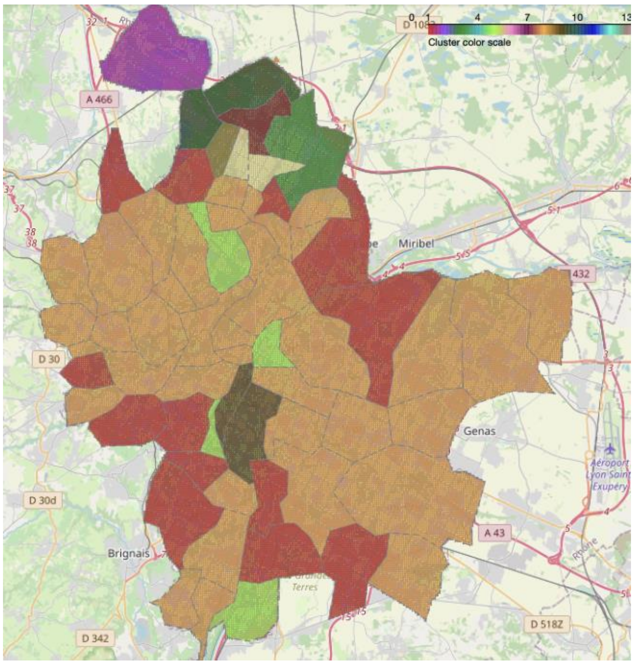


Figure 1. An example of spatial clustering for the city of Lyon, based on time-pattern similarity. Only one application is considered.

AKNOWLEDGEMENT

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU Project: “SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics” – Prot. IR0000013 – Call n. 3264 of 12/28/2021.

REFERENCES

[1] Martínez-Durive O E, Mishra S, Ziemlicki C, Rubrichi S, Smoreda Z, Fiore M. The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv:2305.06933 [cs.NI]. 2023.

Using Service-Level Mobile Internet Traffic Data to Estimate Noise Impact of Flight Trajectories on the Population

Emir Ganić[*], Simon Staudinger[†], Christoph G. Schuetz[†], and Claudia Vinueza[†]
[*]University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, Serbia
e.ganic@sf.bg.ac.rs
[†]Johannes Kepler University Linz, Linz, Austria
{staudinger,schuetz,vinueza}@dke.uni-linz.ac.at

I. Problem Statement

In pursuit of a more environmentally friendly and sustainable—"greener"—aviation industry, the impact of air traffic on the population has caught the attention of policy-makers, regulatory bodies, and researchers. The aviation industry, like any industry, does not act within a vacuum but requires a *social license* to be able to offer services in a sustainable manner. Thus, the negative impacts of air traffic and airports must be brought down to levels accepted by society. In this regard, despite increased efforts by the air traffic industry in general and airport operators in particular, noise emissions remain a key concern for inhabitants of noise-affected areas surrounding airports.

In order to minimize the impact of noise on the general population, air traffic controllers and airport operators require knowledge about the number of people affected by aircraft noise. Estimates of the number of people affected by aircraft noise are traditionally obtained using yearly aircraft noise exposure data and census data, yielding *noise contour maps*. The implicit underlying assumption is that people spend their daily lives mostly at their home addresses. Any intervention to reduce noise impact based on that assumption will neither reflect the realities of the people living in noise-affected areas around an airport who commute to work into another area nor of those people who commute into the noise-affected areas for work, education, and other purposes.

The analysis of mobility patterns in the population, which are not accounted for in contour maps, may yield a more accurate representation of the spatial and temporal aspects of sound in people's daily lives. Measures towards reducing the impact of aircraft noise on the population based on such mobility patterns promise to be more effective than those based on contour maps relying on census data.

II. Methodology

We propose to use the NetMob23 dataset [1] for the analysis of mobility patterns in the population around airports. Hence, traffic data of various internet services generated in different geographic tiles around French metropolitan areas allow to obtain an approximation of the number of people present in a certain tile at a certain time. Using these data, the noise impact of actual flights on the general population can be calculated and a what-if analysis using common departure and arrival routes for the respective airports can be conducted. In particular, we conducted a case study regarding the different noise impacts of selected flight trajectories on different days. In the future, a decision support system may propose alternative routes that may reduce the noise impact of flights given different mobility patterns.

We obtained historical flight data and information about departure and arrival routes at French airports from the OpenSky network [2], which is an online repository for ADS-B data (Automatic Dependent Surveillance – Broadcast), tracking aircraft movement. The movement data can be combined with further information about the aircraft used to operate the flight in order to more accurately estimate the noise impact of a given flight.

We base our work on previous, similar studies [3]–[6], albeit with different datasets, both in terms of spatio-temporal focus and the semantics of the data. Neither of those studies employed actual mobile phone data but relied on data collected by national statistics offices, such as daily passenger mobility surveys. We also used different metrics due to the assumption that the population present in a certain geographic tile during a specific time period is proportional to the service-level mobile internet traffic in that same geographic tile and time period.

III. Case Study

To demonstrate our approach, we have selected Lyon–Saint Exupéry Airport that serves the Lyon metropolitan area, which is covered by the NetMob23 dataset. There are 67 different communes (French local administrative zones) in Lyon, that are further divided into 512 IRIS zones (Figure 1). IRIS represents a fine-grained territorial subdivision of France, where in each IRIS zone there is around 2000 inhabitants.

We calculated the *Noise Impact Index* (NII) for five different flight trajectories, which are those of actual flights having taken place on 16th March 2019, four of which were departures and one of which was an arrival, using three different types of

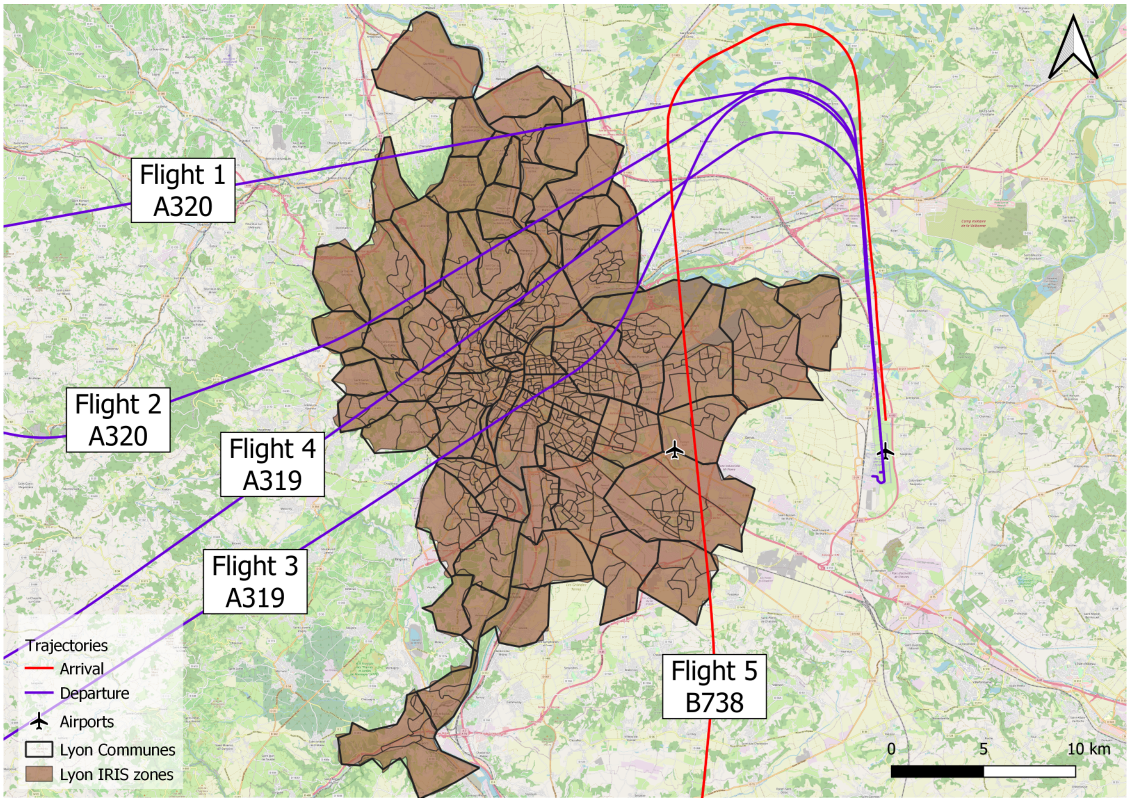


Fig. 1. Lyon communities, IRIS zones, and trajectories of selected flights

aircraft, namely Airbus 319, Airbus A320, and Boeing 737-800 (Figure 1). For each trajectory, we compared the changes in noise impact between different days from 16th–22nd March 2019 due to the different number of people affected in the areas passed by the aircraft. Hence, using the mobile internet traffic as an approximation for the actual population present in a certain area during a time period, we show that the same trajectory may have a considerably different noise impact on different days of the week. The mobile internet traffic could therefore be used by air traffic controllers to select the route for a flight with the least noise impact.

IV. Future Work

The analysis we conducted for the NetMob 2023 Data Challenge is a preliminary study for developing a decision support system guiding air traffic controllers in selecting routes with minimal environmental impact. Incorporating movement data of the general population into the systems employed by air traffic controllers for routing flights has the potential to considerably improve the quality of life of the population affected by air traffic. Hence, we envision a system that incorporates real-time mobility data of the general population into the route planning process for flights to minimize noise impact while also taking into account fuel consumption. Using the NetMob23 dataset we intend to demonstrate the usefulness of such a system and motivate further development.

Acknowledgment

This research was supported by the Serbian Ministry of Science, Technological Development and Innovations and the Austrian Federal Ministry of Education, Science and Research as part of the bilateral scientific and technological cooperation project “Research collaboration towards Greener Aviation” in the period 2022–2024.

References

[1] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, “The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography,” 2023.

[2] “The OpenSky network – free ADS-B and Mode S data for research,” https://opensky-network.org/.

[3] E. Ganić, O. Babić, M. Čangalović, and M. Stanojević, “Air traffic assignment to reduce population noise exposure using activity-based approach,” *Transportation Research Part D: Transport and Environment*, vol. 63, pp. 58–71, 2018, https://doi.org/10.1016/j.trd.2018.04.012.

[4] V. Ho-Huu, E. Ganić, S. Hartjes, O. Babić, and R. Curran, “Air traffic assignment based on daily population mobility to reduce aircraft noise effects and fuel consumption,” *Transportation Research Part D: Transport and Environment*, vol. 72, pp. 127–147, 2019, https://doi.org/10.1016/j.trd.2019.04.007.

[5] E. Ganić, N. van Oosten, L. Meliveo, S. Jeram, T. Louf, and J. J. Ramasco, “Dynamic noise maps for ljubljana airport,” *10th SESAR Innovation Days, Virtual conference*, 2020, https://www.sesarju.eu/sesarinnovationdays.

[6] E. Ganić, F. Rajé, and N. van Oosten, “New perspectives on spatial and temporal aspects of aircraft noise: Dynamic noise maps for heathrow airport,” *Journal of Transport Geography*, vol. 106, p. 103527, 2023, https://doi.org/10.1016/j.jtrangeo.2022.103527.

Network Analysis to Manage Network's Zero-Touch Configuration

Mihaela I. Chidean^a, Luis Ignacio Jiménez Gil^b, Javier Carmona-Murillo^c, David Cortés Polo^{c*}
^{a}*Dept. of Signal Theory and Communication, Univ. Rey Juan Carlos, Spain*
^{b}*Dept. of Computer Science, University of Valladolid, Spain*
^{c}*Dept. of Computing and Telematics Engineering, Univ. de Extremadura, Spain*
mihaela.chidean@urjc.es nacho.jimenez@uva.es jcarmur@unex.es *dcorpol@unex.es (corresponding)

***Abstract*—The challenge aims at deriving new knowledge from the analysis, characterization, modelling and cross-correlation of an original high-resolution dataset describing mobile service usage in multiple regions of France. This work proposes two approaches to analyse the NetMob23 dataset and extract knowledge about the services and their characteristics. This can be used by the network operators to improve services and optimize network resource allocation by means of the information extracted from the network. This approach is known as Zero-Touch Configuration, and it is one of the expected services that will be implemented in Beyond 5G and 6G.**

***Index Terms*—OPNA, L-Moments, Network Analysis, Zero-Touch Configuration.**

I. Introduction

The exponential growth of network devices in recent years has led to a need to automate as many management tasks as possible. This has also led to increased availability of network data and metadata. Machine learning techniques are expected to be used at all network levels in 5G and Beyond networks, especially using geolocated data sets. Geolocated data sets like NetMob23 can be also exploited. In these cases, data can be arranged in a three-dimensional data cube (two dimensions for the spatial location and one for the network feature). In addition, the dataset also includes temporary information establishing an order of occurrence. Spatial and temporal information in the network services metadata enables the usage of novel information extraction techniques, to look at network information from a different perspective by recognizing user behavior patterns.

However, this scenario requires novel network analysis methodologies to take advantage of all this available data, especially for the network usage pattern. In this work, we propose two different methods to analyze the information contained in the dataset. One uses a novel statistical theory to analyze the data. The second one is based on the analysis of each zone as a combination of the services used.

II. Methods

A. Approach 1: Lmom

The first considered approach is based on the L-moment statistical theory [1]. This approach has already shown promising results, in works such as network data analysis for attack detection [2] and, bitrate analysis in cellular networks for different user mobility patterns [3] and applications [4]. In [3] and [4], one of the issues was the dataset size to analyze in detail multiple situations and cases, the drawback clearly solved by the NetMob23 dataset [5].

The L-moment theory [1] is an alternative statistical theory to the classical one, where the moments are computed using linear combinations of the expected values of order statistics. The main benefits of the L-moments is their direct result interpretation, the unbiased estimators, the outlier robustness and the low sampling variability, even for low sample size situations [1], [6]. L-moments are therefore a powerful tool in distributional analysis, specially for data with a large range, variation, skewness, and outliers [6], [7], being all these characteristics fulfilled by multiple network variables [2], [8].

B. Approach 2: Orthogonal Network Projection (OPNA)

The second one is based on the extraction of extreme points using the orthogonal projection. This method decreases the complexity of the classification algorithm to obtain key information about network usage, characterizing the behavior of a zone in terms of the services used by the clients of the network. This approach is focused on the linear model approach, which considers each sub-area characterization a linear combination of different unknown network *comportments* to be unique descriptions of a particular activity in the network. In this paper, we use the Orthogonal Network Projection (OPNA) [9] technique, that extracts and represents these *comportments*, thereby reduces the information that needs to be managed and applied according to different aims. This approach has already been proven for notwork anomaly analysis, in combination with the usage of the Median Absolute Deviation (MAD) metric [10]. This approach allows the detection of higher usage demands of the users to the network.

III. Preliminary results

A. Approach 1: Lmom

For this approach, we considered the traffic dataset for the city of Nancy during May and for all available services. L-moment ratios τ_3 and τ_4 are estimated using $n = 5000$ samples. Fig. 1 shows the LmomRD obtained for the downlink for (a) the 68 available services and (b) the instant messaging services. We can observe positive L-asymmetry and L-kurtosis for all services, spanning a wide range of values. The values of τ_3 and τ_4 are located in specific regions, revealing differences in the

This research was funded in part by TED2021-131699B-I00/ AEI/10.13039/501100011033/ European Union NextGenerationEU/PRTR, by the Univ. Rey Juan Carlos Program for Research Promotion and Development (Ref. F920 and "AYUDA PUENTE 2022, URJC" Ref. F931).

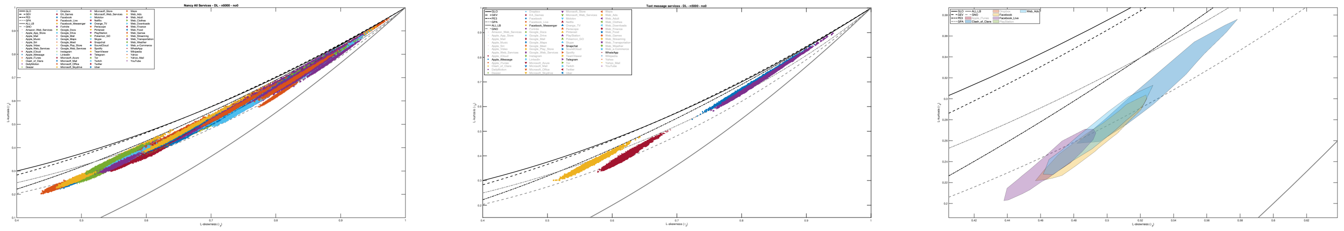


Fig. 1: (a) LmomRD for downlink for all available services. (b) Lmom for downlink for text message services. (c) Area that represents the τ_3 and τ_4 of selected services in order to show the existing overlap.

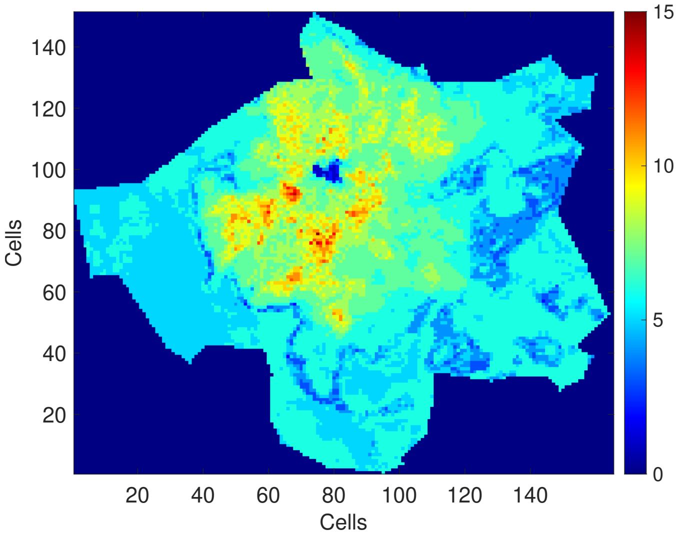


Fig. 2: Cell classification in terms of the number of events that requires an increase in network resources usage.

statistical behavior of each service, as expected and shown in previous works like [4]. Moreover, apparently similar services have different statistical behavior. Fig. 1(c) is focused on the "Facebook Live" service and shows the exact overlap in terms of τ_3 and τ_4 with other services, indicating that the usage of the L-moment statistic has great potential for automatic traffic classification, as well as for network usage pattern detection.

B. Approach 2: OPNA methodology

For this approach, also the city of Nancy was analyzed, now for the complete two-month time interval. The OPNA technique is employed extract the "comportments" that characterize the network usage of each cell and the MAD is computed to determine the average cell usage as well as the time instants when applications require higher bandwidth. Fig. 2 shows the categorization of each cell based on the number of events that requires an increase in network resources usage. This classification distinguishes cells with more consistent changes, depicted in reddish colors, from those with minimal changes, represented in blueish colors. This information is crucial for network operators for user behaviors analysis and network resources optimization, leading to e.g. a situation where active network equipment's could be dynamically adjusted (turned on/off) in terms of newtork usage and using the 5G facilitating technologies such as SDN and NFV, reducing therefore the network maintenance costs and increasing the energy efficiency.

IV. Discussion

As can be observed, both methods have obtained promising results, and multiple research directions were opened from the obtained results.

Regarding the first approach, we can conclude that the value of n is rather higher than previous works [2]–[4], [8], however, recall the dataset size and magnitude. Rather than a drawback due to the memory requirements, this is a feature as more accurate estimates and conclusions are obtained. In a real-time scenario, the value for n will be significantly reduced to profit all the L-moment theory benefits. The results shown that there are services with heavily-tail statistical behavior (high τ_3 and τ_4). The information regarding the statistical behavior of each service as well as the overlap between them will allow the proposal of automatic classification algorithms as well as service customized and more efficient network management and resource allocation algorithms.

The results obtained using the second approach allows us to use the novel methodology OPNA to analyze all the services at the same time, this can allow us to detect anomalies or relevant information without matching the information of each service. As can be observed, experimental results show how both proposed methodologies select and classify network behavior patterns using a simple classification algorithm and how these patterns could be used to find, for instance, anomalies in the network, track human mobility, undertake network planning, detect events in the network, etc.

References

[1] J. R. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *J. R. Stat. Soc.: Series B (Methodological)*, vol. 52, no. 1, pp. 105–124, 1990.

[2] J. Galeano-Brajones *et al.*, "A novel approach for flow analysis in software-based networks using l-moments theory," *Computer Communications*, vol. 201, pp. 116–122, 2023.

[3] D. Cortés Polo *et al.*, "Análisis de tasa de datos en redes móviles en función del grado de movilidad empleando diagramas de L-momentos estándar," in *URSI 2023, Cáceres, Spain*, 2023.

[4] J. Ortega *et al.*, "Bitrate Analysis in 5G Networks for Video Streaming Services Using L-Moment Ratio Diagrams," in *IEEE EUROCON*, 2023.

[5] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," 2023.

[6] W. H. Asquith, "Univariate Distributional Analysis with L-moment Statistics using R," Ph.D. dissertation, Texas Tech University, 2011.

[7] R. M. Vogel and N. M. Fennessey, "L moment diagrams should replace product moment diagrams," *Water Resour. Res.*, vol. 29(6), 1993.

[8] M. I. Chidean *et al.*, "Network Traffic Characterization Using L-moment Ratio Diagrams," in *IOTSMS 2019*. IEEE, 2019, pp. 555–560.

[9] D. Cortés-Polo *et al.*, "A novel methodology based on orthogonal projections for a mobile network data set analysis," *IEEE Access*, 2019.

[10] D. Cortes-Polo *et al.*, "Orthogonal projection for anomaly detection in networking datasets," *J. Ambient Intell. Humaniz.*, vol. 14(6), 2023.

Unveiling social vibrancy in urban spaces with app usage

Thomas Collins*,[1], Diogo Pacheco[1], Riccardo Di Clemente[2,3], and Federico Botta[1,2]

[1] *Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United Kingdom.*
[2] *The Alan Turing Institute, London, NW1 2DB, United Kingdom.*
[3] *Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.*
*trc207@exeter.ac.uk

Keywords: (1) Urban vibrancy, (2) Mobile app activity, (3) Spatial patterns (4) Urban spaces, (5) Social fabric.

Summary

Urban vibrancy, an important metric of a city's vitality, offers insights into the dynamics of urban space utilization. Advances in computational techniques and access to extensive data sets now enable a more comprehensive understanding of this phenomenon. Leveraging these methods is essential to deepen our understanding of urban vibrancy, particularly concerning socio-spatial segregation. Our analysis employs the NetMob23 data set [1], which captures spatiotemporal app usage across France's largest cities. Building upon previous studies [2, 3], we explore the relationship between the usage of a wide range of apps, the urban environments in which they are used, and some socio-demographic indicators of people living in those areas.

We draw data from three sources: (1) the NetMob23 data set, (2) *IRIS2000* [4], and (3) *OpenStreetMap*. NetMob23 tracks mobile app usage across 20 French cities, generating digital signatures. *IRIS2000* contains population and education data for city residents, while *OpenStreetMap* [5] provides crowdsourced geospatial information. Our analysis focuses on Paris, Marseille, and Lyon, France's three largest metropolitan areas.

We derive 'digital signatures' from mobile network traffic data. We interpolate the data spatially from $100 \times 100\,\mathrm{m}^2$ cells to the IRIS2000 regions. We categorize app data into *Apple Store* classifications (e.g., 'Advertising', 'Messaging', 'Productivity') and standardise the data in each cell so that we can compare the app usage across cells with differing overall usage volumes. We aggregate the data into 'Weekdays' (Monday to Thursday) and 'Weekends' (Friday to Sunday).

We retrieve urban feature data from *OpenStreetMap* ('OSM'): (1) 'Points of Interest' ('POI') impacting segregation [6] and (2) socially significant 'third places' [7]. From these features, we construct a range of variables used in our analysis (1) POI density and (2) diversity, (3) third place density and (4) diversity, and (5) the dependency ratio, (6) total enrollment, and (7) total out-of-education as socioeconomic indicators.

In our analysis, we study how urban features and demographic groups are distributed across clusters arising from similarities in app usage. We cluster the NetMob23 mobile network traffic data using HDBSCAN, investigating the emergence of distinct app usage clusters, and compare those clusters with respect to their urban features and socio-demographic information. We also investigate whether urban features and demographic groups are related to individual app groups, to better understand the relationship between urban environments, socio-demographics and our digital behaviour. We fit a series of spatial regression models for each city, studying the 'total effects', encompassing direct and indirect impacts, of the urban and demographic variables included in our analysis.

In the analysis, all three cities exhibit clusters during both weekdays and weekends. Focusing on Paris, we observe different clusters for weekdays and weekends, although some clusters (e.g., cluster one and cluster three; see Figure 1 A) are shared. We study the distribution of urban and demographic features in the clusters, emphasising that the clusters were detected only using the NetMob23 apps data. Visual inspection suggests that these are generally similar between weekdays and weekends. However, differences emerge, particularly in education variables. Population dependency ratio clusters exhibit more dispersion on weekends, possibly due to varying behaviours.

POI diversity is more skewed on weekdays, likely reflecting work-related activities near POIs (Figure 1 B).

Spatial modelling identifies differences and similarities between weekdays, weekends, and cities. In Paris, the professional networking model exhibits a negative relationship with education enrollment on weekend days ($Total$ = -3181.34, $p = 0.02$, pr2 = 0.98), not on weekdays, suggesting leisure-time professional development. However, a positive relationship with the population dependency ratio on weekdays ($Total = 3989.16$, $p = 0.01$, pr2 = 0.97) and weekends ($Total = 2230.39$, $p = 0.04$, pr2 = 0.98) implies equal professional development opportunities throughout the week.

This study leverages extensive app usage data to explore urban vibrancy and its relationship with various app usage features. Including urban and socioeconomic indicators allows us to investigate how the social fabric of cities relates to our digital behaviour across a wide range of app categories, unveiling an interesting interplay between how we use apps, the environment in which we live, and different demographic groups.

In conclusion, our results emphasise the value of computational approaches in comprehending urban environments and integrating social aspects in the computational study of cities.

