Editors:

Francesco Calabrese
Esteban Moro
Vincent Blondel
Alex 'Sandy' Pentland

5-7 APRIL 2017
VODAFONE VILLAGE
MILAN

NetMob

BOOK OF ABSTRACTS

POSTER

vodafone

Universidad Carlos III de Madrid

UCL Université catholique de Louvain

INDEX

POSTER

POSTER SESSION 1 5 APRIL 2017

POSTER SESSION 2 6 APRIL 2017

5 APRIL 2017

POSTER
SESSION 1

Churn Prediction in the Telecommunication Industry using Social Network Analytics

María Óskarsdóttir*, Cristián Bravo†, Wouter Verbeke‡, Carlos Sarraute§, Bart Baesens*† and Jan Vanthienen*
*Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium,
Email: {maria.oskarsdottir, bart.baesens, jan.vanthienen}@kuleuven.be
†Southampton Business School, University of Southampton, United Kingdom, Email: c.bravo@soton.ac.uk
‡Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium,
Email: wouter.verbeke@vub.ac.be
§Grandata Labs, Argentina, Email: charles@grandata.com

***Abstract*—Identifying potential churners is important in competitive and saturated markets such as the telecommunication industry. Research has shown, that performance of models that predict customer churn is increased when social network methods are applied. Studies also indicate that relational learners, applied to customer call networks, are capable of predicting churn accurately. In this research, the difference in performance of a variety of relational learners is tested by applying them to a number of CDR datasets from around the world. In addition, performance of relational classifiers and collective inference methods is compared and the performance of models which combine relational learners with mainstream classifying methods, such as logistic regression, is studied. Our results show that collective inference methods do not improve the performance of relational classifiers and that the scores of relational learners combined with other customer features result in the best performing models.**

I. Introduction

In competitive and saturated markets, such as mobile telecommunications, identifying potential churners quickly and effectively is important. Telecommunication providers (telcos) are increasingly making use of network features, in addition to regular customer features, to capture behavior and interactions of customers when building churn prediction models, as these have been shown to add valuable information [1], [2]. Call detail records (CDR) are aggregated to build call networks, from which the network features are extracted. Alternatively, churn propensity can be inferred from call networks directly by means of relational learners [3], [4]. These methods simulate how churners affect the other customers, by propagating 'churn energy' through the network. The result is a score for each customer, that can either be used as indicator of churn directly or to enrich the customer dataset before building models using classical binary classifiers. In this study, we evaluate the predictive performance of various relational learners when used on their own and when combined with classical classifiers. In particular we

- test the difference in performance of a selection of 24 relational learners when predicting churn in mobile networks,
- study the effect of the two components of relational learners, relational classifiers and collective inference methods, and
- combine relational learners scores with other customer features when building churn prediction models using binary classifiers, to determine which sets of features provide the best performance.

To empirically evaluate the results, the methods were applied to eight distinct CDR datasets from around the world. This high number of datasets allows us to draw conclusions regarding the statistical significance of observed differences and provide conclusive answers.

II. Methodology

A. Learning in Networks

Relational learning in network data can be used to classify interlinked nodes in partially labelled networks, as was demonstrated in the network learning framework and toolkit NetKit [3]. Assuming that each node can belong to one of n classes and that the nodes have been partially classified, that information together with structure of the network, such as links and weights, can be used to infer classes for the unknown nodes.

Typically, relational learners are composed of two types of methods, namely relational classifiers and collective inference methods. Relational classifiers assign a class or score to each node in the network, depending on which class the neighboring nodes belong to and the weights of links between them. In this study, we use the weighted vote relational neighbor classifier (wvrn), the class distribution relational neighbor classifier (cdrn), the network-only link-based classifier (nlb) and the spreading activation relational classifier (spaRC), which is the classifier part of the spreading activation method [4]. In contrast, collective inference methods iteratively apply a relational classifier to stabilize the inferencing process [5]. Here, we've used Gibbs sampling (gibbs), iterative classification (ic), relaxation labelling (rl), relaxation labelling with simulated annealing (rlsa) and spreading activation collective inference method (spaCI), which together with spaRC makes up the spreading activation method [4]. We refer to [2] for a description of all these methods. To make up a relational learner, each relational classifier can be applied on its own or combined with a collective inference method, which results in 24 different methods.

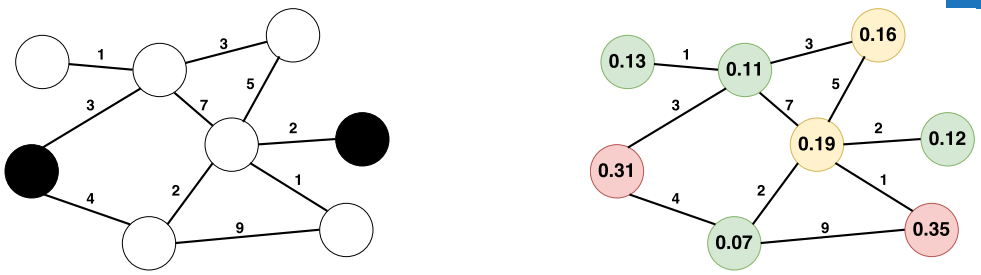


Fig. 1. The figure shows an example of an application of a relational learner. The figure on the left, displays a graph with seven customers, of which two have churned (black) and five have not churned (white). The figure on the right shows the same network after the RL has been applied. Each customer now has a score or probability of churning.

Although NetKit can be used for any kind of network, the specific application for customer churn in telco requires some adjustments [2]. In particular, the relational learners will produce a score between 0 and 1, since there are only two classes, churn (1) and non-churn (0). For churn, prediction are typically made into the future, where all labels are unknown. The learners were therefore applied to networks at time t, assuming that the churn status of all customers was known, to make prediction for the following time period $t+1$.

III. Datasets and Experimental Setup

TABLE I
DESCRIPTIONS OF DATASETS

ID	Origin	Year	#Customers	Churn Rate	Sparsity
1	Belgium	2010	$1.41 \cdot 10^6$	4.4%	$7.93 \cdot 10^{-7}$
2	Belgium	2010	$1.21 \cdot 10^6$	0.84%	$2.20 \cdot 10^{-6}$
3	North-America	2015	$1.57 \cdot 10^6$	0.71%	$3.14 \cdot 10^{-6}$
4	North-America	2015	$1.32 \cdot 10^6$	2.5%	$1.69 \cdot 10^{-6}$
5	Europe	2009	$4.33 \cdot 10^6$	8.5%	$9.42 \cdot 10^{-7}$
6	Europe	2008	$4.52 \cdot 10^6$	3.5%	$9.44 \cdot 10^{-7}$
7	Belgium	2012	$1.70 \cdot 10^5$	8.3%	$1.86 \cdot 10^{-5}$
8	Iceland	2015	$9.36 \cdot 10^4$	2.3%	$1.04 \cdot 10^{-4}$

Table I shows a summary of the eight distinct CDR datasets that were used in the study. Originating from around the world and ranging from 2008 to 2015, they vary in size, churn rates and sparsity. Each dataset spanned six months of call detail records, some including text messages in addition to phone calls. To ensure comparability of results, all datasets were preprocessed in the same way. Phone calls lasting less than a few seconds were disregarded and the activity of customers used to determine their churn status. In that regard, day of churn was defined as the first of 30 consecutive days without any activity. Subsequently, the datasets were used to build both long and short term networks, covering three and one month, respectively. The weights between customers in the networks were defined in two ways, with total duration of phonecalls between customers and the total number of phonecalls, aggregated over the given time period. Finally, churners in each network were labelled using their day of churn. In total there were four training networks for each dataset, and each of the two dozen relational learners was applied to all of them.

The model building process was twofold. On one hand, the scores produced by the relational learners were viewed as probabilities of churn. On the other hand, in a featurization process, common network features were extracted from each of the four networks together with RMF variables from the CDR datasets. We refer to these variables as Network Only features. Subsequently, churn prediction models were built using three sets of features: Network Only features, scores of the relational learners and both of these together. The binary classifiers which were selected to build the models were Logistic Regression, Neural Networks and Random Forest because of their high predictive performance and popularity in the industry [6].

Predictions were made for the month following the train periods, and was the same for each dataset. Model performance was evaluated using lift at 10% and AUC.

IV. Results

For the empirical evaluation of our results we follow the guidelines for statistical comparisons of classifiers over multiple datasets presented by [7]. To test for significant differences between the performance of the 24 relational learners, Friedman tests were applied to the rankings of both AUC and lift. Both tests were rejected, and so was the null hypothesis of equal average ranks, which means that there is a difference in performance between some of the relational learners. We refer to [8] for more details on these differences. Friedman tests were again applied to the performances of the relational classifiers and collective inference methods separately. These tests were also rejected, meaning that there are significant differences in performance, which were further explored using a post-hoc Nemenyi test. Figure 2 shows a comparison of the difference in performance measured in AUC of the relational classifiers and the collective inference methods. Blue boxes indicate differences that are significant. Performance measured in lift at 10% gave the same results. Regarding the relational classifiers, the only one that is always significantly different is the network-only link-based classifier, which consistently outperforms the other classifiers. This classifier builds a logistic regression model using information from the nodes' neighborhoods that is then used to classify the unknown nodes in the network [9]. For the collective inference methods, the

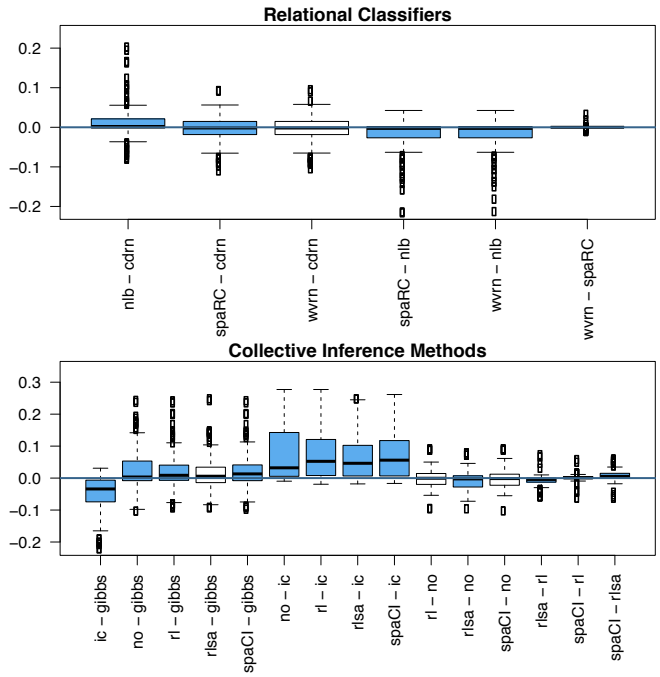


Fig. 2. Differences of Relational Classifiers and Collective Inferences. Blue boxes indicate differences that are significant.

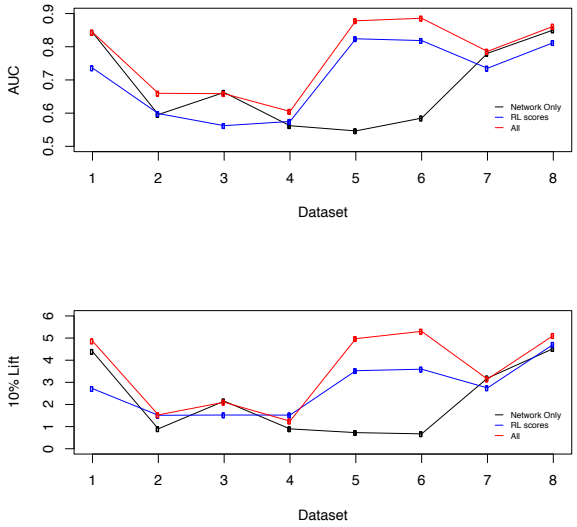


Fig. 3. Performance of logistic regression models using three sets of features from the eight datasets, measured in AUC and lift at 10%.

lower figure in figure 2 indicates that the iterative classification method always performs significantly worse than the other methods and that using no collective inference method always results in higher performance. This result was confirmed using Kruskal-Wallis test, which was rejected with a p-value of less than 0.01.

The performance of classical classifiers with different sets of features was also compared. First, Friedman tests were applied to the results ranked by model performance in AUC and lift. The tests were rejected, confirming a difference in average ranks in the performance of models using the three sets of features. A further exploration of these differences showed that models built with both sets of features performed significantly better than models using only one set of features, and that there was not a significant difference between the models which used only one set of features. The performance of models built using logistic regression for each of the eight datasets can be seen in figure 3, where the black line indicates models with network variables only, the blue line models with relational learner scores only and the red line models with both sets of features combined.

V. Conclusion

In the study we compared different ways of applying social network analytics methods when predicting churn in telco and evaluated our results empirically by applying them to a number of CDR datasets. According to our results, combining collective inference methods with relational classifiers does not improve performance, and the best performing relational classifier is the network-only link-based classifier [9]. In addition we showed that combining relational learner scores with other customer features in a binary classifier produces models with higher performance than using either of these features on their own.

References

[1] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, "Combining local and social network classifiers to improve churn prediction," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 651–658.

[2] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.

[3] S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *The Journal of Machine Learning Research*, vol. 8, pp. 935–983, 2007.

[4] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, 2008, pp. 668–677.

[5] D. Jensen, J. Neville, and B. Gallagher, "Why collective inference improves relational classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 593–598.

[6] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," *European Journal of Operational Research*, vol. 238, no. 2, pp. 505–513, 2014.

[7] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[8] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "A comparative study of social network classifiers for predicting churn in the telecommunication industry," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 1151–1158.

[9] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, vol. 3, 2003, pp. 496–503.

Joint Spatial and Temporal Classification of Mobile Traffic Demands

Angelo Furno*†, Marco Fiore‡, Razvan Stanica*
* Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, F-69621, Villeurbanne, France – name.surname@inria.fr
† IFSTTAR-ENTPE, Université de Lyon, F-69675 Bron, France – angelo.furno@ifsttar.fr
‡ CNR-IEIIT, Corso Duca degli Abruzzi 24, 10129 Torino, Italy – marco.fiore@ieiit.cnr.it

***Abstract*—We present an original approach to infer both spatial and temporal structures hidden in mobile traffic demands, by tailoring Exploratory Factor Analysis (EFA) techniques to the context of mobile phone datasets. Casting our approach to the time or space dimensions of such datasets allows solving different problems in mobile traffic analysis, i.e., network activity profiling and land use detection.**

I. Context

The surge in mobile data traffic –estimated globally at 3.7 exabytes in 2015, with a 74% increase over 2014 and an overall 4,000-fold growth over the past ten years [1]– has fostered the interest of the computer network community towards better understanding the dynamics of the mobile demand. Indeed, a proper characterization of how mobile services are consumed by subscribers can enable an informed, more efficient tailoring of network resource planning and management to the end users' needs. Knowledge mining of real-world datasets has revealed important features of mobile traffic. Examples include a strong temporal periodicity and geographic locality that enable effective prediction of the demand; the appearance of significant fluctuations induced by social events with the consequent need for dedicated resource management policies; or, a neat heterogeneity of the capacity consumed by subscribers that is however captured by a limited number of typical profiles, which enables the informed tuning of traffic plans. A survey of these results is in [2].

Our work focuses on the problem of classification, i.e., finding hidden regular structures in the network-wide aggregate traffic generated by mobile users. The problem can be cast in the temporal or spatial dimensions, both of which are relevant to network operation management.

- In the temporal dimension, the problem is that of network activity profiling, i.e., classifying together time periods that show a similar, stable spatial distribution of the mobile traffic demand. Network activity profiles find applications in cognitive networking [3], as they can drive the establishment and relocation of resources in concert with the temporal variations in the mobile demand [4].
- In the spatial dimension, the problem is that of land use detection, i.e., the decomposition of a geographical area into zones where the mobile traffic dynamics are homogeneous over time. These zones typically correspond to land uses, i.e., the combination of urban infrastructures and predominant undertakings of people at a location. This knowledge can support the dynamic allocation of capacity at individual base stations, and help mitigating high fluctuations of resource needs in small-sized network areas [4]. In addition, land use detection has applications in geoinformatics, as an effective way to automatically label the urban tissue, at lower cost and with higher accuracy than traditional survey methods.

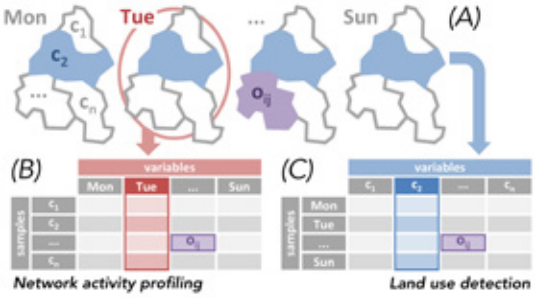


Fig. 1. Mobile demand classification with EFA in a toy scenario. (A) The one-week demand in the target region is aggregated daily with respect to a spatial tessellation of n cells. The resulting demand in the i-th cell during the j-th day is EFA observation o_{ij}. (B) Network activity profiling: days are the EFA variables, each characterized by a set of observations over the cell samples. (C) Land use detection: cells are the EFA variables, each characterized by a set of observations over the daily samples. Figure best viewed in colors.

We present an original methodology for the joint spatiotemporal classification of the aggregate demand that a mobile network has to serve. Our solution stems from Exploratory Factor Analysis (EFA), a well established instrument in psychology research. EFA aims at identifying, in a fully automated way, latent factors that cause the dynamics observed in the data. When tailored to the specific use case of mobile traffic classification, EFA offers the possibility of exploring the space and time dimensions of the data at once. Moreover, EFA allows immediate extrapolation of the structures hidden in the secondary dimension of each problems. In other words, it provides, at no additional cost, knowledge of the spatial patterns that characterize each network activity profile, and of the precise temporal dynamics that distinguish each land use.

II. EFA for mobile demand classification

Given a set of observed variables of interest, EFA is formally defined as *"a model of hypothetical component variables (named factors) that explain the linear relationships existing between observed variables"* [5]. EFA can be tailored to the classification problems of network activity profiling and

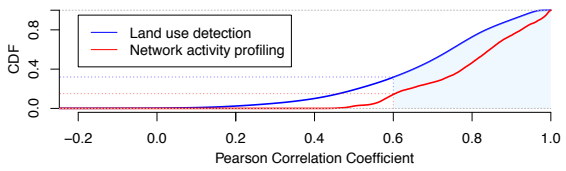


Fig. 2. Distributions of the Pearson correlation coefficient computed between all pairs of EFA variables in the two mobile demand classification problems.

land use detection, as exemplified in Fig. 1. In both cases, the input is an aggregate representation of the communication activity of mobile subscribers in the geographical region of interest. This definition of input is general and can accommodate any level of spatial and temporal aggregation, as well as any notion of mobile user activity (e.g., voice, text, data).

Network activity profiling. We model *time intervals* as the EFA variables. Each variable is thus described by the mobile traffic demand (i.e., the EFA observations) recorded over all spatial cells during a given time interval, as shown in Fig. 1. In this EFA configuration, the common factors seeked by EFA are temporal structures that explain at what time instants the spatial distribution of the mobile demand is comparable: these structures are precisely network activity profiles.

An important remark is that, here, spatial cells map to EFA samples: hence, EFA scores relate cells to temporal profiles, revealing which geographical areas are important for a given temporal profile. This allows the joint inspection of the classification results in the space and time dimensions.

Land use detection. EFA variables correspond to *geographical locations*. Each variable consists in the mobile traffic demand (i.e., the EFA observations) recorded at a specific cell through the complete monitoring period, as in Fig. 1. In this EFA configuration, the EFA common factors represent structures in the geographical space that explain in what areas the mobile demand follows similar temporal dynamics: by definition, such areas correspond to land use classes.

Interestingly, time intervals become now the EFA samples. Therefore, EFA scores point out the time periods when the mobile demand is especially distinctive within each land use. This offers an unprecedented spatiotemporal perspective on land uses, and showcases again the potential of EFA for the concurrent spatiotemporal analysis of mobile traffic data.

Finally, we recall that EFA builds on the hypotheses of linearity of the functional relationships among the observed variables and the unknown hidden factors. In order to verify the validity of such an assimption in the case of mobile traffic data, we run the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy on the two datasets used for our performance evaluation in Sec. III. In both our classification problem formulations, and for all datasets, KMO returns values around 0.99, indicating a high suitability of the data to EFA. As additional checks, we verify: *(i)* the linearity of all pairwise relationships between EFA variables in the two mobile demand classification problems, finding strong correlation in 70–80% of cases, as shown in Fig. 2; *(ii)* the sample-to-variable ratio, finding that it is always much larger than one, i.e., a good rule of thumb for a meaningful factor analysis.

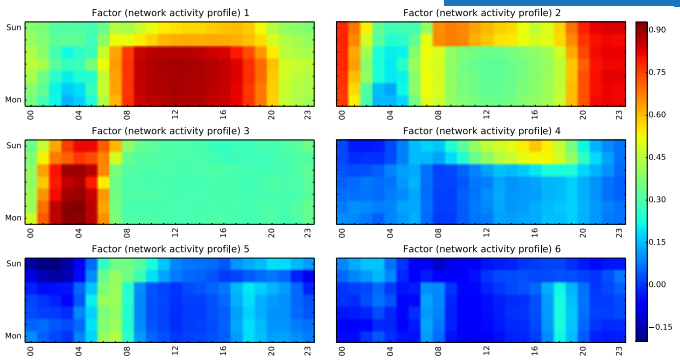


Fig. 3. Network activity profiling. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) over the median week in the TIM-2013 dataset. Loadings of the 24×7 hours of the week (i.e., EFA variables) on the six profiles (i.e., EFA factors). Figure best viewed in colors.

III. Selected Results

We demonstrate the quality of the spatiotemporal classification provided by EFA with real-world mobile traffic data recorded by national operators in two major European cities.

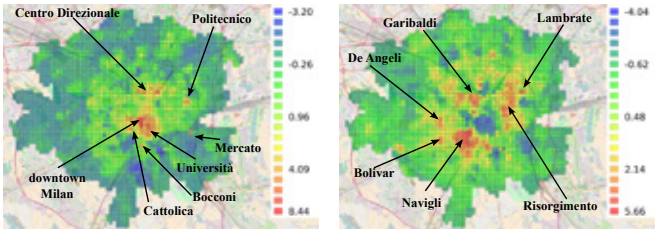
TIM-2013 dataset. The data was released by TIM as part of their Big Data Challenge. The dataset describes the mobile traffic generated by subscribers in the Milan conurbation over a two-month period spanning November and December 2013. The data includes voice, text, and Internet traffic of approximately 400,000 users, aggregated during time intervals of 10 minutes. Traffic volumes are georeferenced with respect to a regular-grid space tessellation of 235×235-m^2 cells.

Orange-2014 dataset. This dataset consists of Call Detail Records (CDR) collected for billing purposes by the operator. CDR describe hourly volumes of voice and text activity in the Paris metropolitan region, on a per-antenna basis. The data was collected from a sample of 100,000 users in September, October and November 2014. We employ a standard Voronoi tessellation to represent the spatial coverage of cells and the geographical distribution of traffic.

A. Network activity profiling

We investigate the performance of EFA for the profiling of network-wide mobile traffic activity over time. Let us first focus on a one-week period that is representative of the typical communication activity recorded in the TIM-2013 dataset. To that end, we condense the two months of data into one single *median week*, so as to mitigate potential classification biases due to outlying behaviors. For each cell in the Milan area: (i) we aggregate the demand of incoming/outgoing calls and texts on a hourly basis; (ii) we associate to every hour of the week the median value of all corresponding hourly demands (e.g., all Mondays at 8 am).

The network activity profiles (i.e., EFA factors) identified by EFA in the TIM-2013 median week are portrayed in Fig. 3. Each plot shows the loadings of every hour of the week (i.e., every EFA variable) on a specific profile, according to the color range on the right of the figure. For example, hours from 8 am to 6 pm on Monday to Friday have very high loadings on factor 1, and low loadings on all other factors: hence, the first

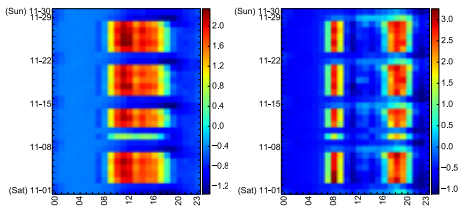


(a) Factor (network activity profile) 1 (b) Factor (network activity profile) 2

Fig. 4. Network activity profiling. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) over the median week in the TIM-2013 dataset. Thurstone's scores of the 434 cells (i.e., EFA samples) on the first two profiles (i.e., EFA factors). Figure best viewed in colors.

(a) Factor (land use class) 2 (b) Factor (land use class) 3

Fig. 5. Land use detection. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) in the Orange-2014 dataset. Loadings of the 1596 Voronoi cells (i.e., EFA variables) on two (out of fourteen) representative classes (i.e., EFA factors). Figure best viewed in colors.



(a) Factor 2 (b) Factor 3

Fig. 6. Land use detection. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) in the Orange-2014 dataset in November. Thurstone's scores of the 91×24 hours (i.e., EFA samples) on two (out of fourteen) classes (i.e., EFA factors). Figure best viewed in colors.

network activity profile characterizes the work hours. Overall, EFA allows identifying the following profiles in Milan: (1) working hours; (2) relax hours in the evenings and weekend mornings; (3) overnight hours; (4) weekend afternoon hours; (5,6) morning and afternoon hours corresponding to the start and end of the work activity.

EFA also lets us understand which geographical cells (i.e., EFA samples) are the most relevant to a specific profile (i.e., EFA factor). In other words, scores tell us *where* the mobile communication activity that characterizes a profile takes place. Fig. 4 shows the scores, estimated via Thurstone's regression, of all geographical cells in Milan on the six network activity profiles. As an example, we can remark how the cells loaded by the first profile clearly highlight downtown Milan, where the business district and most offices are located. Additional geographical areas that have high scores on this profile are university campuses (Politecnico di Milano, Università di Milano, Bocconi, Cattolica) and commercial zones (public entrance to Mercato Ortofrutticolo). Clearly, these are the locations where mobile communication activities surge during the work hours, i.e., those hours that have high loadings on the first profile.

Equivalent analyses are possible for all of the other profiles. For instance, during evening and weekend mornings (profile 2), the network activity is much reduced in the city center, as shown in Fig. 4b. Prevalent areas are within the inner city beltway and characterized by a dense presence of bars, restaurants, and clubs (Navigli, Lambrate, Porta Garibaldi, Piazza Bolivar) or by a strictly residential nature (Risorgimento, De Angelis).

B. Land use detection

When considering classifications in the spatial dimension, Fig. 5 shows two out of fourteen land use classes detected by EFA in the Orange-2014 dataset. For the sake of clarity, in each plot we only show cells (i.e., EFA variables) that have a high loading on the class (i.e., EFA factor) the map refers to. The factor in Fig. 5a corresponds to business areas (e.g., downtown Paris, La Defense or Issy-les-Moulineaux). The factor in Fig. 5b is less trivial, and characterizes very well (93% precision with 96% recall) the deployment of subway stations in the city (black dots in Fig. 5b). This shows how EFA can reveal many microscopic land uses where the mobile traffic demand follows distinctive dynamics.

EFA also returns scores that indicate the relevance of time (i.e., EFA samples) to each class (i.e., EFA factor). Thus, scores tell us *when* each land use class shows an especially remarkable mobile communication activity. Fig. 6 shows the hourly scores during one month of Orange-2014 data on the land use classes presented above. As expected, the working hours are most relevant to business areas in factor 2, shown in Fig. 6a: interestingly, morning hours seem more concerned than afternoon ones. One can also easily detect that the typical working day spans from 8 am to 6-7 pm: these are the morning and afternoon commuting times when the subway stations of factor 3 experience exceptionally high loads, as displayed in Fig. 6b. In all plots it is easy to spot an irregularity, due to a public holiday (Armistice day, November 11).

IV. Conclusions and Perspectives

We propose an original approach to the spatiotemporal classification of mobile traffic data, based on Exploratory Factor Analysis (EFA). Tests with heterogeneous real-world datasets demonstrate the versatility of EFA, which provides supplementary knowledge (i.e., the geographical perspective of profiles and the temporal view of land uses) that proves paramount to the interpretation of the results.

References

[1] Cisco VNI Forecast, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," 2016.
[2] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Large-scale Mobile Traffic Analysis: A Survey," *IEEE Communications Surveys & Tutorials*, 18(1), 2016.
[3] R.W. Thomas, L.A. Da Silva, A.B. MacKenzie, "Cognitive networks," *IEEE DySPAN*, 2005.
[4] H. Assem, T. Sandra Buda, L. Xu, "Initial use cases, scenarios and requirements," *H2020 5G-PPP CogNet*, Deliverable D2.1, 2015.
[5] S.A. Mulaik, *Foundations of Factor Analysis*, CRC Press, 2009.

Understanding mall visiting patterns and mobility in Santiago de Chile with CDR data

M. G. Beiró, C. Cattuto,[1] L. Ferres, E. Graells-Garrido, L. Bravo and D. Caro[2]

1. ISI Foundation, Turin, Italy
2. IDS, Faculty of Engineering, U. Del Desarrollo & Telefónica R&D, Santiago, Chile

1. Introduction

Shopping malls became popular in Chile in the early 80's, and are one of the first signs of the globalization and liberalization of consumption in Latin America. They are currently one of the main elements of Chilean popular culture, influencing people's mobility, daily activity, social mixing and segregation processes [1, 2].

In this work we analyze the mobility patterns of people going to malls in the Chilean capital city, Santiago, in order to determine which factors influence mall election and what types of social mixing are found in malls. We used Call Detail Records (CDRs) provided by Telefónica R&D. Our study comprises 16 malls and 481 cellphone towers inside those malls. The dataset consists of 1,023,118 unique mobile devices (i.e., containing a Movistar/Telefónica SIM card), of which we identified 942,091 as visitors who went to malls 1,471,637 times in the course of a single month, August 2016.

2. Methodology and results

Device filtering. In an initial pre-processing step we designed an algorithm filtering non-visitor devices (in particular, SIM cards associated to card readers, and those associated to people who are not visitors, such as employees or providers). This algorithm is based on the frequency of appearances in the mall network. If a device is there every single day or at late night hours, it's not a "visitor".

Home location determination. For those visitors identified by the method above, we determined the general area where they live by finding the first (before 8AM) and last (after 10PM) tower to which they connected each day of the month. We then computed, for each user: (i) The number of times in which we detected a first/last tower connection and (ii) the relative frequency of the most observed tower (i.e., the number of times in which the user's first or last tower was that tower, divided by the previous quantity). Results show that more than 70 % of people were connected for at least 80 % of days. About half the users repeated their most common first-last tower at least 60 % of the days in which they are seen. In order to ensure a good confidence for the home location determination, we only kept these last consumers.

Antenna Filtering. We started with a database of 13,733 antennas in all the province of Santiago. Malls have their own cellphone towers and antennas within their walls. Since they are usually low power, connections to those antennas are mostly established from inside the mall. We identified these antennas by (i) finding antennas with the word "mall" in a human-targeted description field in the dataset, and (ii) by drawing each mall's polygons and using the towers' coordinates to know which ones were contained by them.

After this preprocessing we built a mall interaction network (Figure 1) identifying people who visited several malls during the month. The network was built as a weighted graph whose 16 nodes are malls and the weights in the edges represent the number of people visiting both malls at least once during the month. As some malls have more inherent activity than others, we normalize these weights by the total number of visits of each mall. Then, the network becomes directed and asymmetric, and the weight of the interaction $A \to B$ will represents the probability that a visitor goes to mall B, given that he visits mall A.

Figura 1: Mall co-visitation network.

From this network we extracted clusters of malls with similar characteristics in terms of client profile and location. The Human Development Index (HDI) at the *comuna* level was obtained from [3], and we used the HDI of the home *comuna* as a proxy for socio-economic status. Figure 2 plots the distribution of the HDI, computed through a kernel density estimation, for the clients of each mall. We also computed the degree of social mixing in each mall, in terms of the entropy of the distribution of their clients' socio-economic profile. We detected three clusters with different client profiles:

- A first cluster comprises the malls Plaza Tobalaba, Líder Puente Alto, Jumbo Departamental and Mall Plaza Vespucio in the south-eastern area of the city. These malls mix people from different socio-economic levels: the middle-class *comunas* of La Florida and Macul and the low-class *comunas* of La Pintana, La Granja and San Joaquín. Interestingly, these people are the ones that travel less from their home to the mall, with average distances ranging from 3.64 km to 4.93 km.

- In the western cluster, involving Mall Arauco Maipú, Plaza Oeste, Mall Paseo Estacin and Plaza Alameda, we find low-income consumers from the *comunas* of Maipú, Lo Prado, Pudahuel and Estación Central. However, they also capture middle-class people from the Santiago and San Miguel *comunas*. In this cluster people travel more (from 4.90 km to 6.83 km) from their homes to the mall.

- Though the rest of the malls constitute a third large cluster, there are some interesting differences in their visitors' profiles. In particular, the north-eastern malls of Lder La Dehesa, Apumanque, Alto Las Condes and Parque Arauco are exclusive for high-income people from Las Condes, Vitacura, Providencia and Lo Barnechea. In order to arrive there, people travel between 5.76 km to 7.00 km. Interestingly, people do not mix here, as people from low income *comunas* do not arrive. Instead, the Costanera Center, Plaza Alameda and Panorámico are the most heterogeneous ones, receiving people from all the city, and from average distances between 6.83 km and 7.36 km. Together with Plaza Egaña, they are the 4 malls with largest heterogeneity among their visitors, in terms of the entropy of their distribution of socio-economic levels. Finally, Plaza Norte is a very particular mall, because it is the only one which covers the northwestern area of the city. Because of this, people travel an average distance of 7.86 km to arrive here (the largest among all malls), and the mall mixes people from low (Conchalí and Renca) and middle socio-economic levels (Quilicura).

3. Conclusion

We observed that mall choice is mainly determined by two factors: distance and socio-economic level. People from poor peripheral areas have their own local malls which they visit and they do not travel much, while people closer to the city center, from low and middle classes, are willing to travel more in order to arrive to central, heterogeneous malls in which people mix socially. People from rich classes in the northeast of the city are not so willing to move, and they usually visit high-target malls which are more exclusive and homogeneous in their visitor profile. Population "mixing" patterns is an important topic and we we plan to extend this work by analyzing how they are induced by other retail "hubs" and other socio-economic factors.

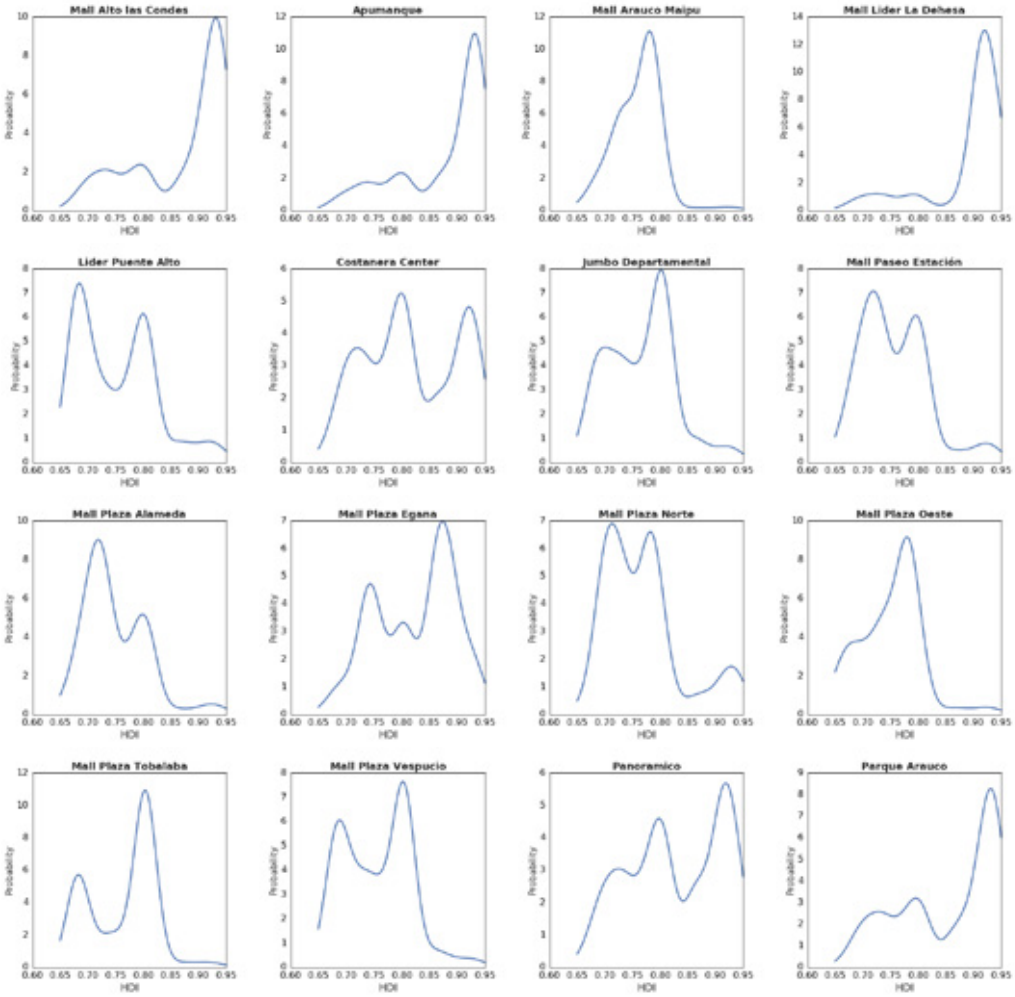


Figura 2: Probability distribution of the HDI (Human Development Index) of the origin *comunas* for the customers of each mall.

