NetMob

Book of Abstracts 2019

The main conference on the scientific analysis
of mobile phone datasets
8-10 July 2019 Mathematical Institute,
Oxford University, Oxford, UK

Organisation:
Renaud Lambiotte (Oxford University)
Esteban Moro (Universidad Carlos III de Madrid)
Vincent Blondel (UCLouvain)
Alex 'Sandy' Pentland (MIT)

Site:
http://netmob.org/

OXFORD
UNIVERSITY PRESS

Science Advances
AAAS

Complexity Centre

Session Economic

Economical Segregation of Encounter Networks in cities

Esteban Moro,[1, 2, *] Dan Calacci,[2] Xiaowen Dong,[3, 2, †] and Alex Pentland[2]

[1] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain*
[2] *Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA*
[3] *Department of Engineering Science, University of Oxford, Oxford, UK*

Segregation is one of the most important population processes in cities: in the US one in five city dwellers live in a very income-segregated community [1]. Social or income segregation is a spatial process and most work has focused mainly in residential segregation, i.e. on the basis of places of residence. But due to increases in mobility of people today, segregation is a process than goes beyond home or work places [2, 4]. Furthermore, as Ray Oldenburg argued [3], third places (not home or work) in which people mix are important for civil society, democracy or civic engagement. But are there still enough third places in our cities to do that? How large and how diverse is the fraction of people that we encounter everyday? How is segregated is our mobility?

To answer that question we study a unique database of 3 billion location events of 329k users in the Boston metropolitan area during 6 months. Using those data we identify 10k different places where people of different economical backgrounds mix and construct the bipartite network that explains the mobility of individuals and places in the network (see figure 1). When we project the bipartite network onto places, we found that most of the places in which there is a lot of economical mixing (third places) are related to shopping and leisure activities and that patterns of mixing depend on the actual character of the place and the surrounding area [5]. On the other side, when projecting on individuals, we were able to construct the high-frequency encounter network due to their mobility. We find that those encounter networks [6] have a larger degree of economical segregation (large assortativity in income), but much less that what residential suggests, showing the important role of mobility to explain social segregation in our cities. Beyond the static view of those networks, we test the resilience of the segregation networks towards the removal of those social mixing (3rd places) finding that, in most cases, the economical segregation of cities severely depends on the existence of those places. We discuss the implications of our results in the context of future development of areas and in the ever-changing evolution of our cities.

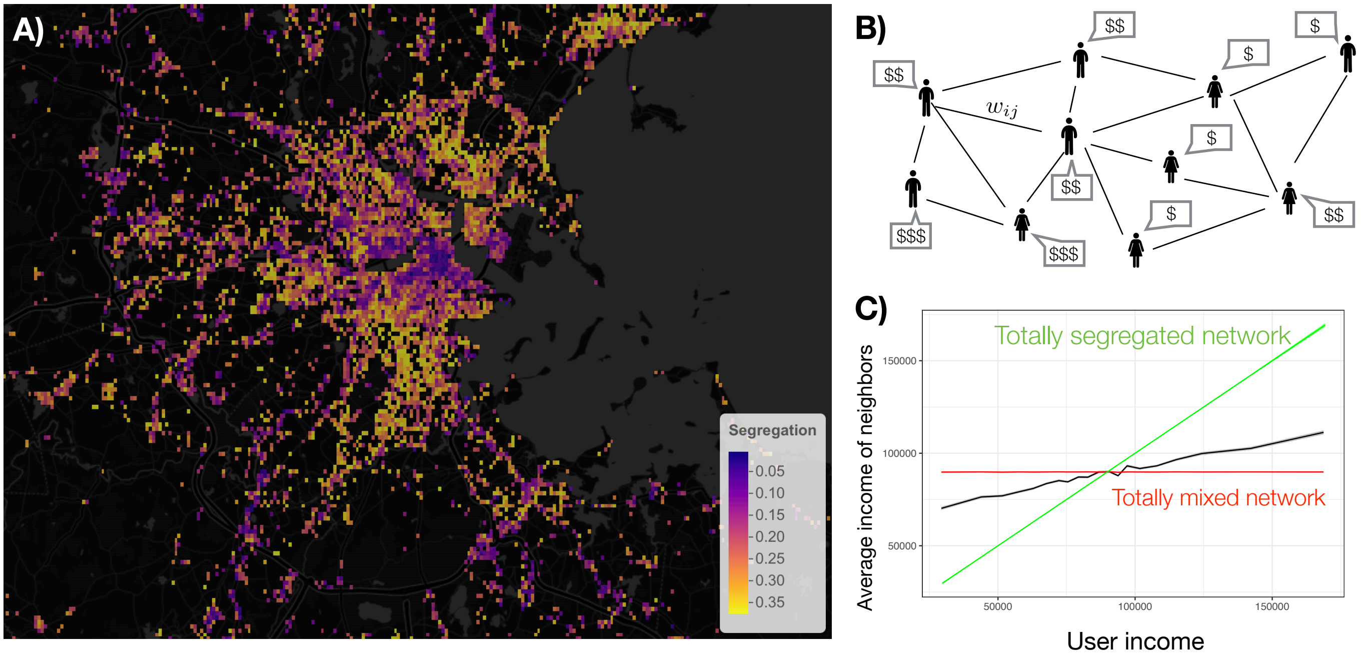


FIG. 1. A) Mobility segregation map around Boston. Each place in which there is at least encounters by 20 people is shown with a color proportional to the economical segregation (larger values are more segregated). B) Encounter/segregation network obtained from the bipartite individuals-places, where w_{ij} is the link weight between two individuals that measures the strength (duration) of the encounter. C) Average income of the neighbors of an individual as a function of her income. Real data. (black) is compared with the full mixed network (red, individual neighbors have random income) and the fully seggregated network (green, individual only meets neighbors of similar income).

[*] Corresponding author: emoro@math.uc3m.es
[†] Corresponding author: xdong@robots.ox.ac.uk

[1] Sampson, R. J. (2017). Urban sustainability in an age of enduring inequalities: Advancing theory and ecometrics for the 21st-century city. PNAS, 536(7617), 201614433.
[2] Wissink, B., Schwanen, T., & van Kempen, R. (2016). Beyond residential segregation: Introduction. Cities, 59, 126?130.
[3] Oldenburg, Ray. The great good place: Caf, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day. Paragon House Publishers, 1989.
[4] Silm, S., & Ahas, R. (2014). The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset. Social Science Research, 47, 30?43.
[5] De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., & Lepri, B. (2016, March 13). The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. arXiv.org.
[6] Sun, L., Axhausen, K. W., Lee, D.-H., & Huang, X. (2013). PNAS 110(34), 13774?13779.

Netflix or Youtube? Regional income patterns of mobile service consumption

Iñaki Ucar*, Marco Gramaglia*, Marco Fiore†, Zbigniew Smoreda‡ and Esteban Moro*
*Universidad Carlos III de Madrid †CNR-IEIIT ‡Orange Labs - SENSE

Introduction: Among the countless use cases that the analysis of the *digital footprints* created by the widespread adoption of fully network connected and portable devices allows to explore, a very promising but yet partially unexplored one is the study of socio-economic development of the society with respect to their interaction with the technology.

Motivated also by the pioneering works in [1], [2], in this one, we try to go one step further in this direction by using and unprecedented data source that allows us to capture a richer variety of features. As the current digital society is shaped around the Internet and their applications, we believe that a data set containing fine-grained information of mobile applications usage such as the one analyzed here will help to unveil currently hidden socio-economic aspects at very different geographical levels (e.g., city, region, country).

The data set analyzed here records the mobile traffic exchanged by users of the largest mobile network operator in a major European country. The data set consists of around 3.7 billion timestamped records, collected between May and June 2017, with a temporal granularity of 5 minutes. Data is aggregated at the Base Station (BS) level, and reports the total traffic (for uplink and downlink) for a particular mobile application. Applications are grouped into 40 categories that include very popular ones, such as Youtube, Facebook or Netflix, and others that identify the type of device, such as Apple Store or Google Play.

Methods and Results: On the one hand, traffic data are gathered per BS, for which we compute the Voronoi tessellation around them (BS zones). On the other hand, geographical, economic and demographic data are gathered per statistical zone. We only consider those for which we have income data (around 12k zones), i.e., populated urban areas (income for zones with less than 1000 inhabitants is not publicly available). Therefore, it is reasonable to assume that the traffic data generated in each BS zone was approximately evenly distributed, and thus we use a real weighted interpolation to map traffic data counts from BS to our target statistical zones.

As our goal is to infer relationship between the median income in a given area from the usage pattern of the 40 defined traffic categories. First of all, we aggregate traffic by hour and perform a hierarchical clustering of the usage for each hour of the day during weekdays. This clustering reveals a clear structure: hours from 8 to 17, or "working hours", are highly correlated. Instead, we are interested in "home hours", from 18 to 7 h, so that we can minimize the effect of people moving across areas during working hours as demographic information matches the inhabitants only.

Following this hierarchical clustering, we aggregate downlink traffic per category and area for home hours during weekdays, and normalize the usage by population for each area to obtain the activity in bytes/inhabitant. Then, we compute the Revealed Comparative Advantage (RCA) index of such activity, which measures the relative importance of a category in a certain area compared to the total importance in all areas. Finally, we re-arrange the data into 41 explanatory variables (RCA per traffic category and the total downlink bytes count) for each zone.

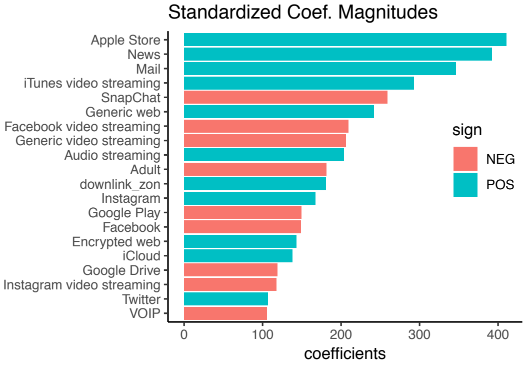


Fig. 1. The GLM standard coefficients

A penalized Generalized Linear Model (GLM) reports an $r^2 = 0.42$, which suggests that mobile applications usage is a very valuable data source for analyzing and predicting socio-economic development. Moreover, the analysis of the magnitudes of the standard coefficients show interesting insights. Our results indicate that the usage of Apple devices (identified by categories such as Apple Store, iCloud and iTunes video streaming) has a positive correlation with the consumption of News, Mail, Instagram and Generic Web content, and all these features have a positive impact in the prediction of the median income. On the other hand, the usage of Android devices (identified by categories such as Google Play) has a positive correlation with the consumption of Facebook, SnapChat and Adult content, and these features have a negative impact in the prediction of the median income.

Limitations and Future Work: Currently, this study is constrained to populated urban areas due to a limitation in the public availability of income data. We only consider coarse aggregates, thus losing the fine-grained features that may be available in the complete time series. However, the achieved correlation is inline with the one obtained by [1], [2]. Furthermore, we are not controlling factors such as tourist areas and other effects that may be adding noise to the data traffic gathered in certain zones. Despite this, these results open up the opportunity to analyze socio-economic development with a rich source of data with a promising predictive power.

References

[1] A. Llorente, *et al.* "Social media fingerprints of unemployment." PloS one 10.5 (2015): e0128692.

[2] L. Pappalardo, *et al.* "An analytical framework to nowcast well-being using mobile phone data." International Journal of Data Science and Analytics 2.1-2 (2016): 75-92.