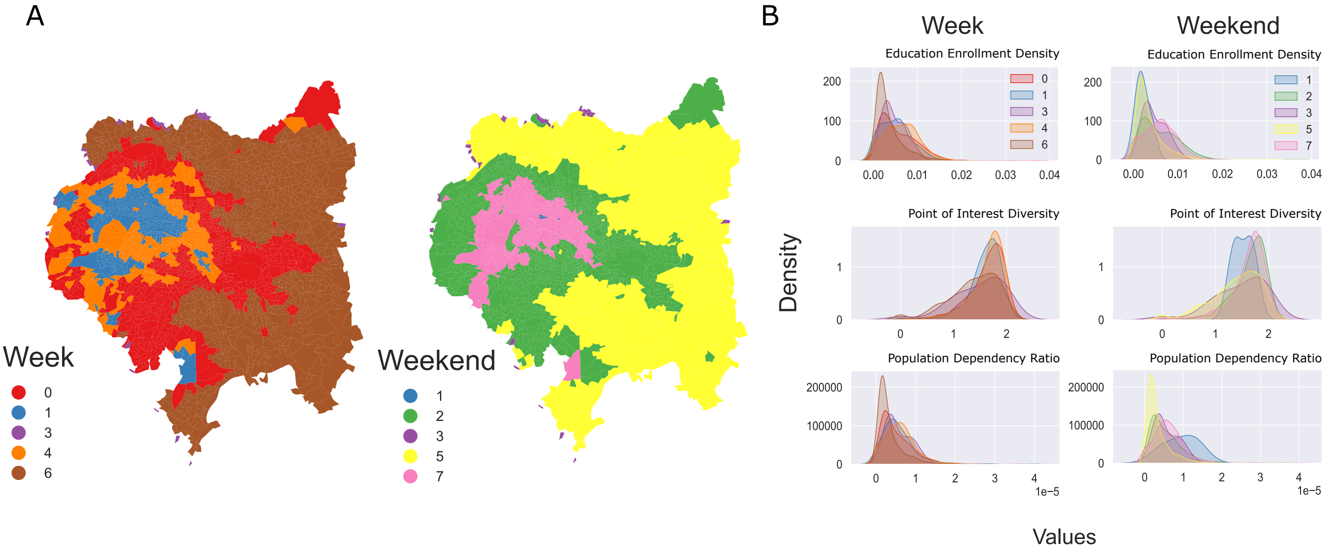


Figure 1: Clustered app usage in Paris across the week (Monday–Thursday) and weekend (Friday–Sunday) is shown as (A) cluster maps and (B) density histograms for three of the seven variables in our analysis: education enrollment, the dependency ratio, and ‘Points of Interest’ diversity. Note that these variables were not used to generate the clusters, which were found purely based on the app usage in each cell.

[1] Orlando E. Martínez-Durive et al. *The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography.* May 2023. arXiv: 2305.06933 [cs].

[2] Federico Botta and Mario Gutiérrez-Roig. “Modelling Urban Vibrancy with Mobile Phone and OpenStreetMap Data”. In: *PLoS ONE* 16.6 June (2021), pp. 1–19.

[3] Thomas R. Collins et al. *Spatiotemporal Gender Differences in Urban Vibrancy.* Apr. 2023. arXiv: 2304.12840 [physics].

[4] The National Institute of Statistics and Economic Studies. *The National Institute of Statistics and Economic Studies.* https://www.insee.fr/en/metadonnees/definition/c1523.

[5] OSM. *OpenStreetMap Contributors.* https://www.openstreetmap.org/. 2017.

[6] Esteban Moro et al. “Mobility Patterns Are Associated with Experienced Income Segregation in Large US Cities”. In: *Nature Communications* 12.1 (July 2021), p. 4633.

[7] Ramon Oldenburg and Dennis Brissett. “The Third Place”. In: *Qualitative Sociology* 5.4 (1982), pp. 265–284.

Mobility data generation based on tabular GANs models

Sara Kassan
Orange Labs
Belfort, France
sara.kassan@orange.com

Bechir Mnakri
Orange Labs
Belfort, France
bechir.mnakri@gmail.com

Frederic Guyard
Orange Labs
Sophia Antipolis, France
frederic.guyard@orange.com

Tamara Tosic
Orange Labs
Sophia Antipolis, France
tamara.tosic@orange.com

Thierry Nagellen
Orange Labs
Sophia Antipolis, France
thierry.nagellen@orange.com

I. INTRODUCTION

The significantly increased mobility of population and goods leads to several problems such as congestion and increased pollution. To limit these problems, it is necessary to have a good road network management and planning. A continuous monitoring for all roadway links is required to identify the features of traffic in the road network. However, this involves a high cost of installation and maintenance of road infrastructure. Hence, mobile telephone systems are considered as a promising technology for mobile traffic data collection system which can characterize traffic in a real time. In other hands, privacy is an aspect to be mentioned in these systems. All these falls within the legal framework as they are governed by regulations in charge of protecting the privacy of phone subscribers. The phone position data would be received aggregated and handled in anonymous manner to respect current regulations. The data collected is pseudo-anonymised and needs a long process for permission to access. Therefore, different methods and presented in the literature that use real traffic mobility pseudo-anonymised data to generate synthetic traffic mobility data. Privacy concerns today, however, have restricted sharing of such dataset. This has led to the development of synthetic traffic generators.

Several approaches have emerged to create synthetic data, including Generative Adversarial Neural Networks (GANs), Variational Autoencoders as well as Auto-regressive Networks. We propose a model based on tabular conditional GANs models to generate synthetic traffic mobility data using NetMob datasets.

GAN, as represented in figure 1, is a type of of machine learning model that consists of two neural networks: a generator and a discriminator. The generator network takes random noise as input and tries to generate synthetic data that are similar to the real data from the training set. In the other hand, the discriminator network, takes both real data features from the training set and the generated samples from the generator network as input. Its task is to distinguish between the real and fake samples.

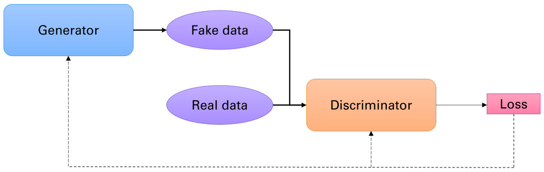


Fig. 1. The architecture of GAN model

A. Data preparation

Normalization and smoothing

The inputs of the neural network will be in the form of vectors containing the time series of different places presented by polygons and the traffic of applications for each polygon. These vectors correspond directly to the rows of Netmob dataset. In this part, there is no manipulation to be performed on the structure of the data. In order to improve the performance of the model, we want to normalize the values for each attribute. Therefore, we apply a **min-max** normalization, which consists of a simple operation to place all our values between -1 and 1 :

$$X_{\text{normalised}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \times 2 - 1 \tag{1}$$

where $X_{\max}$ and $X_{\min}$ are maximal and minimal attributes.

B. Data Description: Youtube Traffic in Nice

For this subset extracted from Nice, we present the traffic of Youtube service on the map in Figure 2.

Figure 2 shows that the traffic of Youtube service is not very important between $02:00$ AM and $05:00$ AM at the morning. After $05:00$ AM, the traffic of Youtube service becomes more and more highly important. We can see that Youtube service is highly used in this subset of Nice for different time frames.

C. Proposed model based on tabular GANs model

Discriminator architecture: In order to distinguish between real data and synthetic data, the Discriminator of this model is based on a succession of 3 Dense layers. The first two

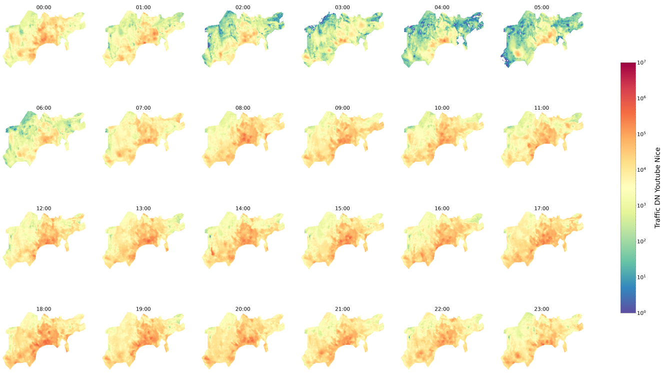


Fig. 2. Traffic of Youtube service in the region of Nice for different time frames.

layers have a drop-out about 20% and an activation function **Leaky ReLU**, presented by :

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{cases} \tag{2}$$

Where α is a coefficient between 0 and 1. The activation function is used due to its simplicity. It permits quick convergence with reducing the problem of gradient desperation.
The third Dense layer has 1 neuron and its role is to classify data. The value done by this layer indicates if the data is real or synthetic.
Generator architecture: The generator is based on the architecture is based on 3 dense layers and skip-connections that reshape the noise data to pass from a layer to another one. The two first layers are Dense layers with a batch-normalisation that permits a performed training more stable of the network and an activation function **ReLU** similar to Leaky ReLU with $\alpha = 0$. The third layer has for each attribute contained in our initial data table, an activation function hyperbolic tangent. This function can be defined by the mathematical equation as follows:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

The values obtained are between -1 and 1. It is a normalization min-max.

II. Experiments

After the training, we produce a few samples and notice that the plots of the synthetic time series bear resemblance with the real time series, as shown in figure 3. Specifically, we can see that the real and synthetic series both have low values in the earlier hours of the day, and then go up after $07:00$ AM, while also maintaining irregular oscillations throughout the day. We believe that, although the global trends of the time-series could be captured with a knowledge-based method, our data-driven Deep Learning method allows to capture the underlying statistical rules of these irregular oscillations, thus allowing the generation of highly realistic data that could be exploited unrestrained by privacy concerns.

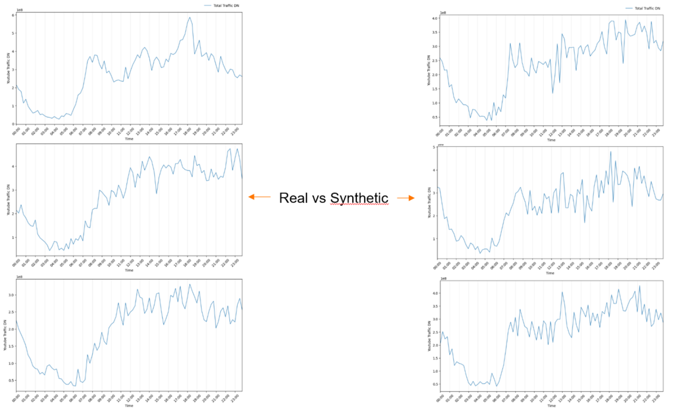


Fig. 3. Comparison between randomly drawn samples from the real data (left) and synthetic data (right).

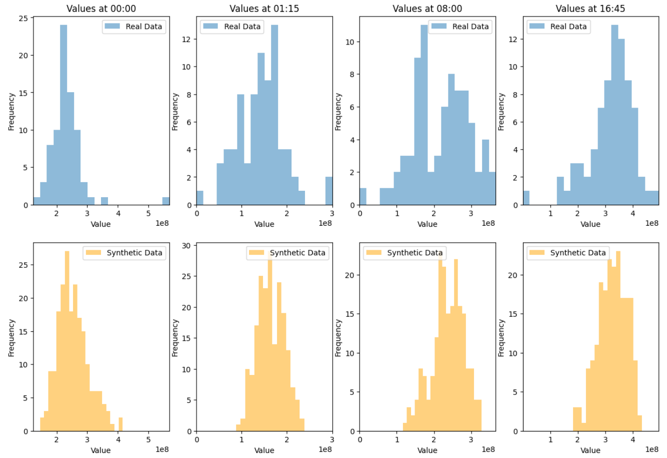


Fig. 4. Comparison of the histograms for 4 time frames. On top, the distribution of the real data, and bottom, the distribution of the synthetic data for the corresponding time frame and with aligned x axis.

A. *Statistical Metrics evaluation*

Among the statistical ways of validation of our model, we also have the histograms of the values at each time frame, as represented in figure 4, for real data and synthetic data generated by our current model. Similar histograms convey that, if we produce numerous samples from our model, we would cover, for each time frame, a realistic range of values.

III. CONCLUSION

This paper proposes an adapted conditional tabular GANs model based on unsupervised approaches to generate synthetic traffic mobility data. This model ensures performance enhancement in terms of quality of data generated. Besides, statistical metrics are used to validate the proposal model. The limitation of traffic mobility data could be further resolved by our proposal model. Moreover, an application of our proposal model on other type of data could be interesting to validate the flexibility of the model.

Seasonal Shifts in Mobile Data Consumption: Investigating the Effect of Daylight Variation on Application Usage Patterns

E. Kotov*[1], O. Hexel*[2], T. Theile*[3], E. Jacobs*[4], D. Perrotta*[5], J. Kim*[6], E. Zagheni*[7]

** Department of Digital and Computational Demography, Max Planck Institute for Demographic Research*

[1] corresponding author: kotov@demogr.mpg.de ; [2] hexel@demog.mpg.de ; [3] theile@demog.mpg.de ; [4] jacobs@demog.mpg.de ; [5] perrotta@demog.mpg.de ; [6] kim@demog.mpg.de ; [7] zagheni@demog.mpg.de

Our objective in this study is to investigate whether mobile data consumption patterns vary based on the duration of day and night. Is it common for individuals to use their mobile devices and specific applications more frequently during periods with less daylight? Or do they maintain their routines and opt for sleep or other non-mobile activities throughout the year?

To the best of our knowledge, there is no research on how smartphone usage patterns may depend on the length of daylight and nighttime and, therefore, on the availability of competing internet activities (e.g., reading books, spending time outdoors, and using other non-handheld devices) and circadian rhythm. Li et al. [1] found that individuals reproduce the same app usage patterns daily, with some variations between weekdays and weekends. Moreover, they found distinct patterns of night vs day activities, for example, overall social and shopping activity was peaking during the day and falling at night, while there were also clusters of nocturnal users engaging in entertainment activities or socializing on the Internet late at night. However, the study only analyzed data for a single week. There is little research on how this behavior may be changing over long periods [2]. One study [2] investigates change in patterns over multiple years, however, it compares the change between years, not between seasons. Most studies that focus on the specific time of app usage are limited to a few weeks [3,4], or months [5]. Some studies that do analyze longer periods focus on location [4,6], overall usage patterns [5,7], and app-specific usage time [8]. None of the long-term app usage studies look at the changes in usage by time of day and between seasons.

We, therefore, analyze the influence of available daylight on the specific timing of app and service usage via cellular data by tracing long-term shifts in this usage. Some studies show that seasonal change affects the circadian rhythm of humans [9,10]. When controlling for the ambient environment and engagement in social activities, individuals naturally went to sleep earlier in summer and later in winter [11]. We therefore expect to see earlier evening mobile phone data usage peaks for certain apps at the end of May compared to mid-March. We use the `NetMob23` dataset [12] to investigate this question in a single country. Within the timespan of 77 consecutive days available for analysis, daylight increases in Lille (the northmost city in the dataset) by 4 hours 15 minutes, and in Marseille (the southernmost city) by 3 hours 28 minutes. The difference in daylight between the north and south of France is, therefore, 47 minutes. This difference will allow us to estimate the effect of the length of daylight and timing of the sunset on app usage patterns. If there is an effect on mobile phone usage, observing the cities in France we should see gradual changes from North to South time shifts of mobile app and service usage.

Materials & Methods. We compare the app usage patterns between weeks 12 and 22 and over all eleven weeks. Data is aggregated over each working week by taking the mean of the total per app traffic for every time interval. This results in an average weekday usage for each week per 15-minute interval per city. The focus is at the city level to estimate the population-level effect, as users may be moving throughout the day and data usage activity cannot be attributed to any single individual, and usage at any location cannot be linked directly and solely to residents, as they may be on Wi-Fi [13–15]. We also break up each city into high-activity and low-activity zones and repeat the aggregation described above. This way for each city we are capturing possible differences between the areas which are predominantly residential and the areas that may be influenced from week 12 to week 22 by seasonal changes, such as tourism and late-night activities. The delimitation into zones for each city is performed by finding the largest and most central cluster of points of interest from the Overture Maps points of interest dataset [16] using DBSCAN algorithm (see Fig. 1).

To perform the comparisons we convert traffic usage shapes into density distributions. Timestamps of 15-minute intervals serve as observations, while traffic volume is regarded as weights of those intervals. For each app or service we take week 12 as a reference distribution and repartition week 22 usage according to the deciles of the week 12 (see Fig. 2). This approach allows us to use a single framework to compare all services and groups of services across cities disregarding the differences in the absolute volume of traffic for specific apps or services. To check if the usage of a specific app or service has shifted in time, we use two metrics: (1) change of the median (i.e. at what time of day half of the daily traffic for a particular app or service has been used) and (2) change of interquartile range (i.e. is traffic getting more or less concentrated in a particular time of the day).

Results. Our results (see) demonstrate most app and service usage patterns change over eleven weeks. Some services see the shifts in timing of usage for as much as 2 hours in either direction. The change in usage is mostly consistent within a service across all cities and does not depend much on geographical location, as we do not see the expected pattern of gradual shifts in usage from North to South. Therefore there is no evidence that increasing length of daylight has any effect on the usage patterns of mobile apps and services. Even though the change in timing of app usage is happening, it can be attributed to other factors, such as global changes in popularity of the services, or a general trend in increase of data consumption worldwide.

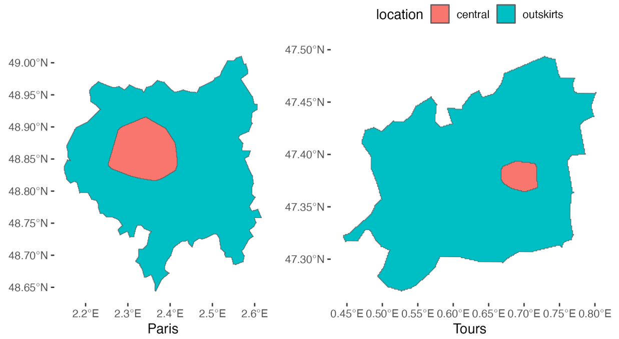


Fig. 1 Most active locations in cities identified through DBSCAN clustering of Overture Maps points of interest

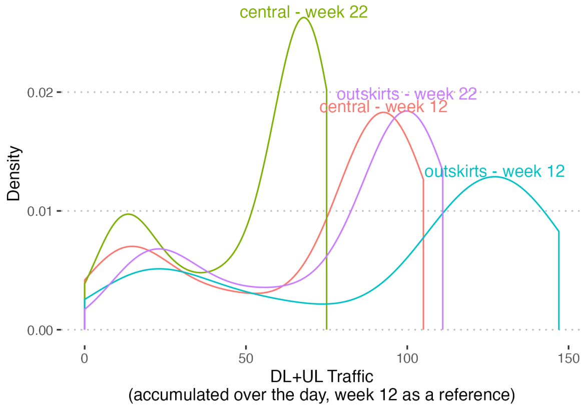


Fig. 2 Week 22 traffic repartitioned into week 12 deciles by location (all cities, all services)

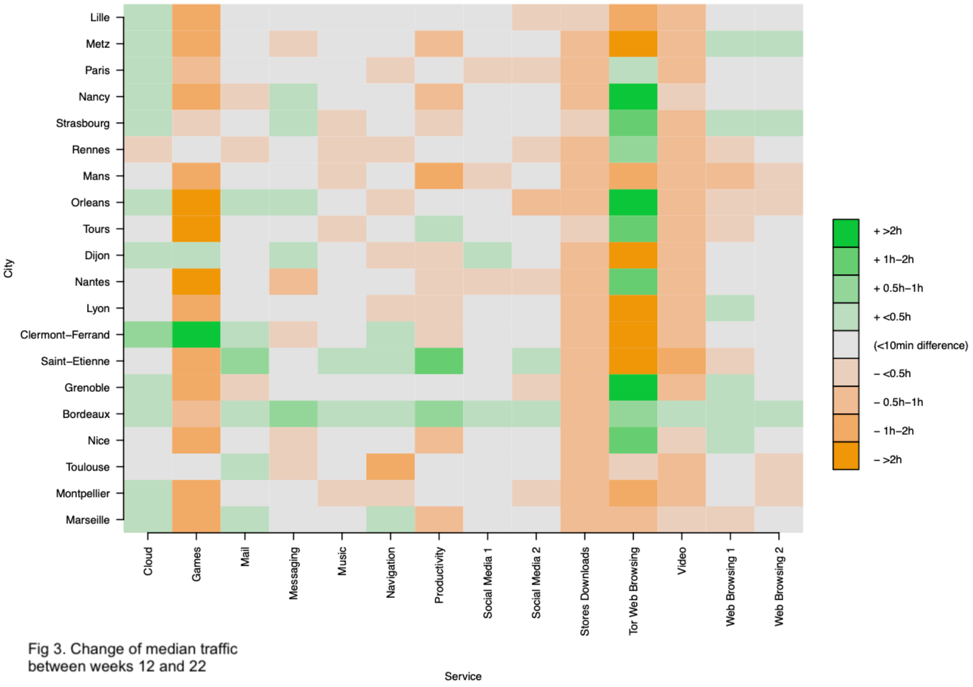


Fig 3. Change of median traffic between weeks 12 and 22

References

1. Li T, Li Y, Hoque MA, Xia T, Tarkoma S, Hui P. To What Extent We Repeat Ourselves? Discovering Daily Activity Patterns Across Mobile App Usage. IEEE Transactions on Mobile Computing. 2022;21:1492–507.
2. Li T, Zhang M, Cao H, Li Y, Tarkoma S, Hui P. ”What Apps Did You Use?”: Understanding the Long-term Evolution of Mobile App Usage. Proceedings of The Web Conference 2020 [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2023 Jul 6]. p. 66–76. Available from: https://dl.acm.org/doi/10.1145/3366423.3380095
3. Van Canneyt S, Bron M, Haines A, Lalmas M. Describing Patterns and Disruptions in Large Scale Mobile App Usage Data. Proceedings of the 26th International Conference on World Wide Web Companion [Internet]. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2017 [cited 2023 Jul 6]. p. 1579–84. Available from: https://dl.acm.org/doi/10.1145/3041021.3051113
4. Wang H, Li Y, Zeng S, Wang G, Zhang P, Hui P, et al. Modeling Spatio-Temporal App Usage for a Large User Population. Proc ACM Interact Mob Wearable Ubiquitous Technol [Internet]. 2019 [cited 2023 Jul 6];3:27:1-27:23. Available from: https://dl.acm.org/doi/10.1145/3314414
5. Do T-M-T, Gatica-Perez D. By their apps you shall understand them: mining large-scale patterns of mobile phone usage. Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia [Internet]. New York, NY, USA: Association for Computing Machinery; 2010 [cited 2023 Jul 6]. p. 1–10. Available from: https://dl.acm.org/doi/10.1145/1899475.1899502
6. De Nadai M, Cardoso A, Lima A, Lepri B, Oliver N. Strategies and limitations in app usage and human mobility. Sci Rep [Internet]. 2019 [cited 2023 Jul 6];9:10935. Available from: https://www.nature.com/articles/s41598-019-47493-x
7. Li H, Lu X, Liu X, Xie T, Bian K, Lin FX, et al. Characterizing Smartphone Usage Patterns from Millions of Android Users. Proceedings of the 2015 Internet Measurement Conference [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2023 Jul 6]. p. 459–72. Available from: https://dl.acm.org/doi/10.1145/2815675.2815686
8. Liao Z-X, Pan Y-C, Peng W-C, Lei P-R. On mining mobile apps usage behavior for predicting apps usage in smartphones. Proceedings of the 22nd ACM international conference on Information & Knowledge Management [Internet]. New York, NY, USA: Association for Computing Machinery; 2013 [cited 2023 Jul 6]. p. 609–18. Available from: https://dl.acm.org/doi/10.1145/2505515.2505529
9. Mattingly SM, Grover T, Martinez GJ, Aledavood T, Robles-Granda P, Nies K, et al. The effects of seasons and weather on sleep patterns measured through longitudinal multimodal sensing. NPJ Digit Med. 2021;4:76.
10. Dunster GP, Hua I, Grahe A, Fleischer JG, Panda S, Wright KP, et al. Daytime light exposure is a strong predictor of seasonal variation in sleep and circadian timing of university students. Journal of Pineal Research [Internet]. 2023 [cited 2023 Jul 8];74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/jpi.12843
11. Honma K, Honma S, Kohsaka M, Fukuda N. Seasonal variation in the human circadian rhythm: dissociation between sleep and temperature rhythm. Am J Physiol. 1992;262:R885-891.
12. Martínez-Durive OE, Mishra S, Ziemlicki C, Rubrichi S, Smoreda Z, Fiore M. The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography [Internet]. arXiv; 2023 [cited 2023 May 12]. Available from: http://arxiv.org/abs/2305.06933
13. de Reuver M, Bouwman H. Preferences in data usage and the relation to the use of mobile applications. Calgary: International Telecommunications Society (ITS); 2014 [cited 2023 Jun 29]. Available from: https://www.econstor.eu/handle/10419/101437
14. Hyun J, Won Y, Nahm DS-C, Hong JW-K. Measuring auto switch between Wi-Fi and mobile data networks in an urban area. 2016 12th International Conference on Network and Service Management (CNSM). 2016. p. 287–91.
15. Walelgne EA, Asrese AS, Manner J, Bajpai V, Ott J. Understanding Data Usage Patterns of Geographically Diverse Mobile Users. IEEE Transactions on Network and Service Management. 2021;18:3798–812.
16. Overture Maps Foundation. Overture Map Dataset [Internet]. 2023 [cited 2023 Aug 30]. Available from: https://overturemaps.org/

Mobile Service Traffic Prediction based on Graph Spatial-Temporal Network and Cross-service Collaboration

Zhiying Feng, Xu Chen
School of Computer Science and Engineering, Sun Yat-sen University

I. Mobile Service Traffic Prediction with Cross-service Collaboration

A. Problem Formulation

In this research, the geographical area of a city can be divided into a M grids, where the size of each grid is 100 meters by 100 meters and these regions are denoted as $\mathcal{M} = \{1, 2, ..., M\}$. The set of mobile services are denoted as $\mathcal{S} = \{1, 2, ..., S\}$. For each region v, the traffic volume of mobile service s at time slot t can be represented as $y_{v,s}^{t}$. We aim to predict the traffic demand of mobile service s in region v at time slot $t+1$ given the previous P observed traffic values of service s in region v, $\mathbf{y}_{v,s}^{t-P+1:t}$, the previous P observed traffic values of other relative services in region v, $\mathbf{y}_{v,r\in\mathcal{S}\backslash s}^{t-P+1:t}$, the previous P observed traffic values of service s in other regions, $\mathbf{y}_{u\in\mathcal{M}\backslash v,s}^{t-P+1:t}$, and the previous P observed traffic values of other relative services in other regions, $\mathbf{y}_{u\in\mathcal{M}\backslash v,r\in\mathcal{S}\backslash s}^{t-P+1:t}$. Our mobile service traffic prediction problem can be formulated as

$$y_{v,s}^{t+1} = \mathcal{F}(\mathbf{y}_{v,s}^{t-P+1:t}, \mathbf{y}_{v,r\in\mathcal{S}\backslash s}^{t-P+1:t}, \mathbf{y}_{u\in\mathcal{M}\backslash v,s}^{t-P+1:t}, \mathbf{y}_{u\in\mathcal{M}\backslash v,r\in\mathcal{S}\backslash s}^{t-P+1:t}). \tag{1}$$

B. Construction of Spatial Relation Graph

The grid structure needs to be transformed into a graph structure based on certain criteria. In this research, we consider two criteria: 1) the distance of geographical location; 2) the similarity of historical time series. For $v \in \mathcal{M}$, the regions of K_{geo} closest to v will be established edge with v. Besides, by calculating the similarity of historical time series data of different regions, the K_{sim} regions with the highest similarity to v are obtained, and then they are established edge with v. Therefore, each region will be connected to at most $K_{\text{geo}} + K_{\text{sim}}$ other regions.