[1] Salcedo R., De Simone, L., Los Malls En Chile: 30 años, Cámara Chilena de Centros Comerciales, Santiago de Chile, 2013.
[2] Davila A., The Spatial and Class Politics of Shopping Malls in Latin America, Univ. of California Press, 2016.
[3] Aedo A., Delgado I., Díaz R., Godoy S., Márquez R., Melis F., Palma A., Salfate V., Sierra M.L., Las trayectorias del Desarrollo Humano en las comunas de Chile (1994–2003), Temas de Desarrollo Humano Sustentable, 11, 2004.

Developing and Deploying a Taxi Price Comparison Mobile App in the Wild: Insights and Challenges

Anastasios Noulas*, Vsevolod Salnikov†, Desislava Hristova‡, Cecilia Mascolo‡ and Renaud Lambiotte†
*Data Science Institute, Lancaster University, a.noulas@lancaster.ac.uk
†Department of Mathematics, University of Namur, name.surname@unamur.be
‡Computer Laboratory, University of Cambridge, name.surname@cl.cam.ac.uk

I. Summary

In response to the growing complexity of taxi transport dynamics, which affects a growing number of cities around the world [4], we describe the process of development of OpenStreetCab, a mobile application that aims to assist users in choosing a taxi provider in a city in real time, offering estimates on taxi prices. We reflect on our design decisions and discuss the application's usage and pricing statistics between two cities and two taxi providers. We provide a validation of the app's price estimates and a comprehensive study of price and journey time measurement through a real world experiment which compared taxi providers in the city of London.

II. Application and System Design

We have first launched the app in New York City in March 2015, and subsequently in London in the very beginning of January 2016. As mentioned above, users that install the app need first to select their city of interest (London or New York). Subsequently in the journey query submission screen they can specify their trip's origin and destination. We provide two functionalities to enable user localisation: first, a button next to the origin input tab that automatically sets the origin address, given the user's geographic location (through GPS / WiFi sensing), and, second, a text-input geocoding that parses user input and matches it to the most similar address name.

After setting the origin and destination addresses for a journey the user can press a button, 'Uber or Yellow?' in New York or 'Uber or Black' for London for comparison between Uber X and Yellow Cabs or Black Cabs respectively. This will push the input query to our server where Uber prices are compared to the competing local provider. Next, the user is presented with a screen where price estimates are provided, including an indication on the price difference ('Savings'), with an additional projection of a colored header at the top of the screen clearly indicating the taxi provider for which the estimate is lower (e.g., Yellow for yellow taxis in New York). In addition to the user input, data is collected on the time of the user query: a GPS sample of the user's current location and the application installation unique identification number. The latter has been useful to associate users with submitted queries over time, as we required no registration information for our users.

Uber prices for the journey are collected through the Uber developer API [1]. The API returns two values, min and max, that define a price range for the costs of its Uber X service. Next, the mean estimate is calculated from these values, rounding to the closest integer value. We chose to provide the mean as opposed to ranges, as in a list of a few providers it would be easier to compare on a single value as opposed to a range. Traditional taxi providers do not typically provide APIs on pricing. Instead, different taxi companies use different tariff schemes. We therefore combine information on tariffs for Yellow and Blacks Cabs in New York and London respectively, with routing information offered by HERE Maps [1]. HERE Maps return a shortest, in terms of time duration, routing path that is sensitive to traffic information the company gathers from a variety of sources. We then simulate the taxi's meter along the route and estimate the price of a journey according to the tariff information in each city. Black Cabs in London feature a more complex tariff logic [2] than the Yellow Taxi company in New York [3]. In principle, tariff schemes apply a flat cost known as *flag* in the beginning of the journey when passenger boards and then the price meter increases as a function of time and distance. For example, a rule may suggest that fare increases by a fixed ammount (e.g. X U.S. Dollars) after Y meters or Z seconds (whatever comes first). HERE Maps returns the routes as a set of segments, technically referred in the system as *manoeuvres*. For each route segment there is information on the length in kilometers and the typical driving time taken to drive on the segment. We exploit this information to increment the fare of the journey according to the tariff rules. Tariff rules depend also on time (e.g. morning versus night) and dates (e.g. holidays versus regular days) and we have integrated this aspect of pricing into OpenStreetCab as well. What is more, special destination or origin points such as airports or train stations can imply additional costs as well as costs that are specific to the route of the journey taken such as tolls. As currently there is no system that provides such information on routes, we have relied on keeping our system to date through manual labour and very critically on user feedback.

III. User Growth and Application Statistics

Overall, since the launch of the app in March 2015, more than 13,000 users have installed it in the two cities with around 75% of all installs taking place on an iOS platform

[1] https://here.com/
[2] https://tfl.gov.uk/modes/taxis-and-minicabs/taxi-fares
[3] http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml

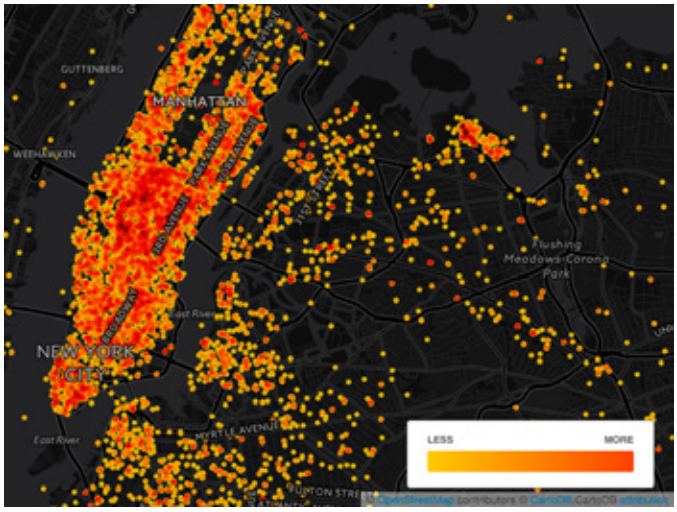


Fig. 1: Query distribution in New York considering journey origin.

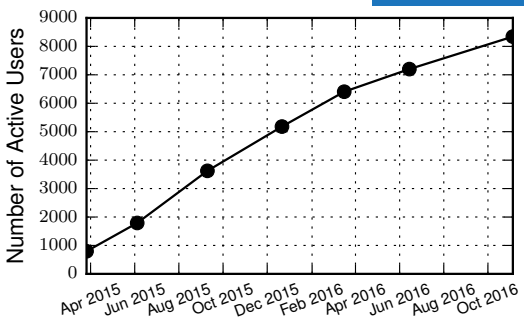


Fig. 2: User growth over time considering users with at least one query since install.

		Application Statistics	
City	*iOS installs*	*Android installs*	*Queries*
New York	9340	3095	25804
London	1030	345	3371

TABLE I: Summary of application statistics across platforms and cities.

and the rest on Android. In Figure 2 we present the number of OpenStreetCabs users that have submitted at least one journey query. Approximately 8000 users have submitted a query, more than 70% of those ever installed the application. Usage trends vary seasonally, but the number of total users with at least one query every three months is in the range of 1500 to 2000. The average number of queries per user is equal to 3.12 with almost 350 user having submitted 10 queries or more. In Table I we provide a summary of the statistics by city together with the total number of journey queries submitted. Regarding the number of queries, in New York there were a total of 25,804 queries submitted to our server. The geographic dispersion of user queries is shown on the map of New York in Figure 1, where the heatmap shows the spatial variations in query frequency. As expected most activity is concentrated in Manhattan with occasional hotspots in peripheral areas that include New York's La Guardia airport.

We have measured an average saving of 8 U.S. Dollars per journey considering the mean difference between provider prices in each query. This corresponds to total potential savings of almost 206,000 U.S. Dollars for the app's users assuming that they always choose the cheapest provider. The number of queries in London are 3,371 with potential savings of 12,405 British Pounds on an average price difference of 3.68 GBP (Great British Pounds). While this number may not be reflective of the real amount of money saved, since users may not pick always the cheapest provider (e.g. due to personal criteria regarding service quality), its scale is indicative of the potential financial impact that similar apps can have on the taxi market.

IV. Taxi Experiments in the Wild

In order to validate the app's price estimates we ran a three day experiment on the ground in the city of London. Beyond validating prices, we took this opportunity to measure journey times and routing behavior for the two competing taxi providers in the city of London: Uber – focusing on their basic Uber X service, and the city's traditional Black Cab service. There are some well known differences between the two services which we take into consideration in analyzing the output of our experiments. To acquire a license in London, Black Cab drivers need to attend a school that takes about three years to complete and pass *The Knowledge* [3] test that thoroughly examines the ability of drivers to know by heart the whereabouts of a large number of streets and points of interest in central London. Notably, medical tests on these drivers have suggested that their training and profession results to a larger number of cells in the hippocampus region of the brain which is the region that hosts the spatial navigation mechanism for mammals [2]. Another advantage of the Black Cab service is that they are licensed with Transportation for London, which means they can use bus lanes across the city. On the other hand, Uber drivers do not receive any special training and rely exclusively on their navigation system. These differences are noticeable to users of the two services in the city but no quantifiable data-driven insights exist on these differences until now.

A. Experimental Setup and Conduct

The experiment took place in London over three consecutive days in February 2016. Two researchers performed 29 side-by-side journeys comparing the prices, times and routes between Uber X and Black Cab in London. Using an in-built route tracking functionality (in the latest version of the app online the feature is enabled for all users), the GPS coordinates of trajectories followed by each provider were recorded along with their respective timestamps, start and stop journey times and price estimates from the app. Black Cab and Uber receipts were collected in the end of each trip so estimates could be compared to actual prices.

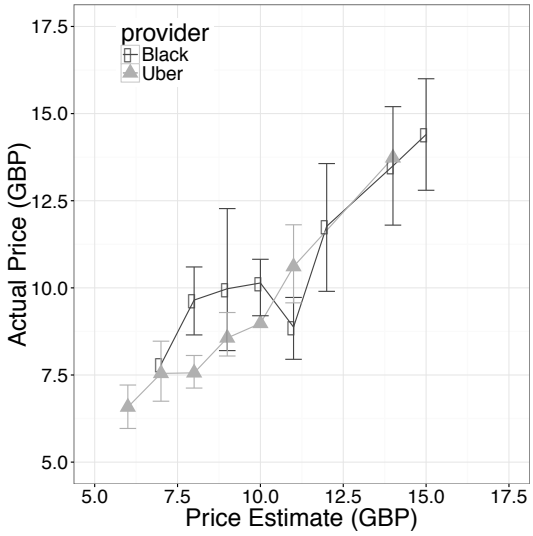
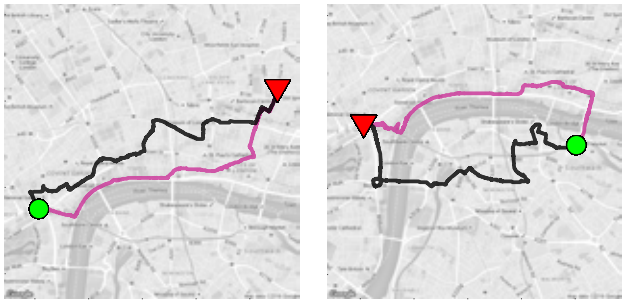


Fig. 3: Application price estimates versus average actual amounts paid in London. The errors bars correspond to standard errors.



(a) Trafalgar Square - Old Street

(b) London Bridge - Trafalgar Square

Fig. 4: Taxi provider trajectories in two areas of London. Black Cab in black color and Uber X in pink. Origins are marked with a Green circle and Destinations with a Red triangle.

B. Experimental Results

Price Estimates: For every journey with an Uber X or a Black Cab in the experiment, we have compared our application's estimate measured as described in Section User Growth and Application Statistics against the actual price charged by the provider. In Figure 3 we plot the mean actual price charged for a given price estimate and the corresponding standard error. We consider the overall estimates for both providers satisfactory, yet deviations exist. For Black Cabs deviations were higher for journeys that cost between 7 and 9 GBP. In the case of Uber, estimates tend to be more stable, however, deviations still remain.

Provider Comparison: We have empirically observed significant variations in terms of how the two providers compare in terms of actual and estimated prices, with routing choices being the most probable reason for these deviations. In Figure 4

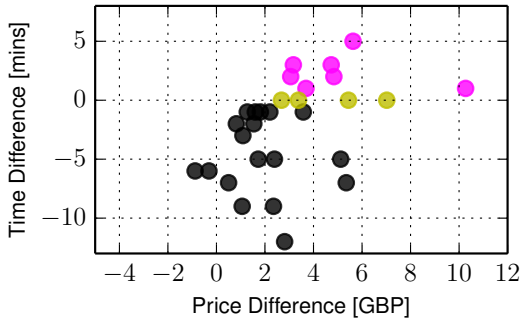


Fig. 5: Price versus time differences where price difference is defined as Black Cab price minus Uber price in GBP and time difference as Black Cab journey time minus Uber X journey time in minutes. Black colored circles correspond to faster journey times for Black Cabs, pink for Uber X and yellow for ties.

we show two characteristic journeys where routes had very little geographic or no overlap at all between the two providers. Black cab drivers tend to take more complex routes in terms of picking side streets as opposed to larger main streets that are recommended more often by GPS navigation systems as part of shortest path routing. As already implied in previous sections by tariffs applied Black Cabs were in general more expensive. Uber X would cost on average 74% of a Black Cab's journey price.

Nevertheless, Black Cabs were faster and took on average 88% of an Uber's trip duration, where average journey time has been 14.06 minutes Black Cabs and 16.34 minutes for Uber (Uber or Black Cab waiting times excluded). Out of the 29 journeys, Black Cabs were faster in 18 cases, there were 4 ties and Uber X was faster in 7 instances. Figure 5 presents a scatter plot reflecting the relationship between price and time differences. The faster Black Cabs have been, as one would expect by definition of the pricing schemes that depend on time in addition to route length, the smaller the price difference. Further, when Black Cabs have been faster, in almost half of the occasions (10 times) they have been faster by 5 minutes or more.

References

[1] Uber API, 2016. https://developer.uber.com/.
[2] E A Maguire, D G Gadian, I S Johnsrude, C D Good, J Ashburner, R SJ Frackowiak, and C D Frith. Navigation-related structural change in the hippocampi of taxi drivers. *PNAS*, 2000.
[3] The Public Carriage Office. The knowledge, 2016. www.the-london-taxi.com/Public_Carriage_Office.
[4] Uber. Uber list of cities, 2016. https://www.uber.com/cities/.

THE COST OF TAXING NETWORK GOODS: EVIDENCE FROM MOBILE PHONES IN RWANDA

DANIEL BJÖRKEGREN*

BROWN UNIVERSITY

Developing countries heavily tax mobile phones. However, these taxes can stunt growing networks. I use a method to estimate and simulate the adoption of mobile phones, using transaction data from nearly the entire network of Rwandan mobile phone subscribers over 4.5 years. Standard metrics that ignore network effects would suggest handset taxes are desirable for governments, with a welfare cost under $1.22 per dollar of government revenue. I find that these grossly underestimate the cost of taxing a growing network—accounting for network effects, the true cost exceeds $2.93: much more costly than other taxes. Further, baseline taxes heavily burden the poor: the lowest half of users receive only 2% of consumer surplus, but account for 19-20% of government revenue. Had the government shifted taxes entirely from handsets to usage it could have increased the consumer surplus of the lowest half of users by 38%, without substantially lowering the surplus of the upper half of users and while raising more government revenue.

Full version of paper: http://dan.bjorkegren.com

1. Introduction

Generating public revenue is a perennial challenge for developing countries, where collection costs are first order. Many governments are confined to a small set of feasible instruments which can be distortionary (Gordon and Li, 2009). However, even in countries with very little other capacity to collect revenue, telecom represents a thriving sector operated by a few formal firms that can easily be taxed.

Developing country governments recognize this convenient source of revenue: the mobile industry contributed an average of 7% of government revenue in sub-Saharan Africa as early as 2007 (GSMA, 2012).[1] In addition to standard taxes, governments charge spectrum license fees and specific taxes on telecom equipment, mobile handsets, and airtime. While it is clear that this emerging sector provides a public finance opportunity for poor countries, it is unclear how to best exploit this opportunity. There is a widespread concern that countries may continue to tax telecom heavily in the short term at the expense of long term growth; the former Director of ICT at the World Bank, Mohsen Khalil, cautions: "the indirect benefits to the economy of having affordable access to telecommunications services far outweigh any short-term benefit to the budget."

Developing countries thus face a tension between generating revenue and extending service, particularly to rural and low income areas ('a paramount concern' in the words of former World Bank ICT Director Mohsen Khalil). Governments typically manage this tension with a set of telecom-specific taxes, and regulations and programs that encourage access to the rural poor. However, there is little evidence to guide the design of these policies, and standard approaches that do not account for network effects can give misleading estimates.

While theoretical work provides intuition about network effects, there is little empirical work to guide policy choices.[2] Empirical work has been limited for three reasons. It is costly to measure an entire network using traditional data sources. It is also difficult to identify network effects: one individual may adopt after a contact adopts because the contact provides network benefits, or because connected individuals share similar traits or are exposed to similar environments. And even if these two issues are overcome, it is difficult to evaluate policies, which can cause effects to ripple through the entire network. As a result, there remain open questions about how to design policies that better capture the spillover benefits associated with network effects, as well as policies that overcome suboptimal provision arising from high concentrations in industries providing network goods.

In this project, I overcome previous limitations using an empirical approach further detailed in Bjorkegren (2015) and 5.3 billion transaction records from Rwanda's dominant mobile phone operator as the network expanded from 300,000 to 1.5 million subscribers. I estimate a structural model of demand for mobile phones, and then use this model to simulate the effects of alternate tax policies.

2. Model

My method has three steps:

First, acknowledging that the utility of owning a mobile phone is derived from its usage, I model the utility of using a phone. I observe every connection between subscribers, as well as the calls placed across

*E-mail: danbjork@brown.edu, Web: http://dan.bjorkegren.com

Revision December 28, 2016. Preliminary and incomplete. I am grateful to Michael Kremer, Greg Lewis, and Ariel Pakes for guidance and encouragement. Thank you to Nathan Eagle for providing access to the data, computing facilities, and helpful conversations. In Rwanda, I thank the staff of my telecom partner and government agencies. This work was supported by the Stanford Institute for Economic Policy Research through the Shultz Fellowship in Economic Policy.

[1] For a sample of 19 countries from which data is available.

[2] Early theoretical work includes Rohlfs (1974), Katz and Shapiro (1985), and Farrell and Saloner (1985). Most empirical work on network goods measures the extent of network effects; see for example Saloner and Shepard (1995), Goolsbee and Klenow (2002), and Tucker (2008). The paper closest in spirit to this one is Ryan and Tucker (2012), which estimates the adoption of a videoconferencing system over a small corporate network, and evaluates policies of seeding adoption.

each connection. Because 99% of accounts are prepaid and the person placing a call pays for it by the second, a subscriber must value a connection at least as much as the cost of calls placed across it. Because calling prices changed over this period, I can estimate the underlying demand curve for communication across each link, and thus the value of each connection.

Let G be the communication graph. Each individual i has a set of contacts $G_i \subset G$, where a directed link $ij \in G$ indicates that i has a potential desire to call j over the mobile phone network. Let S_t be the subset of nodes subscribing in month t. At each period t, individual i can call any contact j that currently subscribes, $j \in G_i \cap S_t$, to receive utility u_{ijt}. Each month, i draws a communication shock ϵ_{ijt} representing a desire to call contact j. Given the shock, i chooses a total duration $d \geq 0$ for that month, solving:

$$u_{ijt} = \max_{d \geq 0} v_{ij}(d, \epsilon_{ijt}) - c_{ijt} d$$

where $v(d, \epsilon)$ represents the benefit of making calls of a total duration of d and c_{ijt} represents the per-second cost.

Second, I model the decision to adopt a mobile phone. The utility of having a phone in a given period is given by the utility of communicating with contacts that have phones: each month i is on the network, he receives expected utility:

$$u_{it} = \sum_{j \in G_i \cap S_t} Eu_{ijt}(p_t, \boldsymbol{\phi}_t) + w \cdot Eu_{jit}(p_t, \boldsymbol{\phi}_t) + \eta_i$$

where u_{ijt} represents calls from i to j (which i pays for), u_{jit} represents calls from j to i (which j pays for), and $w \in \{0, 1\}$ specifies whether recipients value incoming calls. Individual i chooses when to adopt by weighing the discounted stream of these benefits against the declining price of a handset, which is represented by the price index $p_t^{handset}$. Then, i considers the utility of adopting at time τ to be:

$$U_i^\tau = \sum_{t=\tau}^{\infty} \delta^t Eu_{it}(p_t, \boldsymbol{\phi}_t) - \delta^\tau \beta_{price} p_\tau^{handset}$$

I estimate the parameters of this model using maximum likelihood and moment inequalities.

Finally, to evaluate the impact of policies, I use a simulation method that allows each individual to react directly to a policy change, and to each other's responses, capturing effects that ripple through the network and across physical space.

3. Results

I use this approach to evaluate alternate tax policies. At baseline, Rwanda taxed handsets at 48% and airtime at 23%. I find:

Taxing a growing network imposes a substantial welfare cost. I find an average welfare cost of \$2.93 (\$3.14) for each dollar of government revenue raised. This is a higher cost than estimates of marginal cost of public funds from the literature, of 1.21 for sub-Saharan Africa and 1.37 for Rwanda (Auriol and Warlters, 2012), suggesting it would be preferable to use alternative instruments to raise these revenues. In this model telecoms earn no revenue from handset sales, so the entire effect on telecom revenue is driven indirectly, by reduced usage.

Standard, nonnetwork estimates grossly underestimate the welfare costs of handset taxes in a growing network. Network ripple effects account for up to 63% of the effect of handset taxation on telecom revenues. Additionally, ripple effects generate additional government revenue and consumer surplus that would be neglected by a model that only considered individual responses. A naïve estimate that treated phones like standard goods would suggest the average cost of raising a dollar of government revenue from handset taxes would be much lower–\$1.22 in the lower equilibrium and \$1.04 in the upper. Under these estimates handset taxes would have looked more attractive than other tax instruments, as reported by Auriol and Warlters (2012).

The welfare costs of taxation may be heterogeneous over time as the network expands. I consider the effect of taxes during different time periods. This comparison combines two differences—the pace of potential adoption differs at different time periods, but also I observe the dynamic effects of early taxes over a longer time horizon. These effects combine to suggest that welfare costs during the early period of adoption are very high. The welfare cost per dollar raised for taxing handsets from 2007-2009 is \$2.43 per dollar raised (\$1.85). If taxes during that period were lifted, the welfare cost for taxing handsets from 2005-2007 would be \$5.21 (\$13.27). Similarly, the welfare cost for taxing usage, assuming complete passthrough, is \$2.56 (\$2.35) from 2007-2009, and if the tax during that period were lifted, \$4.25 (\$5.16) from 2005-2007.

Usage taxes may impose a similar welfare cost. While handsets are offered by a competitive market so that taxes will be passed through, the telecom may choose whether to pass changes in usage taxes through to consumers. If there is no passthrough, a change in usage tax directly takes revenue from the telecom without distortions. If there is complete passthrough, usage taxes cause distortions similar in magnitude to handset taxes.

Handset taxes impose a high cost on poor users. Although I do not observe household income or consumption, representative survey data suggest that consumers with lower airtime usage have lower

consumption per capita.[3] I show revenues and consumer surplus for the entire sample, and then for the top half of users (above median average daily duration) and bottom half (below). Under the baseline tax regime, the bottom 50% of users account for only 7% of firm revenue and receive only 2% of consumer surplus, but account for 19-20% of government revenue. Since all users must pay the fixed cost of a handset to join the network regardless of usage, poor users end up paying a substantial portion of the tax burden. Eliminating handset taxes would raise the surplus obtained by these consumers by 56%.

Shifting taxes from handsets to usage would have improved welfare of poor phone owners by at least 38%. Earlier results suggest that it would improve welfare to eliminate these taxes and recoup government revenues through other instruments. However, if the government needed to earn a fixed amount of revenue from telecom, it could shift from handset to usage taxes. If the government eliminated handset taxes but raised usage taxes to 30%, it would earn more revenue ($1.79m or $3.61m if usage taxes are passed through; $3.3m or $4.88m if not) and increase the consumer surplus accruing to the bottom half of users by at least 38%, without substantial harm to the top half of consumers (if usage taxes are passed through, their welfare changes by -$0.55m in the low equilibrium or +$0.12m in the high; if not, their welfare increases by at least $18.45m in both equilibria). Doing so would decrease firm revenues, by $4.81m ($3.41m) if usage taxes are passed through and $1.29m ($0.45m) if not. Since potential adopters who are not in my data are likely to be light users, these results likely understate the full distributional impact of a change in policy. The Rwandan government ultimately did lower handset taxes and raise usage taxes in 2010; these estimates suggest there would have been substantial distributional gains to shifting these taxes as early as 2005.

4. Conclusion

This project uses a model to estimate and simulate the adoption of a network good to analyze mobile phone taxation. Mobile phones have been successful among the poor in part because usage charges are primarily marginal; these simulations suggest that governments can encourage adoption among the poor by taxing on the margin of usage rather than adoption. The result could be reversed for technologies like smartphones if there is a wider range of handset qualities available and users upgrade at different frequencies. These results all describe a growing network; optimal taxes are likely to differ once a network is mature.

For more details, find the full paper at http://dan.bjorkegren.com

[3]The 2010 Rwandan EICV suggests that, among households with mobile phones, the households with the top half of airtime expenditure have 3.7 times more overall consumption than those in the bottom half.

References

Auriol, E. and M. Warlters (2012): "The marginal cost of public funds and tax reform in Africa," *Journal of Development Economics*, 97, 58–72.

Bjorkegren, D. (2015): "The Adoption of Network Goods: Evidence from the Spread of Mobile Phones in Rwanda," .

Farrell, J. and G. Saloner (1985): "Standardization, Compatibility, and Innovation," *The RAND Journal of Economics*, 16, 70–83.

Goolsbee, A. and P. J. Klenow (2002): "Evidence on Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 45, 317–343.

Gordon, R. and W. Li (2009): "Tax structures in developing countries: Many puzzles and a possible explanation," *Journal of Public Economics*, 93, 855–866.

GSMA (2012): "Taxation and the Growth of Mobile Services in Sub-Saharan Africa," .

Katz, M. L. and C. Shapiro (1985): "Network Externalities, Competition, and Compatibility," *The American Economic Review*, 75, 424–440.

Rohlfs, J. (1974): "A Theory of Interdependent Demand for a Communications Service," *The Bell Journal of Economics and Management Science*, 5, 16–37.

Ryan, S. P. and C. Tucker (2012): "Heterogeneity and the dynamics of technology adoption," *Quantitative Marketing and Economics*, 10, 63–109.

Saloner, G. and A. Shepard (1995): "Adoption of Technologies with Network Effects: An Empirical Examination of the Adoption of Automated Teller Machines," *The RAND Journal of Economics*, 26, 479–501.

Tucker, C. (2008): "Identifying Formal and Informal Influence in Technology Adoption with Network Externalities," *Management Science*, 54, 2024–2038.

PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets

Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, Herve Rivano, Razvan Stanica
University of Lyon, LIRIS, CNRS, INSA-Lyon, UMR5205, EVS, UMR5600, INRIA, CITI, ENSEEIHT, ENTPE, France
{sonia.benmokhtar, antoine.boutet, olivier.brette, lionel.brunie, stephane.dalu, mathieu.cunche, razvan.stanica, vincent.primault}
@insa-lyon.fr ; patrice.raveneau@univ-lr.fr ; herve.rivano@inria.fr ;
{louafi.bouzouina, patrick.bonnel}@entpe.fr

***Abstract*—The wide adoption of mobile devices has created unprecedented opportunities to collect mobility traces and make them available for the research community to conduct interdisciplinary research. However, mobility traces available in the public domain are usually restricted to traces resulting from a single sensor (e.g., either GPS, GSM or WiFi). In this paper, we present the PRIVA'MOV dataset, a novel dataset collected in the city of Lyon, France on which user mobility has been collected using multiple sensors. More precisely, this dataset contains mobility traces of about 100 persons including university students, staff and their family members over 15 months collected through the GPS, WiFi, GSM, and accelerometer sensors. We provide in this paper both a quantitative and a preliminary qualitative analysis of this dataset. Specifically, we report the number of visited points of interests, GSM antennas and WiFi hotspots and their distribution across the various users. We finally analyse the uniqueness of human mobility by considering the various sensors.**

I. Introduction

The large adoption of mobile devices combined to their embedded localisation capabilities opens novel opportunities to provide mobility traces to the research community at large. Classical usages of mobility datasets include the analysis of user mobility patterns and their regularity [1], discovering hot places in a city [2] or studying privacy threats due to the disclosure of mobility data [3].

However most of the available datasets (e.g., the Cabspotting [4], the Geolife [5], or the T-Drive [6] datasets to named a few) contain data collections coming from only one sensor (i.e., the GPS, GSM or WiFi sensor). While single sensor mobility traces have been widely used in the literature to answer classical research questions, the availability of mobility data coming from multiples sensors opens the door for richer studies. This include comparative studies (where the data provided by a single sensor is compared to the data provided by another sensor) and compositional studies (where the data provided by a given sensor complements the data provided by another sensor). For instance, one may perform a comparative study of personal data leakage due to the sharing of the data provided by a given type of sensor versus the one due to the sharing of another type of sensor. On the other hand, one may combine the data provided by multiple sensors to increase the precision of user mobility data. For instance, one may use the WiFi and GSM mobility data to remove erroneous localisations provided by the GPS sensor.

We present in this paper the PRIVA'MOV dataset that contains the mobility traces of 100 users around the city of Lyon collected using various sensors, namely the GPS, WiFi, GSM, and accelerometer sensors. This dataset has been collected from October 2014 to January 2016 by equipping volunteer students from three universities, staff members and sometimes their relatives with smartphones on which a crowdsensing application was periodically collecting records from the above mentioned sensors.

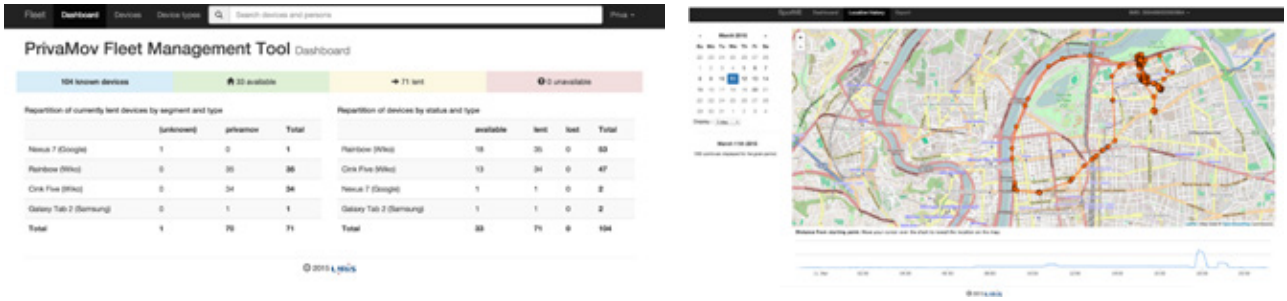
In the remaining of this paper we present the data collection process (Section II), a quantitative analysis of the resulting dataset (Section III) as well as a preliminary qualitative analysis of its records (Section IV). More precisely, we analyse the relation between meaningful locations of the city and users and the uniqueness of mobility traces of users considering the various sensors. We finally draw our conclusions and future research directions (Section V).

II. The PRIVA'MOV CrowdSensing Campaign

The PRIVA'MOV dataset has been collected during a crowdsensing campaign that took place in the city of Lyon from October 2014 to January 2016 in the context of PRIVA'MOV project funded by the LABEX IMU[1] funding agency. In the context of this project 100 smartphones (52 Wiko Rainbow and 48 Wiko Cink 5) have been equipped with a crowdsensing application and distributed to student, staff members and their relatives from three universities: INSA Lyon, ENS Lyon and Université Claude Bernard, Lyon 1. Volunteers were asked to use the PRIVA'MOV phone as their primary phone and to carry it during their daily activities. The crowdsensing application was a modified version of the application developed in the Funf project [2]. A complementary fleet management web application and a trace visualisation tool depicted in Figure 1a and Figure 1b, respectively, have also been developed.

The crowdsensing application has been configured to collect the data every time the system used them (e.g., change of location, new WiFi scan). In order to save battery, the collected data were uploaded to the server only when the smartphone was connected to a WiFi network. The resulting dataset is described in the following section.

[1]LABEX IMU: http://imu.universite-lyon.fr
[2]Funf: http://funf.org



(a) Fleet management portal (b) Visualisation tool

Fig. 1. PRIVA'MOV Fleet management portal (Figure 1a) ; PRIVA'MOV trace visualisation tool (Figure 1b).

III. PRIVA'MOV QUANTITATIVE ANALYSIS

We describe in this section a quantitative analysis of the PRIVA'MOV dataset. Table I shows the number of records collected by the various sensors in the overall dataset.

Sensor	Number of Records
WiFi	25,655,480
Cellular	8,076,512
GPS	156,041,576
Accelerometer	90,066,831
Battery	7,008,504

TABLE I. THE PRIVA'MOV DATASET CONTAINS MOBILITY DATA COLLECTIONS CAPTURED THROUGH DIFFERENT SENSORS.

From these data collections, it is then possible to extract mobility traces. We call a mobility trace a list of spatio-temporal points belonging to a given user. This can be done in the WiFi and GSM data collections by associating a GPS location to each WiFi and GSM antenna by relying on public datasets such as WiGLE [3] or Google [4].

In the PRIVA'MOV dataset we did not perform the mapping of WiFi and GSM antennas to GPS locations. Instead, we use the unique identifier of the cellular antenna to which the user is connected (respectively the MAC address of the WiFi access point to which the user is attached, or discovered through a periodic WiFi scan) as a spacial indicator of the user location.

In addition, mobility traces built from the GPS data collections could be enriched with user Points of Interest (POIs). A POI is a meaningful location where the user has marked a significant stop. To compute POIs, we used a methodology similar to [7]. The idea behind this method is to identify restricted areas where users stay more than a specific duration. More precisely, POIs can be extracted using a simple spatio-temporal clustering algorithm parametrised with a maximum POI diameter d and a minimum stay time t. This POIs extraction is done in two clustering steps, the first one identifies POIs for each user and the second one assigns identifiers to unique POIs, thus allowing to identify POIs shared by several users. For instance, Figure 2 illustrates the POIs of users in the Lyon sub-area for a diameter of 250 meters ($d = 250$) and a stay time of 30 minutes ($t = 30$).

[3]WiGLE: Wireless Network Mapping, http://wigle.net

[4]Google Maps Geolocation API: https://developers.google.com/maps

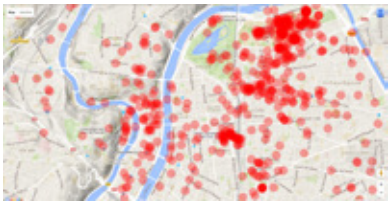
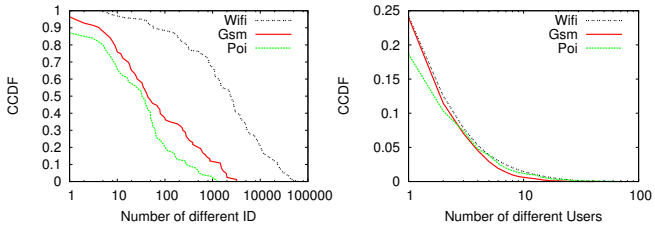


Fig. 2. Location of Points Of Interests (POIs) across the Lyon area.

In the PRIVA'MOV dataset, we extracted from the GPS data collections the set of all POIs and assigned an identifier to each unique POI inside the whole dataset. Then, similarly to the GSM and WiFi datasets we used unique POI identifiers as a spatial component.

Let us consider the extracted mobility traces as explained above (i.e., using cellular antenna IDs, WiFi mac addresses, and POI IDs from the data collections). To analyse the resulting mobility traces, Figure 3 shows the Complementary Cumulative Distribution Function (CCDF, defined as $P(X > x)$). Figures 3a shows the number of unique GSM antennas, WiFi access points and POIs per user. Figures 3b depicts the number of unique users identified per GSM antenna, WiFi access point and POI. These tail distributions show that most of GSM antennas, WiFi access points and POIs have been visited only by one user. For instance, 75% of the WiFi access points have been seen by only one user. Conversely, mobility traces of most of the users are composed of several GSM antennas, WiFi access points and POIs. Finally, these figures show that users have discovered more WiFi access points than GSM antennas, and the number of POIs is lower than the two others.



(a) # of Identifiers (b) # of Users

Fig. 3. The tail distributions of the number of different GSM antennas, WiFi access points, and POI per user (Figure 3a) ; and the number of unique users per GSM antenna, WiFi access point, and POI (Figure 3b).

IV. PRIVA'MOV QUALITATIVE ANALYSIS

In this section, we report two qualitative experiments conducted on the PRIVA'MOV dataset. Specifically, we first analyse the spatial relationship between POIs inferred from the activity of users. Then we study the uniqueness of user mobility traces over the three data types.

A. POI relationship inference

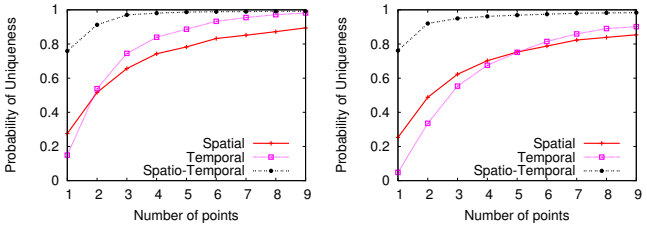
In this experiment, we analysed the relationship between POIs appearing in the dataset. Specifically, we consider that two POIs are related if there exist at least one user that visited the two POIs. Figure 4 shows the resulting graph obtained by extracting the relationship between all the POIs of the PRIVA'MOV dataset. From this figure, we observe that while some POIs have been widely visited by many users (at the bottom left), some places have been only visited by sub groups of users (clusters of point at the top right).