Email addresses: {inaki.ucar,marco.gramaglia,esteban.moro}@uc3m.es, marco.fiore@ieiit.cnr.it, zbigniew.smoreda@orange.com

Enhancing financial inclusion with mobile phone data

María Óskarsdóttir*, Cristián Bravo†, 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
§Grandata Labs, Argentina, Email: charles@grandata.com

Financial inclusion is a globally recognized problem with large parts of the population having limited access to useful and affordable financial products and services, such as transactions, savings, credits and insurance [1]. In particular, over 2 billion people worldwide do not have a regular bank account and hence no means of obtaining credit in the traditional way since their –nonexistent– banking history cannot be used to assess their creditworthiness. At the same time, cell phone ownership in developing countries is increasing as reports from developing countries show [2], [3]. In this research, we demonstrate how mobile phone data can be used in credit risk models, that, in some cases, result in more accurate credit scores than models built with traditional bank history data [4].

We use a combination of data sets consisting of both customer bank data and mobile phone data, which allows us to construct social networks based on the exchange of phone calls between people. Thus we have the opportunity to explore the potential of enriching traditional credit scoring models with social network effects reflecting calling behavior. Our goal is to build credit scoring models for bank customer applying for credit cards. For each applicant, we extract several features which fall into one of the following groups:

1) Socio-Demographic: traditional bank history features including bank history, income and debit account behavior.
2) Calling Behavior: aggregated values for number and duration of phone calls made and received on different days and at different times of the day. These features characterize the intensity, diversity and timing of the applicants' calling behavior.
3) Link-Based: features that represent the individuals in an applicant's nearest neighborhood, such as the number of neighbors that have payment arrears.
4) Influence Score: the scores each applicant obtained after two distinct influence propagation algorithms are applied to the network. Credit card holders who already defaulted are used in lieu of information source.

Furthermore, we use the bank data to label the credit card applicants as defaulters and non-defaulters, depending on their payment arrears during the twelve months after receiving the card.

We build credit scoring models with all the extracted features using random forests. Thus, we are able to identify which features are most important for the model performance. The models are evaluated from a statistical perspective using the commonly applied AUC measure and from an economic perspective using the Expected Maximum Profit measure for credit scoring [5]. It has the advantage of considering the expected losses and operational income generated by a loan, and is thus tailored towards the business goal of credit scoring. Moreover, when applied to credit scoring models, it facilitates computing the model value and allows us to identify which features are most favorable in terms of profit.

Our results show that combining mobile phone data with traditional data in credit scoring models significantly increases their performance when measured in AUC. In terms of profit, the best model is the one built with only calling behavior features. In addition, the calling behavior features are the most predictive in other models, both in terms of statistical and economic performance This is an interesting result since it indicates that people's phone usage can be used as the sole data source when deciding whether they should be granted a credit. Mobile phone data is a powerful source of information for credit scoring. In the context of positive information, it has the potential to enhance financial inclusion by increasing the access to financing to borrowers who would otherwise be out of options.

References

[1] World Bank (2018), *Financial Inclusion*, viewed 27. April 2018, URL: http://www.worldbank.org/en/topic/financialinclusion

[2] Pew Research Center, April, 2015,"Cell Phones in Africa: Communication Lifeline", accessed 27 April 2018, URL: http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/

[3] De Luna-Martinez, J. 2017. *Financial inclusion in Malaysia : distilling lessons for other countries (English)*. Washington, D.C. : World Bank Group. Accessed 27 April 2018. URL: http://documents.worldbank.org/curated/en/703901495196244578/Financial-inclusion-in-Malaysia-distilling-lessons-for-other-countries

[4] Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J. and Baesens, B., 2019. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74, pp.26-39.

[5] Verbraken, T., Verbeke, W. and Baesens, B., 2013. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), pp.961-973.

Session Social Good

FlowKit: An open-source CDR analysis toolkit for humanitarian and development purposes

Daniel Power, Martin Thom, Jonathan Gray, Maximilian Albert, Chris Brooks, Sophie Delaporte,
Tracey Li, James Harrison, Joshua Greenhalgh, Nick Thorne and Linus Bengtsson
Flowminder Foundation

Corresponding author email: flowkit@flowminder.org

Call Detail Records (CDRs) and other mobile phone data can provide a dynamic, near real-time picture of the mobility and movement of millions of people across entire countries and, in combination with other data sources, can also provide information on people's characteristics, such as the socioeconomic status of individuals and communities [1]. With the right tools, actionable insights can be derived from this data and made available to decision-makers in a timely and accessible way. While the commercial world routinely uses mobile data to do everything from targeting food purchases to optimizing one's route to work, the humanitarian and development sector lags behind in optimizing service delivery with mobile network data. Despite the fact that 5 billion people are currently connected to the mobile internet, the data generated from cell phone use is still a novelty for most humanitarian and development organizations due to policy concerns about user safety and the lack of turnkey product solutions.

To address a key part of this problem, we have created FlowKit [2]: a suite of software tools that help humanitarian and development organizations access and analyze mobile data. FlowKit is targeted at the distinct needs of humanitarian and development practitioners in three key ways:

1. FlowKit provides a mobile data analysis toolkit informed by the Flowminder team's experience producing insights from mobile phone data to support the humanitarian and development communities. Because FlowKit has been developed with a strong understanding of the needs and priorities of collaborating partners — humanitarian and development organization and MNOs — it is able to facilitate collaboration and help form strong partnerships. It provides the necessary features and support to allow all parties to engage in an easy, low-cost manner. The implementation of this toolkit can be modified and customized according to the needs and constraints of a range of different end users.
2. Complications around security and privacy tend to be the single biggest blocker to this type of data being used by humanitarian and development organizations. FlowKit facilitates secure and GDPR-compliant data access, and also provides processing with built-in quality assurance, lowering overhead costs for users and smoothing the path to data analysis.
3. FlowKit is released under the MPLv2 open source license [3], thus promoting the transparency of its techniques as well as future sustainability by encouraging additional development and contributions from others working on similar challenges.

FlowKit will continue to grow and evolve with the addition of a modular CDR QA suite, additions and enhancements to the API, stronger connectivity to other data platforms, and support for more data analysis environments. We will present an overview of the features FlowKit provides for data access and analysis.

[1] J. E. Steele et al., Journal of the Royal Society, Interface 14(127), 20160690 (2017)
[2] FlowKit, https://flowminder.github.io/FlowKit/
[3] Mozilla Public License, version 2.0, https://www.mozilla.org/enUS/MPL/2.0/

Mapping geographical communities using Call Detail Records and OpenStreetMap

Joanna Wilkin [1, *], Dr Eloise Biggs [2] and Professor Andrew J Tatem [1]

[1] WorldPop, Geography and Environmental Science, University of Southampton
[2] School of Agriculture and Environment, University of Western Australia
* Correspondence: jw3e15@soton.ac.uk

Abstract

Mobile phone call detail records (CDRs) form an integral part of the UN's proposed data revolution, which aims in part to mobilise the use of data, traditional and innovative, to help monitor and achieve the recently agreed Sustainable Development Goals. Within these goals, the scale of the community features highly; Goal 11, for example, aims to foster sustainable communities. Despite this focus on the community within sustainable development, CDR analyses are not yet provided at this scale. Typically, CDRs are spatially aggregated at the coverage level of the cell tower or aggregated further to an appropriate administrative region. This paper explores the possibility of refining the spatial resolution of CDRs to the community scale, whilst preserving the required anonymity of the users. Utilising emerging literature on social spatial networks, where communities 'adhere to the old-school definition encompassing shared area and based on social ties' (Sarkar et al, 2016[1]), and building upon kernel density community mapping methods used by Comber et al (2012)[2], Gao et al. (2013)[3] and Shi et al. (2015)[4], this paper tests whether geographical communities can be identified within CDR data. Already, big spatial network datasets have shown how geography is a compartmentalizing factor, revealing distinct spatially-embedded social networks identified as communities. The next step, assigning these communities with a geographical reference is yet to be trialled. By integrating CDR-detected communities with OpenStreetMap data, this paper tests whether they can be associated with a known geographical community. Pilot testing on CDRs from Nepal has indicated that towns and cities can be geographically matched to specific CDR-detected communities. This geo-referencing enables each community member's data to be attributed to that community, rather than a cell tower. As a result, aggregation can occur at the community level, providing a more appropriate scale to be used within sustainable development work. For example, the next step for this research will be to extract social network metrics for each detected community to be used in a community resilience index, similar to Cueto et al (2017)[5]. Further testing of this innovative geo-referencing methodology within four districts of Nepal is underway and will be complete by the conference.

Reference List

1. Sarkar, D., Sieber, R. & Sengupta, R. GIScience considerations in spatial social networks. in *The Annual International Conference on Geographic Information Science* 85–98 (Springer, 2016).
2. Comber, A. *et al.* Exploring the geography of communities in social networks. in
3. Gao, S., Liu, Y., Wang, Y. & Ma, X. *Discovering spatial interaction communities from mobile phone data. Transactions in GIS* **17,** (2013).
4. Shi, L., Chi, G., Liu, X. & Liu, Y. *Human mobility patterns in different communities: a mobile phone data-based social network approach. Annals of GIS* **21,** (2015).
5. Cueto, D. N., Villalta, I. V. & Bernal, M. N. Resilience to disasters and social capital. Analysis of social networks in peripheral neighborhoods of the city of Cusco, Peru. *Boletín la Asoc. Geógrafos Españoles* **74,** (2017).

CubicA – A recommender system improving information delivery to smallholder farmers through IVR

Christophe Bocquet[1], Kristýna Tomšů[1], Lucio Melito[1], Denys Sementsov[1], Jamie Arkin[2], Leah Newman[2], William Tinzaara[3], Vincent Johnson[3], Jérôme Urbain[1]

[1]Dalberg Data Insights, [2]Viamo, [3]Bioversity International

In Uganda, with a population expected to increase from 40M today to 100M in 2050, agriculture provides a livelihood to 77% of the population (FAO, 2018). However crops are afflicted by several challenges including unpredictable and remote markets, low financial inclusion, pests and diseases, environmental disasters and natural resource depletion. To better address those challenges, smallholder farmers need access to the right information at the right time. Today, an average of 40.000 Ugandans call Viamo's 321-Service, an interactive voice response (IVR) system, every month to learn about agricultural practices and access weather forecasts. The service works like a voicemail box: anyone can simply call into this hotline for free and listen to the information menu anywhere, anytime, in their local language, on any mobile device. While the 321-Service continues to expand every day, IVR systems are built as a one-size-fits-all. Therefore callers are directed through multiple layers of listen-and-choose steps in order to access one of hundreds of messages available on the service, which can be complicated, time consuming, and tedious.

CubicA is an artificial intelligence (AI) and a big data analytics module developed to improve the information delivery of IVR systems ensuring callers receive the right message at their time of need. The idea is to shift the current system towards a real search engine. Through the integration of machine learning leveraging previous caller data to develop a recommender system along with satellite imagery, meteorological data and other data points to identify underperforming lands for instance, CubicA will (i) build live profiles for callers and (ii) optimize the current structure and content delivery beyond traditional information stylization. The existing top-down elementary and rather static information service will be enriched and extended to also produce bottom up insights to generate targeted content and better user experience. We believe that such a tool will help farmers to both (i) make strategic decisions as well as (ii) anticipate, identify and monitor pests and diseases. Eventually, crowdsourced data produced by the tool should also create value for government ministries & NGOs, input & finance providers, and producers who strengthen the agricultural sector and are tasked with guaranteeing food security.