C. Spatial-temporal Feature Extractor

In order to extract the spatial-temporal feature of each region in the graph, we design a spatial-temporal feature extractor.

We first use the temporal convolutional layers to extract the temporal features of nodes. Lea *et al.* first proposed Temporal Convolutional Network (TCN) to replace RNN to handle the time series data in 2016. Just like processing images with convolutional layers, we can view time series data as a one-channel and one-dimensional image. we combine 3 temporal convolutional layers as a whole to form a temporal-convolutional block. For a region v and service s, given the input $\mathbf{x}_{v,s}$, the output of temporal-convolutional block can be formulated as:

$$o_{v,s}^{\text{TC}} = \sigma_{\text{relu}}\Bigg(f_{\text{TC3}}\Big(f_{\text{TC1}}\left(\mathbf{x}_{v,s};\theta_{\text{TC1}}\right) + \sigma_{\text{sigmoid}}\big(f_{\text{TC2}}\left(\mathbf{x}_{v,s};\theta_{\text{TC2}}\right)\big);\theta_{\text{TC3}}\Big)\Bigg) \tag{2}$$

where θ_{TC1}, θ_{TC2} and θ_{TC3} are the trainable parameters of these three temporal convolutional layers, σ_{relu} and σ_{sigmoid} are the relu and sigmoid activation function.

After obtaining the temporal feature of each node, we use graph convolutional network (GCN) to extract the spatial features of nodes. For service s, given the adjacency matrix $\mathbf{A}_s$ and the identity matrix $\mathbf{I}_M$, the feature extraction formula of GCN is as follows:

$$\mathbf{H}_s^{(l+1)} = \sigma(\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\widetilde{\mathbf{A}}_s\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\mathbf{H}_s^{(l)}\mathbf{W}_s^{(l)}), \tag{3}$$

where $\mathbf{H}_s^{(l)}$ is the output of lth layer, $\mathbf{H}_s^{(0)} = \{o_{1,s}^{\text{TC}}, o_{2,s}^{\text{TC}}, ..., o_{M,s}^{\text{TC}}\}$, $\mathbf{W}_s^{(l)}$ is a matrix with trainable parameters, $\widetilde{\mathbf{A}}_s = \mathbf{A}_s + \mathbf{I}_M$, $\widetilde{\mathbf{D}}_s$ is the degree matrix and its diagonal element $\widetilde{D}_{s_{ii}} = \sum_j \widetilde{A}_{s_{ij}}$.

After L-layer processing, we get $\mathbf{H}_s^{(L)}$. The feature of service s in region v in the Lth layer can be expressed as $\mathbf{h}_{v,s}^{(L)}$.

We combine two graph convolution layers and one temporal-convolutional block to form a spatio-temporal block. We stack several spatio-temporal blocks as a spatial-temporal feature extractor to extract the node embedding.

D. Attention-based Cross-Service Correlation Calculation

Collecting traffic data for all mobile services in each region is expensive, and data integrity is often difficult to guarantee. This brings us to the idea that when the traffic data of a certain mobile service in a certain region is missing, the traffic data of other mobile services can be used to compensate for this lack. The traffic data of different mobile services in the same area are correlated. For example, users use multiple office-related mobile services at the same time during working hours. We introduce an attention mechanism to automatically learn this correlation. The attention mechanisms determine the importance of different mobile services. For service s in region v, the query vector, the key vector, and the value vector are calculated as follows:

$$\mathbf{q}_{v,s} = \mathbf{W}_{\text{query}}^{(s)} \mathbf{e}_{v,s}^{t-P+1:t}, \tag{4}$$

$$\mathbf{k}_{v,s} = \mathbf{W}_{\text{key}}^{(s)} \mathbf{e}_{v,s}^{t-P+1:t}, \tag{5}$$

$$\mathbf{v}_{v,s} = \mathbf{W}_{\text{value}}^{(s)} \mathbf{e}_{v,s}^{t-P+1:t}, \tag{6}$$

where $\mathbf{W}_{\text{query}}^{(s)}$, $\mathbf{W}_{\text{key}}^{(s)}$, and $\mathbf{W}_{\text{value}}^{(s)} \in \mathbb{R}^{d \times P}$ are trainable matrix, $\mathbf{e}_{v,s}^{t-P+1:t}$ is the node embedding after the extraction of spatial-temporal feature extractor.

The attention coefficient of service r towards service s is calculated as follows:

$$\alpha_{v,s,r}^{t} = \frac{\mathbf{q}_{v,s}^{\mathrm{T}} \mathbf{k}_{v,r}}{\sqrt{d}}, \tag{7}$$

$$\hat{\alpha}_{v,s,r}^{t} = \frac{\exp(\alpha_{v,s,r}^{t})}{\sum_{i \in \mathcal{S}} \exp(\alpha_{v,s,i}^{t})}, \tag{8}$$

The impact of all other mobile services on a specific service s is calculated by weighted aggregation as follows:

$$\mathbf{x}_{v,s}^{t-P+1:t} = \sum_{r \in \mathcal{S}} \hat{\alpha}_{v,s,r}^{t} \mathbf{v}_{v,r}, \tag{9}$$

where $\mathbf{x}_{v,s}^{t-P+1:t}$ is the cross-service correlation output of other services towards service s.

E. Graph-Attention-based Mobile Traffic Prediction

Our proposed mobile traffic prediction scheme can consider the temporal correlation, spatial correlation and cross-service correlation synchronously. For the service s in region v, the cross-service correlation output $\mathbf{x}_{v,s}^{t-P+1:t}$ and the node embedding $\mathbf{e}_{v,s}^{t-P+1:t}$ are concatenated and then fed into the MLP to predict the mobile traffic demand $y_{v,s}^{t+1}$.

$$y_{v,s}^{t+1} = \mathrm{MLP}(\mathbf{x}_{v,s}^{t-P+1:t}; \mathbf{e}_{v,s}^{t-P+1:t}), \tag{10}$$

F. Collaborative Personalized Cross-Service Learning

The amount of data for a single mobile service is limited. In order to improve the generalization of the local model, we use the federated learning mechanism to achieve the sharing of model parameters. To further improve the prediction performance, we would like to account for the specific characteristics of individual service, which is called personalization in machine learning.

As mentioned above, each service processes a local prediction model, which contains the spatial-temporal feature extractor, $\mathbf{W}_{\text{query}}^{(s)}$, $\mathbf{W}_{\text{key}}^{(s)}$, $\mathbf{W}_{\text{value}}^{(s)}$, and the fully connected layer. Follow the idea that the closer the model parameters are to the output layer, the higher the degree of personalization. Therefore, we consider the parameters of the spatial-temporal feature extractor as the global parameters and consider $\mathbf{W}_{\text{query}}^{(s)}$, $\mathbf{W}_{\text{key}}^{(s)}$, $\mathbf{W}_{\text{value}}^{(s)}$ and the fully connected layer as the personalized parameters. The global parameters of all services are aggregated in every training rounds. The personalized parameters are kept unchanged during the global aggregation.

Managing Network Performance with Data Driven Strategies

Veena B. Mendiratta
Northwestern University
Evanston, IL, USA
veena.mendiratta@northwestern.edu

Mrinmoy Bhattacharjee
Nokia
Naperville, IL, USA
mrinmoy.bhattacharjee@nokia.com

Anvesh Chamanchula
Nokia
Naperville, IL, USA
anvesh.chamanchula@nokia.com

***Abstract*- This paper presents a preliminary exploratory data analysis for mobile applications; and presents the analyses and algorithms we propose to develop to manage mobile network performance driven by near-real-time data intelligence. The traffic behavior of select applications across cities is analysed and network management strategies like bandwidth control, slice allocation are discussed.**

Keywords: bandwidth management, network slice, machine learning, data-driven

I. Introduction

Next generation (5G and beyond) mobile networks have diverse service requirements, and heterogeneity in applications and devices, and are, hence, evolving into very complex systems. This requires proactive data analytics approaches for managing these networks to include descriptive, diagnostic, predictive and prescriptive analytics. Traditional centrally-managed approaches that are reactive, and conventional data analysis tools that have limited capability (space and time) are not adequate to meet the needs of future complex networks regarding operation and optimization cost effectively [1]

There is considerable work in the area of applying AI and machine learning techniques to different aspects of wireless networks [2]. Our focus too is on managing wireless network performance through data-driven strategies, and, therefore, we will focus on work addressing network performance. We propose analyses and algorithms, based on our expertise in the telecom domain, to manage network performance through resource allocations driven by near-real-time data intelligence. Our aim in this paper is to propose well defined use cases that can be used to forecast bandwidth allocation through the application of data analytics and machine learning.

II. Approach

To keep the scope of work manageable, three representative cities are selected for analysis, namely, Paris, Lille and Marseille. For each of these three cities, two services in each of four categories are selected for analysis as follows:

- Social Media: Instagram, Snapchat
- Office Collaboration: Microsoft Mail, Google Mail
- Storage: Apple iCloud, Google Drive
- Entertainment/Streaming: Netflix, YouTube

The exploratory data analysis (EDA) was performed on these services to understand the individual service behavior, trends, their effect on the network as a category and also comparing the services based on their up- link/downlink traffic characteristics. Next, we analyze how to translate these insights into network management actions such as bandwidth management, networking slicing for different services, etc. A forecasting model may be needed to deploy these management actions.

More specifically, we perform the following tasks with the data:

- The main aim of our research is to be able to come up with a way to allocate network bandwidth for different categories of usage and dynamically change allocations as per demand, which is akin to the concept of network slicing in 5G technology. As a future study we plan to add additional services to these categories for a more complete analysis and forecasting.
- Using bandwidth usage amount of these categories of services over the course of a day and a week, we intend to develop a model that can prescribe allocation of spectrum assuming a maximum available bandwidth in the CSP's network. We present a study of how best to manage the bandwidth based on: day of week, time of day, and the remaining bandwidth available.
- In future we intend to compare the bandwidth usage between the selected cities to analyze the differences and attempt to identify reasons for the differences. We also plan to include additional services in the future study for completeness.
- Based on continuation of this study, we also plan to develop a forecasting model to introduce these management actions in the future.

III. Analysis

In this section we discuss how the insights gained from the data analysis performed on the select cities and service categories can be translated into network management actions such as bandwidth management and network slicing for different services, etc. Also, a forecasting model may be needed to deploy these management actions.

Fig. 1, shows for the Paris site, bandwidth usage by each of the four categories along with the total bandwidth usage over the roughly 2.5 months interval for which data is available.

The maximum bandwidth usage in this site from the four categories of services is roughly between 600K to 700K units, with the maximum contribution coming from Social Media (~55-60%), followed by Entertainment/Streaming (~30%), Storage (~15%) and Office collaboration (<5%). Fig. 2 shows the same in terms of percentage contribution to bandwidth usage.

We observe very similar contribution to total bandwidth usage from these four categories of services in Lille and Marseille too though the total bandwidth usage is lower in both Lille and Marseille as compared to Paris. This analysis shows the amount of bandwidth that can be set aside for each service category by the CSP.

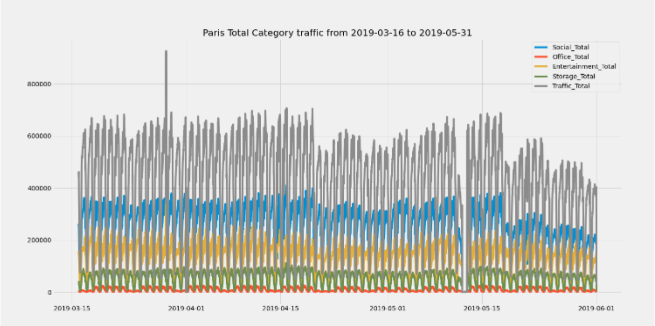


Fig. 1: Paris Total Bandwidth usage by each service category

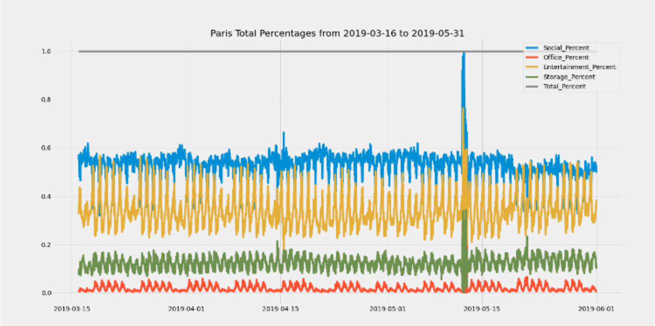


Fig. 2: Paris Bandwidth occupancy percentage by each service category

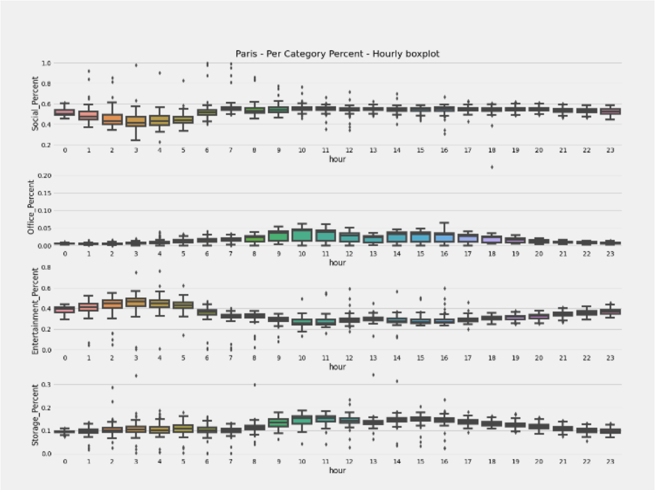


Fig. 3: Paris Hourly Bandwidth occupancy percentage by each service category

Secondly, analyzing the hourly trends, the bandwidth reservation for each service type can be further fine-tuned as shown in Fig. 3. This analysis shows that in 5G networks separate network slices can be created to reserve bandwidth for these service categories.

Table 1 summarizes bandwidth allocation per service type per hour of the day in percentage of total bandwidth. Furthermore, analyzing the percentage contributions from various services within the same category, for example, Instagram and Snapchat, bandwidth can be further split and reserved for different service types within a service category. CSP can reserve bandwidth separately for Instagram and Snapchat within the Social Media slice.

IV. Conclusions

In this preliminary work we have analyzed the data provided by the challenge for selected services and cities. This has provided us with insights for managing network performance with data driven strategies.

First, we propose that in 5G networks separate network slices be created to reserve bandwidth for different (as defined in this paper) service categories. We propose that total available bandwidth of the CSP can be allocated per service categories. This allocation should be data driven as shown in this paper. Bandwidth allocation per service category can further be finetuned over the course of the day as summarized in Table 1 since the bandwidth requirement is not the same throughout the day.

Furthermore, looking at the difference in occupancies between different services within each service category, we propose that the allocated bandwidth be further split and managed among these services.

To perform all these allocations efficiently and accurately, an AI and Machine Learning (ML) pipeline and models should be used. Continuous data collection is required that should be fed into this ML pipeline to train models that can be used to automate bandwidth management. As part of future study with this data, we propose to build AI and ML models to dynamically learn the bandwidth usage trends by each service and subsequently forecast bandwidth need by hour.

Table 1: Bandwidth Allocation per Service Category per Hour

Service	Social Media	Entertainment	Storage	Collaboration
Overall	55-60	~40	~15	~5
Hour 0	50	40	10	<5
1	50	40	10	<5
2 - 5	40	50	10	<5
6	60	30	10	5
7	55	30	10	5
8	55	30	10	5
9 - 19	55	30	15	5
20	55	35	15	5
21 - 23	55	40	10	<5

V. References

[1] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[2] D. Raca, A. H. Zahran, C. J. Sreenan, R. K. Sinha, E. Halepovic, R. Jana, and V. Gopalakrishnan, "On leveraging machine and deep learning for throughput prediction in cellular networks: Design, performance, and challenges," *IEEE Communications Magazine*, vol. 58, pp. 11–17, March 2020.

Characterizing and Comparing Cellular Network Traffic in French Cities

Alberto Blanc
IMT Atlantique
IRISA
Rennes, France
alberto.blanc@imt-atlantique.fr

Xavier Lagrange
IMT Atlantique
IRISA
Rennes, France
xavier.lagrange@imt-atlantique.fr

***Abstract*—Characterizing mobile networks' data traffic in the time and space domains is critical in improving how these networks are dynamically configured. In this paper, we study the NetMob 2023 Data Challenge dataset. We first establish the existence of a weekly pattern for the total traffic in each major city in France. We show that the daily pattern is essentially the same from Monday till Thursday but with more traffic on Wednesday. Some deviations are clearly visible on Friday. Using the IRIS zones, we try to group zones exhibiting similar patterns. We find that four or five different clusters can best characterize the traffic in most cities. Identifying these clusters is a valuable step in optimizing the placement of network functions. Finally, we analyze the percentage of traffic generated by different applications, showing that, while the overall percentage of traffic generated by an application is similar in each city, the spatial diversity for a given application varies from city to city.**

I. Introduction

We study the NetMob 2023 Data Challenge dataset [1] by analyzing the spatial and temporal diversity of the traffic and by characterizing the network traffic and its composition (e.g., daily patterns and which applications generate more data). Studying traffic's spatial and temporal distribution helps to implement efficient energy-saving strategies and place network functions. We are particularly interested in the similarities and differences between cities, as these can help determine whether any given solution can be applied to different cities or need to be adapted to the specificities of each city.

We use the *IRIS* zones defined by the French National Statistical Agency (*INSEE*) to group the $100\,\mathrm{m} \times 100\,\mathrm{m}$ tiles of the dataset. INSEE divides all the towns with more than $10\,000$ inhabitants and most of those with 5000 to $10\,000$ into homogeneous units. Units are of three different types: residential, business, and miscellaneous. Each residential unit has between 1800 and 5000 inhabitants. According to the INSEE website: "These units must respect geographic and demographic criteria and have borders which are clearly identifiable and stable in the long term." As IRIS zones reflect and respect local features, such as major roads, rivers, etc., and are stable in the long term, they are a reasonable basis for comparing the traffic in different cities.

We attribute each tile of the dataset to the IRIS zone that contains its center. This way, each tile is attributed to a single zone to avoid double counting the traffic. The center of some tiles on the border of each metropolitan area does not fall within an IRIS zone of the towns in the metro area. We attribute these tiles to the closest IRIS zone. We then sum the traffic off all the tiles attributed to each IRIS zone to obtain a new set of time series, one for each IRIS, application, and direction (uplink or downlink). All the results presented in this work are based on the downlink traffic only.

II. Total Traffic

We start by analyzing the total traffic in each metropolitan area, i.e., the sum of the traffic generated by all the applications in the dataset. We try to identify daily patterns and to identify the difference and similarities between different cities in France. The traffic shows a clear daily pattern, with very little traffic in the early hours of the morning, sharply increasing around 7 a.m., reaching a peak right after noon, followed by two other peaks, usually around 6 p.m., and 9 p.m., and a sharp drop around midnight. The most striking feature is that the midday peak is present in all the cities at the same time and every week.

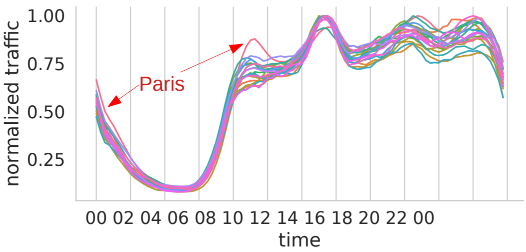


Fig. 1: Total median normalized traffic for all the cities on Tuesdays

To better compare the shape of the daily patterns for working days in different cities, we proceed as follows: first, we compute the median value for every day of the week (i.e., we compute the median of all the samples on a Monday at midnight, 12:15 a.m., 12:30 a.m., etc.). Then, we compute the rolling mean over the last four samples so that each value corresponds to the average of the median in the previous 45 minutes. In addition, when relevant, we further normalize by the maximum over the studied period to easily compare the shape of the traffic for cities with significantly different amounts of total traffic.

Fig. 1 shows the rolling mean of the median values for all the Tuesdays in the dataset for all the cities. It is remarkable how the first hours of the day are in synch between different cities: the rapid decrease around midnight and the rapid increase around 7 a.m. take place almost simultaneously for all the cities in the dataset. As noted above, the peak at noon is also remarkably similar in timing and shape for all the cities. The only exceptions are Nice and Paris, where the midday peak is not the highest in the day; in both cases, the peak is later in the afternoon. In Paris, there is also a significant peak around 9 a.m. (red curve in Fig. 1), possibly linked to the overwhelming popularity of public transportation in the capital. Note that the traffic pattern in Paris is shifted by roughly 45 minutes.

We also compare the daily patterns of different days. In most cases, we find that Wednesday is the day with the most traffic. Most cities show small differences between different days, with Friday often being more different.

A. *Clustering IRIS zones*

Given that IRIS zones have different sizes, populations, and traffic, it is not immediately clear if they are a good fit for analyzing the dataset. At the same time, IRIS zones have the advantage of reflecting the underlying characteristics of each town and have been defined with great care to ensure that the area within each one is homogeneous (e.g., similar housing type and density, presence of commercial or industrial activities, parks). As a first test to determine whether IRIS zones reflect in their traffic the underlying characteristics of the town, we have used the K-Means clustering algorithm with Dynamic Time Wrapping (DTW) of the tslearn[1] Python library to cluster the time series of the rolling mean of the median.

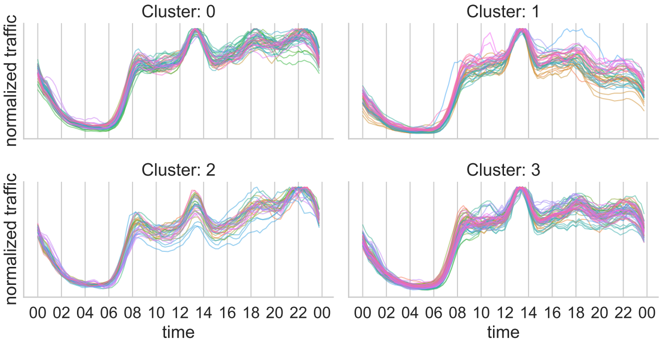


Fig. 2: The normalized traffic for the clusters in Rennes

Four clusters are a good choice for Rennes (see Fig. 2 and 3):

Cluster 0 has peaks at 6 p.m. and 10 p.m. of roughly the same size as the one at noon;

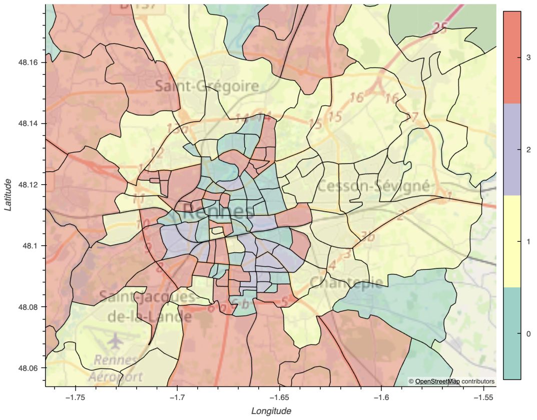


Fig. 3: The clusters in the center of Rennes

Cluster 1 has a peak at noon and then smaller peaks at 6 and 10 p.m., with the peak at 10 p.m. smaller than the one at 6 p.m.;

Cluster 2 has the highest peaks at 10 p.m. and noon, of roughly the same size, while the one at 6 p.m. is smaller;

Cluster 3 has peaks at 6 and 10 p.m. that are roughly the same size but smaller than at noon.

Clusters 1 and 2 show an upward trend during the day; cluster 0 has a roughly flat trend in the afternoon, while cluster 3 has an increasing trend in the morning and a decreasing one in the afternoon.

Cluster 0 covers predominantly residential zones in the periphery and the city of Rennes. Cluster 1 corresponds to IRIS zones classed as ”industrial/commercial” and to less densely populated areas on the periphery with significant commercial activities and office buildings. Cluster 2 is similar to cluster 1 but with an upward trend during the afternoon, indicating a stronger residential component. Cluster 3 covers zones with a mix of residential and commercial activities in the center and in the periphery.

III. Conclusions and future work

Thanks to the NetMob data challenge dataset, we have highlighted the similarities and differences in the traffic patterns in different cities. The most striking similarity is a peak in traffic at the beginning of the afternoon in all the cities in the dataset and at almost the same time.

Given the size and richness of the dataset, there are many possible extensions to this work. For example, we would like to exploit the spatial diversity of the traffic in each IRIS zone to propose an algorithm to place Network Functions so that the load on each one of them is as constant as possible (the idea being that, at least in some case, IRIS zones have complementary traffic patterns).

References

[1] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

[1] https://github.com/tslearn-team/tslearn

Generating Application Traffic Data from Tax Data: A Synthetic Data Approach - French Data Analysis

Tamer Arafa [1] tarafa@nu.edu.eg
Noha Gamal [1] ngamal@nu.edu.eg
Ghada Khoriba [1] GhadaKhoriba@nu.edu.eg
Mina Atef Yousef [1] MYousef@nu.edu.eg
Ahmed El-Mahdy [1] aelmahdy@nu.edu.eg

[1] School of Information Technology and Computer Science, Nile University, Giza, Egypt.

Introduction:

The availability of diverse and representative application traffic data plays a vital role in developing and evaluating algorithms and systems. However, can be difficult and constrained in terms of quantity, variety, and privacy issues. In this project, we suggest a cutting-edge method for producing fake application traffic data from tax data as a source. We seek to overcome the limitations of real data and facilitate research in network traffic analysis and modelling by utilising the correlation between tax data and application network data.

Methodology/Approach:

In our method, generative models are used to create application traffic data from the parameters derived from tax data. To capture the intricate patterns and correlations present in both tax and application traffic data, we use methods such as fully connected and convolutional Generative Adversarial Networks (GANs), variational autoencoders, and generative moment matching networks. With the help of these models, we can generate realistic synthetic traffic data that exhibits similar characteristics to real-world application network traffic. Zhang, et al. used deep generative adversarial networks to generate synthetic dataset by [1]. Charitou, et al. used GANs to generate synthetic Data for Fraud Detection [2].