Fig. 4. Relationship between points of interest inferred from the mobility of users (two points of interest are connected if at least one user has visited these places).

B. Uniqueness of human mobility

To quantify the uniqueness of mobility traces, we use the methodology proposed by De Montjoye and all in [8]. More precisely, for each mobility trace T, we evaluate the uniqueness of a given sub-trace I_p of p randomly chosen spatio-temporal points. A sub-trace I_p is said to be unique if only one user has $I_p \in T$. To measure this uniqueness, we performed a brute force search of users who have the p points composing I_p in their mobility trace T. The size of this set of users sharing the same I_p, noted k, characterizes the uniqueness of the sub-trace I_p. If $k = 1$, the sub-trace is unique. The uniqueness of traces is estimated as the percentage of 2500 random sub-traces that are unique given the p points composing them. We use the same methodology to evaluate the uniqueness of spatial or temporal only mobility traces. In this case, the sub-trace I_p contains spatial or temporal points, respectively.

We report the uniqueness of mobility traces built from the GPS and the WiFi traces. Figure 5 depicts the probability to be unique according to the number of points in the considered sub-trace. The evaluation reports results for spatial, temporal, and spatio-temporal traces. As shown in these figures, the



(a) GSM (b) WiFi

Fig. 5. Four spatio-temporal points are enough to uniquely identify 97% of the individuals.

results for spatio-temporal mobility traces from GSM and WiFi show a strong uniqueness. More precisely, four spatio-temporal points are enough to identify on average 97% of the users. This high uniqueness is the result of combining the temporal and the spatial mobility information of users, which are discriminative enough to uniquely identify them. Although this analysis uses a smaller set of data, this result comforts and generalizes the previous study performed in [8] on call logs.

V. CONCLUSIONS

To address the lack of multi-sensor mobility datasets, the project PRIVA'MOV developed and deployed a crowdsensing platform to collect mobility traces from a sample of real users equipped with mobile phones. We presented the resulting dataset and reported quantitative and qualitative experiments over this dataset, which show the potential of using the latter for answering novel research questions. The PRIVA'MOV datasets is available for research community upon request and under a set of usage conditions. Our future work targets the data collection of multi-sensor information over a larger population of users with a real time data analysis and a validation tool to ask the users to validate or not the inferred information (i.e., leaked personal information, mobility behavior).

REFERENCES

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[2] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *KDD*, 2012, pp. 186–194.

[3] S. Gambs, M.-O. Killijian, and M. N. n. del Prado Cortez, "Show me how you move and i will tell you who you are," in *SPRINGL*, New York, NY, USA, 2010, pp. 34–41.

[4] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset epfl/mobility (v. 2009-02-24)," Downloaded from http://crawdad.org/epfl/mobility/20090224, Feb. 2009.

[5] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining Interesting Locations and Travel Sequences from GPS Trajectories," in *WWW*. ACM, 2009, pp. 791–800.

[6] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with Knowledge from the Physical World," in *KDD*, 2011, pp. 316–324.

[7] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in *WWW*, 2008, pp. 247–256.

[8] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, 2013.

Developing a mobility monitoring application hand in hand with the end user

Jérôme Urbain, Nicolas Snel
Dalberg Data Insights; {jerome.urbain, nicolas.snel}@dalberg.com

1. Introduction

Nowadays over 50% of the world's population is living in urban areas. This proportion is even expected to increase to 66% over the coming 35 years [1]. Consequently, it is urgent to act on mobility issues all around the globe.

It is however difficult, particularly in developing regions, to gather representative data of population's movements in order to plan road infrastructure, set up public transports or react to abnormal events [2]. Generally, the only data available come from surveys conducted every 5 or 10 years. Alternatives are hence explored. Among them, Call Detail Records (CDRs) and data connection logs collected by telecom operators offer opportunities to systematically track the movements of a significant proportion of the population.

Mobility studies from Telecom Operators' data have largely increased in the past years. Insightful results have been published (e.g., [3]–[6]). Nevertheless, these analyses are frequently conducted a) over a limited time frame (the period of the data shared by the Mobile Operator) and b) independently of their potential end-users – the focus being on improving technical methodology rather than ensuring wide usage of the algorithms. Here, we take another approach. We have been mandated to develop mobility applications that are used by the end user on a day-to-day basis. Namely, our end user is the Kampala Capital City Authority (KCCA), which is responsible for the development and management of Uganda's capital. Hence it is critical for us to understand their work, their pains and how we can improve their situation. We will present here the process that was followed, focusing on the collaborative design methodology.

KCCA was involved from scratch in the development of our applications. We first discussed an initial idea, then drafted a mock-up of each application to validate it with them. Once the look and feel had been validated, we started developing the application (both the back-end computations and the data visualization) in an incremental way, continuously asking for KCCA's feedback through the process. That agile methodology enabled us to stay close to KCCA's needs and to build together the next features.

Two key elements have been identified to drive daily application usage. First, the application must be convenient to use: it must be straightforward to understand the data displayed. This puts a tremendous importance on data visualization. Second, the application must display the latest data available: it is important the user accesses every day data that reflects the current situation, so (s)he can take actions accordingly and with confidence that those will bring the expected impact. It obviously affects the way data is computed: for instance, all daily computations must run in maximum 24 hours to ensure we stay up-to-date.

Following our discussions with KCCA, we have started the development of 2 applications: 1) detailed Origin-Destination maps; 2) visualization of traffic evolution throughout a typical day. For both applications, our input data consists in the location (cell tower) of subscribers. These are extracted from CDRs and data logs provided daily by Airtel Uganda, the second largest operator with over 8 million customers (~43% market share).

2. Detailed Origin-Destination

2.a Back-end computations

The detailed Origin-Destination (O-D) application enables to identify regular commutes, which is essential to design long-term mobility master plans. In agreement with KCCA, we have distinguished 4 periods of the day (instead of the usual home and work hours): night (9PM to 5AM), morning (5AM to 11AM), midday (11AM to 4PM) and evening (4PM to 9PM). The most used site of each subscriber is computed at the beginning and end of each of these periods over the last 30 days. The O-D matrices during each period are inferred from the corresponding most used sites. These numbers will be validated in the coming weeks by comparing them to survey results that will be available soon. Besides O-D matrices, we also estimate the typical travel time between each O-D pair and compute the evolution of the flow of people between one O-D pair over the course of a day. The methodology here is to look at trips performed between stable locations (the person remained at least 60 minutes at the origin and destination sites) and to count one person in movement between the last time seen at departure and the first time seen at arrival.

2.b Visualization

As shown on Figure 1, flows of people are displayed on an interactive colored map that allows users to quickly identify areas of interest. Clicking on one area (or a combination of areas) filters the whole application with respect to that area. To further ease the identification of important trips, flows are also displayed via lists and matrices. Users can control the geographical level of aggregation they would like to see (here sub-county or parish), whether they want to see the suburbs or only the city center, the period of the day and the type of day (business day, Saturday or Sunday). The distribution over time of the flow of people between the selected O-D pair is displayed below (see Figure 2). This enables to identify the peak times in the morning and in the evening.

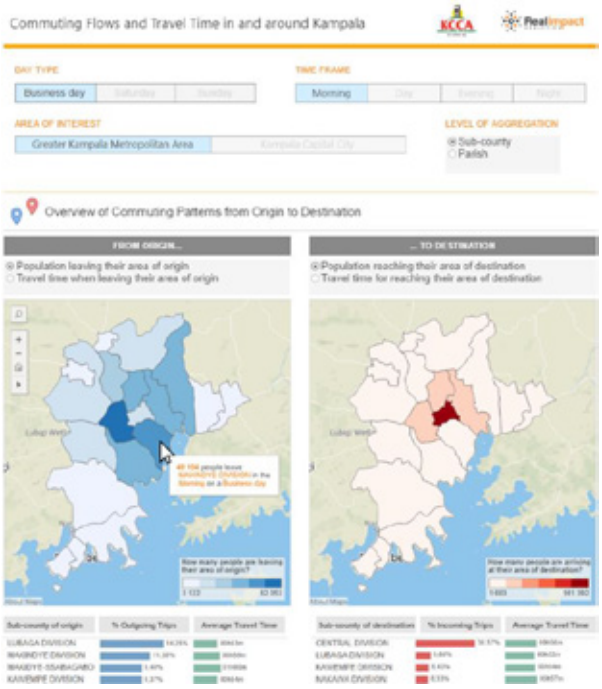


Figure 1: Detailed Origin-Destination flows

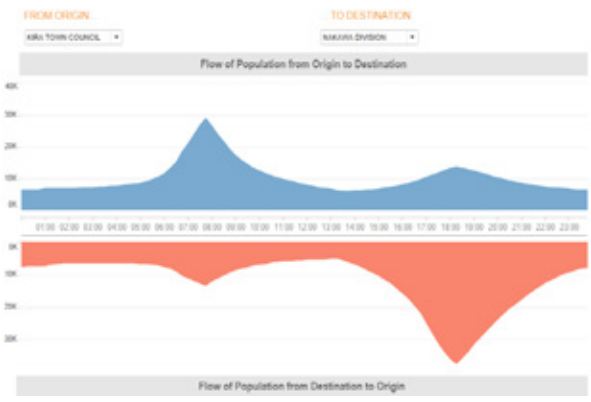


Figure 2: Flow distribution over time

3. Traffic evolution

3.a Back-end computations

It is particularly important for KCCA to visualize the evolution of traffic in the city throughout a typical day. This allows for short-term corrective measures (e.g., regulating the traffic at one junction or diverting traffic from one road). To compute traffic evolution, the following process has been implemented. Every 15 minutes, the last known position of all subscribers is updated. When there is no new location, people are assumed to have stayed in the last location we have seen them. Then, traffic is estimated: a subscriber S is considered incoming to an area X if X differs from her/his previous location (Y). Symmetrically, S will be considered outgoing from Y in that case. A subscriber is considered in traffic internally to an area X if (s)he was incoming to X during the last Q quarters of hour (Q can be customized and is set to 2 by default) and has had no other position since then. The 15-minutes daily traffic values can be aggregated over N days to obtain the average traffic.

3.b Visualization

Traffic is displayed on a colored map (see Figure 3). Users can decide to focus on incoming, outgoing or external traffic and select the type of day. It is possible to select a specific time for analysis or to play the evolution of traffic 15 minutes by 15 minutes. Following a demand from KCCA, traffic is displayed on roads rather than cell tower Voronois. Currently we uniformly distribute the traffic value among the main roads of the Voronoi.

Figure 3: City traffic evolution

4. Future work

Our approach differs from usual scientific research. We are developing our application in close collaboration with its end user. This is bringing initial focus on data visualization. Yet, after successfully iterating on what KCCA would like our applications to display and populating those applications with initial data, we can identify shortcomings of the implemented methodologies and improve them. We are listing future work below, and will extend that list after validating the O-D matrices thanks to survey data we are about to receive.

First, we want to address instabilities in travel time estimations, likely due to lack of user activities at certain moments (e.g., night). Peak detection could be implemented, and unreliable estimates could be corrected considering the number of active people at that time.

Second, KCCA and other potential end users request other geographical groups than what we naturally obtain from Telecom Operators. Areas covered by cell towers do not mean anything to many end users. Sometimes, even the administrative levels we use for aggregation do not correspond to operating areas of the end users. As already stated, KCAA has asked us to display information on road maps. We aim at improving the current uniform road repartition with a more advanced method to better estimate the actual roads taken by subscribers (e.g., [7]).

Finally, considering KCCA's latest feedback, we will complement the current applications with an impact assessment application. The objective is to allow for traffic comparisons between two periods, understand the influence of events (elections, concert, …) and hence better predict the impact of future events. As a first step we will flag abnormal events, following the methodology described in [6]; in a second stage we will automatically segment long periods of traffic (e.g., 60 days), inspired for instance from audio segmentation algorithms ([8]–[10]). That will allow the end users to visualize significant changes in traffic and relate them to events they are aware of.

Bibliography

[1] United Nations, "World's population increasingly urban with more than half living in urban areas," 2014. [Online]. Available: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html. [Accessed: 27-Jan-2017].

[2] S. Lokanathan, G. E. Kreindler, N. H. N. De Silva, Y. Miyauchi, D. Dhananjaya, and R. Samarajiva, "The Potential of Mobile Network Big Data as a Tool in Colombo ' s Transportation and Urban Planning," *Inf. Technol. Int. Dev. [Special Issue]*, vol. 12, no. 2, pp. 63–73, 2016.

[3] A. J. Tatem, Z. Huang, C. Narib, U. Kumar, D. Kandula, D. K. Pindolia, D. L. Smith, J. M. Cohen, B. Graupe, P. Uusiku, and C. Lourenço, "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning.," *Malar. J.*, vol. 13, no. 1, p. 52, 2014.

[4] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the impact of human mobility on malaria," vol. 338, no. 6104, pp. 267–270, 2013.

[5] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee, "Impact of human mobility on the emergence of dengue epidemics in Pakistan," *Proc. Natl. Acad. Sci.*, vol. 112, no. 38, pp. 11887–11892, 2015.

[6] Z. Zhou and C. Volinsky, "Quantifying urban traffic anomalies," in *Bloomberg Data for Good Exchange Conference*, 2016.

[7] M. Chen, "A study of transportation network design problems," University of Manitoba,Winnipeg, Canada, 1991.

[8] S. S. Cheng, H. M. Wang, and H. C. Fu, "BIC-based audio segmentation by divide-and-conquer," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4841–4844.

[9] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, 2000.

[10] M. Cettolo and M. Vescovi, "Efficient Audio Segmentation Algorithms Based on the BIC," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 537–540, 2003.

Effects of Network Architecture on Model Performance when Predicting Churn in Telco

María Óskarsdóttir*, Cristián Bravo†, Wouter Verbeke‡, Bart Baesens*† and Jan Vanthienen*
*Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
†Southampton Business School, University of Southampton, United Kingdom
‡Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium

***Abstract*—Research on social network analytics (SNA) has advanced significantly in recent years. Before networks can be analyzed they need to be built from the available data, which is a non-trivial task. In addition, the results of the analysis can depend on the network architecture itself. In this study, we demonstrate this effect in the case of customer churn in telco. By building numerous networks using different definitions of weights and edges and applying a relational classifier to predict churn, we compare the performance of the resulting models. Our results imply that the network architecture does in fact have a great effect on the performance of such models.**

I. Introduction

Research on social network analytics (SNA) has advanced significantly in recent years. With applications in both social sciences and business communities, state-of-the-art techniques from graph theory and machine learning are applied to a variation of social networks that need to be defined given the available relational data, which can be both challenging and time consuming.

In the case of SNA in the telecommunication industry, the data source is usually call detail records (CDR). These phone logs have been a great source for research in various disciplines over the past years, as an extensive overview of the current literature in large-scale mobile traffic analysis shows [1]. When building networks from CDR data, they need to be filtered and aggregated in an intelligent way in order to represent the activity and interactions of customers accurately. The many possibilities of defining edges and weights together with other factors, need to be taken into consideration when the network is built. The impact of the architecture of networks on the techniques which are applied to them, and the resulting findings have been scarcely studied to the best of our knowledge.

In this study, we explore this potential effect in the case of customer churn in telco, by building more than five hundred networks with varying definitions of edges and weights and aggregating them in numerous ways. The number of networks can increase very rapidly when taking all possibilities into account, and as a result, it becomes computationally infeasible to compare all model combinations. Instead, we will exploit the result of our previous benchmarking study for 'algorithm selection' [2] and use the network only link based classifier [3] as a proxy method to predict churn in the different types of networks and compare the performance of the resulting models. As a result, we hope to guide further studies on how to optimize network architecture when predicting churn in telco.

TABLE I
Segmentation of Networks

Day	Part of Week	Time of Day	Combination Part of Week	Combination Time of Day
Monday Tuesday Wednesday Thursday Friday Saturday Sunday	Working days Weekend	Day Evening Night	$\frac{1}{2}$WD+WE WD+$\frac{1}{2}$WE $\frac{1}{3}$WD+WE WD+$\frac{1}{3}$WE	$\frac{1}{2}$Day+Evening Day+$\frac{1}{2}$Evening $\frac{1}{3}$Day+Evening Day+$\frac{1}{3}$Evening

II. Defining the Network Architecture

The dataset used in this study are call detail records origination from a telecommunication provider in Belgium. The dataset spans six months of phone logs between 1.2 million customers with postpaid contracts and churn rate of under a percent.

When building the networks, phonecalls lasting less than a few seconds were removed, as is common in the literature [4]. The edges of the network were defined in three ways: incoming, representing all phonecalls made to a customer; outgoing, the phone calls made by a customer; and undirectional, when the previous distinction is not made. The weights of the edges were defined in numerous ways. Firstly, we consider binary weights, with an indicator of whether a phone call was made or not. Secondly, length or the total duration of all phonecalls during the period were aggregated. Finally, since the CDR data had information about the date and time of each phone call it was used to segment the data, thus building separate networks for phone calls made on each day of the week, during working days (WD) and weekends (WE) and during the day, evening and night. In addition, parts of the day and parts of the week were combined in various ways to build more networks. Table I shows a summary of all the networks that were built in this way, using duration of phonecalls. Additionally, since some studies [5], [6] remove connections that are not reciprocal before applying SNA, we included this possibility in our exploration, by repeating the whole setup with non-reciprocal calls removed. In total over five hundred networks were built for the one dataset.

All networks were built using the same month of data and the proxy classifier subsequently applied to predict churn in

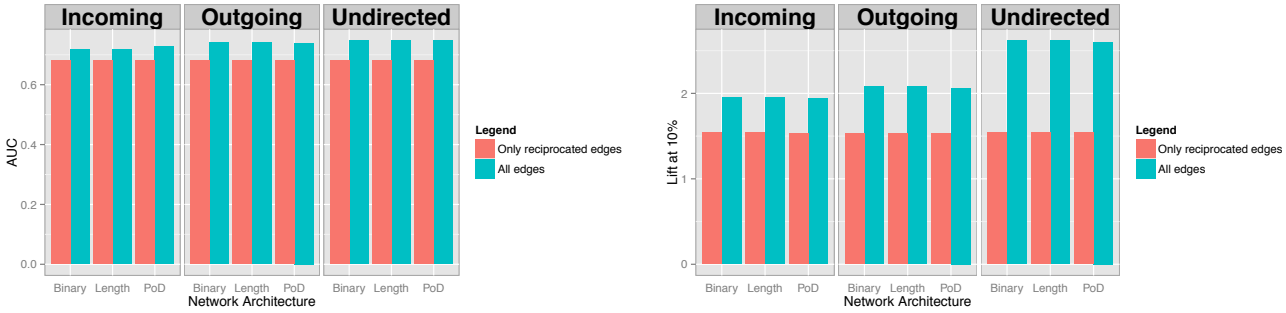


Fig. 1. The figure shows the performance of the classifier when applied to the different networks.

the following month. Finally, the predictions were compared to actual churn labels.

III. Results of Network Construction

Figure 1 shows a comparison of the performance between the various networks measured in AUC and lift at 10%. The two measures show more or less consistent results.

For comprehensibility, we only show three representation of weights in the networks, namely binary weights, total duration of phonecalls and total duration of phonecalls weighted by the best combination of time of day. The best combination of time of day is the first one in table I which is $\frac{1}{2}\text{Day} + \text{Evening}$.

The first observation we make, is the performance of networks with and without non-reciprocated edges, as is evident by the difference between the green and red bars in the figures. The clear difference between these results is that by removing the non-reciprocal connections, the performance has decreased significantly.

A second interesting behavior we can extract, is the difference between performance on undirected, outgoing, and incoming networks. It is clear the undirected outperforms the other two. However, it is not clear whether the outgoing or the incoming network is better. This result is contradictory to the findings of [5], where the outgoing edges were shown to have higher correlation with churn.

When comparing the definition of the weights we see that sometimes the performance on the whole network tends to be slightly better than on the time of day combination network. There is a slight increase in predictive capability by segmenting the network by part-of-day, with evenings more important than daytime. These results are consistent with the conclusion from our previous results: calling the close circle of the user (family and close friends) tends to be the most relevant factor when spreading churn influence [2]. These calls are bound to occur with greater chance during the evenings (after work). Regarding the weights of the edges, the performance on the network with binary weights is almost consistently better or at least as good as on the networks with length. This indicated that the simplest variant, binary weights, will result in models that are just as good. This behavior hints at the importance of the existence of connectivity rather than the intensity of the communication between two users. As mentioned before, it is the close circle of the user that are more at risk of churn when the user does, so the length of the call would be a poor proxy of the intensity of the relationship and as such binary weights that reflect whether there is a connection or not, would be sufficient to represent this relationship.

IV. Conclusion

When defining network architecture with CDR data many decisions need to be made. As our results suggest, these decisions can have a great effect on the performance of models which are applied to the networks. We have shown that a significant increase in performance can be achieved when using a best performing network, which, as our results imply, is one that is constructed with binary weights and undirected edges. As a result, the modeller is responsible for correctly defining the best network for the problem that is being tackled, as failure to do so will result in less predictive capability and therefore less potential gains. In a future study we would like to take into account more datasets and other classifiers, to test the statistical significance of these observed differences in performance.

References

[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2015.

[2] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "A comparative study of social network classifiers for predicting churn in the telecommunication industry," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 1151–1158.

[3] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, vol. 3, 2003, pp. 496–503.

[4] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.

[5] M. Haenlein, "Social interactions in customer churn decisions: The impact of relationship directionality," *International Journal of Research in Marketing*, vol. 30, no. 3, pp. 236–248, 2013.

[6] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, 2008, pp. 668–677.

A Neural Network Framework for Next Place Prediction

Chad Langford[1]
chad.langford@sbg.ac.at

Euro Beinat[1,2]
euro@beinat.net

[1]Department of Geoinformatics – Z_GIS, University of Salzburg, Austria
[2]CS Research Foundation, Amsterdam, The Netherlands

1. Introduction

The goal of this paper is to create a framework for forecasting human mobility that utilizes the unique architecture found within long short-term memory (LSTM) neural networks. This particular neural network architecture is utilized due to its strength of being able to learn the importance of long-term dependencies, which within the lens of human mobility prediction means learning what the next most likely place will be given the temporal context of locations observed beyond the recent past. Using a dataset of call detail records (CDR) from a European country for roamers, issues of data structure, sampling, and quality are considered before determining the temporal extent to which past locations provide information that is able to improve the accuracy of next place prediction. The results will show that prediction accuracy grows with the window of historical context used, until a point at which the importance of the long-term locational dependencies diminishes.

2. Data

The problem of forecasting is fundamentally one of classification. The neural network learns to classify a past trajectory as the most likely next location. How the network learns is partially based on how the training data is structured and fed into the model, a focus of this paper. Thus, when constructing a LSTM model, careful consideration must be given to the structure, sampling method, and quality of the data.

2.1 Data structure

The data is transformed from a CDR format to a list of trips. Each trip is defined as consecutive hourly data for a user, without a break in data availability. Each cellular tower is encoded as a unique Unicode symbol, which are used to create a string of characters that represent a trip. An example trip may look like: AAABCAADAAAAEFF, where each character represents a cellular tower the user was connected to for the most time during that particular hour. By constructing the data in this format, we facilitate the LSTM network's ability to learn long-term dependencies between locations across many trips.

2.2 Data sampling

We choose to predict for the 1000 longest trips within the data. For each of these test traces, the first 80% of the trip (considered the history) is used to sample the appropriate data, while the latter 20% of the trip (considered the future) is predicted for. The sampling method used is a hierarchical geographical intersection. For example, if we are interested in using a three-hour window

of past locations to predict the next hour's location, then the historic data (the first 80%) is broken into trip segments of length 3+1. This is necessary to ensure that we sample data that includes enough information temporally to use a three-hour window of past locations to predict the next location. The hierarchical geographical intersection works by finding training traces that have a common trip segment with the historic test trace at various temporal scales.

Test trace: AAABCAADAAAAEFF.

The underlined portion of the trace makes up the first 80%, which is considered the user's history. The portion that is not underlined is predicted for. If we are interested in using a temporal window of three hours to predict for the next location, then the history would be broken into the following trip segments of length four: [AAAB], [AABC], [ABCA], [BCAA], [CAAD], [AADA], [ADAA], [DAAA], and [AAAA]. All training trips, those that are not part of the 1000 longest trips, are searched through to find traces that contain at least one of the trip segments. Training trips that share a common trip segment are added to a list of trips that will be used for training the LSTM. If any trip segments return zero trips with a matching segment, it is further broken down into smaller trip segments before another search is performed. If the trip segment [AABC] was not found anywhere else within the training data, the segment would be made into two segments [AAB] and [ABC] that are shorter by one hour, and the search would be performed again. This hierarchical sampling is done to ensure that the LSTM has as much information as possible to learn from, even if it does not exactly match the temporal scale we are interested in. The hierarchical downsizing of a path segment will continue as long as the path segment is not represented in the training data, until only a single point is sampled for.

2.3 Data quality

The LSTM is robust against the most common quality flaw found within cellular telecommunications data: antenna jumping. The model will not be negatively impacted by apparent shifts in location that occur when a cellular tower's capacity is saturated. This often appears within a trajectory as a sudden change of position across a large distance. The model will give less weight to these changes in location since these geographical shifts are irregular. The structure of the data also assists in allowing the model to differentiate legitimate transitions between locations and those due to the cellular network's protocol for handling capacity saturation. Since the data is structured into trips, rather than into 24 hour blocks for each day, the LSTM will not learn these locational shifts—even when they're more common at certain times of the day.

There is a large number of trips that are not used within the current framework as a direct result of the historical window size used. If we are interested in using the last three hours to predict the next location, then trips that are smaller than the largest path segment cannot be used. This is because if a trip has only two locations, it is not possible to use the past three hours to determine the last location. More than 50% of the available data consists of trips with a length of two. While it is important to note this, it does not necessarily have a negative impact on the results. The overall volume of data and range of complexity for longer trips ensures that the information found within the shortest trips is not lost when only using longer trips.

3. Model

The training trips gathered by the hierarchical geographical sampling method are fed into the LSTM in batches of segment paths. The weights within the neural network are updated after each batch of trip segments. The experiment will be run for historical windows of various sizes (1-10) In order to determine when the importance of the long-term locational dependencies diminishes. It is expected that the larger historical window sizes will predict the next location better than a smaller historical window size. The maximum window size used must be balanced with the amount of data available, since larger window sizes will inherently have less training data available. After training, each unique combination of locations will have its own probability distribution for the next likely location. While it is most likely that the location with the highest probability is chosen as the next location, there is some chance (depending on the probability distribution) that a given unique historical context will not always produce the same location as a result.

4. Results

While the experiment is not yet complete, preliminary results have favorably shown that it is likely that the long-term dependence of sequential locations is able to improve prediction accuracy. This experiment focuses explicitly on the historical window size used and the subsequent prediction accuracy of the model, while other important factors such as data structure, sampling, and quality are controlled for in a logical manner. This experiment is not designed to explore the many variables surrounding the LSTM model architecture. The information gained through the results in this experiment will be generalized, and will require further investigation to determine if the results hold true for a range of spatial and temporal aggregations of the mobility information.

Acknowledgements:

This work was supported by the Austrian Science Fund (FWF) through the Doctoral College GIScience (DK W1237-N23), Department of Geoinformatics - Z_GIS, University of Salzburg, Austria. Additionally, this experiment would not have been possible without the support of the CS Research Foundation, Amsterdam, The Netherlands (www.collectivesensing.org) who provided the anonymized mobility data and related support.

Explorative analyse of two Italian cities: Turin and Venice for diversity

Didem Gundogdu,[1,2], Matteo Moretti,[3] and Bruno Lepri,[1*]
gundogdu@fbk.eu, MatteoMaria.Moretti@unibz.it, lepri@fbk.eu
[1]Fondazione Bruno Kessler, Trento, Italy
[2]DISI, University of Trento, Italy,
[3]Faculty of Design and Arts of the Free University of Bozen-Bolzano, Italy

1. ABSTRACT

In our paper we explore two different Italian cities for the diversity in population. Turin is the old capital city of Italy, drives Italian automotive industry and hosts the well known university. Hence Venice is one of the most visited touristic cities in the world. Beside from commonly used census data, we analysed mobile phone datasets to examine the distribution of different communities among the city. The presence of immigrants and tourists in various locations during day and night would shape the heterogeneity of a city's ethnic and cultural dynamics. The diversity and integrity of different communities may resolve in healthy economical growth for cities. In that paper we try to answer if we can foresee the diversity of a city under the light of mobile phone usage.

2. INTRODUCTION

Contemporary networked cities are "constituted by flows of people, vehicles, and information" [17], and yet data about these flows are increasingly difficult to collect and analyse using traditional social science research methods [18]. A possible help come from the track and analysis of the mobile phone communications, that allows the gathering of several further information that open the door on a new paradigm of understandings about the citizens and their behaviours across time and space [3, 4]. The applicability of these methods can provide urban planners, municipalities and public institution with useful information on how different groups behave in the city. Knowing who populate different part of the city at different times, support and improve the awareness during the policy making processes, and consequently the well-being of the city and its citizens as well. From that perspective, understanding the mobility behaviour of minority and immigrant groups in society, plays a crucial role as agent of change to enable the adoption of targeted policies. Latter it might helps the prevention of urban segregation phenomena.

According to Lee, migration is defined as a permanent or semipermanent change of residence [11]. Hence, we use word *immigrant* as, a person who changed his origin country and move to destination country and declare his presence to legal authorities.

In this paper we explore two months of spatio-temporal activity patterns from two cities in Italy, Turin and Venice, with data collected from the Telecom Italia Mobile (TIM) phone network. A preliminary assumption led our work, such as the country code of the incoming or outgoing calls defines the nationality of the citizen. Following this principle we were able to identify the different ethnic groups in the city, where they work and commute within the city, similar to Girardin et al. performed with TIM data and Flickr dataset taken from Rome, Italy [8]. However we used Census data to understand tourists' call from immigrants' call. Apart from Giardini et al. our research focus on the ethnic groups, not only the tourists, in order to produce a further detailed portrait of the city, that moves beyond the limits of census data.

3. RELATED WORK

The urban planners underpinned the importance of diversity more than a decade ago [21, 19]. The heterogeneity of the communities foster creativity [14], it can encourage tolerance [7], boost economical growth [15], could create secure environments to live. According to Jane Jacobs the diversity of a city should be evaluated in different aspects, not only socially but also economically as well, the important thing is that all these parameters have to cohere in a harmonious way [10, 9]. The recent work from De Nadai et al, proved that it is not just a theory but a fact, by using mobile phone usage data [5].

The segregation of different ethnic communities among a city can be a good indicator, to understand how secure and healthy a community is. Beside from the spatial distance, social integrity is also an important parameter to evaluate [2]. The social integrity of a city can also be increased, if the urban planners create spaces for different socio-ethnic background people in the city, such as Trafalgar Square in London [21]. In the detailed work on diversity of the cities from Fainstein there is a difference in diversity among major cities [7], such as Amsterdam is more successful to create diversity rather than London and New York. In means of mobility, immigrants travel patterns are quite different from natives however this behavior assimilate in five year time according to a study by Gil Tal and Susan Handy [20]. This kind of behavior change is also in parallel within *spatial assimilation theory*, which propose the change of the new incomers residence according to economical constraints rather than social [12, 13].

Using mobile phone datasets as a proxy to census data is proposed by several studies [6, 16], however neither of these studies aimed to discover the immigrant populations among the city.

4. DATA

In that section we introduce the two datasets we use in our work. The first includes the call detail records and second the census data from Turin and Venice Municipalities from 2015 [1].

Telecom Italia organized Big Data Challenge in 2014 and 2015. We are planning to use the recent data from 2015. The data collection period covers 2 months from 01 March 2015 to 31 April 2015. The data is collected in 7 cities of Italy: Milan, Rome, Turin, Bari, Napels, Venice, Florence [1]. Call detail records (CDR) are mainly used for billing purposes by telecommunication operators.

The dataset includes incoming and outgoing SMS, incoming and outgoing calls and internet usage of subscribers, with the country code. In order to keep the data privacy the data is aggregated both spatially and temporally. Each city is divided into grids, with different in size. Temporal aggregation includes the aggregated activities 15 minute time windows.

The census data are obtained from Municipality of Turin and Venice, with the precision of administrative level for each nationality present in the city. Venice is more cosmopolitan city than Turin with 139 different nationalities all around the world, comparing to 50 in Turin.

[1]www.comune.torino.it/statistica/dati/stranieriterr.htm

5. METHODOLOGY

In our paper, we named tourists, as person who are in that city for a short period of time, either for touristic or business purposes. Immigrants are defined as, people who lived in that country, originating from other nations, and registered in the community. Beside unregistered immigrants are the ones who not legally live in that country therefore not present in the census, but possibly present in CDR. We elaborate the CDR dataset under this perspective and classify the nations either immigrant or tourists.

In CDR dataset, the country code that the call initiated or received are evaluated as the country of origin of the user.

First we compare the distributions of each nationality present both in census and CDR data as a bipartite network, as shown in Figure 1. On the left incoming calls are grouped according which country code it is initiated, in the middle the presence of this nation in the census, and on the right the number of outgoing calls for the same country. All values are normalized within itself ranging from 0 to 1. The countries are listed from top to bottom, as the presence in the corresponding dataset higher to lower respectively, shown in Figure 1. The degree α, corresponds to the degree value in the incoming call and population density, β corresponds to the degree value to population density. The overall degree (γ) is the sum of α and β.

In Figure 2 shows the degree (γ) and population densities of each country for Venice and Turin respectively.

We classified tourists and immigrants by applying linear discriminant analysis (LDA) to the calculated degrees γ from incoming calls to presence in the population and outgoing calls to presence in the population, as shown in Figure 2. The countries on the upper part of the line are more likely to be *tourists*, the ones through the line and below are taken as *immigrants*. From that formulation nationalities with higher in incoming and outgoing calls but low presence in census, are likely to be *tourist*, and the lower in both incoming and outgoing contrary to their high presence in census are likely to resident communities which are *immigrants*.

6. RESULTS AND DISCUSSION

In Turin, the immigrants represents the 15.4% of the citizenships (ISTAT 2015), in which the majority (39.5%) are from Romania. Their presence in the call dataset is proportional to their presence in the census data, this may be because they prefer using TIM as mobile phone operators, and use to communicate with their roots. In March 2013 TIM made a campaign to attract minorities, however this does not explain the rest of the other nationalities with negative correlation between census presence and number of calls. On the contrary, France, Poland, Swiss, Germany, Russia, Spain, UK show high presence in CDR dataset even with low presence in census. The possible explanations for that, there are more people living unregistered in Turin, or they can be tourists.

According to the CDR data, there are some communities prefer to live together like Romanians, Albanian and Moldavian. However Romanians with very high presence in the CDR dataset, they prefer to live outskirts of the city. However we do not confirm these finding from the census, as it includes only the inner city. Similar spatial sparsity in residence also valid for Italians. This can be explained by *spatial assimilation theory* introduced by Massey and Denton [12, 13]. Stating that after several years in the destination country immigrants reside in the city according to economical parameters.

One of the findings from that analysis is that, geographically close communities in origin are prefer to stay closer in the city as well. For African countries Morocco, Egypt and Nigeria basically present inside the Moroccan community as seen in Figure 3, on the left Egypt communities' night presence is shown for Turin, on the right Moroccan community is shown. Similar to African communities, Asian communities (China, Philippines) prefer to be closer as well.

7. CONCLUSION AND FUTURE WORK

In that work we detect different communities from two different Italian cities by using CDR data. We found out that similar communities tend to live together. Communities with high presence in census like Romania and Morocco, the residential areas are sparse all around the city. This may be due to the *spatial assimilation theory*.

The degree versus population distribution among the two cities seems to have a similar trend, this analysis can be performed for different cities in different countries, to generalize this tendency.

Possible future development might include the extension of the research from a market point of view, addressing how migrant flows affect the residential activity of the city and in which way it affects the native citizens behaviours, supporting the prediction for instance of possible gentrification phenomena.

A generative mathematical model can be applied to interpret the spatial distribution of communities, to avoid the missing data.

8. REFERENCES

[1] G Barlacchi, M De Nadai, R Larcher, A Casella, C Chitic, G Torrisi, F Antonelli, A Vespignani, A Pentland, and B Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2, 2015.

[2] Ravi Bhavnani, Karsten Donnay, Dan Miodownik, Maayan Mor, and Dirk Helbing. Group segregation and urban violence. *American Journal of Political Science*, 58(1):226–245, 2014.

[3] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1, 2015.

[4] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[5] M De Nadai, J Staiano, R Larcher, N Sebe, D Quercia, and B Lepri. The death and life of great italian cities: A mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*, pages 413–423, 2016.

[6] 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*, 111(45):15888–15893, 2014.

[7] Susan S Fainstein. Cities and diversity should we want it? can we plan for it? *Urban affairs review*, 41(1):3–19, 2005.

[8] F Girardin, F Calabrese, F Dal Fiore, C Ratti, and J Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4):36–43, 2008.

[9] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[10] Jane Jacobs. Thedeath and life of great american cities. *Jonathan Cape. London, UK*, 1961.

[11] Everett S Lee. A theory of migration. *Demography*, 3(1):47–57, 1966.