The team is developing the recommender system that will make Viamo's existing IVR service more dynamic over time and customized to different users' needs. The development of the Recommender System represents the cornerstone of CubicA's technical aspect. With more than two years of user data from Viamo's existing IVR system (including 2 million unique users, 8 million user journeys including over 95 million interactions with database of 39 thousand items) combined with external big data sources, there is an abundance of data that can be exploited to infer people's needs and wants even before they provide the system with this information.

Our work is focusing on two main lines of inquiry: classical Latent Factor Models augmented by contextual data, and neural networks that are particularly apt at dealing with sequential instructions.

- The former models try to learn lower-dimensional embeddings of people's preferences, inferred from the implicit feedback they have provided by listening to a particular message among the choices they were presented with. They can be enhanced by including information not normally available in their standard representation, like the weather situation their region was experiencing at the time, or machine-understandable representations of the text users listened to.

- Deep recurrent neural networks, on the other hand, are perfect at capturing temporal characteristics implicit in the data (i.e. in its simplest form message B always being listened after message A. They are also ideal for scenarios that might require repeated recommendations.

The testing of these models is accompanied by the usage of relevant evaluation metrics (e.g. precision at K, reciprocal rank etc.) and by ablation studies aimed at understanding the contribution of each component to the final predictions. Baseline benchmarking already shows results exciding initial expectations.

https://bigdata.cgiar.org/inspire/inspire-challenge-2018/cubica-the-new-farmer-advisory-app/

Urban poverty mapping in Haiti

Using linked household survey and Call Data Records for high-resolution estimation

Guilherme A. Zagatti*, Claudio Bosco, Samantha K. Watson, Nancy Lozano-Gracia, Emilie Perge, Sering Touray, Linus Bengtsson
Flowminder Foundation and The World Bank Group
* Corresponding author email: guilherme.zagatti@flowminder.org

According to the World Bank Development Indicators, Haiti is the poorest country in the American continent with an estimated poverty headcount of 23.5 percent in 2012. While the country has many development challenges ahead, there is significant deficit in data to inform national and local policy-making, with the latest nationwide representative household survey only completed in 2012 from which the poverty rate above originates.

Call Detail Records (CDR) present an opportunity for population characteristic mapping estimation, with particular regards to poverty. Indeed, a number of prior studies have demonstrated the feasibility of using CDR for high-resolution poverty mapping estimation. Blumenstock et al.[1] uses CDR to predict poverty rates captured from a cell phone survey representative of the subscriber base, showing that their model would more accurately predict subsequent DHS surveys than using past ones in Rwanda. On the other hand, Steele et al.[2] shows that even in the absence of respondent-level data it is possible to estimate poverty using CDR.

The present study builds on the promises of these early findings. We conduct a stratified sample household-survey that captures household characteristics associated with poverty as well as cellphone usage patterns in the municipality of Cap-Haitien. Poverty is estimated using a selected number of indicators via the SWIFT methodology developed by the World Bank. With the spartcity of studies linking individual survey data with CDR, this study innovates exactly on that, providing the opportunity to estimate poverty at the subscriber level and to model individual heterogeneity more accurately. It also allows us to understand the representativeness of CDR-based estimates with regards to the complete population. Finally, since the SWIFT indicator attempts to capture income-based poverty, we investigate whether CDR can estimate consumption-based indicators of poverty as accurately as asset based ones. Steele et al. reports wide disparities in the accuracy of prediction of asset-based indicators versus income- and expenditure- ones, with significant higher levels of accuracy for predicting the former.

The research is still ongoing. While the data has already been captured and processed, modelling and analysis of the main findings are planned to be completed by the end of April 2019. We plan to proceed by estimating home locations following Zagatti et al.[3] and to estimate poverty using latent Gaussian models within a Bayesian framework. CDR and geospatial covariates of interest are linked to a structured additive predictor via a link function, while the correlation induced by geography and social links are modelled, respectively, via a Gaussian process and by extending the classical Besag model to multigraphs. Latent Gaussian models are simpler to implement, easier to interpret and provide confidence intervals which are extremely important to evaluate our degree of uncertainty.

This paper is the result of research commissioned by The World Bank Group, and performed jointly by Flowminder, the World Bank and Digicel.

[1] Blumenstock et al. 2015. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076.
[2] Steele, J.E. et al. 2017. Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface 14, 20160690. https://doi.org/10.1098/rsif.2016.0690
[3] Zagatti, G.A. A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. Development Engineering 3, 133–165. https://doi.org/https://doi.org/10.1016/j.deveng.2018.03.002

Maximum likelihood reconstruction of population densities from mobile signalling data

Jan van der Laan[*], Edwin de Jonge[†]

Introduction

For policymakers, (local) governments, and emergency services, it is important to have detailed and timely estimates of the spatial density of the number of persons. Therefore Statistics Netherlands started the Real-time Population Statistics project which aims to get (almost) real-time estimates of the population density from aggregated and anonymised signalling data. In this paper we will present a method to obtain maximum likelihood estimates of the population density using signalling data.

Methods

We will assume that the geographic area of interest is divided into a number of subregions j ($j = 1, 2, \ldots, N_p$) which we will call pixels (although they do not necessarily have to be rectangular and of equal size). The goal is to estimate the number of devices x_j in each of the pixels from the number of devices y_i connected to (generating an event with) antenna i ($i = 1, 2, \ldots, N_a$). To estimate the number of devices in a pixel we need the probability that a device in pixel j connects with antenna i, H_{ij}. In our case we used an antenna model to estimate the signal strength at each pixel (using properties of the antenna) from which the probability of connecting to that antenna is derived (Salgado et al., 2018a,b). The expected value of y_i is given by

$$\lambda_i = \sum_j H_{ij} x_j. \tag{1}$$

The y_i are independently Poisson distributed. This leads to the same likelihood function as discussed in Shepp and Vardi (1982). They derive an expectation maximisation (EM) algorithm to obtain estimates of the density, which in our case is equal to:

$$\hat{x}_j^{(n+1)} = \frac{\hat{x}_j^{(n)}}{\sum_i H_{ij}} \sum_i \frac{H_{ij} y_i}{\sum_k H_{ik} \hat{x}^{(n)}}. \tag{2}$$

This is an iterative method that updates the estimate of the population density in each step. The likelihood increases with each step.

Results

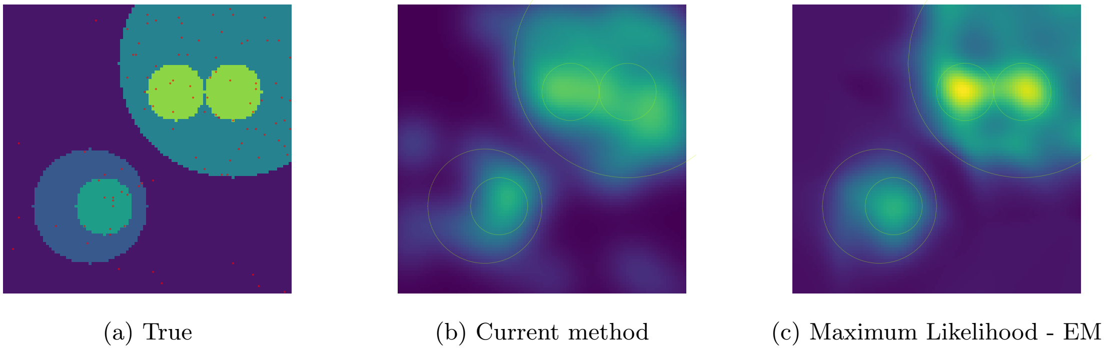
Figure 1 shows the results of applying the method to simulated data. With the EM-algorithm the estimates are more localised and the two hotspots top left are better reconstructed. We are currently applying the method to actual signalling data and comparing the results to official population estimates. The results of these analyses will be presented at the conference.

References

Salgado, D. et al. (2018a). Proposed elements for a methodological framework for the production of official statistics with mobile phone data, ESSnet Big Data, WP5, deliverable 5.3. Technical report, Eurostat.

Salgado, D. et al. (2018b). Some it elements for the use of mobile phone data in the production of official statistics, ESSnet Big Data, WP5, deliverable 5.4. Technical report, Eurostat.

Shepp, L. and Y. Vardi (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging MI-1*(2).



(a) True (b) Current method (c) Maximum Likelihood - EM

Figure 1: Comparison of the two types of estimators to the true population density (red dots are antennas) for a simulated data set. In this example the MSE is reduced by 30%.

[*] Statistics Netherlands, dj.vanderlaan@cbs.nl

[†] Statistics Netherlands, e.dejonge@cbs.nl

AIDA, the Action Insights Data Platform

Rositsa Zaimova, Kristýna Tomšů, Astrid Van Lierde, Jérôme Urbain, Lucio Melito, Christophe Bocquet, Denys Sementsov, Damien Jacques, Frédéric Pivetta

Dalberg Data Insights, Brussels, Belgium, jerome.urbain@dalberg.com

Introduction. Relevant, accurate and timely data are needed to track progress towards the Sustainable Development Goals (SDGs), support policy decision and promote sustainable development. Yet, collecting data for official statistics is expensive and available data sources in developing countries are often outdated or lacking granularity. The Action Insights Data Platform (AIDA) developed by Dalberg Data Insights (DDI) aims to bridge the resulting knowledge gap by harnessing the unique features of data from the private sector (mobile phone data, satellite imagery, traffic sensors…) to develop and test new methods that can supplement official statistics in developing and emerging countries.

Objective. In partnership with the United Nations Capital Development Fund (UNCDF) and the Belgian Government, DDI has been developing technical tools, operational processes, legal frameworks as well as business and commercial models to allow public and private stakeholders to leverage both data and algorithms in emerging and developing countries.

Partnerships. This approach has required to develop specific processes ensuring privacy and creating a trust environment. The different stakeholders of AIDA are Data Providers (both public and private), Developers (e.g., DDI, technical teams of local regulators), End-Users and Funders. AIDA is a hybrid open/closed platform:

- Some data are public and open, while others are private
- Some algorithms and visualization tools are open, while others are closed depending on the development clauses

On the legal side, we have developed first sets of contracts, for both the data holders and the end-users, to clarify the processes, commercial model, responsibilities and scope. AIDA has already demonstrated how to generate economic value and social impact during its pilot phase; and developed a pricing structure depending on the project goals.

The platform was piloted in Uganda and has now been deployed in 7 countries. AIDA provides data insights in various topics such as urban planning, food security, financial inclusion, energy, youth unemployment, gender inequalities and public health.

For each use case, we worked together with the end-users not only to define the relevant metrics, but also to build interactive dashboards to act on the insights. AIDA hence offers an end-to-end solution, from data access to data visualization and covers all steps of a data driven policy process including baseline data collection, decision support and impact assessment tools. The dashboards are available only to the end-users, through a secured web portal.

Data Access. For each deployment of AIDA, we have secured a data access agreement with at least one major private data provider (e.g., a Mobile Network Operator). The private data access has taken place in 2 forms: a) DDI installed a server within the data provider's premises to obtain anonymized data, process and aggregate them on site, behind the data provider's firewall; or b) the data provider directly sends aggregated data to AIDA. In any case, individual data stays where it has been collected; only aggregated data is securely transferred to the visualization layer of the platform. As such, AIDA is in line with the latest developments in individual data access and privacy protection (e.g., the OPAL project).

Algorithms. The base algorithms provided by AIDA cover i) mobility (number of people, migration patterns, travel time and traffic estimation, etc.); ii) mobile money analysis (activity in each area per transaction type, with the possibility to focus on some groups of interest); iii) social graphs; and v) infectious disease risk monitoring. Machine learning algorithms were leveraged in the platform, to predict a) the gender and the employment status of mobile phone subscribers; b) the incidence of infectious diseases; and c) crop yield from satellite images. Algorithms have been implemented in python, R and anatella (an ETL software with a Graphical User Interface).