Data/Experimental Setup:

In our analysis, we utilize geolocalized data provided by the INSEE (Institut National de la Statistique et des Etudes Economiques) derived from tax declarations filed in 2016 for the year 2015 [3]. This dataset is structured into 200-meter boxes and provides average yearly incomes, local population, number of families, living area, home ownership, and various other information. Additionally, we use high-resolution geo-referenced data for 68 popular mobile services across 20 metropolitan areas in France, during a continuous monitoring period of 77 days in 2019 [4]. This data was used in several studies, similar to the study by Nirbhay, el al. Income Inequalities correlation to City Size [5]. We analyzed both datasets (INSEE taxes data, and NetMob23 application traffic data) to confirm the viability of our assumptions. By taking Lyon as an example, we found the exact communes in both datasets, with matching IDs. We also plotted the regions covered by the tiles of both datasets and they are matching.

Our generative models receive input from the tax data parameters, including demographics, income levels, and population density as well as corresponding application traffic data and household. We preprocess the data to preserve privacy and remove any personally identifiable information from the data while maintaining the statistical properties required to produce accurate application traffic data. The model is trained no this data to be able to generate synthetic application traffic data for future years, using tax data generated on an annual basis.

Results/Findings:
We demonstrate the efficiency of our approach in producing synthetic application traffic data through our experiments. We compare the generated data to actual application traffic datasets and find that the statistical characteristics and patterns in the synthetic data closely match those in the original data. We use quantitative metrics, such as traffic volume, patterns, and distribution, to analyse the generated traffic data, and we compare the results to actual data benchmarks. Our findings demonstrate that the generative models successfully capture the relationship between application traffic and tax data, resulting in synthetic data that faithfully mimics underlying traffic patterns.

Discussion/Implications:
The successful generation of synthetic application traffic data from tax data has significant implications for various domains. It enables researchers and practitioners to access diverse and scalable application traffic datasets, which can be used for performance assessment, anomaly detection, traffic modelling, and prediction is made possible for researchers and practitioners. Additionally, the privacy-preserving nature of the generation of synthetic data ensures the security of sensitive data present in actual application traffic data. The research results from this project advance network traffic analysis and modelling, enabling more thorough research and the creation of novel solutions.

Conclusion:
In this project, generative models are used to generate synthetic application traffic data from tax data. Researchers can get around data constraints and advance their research in network traffic analysis thanks to the generated synthetic data's resemblance to real-world application network traffic. Synthetic data generation ensures data confidentiality while protecting user privacy and offering insightful information about traffic patterns. The goal of further improving the quality and applicability of the synthetic application traffic data is to refine the generative models, add more parameters, and broaden the evaluation metrics.

References:

[1] Chi Z, et al. "Generative adversarial network for synthetic time series data generation in smart grids." 2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm). IEEE, 2018.

[2] Charitou C, Dragicevic S, Garcez AD. Synthetic Data Generation for Fraud Detection using GANs. arXiv preprint arXiv:2109.12546. 2021 Sep 26.

[3] Donnees C, Revenus, pauvrete et niveau de vie en 2015 https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305#documentation.

[4] Martínez-Durive O E, Mishra S, Ziemlicki C, Rubrichi S, Smoreda Z, Fiore M. The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv:2305.06933 [cs.NI]. 2023.

[5] Patil N, Nadal JP, Bouchaud JP. Income Inequalities Increase with City Size: Evidence from French Data. arXiv preprint arXiv:2305.12864. 2023 May 22.

Finding unusual spatiotemporal traffic volumes using anomaly detection method

Rei Kemmi and Akihiro Fujihara
Department of Information and Communication Systems Engineering,
Faculty of Engineering, Chiba Institute of Technology, Japan
s20A5029PJ@s.chibakoudai.jp, akihiro.fujihara@p.chibakoudai.jp

1. Introduction

With the spread of smartphones and other mobile devices, it is possible to detect abnormal events such as incidents and disasters by machine learning the dynamic changes of traffic data. Therefore, we considered to find abnormal events by detecting anomalies in traffic volume using the mobile traffic data. Our preliminary research shows that there is some normality in the statistical distribution of traffic volume. In this case, it becomes possible to detect anomalies by an unsupervised learning using the Hotelling's T-squared test statistics [1].

2. Method

Our preliminary research shows that there is some normality in the statistical distribution of traffic volume. In this case, it becomes possible to detect anomalies by an unsupervised learning using the Hotelling's T-squared test statistics [1]. In this method, we assume that the a given observed data is independently and identically distributed. Then, we calculate an anomaly index representing how unusual the observed data is. This index is described as the following equation,

$$a(x) = \frac{1}{\sigma^2}(x-\mu)^2 = \left(\frac{x-\mu}{\sigma}\right)^2, \tag{1}$$

where x is the observed value, μ is the mean value of x, and σ is the standard deviation of x. This index follows the χ-squared distribution with one degree of freedom under the above assumption. A certain threshold is determined, and if the anomaly index exceeds the threshold, the observed value is considered to be abnormal.

3. Results

We analyzed an incident that a fire broke out at Notre Dame Cathedral in central Paris in the evening of April 15, 2019 [2]. We calculated the anomaly index and show its time evolutions of uplink traffic volumes during April 15-16 in Figures 1 and 2. We confirmed that the anomaly has increased sharply since around the evening of April 15. Interestingly, the increase clearly continues on the next day, but it gradually decays into the night. We visualized spatiotemporal anomaly patterns as shown in [3], where each color means $0 \leq a(x) < 2^2$(blue), $2^2 \leq a(x) < 3^2$(light blue), $3^2 \leq a(x) < 4^2$(green), $4^2 \leq a(x) < 5^2$(yellow), $5^2 \leq a(x) < 6^2$(orange), and $a(x) \geq 6^2$(red). In Twitter traffic volumes, we observed a phenomenon in which a highly anomalous region explosively spread throughout Paris in the evening of April 15. As shown in Fig. 3, the number of anomaly areas with red color drastically increase during the incident. On the other hand, we did not observe a similar phenomenon in Google Maps traffic volume. This result suggests that the anomaly indices of traffic volumes are different between services.

We also investigated statistical properties of the number of anomaly areas and their sizes. We show the results in Twitter uplink traffic volume in Figs. 4-6.

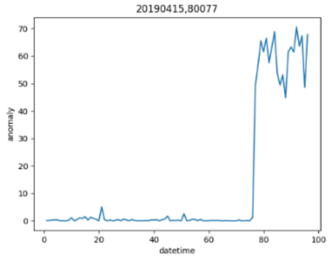


Figure 1: Anomaly index of uplink traffic volume on April 15

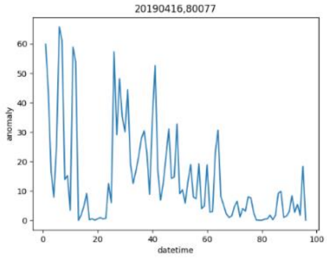


Figure 2: Anomaly index of uplink traffic volume on April 16

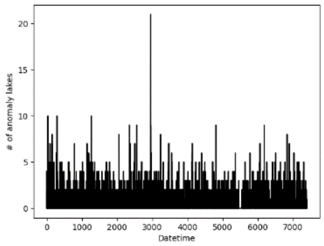


Figure 3: Number of high anomaly areas with red color in Twitter uplink

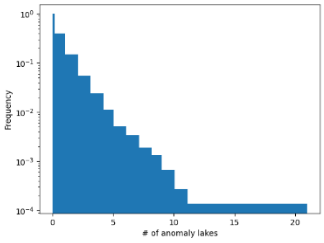


Figure 4: Histogram (CCDF) of # of anomaly areas

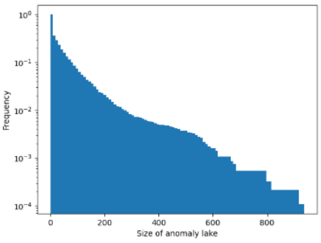


Figure 5: Histogram (CCDF) of the size of anomaly area

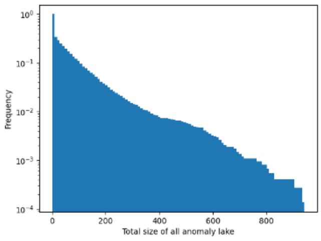


Figure 6: Histogram (CCDF) of the total size of anomaly area

4. Conclusion

The impact of event can be evaluated through anomaly index by Hotelling's T^2method. For future work, it is interesting to categorize events using the size and duration of the anomaly areas. It is also important to investigate the correlation between services.

References

[1] H. Hotelling, "The generalization of Student's ratio," Ann. Math. Statist. 2 (3): 360–378 (1931).

[2] BBC NEWS, "Notre-Dame: Massive fire ravages Paris cathedral"
https://www.bbc.com/news/world-europe-47941794 (Accessed: June 18, 2023).

[3] Visualizations of traffic anomaly in Paris:
https://www.youtube.com/watch?v=IuJBxQ05HZM
https://www.youtube.com/watch?v=fBsGoooP90A
https://www.youtube.com/watch?v=So_M3Dw5aIg
https://www.youtube.com/watch?v=UuWn9QLqlvo
https://www.youtube.com/watch?v=jvuIyBMALZY (Accessed: September 22, 2023)

Data in the woods: analyzing unpopulated areas frequentation and users with mobile data information

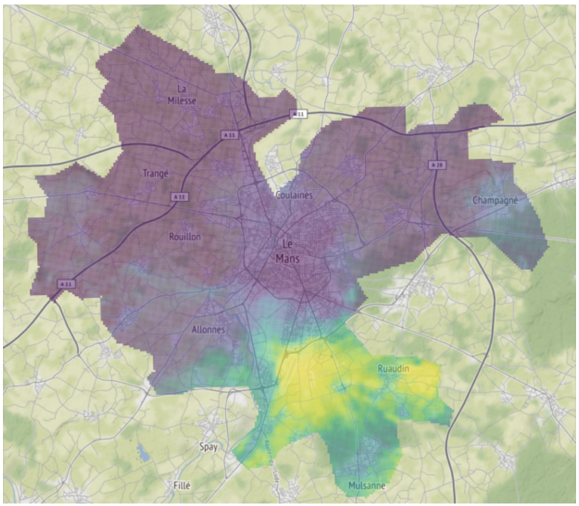
François-Michel Le Tourneau
CNRS – UMR 8586 PRODIG francois-michel.le-tourneau@cnrs.fr

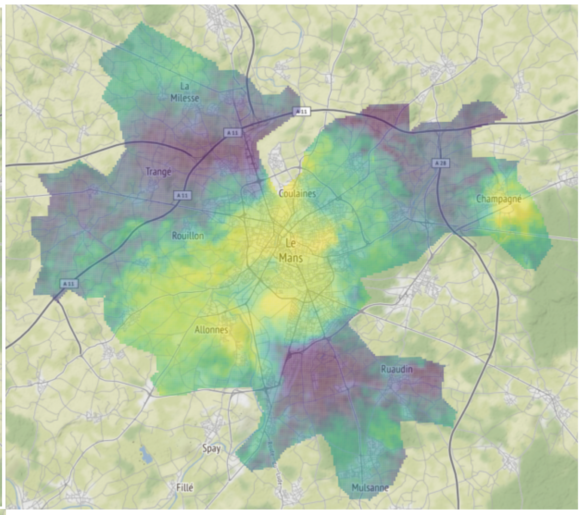
Laetitia Gauvin
IRD– UMR 215 PRODIG laetitia.gauvin@ird.fr

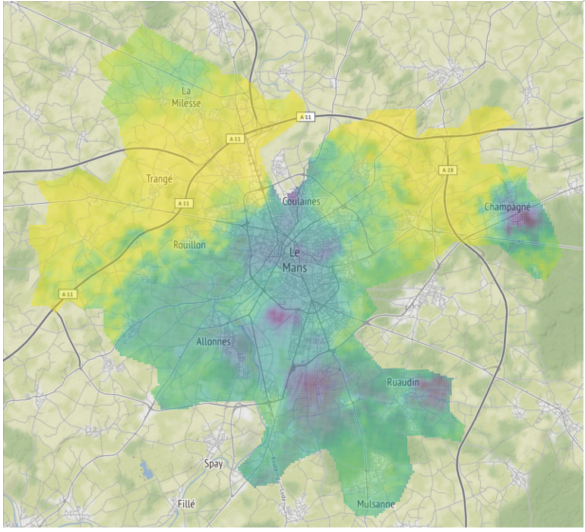
Introduction

In France, much like in numerous other countries, a substantial portion of the land remains devoid of permanent residential settlements. However, this does not imply that these areas are deserted or unused. Unpopulated regions encompass diverse landscapes, including agricultural fields, recreational spaces, industrial zones, and even natural wilderness areas. Census data is used to characterize areas based on resident population but provides little information on sparse or unpopulated regions. However, there is a need to describe the number of visitors and the use of these spaces, for instance to promote sustainable tourism by monitoring the number of visits, or conversely to promote less visited regions, which could help relieve the stress for the most visited areas. The study aims to identify distinct patterns of mobile app consumption in unpopulated areas and compare them to patterns observed in populated areas, creating a dynamic map of human activity complementing census data. Our study endeavors to discern variances in mobile data usage between these unpopulated expanses and densely inhabited regions, thereby constructing alternative characterizations of their usage and visitation patterns.

To this aim, we perform a detailed analysis of time-and space-resolved app usage data, using a multidimensional dimension-reduction technique called Nonnegative Tensor Factorization. This allows us to extract typical patterns of app usage, and induces a soft clustering of cells into units that appear to match several types of land usage.







The city of Le Mans provides a good example of what may be found. It features a substantial proportion of unpopulated or sparsely populated areas. We find that the emerging structures are not trivial concentric clusters separating the city center and the surrounding areas. Notable structures include:

A central zone displaying higher activity, encompassing Champagné, a small urban area slightly outside the city center (11 km), but only during weekends. Champagné is grouped with less populated zones during weekdays.

Northern zones, which are less populated, exhibit a higher usage of the Waze app for driving during weekends.

On weekends, a distinct structure emerges in the southern zone, also less populated. This zone includes both horse and car racetracks and is characterized by increased usage of Apple apps and WhatsApp.

Here we provide a list of the key findings for the study over all the cities and cite a few examples.

City-wise Static Analysis (averaging over time of the day):

- Most cities exhibited a concentric clustering pattern, indicative of differences in population density.
- Weekdays generally yielded a higher number of clusters compared to weekends, signifying more diverse app usage patterns during the week.
- No single app emerged as a strong discriminator among clusters, emphasizing the complexity of app usage patterns.
- Some cities displayed minimal differences in app usage between weekdays and weekends, warranting further investigation.
- The number of clusters varied across cities and did not appear to correlate with city population size.

CIty-wise Time-Resolved Analysis

- The inclusion of the time dimension revealed more subtle patterns, highlighting differences between weekdays and weekends that were previously undetected.
- Bordeaux exhibited a polycentric structure, with distinct app usage profiles during weekdays and weekends.
- Le Mans exhibited a polycentric structure as well and showed a dynamic organization with varying app usage patterns across different zones.
- The time-resolved approach allowed for a better understanding of activity patterns in areas with different population densities.

Country-Wide Analysis:

- Expanding our analysis to the entire country revealed shared app usage patterns among cities.
- However, the signal-to-noise ratio decreased, and only two components were extracted for the whole country.
- The country-wide analysis identified a shared component capturing together many city centers and airports during weekends.
- One weekday component showed similarities between the Old Port of Marseille and the science university campus of Luminy, while in Bordeaux, it grouped together various components identified in city-level analysis.
- The findings emphasized both high-level similarities across the country and the benefits of city-wise analysis for enhanced resolution.

In summary, our analysis demonstrated the value of considering the time dimension and conducting city-specific investigations to better understand mobile app usage patterns in diverse urban environments. The time-resolved approach revealed nuanced behavioral insights and highlighted the complexity of app usage across different unpopulated or sparsely populated regions and across timeframes. More precisely, some neighboring unpopulated areas we grouped in different components, showing different app usage profiles. Some unpopulated areas were also grouped with populated areas, suggesting some possible mobility patterns that may be investigated through app usage analysis. Finally, while country-wide analysis uncovered shared patterns, it underscored the importance of city-level analysis for a more detailed and accurate characterization of activity spaces.

Socio-Economic and Demographic Patterns in Video Streaming: Unveiling City Dynamics in France

Aslıgül Aksan*, İrem Betül Koçak*, and Oğuz Yücel*
aksan20@itu.edu.tr, kocaki@itu.edu.tr, yucelog@itu.edu.tr
* Management Engineering Department, Istanbul Technical University, 34469 Maçka, Istanbul, Turkey.

I. Introduction

This study is part of the challenge of The NetMob23[1], which focuses on cities in France and the internet usage data of different platforms and services.

We aim to understand whether the choice of streaming platforms in two cities in France reflects the socio-economic situation of the people or the demographics of specific communes. While looking at this, we will examine whether there is segregation in the urbanized areas of cities in France. Paris, as France's largest and most diverse city, is our primary focus, with its population of approximately two million residents across 20 districts. We won't consider the entire Paris arrondissements, which exceeded 10 million in 2019[1]. This approach allows for meaningful comparisons with Montpellier, a city with a population similar to some Paris districts, providing insights into the French Riviera's residents compared to Paris.

II. Methodology

The objective of this study is to analyze the usage of different video content platforms, classified as paid and unpaid, in selected cities and the communes of those cities. Since the dataset [2] belongs to 2019, we have neglected YouTube Premium as it was not popular among people at that time. Instead, we have selected YouTube and Dailymotion as representative choices for unpaid video content platforms, including watching videos, TV series, and movies. To represent paid streaming services, we have included Netflix, Apple Video, and Orange TV. In order to simplify the analysis, we have excluded browser streaming or TOR, as these platforms provide both paid and unpaid pirate services.

In this study, downlink data for the mentioned platforms was collected at 15-minute intervals for each day between April 1, 2019, and April 30, 2019. The total daily usage values of these platforms were used to aggregate the points in the cities where each platform had the highest data usage. This was done for two different cities.

We will begin by examining maps of these cities to visualize the dominant platform usage patterns and how these patterns change over the month of April, the time span we choose to analyze, for our exploratory data analysis. Following that, we will employ the statistical tool ANOVA [3] to gain insights into the differences between communes and cities, and subsequently, we will discuss the results of both analyses.

III. Exploratory Data Analysis

We are primarily investigating the usage patterns of these platforms throughout April. The data shows that people use these platforms more frequently during non-work hours, such as lunch breaks and after work. This suggests that these platforms are popular leisure activities for individuals. Interestingly, internet usage on these platforms is generally lower on weekends compared to weekdays. These findings are in line with the Acumen Daily Report on Youth Video Diet, which surveyed individuals aged 13-24. The survey indicated that people in this age group tend to watch videos when they are bored or have free time, with a preference for doing so after work or school and during lunch breaks. Therefore, our findings confirm that this activity is primarily a leisure pursuit, with people choosing to engage in it mainly on weekdays after school or work.

Our analysis aims to extract insights about city demographics and neighborhood socio-economic conditions based on a limited snapshot of people's leisure activities. Our objective is to identify any variations between different days, shedding light on how people respond to daily life and their streaming platform preferences. For instance, we noted a significant uptick in Apple Video usage on April 14, 2019. Further investigation revealed that this spike coincided with the premiere of the final season of HBO's Game of Thrones on Apple Video. This finding illustrates how the hype surrounding a TV series can lead to a remarkable increase in platform usage. It underscores the impact of trends and hype on people's online video-watching habits.

In our study, we assess the similarity between the cities in terms of platform usage by employing the dissimilarity index. Originally introduced by Duncan and Duncan in 1955 to measure ethnic and population disparities[4], this index ranges from zero to one. A value of zero signifies complete integration, while one indicates complete segregation between paid and unpaid platform usage. Over the 30-day period studied, the values in our analysis fall within the range of 0.001 to 0.007. These small values indicate that the cities exhibit a high degree of similarity in their platform usage patterns.

When examining daily changes in platform usage for both cities, we observe that YouTube and Netflix are consistently the most popular platforms. Given the dominance of these platforms, it's challenging to distinguish clear differences between paid and unpaid platforms, as indicated by the dissimilarity index analysis. This pattern also holds true for Apple Video, DailyMotion, and Orange TV. Apple Video falls in the middle in terms of usage in both cities, while DailyMotion and Orange TV are the least used platforms. However, a significant contrast emerges in the daily scales of platform usage between the two cities. This contrast aligns with the differences in population size and city scale.

From a socio-economic perspective, we notice that in economically developed arrondissements of Paris, such as Champs-Élysées, all platforms experience more extensive usage. Conversely, in areas further from the city center, such as La Courneuve, where socio-economic challenges are more pronounced, platform usage is significantly lower. Figures 1 and 2 visually depict the spread of dominant applications across arrondissements in the City of Paris and Montpellier.

When we examine each region of City of Paris in Figure 1, we observe that only three platforms are the most used. YouTube and Netflix stand out as dominant platforms, while Apple Video is

[1] The NetMob 2023 Data Challenge is a competition aiming at deriving and exploiting insights from massive mobile service consumption information. Details can be found at https://netmob2023challenge.networks.imdea.org

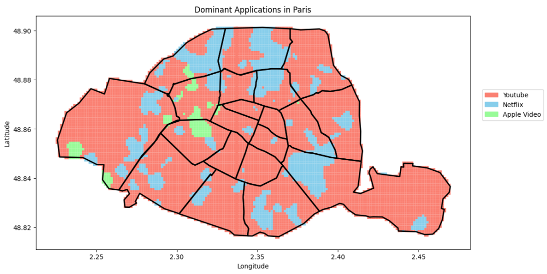


Fig. 1. Dominantly Used Platforms in Paris

scattered specific parts of the central and western of the city. YouTube dominates the in the central and western parts of the City of Paris, while Netflix is the most used platform. The regions where Apple Video is most used are the northwest and central areas. DailyMotion or Orange TV, on the other hand, are not the most used platforms in any region of the city. Similar to Paris, the number of most used platforms is also three in Montpellier, with YouTube and Netflix exhibiting a dominant spread in Figure 2. However, domination of Apple Video is observed in a much smaller number of regions compared to Paris.

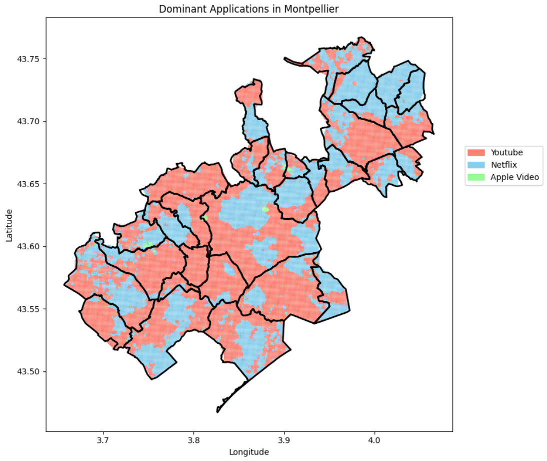


Fig. 2. Dominantly Used Platforms in Montpellier

IV. Anova Analysis

As evident from the map in the preceding section, there is a significant disparity in population between the cities. Therefore, we'll be working with average data for each unit, which is normalized by its respective population. We plan to use ANOVA to detect any differences in the average usage data between the Paris commune with 20 arrondissements and Montpellier. This approach will help us gain insights into the demographics and socio-economic conditions of both similar and distinct regions within the cities.

In this study, an ANOVA analysis was conducted to examine the usage of streaming platforms, categorized as paid or unpaid, across 20 arrondissements in Paris and the city of Montpellier. A total of 21 groups were analyzed initially based on their usage of paid applications. The analysis revealed a significant difference among these groups, with a high F-value of 357,887.347 and a very low p-value of 0.000, indicating strong statistical significance.

To further explore the differences among the groups, a Tukey's HSD test was applied, which showed that most group differences were statistically significant (p is lower than 0.001). This analysis helped identify specific regions that exhibited differences in paid platform usage.

Additionally, when assessing unpaid platform usage, another ANOVA analysis was conducted, revealing significant differences among the regions (p-value = 0.000, F-value = 265,025.66). Subsequently, a post-hoc test was performed to identify regions with distinct unpaid platform usage patterns.

Specific findings include:

1. Arrondissement 7 and Arrondissement 12 showed significant differences in paid platform usage. Although Arrondissement 12 had a larger population, age distributions were relatively similar. Arrondissement 7 had a higher proportion of individuals with higher education.

2. Arrondissement 19 and Arrondissement 20 exhibited almost no difference in paid platform usage. They had similar population distributions, with slight variations in age and education.

3. Arrondissement 14 and Arrondissement 17, which did not differ significantly in unpaid platform usage, displayed significant demographic and economic differences. Arrondissement 14 had more students, while Arrondissement 17 had more employees.

4. Arrondissement 12 closely resembled Montpellier in both unpaid and paid platform usage. Both regions had a young population, but Arrondissement 12 had a higher employment rate.

Overall, the study identified significant variations in platform usage patterns among regions, shedding light on demographic and economic factors influencing these differences.

V. Conclusion

This study focuses exclusively on people's usage of specific streaming platforms to reflect demographics and the socio-economic states of people living in Paris and Montpellier. It sheds light on mobile data usage patterns connected to leisure activities, aiming to pinpoint unique usage areas.

The research highlights that streaming platform choices not only mirror neighborhood socio-economic conditions and demographics but are also influenced by global trends. However, our mid-level study duration offers limited insights into broader neighborhood aspects. To gain a deeper understanding, analyzing yearly and weekly streaming service data could provide insights into how these cities align with global trends, offering clues about age and education levels.

Moreover, an individual service analysis reveals Orange TV's football streaming. Considering the presence of French football teams like PSG, Lille, and Lyon in the 2019 Champions League, and the significance of football as a leisure activity in less privileged neighborhoods, the availability of Orange TV connects to socio-economic situations. This, however, necessitates smaller, more focused studies for meaningful insights.

References

[1] T. N. I. of Statistics and E. Studies. Population en 2019. [Online]. Available: https://www.insee.fr/fr/statistiques/6543200/

[2] O. E. Martinez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," *arXiv preprint arXiv:2305.06933*, 2023.

[3] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[4] O. D. Duncan and B. Duncan, "A methodological analysis of segregation indexes," *American Sociological Review*, vol. 20, no. 2, pp. 210–217, 1955.