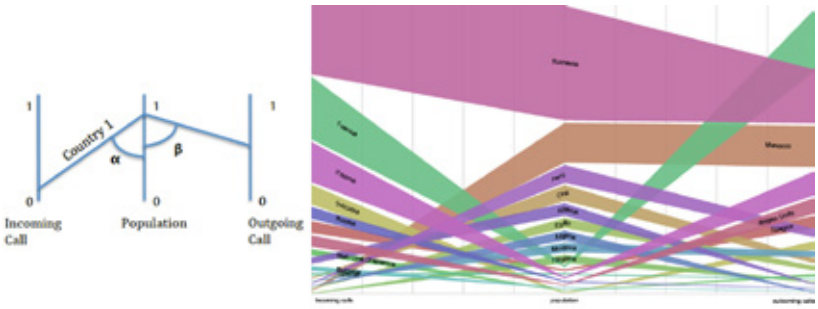


Figure 1: The bipartite network is on the left, on the right bipartite network with data from Turin. On the left the incoming calls coming to Turin from other countries, in the middle the population distribution from census, and on the right the outgoing calls from Turin to other countries.

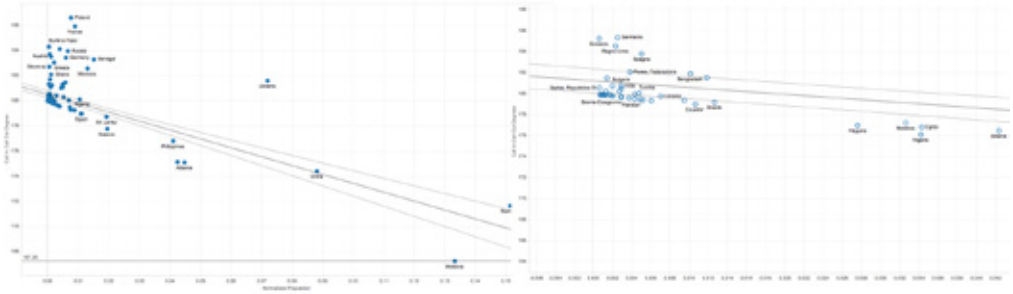


Figure 2: On the left hand, CallIn and CallOut degrees versus foreign population densities from Venice. On the right, Call In and Call Out degrees versus foreign population densities from Turin is shown.

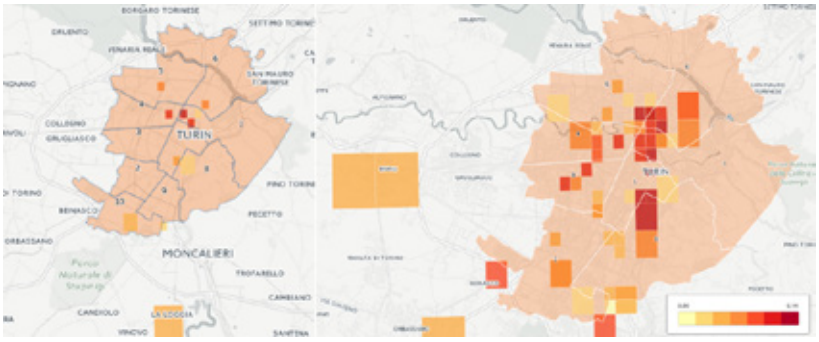


Figure 3: On the left Egyptians night location, on the right Moroccan night locations from Turin.

[12] Douglas S Massey and Nancy A Denton. Spatial assimilation as a socioeconomic outcome. *American sociological review*, pages 94–106, 1985.

[13] John Myles and Feng Hou. Changing colours: Spatial assimilation and new racial minority immigrants. *The Canadian Journal of Sociology*, 29(1):29–58, 2004.

[14] Gianmarco IP Ottaviano and Giovanni Peri. The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography*, 6(1):9–44, 2006.

[15] John M Quigley. Urban diversity and economic growth. *The Journal of Economic Perspectives*, 12(2):127–138, 1998.

[16] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007.

[17] Mimi Sheller and John Urry. *Tourism mobilities: places to play, places in play*. Routledge, 2004.

[18] Noam Shoval and Michal Isaacson. Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1):141–159, 2007.

[19] Michael P Smith. *City, state, and market: The political economy of urban society*. B. Blackwell, 1988.

[20] Gil Tal and Susan Handy. Travel behavior of immigrants: An analysis of the 2001 national household transportation survey. *Transport Policy*, 17(2):85–93, 2010.

[21] Iris Marion Young. *Justice and the Politics of Difference*. Princeton University Press, 2011.

PyMobility: an open source Python package for human mobility analysis and simulation

Luca Pappalardo[1], Gianni Barlacchi[2], and Filippo Simini[3]

[1] Department of Computer Science, University of Pisa, Italy
[2] University of Trento and SKIL-Telecom Italia, Trento, Italy
[3] Department of Mathematical Engineering, University of Bristol, UK

Human mobility modelling is crucial for urban simulation and what-if analysis [1, 2], e.g., simulating changes in urban mobility after the construction of a new infrastructure or when traumatic events occur like epidemic diffusion, terrorist attacks or international events. The developing of generative models that reproduce human mobility patterns in an accurate way is indeed fundamental to design smarter and more sustainable infrastructures, economies, services and cities. Clearly, the first step in human mobility modelling is to understand how people move. The recent availability of big mobility data, such as massive traces from GPS devices and mobile phone networks, allowed the discovery of the quantitative spatio-temporal patterns characterizing human mobility such as the heavy tail distribution in both trip distances [3–5], the characteristic distance traveled by individuals, the so-called radius of gyration [4–6] and the degree of predictability of individuals' movements [7]. Building upon the above findings, many generative human mobility models have been proposed so far which try to reproduce the characteristic properties of human mobility trajectories [8]. The goal of generative models of human mobility is to create a population of synthetic individuals whose mobility patterns are statistically indistinguishable from those of real individuals.

PyMobility is an open-source Python package that provides the implementation of many of the well-known human mobility models and measures. The long-term vision of the project aims at developing a complete package for human mobility analysis and simulation. PyMobility provides an efficient and easy to use implementation of the main collective and individual human mobility models existing in literature, allowing for both the fitting of the parameters from real data and the running of the models for the generation of synthetic spatio-temporal trajectories. Among the collective mobility models, i.e., models generating synthetic fluxes of people between locations on a space, PyMobility implements the gravity model [9, 10], the radiation model [11], and their recent improvements [12]. Among the individuals mobility models, i.e., models generating synthetic trajectories of desired length for a set of agents, PyMobility provides the recent improvements of the classical Exploration and Preferential Return model (s-EPR, r-EPR, d-EPR and recency EPR) [7, 6, 13] as well as the most accurate spatio-temporal generative human mobility models like DITRAS [14] and TimeGeo [15]. PyMobility also allows to compute a large set of measures on mobility data, both at individual (e.g. radius of gyration and user entropy) and collective level (e.g. trips per hour, motifs and origin-destination matrix).

References

1. S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, "Modeling human mobility responses to the large-scale spreading of infectious diseases," *Scientific Reports*, vol. 1, 08 2011.
2. C. Kopp, B. Kochan, M. May, L. Pappalardo, S. Rinzivillo, D. Schulz, and F. Simini, "Evaluation of spatio–temporal microsimulation systems," in *Data on Science and Simulation in Transportation Research* (L. K. D. Janssens, A. Yasar, ed.), IGI Global, 2014.
3. D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–465, 01 2006.
4. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, June 2008.
5. L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti, "Understanding the patterns of car travel," *The European Physical Journal Special Topics*, vol. 215, no. 1, pp. 61–73, 2013.
6. L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabasi, "Returners and explorers dichotomy in human mobility," *Nat Commun*, vol. 6, 09 2015.
7. C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, pp. 818–823, Sept. 2010.
8. D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella, "Human mobility models for opportunistic networks," *IEEE Communications Magazine*, vol. 49, pp. 157–165, December 2011.
9. G. K. Zipf, "The p1p2/d hypothesis: On the intercity movement of persons," *American Sociological Review*, vol. 11, no. 6, pp. 677–686, 1946.
10. W. S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the korean highway," *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48005, 2008.
11. F. Simini, M. C. González, A. Maritan, and A. L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, pp. 96–100, 2012.
12. M. G. Beiró, A. Panisson, M. Tizzoni, and C. Cattuto, "Predicting human mobility through the assimilation of social media traces into mobility models," *EPJ Data Science*, vol. 5, no. 1, p. 30, 2016.
13. H. Barbosa, F. B. de Lima-Neto, A. Evsukoff, and R. Menezes, "The effect of recency to human mobility," *EPJ Data Science*, vol. 4, no. 1, pp. 1–14, 2015.
14. L. Pappalardo and F. Simini, "Modelling individual routines and spatio-temporal trajectories in human mobility," *CoRR*, vol. abs/1607.05952, 2016.
15. Y. Yang, S. Jiang, D. Veneziano, S. Athavale, and M. C. Gonzalez, "Timegeo: a spatiotemporal framework for modeling urban mobility without surveys," *PNAS*, 2016.

An exploratory analysis of ethnic groups in the city of Milan through mobile phone data

Gianni Barlacchi[1,2], Michele Ferretti[4], Bruno Lepri[3], and Fabio Manfredini[5]

[1,2] *University of Trento & SKIL, Telecom Italia, Trento, Italy*
barlacchi@fbk.eu
[3] *Fondazione Bruno Kessler, Trento, Italy*
lepri@fbk.eu
[4] *Department of Geography, King's College London, London, United Kingdom*
michele.ferretti@kcl.ac.uk
[5] *Politecnico di Milano, Dipartimento di Architettura e Studi Urbani, Milano, Italy*
fabio.manfredini@polimi.it

January 31, 2017

1 Introduction:

Since 1999 the Municipality of Milan has registered a more than twofold increase in the number of foreign residents. By the end of 2015 they accounted for 19% of the overall population, corresponding to *circa* 260.000 inhabitants. Their spatial distribution and their country of origin presents specific patterns deriving from historical reasons, availability of low-cost housing and presence of social services.

An inventory of the living population across all the different *Nuclei di Identita' Locale* (NIL), Milan's neighbourhoods, is provided on a yearly basis by the municipal population register. These recordings, although highly valuable, fail to provide actual information on the populations' typologies and densities. They thus necessitate of further integration with alternative datasets [1]. Further, the spatial and temporal dynamics of foreigners' concentration within the city, as well as their actual trends and patterns, are difficult to intercept through conventional data sources [2].

Since the beginning of 2010, the availability of mobile phone data has opened a new field of research at the intersection of new and traditional disciplinary traditions (city planning, sociology, computational geography, urban informatics etc...) aimed at exploring the potential use of mobile phone data for understanding mobility patterns and dynamics [3, 4], to understand structural conditions underlying the quality of urban life [5], and thus ultimately to provide new services for urban populations. [6].

Nowadays, there are almost 6 billion of mobile phone users worldwide [7]. The world coverage raised from 12% of the world population in 2000 up to 96% in 2014 [8], with this trend further accounting for almost the totality of the population in the so-called "developed" countries. Such devices generate an incredible amount of Call Detail Records (CDRs), encompassing informations about how our daily mobile phone usage. Furthermore, they contain location data (e.g. from where we call), a characteristic that makes CDRs an extremely informative source of mobility patterns, and equally of social connections. Mobile phones are in fact considered to be strictly personal items, in contrast with fixed-line phones that are usually shared between more users, e.g., different tenants in a house, colleagues in a office... This characteristic, coupled with their very fine spatial and temporal resolution, renders mobile phone data highly suitable for reconstructing a caller' social network.

2 Challenge:

Mobile phone data could then be applied to study the distribution of ethnic communities in time and in space. Such study might prove relevant for understanding their growing role in the city economy, for monitoring potential risks of spatial segregation, providing new user-centred services or ameliorating the existing. As demonstrated by Hughes et al. [2] in his review of comparative migration estimates, conducted by the means of traditional and new data sources, these would prove to be hard-to-reach goals with methods confined to traditional social research. Other scholars in the field used mobile phone data, or other novel sources, for studying multicultural diversity of cities [9] and for analysing segregation with large-scale spatial network data [10, 11].

Following a similar line of inquiry, the present research focuses on the study of the foreigners' spatial distribution and activity within the city of Milan, *via* the means of telecommunications activities provided by TIM [12]. The dataset containing such activity serves as a measure for the level of interactions occurring between different users within the mobile phone network. The dataset is spatially aggregated using a regular grid and contains information on SMSs and calls received inside each grid square. A valuable information provided by this dataset is the country code of the sim-card connecting to the mobile network. In particular, the data contains a segmentation of the activity, also based on the grid square, the time of the day and the country involved in the communication activities. This information will allow a partial reconstruction of the communications' patterns between users located in Milan and abroad, such as migrant populations.

The data are available on a daily basis for a period of more than two months: from 2013-11-01 to 2013-12-20. The number of calls issued or received could then be correlated with the interactions of mobile phone users within different nations for the aforementioned purposes (e.g. personal communications, business...). Lastly, the gridded data will be aggregated according to the NIL boundaries, to contrasting the official residents data provided by the municipality with the resulting statistics. Such comparison will highlight any parallelisms or discrepancies between the two data sources, further investigating if, and under which manner, mobile phone data can be used *i)* as novel data sources

to describe phenomena for which official statistics do not collect data *ii)* as additional sources complementing the official statistics *iii)* or as alternatives to replace conventional ones [13].

References

[1] Pierre Deville et al. "Dynamic population mapping using mobile phone data". In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 15888–15893.

[2] Christina Hughes et al. *Inferring Migrations : Traditional Methods and New Approaches based on Mobile Phone , Social Media , and other Big Data*. Tech. rep. European Commission, 2016, p. 41.

[3] Marta C González, Cesar a. Hidalgo, and Albert-László Barabási. "Understanding individual human mobility patterns." In: *Nature* 453.7196 (June 2008), pp. 779–82.

[4] Cesar a. Hidalgo and C. Rodriguez-Sickert. "The dynamics of a mobile phone network". In: *Physica A: Statistical Mechanics and its Applications* 387.12 (May 2008), pp. 3017–3024.

[5] Marco De Nadai et al. "The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective". In: *Www* (2016), pp. 413–423. arXiv: `1603.04012`.

[6] Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis". In: *EPJ Data Science* 4.1 (2015), pp. 1–55. arXiv: `arXiv:1502.03406v1`.

[7] Vincent D. Blondel et al. "Data for Development: the D4D Challenge on Mobile Phone Data". In: *CoRR* (2012), pp. 1–10. arXiv: `arXiv:1210.0137v1`.

[8] "The world in 2014 : ICT Facts and Figures. International Telecommunication Union. http://www.itu.int/ (2014)". In: ().

[9] Michela Arnaboldi et al. "Studying Multicultural Diversity of Cities and Neighborhoods through Social Media Language Detection". In: *Tenth International AAAI Conference on Web and Social Media*. 2016, pp. 2–7.

[10] J Blumenstock and L Fratamico. "Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data". In: *Proceedings of the 4th Annual Symposium on Computing for Development, ACM DEV 2013* (2013).

[11] Siiri Silm and Rein Ahas. "The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset". In: *Social Science Research* 47 (2014), pp. 30–43.

[12] Gianni Barlacchi et al. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino." In: *Scientific Data* 2 (2015), p. 150055.

[13] Giulio Barcaroli et al. "Dealing with Big data for Official Statistics : IT Issues". Dublin, Ireland and Manila, Philippines, 2014.

Determining an optimal time window for roaming data for tourism statistics

Martijn Tennekes
Statistics Netherlands
Email: m.tennekes@cbs.nl

May Offermans
Statistics Netherlands
Email: mpw.offermans@cbs.nl

Nico Heerschap
Statistics Netherlands
Email: n.heerschap @cbs.nl

***Abstract*—Mobile phone data can be used for making Official Statistics on various subjects, such as safety, mobility and tourism. One of the major challenges is the protection of privacy when analyzing and processing these data for publication or fundamental research. Using sensitive information sparingly is one of the principles that is often put forward in privacy discussions. The time dimension is key for analyzing Call Detail Records (CDRs) but also for privacy. Therefore, in this paper an optimal time frame is determined for roaming visitors that provides sufficient information to create official tourism statistics of high quality while ensuring privacy. For this study we used anonymized aggregated CDR data from Vodafone in collaboration with Mezuro. The results show that a time window of 15 days is sufficient for official tourism statistics. However, a time window of one month is preferable in order to compare mobile phone based tourism estimations with the current tourism publication numbers.**

I. Introduction

One of the advantages of using Call Detail Records (CDR's) is the enormous detail of information in time and space. This advantage provides new opportunities for National Statistical Institutes like Statistics Netherlands for creating statistics on daytime population and tourism [1-4]. New and more precise information becomes available making it possible to create fast, almost real time statistics on a national level with high regional detail. One of the major challenges when making statistics with an extensive time horizon is the management of privacy and data-access privileges. Since 2009, Statistics Netherlands has collaborated with Mezuro and Vodafone in research on CDR's [5]. No data provided by Vodafone may lead to the identification of a person or company. Anonymization, aggregation of data, and remote systems that leave all personal sensitive data at the providers data center are important measures to ensure privacy. Different privacy rules are applied to ensure that even if information from space and time is combined, no identification is possible. One of these privacy rules in the data provided by Vodafone and Mezuro is that a device receives an anonymized IMSI (International Mobile Subscriber Identity (IMSI) for a period with a maximum of 31 days. After this period, a new anonymized IMSI is created which ensures that the device cannot be tracked for a substantive period of time.

The authors and Statistics Netherlands would like to thank VodafoneZiggo and Mezuro (www.mezuro.nl) for providing data and processing the CDR's. The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

This ensures privacy and provides enough information for statistical research purposes. However, devices that have roaming switched on in the Netherlands, the period was set at 24 hours. The reasons for this shorter, more strict period is the fact that there is no legal framework on collecting data from foreign devices that roam on the Dutch network. Each tourist receives an SMS when it starts roaming on the network. However, form a legal point of view it is impossible to obtain full informed consent or to provide an opt-out option. Therefore, the 24 hour IMSI-hash is introduced. The disadvantage of this 24-hour privacy rule is that the roaming data does not meet the requirements needed for tourism statistics. For example, we can roughly estimate the number of Japanese tourists in Amsterdam within a 24-hour timeframe, but it is not possible to see how long these tourists stay and where they consecutively go when visiting the Netherlands. These statistics contain important information on, for example, national and local tourism planning, public transport and (international) festivals. Determining an optimum period for this privacy rule is therefore important. On the one hand, one wants to produce useful statistics for the general public, government, and industry, and on the other hand, only use information that is necessary to produce these statistics.

II. Data

For this specific analysis, we used two aggregated and anonymized data tables that were extracted from CDRs at the Vodafone data center from all roaming customers on the Vodafone network in the Netherlands during the period from May 1st 2014 to May 31st 2014. The minimum cell count of these CDR tables is 15. Table cells with values lower than 15 are coded as missing to prevent identification. For this 31 day period, the anonymized IMSI numbers were not renewed. Instead, all geolocation data has been left out. The specific descriptions of the two tables that we have used are as follows:

1) The first table contains three columns: first date, last date, and number of foreign roaming devices. The number of foreign users refers to the number of users for which an CDR event is logged every day between (and including) the first and the last date. We refer to this date range as *consecutive stay*. Note that we do not know the length of the consecutive stay if the first date is May 1st or the last date May 31st, because foreign users could

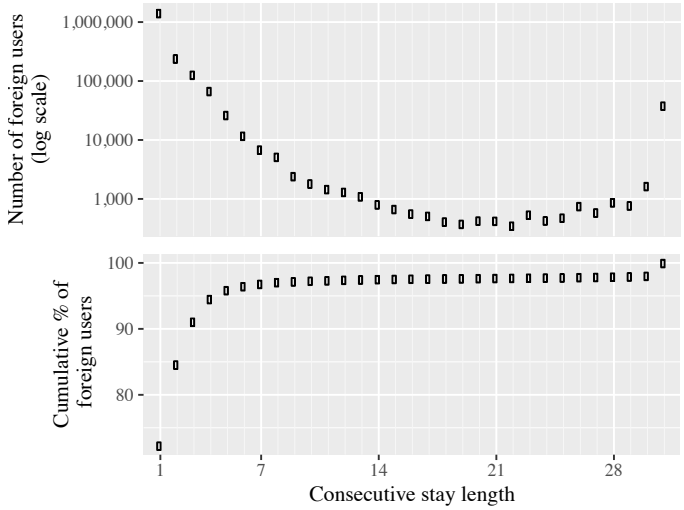


Fig. 1. Number of foreign users per consecutive stay length. Absolute numbers are shown above, cumulative below.

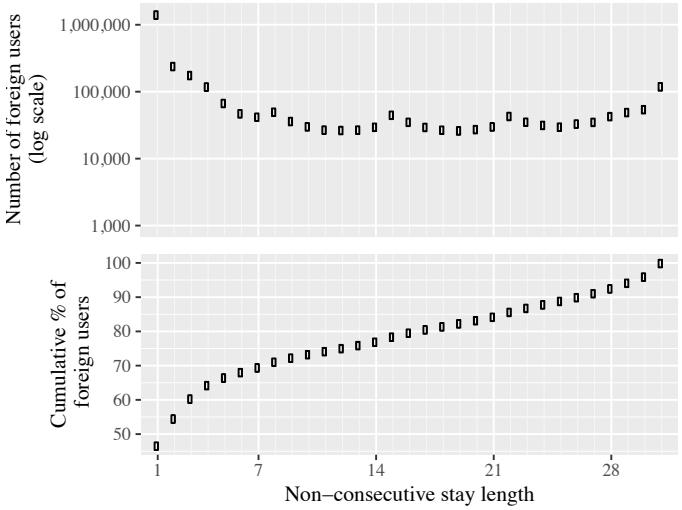


Fig. 3. Number of foreign users per consecutive stay length. Absolute numbers are shown above, cumulative below.

also have been logged before or after the observed time window.

2) The second table contains four columns: first date, last date, number of foreign users, and country of origin. In this table, the number of foreign users refers to the number of users for which the CDR events are logged between (and including) the first and last date, but not necessarily every day in between. Therefore, we refer to this date range as *non-consecutive stay*. To prevent low cell volume, the countries of origin were aggregated as follows: Belgium, Germany, France, United Kingdom, Eastern Europe, Southern Europe, Europe other, Middle-East, China-Japan-Australia-New Zealand, Asia other, North America, South America, and Inter Standard Roaming (ISR). Devices that roam with ISR use a different technology in their country of origin than GSM, such as CDMA. ISR is used in order to make roaming possible in Europe. ISR as a category contains users of different counties, mainly US, Canada, Japan and the Middle-East.

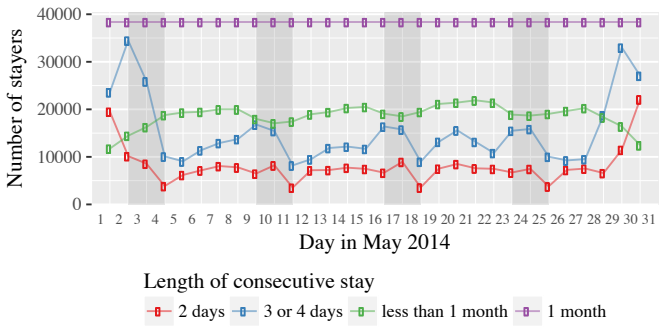


Fig. 2. Consecutive period of foreign users.

III. Results and Discussion

In this study, we focus on the length of stay of foreign roaming devices in the Netherlands. For tourism statistics, it is important to classify foreign users into standard categories, such as tourists, workers that cross the border on a regular basis, and normal border crossings (for example to do shopping). However, it was not possible to classify foreign users in this study, because it is unclear from the used data whether foreigners stay overnight in the Netherlands, or travel back and forth. It can generally be assumed that many people from Belgium and Germany, and, to a lesser extent France, will make day trips, whereas people from other countries will stay overnight in the Netherlands.

Figure 1 shows the number of users per consecutive stay length. For users who were logged on the 1st or the 31st of May, the consecutive stay can be longer, since they could have been in the Netherlands earlier or later than May.

The vast majority of users have only stayed for one day in the Netherlands. The number of users who have stayed between 15 and 30 consecutive days is relatively low. However, there are 38373 regular foreign users (almost 2 percent), probably people who live abroad and work or study in the Netherlands, or people who temporarily live in the Netherlands to work or study.

One of the aims for tourism statistics is to determine the total number of (overnight) stays and day trips of tourists in the Netherlands. In order to obtain those, the absolute numbers shown in Figure 1 will have to be multiplied by the active stay lengths. This will put more emphasis on foreign users who stay for longer periods.

Figure 2 shows the number of (overnight) stays during the observed period, grouped by the total consecutive stay length. For instance, the left-most green dot means that there are 23617 foreign users who have stayed between the 1st and the 2nd of May 2014 in the Netherlands (which can either

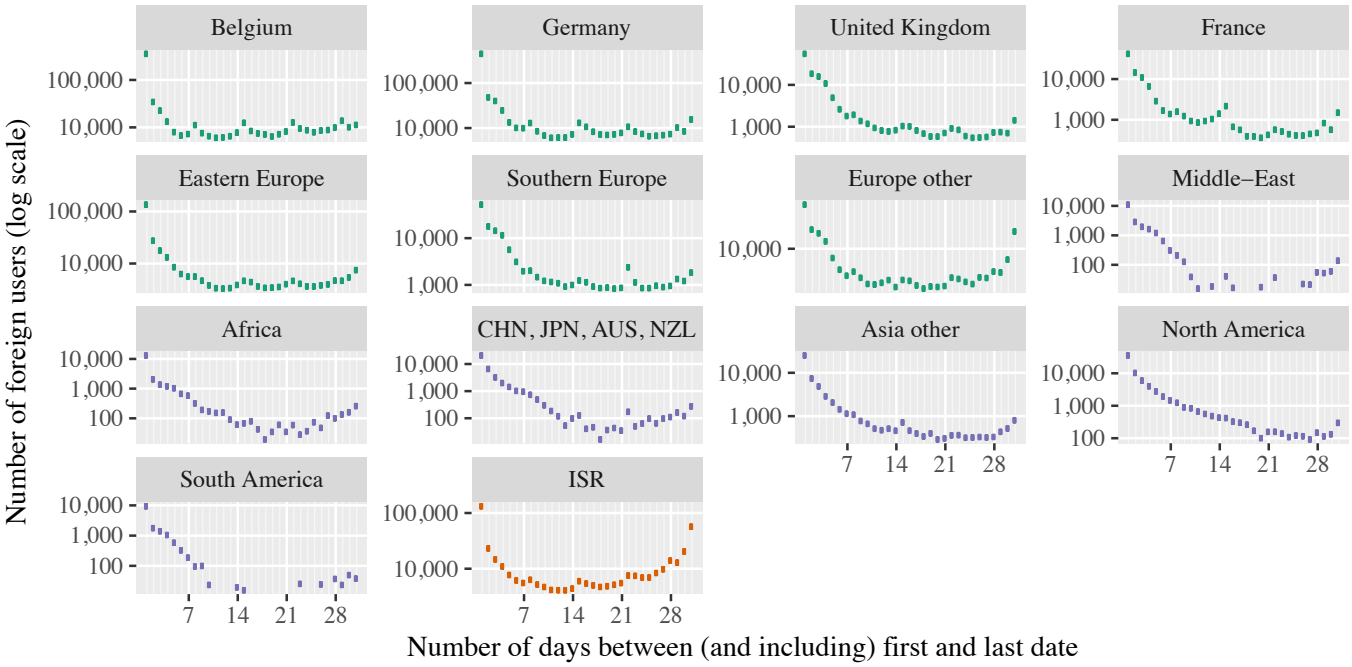


Fig. 4. Consecutive period of foreign users.

be an overnight stay or back and forth traveling) and 3 or 4 consecutive days in total.

Figure 2 also shows a typical tourism pattern. First, there is a holiday period from the 26th April 2014 until the 5th May 2014 of which the first days are visible in this dataset with high frequency of roaming data. These are mainly tourists that stay for 2 days or 3-4 days. Second, there is a weekend pattern (grey area) with again mainly tourists that stay for 2 days and 3-4 days. Many tourists stay from Thursday to Sunday. Third, there are the one-day events. The 29th of May is Ascension Thursday (a National day), and a lot of people also take holiday on the Friday after this Ascension Thursday, again 2 days or 3-4 days. The 5th of May is called Liberation Day (WWII) but is not visible, showing that tourists from other countries are not participating. Also for mothers day (for most countries on the 2nd Sunday of May), no differences are visible.

The number of users per non-consecutive stay length is shown in Figure 3. A stay length of 20 only means that the difference between the first and last date equals 20 days. It does not say anything about how many days in between the user actually stayed in the Netherlands. Observe that 30 percent of the foreign users have stayed non-consecutively for longer than 1 week in the Netherlands. An explanation for this could be the large amount of Belgium and German foreigners who work in the Netherlands or who regularly make day trips to the Netherlands.

Figure 4 shows the number of foreign users per non-consecutive stay length by (grouped) country of origin. The color of the dots correspond to the continent: Europe (green), other (purple) and Inter Standard Roaming Roaming (orange). Small local peaks are visible at 8, 15, and 22 days, which probably corresponds to standard holiday stays of one, two, and three weeks. There are differences visible per country of origin. Germany and Belgium show for example higher rates of visitors that stay short, usually one day , but we will not describe them in detail.

The total number of foreign users per (grouped) country are depicted in Figure 5 shows as expected that the number of German and Belgium users is relatively high. Also, the number of Inter Standard Roaming user is high, especially among regular users.

This dataset contains several limitations. First, in the Netherlands there are four network providers and Vodafone is one of them. In this study, no corrections or extrapolations were

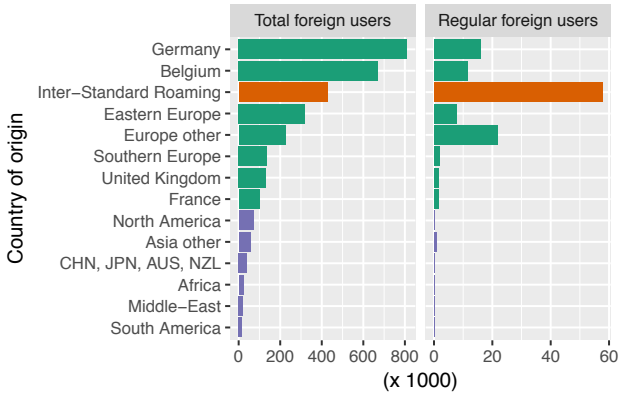


Fig. 5. Consecutive period of foreign users.

made. Second, a time window of 31 days was used. Therefore, the stay length of foreign users that were already in the Netherlands at May 1st or were there at May 31st cannot be determined exactly, since they could have been in the Netherlands earlier or later than May. Third, there was no geolocation provided in this dataset. Therefore, it was not possible to make a clear distinction between tourists on the one hand and border traffic and foreign workers in the Netherlands on the other hand.

IV. Conclusion

Based on the results presented in Figures 1, 2, and 3, it can be concluded that a time window of 15 days is sufficient for covering almost 98% of consecutive stays and 80% of non-consecutive stays. It is advisble to choose the time window larger than the desired stay length to be measured, in order to decrease the probability of overlap.

For tourism statistics, a time window of 1 month is preferable in order to compare the mobile phone based estimations of tourists with the current tourism statistics.

For future research, it is necessary to classify foreign users as tourists, day trip visitors, and cross border workers. For tourism statistics, also the geospatial information is important. Once roaming data is available for time window of 15 days or longer, geospatial information can be used to classify foreign users based on the places they visit.

References

[1] Raun, J. Ahas, R., 2013. Distinguishing tourism destinations with behavioural data. Brussel, New Techniques and Technologies for Statistics, NTTS.

[2] European Commission, 2014. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Eurostat.

[3] Altin, L., Tiru, M., Saluveer, E., Puura, A., 2015. Using Passive Mobile Positioning Data in Tourism and Population Statistics, Brussel, New Techniques and Technologies for Statistics, NTTS.

[4] Meersman, F. de, Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A. Demunter, C., Reis, F., Reuter, H. I., 2016. Assessing the Quality of Mobile Phone Data as a Source of Statistics, Paper for the European Conference on Quality in Official Statistics (Q2016).

[5] Offermans, M, Priem, A, Tennekes, M., 2013. Rapportage project Impact ICT mobiele telefonie. Technical report (in Dutch). Statistics Netherlands.

Measuring transnational population mobility with roaming data

Rein Ahas[1], Siiri Silm[1], Margus Tiru[1,2]

[1]Department of Geography, University of Tartu, Estonia

[2]Positium LBS, Tartu, Estonia

Transnational population mobility can be defined as living and working in two or more countries (Schiller et al 1995). A transnational lifestyle is becoming more and more commonplace in various parts of world thanks to rising population mobility, changing labor market, opening borders, and developing information and communication technologies. Transnationality confronts the historical concept of 'Nation State', which approaches the world from an idealistic, citizenship and territory-based perspective which insists that all permanent residents should be citizens and all citizens should be permanent residents (Mayer et al, 1997). Today we can see that many transnational communities and floating population segments live and work outside of their homelands or apart from their families and homes. Measuring, understanding, and the management of such communities is complicated; we can see that traditional statistics and register data sources are not suitable to describe the transnational community.

In this paper, we are proposing a methodology for measuring the transnational mobility of a population with the help of the roaming databases of mobile network operators. We developed a conceptual and methodological framework for detecting tourists, cross-border commuters, transnationals, and foreign workers from roaming CDR datasets (Eurostat, 2014). We use this methodology for measuring outgoing mobility from Estonia in the databases of the two largest mobile network operators with market shares of 39% and 34%. Methodological challenges are related to integrating data from two mobile network operators; defining trips; and proposing theory based and distribution based parameters for algorithms. The work is related to the development of migration statistics within the framework of the Eurostat BIG Data Task Force.

Table 1. The distribution of outbound visitors according to visitor segments, data from the two biggest mobile network operators in Estonia on 2015.

Visitor segment	Number of persons	Share from outbound visitors	Number of countries visited	Annual average		
				Number of trips	Number of days	Duration of trips
Tourists	592,335	98.2	200	3.6	16.7	5.0

Transnationals	23,587	3.9	101	15.7	158.1	13.3
Cross-border commuters	1,518	0.3	14	85.3	144.7	1.9
Foreign workers	4,366	0.7	61	9.6	306.7	89.4
Total	603,355		200	4.4	25.3	6.0

Results so far show that, in methodological terms, it is possible to distinguish between different transnational segments of outbound travelers. An average outbound traveler carried out a total of 4.4 trips with an average duration for each trip of six days annually, but this averaged result is significantly influenced by high share of tourists in outbound travel (Table). Transnationals carried out 15.7 trips with a duration of each trip lasting 13.3 days. A total of 2.7% of transnationals were connected to two or three foreign countries. Cross-border commuters carried annually out 85.3 trips with average duration of trip 1.9 days and foreign workers 9.6 trips with average duration 89 days. Results show that the most popular countries to be visited by transnationals from Estonia are Estonia's closest neighbors in Scandinavia and Western Europe. One unexpected result was the relatively small proportion of transnationals going to Russia, even if Russia is a very important tourism destination and is closely connected to the Russian minority living in Estonia (forming 27% of the total population). The travel behavior parameters and the social profile of visitors allows us to differentiate between tourists, cross-border commuters, foreign workers, and transnationals.

The results are interesting, but validating our results is a rather complicated business as it is also complicated when it comes to finding alternative sources of data for validation. We discuss the strengths and weaknesses of such data and the methodology from the point of view of the international migration research and statistical system.

The phenomenon of transnational commuting is closely related to the development of ICT and the information society (Nedelcu & Wyss, 2016). Ubiquitous mobile communication data is something that can be a starting point when it comes to broadening the concept of the 'Nation State' with citizens and residents having activities in several countries. The concept of e-residency (Estonian e-Residency 2017) can be one way of developing concept of transnational citizenship.

References

Estonian e-Residency 2017. https://e-estonia.com/e-residents/about/

Eurostat 2014. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report Eurostat Contract No 30501.2012.001-2012.452, Positium LBS, 31p. http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies

Meyer, J. W., Boli, J., Thomas, G. M., & Ramirez, F. O. (1997). World society and the nation-state. American Journal of sociology, 103(1), 144-181.

Nedelcu, M., & Wyss, M. (2016). 'Doing family' through ICT-mediated ordinary co-presence: transnational communication practices of Romanian migrants in Switzerland. Global Networks, 16(2), 202-218.

Schiller, N. G., Basch, L., & Blanc, C. S. (1995). From immigrant to transmigrant: Theorizing transnational migration. Anthropological quarterly, 48-63.

6 APRIL 2017

POSTER
SESSION 2

NetMob 5-7 April 2017 – Vodafone Village – Milan

CLIMATE CHANGE INDUCED MIGRATIONS FROM A CELL PHONE PERSPECTIVE

Sibren Isaacman[1], Vanessa Frias-Martinez[2], Lingzi Hong[2], Enrique Frias-Martinez[3]

[1] Loyola University Maryland; [2] College of Information Studies, University of Maryland;
[3] Telefonica Research, Madrid, Spain

isaacman@cs.loyola.edu {vfrias, lzhong}@umd.edu enrique.friasmartinez@telefonica.com

ABSTRACT

Cell phone traces have been successfully used to study human mobility during natural disasters such as earthquakes and flooding [1,2,3]. Climate change, understood as the change in weather patterns for a long period of time, also has the potential of causing changes in human mobility and cause migrations that have a wider and long standing impact. A *climate migrant* is an individual that is forced to leave their local environment due to long or sudden weather changes. In this study we present initial results of the migrations caused by the severe drought that happened in La Guajira, Colombia, in 2014. Our initial results indicate a linear reduction of the population of 10% during the 6 months considered.