Future Work. Next steps for the platform development are 1) to automate tasks that are currently manual (e.g., data anonymization or debiasing); 2) to standardize some pieces of the code and open them to audit and external contributions.

Exploring the use of mobile phone data for national migration statistics

Shengjie Lai[1,2,3], Elisabeth zu Erbach-Schoenberg[1,2], Carla Pezzulo[1], Nick W Ruktanonchai[1,2], Alessandro Sorichetta[1,2], Jessica Steele[1], Tracey Li[2], Claire A Dooley[1,2], Andrew J Tatem[1,2]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, United Kingdom

[2]Flowminder Foundation, SE-113 55 Stockholm, Sweden

[3]School of Public Health, Fudan University, Key Laboratory of Public Health Safety, Ministry of Education, 130 Dongan Road, Shanghai 200032, China

Shengjie Lai and Elisabeth zu Erbach-Schoenberg contributed equally to this work.

*Correspondence to Shengjie Lai (email: Shengjie.Lai@soton.ac.uk).

ABSTRACT

Statistics on internal migration are important for keeping estimates of subnational population numbers up-to-date as well as urban planning, infrastructure development and impact assessment, among other applications. However, migration flow statistics typically remain constrained by the logistics of infrequent censuses or surveys. The penetration rate of mobile phones is now high across the globe with rapid recent increases in ownership in low-income countries. Analysing the changing spatiotemporal distribution of mobile phone users through anonymized call detail records (CDRs) offers the possibility to measure migration at multiple temporal and spatial scales. Based on a dataset of 72 billion anonymized CDRs in Namibia from October 2010 to April 2014, we explore how internal migration estimates can be derived and modelled from CDRs at subnational and annual scales, and how precision and accuracy of these estimates compare to census-derived migration statistics. We also demonstrate the use of CDRs to assess how migration patterns change over time, with a finer temporal resolution compared to censuses. Moreover, we show how gravity-type spatial interaction models built using CDRs can accurately capture migration flows. Results highlight that estimates of migration flows made using mobile phone data is a promising avenue for complementing more traditional national migration statistics and obtaining more timely and local data.

Employment Demographics of Refugees in Turkey: A Bayesian Probabilistic Approach using Weak Social Science Models[⋆]

Steven Reece[1], Franck Duvell[2], Carlos Vargas-Silva[3] and Zovanga Kone[3]

[1] Machine Learning Research Group, Oxford University and the Alan Turing Institute, London, UK, reece@robots.ox.ac.uk
[2] German Centre for Integration and Migration Research (DeZIM), Berlin, Germany, duvell@dezim-institut.de
[3] Compas, Oxford University, UK, {carlos.vargas-silva,zovanga.kone}@compas.ox.ac.uk

We utilise mobile telephone data provided by Türk Telecom to investigate the economic activities of Syrian refugees in Turkey. To do this we develop a probabilistic model that relates mobile phone usage data to manual and non-manual job classes and use this model to determine the extent, type and spatial dispersion of the refugee labour force in Turkey. Our aim is to classify each phone caller or texter as manual, non-manual or non-worker. We propose weak rules of behaviour from intuition and by interviewing several Turkish and Syrian nationals. These rules are informed hunches and certainly not scientifically proven and hence 'weak'. We determine the efficacy of these rules using a machine learning algorithm that establishes their consistency with the Türk Telecom mobile phone data. This data provides cell tower identifiers used by a group of randomly chosen active users to make phone calls and send texts. The data is timestamped and a particular group of users are observed for a period of 2 weeks. At the end of the two-week period, a fresh sample of active users are drawn at random. Each sample contains 3% of the refugee base plus equal amount of non-refugee users.

Our weak rules are based on the assumption is that manual workers do not use their phone during working hours, 7am to 1pm and 2pm to 8pm, However, non-manual workers may use their phones for business purposes and thus more during working hours. Workers' usage of telephones for private purposes spike around lunch time and after work, hence after 7pm. Calls between 7pm and 11pm are of a private nature. Manual workers may make a call before 7am to determine work availability. We note that non-workers do not use their phone as early in the morning as workers of any kind, and they also use it more evenly across the day. Research suggests that students display specific patterns and rather use their phones in the afternoon and evenings.

We use an extension of an unsupervised Bayesian data analysis technique, the Independent Classifier Combination Algorithm (IBCC), to classify mobile phone users from their mobile phone usage. The IBCC learns the relationship between employment rules and the mobile phone data. The IBCC does not require telephone call/SMS data labelled as *manual* or *non-manual* to train it. Instead it uses the assumption that the majority of the rules are accurate. Our method models both refugee and non-refugee workforces within a single principled Bayesian approach and thus can infer both refugee and non-refugee labour classifications without loss of information. The approach provides both an estimate of each Turk's and refugee's job class, a statement of confidence in the classification as well as the efficacy of the job classification rules. We note that our approach is not biased by the relative number of rules for manual and non-manual work. Furthermore, non-employed phone users are deemed to be those whose behaviours do not conform to manual nor non-manual worker rules.

In this study we found that refugee workers are more likely manual than non-manual workers (in contrast to Turkish workers). Refugees don't normally have permission to work and only have access to informal employment. Our results not only provided country-wide statistics of employment but also gave a detailed breakdown of employment characteristics via heatmaps across Turkey. This information is valuable since it would allow GOs and NGOs to refine and target appropriate policy to generate opportunities and economic integration as well as social mobility specific to each area of Turkey. Our findings with regards to probability of refugees not in work suggest that in most places of high concentration of refugees they are also more probably unemployed than Turks. However, there is one exception and that is the region of Bursa. Here refugees seem much less likely to be unemployed than Turks in the province. Bursa is an industrial hub and its surrounding are a centre of agricultural activities, notably growing of grapes and olives. Thereby, it offers a favorable opportunity structure for refugees who can find employment in manufacturing and/or agriculture.

For further details of our approach and results please see, Steven Reece, Franck Düvell, Carlos Vargas-Silva and Zovanga Kone, *New Approaches to the Study of Spatial Mobility and Economic Integration of Refugees in Turkey*, Data for Refugees Challenge Workshop, 2019. Contact Reece for a pdf copy.

[⋆] Steven Reece was supported through a fellowship at The Alan Turing Institute, sponsored by the Programme on Data-Centric Engineering and funded by Lloyds Register Foundation.

Novel mobility and social network metrics to predict disaster-driven displacements from call detail records

Tracey Li, Jesper Dejby, Maximilian Albert and Véronique Lefebvre*

Flowminder Foundation

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

Over 25 million people are forced to leave their homes every year due to environmental disasters. The majority of these displaced persons are Internally Displaced Persons (IDPs), meaning that they remain within their home country. IDPs are among the most vulnerable people in the world today, as many States affected by natural disasters are unable to support displaced citizens and often lack the capacity to effectively identify and monitor displacements. The Global Humanitarian Overview 2018 showed there is a growing gap between humanitarian needs and the resources and information available to meet them. New approaches to identifying, understanding, and predicting internal displacements are needed in order to better target assistance and make the humanitarian response more proactive.

Call Details Records (CDRs) have been used in numerous studies to estimate population movements in data deficient contexts. We have developed a novel methodology to identify individual disaster-driven displacements from CDRs, which is described in the abstract 'Identification of disaster-driven internal displacements from the time series analysis of call detail records'. We then developed novel metrics to extract information on mobility and social contacts from CDRs in order to analyse how displacement duration and distance are affected by contextual and individual variables.

We present a novel measure of entropy based on meaningful 'stayed' locations and show how it can be used to determine the time at which individual IDPs resettle, either when they return home or resettle at a new location. We then count the number of IDPs who resettle over time (or equivalently the number of IDPs who remain displaced), which provides an indication of the rate at which the affected communities recover and adapt, and thus of disaster resilience. We studied CDRs pertaining to 3 natural disasters (Haiti earthquake 2010, Haiti Hurricane Matthew 2016, and Nepal earthquake 2015) and observed that the fraction of IDPs remaining displaced decays at an exponential rate over time, and at the same rate for the 3 disasters studied. If the study of further disasters leads to the observation of similar exponential decay rates, then it would imply that the number of IDPs at any time can be inferred from an estimate of the initial number of IDPs immediately following the disaster. Alternatively, the method provides a way to monitor disaster resilience and compare recovery rates across disasters.

Secondly, we studied the impact of social contact locations, and places visited prior to the disaster, on the 'choice' of displaced location and displacement distance. We find that good predictors of displacements are the fraction of all social contacts residing within a given distance of an IDP pre-disaster home, and the fraction of all visited places within a given distance of an IDP pre-disaster home. Individuals living by most of their social contacts are more likely to be displaced close to their home than those with a more widespread social network, even in the most severely affected areas. Similarly we observed that individuals who mostly visit places close to their home are more likely to be displaced nearby.

Our results indicate that CDRs can significantly contribute to measuring and predicting displacement durations, distances, and locations of IDPs in post-disaster scenarios. We believe that information and estimates provided by specifically developed CDR analytics, coupled with field data collection and traditional survey methods, can assist the humanitarian response to natural disasters and the subsequent resettlement efforts.

Session Mobility

CDR-based descriptors of seasonal sensitivity and mobility

David Pastor-Escuredo[1,2,*] Federica Carfagna[3] Zbigniew Smoreda[4] Miguel Luengo-Oroz[5] Pedro J. Zufiria[1]
[1] Technical University Madrid, Spain. [2] LifeD Lab. [3] African Risk Capacity, WFP. [4] Orange Labs. [5] United Nations Global Pulse
[*] corresponding: david@lifedlab.org

The use of Machine Learning techniques combined with CDRs enables the profiling of mobility and social behaviors based on the patterns of mobile phone usage. Here we present the integration of profiling of mobility and behavioral (Bandicoot) indicators to automatically estimate seasonal behavior and landmarks of population groups using CDR data available from the Data for Development (D4D) Challenge Senegal. This is an initial step towards a data-driven assessment of vulnerability for resilience building and funding.

We made an analysis at the Senegalese livelihoods level: geographical areas where people share broadly the same patterns of access to food and income, and have the same access to markets. The profiling of the population was done with behavioral Bandicoot descriptors. Here we used six descriptors: *Days-activity, Duration-calls-mean, Duration-calls-std, Entropy-contacts-texts, Entropy-contacts-calls* and *Entropy-places.* Thus, we could relate behavioral patterns with the production system of the country.

We integrated data sources (remote sensing and survey data) to interpret behavioral patterns in terms of the Bandicoot descriptors. The descriptor time series with a monthly resolution featured relevant landmarks in relation to the rainfalls (estimated from NASA-TRMM data) and the most relevant months (field preparation and harvest) of the agricultural calendars of livelihoods (FEWS livelihoods report). The *Duration-calls-mean* and the *Entropy-contacts-texts* reached their maximum during the rainy season as described by the rainfalls, presumably due to communication regarding the fields and the expected yield. During the harvest month, the *Entropy-contacts-calls* spiked, presumably indicating communication regarding trading. The *Entropy-places* showed two peaks during the field preparation and the harvest implying local mobility, potentially due to labor and trading mobility. These landmarks were interpreted to be relevant to monitor seasonal behavior of populations against climate change and droughts.