Graph Neural Network-based Models for Mobile Network Traffic Prediction

Duc-Thinh Ngo[*][†], **Ons Aouedi**[†], **Kandaraj Piamrat**[†], **Thomas Hassan**[*], **Philippe Raipin**[*]
[*] Orange Labs, Cesson-Sévigné, France
[†] Nantes University, École Centrale Nantes, IMT Atlantique, CNRS, INRIA, LS2N, UMR 6004, Nantes, France
{ducthinh.ngo, thomas.hassan, philippe.raipin}@orange.com
{ons.aouedi, kandaraj.piamrat}@inria.fr

1 Introduction

Network traffic prediction is an essential problem in network management. Accurate traffic prediction will provide network operators with insight into the network state and suggest appropriate actions to optimize the network. Similarly, in the context of a zero-touch network, traffic prediction can continuously inform the monitoring components about future network conditions, enabling timely decision-making. Dynamic network slicing serves as a prime application. Within the core of a B5G network where network functions are fully virtualized, network slicing allows network providers to offer their services to diverse clients while maintaining the Quality of Service (QoS) to meet clients' specific requirements. Moreover, dynamic network slicing offers enhanced flexibility for operators to dynamically orchestrate the slicing to address evolving client needs. Consequently, accurate traffic prediction empowers dynamic network slicing by providing valuable insights about the future client usage pattern, facilitating the preparation of slicing configurations to adapt to potential client demands.

Predicting network traffic can be formulated as a time series forecasting problem where one needs to find the best forecasting model using historical traffic data to generate the most accurate future traffic predictions. However, this approach risks losing the spatial information of the data. Considering the network traffic per evolved Node B (eNodeB), e.g., one time series per eNodeB to represent the network traffic of an eNodeB, there exists certainly a spatial correlation between these time series. For instance, geographically proximate eNodeBs can exhibit similar traffic volumes due to comparable population densities. Additionally, depending on the mobility patterns of users during network usage, network traffic state can be *diffused* from one eNodeB to another. Overall, the presence of spatial correlation within the network traffic prediction raises the need of exploiting another dimension of the data beyond the temporal dimension alone. Consequently, spatio-temporal traffic prediction has recently emerged to address network traffic prediction.

Graph Neural Networks (GNNs) play a crucial role in solving the spatio-temporal prediction problem [3–5, 7–9]. Thanks to the invariance to permutation of node orders, GNNs can extract structural information from graphs and embed it within node embeddings. In the context of spatio-temporal prediction, GNNs facilitate the exploitation of the data patterns along the spatial dimension, complementing the temporal characteristics.

Similarly, road traffic prediction has also been formulated as a spatio-temporal forecasting problem. Researchers have developed various frameworks harnessing GNNs to perform multi-step ahead forecasting given historical sub-sequences of transport traffic volume (or average speed) per sensor.

In this data challenge, our objective is to benchmark several spatio-temporal models on the network traffic prediction task using the NetMob23 Data challenge dataset [6]. It is important to note that most of these models have only been benchmarked on road traffic datasets. Although the two scenarios differ, the problems in both can be framed as a spatio-temporal traffic prediction task. However, due to the distinct characteristics of the network traffic dataset, it is necessary to process this data and reframe the problem within the context of spatio-temporal network traffic forecasting. Thus, our contributions in this work are the benchmark of multiple spatio-temporal traffic prediction models, which were mostly evaluated on road traffic data, on the network traffic dataset with different evaluation protocols.

2 Methodology

2.1 Data pre-processing

Firstly, the provided data was in the form of aggregated traffic distributed across regular tiles. While this format allows organizing data in a grid, it does not reflect the realistic scenario where data is collected and aggregated per eNodeB, which is essential for forecasting traffic and performing radio resource management. To address this issue, we combine the spatio-temporal data with the eNodeB map [2] and extract the network traffic only for tiles where eNodeBs are present. Since the methodology to calculate the coverage probability for each tile was not detailed, it is difficult to reverse the calculation to obtain the per-eNodeB traffic. However, we can assume that the generated traffic in tiles attached to eNodeBs maintains the traffic patterns corresponding to those eNodeBs.

Secondly, although diverse mobile application traffic data is provided, it is not necessary to predict the traffic for every one of them. Therefore, we took inspiration from the Quality of Service Class Identifier (QCI) [1] to select 5 specific applications to highlight distinct classes of QoS requirements: (1) Apple Video - conversational voice/video, (2) Fortnite - real-time gaming, (3) Netflix - buffered streaming, (4) Instagram - TCP-based and live streaming and (5) Microsoft Mail - TCP-based application. This selection allows us to capture the varied demands placed on the network infrastructure, considering factors such as peak hours, geographical locations, and different types of content consumption. For example, while Microsoft Mail experiences high demand during work hours and in the workplace, Netflix may exhibit high consumption even at night, on weekends, and in resi-

Table 1: Multi-step prediction results on the aggregated traffic dataset.

	15min			30min			45min		
Metrics	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
HA	18.12	85.62	37.02	18.12	85.62	37.02	18.12	85.62	37.02
VAR	17.47	90.09	33.46	18.19	97.96	35.25	18.09	100.09	35.33
AGCRN	11.76	45.22	24.66	13.11	51.61	27.12	14.07	58.32	28.83
ASTGCN	11.56	37.93	24.00	12.80	52.30	26.60	13.92	46.24	28.44
DCRNN	10.89	38.11	23.26	12.00	42.96	25.48	12.62	46.52	26.67
GMAN	11.18	52.33	23.70	12.04	55.91	25.29	12.43	59.13	26.03
GWN	**10.82**	**34.79**	23.25	**11.86**	38.63	25.32	12.45	41.66	26.37
MTGNN	10.96	34.85	23.25	11.91	**37.45**	**25.28**	12.43	**40.92**	26.33

dential areas. Each application has its unique usage patterns and characteristics, which contribute to the overall diversity of the filtered dataset. By incorporating these diverse applications, our study can provide valuable insights into the challenges and opportunities associated with network traffic prediction and management across different service classes.

We evaluate multiple model architectures using a dual approach. First, we train individual models per framework to predict traffic for five different applications, gauging their ability to generalize across diverse scenarios. Second, we train separate models for each framework to predict aggregated traffic from all five application patterns, testing their capability to handle complex traffic dynamics.

2.2 Graph construction

Similar to road traffic prediction, we have two main strategies for pre-defining the graph for spatio-temporal prediction: (1) spatial proximity graph and (2) temporal similarity graph. The former involves computing the distance between each pair of base stations and then determining the existence of an edge based on a distance threshold. The latter entails extracting a representative pattern from each time series and computing the distance between these patterns. The edges are then defined as in the spatial proximity graph using a threshold or by selecting the k nearest neighbors.

2.3 Spatio-temporal models

In this challenge, we select six spatio-temporal models that have demonstrated robustness in the field of road traffic prediction: DCRNN [5], GWN [7], AGCRN [3], ASTGCN [4], GMAN [9], MTGNN [8].

3 Results

Overall, on one hand, GWN and MTGNN outperform the others according to MAE and RMSE in almost every experiment. On the other hand, ASTGCN coupled with the temporal similarity graph, GMAN with the spatial proximity graph have the least mean average percentage error in predicting traffic of individual applications. Additionally, we also achieved noteworthy results with GMAN on the aggregated traffic data in the 45-minute horizon. In this regard, GMAN's performance is on par with that of MTGNN. It is also noticed that leveraging the temporal similarity graph significantly enhances the MAPE for ASTGCN and DCRNN, leading to substantial improvements in prediction accuracy. However, applying the same strategy to GMAN has the opposite effect, causing a notable decline in prediction performance according to MAPE. We summarize results on the aggregated traffic dataset in the table 1.

References

[1] 3GPP TS 23.203 Policy and Charging Control Architecture V17.2.0, December 2021.

[2] The National Frequency Agency. Cartoradio: The map of radio sites and wave measurements. URL `https://cartoradio.fr`.

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

[4] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 922–929, 2019.

[5] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018.

[6] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

[7] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 1907–1913. AAAI Press, 2019. ISBN 9780999241141.

[8] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 753–763, 2020.

[9] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1234–1241, 2020.

Online Learning for Network Traffic Forecasting

Javier Rivas García[1⋆], Livia Elena Chatzieleftheriou[1,2], Sergi Alcalá-Marín[1,2], Albert Banchs[1,2]
[1] Universidad Carlos III, Madrid, Spain
[2] IMDEA Networks Institute, Spain
⋆ 100406248@alumnos.uc3m.es, {livia.chatzieleftheriou,sergi.alcala,albert.banchs}@imdea.org

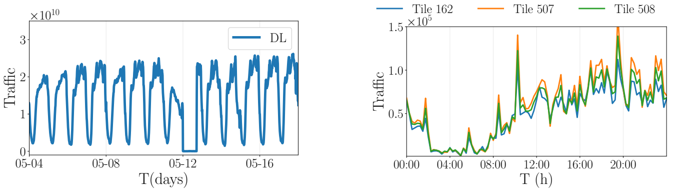
I. Introduction

Traffic forecasting plays a vital role in the efficient management of B5G and 6G networks, given the increasing demand for outstanding performance to ensure the required Quality of Service (QoS) across various scenarios. These scenarios encompass ultra Reliable Low-Latency Communications (uRLLC), enhanced Mobile Broadband (eMBB), and massive Machine-Type Communications. As these networks strive to achieve ambitious goals, rapid traffic growth and the provision of multiple services with distinct requirements are expected. Consequently, traffic patterns and network capacity requirements undergo continuous changes. By employing traffic forecasting, the complexity of upcoming networks can be effectively managed, thereby preempting potential problems and preventing degradation of the user experience. In this context, Online Learning (*OL* emerges as a promising methodology due to its robustness in dynamic environments and computational efficiency, enabling accurate results and facilitating resource allocation and network management tasks.

Challenges. Traffic is unpredictable and statistically difficult to characterize. Lack of available historical data in certain locations complicates more the task as well, because it is impossible to depict traffic patterns in those geographical points. Our purpose is to present a set of techniques to forecast traffic using *OL* algorithms, which converge quickly to the optimal prediction and adapt to any possible context. Thus, we set an *OL* scenario where, contrary to traditional Machine Learning (ML) settings, training and test makes no sense because the algorithms learn from data streams.

Experimental evidence. In Fig. 1a we depict total downlink Instagram traffic in Paris during 2 weeks. We observe that, as expected, the amount of total traffic tends to periodicity as result of human regular activities. However, we cannot leave aside the presence of an unexpected downfall during a certain day. This kind of unpredictable events which appear as downs or peaks can occur from time to time as a result of many different events such as network outages or unexpected excessive high demands, demonstrating the need for algorithms that quickly adapts to this anomalies.

In Fig. 1b we depict downlink Instagram traffic during 1 day in 3 different geographical positions. In the short term, instability and high variability appears from sample to sample. Constant ups and downs without a time correlation makes impossible to characterize the traffic data in any particular way. This makes it perfect to test our algorithms over a challenging environment. Notice traffic patterns similarities between different tiles meaning same applications traffic volume does not change that much between similar geographical zones.



(a) Paris (Total)

(b) Paris (3 discrete tiles)

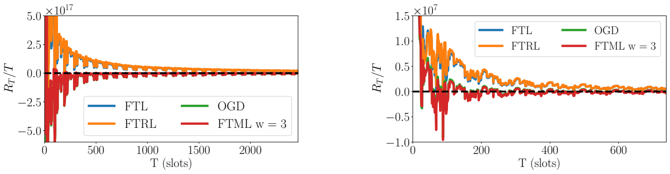
Fig. 1: Instagram network dowlink traffic in Paris (a) total traffic for 2 weeks, and (b) per tile traffic, for 1 day.

Our contributions. Our desire is to ensure computation efficiency for large-scale datasets. Common ML models are less dynamic in front of changes in the data. We design the "Follow The Moving Leader (FTML)", an OL algorithm that considers only recent past samples, adapting the model to the last changes in the data. Despite the periodicity that comes as a result of human activity, we design our algorithms against adversarial chosen data, dismissing any forecast strategy based on statistical models. This ensures a provably good performance even for worst-case scenarios, and demonstrates our predictor's reliability.

II. Online Learning for traffic forecasting

OL is a powerful tool for learning tasks, because it is lightweight and fast, and if we study the geometry of our problem, we can obtain solutions with performance guarantees. *OL* problems, consider a sequence of consecutive rounds t with time horizon, T. Learner chooses $x_t \in S$ every round t, which in our setting captures the predicted traffic. Once, the decision is made, given a cost function f, the actual cost f_t is revealed. The player incurs a cost of $f_t(x_t)$ which measures disparity from prediction and real value .

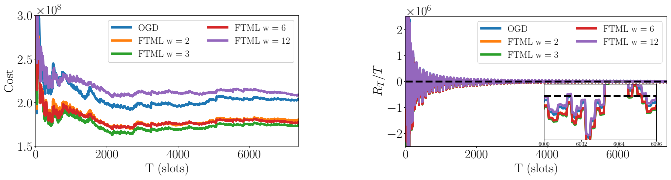
Setting. Our cost function is the *Mean Squared Error* [1]: $MSE(T) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2$. The performance of online algorithms is characterized by their *regret* [1]. Let x_t^* be the optimal decision in hindsight (offline benchmark). Regret is defined as it follows: $Regret(T) := \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_t^*)$, and the difference between the cost incurred by the online decision x_t and the optimal solution in hindsight. The main goal in *OL* problems is to minimize the regret. An OL algorithm performs well if its regret grows sublinearly to T, *i.e.*, if $\lim_{T\to\infty} \frac{Regret(T)}{T} = 0$. This implies, that on average the algorithm performs as good as the offline benchmark.



(a) Total traffic

(b) Tile traffic

Fig. 2: Regret of online algorithms against *OPS* for K=24, for (a) total traffic and (b) tile traffic Instagram Downlink data.



(a) Cost

(b) Regret *OPS* K = 24

Fig. 3: (a) Cost and (b) Regret of *FTML* for multiple windows' size W and OGD for Paris Tile 162 Instagram DL traffic.

When this condition is fulfilled, the algorithm has "no regret" against the offline benchmark.

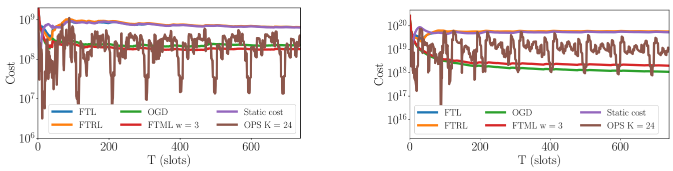
Offline benchmark. We compare against the *Optimal Periodic Static (OPS* benchmark [2]. *OPS* generalizes the state-of-the-art, beholding as special cases, both the static and the dynamic benchmark, which are the most extended benchmarks in OL [1]. It considers a partitioning of the time-horizon in K periods, and for each of them finds the optimal static solution in hindsight.

Online benchmarks. We compare against the following well-known online algorithms [1]. *Follow the Leader (FTL)* predicts what would have been optimal in past rounds, *i.e.*, minimizes the sum of cost functions at previous rounds, *i.e.*, $x_t = \arg\min \sum_{i=1}^{t-1} f_i(x)$. *Follow the Regularized Leader (FTRL)* includes a regularization function $R(x)$ based on the geometry of each problem, to promote stability of the solution. $x_t = \arg\min \sum_{i=1}^{t-1} f_i(x) + \frac{1}{\eta} R(x)$. *Online Gradient Descent (OGD)* is the online version of the "gradient descent" algorithm. In each iteration the algorithm takes a step in against the direction of the gradient, *i.e.*, $y_{t+1} = x_t - \eta_t \nabla f_t(x_t)$, and then projects back to the feasible space, *i.e.*, $x_{t+1} = \Pi_S(y_{t+1})$.

Our algorithm: Follow The Moving Leader (*FTML*). We present a novel algorithm to solve traffic forecast difficulties. The idea is to keep minimizing previous cost functions but considering windows consisting on the most recent W data points, and weighting recent samples exponentially more compared to the past ones. For $\mathbf{w}$ a weight vector and W the number of past samples considered, FTML's prediction is given by $x_t = \arg\min \sum_{i=t-W}^{t-1} f_i(x) * w_i$. *FTML* generalizes state-of-the-art, having as special cases two of the most powerful OL algorithms. More specifically, it reduces to *FTL* and to *OGD* or $W = T$ and $W = 1$ respectively, marking its full potential in settings where data changes.

III. Evaluation

We present the cost and the Regret for the algorithms above. Tests were conducted over Paris Instagram Downlink traffic data considering both the total traffic and the per-tile traffic. Total traffic presents stable patterns with little variations due to specific events, while per-tile traffic shows highly changing dynamic from time slot to time slot. By considering both, we evaluate our algorithm under different conditions.



(a) Cost , Tile traffic

(b) Costs, Total Traffic.

Fig. 4: Cost comparison

From Fig. 2 we can see that the Online algorithms used for the Netmob challenge, perform asymptotically as good as the static decision made in hindsight. However, in very dynamic environments such as network traffic, may not be the optimal performance metric. Considering traffic data periodicity, *OPS* gives an adequate metric to measure our algorithms performance as it permits us to adapt the metric to the data. In Fig 2a we see that the Regret obtained by all algorithms converge to zero quickly against static benchmarks. Contrarily, for a more refined metric such as *OPS* (Fig 2b only OGD and FTML, our algorithm, achieves asymptotic results showing adaptability to changes.

Fig. 3 suggests that *FTML* xhibits similar traits to *OGD* hen the past sample count, W, is low. For $W = 1$, *FTML* ssentially becomes *OGD* However, as the window size increases (e.g., $W = 12$), *FTML* urpasses *OGD* n performance. Adjusting the window size W allows tailored adaptation to specific datasets.

In Fig.4, *OPS* truggles to characterize highly dynamic data compared to stable data. Noiseless data is easier to forecast, and *FTML* chieves accurate predictions with just $W = 1$ sample. Notably, advanced algorithms like *FTL* nd *FTRL* end to stagnate after considering a certain amount of past data, making them less suitable for large-scale periodic data.

IV. Conclusions

We addressed traffic forecasting in the online (OL) context by introducing FTML, a parametric algorithm. FTML efficiently adapts and predicts upcoming traffic based on recent samples, with a parameter controlling the sample count. Its parametric nature merges FTL and OGD algorithms, enabling swift adaptation to traffic changes. FTML's versatility suits various applications and time series, ensuring both robustness and adaptability to evolving data.

V. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreements no.101017109 "DAEMON".

References

[1] E. Hazan, "Introduction to Online Convex Optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2015.

[2] L.E. Chatzieleftheriou, A. Destounis G. Paschos, and I. Koutsopoulos, "Blind Optimal User Association in Small-Cell Networks," in *proc. IEEE International Conference on Computer Communications - INFOCOM*, 2021.

Désordre; a Framework for Real-Time Detection and Prediction of Social Unrest and Catastrophes Using Mobile Data Traffic

Noha Gamal(1)
ngamal@nu.edu.eg

Mina Atef Yousef(1)
Myousef@nu.edu.eg

Ahmed El-Mahdy(1)
aelmahdy@nu.edu.eg

(1)School of Information Technology and Computer Science, Nile University, Giza, Egypt.

Abstract

In the contemporary digital landscape, telecommunications act as the backbone that facilitates our daily interactions and enables societal functionalities. This research proposal introduces "Désordre," an innovative theoretical and computational framework specifically designed to revolutionize the field of telecommunication traffic dynamics. By synergistically melding three pivotal domains—telecommunication network optimization, real-world event-driven social media behavior, and predictive modeling of anomalous traffic patterns—the research seeks to transcend traditional boundaries. Anchored in a multi-layered, interdisciplinary framework, "Désordre" employs state-of-the-art technologies and methodologies, including Fitted-SEIR epidemiological models and advanced anomaly detection algorithms. The project aims to provide actionable intelligence in real-time social event prediction, focusing specifically on the bustling urban landscapes of Paris and Lyon. Through this holistic approach, the research aspires to unveil an unprecedented tapestry of insights into spatiotemporal dynamics, behavioral changes, and the complex interplay of factors that influence telecom traffic.

Introduction

Telecommunications serve as the backbone of modern society, orchestrating a complex symphony of digital interactions [1]. Our research team aims to elevate this symphony by introducing "Désordre," a comprehensive framework that seamlessly weaves together big data analysis, unsupervised machine learning for anomaly detection [2], the mathematical foundations of epidemic theory, and graph theory to offer a holistic approach to mobile data traffic management. The central research question driving this study is multifaceted: How can the spatial dynamics within telecom network infrastructure be effectively harnessed for traffic propagation modeling? What are the parallels between traffic in telecom networks and the spread of epidemics? How can traffic data serve as a resource for disaster response and population mobility monitoring? And what strategies exist for the overall optimization of telecom networks?

Subsystems of the Proposed Framework

Anomaly Detection in Telecom Data Traffic: We focus on analyzing traffic data from social category applications over specific event periods in Paris and Lyon. The goal is to capture traffic anomalies associated with significant urban events [3][4], leveraging the NetMob23 dataset [5].

Epidemiological Modeling for Traffic Analysis: Utilizing agent-based Susceptible-Exposed-Infectious-Removed (ABM-SEIR) models, we aim to understand the dynamics of traffic flow and anomalies within cities.

Visualization of Traffic Transmission Patterns on Network: Advanced visualization techniques will be employed to interpret complex traffic data and the flow of anomalies within the network graph.

Traffic Load Comparison for Optimization: The research extends to comparing traffic load with maximum load constraints, facilitating actionable recommendations for traffic optimization.

Understanding Telecom Traffic Dynamics

Telecommunications today is a complex interplay of various types of cells like microcells, picocells, and femtocells [6], each serving as a critical node in the intricate web of data traffic. Microcells are predominantly found in densely populated urban areas, providing localized coverage and handling high-density traffic. Picocells extend this coverage into indoor spaces like shopping malls and offices, while femtocells [7] are designed for residential use, enhancing coverage within homes. Understanding the role and interaction of these various types of cells is paramount for optimizing network performance [8].

Handover Processes and Traffic Flow

The process of transitioning traffic from mobile phones to macro-cells, involving several key steps, is a well-orchestrated dance that ensures the continuity and efficiency of connectivity [9]. It relies on monitoring signal quality, interference levels, and various other factors to initiate an intricate handover process [10]. For data sessions, technologies like carrier aggregation and intelligent load balancing algorithms are employed to distribute traffic judiciously, especially during peak usage times [11].

Graph Theory and Traffic Propagation

Visualizing the telecom network as a graph provides a unique perspective on how traffic propagates within the network [12]. In this graph, micro-cells serve as the nodes, and their spatial proximity forms the edges. This representation is dynamic, adapting to changes in network configurations and the movement of mobile devices, and offers a visual framework for understanding how traffic flows within the network, highlighting areas of high connectivity and potential congestion points [13].

Epidemiological Models and Traffic Dynamics

We also explore the analogy between the spread of infectious phenomena and telecom traffic. Using epidemiological models such as the SEIR model [14], we can adapt these frameworks to represent the flow of traffic between micro-cells in a telecom network. This novel outlook allows us to explore how epidemiological agent-based modeling can be adapted to represent the flow of traffic, thereby providing insights into traffic patterns and congestion points.

An Integrated Perspective: SNA and Epidemiological Models

Combining Social Network Analysis (SNA) with epidemiological models offers a more holistic approach to understanding traffic spread in a telecom network [15-16]. This integrated perspective enables more efficient resource allocation and suggests strategies for load balancing, routing enhancements, and infrastructure deployment [17].

Methodology

The framework is initiated by setting up the necessary computational environment and loading essential libraries. It aims to address various challenges in the field of mobile data traffic, focusing on anomaly detection, traffic progression modeling, traffic visualization, and network load comparison.

Data Collection and Preprocessing

The initial step involves collecting data related to mobile traffic for a specific time duration, often around 100 days. We focus on social category applications and specifically track events in cities like Paris and Lyon. The data is filtered based on these criteria and the time interval around each event date, providing a targeted dataset for further analysis.

Anomaly Detection

Once the data is prepared, the framework proceeds to the anomaly detection phase. Here, machine learning algorithms, particularly unsupervised learning methods, are employed to identify anomalies in mobile traffic data. Each detected anomaly is tagged with pertinent information such as its geographic location, the time it occurred, and associated traffic volumes. This information is saved for further investigation and analysis.

Traffic Progression Modeling

After identifying anomalies, the framework transitions to the modeling phase. Here, epidemiological models like SEIR are adapted to the context of mobile traffic. Each area in the city under study is labeled based on a 15-

minute entry, and these labels are used to prepare datasets for the SEIR model. Optimization techniques are applied to find the most suitable parameters for these models. Furthermore, time series prediction methods are used to forecast future behavior based on these parameters. A network graph for the city under study is also generated to facilitate agent-based model simulations for more in-depth analysis.

Traffic Visualization

With the models in place and the city network graph prepared, the next step involves visualizing the flow of traffic. Advanced visualization techniques are employed to map out the traffic flows on the network graph, providing a visual representation that can be easily interpreted. These visualizations offer insights into areas of congestion, high traffic volume, and directional movement, among other factors.

Traffic Load Comparison and Optimization

The final step in the framework involves analyzing the traffic load at each node in the network. If the traffic load exceeds predefined maximum constraints, recommendations are made for network upgrades or load balancing algorithms. The outcome of this phase is a set of actionable insights and recommendations aimed at optimizing the network's performance.

Results and Discussions

Groundbreaking Contributions to Anomaly Detection and Predictive Modeling

The results of our study not only meet but exceed the objectives set forth. By merging three core domains—telecommunication network optimization, social media behavior, and predictive modeling of anomalous traffic patterns—we have achieved a pioneering feat in the study of mobile data traffic. Utilizing Fitted-SEIR epidemiological models, advanced anomaly detection algorithms, and quartile statistical methods, our study unveils a comprehensive understanding of spatiotemporal dynamics and behavioral changes in traffic patterns. For instance, our predictive models boast impressive accuracy metrics of 98% for the Notre-Dame Cathedral fire and 96% for the Yellow-Vests Protests. Our system demonstrated exceptional accuracy and real-world utility when it detected anomalous traffic on May 24, 2019, in Lyon. This traffic spike corresponded perfectly with a parcel bomb explosion that injured 14 people as shown in the visualization presented in Fig. 1. This event validates the system's reliability for real-time crisis detection and its value for public safety and proactive network management.

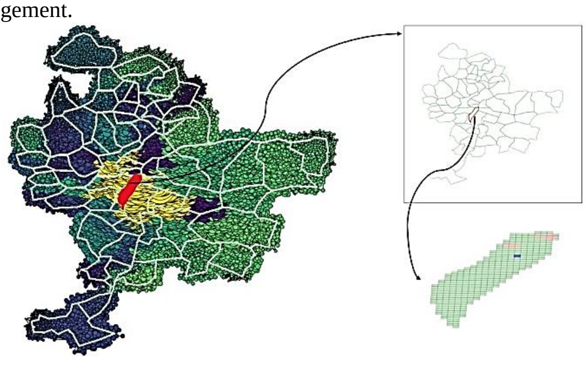


Fig.1 detecting and visualizing the anomalized traffic caused by a barcel bomb explosion emerged in Lyon, France, 24th of May 2019.