La Guajira is a department of Colombia located in the northwest tip of the country and borders Venezuela. According to the UN Office for the Coordination of Humanitarian Affairs (OCHA)[4], since the beginning of 2014 an extreme drought has affected La Guajira. The drought caused a declaration of the state of public calamity in the municipality of Uribia (La Guajira) in February of 2014[5]. It is estimated that around 65,000 people have been affected by the severe droughts in La Guajira. Consequences to the population are mainly regarding malnutrition, especially for infants, with extreme consequences for the agricultural and livestock sectors.

Using a 6-month (December 2013 through May 2014) anonymized CDR dataset from Colombia we wanted to measure the impact, if visible, in the number of inhabitants of both Uribia and La Guajira during the drought period. A home detection algorithm based on when phone calls were made was applied weekly for any cell phone that was present in La Guajira during the period of study. Note that all calls independently of where they were made have been considered when applying the home detection algorithm, i.e. homes were assigned anywhere in Colombia. If there was not enough information to assign a home tower in a given week we assumed the last known home. Figure 1 presents the number of cell phones whose home location was assigned in the towers that cover Uribia and La Guajira from January to June during each week, showing a linear decrease in both cases of 10% of the population during the period of study. Both cases can be model with a linear regression showing an r^2 of 0,78 and 0,93 for Uribia and La Guajira respectively.

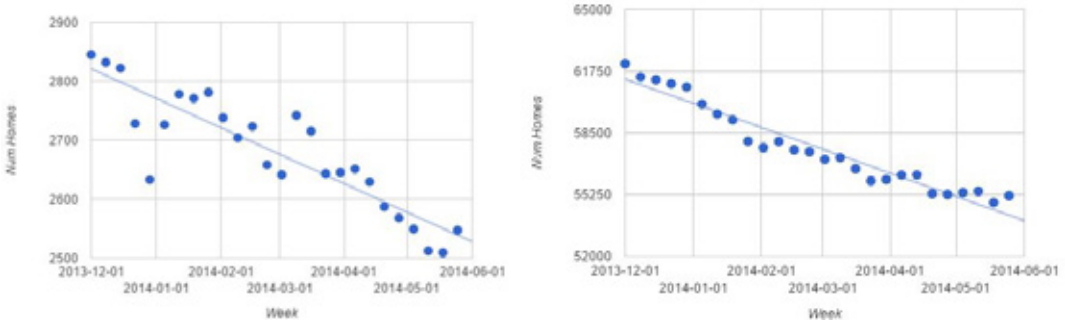


Figure 1. Number of Homes detected in (left) municipality of Uribia and (right) the department of La Guajira from December 2013 to June 2014.

Regarding where those climate migrants go, 90% of them stay in La Guajira relocating to other municipalities where access to help, food and water is probably easier. The other 10% move to other departments. In general, the closer the department to La Guajira the higher the number of people relocating is, i.e. climate migrants move to neighboring departments, with the exception of Bogota. Figure 2 presents in a red color scale where the migrants move from the second week of January 2014.

Figure 2. Departments where climate migrants move in Colombia for the second week of January 2014.

References

[1] Moumni, B., Frias-Martinez, V., & Frias-Martinez, E. (2013, September). Characterizing social response to urban earthquakes using cell-phone network data: the 2012 oaxaca earthquake. In *Pro0. 2013 ACM Conf. Pervasive and Ubiquitous Computing* (pp. 1199-1208). ACM.

[2] Pastor-Escuredo, D., Morales-Guzmán, A., Torres-Fernández, Y., Bauer, J. M., Wadhwa, A., Castro-Correa, C., ... & Oliver, N. (2014, October). Flooding through the lens of mobile phone activity. In *Global Humanitarian Technology Conference (GHTC), 2014 IEEE* (pp. 279-286).

[3] Frias-Martinez, V., Soguero, C., & Frias-Martinez, E. (2012, August). Estimation of urban commuting patterns using cellphone network data. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* (pp. 9-16). ACM.

[4] UN Office for the Coordination of Humanitation Affairs: Colombia Disaster Snapshot 2014 http://reliefweb.int/sites/reliefweb.int/files/resources/DisastersSnapshot_Colombia_April2014_1.pdf

[5] Declaration of Public Calamity in Uribia (in Spanish) February 14th 2014: http://uribia-laguajira.gov.co/apc-aa-files/62353738366136663032666631393236/decreto-calamidad-publica-uribia.pdf

Abstract Submission to NetMob17
Fifth conference on the Analysis of Mobile Phone Datasets

High speed analysis of volatile mobile data applied to road safety

Gómez Castaño, José
CTO INSPIDE jgcasta@inspide.com
GIS Specialist Dpto. Astrofísica y CC de la Atmósfera;
Univ. Complutense de Madrid jgomez03@pdi.ucm.es

Cabrera García, Juan José
CEO INSPIDE jjcabrera@inspide.com

The mobile data analysis focuses on the extraction of information using large amounts of data. These come not only of the CDRs, but from other sources such as social networking or connections from mobile applications. In most cases, this analysis is carried out over a variety period of time and they use the dataset accumulated implementing batch processes. These bear a certain amount of run time depending on the algorithm complexity .

If we want to obtain quality results in certain environments, the processing of the data must be more agile and it have to spend less time. In the field of road safety, for instance, you cannot wait for a process that takes a long time, due to the fact that the traffic is changing, and the information should be provided as soon as possible.

This work explores the analysis of the mobile data taking into account two aspects: High Speed Analysis and Volatility of the Information. The purpose is to gather georeferenced information provided by mobile data, determine the position in relation to the traffic and computes traffic congestion situations. The result is sent to the drivers concerned and to inform Traffic Authorities, in real time. Congestions are bottlenecks and the are a traffic hazard because a reduction of the unexpected speed by drivers and they may be due to multiple causes.

The origin of the information, in this work, are anonymous Twitter messages, and they are obtained using the API provided to developers [API Twitter 2016]. This API allows us to get all the messages with location information associated. We have chosen this set of data for its easy access and its proven usefulness in works such as the of Llorente, et al 2015, Carley, K et al.(2014) and Abassi, A. et al (2015)

Regardless the origin of the messages, the treatment of the information, their volatility and the results are independent of the source data, and you can apply this methodology to any other data.

The concept of volatility of the information is related to the use of the information. So we define time of volatility of the information as the period of time during which a data is useful for a particular purpose. This period depends not only on the data, but also on the processes of calculation required to obtain a useful outcome from them. High Speed Analysis are the processes implemented to achieve this. After this period, the data is considered as expired.

In this work applies this concept of volatility of the information to improve road safety. The receipt of messages is carried out by a series of processes developed in Python that run on a distributed Spark architecture implemented on a Amazon AWS EC2 environment.

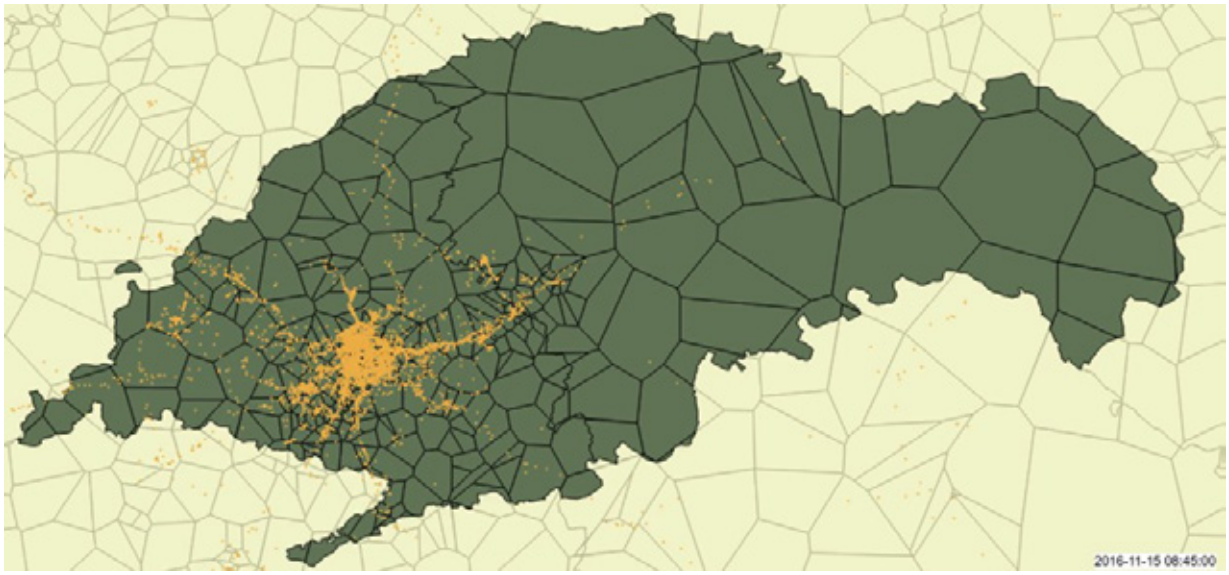
The location of every message received is compared with the geometries of roads stored in a PostGIS spatial database. These geometries (2.6 million) have been previously charged and they come from OpenStreetMap. The set of messages is translated in a geometry of roads for the stretch of the same affected by congestion. This new geometry is used as a final result as a track over the road geometry. The messages are grouped into short periods of time, after which they are discarded and are not valid for a good result. The are expired.

Because the aim is to improve the information to drivers, the system presented in this paper allows us to locate the information panels that are closest to the drivers in the direction of its motion, and it is possible to issue alerts of congestion to these panels. These panels are located over the road to inform the drivers with different notifications abut traffic. We have developed a preprocessing that locates the panels for every of the roads of study.

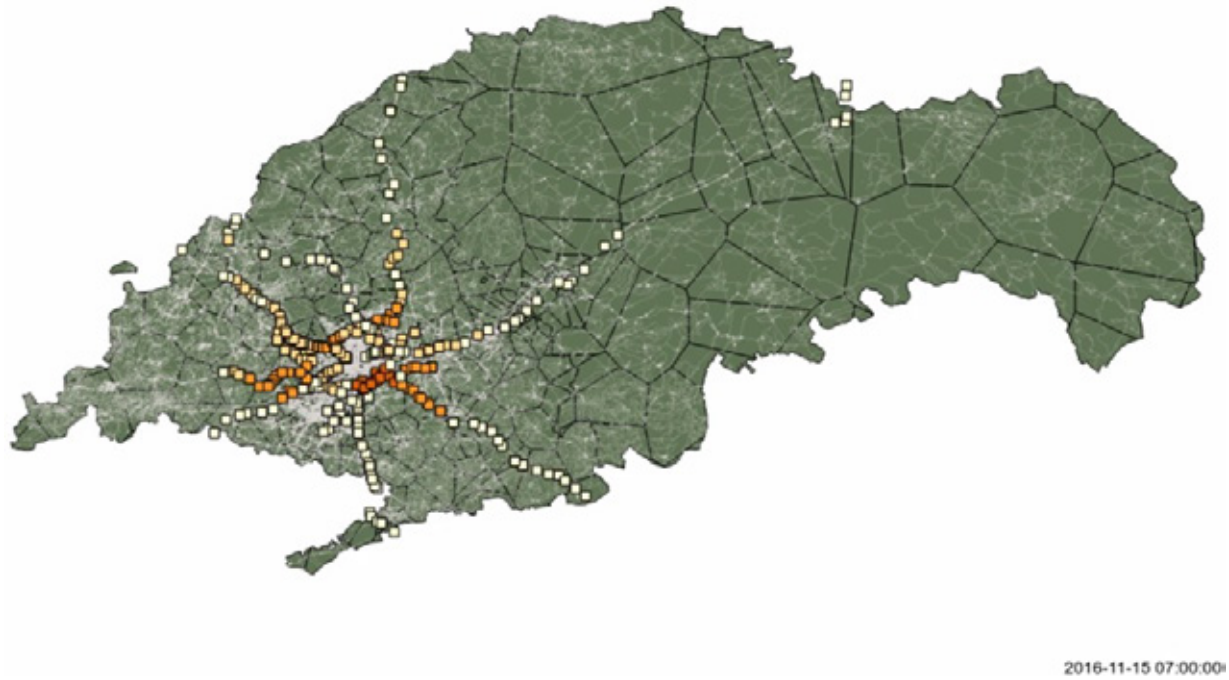
This information is also used by the Singularity platform, which provides the logic of road safety for Comobity application. This is a road safety applications developed for the General Directorate of Traffic [DGT 2015] and Gomez, J, Cabrera, JJ 2016, and it provides road safety information to more than 20,000 drivers in real time since November of 2015.

As a result is to be able to issue notifications and information that are visible by drivers in advance and automatically. As a result, below there is a graph indicating warning levels at different times, and an example of the tweets used. More than 400,000 locations and tweets has been taken into account to generate a near real environment as a test for the scenario over Spain, and more than 2,6 geometries have been used. The expiration time of the volatile data has been evaluated from 1 minute to 15 minutes.

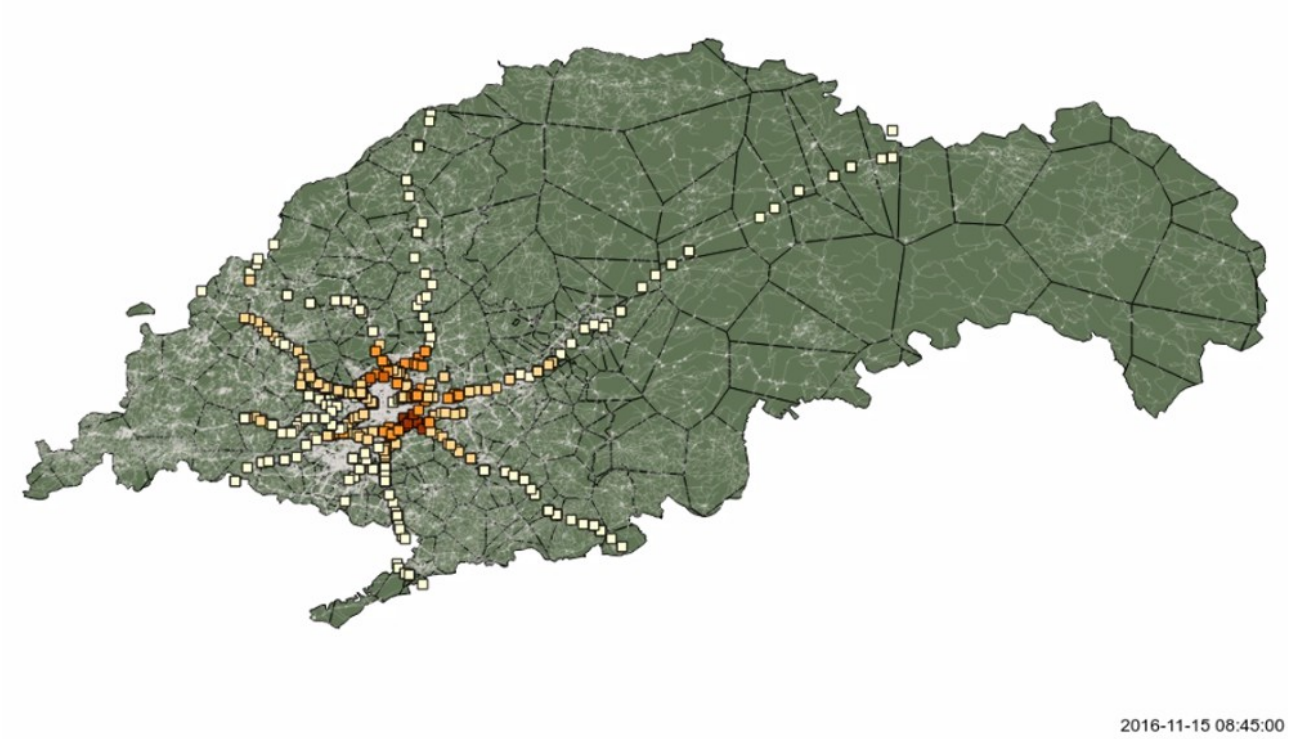
Better information to the drivers has an impact on an improvement in the road safety



tweets distribution at 8:45 over Madrid and Guadalajara provinces



Information level on panels at 7:00 over Madrid and Guadalajara provinces



Information level on panels at 8:45 over Madrid and Guadalajara provinces

References

Abassi, A., Hossein T, Maghrebi, M., Travis, S. 2015, "Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play" Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Bellevue, WA, USA, Nov 2015 ISBN: 978-1-4503-3975-9

API Twitter 2016 https://dev.twitter.com/rest/reference/get/geo/search

Carly, K, Kathleen M. Pfeffer, Jürgen Morstatter, Fred Liu, Hua (2014) "Embassies burning: toward a near-real-time assessment of social media using geo-temporal dynamic network analytics" Social Network Analysis and Mining, 2014, vol 4 issue 1 pag 195

DGT 2015 "Comobity: La app que conecta y protege a los conductores, ciclistas y peatones" http://www.dgt.es/es/prensa/notas-de-prensa/2015/20151112-App-Comobity-DGT.shtml

Gómez, J,Cabrera, JJ, 2016 "Phii y Comobity. Movilidad y Seguridad vial colaborativa" X Jornadas de SIG Libre, Universidad de Girona 2016

Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social Media Fingerprints of Unemployment. PLOS ONE 10(5): e0128692. doi: 10.1371/journal.pone.0128692

Context-Aware Recognition of Physical Activities Using Mobile Devices

Gabriele Civitarese Claudio Bettini
EveryWare Lab, Università degli Studi di Milano
Email: {gabriele.civitarese, claudio.bettini}@unimi.it

I. Introduction

Context-awareness in mobile apps has been the focus of several research efforts in the last years. The activity that a mobile user is performing while using an app is an essential component of context and an enabler for innovative functionalities [1]. While activity recognition is a wide area, in this work we focus on the recognition of simple physical activities (e.g., walking, running, riding a bicycle, ...) by processing data provided by the many sensors that we find inside modern smartphones and smartwatches. This task is not new and solutions are typically based on machine learning applied to data obtained from inertial sensors (e.g., accelerometer, gyroscope, and magnetometer) [2]. The techniques have been available to Android developers as well in the form of the Google Activity Recognition API. Despite the amount of work in the area there is still a lot of space to improve since the set of activities is very restricted, precision is sometimes inadequate, some activities are easily confused with others, and there is a delay in the recognition.

The EveryWare Lab has introduced hybrid recognition techniques and designed COSAR [3], a system that exploited the semantics of the locations where the activities are performed (e.g., home, work, park, street, ...) to refine the statistical classification. Essentially, an ontology which describes the relations between locations and activities is used to exclude from the classification results those activities which have a low probability of being performed in a particular location. In this work we present new results in line with that approach by using state of the art smartphones and smartwatches instead of wearable prototype sensors as in previous work, and we apply new algorithms. We collected a real dataset consisting of several activities performed by different subjects in different locations. The technological setup allows us to evaluate the quality of our recognition, the added value of considering smartwatch data in addition to smartphone, and to compare with the predictions obtained through the Google Activity Recognition API, showing the effectiveness of our approach.

II. Methodology

The general methodology consists in continuously acquiring and preprocessing raw data from sensors in order to classify the performed activities using standard machine learning techniques. Symbolic reasoning is then applied considering contextual information in order to refine the statistical classification. We assume that the user carries the smart-phone in the pants front pocket and the smart-watch at the wrist of the non-dominant arm. It is important to note that our method works even if only one of the two devices is used by the user (obviously with different accuracy levels).

A. Sensing

At the base of our infrastructure we consider a continuously collection of data from sensors equipped in nowadays smartphones and smart-watches using dedicated apps. In particular we consider *inertial sensors* (e.g., accelerometer, magnetometer, gyroscope, ...) to monitor the physical movements of the user and *context sensors* (e.g., GPS, location, light sensor, ...) to capture the context of the surrounding environment.

B. Preprocessing, segmentation and feature extraction

First of all, a median filter is applied on raw signals of inertial sensors to reduce the noise. Magnetometer data is then analyzed to infer the smartphone's orientation in the pocket. We consider "normal position" when the screen is against user's thigh, and when we detect that the magnetometer indicates a different orientation we swap the axis to bring all the smartphone's sensors to the "normal position".

After this preprocessing phase, we align and segment preprocessed data from the different inertial sensors of both devices using temporal sliding windowing. A feature vector consisting of several statistical features on smartphone/smartwatch inertial sensors data is then computed from each window.

C. Hybrid activity recognition

A multinomial classifier (which has been previously trained off-line) is used to obtain, for each feature vector, the probability distribution over all the considered activities. Since statistical classification accuracy is affected by intrinsic noise, we propose a symbolic refinement which aims to improve the recognition accuracy by considering the surrounding context where the activities are performed. In particular, we exploit semantic location information (e.g., home, office, park, street, ...) to exclude from the probability vector generated by the multinomial classifier all the activities which have a low probability of being performed according to the current context (e.g., taking the elevator in the street is very unlikely). The final result of classification is the most probable activity of the refined probability vector. Semantic location can be inferred using reverse geocoding techniques. However, in our preliminary implementation we annotated in the dataset the

semantic locations and we used those information for the experimental evaluation of the method.

III. Preliminary results

A. Experimental setup

Our experimental setup consists in a *LG Nexus 5x* smartphone equipped with 9-axis accelerometer, GPS sensor and several environmental sensors. This device allowed us to acquire data at a 200HZ frequency. We also adopted a *LG G-watch R* smart-watch which includes the same sensors of the above mentioned smart-phone (except for GPS). We developed two Android applications (one for the smart-phone and one for the smart-watch) capable to continuously collect the data from the sensors, allowing the user to introduce activities annotations.

B. Dataset acquisition

We designed the acquisition of a large dataset, in order to obtain a reasonable variability in activity execution. The considered activities are walking, running, go upstairs, go downstairs, still, sitting, elevator up and elevator down. We accurately planned four different scenarios for those physical activities to be executed in different locations: office, home, park and street. Having different locations it is also useful to collect different ways of performing the same activity. Consider, for instance, the activity "walking". This activity can be performed differently indoor and outdoor. We asked to sixteen different volunteers (aged between 20 and 27) to perform the activities of a particular scenario wearing the smartwatch on the non-dominant arm and the smartphone in the pants front pocket (any direction, face-up or face-down). Every activity in the scenarios has been performed for around 2 minutes (in total each scenario duration was 20-25 minutes). In order to collect the ground-truth, the users annotated the activities using a dedicated smartwatch application. We collected in total 6 hours of sensor data. While performing the activities we also had active on the smartphone an app based on Google activity recognition APIs.

C. Results

1) Evaluating the context impact: In Table I it is possible to analyze the comparison between using or not the context information to refine the statistical classification. The metrics indicated with the letter C are computed by considering context (the semantic location). The influence of the symbolic reasoning module is clearly most relevant for those activities which are strongly related with a location (for instance using elevator is very unlikely when being at the park). The improvement in recognizing "walking" is due to the fact that it is often confused with "going upstairs" activity. Since we improved the recognition of the latter, we also recognize better the former.

2) Comparison with Google Activity Recognition API: We finally compare our results with the ones provided by an app using Google APIs. For this comparison we considered the subset of our activities which are also considered by the APIs: walking, running, riding a bicycle, and still. Since the APIs only rely on the smartphone, we compared with our method excluding smartwatch sensor data. In Table II it is possible to analyze the comparison between the results obtained with Google Activity Recognition API (indicated with the letter G) and the ones obtained with our approach. It is easily seen that our methods significantly reduce the number of false positives. Overall, our method improved Google results by 10%.

TABLE I
Comparison of recognition with and without context refinement

Activity	Precision	Recall	F1	Precision C	Recall C	F1 C
Riding Bicycle	0.98	0.92	0.95	**0.99**	**0.94**	**0.96**
Walking	0.87	0.85	0.86	**0.89**	**0.92**	**0.90**
Running	1.00	0.99	0.99	1.00	0.99	0.99
Going upstairs	0.73	**0.80**	0.77	**0.82**	0.77	**0.78**
Going downstairs	0.72	0.71	0.72	**0.85**	**0.76**	**0.80**
Still	0.93	**0.98**	0.95	**0.94**	0.97	0.95
Sitting	0.97	0.99	0.98	0.97	0.99	0.98
Elevator	**0.94**	0.73	0.82	0.90	**0.82**	**0.86**
Total Average	0.91	0.86	0.88	0.91	**0.89**	**0.90**

TABLE II
Comparison of our method with Google AR API

Activity	Precision G	Recall G	F1 G	Precision	Recall	F1
Riding Bicycle	**0.99**	0.70	0.82	0.88	**0.97**	**0.95**
Walking	0.91	0.83	0.87	**0.95**	**0.95**	**0.95**
Running	**0.96**	0.85	0.90	0.94	**0.96**	**0.95**
Still	0.90	0.73	0.81	**0.97**	**0.97**	**0.97**
Total Average	0.93	0.77	0.84	**0.94**	**0.94**	**0.94**

IV. Conclusions and future work

In this work we presented new experimental results on refining statistical activity recognition using current mobile device sensors by considering the semantic location of the user. The results show how the contextual information of semantic location improves activity recognition. A preliminary comparison with Google Activity Recognition API shows how our approach is significantly superior for some activities. However, the comparison results need to be validated on much larger datasets including a large variety of users and situations and removing the assumption that we currently have on the user wearing the smartphone in the pants front pocket.

Acknowledgments

The authors would like to thank Chiara Mariani and Riccardo Presotto for their excellent work in software implementation and data acquisition.

References

[1] J. W. Lockhart, T. Pulickal, and G. M. Weiss, "Applications of mobile activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. ACM, 2012, pp. 1054–1058.

[2] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[3] D. Riboni and C. Bettini, "Cosar: hybrid reasoning for context-aware activity recognition," *Personal Ubiquitous Comput.*, vol. 15, no. 3, pp. 271–289, Mar. 2011.

Mining the Air -- for Research in Social Science and Networking Measurement

Scott Kirkpatrick, Hebrew University ***kirk@cs.huji.ac.il***
Ron Bekkerman and Adi Zmirli, Haifa University ***ron.bekkerman@gmail.com***
Francesco Malandrino, Politecnico de Torino ***francesco.malandrino@polito.it***

Smartphone apps provide a vitally important opportunity for monitoring human mobility, human experience of ubiquitous information aids, and human activity in our increasingly well-instrumented spaces. As wireless data capabilities move steadily up in performance, from 2&3G to 4G (today's LTE) and 5G, it has become more important to measure human activity in this connected world from the phones themselves. The newer protocols serve larger areas than ever before and a wider range of data, not just voice calls, so only the phone can accurately measure its location. Access to the application activity permits not only monitoring the performance and spatial coverage with which the users are served, but as a crowd-sourced, unbiased background source of input on all these subjects, becomes a uniquely valuable resource for input to social science and government as well as telecom providers.

The public also stands to benefit. National and regional regulators tasked to ensure that consumers are getting the communications bandwidth, coverage and capability that were advertised and they paid for, are beginning to use crowd-sourced measurements from the edge to provide public "report cards" of communications quality. We have been working with data captured by applications based on the phones, authorized by their users to capture location information and share it to build a public database of internet access performance. We have used most extensively results from an Israeli startup called WeFi, which observes the category of application in use during some of its measurements and determines upload and download data volumes and rates. We have, in all, about 3 billion measurements from five US cities and their surroundings, Atlanta, Boston, Brooklyn, Los Angeles and San Francisco, for several months in each location during 2014 and 2015.

Our data has been presented in several publications that address issues in mobile network planning and management. One surprising result is the range over which each cell antenna is received. Earlier studies in which the data source is a carrier have used the cell tower locations as a proxy for user location. In this study, which sees all carriers in each city, we first estimated the locations of the cell towers, which were named in each measurement record, as the centroid of all the observations where a tower was seen. Cell dimensions of several km are observed. By contrast, the phone locations are known to GPS accuracy, with errors as little as 10m.

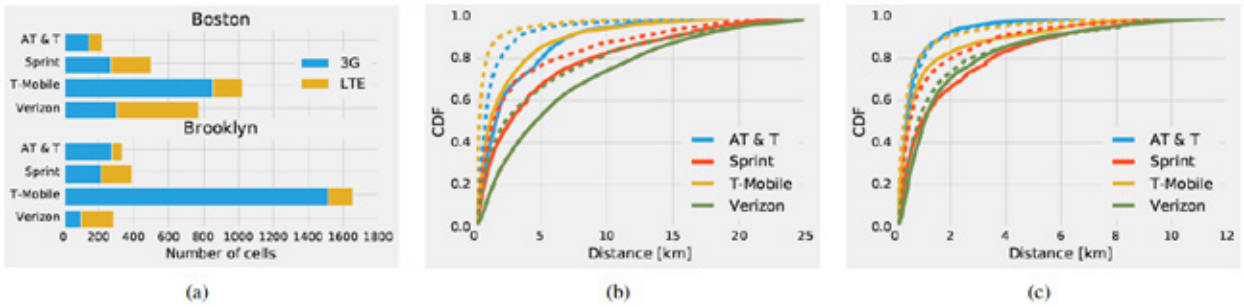


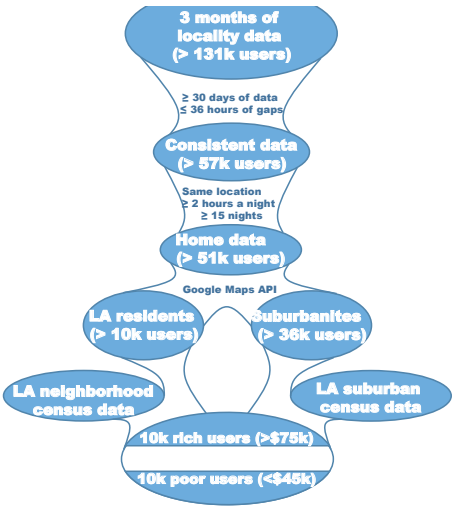
Fig. 1. Number of LTE and 3G cells deployed by each operator (a); distance from user to cell tower in Boston (b) and Brooklyn (c). Solid lines are LTE connections, dashed lines are 3G.

In this report we use unbiased observations made in the background in the course of the daily lives of over a hundred thousand people in and around Los Angeles, sampling roughly 1 per cent of the population, from all economic levels and demographics. The WeFi application monitors location, data connections, and application usage on Android smartphones, but does not capture any content exchanged or any information relating to phone calls. Personally identifying information is removed from the data by hashing the identifiers of the phones. We observe the activities of individual phones, but collect them into aggregated communities sufficient in number to prevent re-identification of individuals. One purpose of this study is to understand how much data is required in order to observe the social behaviors relevant to well-functioning cities. One observation is that the more data we can consider, the finer the scale which we can study without danger of compromising individuals' privacy.

Measurements based on the cellphone can occur whenever there is activity, either data transfer or motion of the user, and thus are much more frequent than the monitoring normally seen in datasets which record the metadata of phone calls. In mobile CDRs and related carrier data the towers see each phone typically 5-20 times a day, with phones in cars seen more often as they change towers during a call. Some of the WeFi data is taken much more frequently. The Android operating system allows the WeFi app to request a measurement when the phone position changes by as little as .0001 in the Lat or Lon coordinate, or on any change in the system's connectivity. As a result we are presented with position information as often as more than a thousand times an hour.

Using Location Data

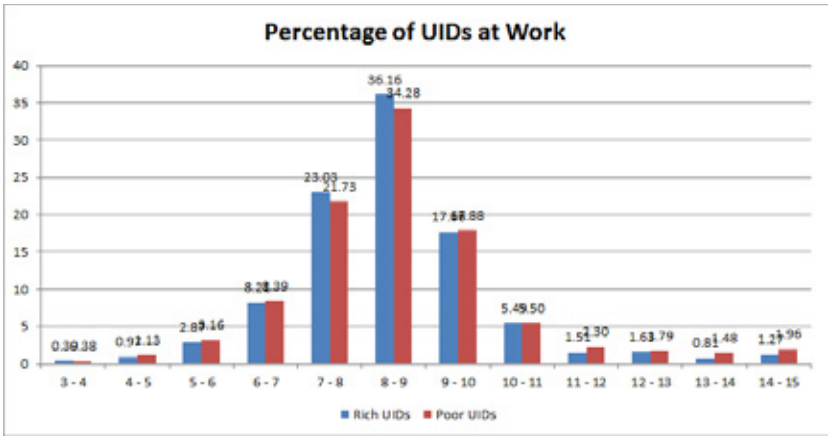
In employing this data for more traditional social science ends we follow a methodology of successive reduction to isolate distinctive communities, then extract their characteristic patterns of commuting, working, shopping, and leisure activities. Our filters are simple and fairly strong. We have used two data sets. In the first, we observe over 131K users, recording their position as lat/lon to a precision of .0001, with a timestamp giving days, minutes and seconds. The 835M measurements in this data set take up about 20 GB. We refer to this data as the location data set. It was gathered during March through May 2015,. The second data set, collected in February, 2015 consisted of 130 GB with 422M measurements, each containing location information plus details of the applications in use, the data connection used, and the amounts of data uploaded or downloaded since the previous measurement. We refer to this data as the application data set. Because these measurements were grouped an hour at a time to simplify their retrieval, the time stamps were only given to the precision of an hour. The application data set contained over 91K distinct users. Each user is only known to us by a random hash of the machine identification of their Android phone. The same hash function was used in



gathering the two data sets, so we can determine that over 67K of the UIDs (the user's randomized identifier) are present in both data sets. This allows us to combine information gathered in each of these two ways. We start our characterization of interesting communities with the location data set.We refine the 57K UIDs by determining the towns or neighborhoods in which they reside. The criterion applied was that a residence is identified when the UID is seen at the same location to an accuracy of .001 in lat/lon two or more hours a night for 15 or more nights within our sample. For more than 51K of our consistent UIDs we can identify such home locations. More than half of the UIDs (>36K) live outside the city of LA proper, and are called suburbanites in this discussion. However, the city of LA is rather porous. Thus quite a few of our suburban districts are governed as separate towns but are contained within the city of LA. More than 10K UIDs reside in neighborhoods in the city proper. We also omitted some UIDs for which the suburb or neighborhood is not unambiguously defined, or for which the residential area population is <5000 people. This left us with 36,531 users who live in 232 independent towns and 10,573 users who live in 83 different neighborhoods within the city.

Next, using census demographic tables of population and median income, we assign to each UID the median income of their home district. A concern is that usage of smart phones might skew our sample of users within the populations of each home town, but with smart phone usage now passing 60-70% of all mobile telephone customers, we do not think this will cause significant bias. We next identify a population of 10,094 UIDs (and their users) from the bottom of the demographics, living in areas with median income < $45K, and another 9780 UIDs, whose users live in the wealthiest areas, with median incomes > $75K. We will distinguish these "rich" and "poor" users when identifying further details of their activities. The two sets make up 20.8% (the "poor" cohort) and 21.5% (the "rich" cohort) of our total sample of users.

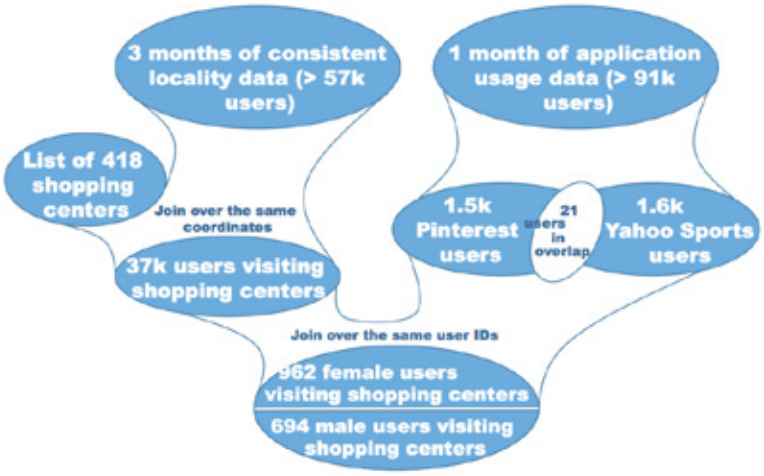
We next determine where our users work. We again start with the users seen more than 30 days, with few gaps, the consistent user set. We identify a stable daytime location, or work location, as a place defined to .001 accuracy in lat and lon, and seen for at least 4 fours per day, on at least 30 workdays.
For almost 25K users, we can find such locations, but for 14K users, this was also their home location. This leaves us with 10,596 commuters, for whom we also know their home location. We still have 21.4% of our sample of users to study. Of these, 2263 (or 20.9% of the commuters) live in our wealthier districts, thus are members of the rich cohort, and 1902 (0r 17.5% of the commuters) live in the poorer neighborhoods. We notice that the second group has dropped by 4% or about one fifth, in their participation, as these members of the poorest cohort do not have fixed locations in which they work during the day. The distribution of hours worked also shows an important difference. Our richer cohort works shorter hours than the poor cohort.



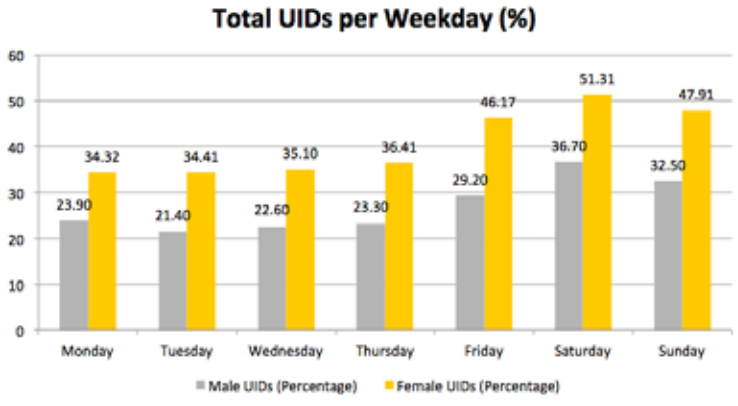
Adding Application Data

To find larger populations and address larger distinctions in behavior, we looked into two things that affect almost everyone: shopping and fast food restaurants. This required some manual effort. Shopping malls were identified with the help of Google Maps API, and resolved into 418 shopping areas. Over 37,000 of our consistent users spent at least 10 minutes and up to 6 hours at one of our shopping areas during the study period. They averaged 1 hour and 7 minutes per visit. Similarly, MacDonald's restaurants were screened to identify 553 with outdoor, separated locations, not inside some larger shopping center. McDonald's visitors spent from 5 minutes to two hours, for an average of 24 minutes in their fast food breaks. In order to know more about who goes to these, when, and why, we need some information about the users' interests. Here the Application data set is useful.