We introduced a *seasonal sensitivity indicator* aggregating Bandicoot indicator landmarks into a feature vector as a behavioral signature of each user. We applied temporal filtering to the Bandicoots time series to select the most relevant time points for each descriptor as variables. The resulting vector was then combined using weights for each variable into a single scalar indicator representing the communication and mobility associated to seasonal dynamics. Each variable affected positively to the indicator so the weights were all positive. Several weight combinations were tested. Thus, the *seasonal sensitivity indicator* was designed as a multidimensional aggregation of landmarks maximizing the amount of activity related to the livelihoods and the environmental conditions. The indicator showed a quasi-gaussian distribution for the Senegalese population. We segmented the distribution into different population groups using unsupervised k-means clustering (k=3): low sensitivity, medium sensitivity and high sensitivity.

We measured the mobility of each classified group by observing which livelihoods were visited through the year by each user, creating a dynamic census in each livelihood zone for the three population groups. Thus, we could explore characteristics and differences across livelihoods. Significant patterns were observed for the low sensitivity population, mostly in rural zones. We consider that tracking variations and mobility of this group along years will allow understanding the vulnerability to climate conditions, production and market prices. Finally, we compared the mobility of the sensitivity profiles with the mobility profiles obtained through direct unsupervised clustering of mobility vectors discovering that the mobility of the low sensitivity population was not captured by the main profiles. This means that the use of Bandicoot indicators is key for data disaggregation for humanitarian purposes. We foresee this framework will be relevant for humanitarian agencies to identify beneficiaries and monitor impact of humanitarian aid.

Anomaly detection of urban dynamics in an extreme weather with mobile GPS data

Yutaro Mishima[1], Atsunori Minamikawa[1]
KDDI Research Inc., Tokyo, Japan
yu-mishima@kddi-research.jp, at-minamikawa@kddi-research.jp

ABSTRACT

These days human dynamics data generated from GPS data or CDRs has an important role in considering urban environments and emergency management during natural disaster [1]. In particular, with alerts of abnormal states of dynamics, i.e. large-scale change of the volume of people, we can find some kind of important information to make management decision, e.g. evacuation spots not designated, so that we can minimize the damage caused by natural disaster. Generally, many anomaly detection techniques, e.g. classification-based, have been proposed [2]. Although anomaly detection techniques in human dynamics data are proposed in related research [3-4], they have challenges. Interpreting "Abnormal state" is difficult in [3] and [4] is not able to recognize the time when abnormal state ends in real time. In this paper, we present an anomaly detection method which can detect anomaly of human dynamics in real-time and indicate that our method enables monitoring wide area (city-level) human dynamics for a months or year and it is useful for emergency management.

DATASET

We use human dynamics data generated from GPS data provided by our mobile users who agree to our license. Human dynamics data consists of the volume of staying people and moving people in 500m x 500m grid-cell in 30 minutes timeslot. The volume is estimated from the ratio of our mobile users to population in Japan. There are 2387 grid-cells which cover whole metropolitan area in Tokyo. The data contains dynamics from 1 May, 2017 to 31 Jan, 2018. Notice that the data is generated so that it does not enable to identity specific individual.

APPROACH

In this section, we introduce the procedure of our anomaly detection method.

1. Define "normal" dynamics.
   i. We divide dynamics data for each grid-cell by date into 8 (2*2*2) day groups according to whether the day, previous day and next day are weekday or holiday.
   ii. Referring to [5], we cluster day by day data for each day group by k-means clustering for excluding abnormal dynamics caused by events etc. We define the cluster the average people volume of which is minimum as "normal" cluster. Then we define mean and standard deviation of all day by day data in cluster for each timeslot as "normal" dynamics of the grid-cell and day group.
2. Define "abnormality" for each grid-cell and timeslot in the following equation.

$$A_{d,t,c} = \left(\frac{v_{d,t,c} - m_{d,t,c}}{s_{d,t,c}}\right)^2$$

where $A_{d,t,c}$ and $v_{d,t,c}$ are abnormality and volume, $m_{d,t,c}$ and $s_{d,t,c}$ are mean and standard deviation as "normal" dynamics in the day group d, the timeslot t and the grid-cell c.

3. Determine if the state $ST_{d,t,c}$ is abnormal or normal in the following.

$$ST_{d,t,c} = \begin{cases} Abnormal, if\ A_{d,t,c} \geq 4\ and\ |v_{d,t,c} - m_{d,t,c}| \geq 1000 \\ \qquad Normal, otherwise \end{cases}$$

We note that the threshold is set temporarily.

EXPERIMENT

First, we determine states for all grid-cell and timeslot (from 1 May, 2017 to 31 Jan, 2018) and find out when many abnormal grid-cells exist and why the anomaly occurs. Next, we generate heatmaps and plot the dynamics in a certain grid-cell in one day of abnormal days excluding long vacations. Finally, we check the difference of dynamics from ones of "normal" days matches with the content of news reported in the day.

RESULTS AND DISCUSSION

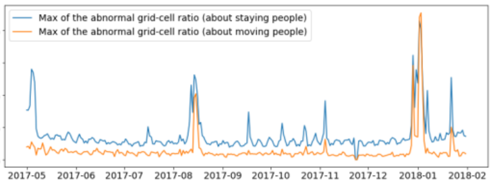


Fig. 1 Max value of abnormal grid-cell ratio for each day

We define abnormal grid-cell ratio as the ratio of number of abnormal grid-cells to the number of all grid-cells for each timeslot. Fig. 1 shows max value of abnormal grid-cell ratio (about staying and moving people) for each day. Although there are some peaks on both of the plot about staying and moving people, almost all of the peaks are caused by vacations, e.g. spring vacation in May, 2017. Exceptions are peaks on 22 Oct, 2017 and 22 Jan, 2018. The former is caused by typhoon and the latter is heavy snow. As shown by Fig. 1, we can quantify the impact which an event causes on human dynamics.

For an example, we show heatmaps below in 22 Jan, 2018, the anomaly ratio on which is largest except for vacations. Heavy snow struck Tokyo in 22 Jan, 2018 and it had a huge impact on many people and public transportation.

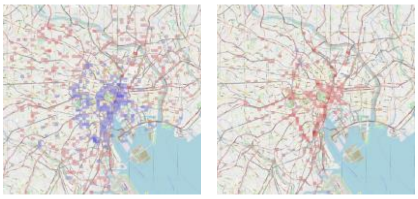


Fig. 2 Heatmaps (staying people/moving people) in 22 Jan, 2018 †1

Red colored cell represents abnormal increase and blue cell represents abnormal decrease. Color depth corresponds to the degree of increase or decrease. Left heatmap represents abnormal state about staying people from 18:00 to 18:30 in 22 Jan, 2018 and right one represents about moving people from 16:00 to 16:30. We can find that many people (probably working people) came home earlier than usual because of heavy snow. Therefore, we can monitor city-level human dynamics and see where significant anomaly occurs.

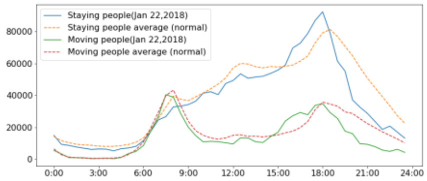


Fig. 3 Time-series plot of a certain grid-cell in 22 Jan, 2018

Fig. 3 shows a time-series plot of a grid-cell (contains Ikebukuro which is one of the major stations in Tokyo) in 22 Jan, 2018. Vertical axis indicates the volume of people. Blue and green solid lines show the volume of staying and moving people, while yellow and red dashed lines show the average volume of staying and moving people in normal dynamics. On the plot the volume of staying people is largest around 18:00 and actually at that time entering station was restricted in Ikebukuro and it was most crowded [6], so we prove that this dynamics trend correctly reflects the actual human dynamics trend even if we gaze at only one grid-cell. Processing time it takes to define normal dynamics from about 9 months dynamics data is about 2600 seconds and one it takes to define abnormality for 1-day data is about 10 seconds with 36 CPUs and 750 GiB of RAM. That processing time is short enough to construct real-time anomaly detection system because we don't have to run the process defining normal dynamics frequently (the system will work even if we run it only once a week).

CONCLUSION

In this paper we present an anomaly detection method in human dynamics data generated from GPS data and show that we can not only monitor wide area human dynamics for a long period in real-time but also observe the human dynamics properly in a single grid-cell (500m x 500m) with some experiments. For future work, we consider more appropriate decision rule whether abnormal or normal and propose more useful information for emergency management, e.g. priority of rescue, using hazard information generated from sensor data.

REFERENCES

[1] Sagl, G., Loidl, M., & Beinat, E. (2012). A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. ISPRS International Journal of Geo-Information, 1(3), 256-271.
[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.
[3] Kamiya, K & Fuse, T. (2015). STATISTICAL ANOMALY DETECTION FOR MONITORING OF HUMAN DYNAMICS. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XL-4/W5. 93-98. 10.5194/isprsarchives-XL-4-W5-93-2015.
[4] Neumann, J., Zao, M., Karatzoglou, A. and Oliver, N. (2013). Event Detection in Communication and Transportation Data. Proceedings of IbPRIA 2013.
[5] AL Alfeo, MGCA Cimino, S Egidi, B Lepri, & G Vaglini. (2017). An emergent strategy for characterizing urban hotspot dynamics via GPS data. – NetMob 2017, 27-29.
[6] Weathernews "1gatu22nichi, nanganteikiatsunitomonaukanto-chihounoooyuki" (Heavy Snow caused by winter extratropical cyclone struck Kanto region)
- https://jp.weathernews.com/news/21459/ (2019/3/8)

†1 © OpenStreetMap contributors, https://www.openstreetmap.org/copyright

Extracting meaningful mobility flows from mobile phone individual trajectory data for epidemic metapopulation models

Giulia Pullano[1,2 *] , Stefania Rubrichi[2], Vittoria Colizza[1]

[1]INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique IPLESP, F75012, Paris, France
[2]SENSE/Orange, XDLab , Chatillion, France
*Presenter: giulia.pullano@inserm.fr

The past few years have seen an increasing interest in the use of Call Detail Records (CDRs), i.e. georeferenced data obtained from cell phones activities. As they contain attributes on where and when activities are made, CDRs allow the extraction of user's individual trajectories in space and time by interpolating displacements based on two consecutive activities. Such high-resolution data have been widely used to characterize individual patterns of mobility and were recently integrated into spatially explicit epidemic models to understand the spread of infectious disease epidemics. These have been applied to a variety of epidemic contexts including influenza, malaria, cholera, schistosomiasis. Little attention has however been devoted to the aggregation process aimed at extracting coupling forces between geographical locations based on individual mobility described by such high-resolution data, and its relevance for spatial infectious disease modeling. This may become particularly important given the large heterogeneities exposed by the analysis of trajectory data, with e.g. individuals spending most of their time in few locations vs. others visiting a large number of locations, or individuals with remarkably reproducible mobility patterns vs. others displaying highly variable ones.

We considered different aggregating procedures and evaluated the resulting coupling forces between locations as well as their impact on the modeled epidemic diffusion once they are integrated in a spatially explicit metapopulation scheme. We selected three commonly used methods at highest, medium, and lowest spatio-temporal resolution that are based on, respectively: (i) the full temporal sequence of a user's displacements between any two consecutive calls (*displacement-based coupling matrix D*); (ii) the number of calls made in a given location (*location-based coupling matrix L*); (iii) the daily most visited location computed on the number of calls (*commuting-based coupling matrix C*). The *D* coupling matrix is able to trace the actual spatial path followed by each user through a sequence of consecutive displacements, however it does not consider the location of residence. The *L* and *C* matrices, instead, couple the residence location of each user with all her visited locations (in *L*), or with her most visited location (in *C*), however loosing the information on the actual trajectory. We applied our study to 1-year CDRs from Senegal, from January to December 2013, including 9,569,425 users, and integrated the coupling matrices in a metapopulation model accounting for disease dispersal due to visitors to a location and to returning residents. The model is stochastic, discrete and non-Markovian, and considers three different synthetic epidemic scenarios – high transmissibility (Ebola-like), moderate transmissibility (influenza-like), and low transmissibility (e.g. when interventions are put in place).