Unveiling Seminal Insights into Spatiotemporal Dynamics

Our results indicate significant differences in traffic patterns between planned and spontaneous events. These findings affirm the versatility of our predictive models while offering actionable insights for telecommunication companies and public safety sectors. The research fills a critical gap in understanding contrasting behaviors between different types of events, thereby accentuating the need for adaptive modeling techniques.

Recognizing Limitations and Pointing Towards Future Research

While our models have shown high scalability and robustness, they are not devoid of limitations. Data sparsity and computational needs are areas that require further exploration. Looking forward, we propose the incorporation of real-time analytics, Internet of Things (IoT) devices, and other data streams to refine the predictive capabilities of our models.

Transformative Implications for Industry and Governance

Our results offer revolutionary implications for both the industry and governance sectors. For telecom companies, this could translate to significant gains in operational efficiency and customer satisfaction. For public authorities, the high predictive accuracy of our models could serve as a crucial early-warning system for proactive crisis management, thereby enhancing public safety.

Concluding Remarks

In summary, the results go beyond academic excellence to serve as a multi-dimensional, actionable tool. By synthesizing multiple research avenues into a unified framework, we have laid a robust foundation for future interdisciplinary research. Our results could revolutionize not just the telecom sector but also offer significant advancements in urban planning and public safety measures.

Conclusion

In conclusion, our research introduces "Désordre," a groundbreaking framework that successfully interweaves telecommunication network optimization, social media behavior, and predictive modeling to offer a holistic approach to managing mobile data traffic. Through advanced algorithms and innovative modeling techniques like the Fitted-SEIR epidemiological models, we have achieved unprecedented accuracy in real-time anomaly detection and predictive analytics. Our work holds transformative implications not just for the telecommunication industry but also for public safety and governance, providing a reliable early-warning system for crisis management. While acknowledging existing limitations, we highlight the potential for future research to further refine and extend the capabilities of our framework. Overall, "Désordre" sets a new paradigm, laying a solid foundation for future interdisciplinary endeavors aimed at optimizing telecom networks and enhancing urban life.

Acknowledgments

The authors would like to extend their heartfelt gratitude to AbdElRahman Yaser (a.saeed@nu.edu.eg) and Karim Ayman (k.ayman@nu.edu.eg) for their invaluable support throughout the course of this research project.

References

[1] Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C., & Zhang, J. C. (2014). What will 5G be? IEEE Journal on selected areas in communications, 32(6), 1065-1082.

[2] Gogoi, P., Bhattacharyya, D. K., Borah, B., & Kalita, J. K. (2011). A survey of outlier detection methods in network anomaly identification. The Computer Journal, 54(4), 570-588.

[3] Tober, T. L. (2019). Legacies and Memories in Movements: Justice and Democracy in Southern Europe.

41] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.

[5] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The NetMob23 Dataset: A High-resolution Multi-Region Service-level Mobile Data Traffic Cartography," May 2023, [Online]. Available: http://arxiv.org/abs/2305.06933.

[6] Alnoman, A., Anpalagan, A. (2017), "Towards the fulfillment of 5G network requirements: technologies and challenges". Telecommun Syst 65, 101–116 (2017)

[7] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan and M. C. Reed, (2012) "Femtocells: Past, Present, and Future," in IEEE Journal on Selected Areas in Communications, vol. 30, no. 3, pp. 497-508,

[8] John Strassner, Barry Menich, and Walter Johnson. 2007. "Providing Seamless Mobility in Wireless Networks Using Autonomic Mechanisms". In Proceedings of the 1st international conference on Autonomous Infrastructure, Management and Security: Inter-Domain Management (AIMS '07). Springer-Verlag, Berlin, Heidelberg, 121–132.

[9] Aljeri, N., & Boukerche, A. (2020). Mobility management in 5G-enabled vehicular networks: Models, protocols, and classification. ACM Computing Surveys (CSUR), 53(5), 1-35.

[10] Tangelapalli, S., & PardhaSaradhi, P. (2020). Handover Techniques Analysis for Dense LTE Network. International Journal of Scientific & Technology Research, 9(01).

[11] Learning-Based SON for LTE/LTE-A Macro-Pico HetNets

[12] Guo, T., & Suárez, A. (2020). Fine-grained frequency reuse in centralized small cell networks. IEEE Transactions on Mobile Computing, 20(7), 2367-2378.

[13] Sireesha, R., Rao, C. S., & Kumar, M. V. (2023). Graph theory-based transformation of existing Distribution network into clusters of multiple micro-grids for reliability enhancement. Materials Today: Proceedings, 80, 2921-2928.

[14] Sintunavarat, W., & Turab, A. (2022). Mathematical analysis of an extended SEIR model of COVID-19 using the ABC-fractional operator. Mathematics and Computers in Simulation, 198, 65-84.

[15] Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. science, 323(5916), 892-895.

[16] Ciocarlie, G. F., Lindqvist, U., Nováczki, S., & Sanneck, H. (2013, October). Detecting anomalies in cellular networks using an ensemble method. In Proceedings of the 9th international conference on network and service management (CNSM 2013) (pp. 171-174). IEEE.

[17] Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. L. (2011, August). Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1100-1108).

Left on read: Human behavior characterization based on messaging service applications

Geymerson S. Ramos[1], Gean Santos[2], Douglas Moura[2], Danilo Fernandes[2], Fabiane Queiroz[2]
Osvaldo A. Rosso[2], Razvan Stanica[1], and Andre L. L. Aquino[2]
[1]Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France [2]LaCCAN Laboratory, Federal University of Alagoas, Brazil.
geymerson.ramos@inria.fr; {gean.santos, douglas.moura, dfc, fabiane.queiroz}@laccan.ufal.br;
oarosso@laccan.ufal.br; razvan.stanica@inria.fr; alla@laccan.ufal.br

I. General Problem and Motivation

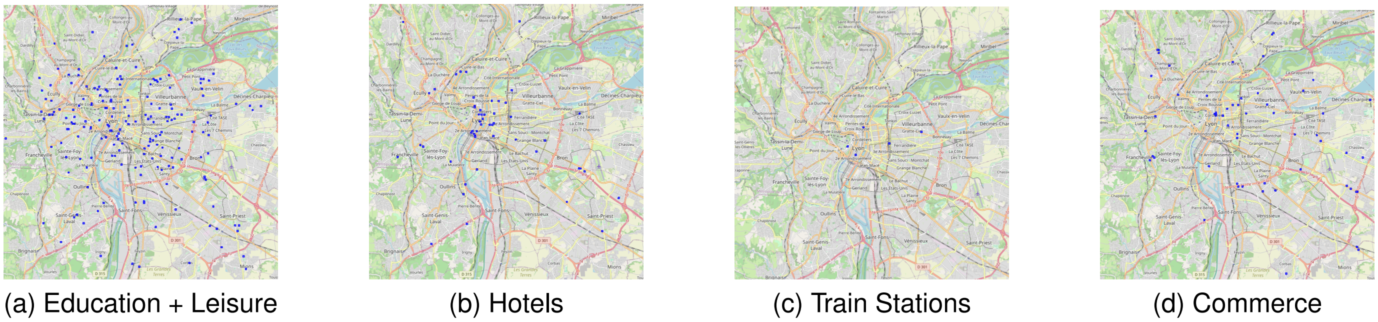
The objective of this study is to analyze mobile data from messaging applications. We aim to understand messaging app usage and gather information that can enhance infrastructure and quality of life within a smart city context. We will conduct our analysis using the NetMob23 dataset [1], which covers the period from March 16, 2019, to May 31, 2019, and provides uplink and downlink data for messaging applications such as WhatsApp, Telegram, Facebook Messenger, and Apple iMessage. Our observations pertain to the Lyon Metropolis in France. The traffic dataset provided is mapped through GeoJSON files using the WGS84 coordinate system. Each feature represents one square cell (tile) covering an area of (100×100) m^2. The Lyon metropolitan area has a total of 54013 tiles, and we used the data provided by the Grand Lyon Portal [2] to label some of these tiles, grouping them in the following classes: C1) *Education Centers* (208 tiles); C2) *Events* (74 tiles); C3) *Commerce* (67 tiles); C4) *Hotels* (58 tiles); C5) *Sports* (57 tiles); C6) *Restaurants* (51 tiles); C7) *Religious Centers* (41 tiles); C8) *Hospitals* (31 tiles); C9) *Train Station* (5 tiles). Figs. 1((a) – (d)) show the tile distribution of 4 different classes across the Lyon metropolitan area. We considered these classes to explore the following hypothesis: *"It is possible to characterize, based on information theory, the tiles where users are more likely to engage in online conversations"*. To verify this hypothesis, we conducted a detailed analysis of the tiles for each class, and looked for similar usage behavior categorized with the Complexity-Entropy Causality Plane (CECP) [3]. Only WhatsApp network traffic was considered in the study because it is the application which generates more traffic.

II. Results

Given that the NetMob dataset [1] provides 77 time series S per tile (one observation for each day), we used WhatsApp uplink data to compute the average traffic time series $\overline{S}_c = \{\overline{x}_{00:00}, \overline{x}_{00:15}, ..., \overline{x}_{23:30}, \overline{x}_{23:45}\}$ as the typical traffic signature for each class. The average network traffic at a specific time is represented by $\overline{x}_{hh:mm}$. We also computed the average traffic per day for each time series, represented by μ_S, and discarded the time series with average daily traffic below the median or above the 75th percentile (third quartile) values among all the averages. Therefore, each class has a representative and unique average time series generated from a set $D = \{S \mid median \leq \mu_S \leq Q_3\}$. This helps to mitigate the impact of anomaly events, outliers, and low traffic days that might not be representative for our analysis. We also make a distinction between weekdays (Monday to Friday) and weekends. This results in an average class behavior, which can be seen in Fig. 2 for data of the *Education Centers* class.

In Fig. 2, we can observe that the traffic typically begins to increase earlier during week days, at around 5:00, as compared to weekends (6:00). It also starts to decrease around 22:30 for weekdays and around 23:30 for weekends. One reason for this difference may be attributed to the weekly responsibilities related to studies and teaching. The *Education Centers* class suggests that people tend to wake up earlier during weekdays to attend schools, universities, and similar institutions, which might also imply exchanging WhatsApp messages. On weekends, we see a higher traffic volume. People typically have a break from these obligations, and they may wake up and go to sleep later as well. For both cases, peak usage occurs between 12:00 and 21:30. The early activity on weekdays seems to occur for most of the analyzed classes, but this difference is reduced for the *Train Stations* class. This class has the highest average traffic volume, as shown in Figure 3 (C9). The peak time traffic for this class appears around 18:00. For train stations, this refers to the afternoon period when there is the highest demand for train services. This is typically when many people are commuting from work and school, a good moment to exchange messages with friends and family. Interestingly, there is no peak in WhatsApp traffic during the morning commute hours, when the train demand is even more significant.

We can take a deeper look at the underlying nature of C9 and the other classes by analyzing the Complexity-Entropy Causality Plane in Fig. 4. The *Commerce* class (C3) has the highest permutation entropy and statistical complexity, which means more randomness and hard to predict usage behavior if compared to the other classes. This behavior usually implies chaotic traffic, with the least existing patterns and minimal information. Uncorrelated stochastic processes have $H \approx 1$ and $C \approx 0$. If we look specifically at the *Train Station* and the *Commerce* classes, one possible explanation to why C3 is more chaotic than C9 is that train stations coordinate travels, and people arrive and leave stations at specific and



(a) Education + Leisure

(b) Hotels

(c) Train Stations

(d) Commerce

Fig. 1. Tile location and distribution for some of the mentioned classes in Lyon metropolitan area.

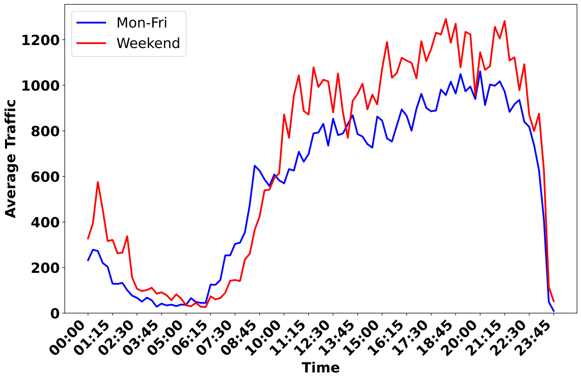


Fig. 2. The average uplink traffic during weekdays and weekends for the *Education Centers* tiles class.

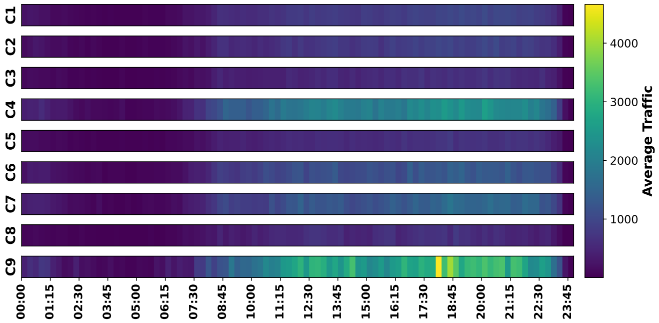


Fig. 3. The average uplink traffic during weekdays for all the classes.

somehow controlled times every day. This kind of behavior can create certain underlying traffic patterns. In the other hand, the *Commerce* class contain very distinct types of business, with distinct client profiles with different daily routines. The clients might go shopping whenever they feel like it, with no specific schedule or planning. It is also worth mentioning that spatial factors might influence user behavior, and the *Commerce* class has significantly more tiles, which are distributed in a greater area, as shown in Fig. 1(d).

III. Conclusion and Future Work

This work analyzed network traffic from messaging apps to identify user texting behavior. We defined classes based on specific locations in the Lyon metropolitan area and computed an average network traffic signature for each class using WhatsApp uplink traffic. The *Train Station* class exhibited the highest traffic volume, and we also identified a peak-time traffic at around 18:00. It is interesting to observe that the morning commuting rush hour does not generate a peak of messaging traffic. Transportation authorities can use this information to imagine adapting evening trains and services to this intense messaging usage. By analyzing Entropy-Complexity features, we observed that the *Commerce* class demonstrated more randomness compared to the other classes. We intend to continue our efforts by investigating temporal features to identify texting behavior during specific time periods and exploring information theory quantifiers such as permutation entropy and statistical complexity.

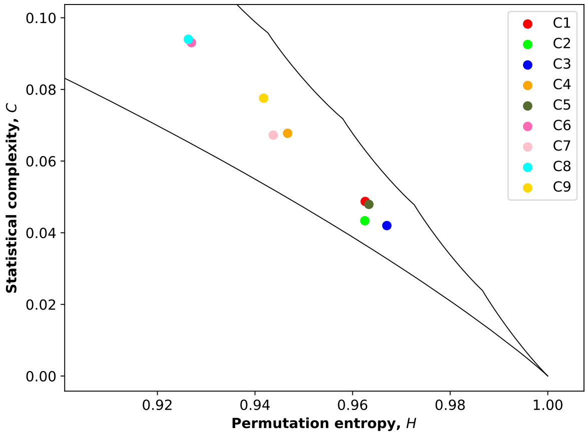


Fig. 4. Complexity-Entropy Causality Plane for the tiles classes.

References

[1] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," 2023.

[2] Data Grand Lyon, https://data.grandlyon.com/portail/fr/accueil, last visited on September 18, 2023.

[3] C. G. Freitas, O. A. Rosso, and A. L. Aquino, "Mapping network traffic dynamics in the complexity-entropy plane," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–6.

Batch Network Slicing for OTT Services in a Demand-Aware SDN Environment

Pooja Premnath
Department of Computer Science and Engineering
SSN College of Engineering,
Tamilnadu, India
pooja211015@ssn.edu.in

Sanjai Balajee Giridharan
Department of Computer Science and Engineering
SSN College of Engineering,
Tamilnadu, India
sanjai2110173@ssn.edu.in

Venkatasai Ojus Yenumulapalli
Department of Computer Science and Engineering
SSN College of Engineering,
Tamilnadu, India
venkatasai2110272@ssn.edu.in

Shahul Hamead Haja Moinudeen
Department of Computer Science and Engineering
SSN College of Engineering,
Tamilnadu, India
shahulhameadh@ssn.edu.in

Dhannya SM
Department of Computer Science and Engineering
SSN College of Engineering,
Tamilnadu, India
dhannyasm@ssn.edu.in

INTRODUCTION

Network slicing is considered a powerful technique for efficient network resource management and provision of differentiated services in modern networks. It enables the partitioning of network infrastructure into virtual slices, each customized for specific applications or services. It allows for efficient resource allocation, flexibility, and improved service quality by providing dedicated resources to each slice. Software Defined Networking (SDN) is the networking paradigm of this era that has centralized control over the network and enables data driven networking. The admission control of network slicing technique has been abstracted as flow rule placement at data planes (SDN switches) [7].

OBJECTIVE

Generally, a less important flow rule gets evicted due to the expiry of idle timeout or by imposing proactive deletion schemes [1]. Reinstallation of an evicted flow rule at a switch requires involvement of the controller, that affects QoS by causing a round trip delay. Effective rule management in SDN flow tables has been seen as one of the potential problems as it strengthens the performance of data plane [2]. Having known the traffic demands, efficient network slicing schemes can be designed to improve the QoS parameters. This paper presents a solution for network slicing, facilitated by Software-Defined Networking (SDN) techniques to efficiently manage network resources, and deliver differentiated services, with a specific focus on addressing the challenges posed by Over-the-Top (OTT) services and their data-intensive flows. OTT services exhibit distinctive characteristics, particularly in terms of traffic patterns and data flow. These services typically involve the transmission of data-intensive content, such as video streaming, large file downloads, and real-time communication applications. These types of flows are called heavy hitters or elephant flows [3].

PROPOSED IDEA

By formulating the problem as a knapsack optimization, this paper aims to maximize the QoS parameters, while meeting the unique demands of OTT traffic. An LSTM model is used to predict the traffic demands of OTT service. Picking the right slice for the predicted demand is considered as a cost function to the Knapsack formulation and, the reduction of delay has been framed as profit function.

KNAPSACK FORMULATION

Our proposal aims to minimize the overall round trip delay, which in turn, is inversely proportional to the traffic demand posed by the OTT services. Clearly, the total traffic demand that can be serviced is upper bounded by the maximum bandwidth. Furthermore, the bandwidth requirement request from each OTT service is a cost that contributes to the bandwidth consumption. Since the objective of the proposal is to minimize the delay incurred, we consider the negative of each delay value, due to which this minimization problem can indeed be viewed as a maximization problem. This enables us to formulate the given problem as a knapsack problem, where we define the profit to be negative of the delay values for each OTT service. Let $S = \{s_1, s_2, \ldots, s_k\}$ be the k OTT services. For each time slot $i \in T$, we consider a cost vector $c_1, c_2 \ldots c_k$ whose values correspond to the predicted traffic generated by $s_1, s_2, \ldots, s_k$ respectively from the deep learning model. Let B denote the maximum bandwidth. Let $D = \{d_1, d_2, \ldots, d_k\}$ be the delays incurred by the k OTT services. Let $D' = \{d_1', d_2', \ldots, d_k'\}$ denote the profit vector corresponding to the set S of k OTT services, where for each $i \in [k]$, $d_i' = -(di)$.

We now present the knapsack formulation:

$$maximize \sum_{i \in [k]} d_i' * x_i \quad (1)$$

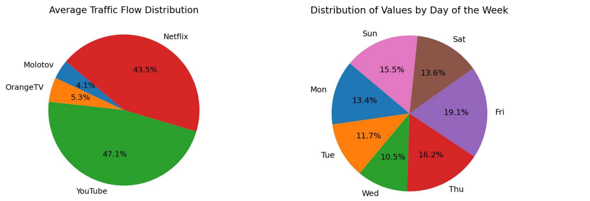
$$subject\ to \sum_{i \in [k]} c_i * x_i \leq B \quad \{2)$$

$$0 \leq x_i \leq 1 \qquad i \in \{1,2,\ldots,k\} \quad \{3\}$$

The vector $x = \{x_1, x_2 \ldots, x_k\}$ is the solution vector from which the flow rules can be generated for each time slot. Observe that the controller does not need to allocate the full bandwidth request from each service, instead it can allocate a fraction of the traffic demand request. This results in a fractional knapsack problem as given above. We know that the fractional knapsack problem can be solved in polynomial time using a simple greedy algorithm. Further, note that a knapsack instance must be solved for each 15 min slot separately for each geospatial id. Existing knapsack-based solutions for network slicing which aims at maximizing the revenue can be studied from [4], [5].

DATA ANALYSIS

The Netmob 2023 Dataset [6] provides information regarding the demands generated by 68 popular mobile services distributed over 20 different regions in France for 77 consecutive days in 2019. The initial phase involved an exhaustive examination of raw traffic data spanning 77 days across various online streaming services—namely, OrangeTV, YouTube, Netflix, and Molotov. Each day's data was segregated into distinct text files, where the first number in each line signified a geospatial identifier. The objective was to glean insights into overall traffic patterns and trends among services. The geospatial identifiers were extracted from each line and aggregated into a list, shedding light on the distribution of regions within the dataset. A pivotal step was to systematically restructure the dataset, in order to carry out efficient analysis of traffic across the given period. Using the extracted geospatial identifiers, we proceeded to reorganize the data into individual CSV (Comma-Separated Values) files, each corresponding to a distinct geospatial region. These CSV files were designed to have two columns: one denoting the date and time, recorded at 15-minute intervals, and the other representing the corresponding traffic values. After initial preprocessing, the traffic distribution based on service and day of the week has been identified by analyzing the data (Fig 1). An LSTM model has been built to predict the traffic demand for the queried OTT service, given a specific time slot. Based on the predicted traffic demand, network slice with proportional bandwidth can be allocated.

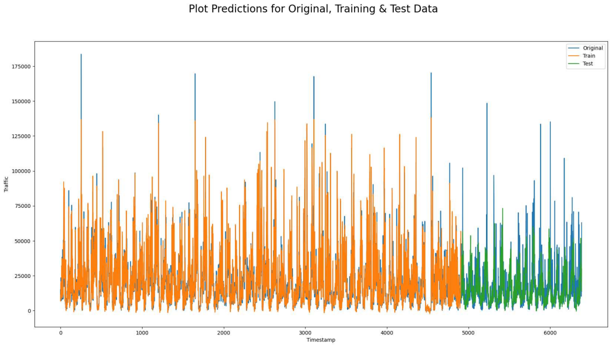


(a) Proportional shares of OTT services

(b) Traffic flow by day of the week

Fig. 1: Pie Charts of OTT Traffic Distribution

The predictions of the LSTM model can be seen in Fig 2. We utilize two key evaluation metrics, namely the coefficient of determination (R2) and the mean squared error (MSE), to comprehensively assess the performance of the LSTM model developed. These metrics are fundamental for evaluating the model's predictive accuracy and overall goodness of fit. (R2), also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that is explained by the model. It ranges from 0 to 1, where a higher (R2) indicates a better fit between the model's predictions and the actual data, with 1 signifying a perfect fit. Conversely, MSE measures the average squared difference between the model's predictions and the true values, providing a more granular understanding of the model's predictive errors. Lower MSE values indicate better predictive performance, as they signify smaller prediction errors. We obtain an MSE of 0.62 and an (R2) value of 0.79, signifying that our model can effectively predict future traffic values.



Architecture of the Model for Prediction of Traffic Value

REFERENCES

[1] Zehua Guo, Ruoyan Liu, Yang Xu, Andrey Gushchin, Anwar Walid, and H.Jonathan Chao. "Star:Preventing flow-table overflow in software-defined networks", *Computer Networks, 125, 04 2017. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.*

[2] Lei Wang, Qing Li, Richard Sinnott, yong Jiangg and Jianping Wu, "An Intelligent rule management scheme for software defined networking", *Computer Networks, Volume 144, 24 October 2018, Pages 77-88*

[3] Simon Bauer, Benedikt Jaeger, Fabian Helfert, Phillipe Barias, Georg Carle "On the evolution of internet flow characteristics,*". ANRW '21: Proceedings of the Applied Networking Research Workshop, July 2021*

[4] Jesutofunmi Ajayi, Antonio Di Maio, Torsten Braun, and Dimitrios Xenakis. "An online multi-dimensional knapsack approach for slice admission control". *IEEE 20th Consumer Communications & Networking Conference (CCNC) pages 152–157, 2023.*

[5] Rajesh Chala, Vyacheslav V. Zalyubovskiy, Syed M. Raza, Hyunseung Choo, Aloknath De, "Network slice admission model: Tradeoff between monetization and rejections*", IEEE Systems Journal, Vol. 14, No. 1, March 2020*

[6] Martínez-Durive O E, Mishra S, Ziemlicki C, Rubrichi S, Smoreda Z, Fiore M. "The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography". *arXiv:2305.06933 [cs.NI]. 2023.*

[7] J. -J. Chen *et al*., "Realizing Dynamic Network Slice Resource Management based on SDN networks," *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, Tainan, Taiwan, 2019, pp. 120-125, doi: 10.1109/ICEA.2019.8858288.

Going Green in RAN Slicing

Hnin Pann Phyu*, Razvan Stanica†, Diala Naboulsi‡, and Gwenael Poitau§
* Département de Génie Logiciel et TI, École de Technologie Supérieure (ÉTS), Montreal, Canada, hnin.pann-phyu.1@ens.etsmtl
†Univ Lyon, INSA Lyon, Inria, CITI, Villeurbanne, France, razvan.stanica@insa-lyon.fr
‡ Département de Génie Logiciel et TI, École de Technologie Supérieure (ÉTS), Montreal, Canada, diala.naboulsi@etsmtl.ca
§Dell Technologies, Ottawa, Canada, gwenael.poitau@dell.com

***Abstract*—Network slicing is essential for transforming future telecommunication networks into versatile service platforms, but it also presents challenges for sustainable network operations. While meeting the requirements of network slices incurs additional energy consumption compared to non-sliced networks, operators strive to offer diverse 5G and beyond services while maintaining energy efficiency. In this study, we address the issue of slice activation/deactivation to reduce energy consumption while maintaining the user quality of service (QoS). We employ Deep Contextual Multi-Armed Bandit and Thompson Sampling Contextual Multi-Armed Bandit agents to make activation/deactivation decisions for individual clusters. Evaluations are performed using the NetMob23 dataset, which captures the spatio-temporal consumption of various mobile services in France. Our simulation results demonstrate that our proposed solutions provide significant reductions in network energy consumption while ensuring the QoS remains at a similar level compared to a scenario where all slice instances are active.**

I. Introduction

The telecommunication industry accounts for approximately 2% of total global carbon emissions [1]. Energy consumption will continue increasing in beyond 5G and 6G networks, where computationally intensive services will be largely deployed. In these future architectures, network slicing allows for the splitting of a physical network into multiple virtual networks, enabling mobile networks to cater to a diverse range of network services [2]. Satisfying the requirements of different network slices, which includes ensuring performance isolation, comes at the cost of increased energy consumption. Meanwhile, we observe that the highest amount of energy is consumed in the radio access network (RAN), accounting approximately for 70% of the overall network energy utilization [3].