An indicator that serves to differentiate populations are the differences between two applications, Pinterest and Yahoo Sports. Pinterest has been found elsewhere to attract about 85% female users. Frequenting Yahoo Sports for results would seem to give a high probability that the user is a male. Almost all of the users we can separate in this way are seen at some point in our data set, shopping at a mall. The Fig at right summarizes the sizes of the communities that we extract in this fashion. We see 1.5K Pinterest users (each of whom has invoked the app >100 times) and 1.6K Yahoo sports users (also calling for the latest results at least 100 times). Many of these users, about 1000 of the likely female UIDs and almost 700 of the sports fans were seen at one of the 418 shopping centers.



Looking at the activity patterns, we see that our female shoppers are seen at the malls almost three times per week, while the males appear less than twice per week. In the longer work from which these examples are drawn we also separate younger and older users, analyze fast food consumption, commuting times and distances.



We have been able to resolve our data down by two and even three levels of filtering, but only because we have a lot of it. Imagine how much could be safely learned if all phones contributed to this information!

User Authentication with Neural Networks Based on CDR Data

Dominik Filipiak Bartosz Perkowski Agata Filipowska

Department of Information Systems, Poznań University of Economics and Business

The paper describes an attempt to measure the performance of neural networks in the behaviour-driven user authentication task. The goal of this experiment was to minimise the loss of a binary classifiers and check how well neural networks perform while authenticating a user based on his actions, namely calls and messages exchanged).

Introduction

It is possible to identify various patterns based on a frequent behaviour for active users. Daily call duration, a number of text messages sent, BTS stations visited etc. define a user profile. However, are these features enough to tell an identity of a given user? How many activities are needed to perform a robust authentication? These questions constitute the fist research problem addressed within the paper. The second research problem concerns the minimal timespan needed for a robust user authentication (the shorter required, the better). This paper describes an attempt to measure the performance of neural networks applied for these tasks. The goal of this experiment was to minimise the loss of a binary classifier and check how well neural networks perform with a different day granulation while authenticating a user. The higher granulation is, the faster an anomaly can be detected, i.e. after shorter time, what is desirable.

The problem of user authentication in various domains (e.g. telecommunication, social networks) relates to profiling. It may be associated with Knowledge Data Discovery model – many steps in KDD model resemble stages associated with the user profiling process (Kanoje, Girase, & Mukhopadhyay, 2015). Recently, researchers' attention is drawn by an emerging field of behavioural profiling, which can be described as a dynamic profiling approach. The number of domains in which such methods are applicable is constantly growing. The aforementioned research activities can be noticed in the domain of online communities (Fernandez, Scharl, Bontcheva, & Alani, 2014), social networks (Abel, Gao, Houben, & Tao, 2011; Xu, Zhang, Wu, & Yang, 2012) or telecommunication (Hohwald, Frías-Martínez, & Oliver, 2010).

Neural networks have already been used to solve various problems with authentication of users. Bagnall (2015) presented an approach, which helps solving the problem of authentication of authors in documents of an unknown authorship. Another example of employment of neural networks is user authentication in the Tor network (Ishitaki, Oda, & Barolli, 2016).

In this paper, we present an approach to the user authentication based on CDR data mimicking data that may be inferred from phone logs using neural networks.

Methodology and Results

The goal of the presented experiment was to check whether it is possible to develop classifiers capable of authentication of different users based on call and message logs. Each classifier was tested against authenticating a single user

✉ dominik.filipiak@ue.poznan.pl

behaviour in a different day granulation setting.

Data. The sample used in the experiment was a Call Detail Records database, containing actions performed by 100,000 users in six consecutive months. For an average user, the mean number of activities performed is 1150. There are three groups of users: these with an average activity (650-1650 actions, 50 randomly chosen users), high activity (1650-5000 actions, 50 randomly chosen users), very high activity (20 randomly chosen users from 2000 the most active ones). We used 10 input variables, representing accordingly: time of the day, total number of calls, total calls' duration, number of text messages, number of visited base transmission stations (BTS), number of BTS used when calling, number of BTS used when texting, number of different contacts, number of contacts called, and number of contacts texted. The last position in a vector represents whether a given record is a true one or a fake one. Since our data sample consists of (arguably) *true* records, we came up with a simple idea to extend our dataset and enrich with *fake* records. To produce fake data, we simply randomly swapped user labels and combined original daily actions with signatures from an another user. A standard approach in the dataset division was used: for the each user and a given granulation, 70% goes to the training sets, whereas the rest is for the testing process. In calculations, days were partitioned to equal parts in several manners ($d = \{1, 2, 3, 4, 6, 8, 12, 24\}$). For example, $d = 1$ means no granulation at all. All the hours of a given day are placed in a single data row. In other example, $d = 3$ means dividing day into three equal parts with regard to hours (0-8, 8-16, 16-24).

Methodology. The algorithm behind every model is Feed-Forward Artificial Neural Network with Backpropagation as a learning model and the sigmoid and ReLU activation functions, Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) and cross-entropy cost function. Classic Stochastic Gradient Descent algorithm was replaced by more recent Adam optimiser (Kingma & Ba, 2014). Each neural network has six layers, each of which is a densely connected layer (which means that every output of the layer is connected with each input in the next layer). The first layer is an input layer with 10 neurones with Rectified Linear Unit (ReLU) activation, each of which corresponds to a specific input variable (a single daily, weekly, monthly statistics related number). The second layer has 30 neurones with ReLU activation. Dropout technique ($p = 0.5$) was used between the first and second layer. The third one has 50 neurones with ReLU activation. The fourth one has 30 neurones with Sigmoid activation. The fifth one consists of 50 neurones with Sigmoid activation. The output layer has only a one neurone, which value ranges between 0 and 1. The closer to 1, the more probable is the data sample is true (which means). Each model was trained during 250 epochs with 10 items in one batch.

Results. We have tested different 960 models (8 different granulation times for 120 users in three groups). As it turned out, the Dropout technique significantly reduced overfitting, since the difference between training and test accuracy is relatively small. Increasing the d value generally results in a drop in accuracy (see Figure 1). Whereas the best score is reached for $d = 1$ (no day partitioning), the other scores smoothly drop, with no strong tendency to plummet in accuracy. It is worth to mention that the fake data are made-up and the results might be different in the real world usage. On the other hand, such data are extremely hard to obtain in the appropriate amount and this is perhaps the only possible option. However, taking the results into account the data need to be combined e.g. with the data on user trajectories or phone environment to enable for authentication with a higher level of confidence.

References

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Analyzing user

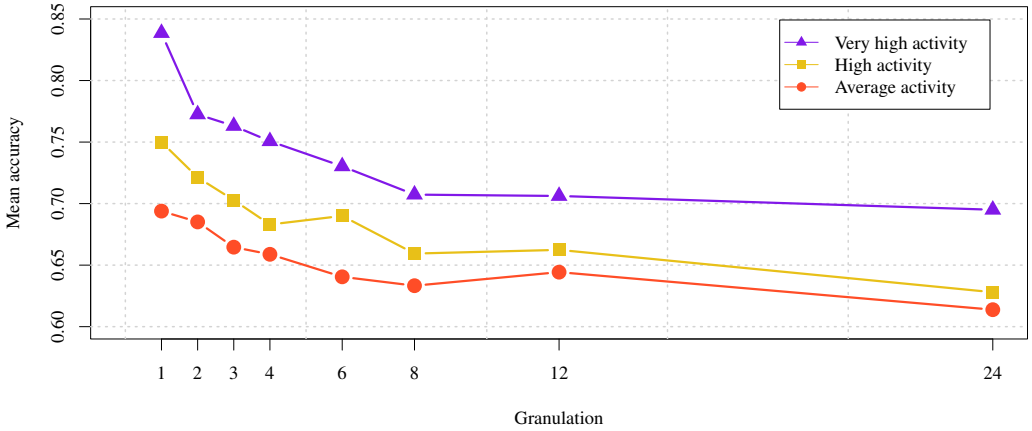


Figure 1. The difference between mean testing accuracies for different granulation (d) values in the tested activity groups

Table 1

Mean accuracy scores for different granulation (d values)

	User activity group		
d	Avg.	High	V. High
1	0.6939	0.7500	0.8386
2	0.6851	0.7213	0.7725
3	0.6646	0.7028	0.7631
4	0.6588	0.6830	0.7507
6	0.6405	0.6900	0.7303
8	0.6333	0.6594	0.7073
12	0.6443	0.6624	0.7062
24	0.6138	0.6280	0.6950

modeling on twitter for personalized news recommendations. In *International conference on user modeling, adaptation, and personalization* (pp. 1–12).

Bagnall, D. (2015). Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

Fernandez, M., Scharl, A., Bontcheva, K., & Alani, H. (2014). User profile modelling in online communities. In *Proceedings of the third international conference on semantic web collaborative spaces-volume 1275* (pp. 1–15).

Hohwald, H., Frías-Martínez, E., & Oliver, N. (2010). User modeling for telecommunication applications: Experiences and practical implications. In *International conference on user modeling, adaptation, and personalization* (pp. 327–338).

Ishitaki, T., Oda, T., & Barolli, L. (2016). A neural network based user identification for tor networks: Data analysis using friedman test. In *2016 30th international conference on advanced information networking and applications workshops (waina)* (pp. 7–13).

Kanoje, S., Girase, S., & Mukhopadhyay, D. (2015). User profiling trends, techniques and applications. *arXiv preprint arXiv:1503.07474*.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Xu, Z., Zhang, Y., Wu, Y., & Yang, Q. (2012). Modeling user posting behavior on social media. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 545–554).

Analysis of Tourist Activity from Cellular Network Data

Marco Mamei
Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
marco.mamei@unimore.it

Massimo Colonna
Engineering & Tilab
Telecom Italia, Italy
massimo.colonna @telecomitalia.it

1. INTRODUCTION

In this work, we explore the use of anonymized Call Detail Records (CDRs) from a cellular network to classify users according to different mobility-behavioral profiles (i.e., residents, commuters, user in-transit, tourists and excursionists). Our focus is primarily on the tourist class and in estimating the tourist presence in an area over a specific period of time.

A number of previous works addressed tourists identification via the analysis of the SIM card registration country and on the basis of the number of days spent in the region [1]. More recent approaches try to detect also national tourists by the classification of their behavioral patterns [2]. In general, classification of users from CDR data is problematic since groundtruth classes are completely missing. Therefore, supervised learning, in which the system learns a classification mechanism from training samples, cannot be applied directly. To tackle this issue two approaches are possible:

1. Apply a clustering algorithm to feature vectors describing users' behavior. Then, attach a label (e.g., tourist) to each cluster on the basis of domain knowledge. This is basically the approach adopted in [2].

2. On the basis of domain knowledge identify class labels for a subset of the feature vectors. Train a supervised classifier on these labeled data. Use the classifier to assign a label to the remaining feature vectors. This is the approach used in this work.

While the two approaches are similar from a functional viewpoint, there are some differences that are at the basis of our contribution: *(i)* supervised classifiers tend to produce better results than clustering algorithms. Moreover, some algorithms (e.g., decision trees) tend to produce more understandable decisions than clustering approaches. *(ii)* From our perspective, it is easier to identify class labels for individual users than to entire clusters. The domain expert in charge of assigning the labels can in fact better analyze the behavior of individual users and establish more precise rules for classifying some of them.

In the following we better describe our training approach and some experimental results.

2. METHODS

We analysed anonymized CDR data comprising SMSs and calls exchanges from multiple tourist cities in Italy. In this abstract we focus on results obtained in Venice. We monitored Venice province (800'000 inhabitants) with the goal of classifying users in Venice historic center (60'000 inhabitants). The monitored period consists of the months of July 2013 and March 2014. The approximate number of CDR users in each period is 600K.

In order to set up a classification approach, we processed CDR data to extract a feature vector compactly describing users' pattern of visit. Specifically, for each user we computed features comprising: *(i)* whether the user is roaming (likely to be foreigner) or not, *(ii)* number of days in which the user generates at least one CDR, *(iii)* number of days in which the user generates at least one CDR from the city (i.e., excluding events generated in the city's neighborhood), *(iv)* number of nights, *(v)* number of weekends, *(vi)* max gap in hours between two subsequent CDR generated in the city

Then we assign profile labels to a subset of the users on the basis of domain knowledge (this will form our training set). Then to train a classifiers on this data and to extend the labeling to all the other users.

Focusing on the standard classes adopted in the analysis of tourists' and visitors' behavior (*www.unwto.org*), we consider 5 main user profiles: **residents:** users living in the city. They show a continuous presence over all the monitored period during the whole day, there included night hours and weekends, **commuters:** people living outside of the city, but working or studying in the city, **people in-transit:** people passing by/driving though the city without stopping for a visit, **tourists:** people traveling to and staying in places outside their usual environment for leisure, business and other purposes, for a limited amount of time and **excursionists:** people visiting the area for leisure, business and other purposes, but who does not stay overnight.

On the basis of domain knowledge, we defined the set

computable rules, and assign profile labels to a subset of the users. For example, we defined Tourists those users in roaming – i.e., foreigner and stay in the city for more that 1 and less than 4 days. We deliberately set up a rather stringent set of rules (e.g., there are also Italian tourists) in order to limit the number of misclassified users. Overall, these rules allow to assign a label to about 30% of the users. This labeled set will be the basis for instrumenting our classification mechanism.

We then set up a supervised classification approach. We used the labeled instances to train and test (in cross validation) the classifier. We then use the trained classifier to assign a label to the users that were not classified with the "domain knowledge" approach described in the previous section. The classifier will generalize the rules to label all the users.

3. EXPERIMENTS

Although the lack of actual groundtruth prevents sound evaluation of our framework, we conducted several experiments to partially evaluate and validate our results.

Classifier Evaluation We run classification experiments on the labeled subset of the users. For computation efficiency we sampled only 5% of the labeled population to create a training set. We trained a model, then used the trained model to classify the whole labeled population.

In Figure 1 we report results comparing different classifiers. We used 1R (i.e., one-level decision tree) as a baseline. Then, we compare C4.5, logistic regression, classification via clustering (using k-means - with 5 clusters - as the clustering algorithm, and also using self-organizing maps - SOM as in [2]). It is possible to see that even the baseline produces good classification results, and more sophisticated algorithms like C4.5 and Logistic regression further improve on that. Viceversa, classification via clustering do not produce good results. On the basis of this result, in the following we adopt C4.5 as our classification mechanism.

Three comments are important with regard to these results: *(i)* These results are obtained with an *artificial* groundtruth. By setting groundtruth with a rule-based procedure like the one described, it is not hard to obtain extremely high accuracies. *(ii)* These results are obtained considering only 5% training data. So results are at least stable across our dataset. *(iii)* The fact that

Venice	C4.5	1R	Logistic	KMeans	SOM
Jul 2013	99.91	91.29	99.94	70.83	81.58
Mar 2014	99.96	91.96	99.5	38.74	61.72

Figure 1: Accuracy with *artificial* groundtruth.

based on clustering supports the approach adopted in this work over clustering-based approaches [2].

Analysis of descriptive statistics. In this section we analyze the resulting classes of users obtained by the classification. Our aim is to find in the statistics describing those classes evidences supporting the goodness of the obtained results. We report some statistics of the users classified as tourists in Venice in July 2013.

Figure 2(a) illustrates how the algorithm classifies users, dividing between ITalian and FOREigner users. Looking at this graphs, it is possible to see that our classifier identifies residents, commuters and people in transit mainly in Italian people (it is worth noting that Venice geographic position very peculiar, and transit through Venice might be biased by cruise ships that visit Venice in the summer). Although the lack of actual groundtruth information impede sound evaluation, the results well conform to commonsense expectations.

Something that might be puzzling at the beginning is the huge number of tourists and excursionists being present. The *wrong* conclusion would be to think that on a typical *day* in Venice there are more tourists and excursionists than residents. In fact, to consider a typical day, we should divide the tourists' and excursionists' numbers by about 15-30 (as they typically visit the city for 1 or 2 days), while leaving the resident column unchanged (as they are always there).

Figure 2(b) illustrates the number of days of presence in the area for a given percentile of tourists. It is possible to see that the majority of tourists remains in the area for less than 2 days, associated to a typical tourist destination (official statistics estimate that tourists stay in the city for 2-3 days).

Figure 2(c) illustrates tourists composition by nationality. It is possible to see that although Italians are the major cut (15%), the remaining 85% of the tourists are likely to be foreigner.

An important aspect of these statistics is that they well illustrate the generalization effect of the classifier. For example, there are indeed Italian tourists and tourists staying in the city for more than 3 days (contrarily to groundtruth rules).

In another analysis we report experiments cross analyzing users from different months and regions. Figure 2(d) compares users from Venice in July 2013 to users from Venice in March 2014. We try to analyze how many users of a given class from July 2013 are in the data also in March 2014. The histogram shows the percentage of users of a given class who are in the city also in the other month. The main assumptions is that most tourists in July 2013 will not visit Venice again in March 2014. Vice versa residents in July 2013 will remain residents also in March 2014. Results are roughly in line

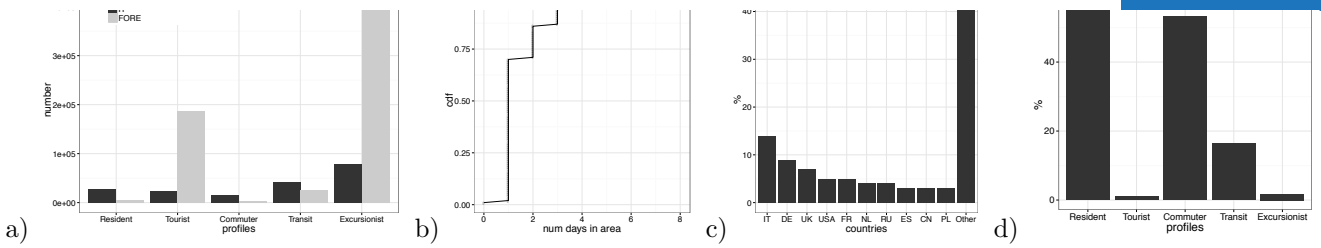


Figure 2: Tourists in Venice in July 2013 . (a) Daily average number of CDR. (b) Number of days in the area. (c) Tourists by country. (d) Cross month analysis: the fraction of users from a given profile in one setting that are also present in the other setting

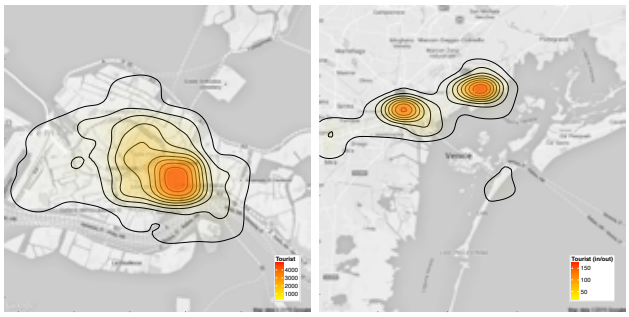


Figure 3: (left) Tourists presence in the area. (right) Tourist entry and exit locations.

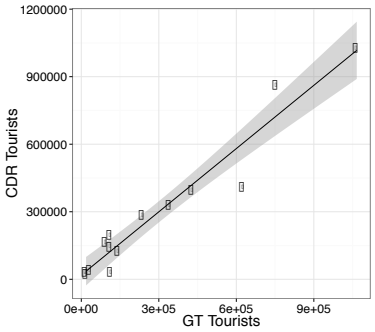


Figure 4: Correlation between official (groundtruth - GT) data and classification results. Pearson = 0.97

with expectations. Only a tiny fraction of July 2013 - tourists and excursionists are in Venice also in March 2014. While a large fraction of residents and commuters are still there.

A careful analysis of these results raises some issues. Considering the "Resident" data, one might ask "*Why the resident bar is not 100%?*". These users are *not* in the data from March 2014. This is rather surprising: they were users who stayed in Venice in July 2013 for almost a month – including nights and weekends, are not there anymore in March 2014. This can be associated to network operator churn and to the fact that our definition of "Resident" includes long-time staying in the city other than persons permanently living in the city. Probably an analysis involving a longer time period (some months) would reduce the percentage of missing residents.

we visualize tourist presence in the area from a spatial perspective. Figure 3(left) shows tourists presence in the area. Figure 3(right) shows tourist entry and exit approximate location (location of the first and of the last CDR in the area). It is possible to see that tourists clusters their stay around Rialto and San Marco areas. It is also possible to see tourists arrivals from Mestre and Venice airport.

Comparison with Other Datasources. In a final experiment, we computed the correlation that exists between the number of tourists from official (i.e., statistical offices) data and our classification. We run our analysis over multiple cities (Venice, Florence, Turin, Lecce) and months and compared the results (see Figure 4). On the basis of this strong correlation, it is possible to create a linear regression model to estimate the number of tourists on the basis of our classification. Fitting the model with least mean square errors, we obtain: $nTourists = 5.39 * estimated + 21.162$ and a median relative absolute error of 27%.

4. REFERENCES

[1] R. Ahas, A. Aasa, A. Roose, U. Mark, S. Silm, Evaluating passive mobile positioning data for tourism surveys: An estonian case study, Tourism Management 23 (3) (2008) 469–486.

[2] B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo, Analysis of gsm calls data for understanding user mobility behavior, in: IEEE Big Data, Santa Clara (CA), USA, 2013.

A neural embedding approach to recommender systems in telecommunication

Nikolaos Lamprou[1], Giovanna Miritello[1]
[1]Vodafone UK

February 6, 2017

Abstract

With the advent of online retail at scale, the increasing demand for improved, personalised and relevant product recommendations has resulted in the development of new exciting techniques in recommender systems. At heart, recommendation tasks rely on some form of similarity measure between products, customers or both. In this work we explore how a product-to-product similarity can be learned by embedding items in a low dimensional space, without any consideration of product or customer details but solely on customer interaction with the products. We then discuss ways of utilising the learned similarities for generating recommendations along with the suitability and limitations of the method.

Different recommendation systems have been developed over time, each with its strengths and weaknesses. Popular methods, such as nearest neighbours or clustering rely on generating quantifiable features based on the characteristics of either the products or the customers, such as screen size or customer age. Product-to-product or customer-to-customer similarity can be computed using these features. However, such methods fail to capture the impact of cultural and social influence, customer predispositions or existing trends. An illustrative example is how unlikely a customer is to switch to a Samsung device when he possesses an iPhone and vice versa, no matter how similar the device features are. More recent approaches, such as collaborative filtering, address this issue by learning a low dimensional embedding of users and items simultaneously. Implicit information about customers and products is captured in the form of latent features. In general, however, recommendations generated by such systems are for specific customers.

The methodology to extend the notion of latent features to a latent representation of products can be borrowed from the field of natural language processing (NLP). Recent research in this field has shown that learning a latent representation of words using neural embedding algorithms is particularly effective in identifying word-word relations and similarities. Although word embeddings were originally introduced by Bengio et al.[1] [2] more than a decade ago, they are one of the most exciting area of research in deep learning at the moment.

A popular algorithm for word embeddings is the Skip-Gram with Negative Sampling (SGNS), commonly known as Word2Vec, a neural word embedding method introduced by Mikolov et. al in [3]. This algorithm builds the embedding using a shallow neural network that is trained to predict a word given its neighbouring words, i.e. its context. The hidden layer of the neural network is a lower dimensional space with respect to the entire word space. Once the neural network is trained, the word embedding is built by using the projection into the hidden layer, where words are transformed into vectors.

This simple embedding is able to generate complex structures in which words similar in meanings and in use are mapped close together. Moreover algebraic relations between words are surprisingly captured such as the famous "king - man + women = queen".

By adapting the SGNS methodology, we construct a framework that produces embedding for items in a latent space. The method is capable of inferring product-to-product relations without considering details about either the product or the end users. A similar approach has been proposed before for computing artist similarity [4] although the nature of the data available is different. Specifically, we have applied this approach to various telco related domains such as the mobile tariff, device and mobile application spaces, to explore whether this technique can shed light on similarities and relations between those products that can be used for recommendations.

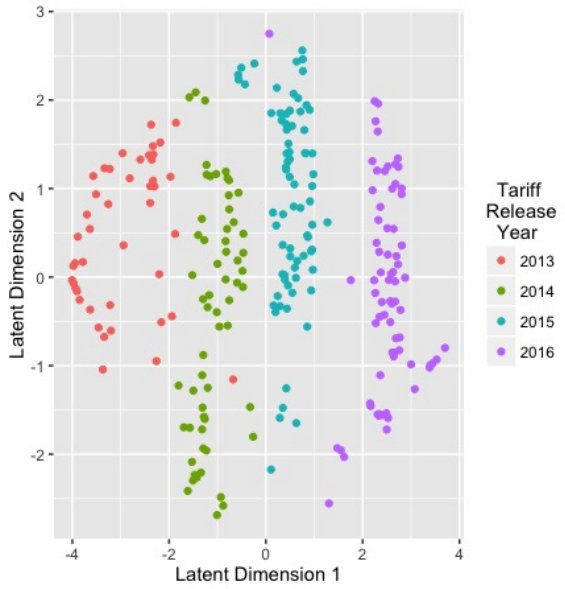


Figure 1: t-SNE embedding for the tariff vectors where each has subsequently been coloured based on release date. Clear clusters are observed to have formed containing tariffs released in different years.

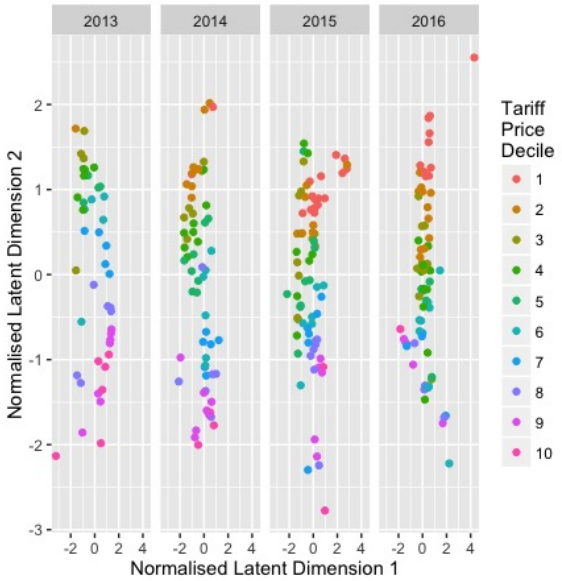


Figure 2: The tariffs are considered separately by release year and coloured according to price. This demonstrates that there is an inherent substructure within the observed main clusters.

In the context of the mobile tariff space, a word corresponds to a tariff and a sentence is made up by the historical transitions a customer has made between different tariffs. Using the corpus built by the overall customers for training the neural network, we create the embedding. Overall, 3.37 million customer transitions were observed between 519 tariffs since 2015.

In order to explore the obtained embeddings we use t-SNE, a powerful dimensionality reduction technique for visualising high-dimensional data. When we inspect these visualisations it becomes apparent that the vectors capture some general, and in fact quite useful, semantic information about tariffs and their relationships to one another. For example, some of the structures that emerge in the tariffs space can be related to price band, type of contract or time since the tariff has become available (see Fig.1 and Fig.2)

Techniques from Word2Vec have been used in some of the recommendation methods to represent text based features. For example, in [5] authors aim to make recommendation to users about which Tumblr blogs to follow by using side features (i.e. likes, re-blogs and tags) as well as past preferences of users. The method does not directly use techniques from natural language processing, but employ Word2Vec to compute vector representation of tags, which are word based features. In [6] authors empirically evaluate three word embedding techniques, namely Latent Semantic Indexing, Random Index- ing and Word2Vec, to make recommendation. They evaluate their proposed method on MovieLens and DBbook datasets. Finally, in [7] authors apply Word2 Vecs skip-gram to recommend check-in locations from Foursquare data to the target users. In this case, non-textual features, namely the past check-ins of the users, are used. They mapped the items in the datasets to textual contents using Wikipedia and used the textual contents for making recommendation.
Our findings on telco related product spaces open up new possibilities of using this embedding technique to generate product recommendations also in the telecommunication world.

References

[1] Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural Probabilistic Language Model. Advances in Neural Information Processing Systems 13 (NIPS'00), MIT Press, 2001.

[2] Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural Probabilistic Language Model. Journal of Machine Learning Research 3 (2003) 1137-1155.

[3] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 2003 (pp. 3111-3119).

[4] Oren Barkan, Noam Koenigstein. Item2Vec: Neural Item Embedding for Collaborative Filtering.

[5] D. Shin, S. Cetintas, and K. Lee, Recommending tumblr blogs to follow with inductive matrix completion, in Poster Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 2014, Foster City, Silicon Valley, CA, USA, October 6-10, 2014, 2014.

[6] C. Musto, G. Semeraro, M. de Gemmis, and P. Lops, Word embed- ding techniques for content-based recommender systems: An empirical evaluation, in Poster Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16, 2015., 2015.

[7] Makbule Gulcin Ozsoy, From Word Embeddings to Item Recommendation arXiv:1601.01356.

Drivers of spatial heterogeneity of HIV prevalence in Senegal: disentangling key features of human activity and mobility

L. Righetto, L. Mari, M. Gatto and R. Casagrandi

The human immunodeficiency virus (HIV) is still the second cause of mortality among infectious diseases at the global scale, with an estimated 1.1 million people dying from AIDS-related illness in 2015; around 36.7 million people are currently infected, with 2.1 new infections in 2015 (UNAIDS Global Statistics 2015). Sub-Saharan Africa alone accounts for more than a half of these figures. Although, in the African scenario, Senegal emerges as one of the countries where HIV has been controlled quite successfully, the spatial variability of the prevalence of this disease is still remarkable. For this reason, in particular, it would be important to single out the main drivers of the diffusion of HIV in the country, including socio-economic, behavioral and cultural factors, as well as human mobility patterns.

In terms of the characterization of spatial processes and social structure, mobile communication data have been used fruitfully in the field of epidemiology. We thus aim to analyse a call detail record (CDR) dataset, acquired via the D4D Senegal Challenge, which was organized by Orange and Sonatel in 2014. We extract an extensive set of spatially aggregated variables that may explain the variability of HIV prevalence at the Health District scale throughout Senegal. These variables encompass several measures of user activity, but they also focus on their spatial displacement and long-term movement. To analyze the data, we plan to use multivariate regression and feature selection methods for problems with high multicollinearity among predictor variables (such as lasso regression). Studies using a similar methodology have been recently performed for Ivory Coast, however they are based on a relatively limited set of mobile communication data (CDRs from the D4D-Ivory Coast challenge 2013).

In addition, we aim to use the available information gathered from the Demographic and Health Survey (DHS) carried out by the United States Agency for International Development (http://www.dhsprogram.com/), in order to bridge knowledge inferred from remotely acquired mobile communication cues with locally collected data on socio-economic structure and behavior. The comparison of these two different data sets is interesting per se, as it provides information of the relationships between communication and mobility features and socio-cultural indicators at local level. When we evaluate their respective significance in defining patterns of HIV prevalence, we also assess whether at least some of the CDR-extracted information can be actually more informative than social, economic and cultural indicators.

To the best of our knowledge, such an endeavour has never been undertaken so far in the context of HIV/AIDS research. Therefore, it might provide essential information on the large-scale drivers of disease transmission, especially regarding the controversial role of human mobility in defining patterns of HIV prevalence. This might help guide future efforts in disease control in Senegal and in other disease-stricken countries.

Automatic stress assessment using smartphone interaction data

Matteo Ciman and Katarzyna Wac
Center for Informatics, University of Geneva
1227 Carouge, Switzerland
Email: name.forename@unige.ch

Abstract—An increasing presence of stress in people' lives has motivated much research efforts focusing on continuous stress assessment methods in individuals, leveraging smartphones and wearable devices. These methods have several drawbacks, i.e., they use invasive external devices or several privacy-related information. This paper presents an approach for stress assessment based on smartphone usage analysis, evaluated with 25 participants, with a final F-measure of 77-88% for a within-subject model and 63-83% with a global population model.

Index Terms—Human-smartphone interaction, stress, smartphone, affective computing, mobile sensing, pervasive computing.

1 Introduction

In the last few decades, the amount of stress experienced by individuals in daily life has dramatically increased. Stress has been strongly correlated with several chronic diseases and health risks, such as hypertension and coronary artery disease [1], cardiovascular disease [2], diabetes and obesity [3]. It is correlated with Heart Rate (HR), Respiration Rate (RR) and Galvanic Skin Response (GSR) values [4] [5]. The development of mobile wearable devices permits to continuously and unobtrusively monitor these psycho-physiological parameters of individuals, enabling practitioners to acquire objective data of their patients [6] [7] [8], without relying on self-reports that are intrinsically biased and prone to errors [9]. However, wearable devices may be still perceived as burden for individuals, hence not being used in regular practice.

In this paper we present a method and associated computational model for stress assessment based solely on human-smartphone interaction analysis, e.g., which does not rely on invasive wearable external sensors, but only on the smartphone data. This method analyses data like which kind of applications are used, for how long, what are the interactions with the screen, etc. We involved 25 participants for a 4-weeks long data acquisition, using Experience Sampling Method (ESM) to gather information about their stress level. The classification models achieved an average F-measure of 77-88% when building a within-subject computational model, or between 63-83% when building a global one.

2 Stress Assessment In-The-Wild

2.1 Methodology

We collected data about daily life routine of the participants, to analyze their behavior in a free environment, without any constraints or particular tasks to perform, installing a stress assessment service on their own smartphone. This service runs in background and gathers information about all the interactions and events that take place on the smartphone. Collected information include:

- events related to the screen (when it is turned on or off, unlocked, etc.);
- the applications used and each interaction duration;
- the physical activity performed by the person (running, walking, etc., using the Google API);
- the number of touches on the screen when the smartphone is unlocked and
- the value of the lightning ambient condition measured by the phone's light sensor.

The data collection lasted four weeks. During this month, we used the Experience Sampling Method (ESM) [10] to collect data about the participants' stress condition, with a notification prompted on the smartphone asking to evaluate stress level in a 5-Likert scale. ESMs were provided every three hours, from 10.30 AM to 8.30 PM, asking for an evaluation of the past three hours. In total, 25 individuals affiliated with the Quality of Life (QoL) Living Lab [11] participated in the experiment and provided their data.

2.2 Results

We analyzed the collected data in two different ways. The first one (Section 2.2.1) evaluated the statistical significance of the extracted features for the different stress levels. The second one (Section 2.2.2) evaluated the accuracy of a classification model build with the collected data and the extracted features. For every feature, we calculated both the average and the standard deviation values for each 3-hours time range. Features values were normalized with a linear transformation inside the interval]0, 1], to combine in the global model values from different participants. The total number of features was 182.

Participants that took part in the data collection provided in total 1630 answers (over a maximum of 2800, representing the 58% of the expected value). On average, each

TABLE 1. DISCRIMINATION POWER: DURATION OF *OnOffScreen* EVENTS.

Stress Level	1	2	3	4	5
1	-	25%	24%	28%	50%
2	-	-	8%	12%	50%
3	-	-	-	6%	50%
4	-	-	-	-	33%
5	-	-	-	-	-

TABLE 2. DISCRIMINATION POWER: DURATION OF *UnlockedScree* EVENTS

Stress Level	1	2	3	4	5
1	-	21%	15%	33%	60%
2	-	-	12%	25%	20%
3	-	-	-	20%	40%
4	-	-	-	-	25%
5	-	-	-	-	-

TABLE 3. DISCRIMINATION POWER: DURATION OF THE USAGE SESSIO

Stress Level	1	2	3	4	5
1	-	25%	15%	21%	60%
2	-	-	4%	22%	40%
3	-	-	-	17%	40%
4	-	-	-	-	25%
5	-	-	-	-	-

participant answered 65.20 ± 20.76 times. The response rate is slightly over 50%. Even if this percentage could appear low, it is necessary to remind that we did not impose to participants to answer each ESM, in order to avoid stressing them, thus collecting biased answers. Each answer to an ESM is the evaluation of the perceived stress, and relates to the previous three hours of the day. Each 3-hours range contains different *usage sessions* (the number changes depending on the behavior of the individual), and other data recorded in the files presented in the previous section. Answers provided by participants were distributed as follows, where 1 means "Very Relaxed" and 5 "Very Stressed": 1=> 46.44% (757/1630); 2 => 32.76% (534/1630); 3 => 16.01% (261/1630); 4 => 4.11% (67/1630); 5 => 0.67% (11/1630).

As we can see, the collected data is strongly unbalanced. This unbalance can reduce performances in the classification task, hence provide biased models. We applied the SMOTE filter [12] to the dataset to resolve the problem. This filter, used before training the classifier, is used to increase the number of instances of the less occurring classes. In this way, we add random examples to the less frequent classes, to obtain a final dataset that is balanced and to perform a more precise training. We applied the SMOTE filter until the difference between the number of instances of all the classes was less than 10. Therefore, we have added 4 instances for class 1, 225 instances for class 2, 499 instances for class 3, 689 instances for class 4 and 749 instances for class 5. The final distribution was: 1: 761, 2: 759, 3: 760, 4: 756, 5: 760. We defined a 3-class problem, where class 1 and class 2 were combined in "Relaxed", class 3 in "Normal" and class 4 and class 5 in "Stressed". The final distribution of the instances was: "Relaxed": 1296, "Normal": 1288 and "Stressed": 1289.