We found that the median coupling probability between any two locations in *D* is two orders of magnitude smaller than the values estimated in *C* and *L*. These discrepancies increase with the distance between the locations, such that long-range coupling is less likely to be captured in *D* compared to the other aggregating approaches. The lower coupling probability measured by *D* results in delayed arrival times of the simulated epidemic ranging from few weeks to almost two years compared to *L* and *C*, and depending on the epidemic scenario. The location-based and commuting-based coupling matrices lead to highly correlated and similar arrival times per location that are consistent with observations based on commuting flows. Most importantly, differences in estimated coupling lead to substantially different paths of invasion of the epidemic in space and time in the country. Such invasion is mainly fragmented into short distance hops of spatial transmission in *D*, whereas a more heterogeneous pattern integrating long-range transmission with local-range dispersal is reproduced by *C* and *L* coupling matrices, consistently with observed epidemic patterns of spatial diffusion.

Our findings indicate that preserving the full resolution of the observed trajectory of individual movements may bias the spatio-temporal diffusion of the simulated epidemic in both the timing and pattern of invasion. Aggregating on visited locations, while loosing all information about their sequence in an individual path, reproduce realistic simulated patterns, but preserving only the most visited location is already enough to reliably model the disease spread. This also suggests that for a range of epidemic contexts leisure activities (i.e. those performed at the various locations except the most visited one) have no significant impact on the spread, and commuting-like mobility is the main driver of disease diffusion. While addressing a specific methodological problem of reducing data resolution, our findings have important implications for (i) the identification of the most relevant locations to be targeted for disease prevention and control, and for (ii) the data and resolution needs in case of an emerging epidemic, a recently much debated topic.

Urban Vibes and Rural Charms: Analysis of Geographic Diversity in Mobile Service Usage at National Scale

Rajkarn Singh, Marco Fiore[†], Mahesh K. Marina, Alessandro Nordio[†] and Alberto Tarable[†]
The University of Edinburgh UK, [†]CNR-IEIIT Italy
Email: r.singh@ed.ac.uk, marco.fiore@ieiit.cnr.it, mahesh@ed.ac.uk, alessandro.nordio@ieiit.cnr.it, alberto.tarable@ieiit.cnr.it

I. Problem and results

As mobile data traffic keeps surging worldwide, knowledge of where, when, how and why mobile services are consumed by network subscribers becomes increasingly relevant across research and technology domains. Still, our comprehension of mobile service adoption is currently limited and many questions remain unanswered. We focus on one such open question, namely: *"how similar (or different) are demands for mobile services across a whole country?"* We answer by analyzing a real-world dataset of mobile network traffic collected by a major operator that describes the demands for individual services in 10,000 *communes* (*i.e.*, administrative areas) in France. Our study yields the following key insights.

- The demand for most popular mobile services is fairly uniform across the whole country, and only a reduced set of peculiar services (mainly operating system updates and long-lived video streaming) yields geographic diversity.
- Just 9 (respectively, 50) service consumption patterns are sufficient to retain 23% (respectively, 35%) of the overall usage diversity, implying that a small number of distinct behaviors is sufficient to characterize the many thousands of areas in the whole of France.
- The spatial distribution of these behaviors correlates well with the urbanization level, ultimately suggesting that the adoption of geographically-diverse mobile applications is linked to a dichotomy of cities and rural areas.

II. Methodology

In order to derive our results, we model the amount of traffic generated by mobile services in different communes as jointly distributed random variables, and adopt an approach that hinges on information theory. We first assess the global geographic diversity of service usages in France, by computing the *mutual information* $I(C; S)$ of mobile service demands ($\mathcal{S}$) and geographical locations ($\mathcal{C}$). The mutual information captures how much can be inferred about the consumed services by knowing the commune, and we find it to be close to zero in our nationwide scenario, implying that mobile services demand distributions are homogeneous across the whole of France.

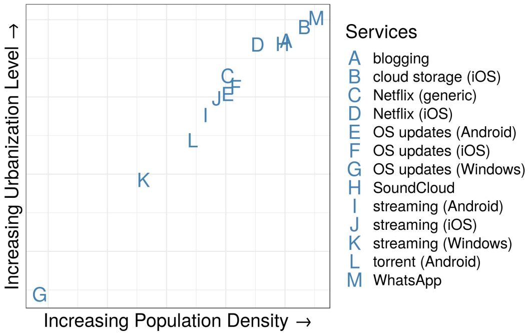


Fig. 1. Relative positioning of informative mobile services in the space of population density and urbanization levels. People living in rural regions of France have a preference to use Windows Mobile devices. Instead, inhabitants of metropolitan areas prefer Apple iPhones, and display a remarkable tendency to significantly use WhatsApp and long-lived streaming services like Netflix.

However, usage distributions are highly skewed, with a few services generating the vast majority of traffic. Thus, the mutual information primarily captures the (absence of) diversity in the consumption of popular apps. We investigate if, among less popular services, there exist some that are *informative*, *i.e.*, are characterized by a non-negligible diversity of usage across geographical areas. In information theoretical terms, this can be formulated as the combinatorial problem of identifying a subset $\mathcal{S}' \subset \mathcal{S}$ of all mobile services whose mutual information with the spatial areas is maximized, *i.e.*,

$$\mathcal{S}'_{\text{opt}} = \arg\max_{\mathcal{S}' \subseteq \mathcal{S}} I(C|_{\mathcal{S}'}; S|_{\mathcal{S}'}).$$

An efficient approximation of the solution is provided by the Blahut-Arimoto algorithm [1], [2], which maximizes $I(C; S)$ by weighting all services according to their informativeness. By applying the approach to our scenario, 13 informative services are singled out. They are listed in Figure 1.

In order to understand *where* the informative services differ in usage, we cluster communes in France based on how their inhabitants consume such services. To this end, we measure the distance between two communes i and i' as the loss of mutual information $d(i, i') = I(K; S) - I(K_{i,i'}; S)$ incurred when they are merged [3]. Here $K_{i,i'}$ is the set of (clusters of) communes in K such that $i \in K$ and $i' \in K$ are merged. We feed such distances to a scalable two-phase greedy hierarchical clustering algorithm, and find that a significant portion of the system information is captured by a small number of clusters, *i.e.*, mobile service consumption patterns. Interestingly, the geographical distribution of these usage patterns over the French territory is strongly correlated with population density and urbanization levels, as in Figure 1.

References

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," IEEE Trans. on Inf. Theory 18(1), 1972.

[2] R. Blahut, "Computation of channel capacity and rate-distortion functions," IEEE Transactions on Information Theory 18(4), 1972.

[3] P.S. Chodrow, "Structure and information in spatial segregation," Proceedings of the National Academy of Sciences 114(44), 2017.

`scikit-mobility`: an open-source Python library for human mobility analysis and simulation

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

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

The availability of geo-spatial mobility data (e.g., GPS traces, call detail records, social media records) is a trend that will grow in the near future. In particular, this will happen when the shift from traditional vehicles to autonomous, self-driving, vehicles, will transform our society, the economy and the environment. For this reason, understanding and simulating human mobility is of paramount importance for many present and future applications, such as traffic forecasting, urban planning, estimating migratory flows, and epidemic modeling, and hence for many actors, from urban planners to decision makers and advertising companies.

`scikit-mobility` is a Python library for human mobility analysis and simulation. In particular, the library allows the user to: (1) analyze mobility data by using the main measures characterizing human mobility patterns (e.g., radius of gyration, daily motifs, mobility entropy, origin-destination matrices); (2) simulate individual and collective mobility by executing the most common human mobility models (e.g., gravity and radiation models, exploration and preferential return model); (3) compare all these models by a set of validation metrics taken from the literature. `scikit-mobility` provides an efficient and easy-to-use implementation, based on the standard Python library numpy, pandas and geopandas, 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, `scikit-mobility` implements the gravity model [1, 2], the radiation model [3], and their recent improvements [4]. Among the individuals mobility models, i.e., models generating synthetic trajectories of desired length for a set of agents, `scikit-mobility` provides the recent improvements of the classical Exploration and Preferential Return model (s-EPR, r-EPR, d-EPR and recency EPR) [5–7] as well as the most accurate spatio-temporal generative human mobility models like DITRAS [8] and TimeGeo [9]. During the talk practical examples on how to use the library on real-world mobility data and generate synthetic trajectories will be shown.

`scikit-mobility` is a starting point for the development of urban simulation and what-if analysis [10, 11], 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.

Session Methods

MobiSenseUs: Predicting Objective and Subjective Well-Being from Aggregate Mobile Data

Martin Hillebrand[1,2], **Imran Khan**[1], **and Nuria Oliver**[1]

[1] Vodafone Research, London, UK
[2] Correspondence: hillebrand_martin@gmx.net

Assessing the socio-economic status and the well-being of a population is of utmost importance for policy and decision makers so they can design appropriate policies and interventions aimed at improving the quality of life of their citizens. Traditional methods to assess such measures consist of surveys which are expensive to obtain and difficult to scale. Hence, the national statistics offices of most developed countries only carry out such surveys every few years.

In the case of the United Kingdom, the Office for National Statistics (ONS) publishes two important metrics: the Index of Multiple Deprivation (IMD), composed of seven variables (income, employment, health, education, barriers to housing and services, crime and living environment) and obtained every three to five years; and the Subjective Well-being (SWB), composed of 4 metrics (life satisfaction, worthwhile, happiness and anxiety) and assessed every year.

In recent years, the wide adoption of mobile phones has led to several research works that estimate socio-economic and well-being indicators from pseudonymized, aggregated mobile data both in developed (1) and developing (2, 3) countries.

The research described in this paper also focuses on the automatic estimation of socio-economic and well-being from mobile human behavioral data, with three key contributions when compared to previous work: Firstly, our analysis covers a wide spectrum of human behavior by including geographically aggregated communication patterns (e.g. number of calls), spatial mobility (e.g. commuting distance) and app usage behavior (e.g. number of apps) based on a pseudonymised dataset of more than one million mobile phone users in the UK over a period of 6 months (Jan-Mar 2017 and Jan-Mar 2018). Secondly, we build predictive models of subjective well-being (SWB) using the mobile behavioral data. And thirdly, we carry out a longitudinal analysis of SWB by applying in 2018 a model that was trained with 2017 data. To the best of our knowledge, we are the first to carry out such a longitudinal analysis.

In terms of ground truth, we use the most recent IMD survey which is from 2015 and the two SWB surveys from 2017 and 2018. Note that SWB is available with more frequency which allows the use of the two most recently published datasets. Our spatial granularity is the UK local authorities (391 local authorities).

Regarding the mobile data, we compute 16 features that characterize the aggregated communication, mobility and app usage behaviors at the local authority level. We train supervised state-of-the-art machine learning models (Gradient Boosted Trees, Support Vector Machines,…) to automatically infer SWB and IMD. We only report the results for Linear Support Vector Machines (using 5-fold cross-validation) using 8 weeks worth of data (mid-Jan to mid-Mar 2017) as this combination consistently outperformed the other models. Figure 1 shows the four quartile classification performance of our models when applied to both the Life Satisfaction facet of SWB and the IMD average score. As seen in the Figure, IMD can be predicted better ($Acc = 66.1\%$) than SWB ($Acc = 46.4\%$). When we only want to separate the top quartile from the lowest quartile, analogous to the methodology applied in (4), accuracies rise to 99% for IMD and 84% for SWB.