In this respect, several research works consider base station sleep schemes to further optimize the energy consumption in 5G networks [4], [5]. Applying these techniques directly in multi-services network slicing environments is more challenging, due to distinct temporal traffic patterns exhibited by different slice instances. Completely shutting down or putting the entire base station into sleep mode could significantly impact the quality of service (QoS) for users in particular slice instances. Moreover, several research works ([6], [7], [8], [9]) consider optimizing the allocation of network slice resources (i.e. radio, CPU, transmission bandwidth and power) with respect to different network domains. However, none of these studies specifically addresses the handling of underutilized slice instances and the ability to deactivate them based on certain conditions.

This motivates us to propose a new approach: dynamically activating and deactivating slice instances based on their traffic patterns to enhance base station energy efficiency. However, deactivating certain slices to save energy might degrade the user QoS, while activating all slices at all times to maximize QoS results in significantly higher energy consumption. Therefore, the energy minimization objective shall be coupled with a QoS maximization objective [10].

To manage the trade-off between the two objectives, we hereby introduce an EcoSlice, which is a slice instance using bare minimum resources and network functions. By that, it incurs lower energy consumption than typical slice instances. We consider the EcoSlice is up and running 24/7 to provide a bare-minimum service. In some conditions, e.g., low traffic demand, operators may switch the users of other slices to this specific EcoSlice, without a significant QoS impact. All in all, in this work, we investigate the slice activation/deactivation problem with the aim of reducing the overall energy consumption while respecting as much as possible the QoS.

II. Methodology

A. System Model

In a time-slotted system, we define τ as the slice activation/deactivation interval (SADI), during which slices remain continuously active or inactive. (De)activation decisions are made at the end of each τ, for the subsequent $\tau + 1$.

Moreover, our system model is composed of: *i)* **Set of slice instances.** This set comprises various virtual slices (e.g. eMBB, URLLC, mMTC, and EcoSlice) associated to base stations. Each slice instance is characterized by a specific Quality of Service Class Identifier (QCI) and its corresponding energy consumption. *ii)* **Set of users:** This set represents the users associated with the slice instances deployed at base stations. Each user is defined by their required delay and traffic load demand. *iii)* **Set of base stations:** This set includes all the base stations. Each base station has a set of possible slice activation/deactivation configurations.

We then define the overall energy consumption of a base station, which is composed of the energy consumption resulting from its associated slice instances (i.e. dynamic energy consumption) and a static energy consumption level. As mentioned, our objective is coupled with ensuring the user QoS. Hence, the user satisfaction is determined based on whether a user's delay requirement is satisfied or not. Finally, we use β as a trade-off parameter between energy consumption and QoS: a larger β gives more weight to the QoS.

B. Proposed Solution

We propose two decentralized approaches: Deep Contextual Multi-Armed Bandit (DCMAB) and Thompson Sampling Contextual Multi-Armed Bandit (Thompson-C). The system architecture of our solutions is depicted in Figure 1. DCMAB and Thompson-C agents operate at the level of individual base stations. Each agent is presented with various configuration options of a base station, each associated with a reward. The agent's task is to select the most appropriate configuration at each time step (τ) to achieve the overall objective. We model the problem of slice activation/deactivation as a contextual multi-armed bandit (MAB) problem.

- **State**: The state space of an agent includes the overall energy consumption and average QoS of the base station.
- **Action**: An action of an agent is to choose which slice instances are active or not, and navigate users to and from an EcoSlice based on the state of their requested slice instance.

- **Reward**: The reward function is designed to find the trade-off between energy consumption and QoS at each base station.

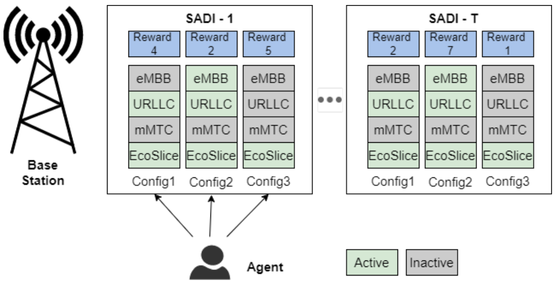


Fig. 1. System architecture.

III. Evaluation

For the simulation, we assume slice instances are deployed on an application-basis, and we consider downlink traffic information of three distinct applications from the NetMob23 dataset: Facebook, Netflix, and Spotify, in the city of Orleans [11]. Our agents operate at the individual base station level, which requires a pre-processing of the dataset. Specifically, we employ hierarchical clustering with the Ward linkage method to group the given tiles in Orleans based on their Pearson correlation values, and since the clustering results exhibit geographic continuity when plotted on the map, we consider each cluster to be equivalent to one base station. We train our models using data from 10 days (March 16-25, 2019).

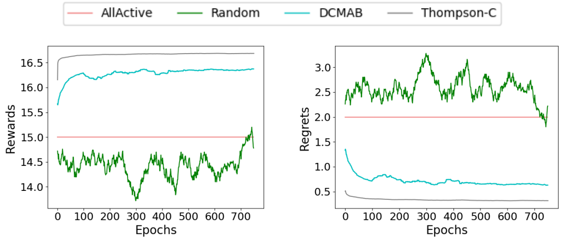


Fig. 2. Reward and regret obtained for $\beta = 5$.

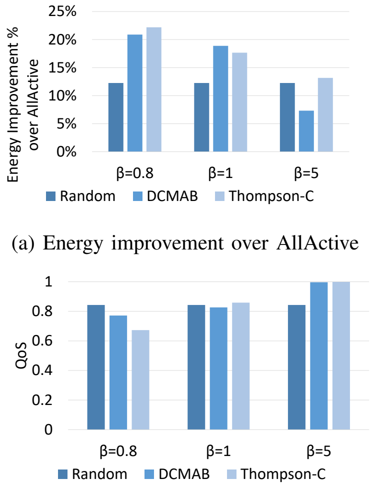
We compare the performance of the proposed DCMAB and Thompson-C solutions with two counterparts: AllActive (all the slice instances are active) and Random. In Figure 2 (obtained for $\beta = 5$), the DCMAB and Thompson-C agents exhibit better reward and regret trends than AllActive and Random strategies.

As further depicted in Figure 3, the energy improvement of DCMAB and Thompson-C compared with the current standard AllActive approach can reach 20% for different β values. The highest gain can be seen at $\beta = 0.8$ but at the expense of a QoS degradation. It is important to note that reducing the energy consumption involves some compromise on QoS. However, as illustrated in Figure 3 (b), when $\beta = 5$, our agents ensure the same level of QoS as the AllActive solution, while providing significant energy gains.

In all cases, Thompson-C outperforms DCMAB, but this comes with a much higher computational cost, as the execution time for Thompson-C is approximately 100 times larger. That being said, Thomson-C can be a favorable solution for a system with no computing time constraints, while DCMAB is better suited for real-time decision-making systems.

IV. Conclusion

Our work focuses on enhancing energy efficiency in RAN slicing by addressing the slice activation/deactivation problem. We propose state-aware MAB approaches, namely DCMAB and Thompson-C, where an agent aims to activate the optimal slice instances while maintaining a given QoS level. Our results, based on the NetMob23 dataset, demonstrate that our proposed solutions significantly reduce energy consumption at the base station level while preserving QoS.



(a) Energy improvement over AllActive

(b) QoS

Fig. 3. Energy and QoS based on different β values.

References

[1] B. Cubukcuoglu, "The Importance of Environmental Sustainability in Telecom Service Providers' Strategy," *Risk, Reliability and Sustainable Remediation in the Field of Civil and Environmental Engineering*, pp. 249–254, 2022.

[2] H. P. Phyu, D. Naboulsi, and R. Stanica, "Machine learning in network slicing—a survey," *IEEE Access*, vol. 11, pp. 39 123–39 153, 2023.

[3] N. Piovesan, D. López-Pérez, A. D. Domenico, X. Geng, H. Bao, and M. Debbah, "Machine Learning and Analytical Power Consumption Models for 5G Base Stations," *IEEE Communications Magazine*, vol. 60, no. 10, October 2022.

[4] F. Han, Z. Safar, and K. J. Liu, "Energy-Efficient Base-Station Cooperative Operation with Guaranteed QoS," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3505–3517, 2013.

[5] M. Feng, S. Mao, and T. Jiang, "Base Station On-Off Switching in 5G Wireless Networks: Approaches and Challenges," *Wireless Communications*, vol. 24, no. 4, p. 46–54, January 2017.

[6] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Energy-Efficient Deep Reinforcement Learning Assisted Resource Allocation for 5G-RAN Slicing," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 856–871, 2022.

[7] F. Rezazadeh, H. Chergui, L. Christofi, and C. Verikoukis, "Actor-Critic-Based Learning for Zero-touch Joint Resource and Energy Control in Network Slicing," *Proc. IEEE International Conference on Communications (ICC)*, 2021.

[8] O. Akin, U. C. Gulmez, O. Sazak, O. U. Yagmur, and P. Angin, "GreenSlice: An Energy-Efficient Secure Network Slicing Framework," *Journal of Internet Services and Information Security*, vol. 12, no. 1, pp. 57–71, 2022.

[9] H. Chergui, L. Blanco, L. A. Garrido, K. Ramantas, S. Kuklinski, A. Ksentini, and C. Verikoukis, "Zero-Touch AI-Driven Distributed Management for Energy-Efficient 6G Massive Network Slicing," *IEEE Network*, vol. 35, no. 6, pp. 43–49, 2021.

[10] A. Chatzipapas, S. Alouf, and V. Mancuso, "On the Minimization of Power Consumption in Base Stations using On/Off Power Amplifiers," *Proc. IEEE Online Conference on Green Communications (GreenCom)*, pp. 18–23, 2011.

[11] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," 2023.

Urban life and wellbeing assessed by mobile usage

Csaba I. Sidló, Ferenc Béres, Júlia Tompa, Benedek Székács, Domokos M. Kelen, Gábor Lukács, Katalin Hum, and András A. Benczúr

Institute for Computer Science and Control, Hungarian Research Network
Email: {sidlo, fberes, kdomokos, lukacsg, hum, benczur}@sztaki.hun-ren.hu

I. Introduction

Enhancing wellness, promoting livability, and improving the quality of urban life are fundamental objectives of urban development. Detailed mobile network data provides a significant opportunity to understand the hidden organization of urban systems and society.

We focus our research on urban density, proximity, and the role of mobility. We investigate the 15-minute city as an urban planning concept, a strategy that Paris and other French cities have already started to implement [5]. These cities try fulfilling six essential functions within a 15-minute walk or bike ride: living, working, commerce, healthcare, education and entertainment [4]. We address several potential research questions regarding this relatively new concept, including the measurability of related socio-demographic implications and relationships and the adaptability to various urban structures.

Mobile network data can be used to infer various socio-demographic and mobility patterns. The geographical features of the 15-minute city and mobile network traffic of the NetMob23 dataset [3] are both strongly linked to the socio-demographic attributes of urban areas. As such, we undertook exploratory analysis to understand how network traffic data can supplement the 15-minute city principles and enrich existing urban models. We deployed explainable AI (XAI) [1] to investigate the relationships among socio-demographics, mobile usage, and city structure features.

II. Socio-demography and mobile usage

We conducted comprehensive data collection from various external sources[1] to examine the relationship between mobile network traffic and socio-demographic indices as shown in Table I. Our particular interest lies in analyzing traffic patterns among communities with different living standards. The use of SHAP [2] to explain our predictive models, as shown in Fig. 1, provides further insights into the relation between annual income and mobile traffic.

III. Exploring the 15-minute city

OpenStreetMap[2] (OSM) collects and categorizes POIs, enabling the identification of urban usage patterns, including the features of the 15-minute city concept. We use NetMob23 data to extend OSM features by app usage patterns.

Variable	Description
DEC_MED19	Median declared income per consumption unit
DEC_PCHO19	Share of income from unemployment benefits (%)
P19_POP0014	Share of age 0-14 in population (%)
P19_POP1529	Share of age 15-29 in population (%)
C19_POP15P_CS3	Share of higher intellectual professions (%)
C19_POP15P_CS6	Share of workers in population (%)

TABLE I: Selected socio-demographic variables in 2019 obtained from INSEE that we explain by mobile traffic data.

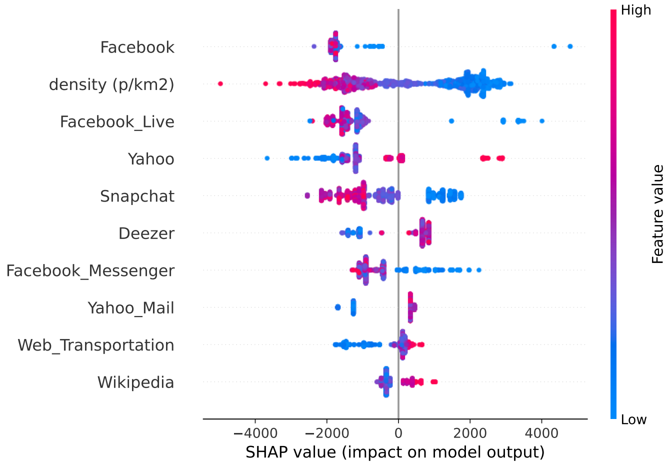


Fig. 1: SHAP explanation for a CatBoost regression model predicting annual income solely from all available features for IRIS regions.

Our investigation, exploration and explanation are based on explainable modeling of the sociodemographic indicators using OpenStreetMap and NetMob23 data. For modeling, we use the CatBoost [6] gradient boosting implementation.

Using detailed mobile network traffic data, we can uncover new or more detailed patterns associated with the 15-minute city concept. When explaining models predicting social attributes, such as annual income in Fig. 2, we can see 15-minute POIs and services (atm, bank, optician) that are related to the daily routine of the wealthier population. On the other hand, SHAP explanations also show that a high frequency of public transport stations or playgrounds is not associated with wealthy neighborhoods.

In order to separate the explanatory power of OSM for sociodemography from its causal effect on mobile app usage, we deployed our recent theory on asymmetric SHAP for root cause analysis [1], In Fig. 3, we show variance reduction,

[1] Revenus, pauvreté et niveau de vie en 2019 (Iris): https://www.insee.fr/fr/statistiques/6049648; Population en 2019 - https://www.insee.fr/fr/statistiques/6456157?sommaire=6456166

[2] OpenStreetMap POIs and categories: https://www.openstreetmap.org

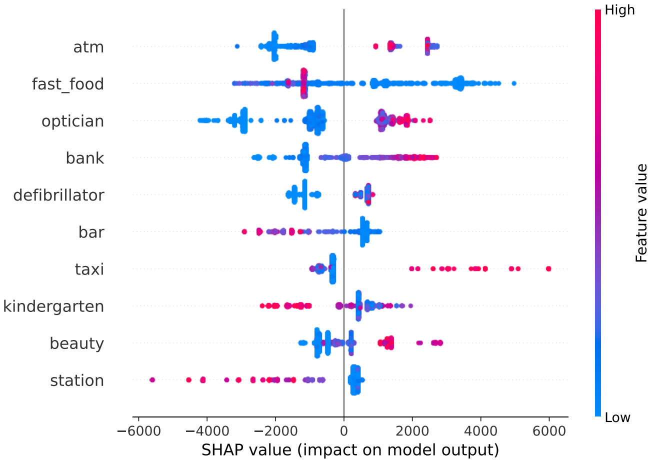
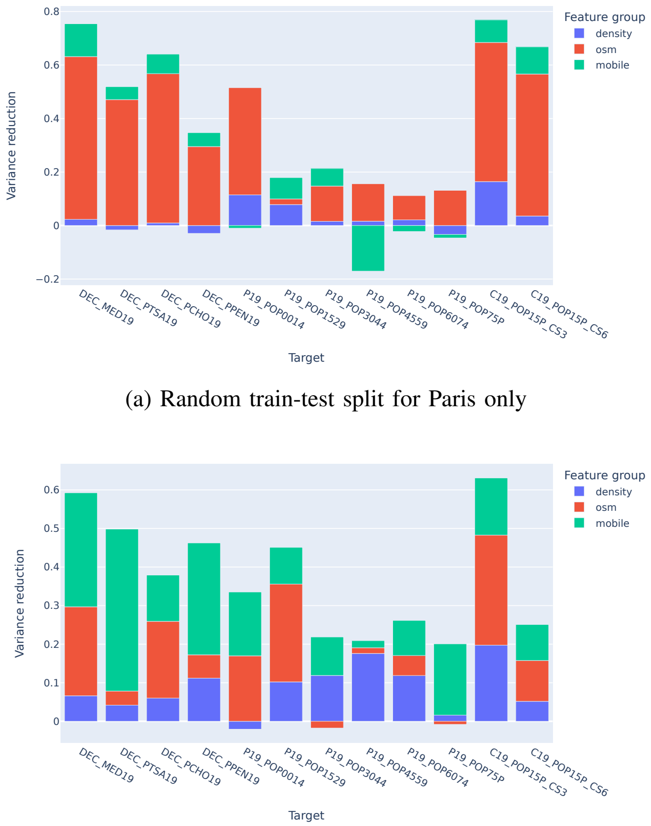


Fig. 2: Asymmetric SHAP explanation for a CatBoost regression model predicting annual income solely from OSM.



(a) Random train-test split for Paris only

(b) Training and validation across cities.

Fig. 3: Impact of different feature groups for the prediction of various sociodemographic indicators of Table I.

by attributing the contribution of the explanatory variables in causal order, starting with population density, followed by PoI and finally mobile app usage features.

As part of our results, we developed a Streamlit-based dashboard that can be used to explore our data collection and explanatory models. Over the interface, one may experiment with different target variables, feature groups, as well as model types and hyperparameters.

IV. Summary

In our experiment, we effectively assessed urban life quality by combining NetMob23 mobile app usage data with open population census and street map information, primarily focusing on sociodemographic indicators like median income and unemployment rates. Notably, our findings demonstrated that NetMob23 app usage data offered explanatory power for sociodemographic characteristics at the neighborhood level (IRIS). Specific apps such as LinkedIn and Wikipedia were linked to higher education levels and increased median income, while others like Facebook and Messenger also played significant roles, though with less apparent reasons. Our models showcased substantial variance reduction, affirming the potential to explain urban well-being, but the contribution of urban Points of Interest (PoI) features to life quality was relatively weaker compared to mobile app usage when models were trained and evaluated across different cities. We used SHAP theory to dissect the causal impact of these variables, concluding that PoI features may exert an indirect influence on mobile app usage patterns.

Due to the limited time of the Challenge, several directions are yet to be explored, such as the explanation of sociodemographic changes between 2019 and 2020. Our resources will be (partially or fully) available for future research upon approval of the Organizers.

References

[1] Domokos Miklós Kelen, Péter Kersch, András Benczúr, et al. “Causal Explanations for Performance in Radio Networks”. In: *CEUR WORKSHOP PROCEEDINGS*. Vol. 3189. 2022.

[2] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[3] Orlando E Martínez-Durive et al. *The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography*. 2023. arXiv: 2305.06933 [cs.NI].

[4] Carlos Moreno et al. “Introducing the “15-Minute City”: Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities”. In: *Smart Cities* 4.1 (2021), pp. 93–111. ISSN: 2624-6511. DOI: 10.3390/smartcities4010006. URL: https://www.mdpi.com/2624-6511/4/1/6.

[5] Georgia Pozoukidou and Zoi Chatziyiannaki. “15-Minute City: Decomposing the New Urban Planning Eutopia”. In: *Sustainability* 13.2 (2021). ISSN: 2071-1050. DOI: 10.3390/su13020928. URL: https://www.mdpi.com/2071-1050/13/2/928.

[6] Liudmila Prokhorenkova et al. “CatBoost: unbiased boosting with categorical features”. In: *Advances in neural information processing systems* 31 (2018).

Synthetic Network Traffic Data Generation using Deep Generative Models

Yanbo PANG[1] Kunyi ZHANG[2] Pierre FERRY[2]

[1]Center for Spatial Science The University of Tokyo, Ce509, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan

[2]Department of Civil Engineering The University of Tokyo, Ce509, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan

pybdtc@iis.u-tokyo.ac, kyzhang@iis.u-tokyo.ac.jp, ferryp@iis.u-tokyo.ac.jp

Abstract – Our team is eager to address the gap in accessible, reliable synthetic mobile data traffic datasets. We cast the problem of simulating network traffic usage (uplink/downlink) scenarios in a generative modeling framework, aiming to create datasets that could fuel innovation and provide reproducibility in research. Our preliminary analysis of the provided dataset has generated promising leads, which we will explore further using deep generative models with the expressive power of neural networks.

1 Introduction

The lack of uniform access to mobile data traffic in the research community inhibits innovation and verifiability. Our study aims to bridge this gap, deploying a generative modeling framework for simulating network traffic usage scenarios. To achieve this goal, we base our work on the NetMob23 dataset, a significant advancement in mobile network data availability for researchers. Unlike previous datasets focused on Call Detail Records, NetMob23 offers comprehensive 4G data traffic information. It encompasses traffic across 20 metropolitan areas in France and offers rich data on 68 popular mobile services usage. The original generation process of NetMob23 abandons the traditional approach of using Voronoi tessellations for antenna coverage, instead mapping data traffic to over 870,000 high-resolution grids, resulting in over 440 billion data points. This innovative dataset sets the stage for novel explorations of mobile network traffic. Therefore, generating a synthetic dataset based on this resource becomes critical to enable more extensive studies on mobile network traffic without access constraints.

In recent years, significant progress in deep generative models for synthetic data generation has been observed across various domains, generating synthetic images, text, music, sensory data, electronic health records, mobility trajectories, and financial time series with impressive fidelity. In the context of our research, these advancements provide the tools needed to generate synthetic datasets that closely mimic the rich, high-dimensional data offered by NetMob23. They can handle large, highly correlated datasets and excel at learning intricate patterns, effectively capturing the underlying data distribution to produce realistic synthetic data.

2 Explanatory Data Analysis

The NetMob23 dataset uncovers compelling spatial-temporal patterns in mobile service usage. Observable patterns reflecting daily human routines emerge from the data, with discernible peaks and dips in usage corresponding to different times of the day and varying between weekdays and weekends. Moreover, spatial patterns reveal a diverse geographical distribution of traffic. Differences are noted in usage across regions, ranging from a widespread city distribution to concentrated usage in specific zones like business districts or tourist spots.

To delve deeper into these patterns, we applied outlier detection methods such as Inter Quartile Range, Z-score, and Isolation Forest to the raw traffic volume for Instagram (UL) and Google Maps (DL). In Figure 1, our results showed a clear spatial fracture between dense and sparsely populated areas, and an exponential relationship between average traffic volume and population density. Normalizing the traffic volume by population density revealed activity zones that consistently overlap non-residential IRIS zones. However, artifacts remain in the residential areas, necessitating a refinement of soil usage.

The application of Earth Mover's Distance (EMD) as our disparity metric of the spatial-temporal distribution of traffic volume helped categorize mobile services into groups. From Figure 2, we observed that network traffic data is highly affected by multiple context and semantics features including application type, geographic factors, time factors, social economics, and telecommunications.

3 Proposed Method and Approach

These findings underline the need for a sophisticated generative model to reproduce the complexity of the dataset faithfully. Acknowledging this, we propose a two-stage generative model that incorporates contextual information. This model is inspired by text-conditional generation problems such as text-

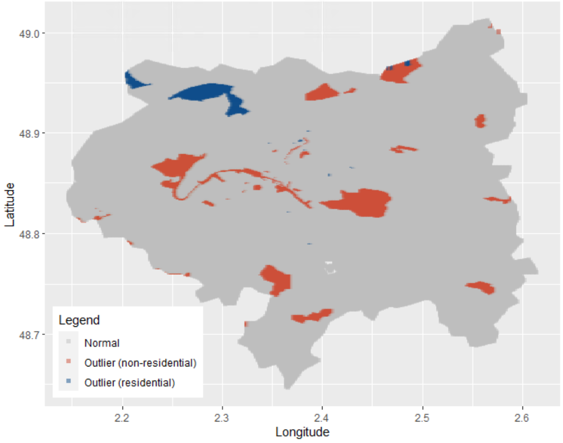


Figure 1: Isolation Forest 6% outliers (on Paris' Instagram UL traffic volume normalised by population density)

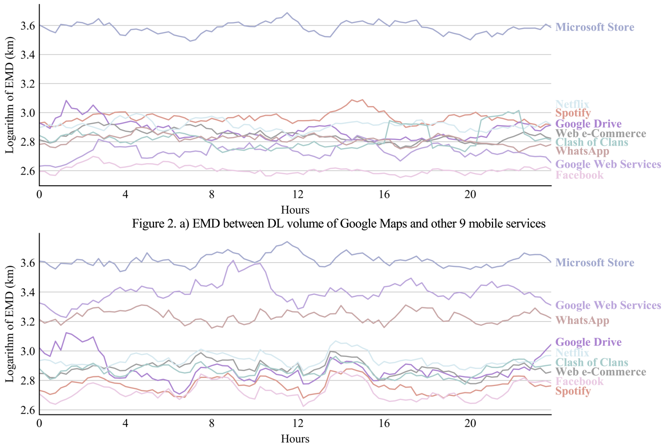


Figure 2. a) EMD between DL volume of Google Maps and other 9 mobile services

Figure 2. b) EMD between UL volume of Google Maps and other 9 mobile services

to-image and text-to-music. It consists of a prior encoder that aggregates background knowledge of the target region and application, and a generator that produces network traffic data based on the prior encoder's output. This comprehensive approach aims to faithfully reproduce the variable nature of user behavior across time and space that is evident in the original data.

Nonetheless, there are two major challenges. Firstly, unlike text-to-image or text-to-music tasks, where existing pre-trained language models such as BERT or transformers can efficiently extract information from text, we need to develop a novel embedding model to aggregate contextual information from application type, geographic factors, time factors, social economics, and telecommunications. Secondly, the network traffic data varies across more than 20 metropolitan areas and follows patterns reflecting daily human routines, with discernible peaks and dips in usage that correspond to different times of the day and vary between weekdays and weekends. The success of this study relies on selecting a powerful generator that can learn the network traffic data representation and jointly train the generator to reproduce the network traffic data.