2.2.1 Statistical Analysis

The first step of data analysis aimed at understanding which features alone are statistically significant for stress condition of the user. A two-sample, one-tailed t-test was performed for each of the calculated features (significance value set at $p = 0.05$). To obtain a broad overview of the significance power, we evaluated all the features against all the possible stress levels ($1 \div 5$, thus not considering the 3-class problem). We present the results in the form of a table, where each column and each row represents a stress value. Calling i the index for the rows and j the index for the column, the cell x_{ij} indicates the percentage of participants for which that particular feature was statistically significant, i.e., had a discriminatory power for stress state assessment. The analysis was made comparing all the possible stress levels reported by participants.

percentages of discrimination are achieved when analyzing statistical significance between stress level 1 and 5, or between stress level 2 and 5. This means that, given the extremity ranks of stress level, there is a statistically significant difference on how people interact with their smartphone, i.e., how many times they turn on and off the screen without unlocking it (Table 1) or how much time they use the smartphone (Table 2 and Table 3), or for how much time the Social application category is used (Table 4).

2.2.2 Classification Task Analysis

We used all the features together to build a model for stress assessment. We used standard classification algorithms: Decision Tree (DT), k-Nearest Neighborhood (kNN), Support Vector Machine (SVM) and Neural Network (NN) algorithms available with the Weka Software [13]. We compared the results of these algorithms with the baseline calculated using the ZeroR classifier, which merely predicts the majority category. We used 10-Fold Cross-Validation to evaluate the within-subject model, and Leave-One-Out for the global one, and we calculated the micro F-measure to evaluate the goodness of the classification model [14]. In Tables 5 and 6 and Figure 1 we present the micro F-measure of the classification task for the 5-class and the 3-class classification problem. All the classification performances achieved by the learning algorithms outperformed the results of the ZeroR algorithm baseline. The results show that for the 5-class classification problem, it is possible to assess stress level with an F-measure on average between 77-80% for the within subject model, or between 69-74% for the global model. On the other side, with the 3-class classification problem, F-measure is on average between 86-88% for the within subject model, or between 63-83% for the global model.

TABLE 4. DISCRIMINATION POWER: TIMING OF SOCIAL APPLICATIO CATEGORY

Stress Level	1	2	3	4	5
1	-	17%	21%	15%	33%
2	-	-	9%	8%	67%
3	-	-	-	23%	67%
4	-	-	-	-	100%
5	-	-	-	-	-

TABLE 5. F-MEASURE FOR THE 5-CLASS PROBLEM.

TABLE 6. F-MEASURE FOR THE 3-CLASS PROBLEM.

Model	ZeroR	DT	kNN	SVM	NN
Average	0.39 ± 0.09	0.86 ± 0.12	0.88 ± 0.09	0.86 ± 0.09	0.88 ± 0.09
Global	0.39	0.81	0.83	0.63	0.70

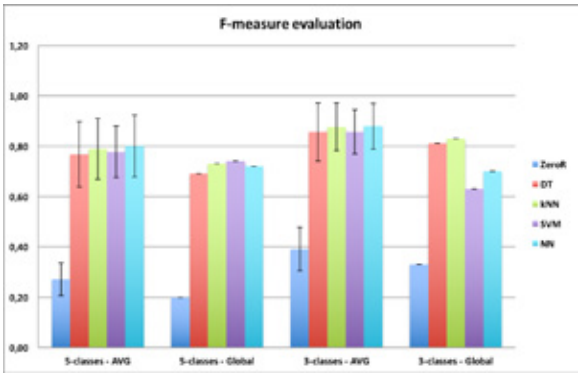


Figure 1. F-measure of the algorithms for all the classification models.

To create a rank of the classification algorithm and to understand which is the one with the best performances, a paired, one-tailed t-test was performed between all the classification models for all the participants, both for the 5-class and the 3-class classification problem, with the adjustment of the p-value for multiple comparisons. For the 5-class problem, NN was statistically significantly better than DT (p-value=0.004) and than SVM (p-value=0.03). All the other models were not statistically different. For the 3-class classification problem, NN was significantly better than SVM (p-value=0.009) and kNN was significantly better than SVM (p-value=0.04). Looking for a generalization of these results, NN and kNN are the two learning algorithms with better performances, followed by DT and SVM.

Finally, we analysed the most predictive features when building the classification models, evaluating the *Information Gain Ratio*. Table 7 reports the five most predictive features both for the 5-class and the 3-class classification task. Casual, Word and Puzzle relate to mobile game applications, while Music&Audio relates to applications used to listen to music.

TABLE 7. MOST SIGNIFICANT FEATURES FOR THE CLASSIFICATION.

Rank	5-class problem	3-class problem
1	Casual Influence	Word Influence
2	Casual Timing Influence	Score sum activity
3	Puzzle Influence	Music&Audio Timing Influence
4	Word Influence	Word Timing Influence
5	Word Timing Influence	Music&Audio Influence Influence

This ranking shows that the type of application individuals use is strongly correlated with their stress level. Decision Tree construction gives an idea of how stress is related to feature values and individual's behavior. In particular, it is possible to notice that the higher frequency usage of the Word gaming application category is associated with lower stress values. The same happens with the amount of physical activity performed during the day, where a smaller amount is frequently associated with higher stress values. Moreover, lower Puzzle game application category frequency, in combination with low usage of Word and Casual gaming application categories, is associated with higher stress levels.

3 CONCLUSIONS AND FUTURE WORKS

The presence and influence of stress in individuals' life strongly increased in the last years. Prolonged exposure to stress situations can lead to several adverse effects, e.g., cardiovascular diseases, diabetes, and obesity. In this paper we propose a method for stress assessment that relies only on the analysis of how people use the smartphone, and do not acquire personal, privacy-sensitive information. The models reached an average F-measure between 77-88% for the within-subject model, or between 63-83% for the global model. Most predictive features for stress related to the use of games applications (Word, Puzzle and Casual) and Music&Audio applications. Given the conducted experiments, we showed that it is possible to reach accurate classification for stress assessment using only interaction data and without relying on privacy-related or invasive information, thus potentially increasing user acceptance. In the future, we plan to develop a background service that continuously and in real-time assesses stress state of the user.

REFERENCES

[1] T. Pickering, "Mental stress as a causal factor in the development of hypertension and cardiovascular disease," *Current Hypertension Reports*, vol. 3, 2001.
[2] P. L. Schnall, P. A. Landsbergis, and D. Baker, "Job strain and cardiovascular disease," *Annual review of public health*, 1994.
[3] M. F. Dallman, N. Pecoraro, S. F. Akana, S. E. la Fleur, F. Gomez, H. Houshyar, M. E. Bell, S. Bhatnagar, K. D. Laugero, and S. Manalo, "Chronic stress and obesity: A new view of "comfort food"," *Proceedings of the National Academy of Sciences*, 2003.
[4] A. de Santos Sierra, C. Avila, J. Guerra Casanova, and G. del Pozo, "A Stress-Detection System Based on Physiological Signals and Fuzzy Logic," *IEEE Transactions on Industrial Electronics*, 2011.
[5] T. G. Vrijkotte, L. J. Van Doornen, and E. J. De Geus, "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability," *Hypertension*, 2000.
[6] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra Short Term Analysis of Heart Rate Variability for Monitoring Mental Stress in Mobile Settings," in *Proceedings of the IEEE EMBS*, 2007.
[7] A. Muaremi, B. Arnrich, and G. Trster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, 2013.
[8] E. Jovanov, A. O'Donnell Lords, D. Raskovic, P. Cox, R. Adhami, and F. Andrasik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE EMB Magazine*, 2003.
[9] R. A. Bryant, M. L. Moulds, and R. M. Guthrie, "Acute stress disorder scale: a self-report measure of acute stress disorder." *Psychological Assessment*.
[10] R. Larson and Csikszentmihalyi, "The experience sampling method," *New Directions for Methodology of Social & Behavioral Science*, 1983.
[11] K. Wac, M. Gustarini, J. Marchanoff, M. A. Fanourakis, C. Tsiourti, M. Ciman, J. Hausmann, and G. P. Lorente, "mQoL: Experiences of the 'Mobile Communications and Computing for Quality of Life' Living Lab," in *Proceedings of The Living Lab*.
[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 2002.
[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, 2009.
[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management: an international Journal*, 2009.

A New Point Process Model for the Spatial Distribution of Cell Towers

Ezequiel Fattori*, Pablo Groisman*, Martin Minnoni† Carlos Sarraute†
*FCEyN, Universidad de Buenos Aires, Argentina. Email: *ezequieltori@hotmail.com, pgroisma@dm.uba.ar*
†Grandata Labs, 550 15th Street, San Francisco, CA, USA. Email: {*martin, charles*}*@grandata.com*

***Abstract*—We introduce a model for the spatial distribution of cell towers in the city of Buenos Aires (CABA). After showing that the Complete Spatial Randomness (homogeneous Poisson distribution) hypothesis does not hold, we propose a model in which each site is distributed according to a bivariate Gaussian variable with mean given by the barycenter of its neighbors in the Delaunay triangulation. We show that this model is suitable, and can be used to generate a synthetic distribution of cell towers.**

I. Introduction

The analysis of mobile phone datasets, which contain information about the mobility and interactions of individuals at unprecedented scales, is a recent field of study with published works starting in 2005 and an increasing number of publications from 2012 onwards [1].

An important problem of the data collection process is the preservation of subscriber privacy. Mobile traffic datasets contains sensitive information on subscribers, whose privacy needs to be protected [2]. Limitations to the current approaches to the anonymization of datasets are shown by [3].

One promising approach to avoid privacy issues is to generate synthetic datasets. Realistic records that do not contain personal information can be freely distributed, and used by any entity needing to perform analysis [4]. In this work, we tackle the problem of generating a synthetic distribution of *cell towers* (also called *cell sites* or *base stations*), which can be used as a foundation for a synthetic mobile phone traffic generator.

Several articles deal with the complete spatial randomness hypothesis [5], [6] to model the distribution of cell sites, mostly due to its tractability. Although, it has been shown that this kind of point processes is not suitable in realistic situations [7]. Typical alternatives suggest the use of Matérn processes, determinantal processes, Poisson hard-core processes, Strauss processes, and Perturbed triangular lattice models, but few have been contrasted with real data [7].

In this article we use real data to test the homogeneous as well as non-homogeneous *Poisson point process* (PPP) hypothesis and we conclude that none of them are realistic models. We propose a new model based on the Harmonic deformation of the Delaunay triangulation of a PPP [8] and we contrast the model with real data. This model turns out to be suitable.

II. Complete Spatial Randomness

The mathematical object used to model the distribution of cell towers is a *point process*. A point process is a random variable $X: \Omega \to \mathcal{N}$ that draws finite points in each bounded region of the space. When we face an instance of a point process we compare it with a benchmark model which exhibits *complete spatial randomness* (CSR). Given a point process X and a subset $S \subset \mathbb{R}^2$, $N(S)$ denotes the number of points of X contained in S and $\mu(S) := E(N(S))$ is the *intensity measure*. We say that X is distributed according to a PPP of intensity $\rho: \mathbb{R} \to \mathbb{R}_{\geq 0}$ if the following properties are satisfied:

1) $\mu(S) = \int_S \rho(x)\,dx$ for any square $S = [a, b] \times [c, d]$.
2) For any square S with $\mu(S) < \infty$, $N(S) \sim \mathcal{P}(\mu(S))$. The number of points in S is distributed according to a Poisson distribution with mean $\mu(S)$.
3) For any n and any square S with $0 < \mu(S) < \infty$, conditional on $N(S) = n$, the n points within S are independent and have density $\rho(\cdot)/\mu(S)$.

When $\rho(x) \equiv \rho > 0$ is a constant function we say that X is an homogeneous Poisson point process.

III. Nonparametric Estimation of Intensity Functions

In the case of a homogeneous Poisson process in a bounded set S, there is an unbiased estimate of its intensity, the maximum likelihood estimate given by $\hat{\rho} = N(S)/Area(S)$. The case of our network is different, since a greater concentration of cell sites is observed at least in the city centre. When homogeneity is suspected not to hold, a non-parametric kernel estimate of the intensity function should be used:

$$\hat{\rho}(x) = \sum_{\zeta \in X} k_b(x - \zeta) \tag{1}$$

Here k_b represents a volume preserving scaling of a *kernel function* k (which could be a multivariate Gaussian).

This estimate is usually sensitive to the choice of the bandwidth b, while the choice of k is less important. We choose the bandwidth b in order to improve the estimation. The distance $\|\rho - \hat{\rho}_b\|_2$ which measures the dissimilarity between the real intensity ρ and the estimate $\hat{\rho}_b$ is to be minimized. A good choice for b is the one that minimizes the following quantity:

$$L(\hat{\rho}_b) = \|\hat{\rho}_b\|_2^2 - 2\int_{\mathbb{R}^2} \rho(x)\hat{\rho}_b(x)dx. \tag{2}$$

Since this formula involves ρ, which is unknown, it can only be estimated. The estimator will be:

$$\hat{L}(\hat{\rho}_b) = \frac{1}{|X|}\sum_{\zeta \in X}\left[\|\hat{\rho}_b^{\zeta}\|_2^2 - 2\hat{\rho}_b^{\zeta}(\zeta)\right] \tag{3}$$

where $\hat{\rho}_b^{\zeta}$ denotes the intensity estimator built from $X - \{\zeta\}$. This estimator is known as *Left One Out Cross Validation*. The variance and the bias of the estimator can be approximately calculated by bootstrapping. We will settle for simply choosing the value of b with less estimated risk.

Our experiments show that the estimated optimal value of b lies around 1.6 in the chosen scale units. Fig. 1 displays three different bandwidth choices for the city of Buenos Aires.

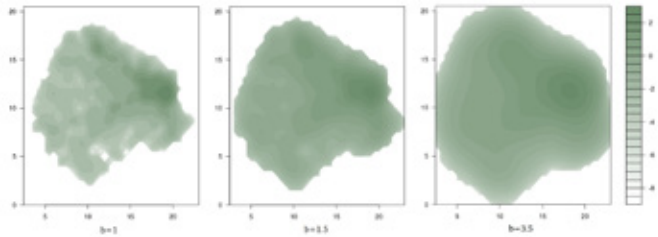


Fig. 1. Contour plots of intensity function $\hat{\rho}_b$ in logaritmic scale.

IV. Simulation of Poisson Point Processes

By simulation of a point process, we understand the generation of an instance of the process. For the same process there may exist various possible algorithms that simulate it.

The simplest case is the simulation of a homogeneous Poisson process in a box. Let ρ be its intensity and $[a,b]\times[c,d]$ a box in $\mathbb{R}^2$. The number of point occurrences must be first drawn from a Poisson distribution with mean $\rho.(b-a).(d-c)$. Once the number n of occurrences is established, the process is the result of putting together n independent draws from the uniform distribution $U([a,b]\times[c,d])$.

For the simulation of non-homogeneous point process, the box is first partitioned into a grid of $m\times m$ rectangles, in each of whom the intensity function is assumed to be approximately constant. These rectangles form $m\times m$ independent Poisson processes, each of whom is simulated like the case handled above.

V. Complete Spatial Randomness Testing

Testing the CSR hypothesis is an important part of the analysis. If the hypothesis is accepted, then it is not possible to find interesting interactions between the points based on the geometry observed. This analysis provides also information on the direction of the deviation from CSR.

In our case, we suspect that the process of cell sites is not CSR, because it has a clear non-homogeneity. It is still necessary to determine if the non-homogeneous Poisson hypothesis holds. Under the Poisson hypothesis, if we apply a *thinning transformation* to the process, the resulting process should be Poisson homogeneous, that is, CSR. Then we can apply a test for CSR, and if the test rejects CSR for the transformed data, then it also rejects the Poisson hypothesis for the original data.

A thinning transformation consists of a random drop of points of the process. A function $p(x)$ with range $[0,1]$ is defined, and each point x is dropped with probability $1-p(x)$.

Fig. 2. Distribution of cellphone sites before and after homogenization.

A thinned Poisson process is also a Poisson process. If the intensity of the original process is $\lambda(x)$, the new intensity becomes $\lambda(x).p(x)$. So that if we apply $p(x) = 1/\lambda(x)$, assuming $\lambda(x) > 1$ everywhere, the process becomes homogeneous.

Many statistics for testing CSR are available, one is *Ripley's K function*. The K function is defined as $K(t) = \lambda^{-1}E[N_0(t)]$ where $N_0(t)$ represents the total count of points of X at distance less than t from the origin, which in the Poisson homogeneous case should be the same as $N_p(t)$ for any p.

To correct biases caused by edge-effects, we use the following estimate due to Ripley:

$$\hat{K}(t) = \frac{1}{\lambda n}\sum_i\sum_{j\neq i} I(r_{ij} < t)w_{ij}^{-1} \qquad (4)$$

where $w(x,r)$ stands for the proportion of the disk $D(x,r)$ contained in S, and w_{ij} represents $w(x_i,r_{ij})$. The K function of our *thinned* process is compared with the (theoretically calculated) expectation of a Poisson homogeneous process.

Our experiments show that the process lacks small distances between points, which indicates a departure from the Poisson hypothesis. Our process shows *repulsion* between points.

VI. Harmonic Deformation of Delaunay Triangulation

Ferrari, Groisman and Grisi proved the existence and constructed a point process model that fits well the distribution of cell sites [8].

Let X be an homogeneous PPP in $\mathbb{R}^2$. The *Voronoi cell* associated to each point p of X is the region made up of all points in $\mathbb{R}^2$ which are closer to p than to any other point of X. The *Voronoi neighbors* of p are all the points of X owners of the adjacent cells. The *Delaunay triangulation* of X is the graph built from all points of X and all edges made up of pairs of Voronoi neighbors.

Let $H : X \to \mathbb{R}^2$ be a function such that for each p in X, $H(p)$ is located at the barycenter of $\{H(q)|q \leftrightarrow p\}$. Here $q \leftrightarrow p$ denotes that p and q are Delaunay neighbors. Such a function is called a "Harmonic function", and its existence was proved in [8]. It represents a position change made to every point of X such that every point is relocated at the barycenter of its, also relocated, Delaunay neighbors (as illustrated by Fig. 3).

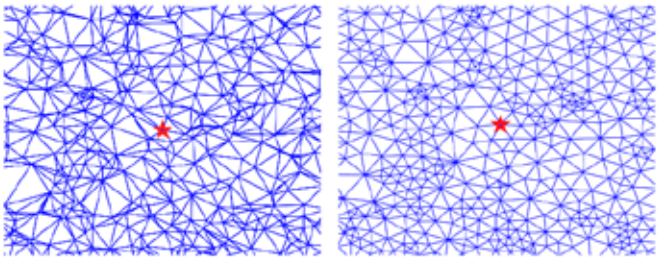


Fig. 3. Delaunay triangulation of an homogeneous PPP (left) and its harmonic deformation (right). The star indicates the origin (left) and $H(0)$ (right).

We simulate a non-homogeneous PPP according to our estimated intensity $\hat{\rho}$ for the city of Buenos Aires, calculate its Delaunay triangulation and finally apply a function similar to H to locate every point at the barycenter of its neighbors. Since our region is bounded, it is necessary to allow some points to stay fixed by H.

Which points should be kept fixed? Ultimately, we would like to perform the harmonic deformation over our bounded region S, without fixing any point and without altering the original intensity function. Our proposed solution is therefore to extend the intensity function smoothly over a box B containing region S; simulate a PPP according to this intensity in $B \setminus S$; join these points with the ones in S; and finally perform an harmonic deformation in B leaving the points in $B \setminus S$ fixed.

In order to smoothly extend the intensity function, we divide the box B into a grid and assign to each intersection of grid lines a variable ρ_{ij}. We generate and solve a system of linear equations to obtain the extended function (see Fig. 4).

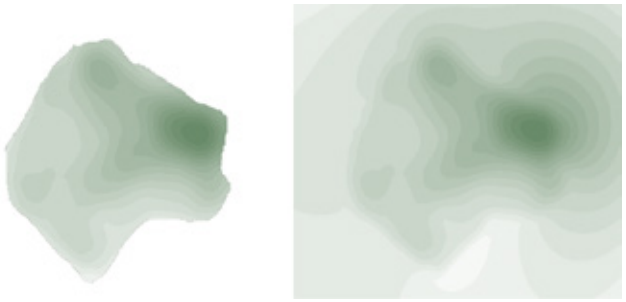


Fig. 4. Smooth extension of the intensity function to a box containing the city of Buenos Aires.

The harmonic deformation can be achieved by an iterative procedure. First the Delaunay triangulation is calculated for the whole box B. Secondly the points within S are copied into a sequence. We iterate over this sequence, in each step relocating the present element at the barycenter of its neighbors. This iteration is performed until a prescribed degree of convergence is achieved.

This model is better than the Poisson model but not realistic enough. In the actual distribution, each cell tower is not located exactly at the barycenter of its neighbors. In order to make the model fit better, we add a *noise parameter* ϵ so that for every $x \in X$, conditioned on the position of its neighbors, the position of x is Gaussian with mean the barycenter μ of the neighbors and variance ϵ.

Fig. 5 displays the final result and assessment via the Ripley's K function. As can be seen, there is no significant deviation between the K function of the actual distribution of cell towers and the final simulated model.

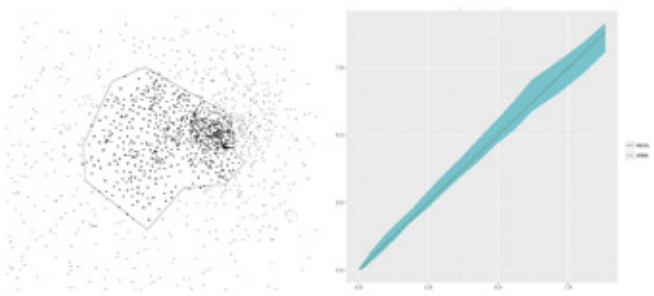


Fig. 5. Example of the final model (left). Upper and lower envelopes of Ripleys K function for 100 simulated (and homogenized) processes from the harmonic deformation with noise model (right).

VII. Conclusions

Using real cell tower location data, we showed that a Poisson point process is not a realistic model for the distribution of cell sites in cellular phone networks. We proposed an alternative point process which fits the real distribution.

We used real data for the location of cell towers in the city of Buenos Aires to show that this model is adequate. This model can be used to generate a realistic spatial distribution of cell sites, or to simulate the growth of the network. The generated distribution contains no sensitive or proprietary information, and can thus be freely shared with research groups, fostering further research on the subject.

References

[1] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. *Mobile Traffic Analysis: a Survey*. PhD thesis, Université de Lyon; INRIA Grenoble-Rhône-Alpes; INSA Lyon; CNR-IEIIT, 2015.

[2] President's Council of Advisors on Science and Technology. Big Data and Privacy: A Technological Perspective. Technical report, Executive Office of the President, 5 2014.

[3] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[4] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Kolar Purushothama Naveen, and Carlos Sarraute. Measurement-driven mobile data traffic modeling in a large metropolitan area. In *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*, pages 230–235. IEEE, 2015.

[5] Martin Haenggi, Jeffrey G Andrews, François Baccelli, Olivier Dousse, and Massimo Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *Selected Areas in Communications, IEEE Journal on*, 27(7):1029–1046, 2009.

[6] Martin Haenggi. *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

[7] Anjin Guo and Martin Haenggi. Spatial stochastic models and metrics for the structure of base stations in cellular networks. *Wireless Communications, IEEE Transactions on*, 12(11):5800–5812, 2013.

[8] Pablo A Ferrari, Rafael M Grisi, and Pablo Groisman. Harmonic deformation of delaunay triangulations. *Stochastic Processes and their Applications*, 122(5):2185–2210, 2012.

Evolving connectivity graphs in mobile phone data

Olivera Novović, Sanja Brdar, Vladimir Crnojević
BioSense Institute
University of Novi Sad, Serbia
onovovic@gmail.com, {brdars, crnojevic}@uns.ac.rs

1. INTRODUCTION

Connectivity graphs inferred from mobile phone data uncover pulse of human interaction. In the recent years many innovative applications based on this rich data emerged, such as urban sensing, transport planning, social analysis and monitoring epidemics of infectious diseases [1][2]. Anonymous mobile communication data from telecom operators can be utilized for sensing activities occurring within a city and can further fit into wider vision of smart cities that aims at monitoring and optimizing urban landscapes. Several studies explored mobile phone data in the context of urban sensing. Cici et al. analyzed aggregated cell phone activity per unit area that allowed them to detect seasonal patterns (weekday/weekend), anomalous activities and to segment a city into distinct clusters [3]. In another study, authors examined interactions among city inhabitants and visitors and identified the city's hotspots [4]. Mobile phone data can be also used to derive city land use information [5].

Here we utilized graph theory to study connectivity patterns on a city scale. We focused on the dominant backbone of networks - the most significant part of overall communication interaction. Pairwise communication was aggregated over spatial units of a city and one day time intervals and analysed throughout two months period. In our graphs nodes are spatial units and links were drawn if communication strength between units was significant. This allows us to study the backbone connectivity graphs as evolving structures and to examine temporal and spatial dynamics within a city. We measured global and local graph properties and here present a part of obtained results.

1. DATA

Mobile phone service providers collect large amount of data for every user interaction. Every time a user makes interaction using mobile phone (SMS or call), one *Call Detail Record* (CDR) is created in Telecom operator database. CDRs used in our research are provided by the Semantics and Knowledge Innovation Lab (SKIL) of Telecom Italia [6]. The records refer to the communication inside city of Milan for a time period of two months (November and December 2013). Telecommunication interaction between mobile phone users is managed by Radio Base Stations (RBS) that are assigned by the operator. Every RBS has unique id, location and coverage map that provide approximate user's geographical location. CDRs contain the time of the interaction and the RBS which handled it. In available data collection CDRs are spatially aggregated on the grid containing 10 000 cells and temporally aggregated on time slots of ten minutes. We used telecommunication interactions set that comprises measured intensities between different cells. Only cells that spatially intersect with administrative area of Milan city were selected. The city is divided into 88 administrative zones [7]. The map with zones is presented in Fig 1, where each is represented by unique id placed in its center.

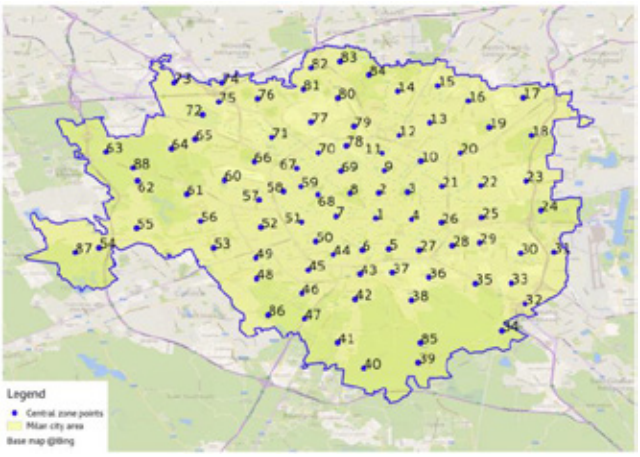


Fig. 1. Centres of 88 adninistartive zones of Milan.

2. CONNECTIVITY GRAPHS

To create connectivity graphs, we further aggregated communication from grid cells to 88 zones of Milan. Each grid cell is assigned to corresponding zone, thus our graphs refer to zones interactions. The connectivity matrices for 61 days were made in pairwise manner. Matrix element on the position $[i, j]$ represents aggregated communication strength between zone i and zone j. After creating the connectivity matrices the filtering was performed to eliminate weak links. The strong and weak links distinction was made by equation 1 by calculating the significance of links α_{ij} [8]:

$$\alpha_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}\,dx < \alpha, \qquad (1)$$

where α denotes significance threshold, p_{ij} is the probability of having link between nodes i and j, and k is the number of nodes. In our case $k = 88$ and α was set to 0.05. After the weak links were eliminated, graph structure for each day was created from remaining links in connectivity matrix. The final graphs are sparse and suitable for visual analytics. We presented links with QGIS and selected four typical graphs (Fig 2.). The first is typical weekday, where the strongest communication links tend to appear and concentrate near city center and maximum communication strength is much higher than observed on the weekends. The second is the holyday (Friday 2013-11-01), where the strongest communication

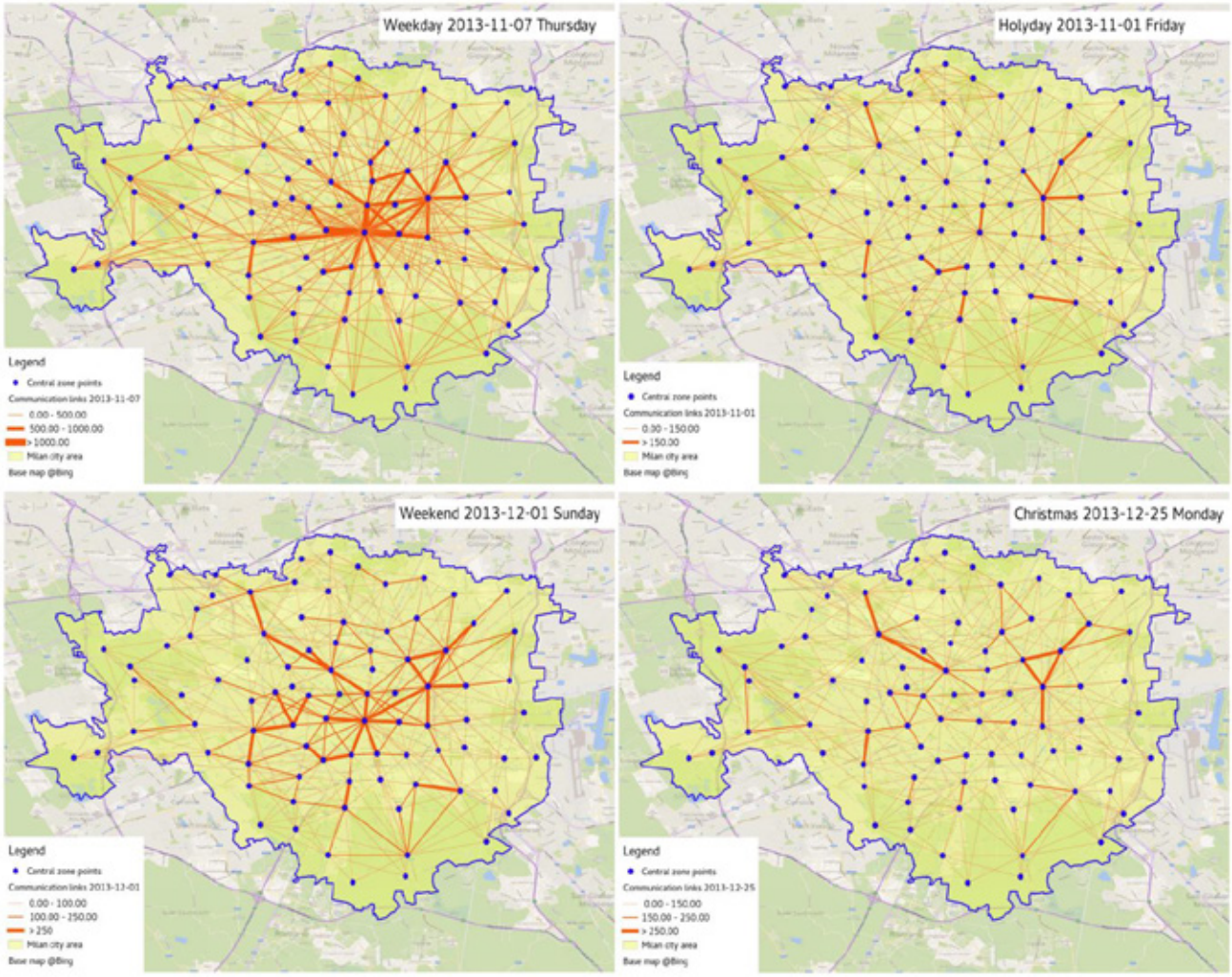


Fig. 2. Bacbone connectivity graphs between zones of Milan

links tend to disperse and the maximum communication strength is very low. The third graph presents typical weekend. The strongest communication links tend to disperse across city but the maximum communication strength is higher than on the holyday. Finally, the Christmas day is presented in the fourth graph. Its structure is similar as the one presented for another holiday. The strongest links are dispersed across residential parts of the city and the overall communication strength is low, which is typical for holydays.

3. GRAPHS PROPERTIES

Along with visual inspection of graphs across two month period we quantified graphs properties and did deeper analysis of their changes during time and identified interesting weekday/weekend distinctions. We performed both, global and local, graphs analysis [9].

Global graph properties provide information on a global structure and further allow comparisons among graphs. We calculated the number of edges, maximum weight, radius, diameter, max clique size, average clustering for all inferred graphs. We compared global graph properties on different day types: weekdays and weekends and discovered differences. We observed that number of edges is higher on the weekdays than on weekends (Fig 3). Distributions of measured diameters unveil that weekday graphs have lower diameter compared to weekend, indicating faster information flow during work days and larger "connectivity distance" in weekends.

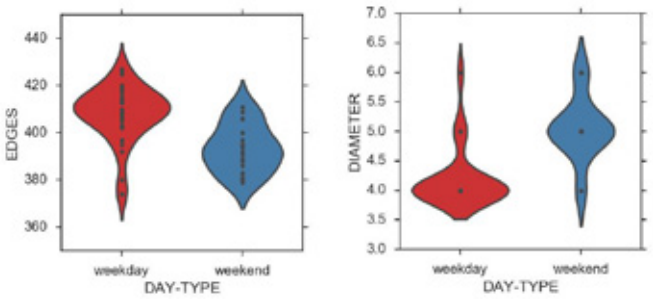


Fig. 3. Distribution of number of edges and diameter in different day types

Local graph properties uncover localized patterns in the graphs. We measured numerous local properties such as clustering coefficient, node degree, page rank, betwenness centrality, etc. Here we selected three nodes (Zone 1, 71 and 23, see Fig 1.) and presented changes in time of betwenness

centrality (Fig 4.) and PageRank (Fig 5.). Betweenness centrality calculates the number of shortest paths that pass through examined node and PageRank determines the relevance or influence node in graph. Zone 1 that encompasses city center has the largest betweenness centrality and PageRank and we can notice clear weekday-weekend pattern. Zone 71, has opposite pattern, it increases on weekends and drops on workdays. The strong links involving Zone 71 are also visible in Fig 2. during weekends and holidays. Interestingly Zone 23, that covers part of Lambrate district has unusual jump in betweenes centrality at 15th December, that could be due to event *The Lambrate Bicycle Film Festival*. PageRank pattern for Zone 23 differs considerably form betweenes centrality implying that different graph properties can provide complementary information.

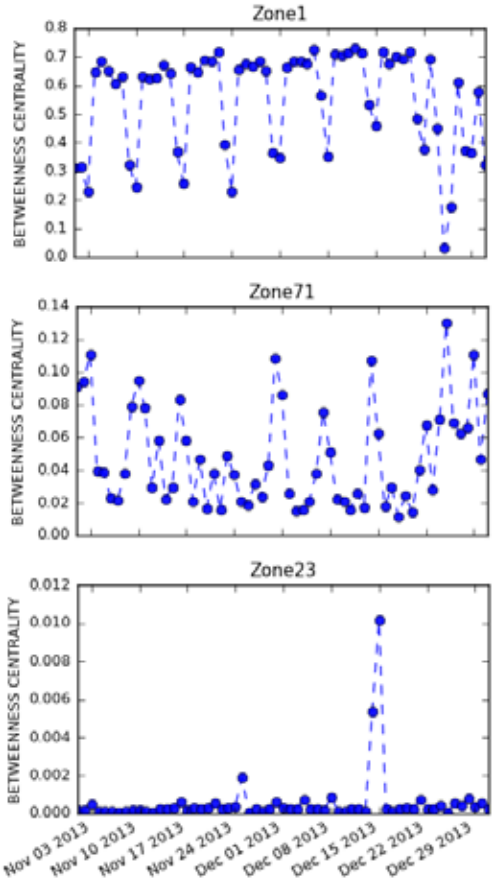


Fig. 4. Betweenness centrality of zones 1, 71 and 23

4. CONCLUSIONS AND FUTURE WORK

Our analysis of the connectivity backbone networks in the city provided new insights into social interactions and their changes across the city zones in different day types. Through the lenses of graph theory we discovered properties that can serve for detecting the patterns and deviations from typical observations. Our future work will include more graph-based formalism in identifying strong temporally consistent links, patterns of change and evolving graph sequences [10] as well as unveiling underlying social pulse that is reflected in mobile phone data.