The better performance for IMD might be due to the fact that economic activities are better reflected in the behavioral patterns captured by the mobile phone data than self-reported subjective well-being. Post-hoc partial correlational analyses reveal that third variables like median age and population density of a region account for a substantial share of the correlation between phone behavior features and well-being targets.

While spatial mobility features are particularly important to predict IMD, communication features play the most important role for SWB. This is consistent with the fact that social relatedness is a key element to mental well being (5).

Next, we aim to predict SWB from the available mobile data instead of conducting expensive surveys (6) thus assessing the potential value of mobile data to replace or complement existing methodologies. We build a regression model with the mobile features and SWB data from 2017 and apply the same model to the mobile features from 2018 to automatically infer SWB in 2018. Note that we are unable to carry out the same analysis for IMD as we only have one sample of IMD for 2015. We obtain a $R^2 = 18\%$ (compared to $R^2 = 27\%$ in 2017), showing that a model trained with data from the previous year could be re-used the next year.

Our empirical findings allow a direct comparison between different aspects of human behavior and their impact on the objective and subjective well-being of a region. In addition, we show that the correlational stability between mobile features and well-being targets allows the re-use of trained models across years.

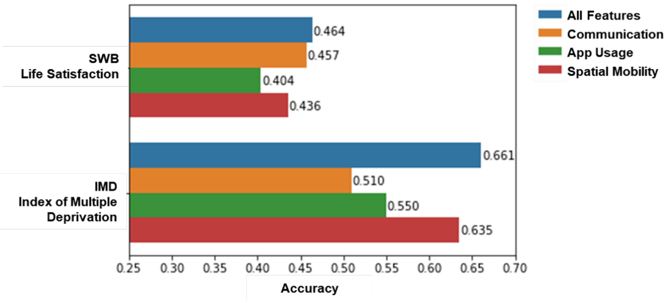


Fig. 1. Accuracy of a four quartiles classification task with different sets of features

Bibliography

1. Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328 (5981):1029–1031, 2010. ISSN 0036-8075. doi: 10.1126/science.1186605.
2. Joshua Blumenstock and Nathan Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM, 2010.
3. Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 377–388. Springer, 2011.
4. Chris Smith-Clarke, Daniele Quercia, and Licia Capra. Finger on the pulse: identifying deprivation using transit flow analysis. In *CSCW*, 2013.
5. Edward L Deci and Richard M Ryan. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian psychology/Psychologie canadienne*, 49(3):182, 2008.
6. Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '15, pages 254–264, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675233.

Road crash hotspot identification through bystander Tweets

Guadalupe Bedoya[a], Arianna Legovini[a], Robert Marty[a], Sveta Milusheva[a], Elizabeth Resor[b], Sarah Williams[c] [1]

Introduction

Road injuries are the 8th leading cause of death globally. This development challenge is especially prominent in sub-Saharan Africa, with Kenya being one of the worst affected countries. It ranks among the top countries in death rate due to road traffic crashes (RTC) in the world with 29 deaths per 100,000. Yet, the available statistics are unreliable: the WHO estimates that the number of RTC fatalities is 3.7 times the official registry, a challenge faced in most low-income countries. Better data and scientific analysis on where crashes happen, and factors associated with crash hotspots are needed to improve evidence and policy to address the growing road safety issues in low-income countries.

Leveraging on crowdsourcing, machine learning and a unique multi-year geolocated crash police dataset we built for Nairobi, this paper demonstrates the use of mobile phone applications for informing road safety. We build on existing literature that geocodes events from tweets (Al-Olimat et al 2017, Gelernter and Mushegian 2011, Gelernter and Balaji 2013, Gu et al 2016, Middleton et al 2014, Qian 2016, Zhang and Gelernter 2014)[2], innovating by developing additional techniques to clean a gazetteer of landmarks and developing heuristics to prioritize certain landmarks over others if multiple are found in a tweet. This is the first application of these tools for road safety in a developing country.

Data

The crowdsourced data comes from Ma3Route, a mobile/web/SMS platform that crowd-sources transport data and provides users with information including traffic and accidents with around 500,000 subscribers in Nairobi. Individuals can either directly Tweet to Ma3Route or use the Ma3Route mobile phone application to post when they see a road traffic crash (or traffic). All 734,795 Ma3Route tweets from 2012-2018 are scraped and used for this analysis. The Police data consist of paper reports of crashes aggregated from the 14 police stations in Nairobi for 2011-2018. 11,186 police reports are manually digitized and geolocated based on the location from the report.

Methodology

We use a two-step machine learning algorithm for identification of crash locations from crowd-sourcing (similar to Qian 2016). The first step categorizes tweets to identify those that relate to a crash based on a dictionary of words associated with crashes. This initial dictionary was developed using a training dataset of tweets identified as relating to crashes. In order to geolocate crash locations, ideally we would use the GPS associated with the tweet, but due to the heavy battery usage and data costs associated with turning GPS on, in 99.9% of cases there is no GPS location. Therefore, in step two, using a dictionary of landmarks in Nairobi developed by (1) scraping landmark names and locations from Google Maps and (2) simplifying and cleaning landmark names (building on techniques outlined in Al-Olimat et al. 2017), we identify the landmarks in the tweet that allow us to geolocate the crash. If multiple landmarks are found in a tweet, we employ a series of heuristics to choose which landmark to geocode.

To assess the quality of the data and precision of the algorithm we (i) built a truth dataset using double manual entry of tweets reported in an entire year, and (ii) worked with a motorcycle delivery mobile phone platform to verify crashes identified via Twitter. For the physical verification, tweets were processed through the algorithm in real-time and triggered a "delivery" in the platform that alerted the closest motorcycle driver to go to the location of the crash location identified by the algorithm. We compare the crowdsourced data and police data, for the years they overlap. We develop a clustering algorithm to identify crashes in the police data and in the crowdsourced data that are the same based on location and timing.

Results

Using crowdsourcing, we identify 5,727 unique crashes from 2012-2018. Our verification processes have shown that our algorithm was able to geocode 68% of crash-related tweets, where 93% of tweets were physically verified to have had a car crash near the estimated location. Comparing the crowdsourced data with police data, we find that 33% percent of crowdsourced cases are new cases not found in the police data. Our approach was able to capture many crashes not identified by administrative data sources. Our study shows how the use of machine learning, crowdsourcing and data analytics can contribute to cost-effective improvements in road crash data.

[1] We thank Shin Bin Tan, Andres Achury and Scott Margeson for their work on the first version of the algorithm, and Robert Tenorio for his excellent support in Nairobi to build the Police dataset and managing the process to physically verify crashes derived from crowdsourcing. a) World Bank b) UC Berkeley c) Massachusetts Institute of Technology; Contact: smilusheva@worldbank.org; gbedoya@worldbank.org.

[2] Al-Olimat H, Thirunarayan K, Shalin V and Sheth A. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. CoRR 2017.; Gelernter J and Balaji S. An algorithm for local geoparsing of microtext. GeoInformatica. 2013;17(4) 635-667.; Gelernter J and Mushegian N. Geo-parsing Messages from Microtext. Transactions in GIS. 2011;15(6): 753-773.; Middleton S, Middleton L, Modafferi S. Real-Time Crisis Mapping of Natural Disasters Using Social-Media. IEEE Intelligent Systems 2014;29(2), 9-17.; Qian, Z. Real-time Incident Detection Using Social Media Data. The Pennsylvania Department of Transportation 2016.; Gu Y, Qian Z, Chen F. From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies 2016;67, 321-342.; Zhang W, Gelernter J. Geocoding location expressions in Twitter messages: A preference learning method. JOSIS 2014;9.

Inactivity prediction using Convolutional Neural Networks

Elisabetta Alfonsetti, Paola Jafrancesco, Francesco Calabrese, Carlo Baldassi, Riccardo Zecchina
Vodafone Research, Bocconi University
Email: {Elisabetta.Alfonsetti, Paola.Jafrancesco, Francesco.Calabrese}@vodafone.com {carlo.baldassi, riccardo.zecchina}@unibocconi.it

***Abstract*—One of the major challenges of mobile operators is the rate at which customers go into inactivity state. There is not a specific definition on when a customer will become inactive.Indeed, this is not related to a specific event, like mobile number portability or deactivation request [1, 2], but it represents a change in customers' usage behaviour. In particular, in this paper, we refer to an inactive customer as a customer with no outbound calls, sms and data traffic for a period of 30 days. As the days of inactivity increase, it becomes more difficult to reach the customers with prevention activities. Hence, knowing whether a customer is likely to enter into inactivity state is crucial if we want to prevent it with targeted actions. In this paper we show how to address this problem using Deep Convolutional Neural Networks (CNN). In particular, traffic variables related to outbound calls, sms and data traffic are treated as images to perform inactivity prediction by leveraging deep learning architectures prominent in image classification. Experimental results show that CNN outperform traditional approaches.**

Our data corresponds to anonymized usage and profile data from a sample of 200k telecom prepaid customers. We structure the training dataset in two set of features: the temporal and the static ones. The temporal dataset represents information about user activities like data traffic, voice (outgoing and incoming) and sms usage. In order to encode this information as an image we collect 84 days of data - corresponding to roughly 3 months - where each column is a day and each row is a feature, ending up with a two-dimensional array of normalised pixels. The static dataset, instead, contains monthly aggregations of features about customer behaviour such as top up, balance events, digital channels interactions, marketing campaigns, to name a few. The label is defined as shown in Figure 1. Customers who do not perform any outgoing traffic event in the 30 days observation period are labeled as 1. In order to exclude customers who are already partially inactive (and thus difficult to reach with targeted actions), we label as 0 those customers with no traffic events in the last 14 days.

Our network architecture consists of three main components: a set of convolutional layers for temporal features, an identity layer for static features and a final step of fully connected layers, as shown in Figure 2. The first part is made up of 4 convolutional layers, the first layer involves 64 filters of size 7x3, with padding of 1 and stride of 7, followed by a second layer with 64 filters of size 7x7 in order to go across 7 days of customer usage. A third layer with 64 filters of size 1x4 and a last one of 1x3 are then used in order to analyse the temporal data in a monthly and then 3-monthly basis. The output of the convolutional layers is then merged with a static inputs layer containing all the static features and both of them are used as input of a fully connected layer architecture, composed by 2 dense layers of dimension 100 and 80, respectively. The output is a softmax node to predict the final probability of scoring a customer as positive (inactive) or negative (still active). ReLu activation functions and AdaDelta optimizer are selected to train the network. We refer to this entire architecture as CNN-DL, while the part composed by only static inputs layer and the fully connected layers is named as DL architecture.

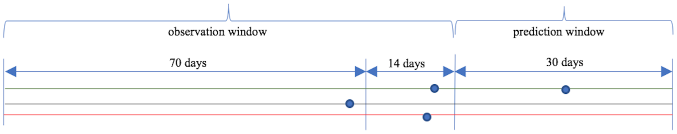


Fig. 1. Three different types of customer behavior where the dots represent outgoing traffic events. The green line represents a customer who is active for all the observation period and is still active in the prediction window, thus he will be labeled 0; the black line is an example of customer to be removed from the training dataset as he is already not doing any traffic 14 days before the prediction window; the red line is an example of customer who will become inactive in the prediction window, so will be labelled as 1.