To tackle these challenges, we plan to experiment with a selection of deep generative models that have demonstrated impressive performances in complex synthetic data generation tasks. These include:

Autoregressive Networks, which predict future data points based on past observations, adding a random error term. They're simple, interpretable, and effective at capturing trends and seasonalities in temporal data, making them a robust initial choice for our study.

Variational Autoencoders are powerful generative models that use neural networks to map inputs to a latent space, with a decoder then generating outputs from this space. Given their probabilistic nature and adeptness at modeling complex, high-dimensional data, they're suitable for our task.

Generative Adversarial Networks (GANs) involve two neural networks - a generator and a discriminator - that compete against each other. The generator creates synthetic data, while the discriminator evaluates its realism. GANs' ability to generate highly realistic data makes them an attractive option for our high-resolution synthetic dataset generation.

Transformers use self-attention mechanisms to understand the context and dependencies in sequential data. Given the sequential nature of network traffic data, Transformers could effectively capture complex dependencies and patterns.

Diffusion Models generate data by reversing a process that transforms the original data into noise. Their ability to model the complex, multi-step process of network traffic data generation makes them an intriguing approach for our task.

Each of these models presents unique strengths that may be advantageous for our task. By comparing their performance on the NetMob dataset, we aim to identify an optimal model that balances representativeness, novelty, realism, diversity, and coherence.

4 Expected Results and Impact

Our project aims to advance state-of-the-art knowledge in synthetic dataset generation for mobile data traffic. The generative models' outputs should meet five criteria: representativeness, novelty, realism, diversity, and coherence. We anticipate that our models will represent the original dataset accurately and generate novel and realistic synthetic data that captures the diversity of mobile data traffic patterns. Furthermore, the generated data should maintain coherence in the context of network traffic.

This work will lead to the development of a publicly available, high-resolution synthetic dataset of mobile data traffic, a substantial contribution to the research community. We also anticipate that our approach and models will enhance capabilities in the synthetic data generation domain, impacting larger projects that require similar datasets. We hope to foster collaboration and further advancements in this field by sharing our codebase and methodologies. Ultimately, our project will push the boundaries of knowledge and techniques in synthetic data generation, providing meaningful impacts in the AI and ML fields.

The online geography of discontent:
social media, urban peripheries and radical voting in French cities

Andrea Musso[1], Gianluca Risi[2], Theodore Tallent[3,4]

Introduction

In the last years, Western democracies have experienced a wave of populism and a surge of political polarization. Among the main attempts to explain this upsurge of political discontent, two have attracted particular attention. On the one side, scholars focusing on the so-called "geography of discontent", explaining the turn towards populism by the tendency for inhabitants of peripheral areas - "which suffered long-term economic and industrial decline, often alongside employment and demographic losses" - to resort to the ballot box to take a revenge against the system that left them behind (Rodriguez-Pose, 2018, p.1). On the other side, social media have also gone under increasing scrutiny for their alleged role in influencing the results of decisive elections and, more generally, for fueling political polarization and voters' radicalization (Kubin and von Sikorski, 2021). However, most studies investigating the role of social media in shaping electoral results and political attitudes (including political polarization) focus on aggregate levels (mostly national) and tend to be blind to potential spatial variations in internet consumption behaviours. In a certain sense, we can say that these two streams of literature have evolved mostly in parallel.

We argue that, as political discontent is on the rise, we need to bring these two perspectives together in order to deeply understand the current phenomenon. In fact, the analysis of political polarization and radical voting cannot remain in one of these two silos, emphasizing either the role of "left-behindness" or the power of social media.

Despite the difficulties involved in examining the interplay between political dissatisfaction, geographical factors, and social media usage, we argue that this matter deserves further investigation, especially at the level of major urban centers. In fact, it is well known that urban centers play a huge role in determining the outcomes of political elections, but they are still an unidentified object even in the *geography of discontent* literature (Dijkstra et al., 2020). As we comment more exhaustively in the full report, these kinds of analyses are often conducted at aggregate level (referring to regions or macro-areas) and usually consider cities to be marble loci of progress and development, even if we know that they hide huge socioeconomic inequalities and rising political discontent within them (Lelo, Risi, 2022).

Summary of analysis

Within this framework, we exploit the NetMob dataset focusing on the twenty biggest cities in France, aiming to build a coherent picture of the relationship between the usage of social media, spatial inequalities, and political attitudes. In doing so, we make good use of the socioeconomic data provided by INSEE (Institut national de la statistique et des études économique), the Internet usage data presented in the NetMob dataset (Martínez-Durive et al., 2023) and the votes expressed in France for the 2019 European elections. Our analysis merges these datasets at the IRIS level, the smallest spatial unit at which the INSEE reports statistics.

[1] *Computational Social Science, ETH Zurich, Zurich, Switzerland (andrea.musso@gess.ethz.ch)*
[2] *Applied Economics, Politecnico di Milano, Department of Architecture, Built Environment and Construction Engineering (ABC), Milan, Italy (gianluca.risi@polimi.it)*
[3] *Political Science, Centre for European Studies and Comparative Politics, CNRS, Sciences Po, Paris, France.*
[4] *Cambridge Centre for Environment, Energy and Natural Resource Governance, Department of Land Economy, University of Cambridge, United Kingdom (theodore.tallent@sciencespo.fr)*

Consequently - through a quantitative analysis - this paper identifies an *online geography of discontent* (France Info, 16 April 2022; Aral & Eckles, 2019), whereby political discontent, socioeconomic conditions (Fujiwara et al. 2021), and the consumption of certain social media (e.g. Facebook, Twitter, YouTube) form a coherent spatial picture, showing that a common pattern between the major urban areas in France does exist. In simple words, there seems to exist a tendency according to which disadvantaged and "left-behind" *peripheral* neighborhoods are likely to vote more for radical parties and use more specific social media.

Furthermore, we then perform several correlational analyses of radical voting and social media consumption, controlling for socioeconomic characteristics of places. In doing so, we identify a significant but small correlation between social media consumption and political discontent, when differences in socioeconomic situations are considered. This seems to suggest that, while social media surely play a role in shaping political attitudes, it is likely that socioeconomic characteristics and spatial "left-behindness" play an even bigger role in fueling political discontent.

Conclusions

We conclude that, although social media likely contribute - and are fed by - political discontent, its effect should not be overestimated. Socially based discontent (i.e. discontent based on material deprivation and lack of resources) is a key contributing factor in shaping political attitudes and social media consumption. Consequently, we argue that the *online geography of discontent* needs to be given increasing relevance , both for the importance of spatial inequalities in fueling political discontent and polarization, and for the significant role that - anyway - social media plays in this triangle.

References

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858-861

Dijkstra, L., Poelman, H., & Rodríguez-Pose, A. (2020). The geography of EU discontent. *Regional Studies*, 54(6), 737–753.

Franceinfo *(2022), INFOGRAPHIES. Macron dans les métropoles, Mélenchon dans les quartiers populaires, Le Pen dans les campagnes… Visualisez la France du premier tour de la présidentielle*, 2022 avril 16

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, *45*(3), 188-206.

Lelo K., Risi G. (2022), Urban Development in Rome: Illegal Housing Expansion, Inequalities and Governance, The regional challenges in the post-Covid era, Collana Scienze Regionali di FrancoAngeli, vol.62, pp. 81-100

Martínez-Durive, O. E., Mishra, S., Ziemlicki, C., Rubrichi, S., Smoreda, Z., & Fiore, M. (2023). The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography. *arXiv preprint arXiv:2305.06933.*

Rodríguez-Pose, A. (2018). The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society*, 11(1), 189–209.

Opening Telecommunication Data for Inclusive Education

Aránzazu San Juan Llano[1,2], Thi-Mai-Trang Nguyen[2,3], Daniel Rodríguez[4], and Matthieu Chardy[1]

[1]*Orange Innovation, Châtillon, France*
[2]*LIP6 - Sorbonne Université, CNRS, Paris, France*
[3]*L2TI - Université Sorbonne Paris Nord, Villetaneuse, France*
[4]*Universidad de Alcalá, Alcalá de Henares, Spain*
aranzazu.sanjuanllano@orange.com, thi-mai-trang.nguyen@lip6.fr , daniel.rodriguezg@uah.es, matthieu.chardy@orange.com

***Abstract*—In this paper, we present the main stages of our research in the context of the NetMob 2023 Data Challenge. Following our global research work in cloud continuum infrastructures for inclusive education, we aim to evaluate the impact of telecommunication infrastructures on digital access from an open data perspective in education. We will analyze the French National Education use case crossing the 4G dataset provided by the French telecommunications operator Orange with open datasets for inclusive education in order to propose the first definition of a new Inclusiveness Education KVI (Key Value Indicator) compatible with the Hexa-X 6G KVI definition.**

***Index Terms*—telecommunications data, open data, digital inclusiveness, digital divide, inclusive education, 6G, KVI**

I. Introduction

Technology represents an excellent opportunity to improve digital equality, but during COVID-19, the digital divide has grown. Besides, disruptive technologies like 5G generate an important societal controversy. To address this issue, Hexa-X presents the 6G technology as a technology centered on societal values that moves the gravity center from traditional hard indicators, the Key Performance Indicators (KPIs), to the soft indicators, the Key Value Indicators (KVIs).

Our research work concerns the impact of telecommunications infrastructures on inclusive education. From a global point of view, we propose a new open infrastructure of telecommunications for education based on the cloud continuum concepts and beyond with explicit consideration of inclusiveness workload and indicators by design, introducing an Inclusiveness Education Key Value Indicator (KVI) definition compatible with the 6G Key Value Indicator concept of Hexa-X[1].

Regarding communications infrastructures for education challenges revealed during the COVID-19 pandemic, we could focus our research, especially on applications with similar characteristics to those used in existing central cloud infrastructures for distance and hybrid learning or future classrooms. [2]

Crossing datasets with a multi-disciplinary approach contributes to reducing the digital divide and digital illiteracy. From this inclusion perspective of designing open infrastructures for education centered on societal values, we would cross some open datasets of the education and telecommunications domains with the provided Orange dataset [1] to define our first version of the proposed Inclusiveness Education KVI.

The remainder of this abstract paper is organized as follows. Section 2 describes our use case, French National Education. Section 3 presents some analyses concerning open education data. Finally, we present the future directions in Section 4.

II. The French National Education use case

The French National Education use case [2] presents some learning, space-temporal, and inclusion characteristics that are interesting for our research.

Concerning learning stages organization[3], the K12 French system is divided into four major stages from 3-5 years named ”maternelle” (preschool [4]), 6-10 years named ”école” (primary school), 11-14 years named ”college” (middle school), and 15-17 years named ”lycée” (high school). Regarding spatial structure, Education in Metropolitan France [5] is divided into three geographical areas: A area, B area, and C area. Every educational area regroups some administrative educational units named ”Academia” (meaning Academy), which concentrate on managing all K-12 education establishments across one city or more. For example, the Paris metropolitan area in the provided 4G Orange dataset includes three academies (Paris, Versailles, and Créteil). Normally, academic holidays[6] take place every six weeks for a period of 16 natural days for an academic area, with a total period of 30 days across all academic areas. A scholar holiday period is divided temporarily and spatially: a scholar holiday period starts on Saturday

[1] *Hexa-X H2020 5G-PPP project. Deliverable D1.2, Expanded 6G vision, use cases, and societal values,* April 2021.

[2] Equivalent in terms of data consumption, latency of other parameters

[3] *Organisation de l’école*. Ministère de l’Éducation Nationale et de la Jeunesse. URL: https://www.education.gouv.fr/organisation-de-l-ecole-12311.

[4] Education is obligatory in France from 3 years old from September 2020.

[5] For Corsica, the overseas territories the calendar could be adapted by the educational authorities

[6] *Vacances scolaires 2018-2019*. Ministère de l’Éducation Nationale et de la Jeunesse. URL: https : / / www . education . gouv . fr / bo / 17 / Hebdo26 / MENE1719943A.htm.

in the A area, and after a period of 7 days, on the next Saturday, the scholar holiday starts in the next academic area following the ordered sequence A area, B area, C area. [7] In terms of educational inclusion, the French National Education has a priority education policy that aims to correct the impact of social and economic inequalities on academic success (RES and RES+) and various units for inclusive education of students with disabilities (ULIS, EREA, and SEGPA).[8]

III. Open data for inclusive education

We are working with the open data repository of the French government dedicated to education [3]. This repository provides open datasets with general information about the schooled population and digital inclusion statistics like the Social Position Index of schools (named after the French acronym, IPS) in France. We were exploring the IPS of different academic stages (primary, middle, and high school) and the IPS of EREA units from the academic year 2018-2019. We were performing preliminary analyses of the primary school IPS dataset and after we extended these analyses to the other stages for all cities in the dataset. Preliminary analysis showed that the IPS index presents a normal distribution. Descriptive statistics of primary schools IPS dataset showed the city with the biggest IPS standard deviation is Marseille, and the city with the smallest IPS standard deviation is Rennes; this means IPS are more spread out for Marseille. There were our first two candidates to deeply analyze the digital education gap in France and its relationship with telecommunication data. We cross the Educational National open data with the traffic files of eight cloud services for these two cities for a laboral day. As represented in Fig. 1, there is no correlation between the IPS and the data consumption on the tiles associated with the set of primary, middle, high and EREA schools of Marseille and Rennes.

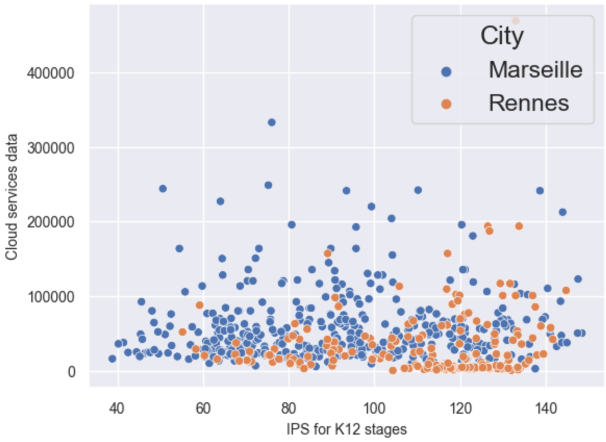


Fig. 1. Relation between the cloud services traffic and the IPS for the K12 stages

IV. Future work

As mentioned in the Introduction section, one of the main goals of our global research about cloud continuum infrastructures for inclusive education is to propose a new open infrastructure education based on 5G technology and to define a new Inclusiveness Education KVI. This KVI indicator would be composed of many soft indicators such as global network availability, mobile network usage, quality of service, or ownership of digital devices close to the Student's Digital Opportunity model proposed by Jim, Evans, and Grant for the United States [4]. We would extend this model to the mobile network and add other soft indicators emerging from intrinsic peculiarities of the French National Education use case, especially regarding inclusion scope.

The main goals of this study are the following :

1) Continue to evaluate the relationship between economic factors like the IPS of a given area and data usage for all applications and by categories of applications.

2) Continue to study the temporal and spatial traffic patterns across the academic areas and scholar or academic holiday periods described below as the first stage of our global research to characterize the mobile traffic for future mobile infrastructures for inclusive education.

3) Evaluate the relationship between 4G network availability and data usage for a set of applications equivalent to educational applications of a traditional central cloud infrastructure for distance or hybrid education for the 20 French cities.

4) Create a first version of our inclusiveness KVI relating the previous observations with the schools located in a given area in order to estimate the digital divide across different geographical levels. Improve the accuracy of this inclusiveness KVI taking into consideration other inclusion open data.

5) Predict the traffic demand for a mobile education infrastructure with Machine learning (ML) based forecasting models.

References

[1] Orlando E. Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. *The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography*. May 11, 2023. arXiv: 2305.06933. URL: http://arxiv.org/abs/2305.06933. preprint.

[2] *Construisons ensemble une école engagée*. Ministère de l'Éducation Nationale et de la Jeunesse. URL: https://www.education.gouv.fr/construisons-ensemble-une-ecole-engagee-221.

[3] *Explore — Éducation Nationale - Accueil*. URL: https://data.education.gouv.fr/explore/?sort=modified.

[4] Cary Jim, Sarah Evans, and Alison Grant. "Multidimensional Approaches to Illustrate to Digital Divide among K-12 Students". In: (July 15, 2021). DOI: 10.29173/iasl8282.

[7] The period of the provided dataset, from 16th March 2019 to 30th May 2019, includes three bank holidays and the Spring break of the 2018-2019 academic year, starting in A area on a Saturday, 13th April, and returning the school the Monday 6th May for the C area.

[8] *École inclusive*. éduscol — Ministère de l'Éducation nationale et de la Jeunesse - Direction générale de l'enseignement scolaire. URL: https://eduscol.education.fr/1137/ecole-inclusive.

Predicting the Productivity Indicators of a Society using Workforce Behavior Dynamics

Noha Gamal[(1)] ngamal@nu.edu.eg
Tamer Arafa[(1)] Tarafa@nu.edu.eg
Mina Youssef[(1)] Myousef@nu.edu.eg
Ahmed El-Mahdy[(1)] aelmahdy@nu.edu.eg

[(1)]School of Information Technology and Computer Science, Nile University, Giza, Egypt.

Abstract:

In an era marked by pervasive digital interconnectivity, comprehending the intricate interplay between workforce behavior and societal productivity indicators is of paramount importance, extending beyond academic curiosity to underpin economic policies, resource allocation, and societal well-being. This study, centered in Lyon, France, leverages diverse data sources, including mobile app usage metrics, to delve into this nexus, striving to serve as a template for similar analyses worldwide.

Introduction:

In a world steeped in digital interconnectedness, unraveling the intricate dynamics of workforce behavior assumes significance beyond academic pursuits. It is a fundamental cornerstone for shaping economic policies, allocating resources judiciously, and enhancing societal well-being. This study, focusing on Lyon, France, delves deep into the relationship between workforce behavior and societal productivity indicators. It leverages an array of data sources, including mobile app usage metrics, to offer insights applicable not only locally but also as a blueprint for analogous analyses in diverse geographical contexts.

Objectives:

The primary objectives of this research encompass three crucial dimensions:

App Impact on Productivity Analysis: This research strives to construct a model capable of accurately gauging the influence of mobile applications on productivity. It employs rigorous analysis of coefficients and feature importance scores to pinpoint the apps with the most substantial impact. This analysis elucidates which applications, upon increased or decreased usage, lead to noticeable changes in productivity—an insight valuable to individuals and organizations seeking to optimize app usage for enhanced productivity.

Categorization of Productivity: Another pivotal research objective involves categorizing mobile applications into two distinct groups: productive and non-productive. This categorization hinges on their discernible effects on the productivity index, offering valuable guidance to users seeking to enhance work efficiency and minimize distractions.

Identification of App Usage Patterns: This research also endeavors to unearth patterns in app usage that correlate with productivity fluctuations. For instance, it may reveal that heightened social media app usage during working hours is inversely linked to productivity. Such findings empower users with self-awareness, enabling informed decisions regarding app usage timing and frequency. Ultimately, this research aspires to elevate productivity by unraveling the intricate relationship between mobile app usage patterns and overall efficiency.

Literature Review:

The study bridges a critical gap in previous research by incorporating mobile application usage as a determinant of workforce behavior and productivity. Traditional models focused on individual attributes, while this research integrates variables such as app usage and social media activity to provide a holistic view of workforce behavior. Productivity indicators, both macro and micro, are explored, with machine learning algorithms employed to predict these indicators.

Contributions:

This paper introduces a pioneering multidimensional framework for dissecting productivity, underpinned by a diverse dataset encompassing mobile app usage metrics, job market statistics, and demographic data. It harnesses advanced machine learning techniques to predict a wide array of productivity indicators, enhancing the applicability of the study to real-world policy formulation. Furthermore, it offers actionable insights and recommendations tailored to policymakers, employers, and urban planners striving to optimize productivity across varied societal contexts.

Methodology and Results:

The methodology entails meticulous data collection, cleaning, and normalization, followed by categorization of app usage and job types. An innovative approach aligns mobile app traffic data with geo-spatial and socio-economic attributes at a granular level as shown in Fig. 1. Advanced machine learning algorithms correlate estimated productivity indicators with actual labor productivity indices. The study examines job openings, app usage, and skill gaps across various geolocations. While no clear correlation is found between app usage time and job openings, the Skill Gap Analysis reveals significant insights. Cities like Paris and Lyon show a demand for administrative and collaborative skills, respectively, based on app usage. Conversely, Marseille and Toulouse indicate a need for social media and general-purpose skills as illustrated in Fig. 2. These findings offer valuable opportunities for targeted app development and skill training programs.

Discussion and Future Directions:

The research presents a comprehensive view of workforce behavior and productivity in Lyon, France. It categorizes app usage, identifies patterns, and correlates them with productivity. The impact of different app categories on productivity is analyzed, revealing both expected and unexpected results. The study opens avenues for future research, including exploring additional variables, experimenting with different machine learning models, and deeper investigations into the relationship between app usage and productivity.

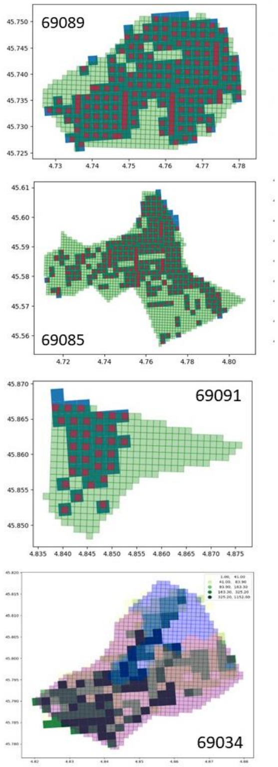


Fig. 1 Traffic Tile to Income Tile Mapping for Some Communes in Lyon

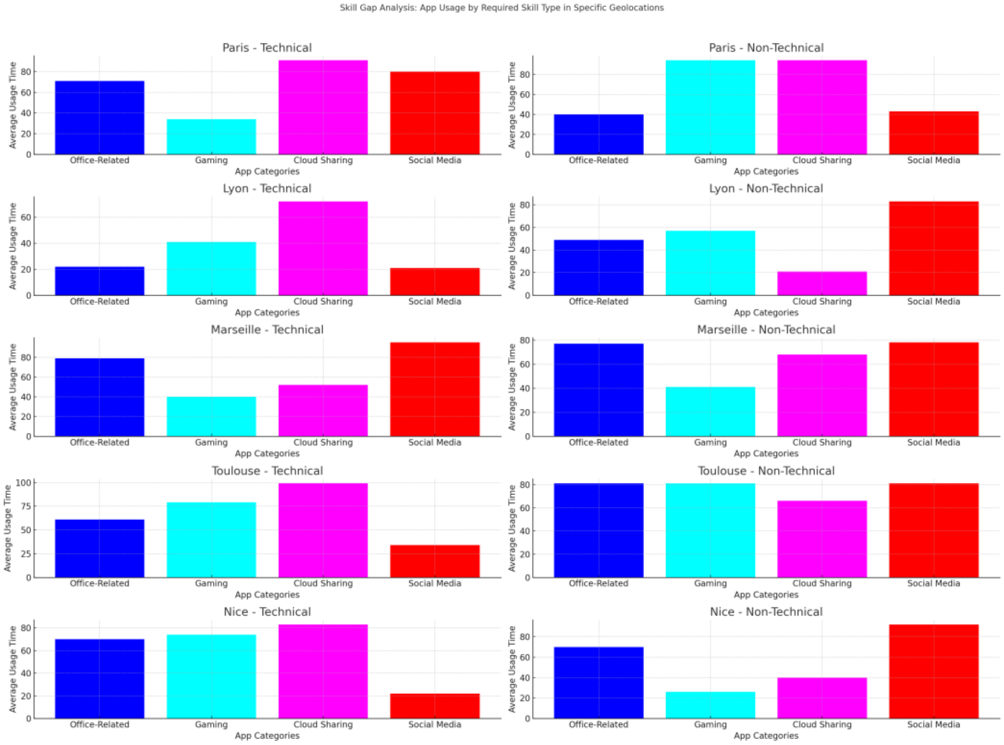


Fig. 2 Skill Gap Analysis

Conclusion:

This study pioneers an interdisciplinary approach to understanding workforce behavior and productivity. By integrating mobile app usage metrics with socio-economic and demographic data, it provides a comprehensive framework for analyzing and predicting productivity at a micro-level. The findings contribute to the broader discourse on leveraging digital data to enhance societal productivity, offering actionable insights for diverse stakeholders.

Acknowledgments

The authors would like to express their profound gratitude to Tawfik Yasser and Al-Amir Hassan. We are indebted to them for their time and efforts. They can be contacted at Tyasser@nu.edu.eg, AlHassan@nu.edu.eg.

References

[1] Smith, J. (2000). A critical survey of empirical methods for evaluating active labor market policies (No. 2000-6). Research Report.

[2] Johnson, M. H., Griffin, R., Csibra, G., Halit, H., Farroni, T., de Haan, M., ... & Richards, J. (2005). The emergence of the social brain network: Evidence from typical and atypical development. Development and psychopathology, 17(3), 599-619.

[3] Davies, N. M., Dickson, M., Davey Smith, G., Van Den Berg, G. J., & Windmeijer, F. (2018). The causal effects of education on health outcomes in the UK Biobank. Nature human behaviour, 2(2), 117-125.

[4] Lee, H. W., & Kim, E. (2020). Workforce diversity and firm performance: Relational coordination as a mediator and structural empowerment and multisource feedback as moderators. Human Resource Management, 59(1), 5-23.

[5] Kuznets, S. (1962). quantitative aspects of the economic growth of nations: VII. the share and structures of consumption. Economic Development and Cultural Change, 10(2, Part 2), 1-92.

[6] Green, K. L., Brown, G. K., Jager-Hyman, S., Cha, J., Steer, R. A., & Beck, A. T. (2015). The predictive validity of the beck depression inventory suicide item. The Journal of clinical psychiatry, 76(12), 15048.

[7] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR), 52(1), 1-38.

[8] Oyserman, D., & Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. Psychological bulletin, 134(2), 311.

[9] Williams, N., & Vorley, T. (2017). Fostering productive entrepreneurship in post-conflict economies: The importance of institutional alignment. Entrepreneurship & Regional Development, 29(5-6), 444-466.

[10] Martínez-Durive, O. E., Mishra, S., Ziemlicki, C., Rubrichi, S., Smoreda, Z., & Fiore, M. (2023). The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv preprint arXiv:2305.06933. Retrieved from https://arxiv.org/abs/2305.06933. [11] https://www.insee.fr/fr/statistiques/7233950, Visited Sept 2023.

[12] https://citypopulation.de, visited date Sept 2023.