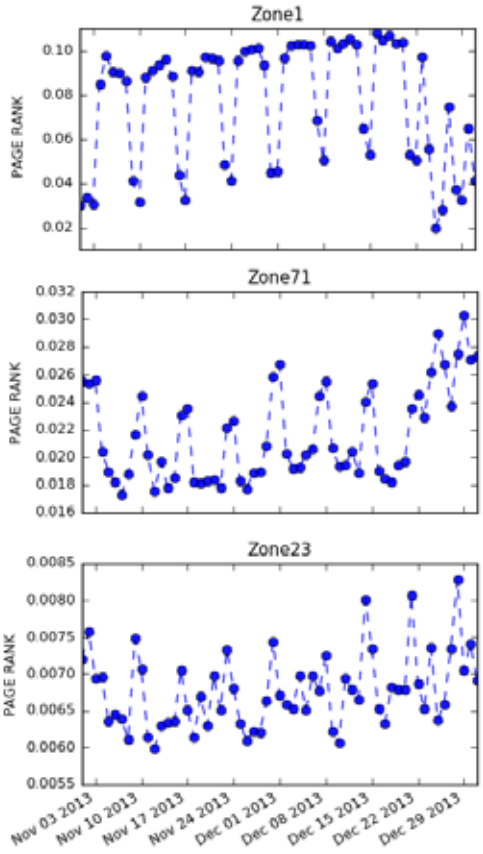


Fig. 5. Pagerank of zones 1, 71 and 23

REFERENCES

[1] V.D. Blondel, A. Decuyper, and G. Krings. "A survey of results on mobile phone datasets analysis." EPJ Data Science vol. 4, no. 1, 2015.

[2] D. Naboulsi, M. Fiore, S. Ribot, S and R. Stanica,. Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys & Tutorials, vol. 18, no.1, pp. 124-161, 2015.

[3] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, On the decomposition of cell phone activity patterns and their connection with urban ecology. In Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 317-326, June 2015.

[4] P. Bajardi, M. Delfino, A. Panisson, G. Petri & M. Tizzoni. Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Science, vol. 4, no. 1, 2015.

[5] T. Pei, S. Sobolevsky, C. Ratti, C., S. L. Shaw, T. Li, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, vol. 28, no. 9, pp. 1988-2007, 2014.

[6] G. Barlacchi et al. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," Scientific data, vol. 2, 2015.

[7] Milan's public data. http://dati.comune.milano.it/, September 2016

[8] M. Á. Serrano, M. Boguná, & A. Vespignani, Extracting the multiscale backbone of complex weighted networks. Proceedings of the national academy of sciences, vol. 106, no. 16, pp. 6483-6488, 2009.

[9] L. D. F., Costa, F. A. Rodrigues, G. Travieso, and P. R Villas Boas.. Characterization of complex networks: A survey of measurements. Advances in physics, vol. 56, no. 1, pp. 167-242, 2007.

[10] A. Kosmatopoulos, K. Giannakopoulou, A.N. Papadopoulos, and K. Tsichlas, An Overview of Methods for Handling Evolving Graph Sequences. In Algorithmic Aspects of Cloud Computing, pp. 181-192, 2016.

Abstract: Using Mobile Phone Signalling Data For Estimating Urban Road Traffic States

Thierry Derrmann, Raphael Frank, Francesco Viti, Thomas Engel
University of Luxembourg
E-Mail: {*firstname.lastname*}@uni.lu

I. Introduction

It is intuitive that there is a causal relationship between human mobility and signaling events in mobile phone networks. Among these events, not only the initiation of calls and data sessions can be used in analyses, but also handovers between different locations that reflect mobility. Various studies have been performed on Location/Tracking Area Updates, which concern idle phones (in a disconnected state) and are useful at larger distances, e.g. on Interstate highways [1], [2], [3]. However, there are – to the best of our knowledge – no studies on the relationship handovers of mobile phones and traffic states in urban environments. In this work, we investigate if handovers can be used as a proxy metric for flows in the underlying road network. More precisely, we show that characteristic profiles of handovers within and between mobile cell clusters exist. We base these profiles on models from road traffic flow theory, and show that they can estimate traffic states in an urban context using a distinct linear relation. The presented traffic state estimation model can be beneficial in areas with strong mobile network coverage but low road traffic counting infrastructure, e.g. in developing countries, but also complement existing transportation systems.

In the following sections, we will present our datasets and the necessary mapping between both domains. Then we introduce the clustering technique to partition both networks jointly and finally present the results of estimating road traffic state with this model.

II. Datasets

A. Mobile Network Signaling Dataset

The mobile dataset contains aggregate data from 1839 3G (UMTS) cells within the country of Luxembourg, 611 of which are located in and around its capital, Luxembourg City, the region relevant to this study. More specifically, the data consists of:

- the count of handovers between cell pairs per hour
- the count of calls initiated from each cell per hour

This data was made available for a single week at the end of September 2016. We organize the handovers in a *handover matrix*, i.e. the weighted, directional adjacency matrix of 3G cells in the study area.

B. Floating-Car Dataset

As ground truth data, we use Floating-Car Data (FCD) collected during the same study week. This is a set of time-stamped location updates and travel speeds which was collected in the area of Luxembourg City and its highway ring, and consists of 600 trips and 220000 location updates. In particular, we are interested in traffic states, i.e. the ratio between actual speeds and the speed limit ($\frac{v}{v_{max}}$).

C. Mapping FCD to the Mobile Network

In order to enable the use of FCD for validation purposes, we need to map the most likely associated mobile network cell to each GPS location entry. We can easily find each location entry's nearest cells by distance. Usually, this is a set of cells, as multiple cells are co-located at a common base station site.

From an FCD trajectory, we can thus identify a sequence of these sets of potentially associated cells, matching to the road path taken. Now, in order to identify the single, most likely visited cell sequence, we use the handover matrix. We choose the most frequent cell transition to be the likely cell pair visited, thus building a chain of visited cells over the entire trip. This means that we get a single likely associated cell for each Floating-Car Data entry, i.e. the cell that the car could have most probably been connected to at its location. This allows to compute road traffic states relative to the connected cell.

III. Clustering

We want to consider internal and exiting handovers for different partitions of the mobile network, and thus have to partition it into mobile cell clusters. In previous work [4], spectral clustering was identified as an adequate method for partitioning the handover matrix into densely connected clusters.

Ji et al. have shown in [5] that partitioning by normalized graph cuts is a valid starting point for finding homogeneous road network partitions. Spectral clustering is a relaxation of normalized cuts, and was effective for clustering the mobile network in a previous study [4] that we performed on simulated handover data.

In this work, we apply spectral clustering to the handover matrix (i.e. the weighted adjacency matrix) of the mobile network cells. The weights in the matrix correspond to the number of handovers between cell pairs. Spectral clustering allows defining the desired number of clusters. We identified the ideal number of clusters with respect to the error in traffic state estimation (cf. next section).

We have found the prediction error resulting from the respective clustering to be lower for the square root of the handover matrix, i.e. by increasing the amount of potential cuts by bringing the weights closer together.

IV. Road Traffic State Regression

We evaluate the predictive power of mobile network signaling data using Floating-Car Data (our ground truth). In particular, we want to estimate for each road network partition the *traffic state* variable, i.e. the ratio between the actual observed link speeds and their respective speed limits.

We want to evaluate if there are profile functions of the mobility inside and between mobile network cell clusters, similar to Macroscopic Fundamental Diagrams (MFD) in road networks. The MFD describes the relationship between outgoing flows and internal density of vehicles in a homogeneous partition of a road network. Similarly, we want to make use of exiting and internal handovers of mobile cell clusters to build such profile functions for mobile cell clusters. Using these functions, we then want to estimate the current degree of saturation of the underlying road network, as more or fewer handovers happen within clusters or across their boundaries.

The final goal then is to find a single regression equation that uses these (previously trained) profile function of mobile cell clusters, along with each cluster's handover and call counts, to estimate the road traffic states. If that is possible, then we have sufficiently characterized the mobility inside the clusters, as the same distinct regressive relation holds for all clusters.

A. Training set

In order to prove the above conjectures, we need to build a training set with the road network traffic state (the response variable) as well as each cluster's hourly statistics and profile function coefficients (as the dependent variables). From this, we can then estimate a global equation for all clusters relating the inputs to the traffic state. Thus, we construct the following training dataset using FCD and mobile network data from Monday through Wednesday of the study week:

To each FCD entry we have associated the current traffic state variable and most likely associated cell. This cell, in turn, maps to a single mobile network cluster as defined by the spectral clustering of the handover matrix. This means that we have a mapping of each FCD entry to its respective mobile network cluster, i.e. we can associate the response variable (traffic state) to the predictors (mobile network statistics).

For each cluster and hour, we compute:

- entering, internal and exiting handovers
- aggregate amount of calls emitted from this cluster
- quadratic fit coefficients of the profile function
- the derivative of the profile function
- the average traffic state ($\frac{v}{v_{max}}$, response variable)

Using these features, we can then evaluate our approach: We study whether the aggregate mobile network statistics and the previously learned profile functions (i.e. coefficients of the convex quadratic curve) can serve as sufficient predictors for the underlying road network traffic state. We found that a quadratic polynomial best fits the relationship between inner and exiting handovers.

B. Validation Set

The validation dataset consists of FCD data and mobile network data from Thursday. The values for a,b and c –i.e. the profile function coefficients– are adopted as learned from the training dataset. They represent the learned profiles from past observations.

Figure 1: MAPE by number of clusters

Figure 2: Test set validation with 28 clusters

C. Regression Results

As described above, we train our profile functions of each clusters on the data from Monday through Wednesday, and evaluate them on data from Thursday. We consider 20 through 130 clusters of the mobile network, all obtained by spectral clustering. We performed a set of 10 clustering and regression runs on every cluster count, to account for the randomness of the clustering initialization. Fig. 1 shows the Mean Absolute Percentage Error (MAPE) with respect to the number of clusters in question, averaged over 10 test runs for each cluster count. The noisy aspect of the scatter is due to the way the respective clustering partitions the road network. We have found a local optimum at 28 clusters. Hence we will continue our study with this value. Fig. 2 shows the scatter between estimated and actual traffic states for the Thursday data with for 28 clusters. Individual points represent a clusters' mean traffic state during an hourly time slot. The green line is the identity line, while the blue line represents the regression trend between prediction and true values. The proximity of both lines indicates a good fit The MAPE amounts to 11.8, which – given the constraints of our data sets – shows that mobile

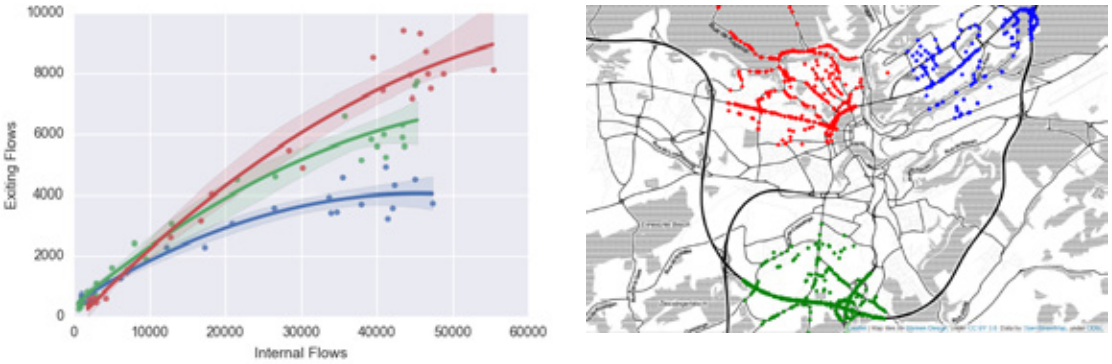


Figure 3: Example profile functions (left) and corresponding road network partitions covered by the cell clusters (right)

network data can be used for estimating road network traffic conditions.

D. Mobile Network profile functions

In Fig. 3, we see three example profile functions and the underlying data scatter (left) and a map of the FCD locations that correspond to the same clusters. We can see that the different clusters correspond to coherent regions of the road network. They were not yet optimized in terms of homogeneity of the underlying road network partitions. However, the resulting clusters' profile functions exhibit low variance, and distinct ratios of internal and exiting handovers that are characteristic of each region.

E. Limitations and Possible Extensions

The main limitation of this work is the temporal aggregation of the mobile phone data set (1 hour granularity). However, we believe that the results above are sufficiently good to show that traffic state estimation is feasible using only mobile phone data. In particular, there is room for improvement of the predictions, using more fine-grained data, a longer training period and multiple radio technologies (2G, 3G and 4G).

Another limitation is the low degree of congestion in the data, which prevents us from directly comparing our profile functions to other models. Generally speaking, the functions follow the free-flow and 'sweet-spot' parts of a Macroscopic Fundamental Diagram (MFD), as known from traffic flow theory [6]. However, for a lack of very severe congestion in the road network, we do not observe the negative slope of the flow ratio characteristic of grid-lock and spill-back phenomena.

V. Conclusion and Future Work

We have shown that profile functions of partitions of mobile networks exist, and that they exhibit predictive power for estimating the road network traffic state. Future work consists in evaluating whether or not these profile functions and Macroscopic Fundamental Diagrams possess a theoretical link, and if that link holds for extreme congestion conditions, i.e. grid-lock. This could be studied further in a simulation setting. We made a first step in this direction with our work in [4], but a more extensive study and comparison to FCD-based MFDs is necessary.

As mentioned above, there are various extensions and directions for future work regarding the results we found. We believe that by using multiple radio technologies and finer-grained data can lead to better models of the underlying topology.Further improvements can be expected from improved clustering algorithms of the mobile network that lead to more homogeneous road network partitions.

The goal of this work was to get a single regression equation mapping to all the clusters, in order to show that mobile network clusters have characteristic profiles with predictive power. We have shown that this is indeed possible, and have reached a MAPE of 11.8% with respect to the true traffic states. It is possible to improve the resulting error by using fixed temporal effects for the typical daily road patterns, more radio access technologies and longer periods of data. Finally, comparing mobile network profile functions directly to the underlying road network fundamental diagrams would be a helpful next step, because this would tighten the theoretical link between both domains.

References

[1] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs, "Cellular data meet vehicular traffic theory: Location area updates and cell transitions for travel time estimation," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 361–370. [Online]. Available: http://doi.acm.org/10.1145/2370216.2370272

[2] K. Hui, C. Wang, and A. Kim, "Investigating the use of anonymous cellular phone data to determine intercity travel volumes and modes," 2017.

[3] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380–391, 2007.

[4] T. Derrmann, R. Frank, and F. Viti, "Towards estimating urban macroscopic fundamental diagrams from mobile phone signaling data: A simulation study," 2017.

[5] Y. Ji and N. Geroliminis, "On the spatial partitioning of urban transportation networks," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1639–1656, 2012.

[6] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 759–770, 2008.

Characterizing Significant Places using Temporal Features from Call Detail Records

Mori KUROKAWA† Hiroki ISHIZUKA† Chihiro ONO†

† KDDI Research, Inc. 15-1-2 Ohara, Fujimino, Saitama, 356-8502

E-mail: † {mo-kurokawa, hk-ishizuka, ono}@kddi-research.jp

Abstract

Location information associated with Call Detail Records has paid attention as a probe for human mobility. In this paper, we evaluate methods to extract temporal features of significant places from location histories and to characterize them with respect to stay purposes. The accuracy of the methods is demonstrated by experiments using actual Call Detail Records and online questionnaire results about visited places collected from 1250 subjects with permission.

Keyword Call Detail Records, Significant Place, Stay Purpose

1. Introduction

Call detail records (CDRs) are seen as an important data source in the inspection of human mobility. CDRs are recorded on telecommunication equipment, mainly to detect and resolve problems with the equipment. Each record is generated through a call, sending of a text message or browsing the Internet via a mobile phone, and the record contains a timestamp and location related to connected base stations.

CDRs of entire mobile phone users enable us to make spatio-temporal statistics of human mobility. These statistics have potential to be utilized for social good, such as controlling QoS of telecommunication, urban planning, and minimizing damages of disasters.

Characterizing human mobility from the viewpoint of "For what they are" is important to enhance comprehension of human mobility. In this paper, we show methods of extracting significant places from location histories of each user and characterizing them. For characterization, we make a classifier of stay purpose classes listed below based on temporal features of significant places.

1) Home
2) Office / School
3) Shopping
4) Meal / Social activity / Entertainment
5) Sightseeing / Leisure
6) The other private purposes
7) The other business purposes

We evaluate our methods using four weeks' CDRs and questionnaires about stay purposes collected from 1250 subjects with permission.

The remainder of the paper is as follows. In section 2, we explain existing and proposed methods. In section 3, we present the results of evaluation. In section 4, we conclude this paper.

2. Methods

In this section, we explain methods to extract significant places and to characterize them. The extraction step is conducted as pre-processing for the characterization step.

2.1. Step 1: Extracting significant places

Among existing studies, some apply time-based clustering: [1] clusters points based on distance between temporally adjacent points and filters small clusters where little time was spent, and [2] clusters points by setting a threshold for switching the counts of cell towers.

In this paper, we apply our two-step clustering method. The first sub-step is for extracting candidate places and the second one is for clustering them to obtain significant places and their time intervals.

In sub-step 1, we apply a sliding time window with width T and shift S, where the first time segment is $[t, t+T)$, the second time segment is $[t + S, t + T + S)$, and so on. We apply Mean-shift [3] to each time segment which includes no less than N points. Mean-shift is a popular mode seeking method and we regard the mode of the location distribution as a candidate significant place if the distribution within the time segment is unimodal. Mean-shift has two parameters: distance thresholds Th_1 and Th_2. The previous one corresponds to a bandwidth of density estimation and the latter one is used for convergence check.

In sub-step 2, we apply Mean-shift again to cluster candidate places to obtain significant places as the clustered places. Finally, we apply rules described below to cluster the time segments and determine time intervals of significant places.

a) If the place in the first time segment and the one in the last time segment are equal, and
b) If the places in the intermediate time segment do not belong to the places other than the one determined in a),

then they belong to a consequent significant place.

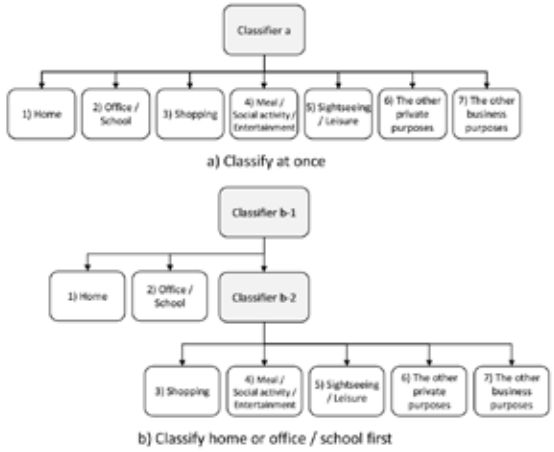


Fig1. Overviews of classifiers

2.2. Step2: Characterizing significant places

Existing studies apply machine learning techniques to predict semantic meanings of significant places. Isaacman et al.[4] extracts temporal features from CDRs and applies logistic regression to predict semantic meanings (home and work). Zhu et al.[5] extracts spatio-temporal features from Nokia Mobile Data Challenge datasets including GPS and accelerometer data and compares several machine learning methods for predicting semantic meanings (ten classes).

In this study, we make a classifier for each user, which predicts one of seven stay purpose classes based on temporal features of significant places. Temporal features consist of 24 features correspondent to 24 hours of day and each feature f[00-23] is defined: If the time interval of the significant place contains the hour of day, it takes 1. Otherwise, it takes 0.

Overviews of the classifier are displayed in Fig 1. Fig 1 a) shows a flat model which classifies seven classes at once. Fig 1 b) shows a hierarchical model which primarily classifies home, office / school or the residuals and secondarily classifies the residuals into the other five classes. Each classifier is Random Forest [6] which is known to perform well in many classification tasks.

Training data for the classifier consist of instances in each of which temporal features of a consequent significant place are labelled by one of seven classes. We assume we have no labelled data of the prediction target user, so we draw labels from semantic information of visited places collected from the other users.

3. Evaluation

In this section, we evaluate our methods in terms of i) predictive accuracy and ii) ranking of important features.

3.1. Collected data

From February 1, 2013 to February 28, 2013, we conducted an experimental survey with 1250 mobile phone users to obtain mobility data including CDRs and online questionnaires about visited places. Examinees are men and women ages 18-60 excluding high-school students, resident of metropolitan and six prefectures of Kanto, Japan. The mean number of records per user per day is 160.6.

Semantic information of visited places was collected by online questionnaires. In prior questionnaires, we collected zipcodes of examinees' home and office / school. During survey period, once in each week, examinees provided information about visited places where they visited within a week and stayed for no less than one hour. The information about visited places includes arrival times, stay durations and stay purposes.

3.2. Data pre-processing

For preparing training and test data, we apply spatio-temporal matching to label the time intervals of significant places with one of seven classes. The way of matching is as follows.

In case of home and office / school: we apply geocoding to zipcodes of them obtained in the prior questionnaires and assign the class to the significant places, if the centers of the places are neighbor of the geocoded lat/lon (within 2 km circle).

In case of shopping and the others: We assign the class to the time intervals of significant places, if the starting times of time intervals are around the arrival times obtained in questionnaires during survey period (within 1 hour difference).

Here, we use only labelled data to which stay purpose classes can be assigned. We do not deduplicate redundantly labelled data to each of which a class different from each other is assigned. In the results of the above processing, the number of resulted effective examinees is 1033.

3.3. Evaluation metrics and results

In section 3.3.1 we describe evaluation results in terms of predictive accuracy and in section 3.3.2 we present ranking of important features.

The parameters of our methods are set as follows.

The width of the sliding time window T=60min
The shift of the sliding time window S=15min
The parameters of Mean Shift Th_f=5km, Th_p=1km
The min points to apply Mean Shift: N=4
The number of trees in Random Forest: 500

3.3.1. Evaluation in predictive accuracy

We conducted five-fold cross validation. For each fold, randomly selected examinees are tested. The test data are labelled data of them and the training data are labelled data of the rest of examinees.

Evaluation metrics are precision, recall and F measure defined as follows.

$$\text{Precision} = \frac{\#\text{Corretly\_Predicted\_and\_Labelled\_Significant\_Places}}{\#\text{Extracted\_Significant\_Places}}$$

$$\text{Recall} = \frac{\#\text{Corretly\_Predicted\_and\_Labelled\_Significant\_Places}}{\#\text{Labelled\_Significant\_Places}}$$

$$\text{F measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 1 shows the results. It indicates the method b) outperforms the method a) for all classes. We think the reason is effect of grouping classes with high frequency (home, office / school or the others). However, the accuracy of 3-7th class is still low.

Table 1. Results of accuracy for each stay purpose

	Method a) classify at once			Method b) classify home or office / school first		
	Precision	Recall	F measure	Precision	Recall	F measure
1	0.788	0.898	0.839	0.814	0.877	0.845
2	0.646	0.800	0.715	0.754	0.714	0.733
3	0.000	0.000	0.000	0.082	0.020	0.032
4	0.083	0.002	0.005	0.185	0.163	0.173
5	0.000	0.000	0.000	0.125	0.035	0.055
6	0.000	0.000	0.000	0.150	0.317	0.204
7	0.000	0.000	0.000	0.095	0.022	0.036

3.3.2. Ranking of important features

Fig 2.1, 2.2 show ranking of important features the scores of which are calculated using Mean Decrease Gini. Here, we chose randomly one trial of cross validation for evaluation.

Fig 2.1 shows results of the first classifier of method b) which classifies home, office / school or the others. It shows top features are those of early morning (f[00-05, 23]). Those reflect temporal features of home well.

Fig 2.2 shows results of the second classifier of method b) which classifies the residual classes. It shows top features are those of daytime (f[08-09, 15-16, 21-22]). Those reflect time intervals when users often move for shopping or the other purposes.

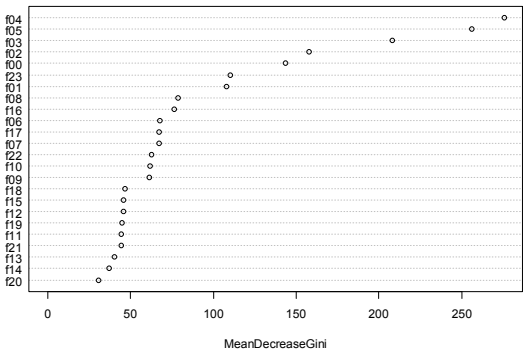


Fig 2.1. Ranking of important features
(Classifier b-1: classifies home, office /school or the others)

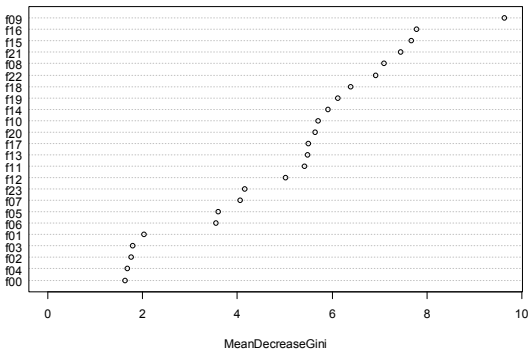


Fig 2.2. Ranking of important features
(Classifier b-2: classifies the other classes)

4. Conclusion

In this paper, we evaluated methods to extract significant places from CDRs and characterize them based on their temporal features. For further study, we will refine our classifiers to improve the predictive accuracy of stay purposes.

References

[1] J. H. Kang, W. Welbourne, B. Stewart, G. Borriello, "Extracting Places from Traces of Locations", Mobile Computing and Communications Review, Vol. 9, No. 3, pp.58-68, 2005.

[2] M.A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users," Proc. of the International Conference on Pervasive and Mobile Computing, Vol.6, No.4, pp.435–454, 2010.

[3] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", IEEE Trans. Pattern Anal. Mach. Intell, Vol. 17, No. 8, pp. 790-799, 1995.

[4] S. Isaacman, R. Becker, R. C´aceres, S.G. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data," Proc. of the 9th International Conference on Pervasive Computing, pp.133–151, 2011.

[5] Y. Zhu, Y. Sun, Y. Wang, "Predicting Semantic Place and Next Place via Mobile Data," the Mobile Data Challenge 2012 (by Nokia) Workshop, June 18-19, 2012.

[6] L. Breiman, "Random Forests," Machine Learning 45 (1): 5–32, 2001.

Estimating the Indicators on Education and Household Characteristics and Expenditure from Mobile Phone Data in Vanuatu

Zakiya A. Pramestri, Muhammad Subair, Imaduddin Amin, Dikara Alkarisya, Muhammad Rheza, Ni Luh Putu Satyaning P.P, Yulistina Riyadi, Mahdan A. Fauzi, Jong Gun Lee

Pulse Lab Jakarta - United Nations Global Pulse

Email: {zakiya.pramestri, muhammad.subair, imaduddin.amin, dikara.alkarisya, muhammad.rheza, ni.paramita, yulistina.riyadi, mahdan.hasan, jonggun.lee}@un.or.id

ABSTRACT

Vanuatu is a developing country and one of the fastest growing economies in the Pacific region. The government has implemented policy initiatives for developing its education and socioeconomic sector so it is important to monitor how the policy actually impacts the livelihood of its citizens. In this study, we aim to test the potential of using mobile phone data as proxies for education, household characteristics, expenditure and income diversity in Vanuatu as a data source to help monitor such proxy indicators in a timely manner.

1 INTRODUCTION

Vanuatu is located in the South Pacific Ocean and has a population of about 290,000. Vanuatu is a Y-shaped archipelago with four main islands and 79 small islands. It has six provinces (**Malampa**, **Penama**, **Sanma**, **Shefa**, **Tafea**, and **Torba**). The capital city **Port Vila** on **Efate** island in **Shefa** province and **Luganville** on **Espiritu Santo** island in **Sanma** are the largest and the second largest cities, respectively.

With regard to education, the government has been providing free primary education since 2010 in response to decreasing enrolment rates. This policy resulted in an increase in the enrolment rate which in 2012 was 88% for primary education. In terms of socioeconomic development, policy initiatives were introduced to promote the growth of agricultural production and tourism as the mainstay of the economy, and telecommunications, airline industries and infrastructure as supporting industries [1]. It was reported in 2010 that households spent 56% of expenditure on food, 14% on household operations and 5% on miscellaneous expenses. There are differences in the categories of expenditure, e.g., people in urban areas spend considerably more cash on food, clothing and transport vs. households in rural areas consume much more of their own home produce.

As one of the fastest growing countries economically in the region, it seems obvious that Vanuatu should understand the dynamics of its society. Among others, a need to monitor key development aspects such as education, housing characteristics and expenditure, has been addressed, but the national census in Vanuatu is carried out every 10 years, while the household expenditure and income survey is conducted every four years. More frequent information is essential to monitor and evaluate policy development to improve the social welfare system and protect vulnerable populations in developing countries such as Vanuatu but this is often quite difficult to achieve.

Recent studies have shown that mobile phone data can be used to understand socioeconomic conditions in the absence of official statistics. Researchers have found that Call Detail Records (CDRs) and airtime credit purchases can be used to infer economic parameters such as wealth, economic diversity, and economic segregation within communities [2]. Another study reveals that poverty indices have good correlation with several indicators derived from mobile phone activity [3, 4]. Furthermore, the authors of [5] confirmed that top up purchases can potentially be used for inferring poverty levels and measuring food consumption at quite a granular level.

In this paper, we investigate how mobile phone data can be used to produce a set of proxies for education, household characteristics, expenditure and income diversity in Vanuatu by comparing the features measured from mobile phone data, such as CDRs and airtime purchase records, with the ground truth data from official statistics. Our preliminary result shows that mobile phone data is highly correlated with many indicators on education, household characteristics and expenditure both at province and island level in Vanuatu. This implies that mobile phone data could be used as a source of near real-time proxy indicators to complement the information obtained from census which would be helpful in supporting the mapping of education and household demographics in Vanuatu.

2 DATA

This section briefly explains how to prepare a list of indicators from our primary and secondary data and Table 2 shows the entire indicators we used in this paper.

2.1 Official Statistics (Ground-truth Data)

We use two different forms of official statistics, (a) National Census of Population and Housing 2009 and (b) Household Income and Expenditure Survey 2010, both conducted by the National Statistics Office in Vanuatu. The first covers all households and individuals throughout

Vanuatu, while the coverage of the second is approximately 82.5% of all households.

- Education: We extract all education-related statistics at island level from the National Census to investigate the population segregation by education attainment, education completion, and literacy.
- Household Characteristics: We extract household electricity usage data and its segregation by the main source of electricity, electronic appliances ownership, motor vehicle and boats ownership and main source of water from the national census and also prepare the data per province.
- Household Expenditure: We extract all relevant statistics from the household expenditure survey at province level to investigate average education expenditure per household, total housing, fuel & lighting, and food expenditure.
- Household Income: We use total monthly income data and its segregation based on the income source including from wages & salaries, cash crops, fruits and vegetables sales, and fishing activities from the household expenditure survey at province level.

2.3 Mobile Phone Data

We produce nine indicators at province and island level, as shown in Table 2, using two types of anonymized mobile phone data from a mobile operator in Vanuatu, (a) 3-week CDRs and (b) 7-week airtime purchase records, collected between November 2016 and January 2017. It is worth noting that, when producing the indicators, per province, we use the mobile information up to the third or fourth most populated cities.

3 PRELIMINARY FINDINGS

Using the Pearson correlation method, we investigate which indicators on the education, household condition, and spending at province and island level from official statistics, could be estimated by the mobile phone data.

3.1 Province Level

One natural pattern from official statistics in Vanuatu is **`Shefa > Sanma > Malampa > Penama > Tafea > Torba`** which is the order proportional to population size. It is not surprising that we find the same pattern from mobile phone data as well, but it is interesting to note that any combination of two indicators, one from the official statistics and another from the mobile data from Table 1, is highly correlated each other (> 0.8).

Official Statistics	Mobile Phone Data
Working population	Total amounts of top-up
Internet usage	Total number of top-up
English literacy	Total num. calls and SMSs
French literacy	Num. unique customers
Bislama literacy	Total number of SMSs
Avg. edu. expenditure per H.H.	Total number of calls
Generator usage	Total duration of calls
Freezer ownership	
Telephone ownership	
Mobile phone ownership	
Boat ownership	
Total Monthly Income (TMI)	
TMI from wages/ salaries	

Table 1 A set of indicators from official statistics and mobile phone data which shows the pattern of `Shefa > Sanma > Malampa > Penama > Tafea > Torba` and is highly correlated each other

An interesting indicator from mobile data showing a different pattern from the indicators from official statistics is **`M3`** or the number of unique customers (SIM cards) who purchased prepaid phone credits, with the following order, **`Sanma > Shefa > Malampa > Penama > Tafea > Torba`** where the highest record is **`Sanma`**, while the province with the highest values from the other indicators is **`Shefa`**. It would be explained with the fact that the **`Sanma`** province is the most popular tourist destination based on Department of Immigration in Vanuatu [6], expecting that not only residents but also tourists top up phone credits, while tourists are not actively calling phones or sending messages.

3.2 Island Level

Among three types of official statistics indicators, Education indicators at island level are most correlated with the indicators from mobile phone data. The average correlation coefficient value of all combinations between 18 Education Indicators and nine Mobile Data Indicators is 0.92, while the one between 16 Household Characteristics (20 Household Ownership) and nine Mobile Data Indicators is 0.76 (0.72, respectively).

Table 3 shows the correlation results (up to 10 most correlated indicators from mobile phone data at island level) but some highlights are like the following.

- Education
  - The indicators most correlated with mobile phone indicators are the literacy rates of Bislama and English, which are two of three national languages in Vanuatu.
  - The population who completed secondary education or above (10-year certificate, college, university, vocational) generally has higher correlation with the mobile data indicators than the population with primary education only.
- Household Characteristics and Expenditure
  - Mobile phone data is highly correlated with the number of households (a) whose main source of light is generator or electricity main grid, (b) whose main source of cooking

Mobile Phone Data Indicators	
M1	Total Number of Calls and SMSs
M2	Total Duration of Calls
M3	Number of Unique Customers from Topup Records
M4	Total Number of Topup Transactions
M5	Number of Unique Customers from CDRs
M6	Total Number of SMSs
M7	Total Number of Calls
M8	Total Amount of Topup
M9	Average Amount of Topup per User
Education	
E1	English literacy
E2	French literacy
E3	Bislama literacy
E4	Other language literacy
E5	Education completed - Some primary education
E6	Education completed - Primary leaving certificate
E7	Education completed - Form 3 certificate
E8	Education completed - Year 10 leaving certificate
E9	Education completed - Senior secondary certificate
E10	Education completed - University entrance
E11	Education completed - College
E12	Education completed - Bachelor degree
E13	Education completed - Master degree
E14	Education completed - Vocational certificate
E15	Education completed - Others
E16	Population currently attending formal educational institution
E17	Population attending formal educational institution - full time
E18	Population attending formal educational institution - part time
Household Characteristics	
HC1	Number of household (# HH) - Rent free
HC2	#HH - Rent
HC3	#HH - Owned
HC4	# HH with main source of lighting from electricity main grid
HC5	# HH with main source of lighting from generator
HC6	# HH with main source of lighting from solar
HC7	# HH with main source of lighting from kerosene
HC8	# HH with main source of lighting from gas
HC9	# HH with main source of lighting from coleman
HC10	# HH with main source of cooking energy from kerosene
HC11	# HH with main source of cooking energy from electricity
HC12	# HH with main source of cooking energy from wood
HC13	# HH with main source of drinking water from piped private
HC14	# HH with main source of drinking water from piped shared
HC15	# HH with main source of drinking water from HH tank
HC16	# HH with main source of drinking water from village tank
Household Ownership	
HO1	# HH owning radio
HO2	# HH owning TV screen
HO3	# HH owning DVD
HO4	# HH owning mower
HO5	# HH owning freezer
HO6	# HH owning computer
HO7	# HH owninginternet
HO8	# HH owninggenerator
HO9	# HH owningmobile phone
HO10	# HH owning telephone
HO11	# HH owning gas stove
HO12	# HH owning boats
HO13	# HH owningmotor vehicle
HO14	# HH owning motorbike
HO15	# HH owning canoe
HO16	# HH having chickens as livestock
HO17	# HH having pigs as livestock
HO18	# HH having horses as livestock
HO19	# HH having cattle as livestock
HO20	# HH having goats as livestock

Table 2 **List of Indicators from Official Statistics and Mobile Phone Data**

Table 3 **Correlation Results between Indicators from Official Statistics and Mobile Phone Data at Island Level (Darker, higher correlation coefficient values. Numbers in cells are Indicator Indexes from Table 2)**

energy is electricity, and who owns electricity appliances, such as radio, TV, and DVD.

- Considering the fact that the electricity supply is dominated by main grids in urban areas, while it is dominated by diesel generation in rural areas, mobile phone data may be able to be used for assessing the electrification profiles in Vanuatu but we will leave this as our future work.

4 SUMMARY

We have presented that mobile phone data could be used for producing a set of proxies for some indicators on education and household characteristics, expenditure, and income diversity in Vanuatu. This is particularly valuable in developing economies, where traditional sources of population data are often scarce but mobile phones are increasingly common as these methods may provide a cost-effective option for measuring the characteristics of populations.

We plan to extend our findings by applying other statistical methods to map those education and socioeconomic features at quite a granular level, for instance at city level where official statistics data are not yet available.

REFERENCES

[1] Republic of Vanuatu. 10-12 February 2010. *Pacific Conference on the Human Face of the Global Economic Crisis Report*. Port Vila, Vanuatu.

[2] Gutierrez T, Krings G, Blondel VD. 2013. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*.

[3] Smith-Clarke C, Mashhadi A, Capra L. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, CHI '14. 511-520.

[4] Mao H, Shuai X, Ahn YY, Bollen J. 2015. Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Cote d'Ivoire. EPJ Data Science 4 (15). doi: 10.1140/epjds/s13688-015-0053-1

[5] A. Decuyper, A. Rutherford, A. Wadhwa, J. M. Bauer, G. Krings, T. Gutierrez, V. D. Blondel, and M. A. Luengo-Oroz. (2014). Estimating food consumption and poverty indices with mobile phone data. *arXiv preprint arXiv:1412.2595*.

[6] Vanuatu National Statistics Office. 2016. Statistics Update: International Arrival Statistics. Retrieved from http://www.vnso.gov.vu/index.php/economic-statistics/tourism-news