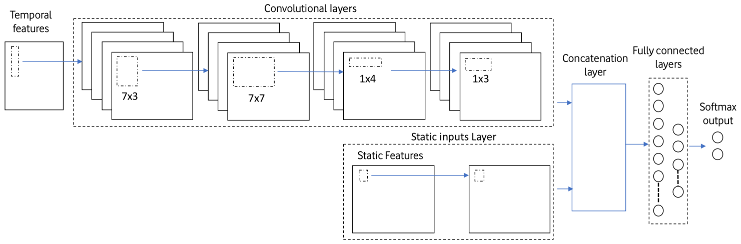


Fig. 2. CNN architecture used for inactivity prediction.

We compared our CNN-DL architecture with an Xgboost model (trained with both static and dynamic features) and with a deep learning (DL) architecture trained only using static features. For the evaluation and comparison of models split the dataset in 80% training 20% test, and use the lift on the top 10% of the population. Results presented in the table below show the improvement obtained using the CNN.

Lift top 10%	Train	Test
XGB	7.03	5.48
DL	6.23	4.51
CNN - DL	9.18	7.15

References

[1] M. F. Abdillah, J. Nasri, and A. Aditsania. Using deep learning to predict customer churn in a mobile telecomunication network. *eProceedings of Engineering*, 3(2), 2016.

[2] C. Yang, X. Shi, L. Jie, and J. Han. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *ACM SIGKDD*, pages 914–922. ACM, 2018.

Measuring the Relationship between Apps and Urban Context

Eduardo Graells-Garrido
Universidad del Desarrollo
Santiago, Chile

Diego Caro
Universidad del Desarrollo
Santiago, Chile

Rossano Schifanella
University of Turin
Turin, Italy

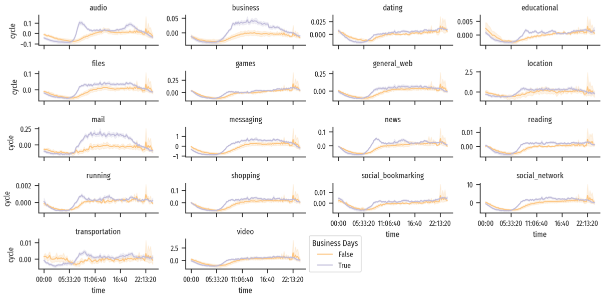


Figure 1: Application traffic during the period of study in Santiago, Chile. The traffic has been de-trended and averaged in two different periods: laboral days and weekends.

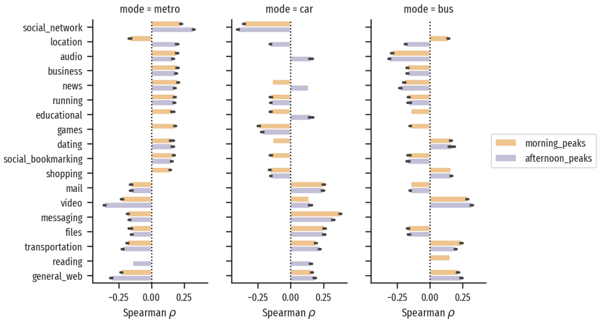
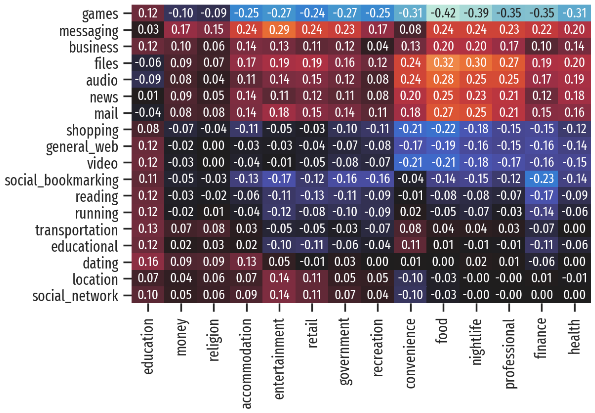


Figure 2: Correlations between application usage rates and association to specific modes of transportation per tower.



	education	money	religion	accommodation	entertainment	retail	government	recreation	convenience	food	nightlife	professional	finance	health
games	0.12	-0.10	-0.09	-0.25	-0.27	-0.24	-0.27	-0.25	-0.31	-0.42	-0.39	-0.35	-0.35	-0.31
messaging	0.03	0.17	0.15	0.24	0.29	0.24	0.23	0.17	0.08	0.24	0.24	0.23	0.22	0.20
business	0.12	0.10	0.06	0.14	0.13	0.11	0.12	0.04	0.13	0.20	0.20	0.17	0.10	0.14
files	-0.06	0.09	0.07	0.17	0.19	0.19	0.16	0.12	0.24	0.32	0.30	0.27	0.19	0.20
audio	-0.09	0.08	0.04	0.11	0.14	0.15	0.12	0.08	0.24	0.28	0.25	0.25	0.17	0.19
news	0.01	0.09	0.05	0.14	0.11	0.12	0.11	0.08	0.20	0.25	0.23	0.21	0.12	0.18
mail	-0.04	0.08	0.08	0.14	0.18	0.15	0.14	0.11	0.18	0.27	0.25	0.21	0.15	0.16
shopping	0.08	-0.07	-0.04	-0.11	-0.05	-0.03	-0.10	-0.11	-0.21	-0.22	-0.18	-0.15	-0.15	-0.12
general_web	0.12	-0.02	0.00	-0.03	-0.03	-0.04	-0.07	-0.08	-0.17	-0.19	-0.16	-0.15	-0.16	-0.14
video	0.12	-0.03	0.00	-0.04	-0.01	-0.05	-0.08	-0.07	-0.21	-0.21	-0.18	-0.17	-0.16	-0.15
social_bookmarking	0.11	-0.05	-0.03	-0.13	-0.17	-0.12	-0.16	-0.16	-0.04	-0.14	-0.15	-0.12	-0.23	-0.14
reading	0.12	-0.03	-0.02	-0.06	-0.11	-0.13	-0.11	-0.09	-0.01	-0.08	-0.08	-0.07	-0.17	-0.09
running	0.12	-0.02	0.01	-0.04	-0.12	-0.08	-0.10	-0.09	0.02	-0.05	-0.07	-0.03	-0.14	-0.06
transportation	0.13	0.07	0.08	0.03	-0.05	-0.05	-0.03	-0.07	0.08	0.04	0.04	0.03	-0.07	0.00
educational	0.12	0.02	0.03	0.02	-0.10	-0.11	-0.06	-0.04	0.11	0.01	-0.01	-0.01	-0.11	-0.06
dating	0.16	0.09	0.09	0.13	0.05	-0.01	0.03	0.00	0.01	0.00	0.02	0.01	-0.06	0.00
location	0.07	0.04	0.06	0.07	0.14	0.11	0.05	0.05	-0.10	-0.03	-0.00	-0.00	0.01	-0.01
social_network	0.10	0.05	0.06	0.09	0.14	0.11	0.07	0.04	-0.10	-0.03	-0.00	-0.00	0.00	0.00

Figure 3: Correlations between application usage rates and the availability of several categories of POIs around towers.

ABSTRACT

The adoption of user-generated content and interaction logs has been prominent in the measurement of how people and places relate [4]. Understanding this relationship can help to design and inform place and policy design. The ubiquity of smart phones, and the wide range of activities supported by mobile applications, may allow to understand how people use their time. However, there is still a gap in the knowledge of what kind of activities are performed in different urban contexts, as there is little information about how such contexts shape mobile application usage.

In this abstract, we propose a study about how the urban context shapes what people do on their mobile phones. As urban context, we consider daily activities such as commuting, and physical places such as Points of Interests (POIs). To analyze these contexts, we rely on a Deep Packet Inspection data set covering two weeks of IP requests in Santiago, Chile [1]. This data set contains the aggregated number of requests per mobile phone tower at the top country-wise 5,000 popular IP addresses by the largest telecommunications operator in the country. We matched those IPs to app. categories to characterize their daily traffic (*c.f.* Fig. 1), finding that app. usage varies with respect to laboral days and weekends (*e.g.*, mail is more used in laboral days), as well as commuting times (*e.g.*, music and news are correlated with rush hours). Then, using the results from prior work that associated mobile phone towers to mode(s) of transportation using XDR [2]), we estimated rank-correlation coefficients between app. usage rates (adjusted by floating population) and mode(s) of transportation (bus, metro, and car), at two times of the day: morning and afternoon rush hours (*c.f.* Fig. 2). Results include patterns such as the usage of social networks and games with metro, and transportation applications with car and bus (*e.g.*, ride hailing and public transport). Finally, we estimated rank-correlations between app. usage rates and POIs available in OpenStreetMap (*c.f.* Fig. 3). Results include that *entertainment* areas are more correlated with messaging apps. than others, and that *food* places are inversely correlated with games, among others.

Our results confirm the intuition that place and time shape what people do on their phones, in line with theories about time use such as *equipped time* [3]. Thus, understanding this behavior may help to design transportation equipment, context-aware recommended systems, as well as to measure what people do with their time.

REFERENCES

[1] E. Graells-Garrido, D. Caro, O. Miranda, R. Schifanella, and O.F. Peredo. 2018. The WWW (and an H) of Mobile Application Usage in the City: The What, Where, When, and How. In *Companion of the The Web Conf. 2018*. Int. World Wide Web Conf. Steering Comm., 1221–1229.

[2] E. Graells-Garrido, D. Caro, and D. Parra. 2018. Inferring modes of transportation using mobile phone data. *EPJ Data Science* 7, 1 (2018), 49.

[3] Juliet Jain and Glenn Lyons. 2008. The gift of travel time. *Journal of transport geography* 16, 2 (2008), 81–89.

[4] D. Quercia, L.M. Aiello, and R. Schifanella. 2018. Diversity of indoor activities and economic development of neighborhoods. *PloS one* 13, 6 (2018), e0198441.

Empirical Evidence For Networks Effects Of Urbanisation, Fertility Transition, And Migration

Tamás Dávid-Barrett[a,b,c,d*]
Sebastian Diaz[a,e]
Loreto Bravo[b]
Carlos Rodriguez-Sickert[a]
János Kertész[f,g,h]

[a] Universidad del Desarrollo, Facultad de Gobierno, CICS, Av. Plaza 680, San Carlos de Apoquindo, Las Condes, Santiago de Chile, 7610658 Chile
[b] Trinity College, University of Oxford, OX1 3BH, Oxford, UK
[c] Kiel Institute for the World Economy, Kiellinie 66, D-24105 Kiel, Germany
[d] Population Research Institute, Väestöliitto, Kalevankatu 16, Helsinki 00101, Finland
[e] Data Science Institute, Universidad de Desarollo, Av. Plaza 680, Las Condes, 7610658, Santiago de Chile, Chile
[f] Central European University, Center for Network Science, Nador u. 9, Budapest, H-1051, Hungary
[g] Department of Computer Science, Aalto University School of Science, P.O.Box 15500, 00076 Finland
[h] Department of Theoretical Physics, Budapest University of Technology and Economics, H1111, Budapest, Hungary

***Corresponding author:**
Tamas David-Barrett, tamas.david-barrett@trinity.ox.ac.uk

A recently published paper (David-Barrett Sci Rep 2019) proposed a mathematical theory that links demographic processes, such as fertility transition, urbanisation, and migration, to a rise in norm violations, and the replacement of traditional social network based norm enforcement with formalised legal institutions. Here we report an empirical test to this theory using a combined mobile phone database, census data, and macro statistics. The size of the database is 5.6 million people, from Chile. We found that the data is consistent with all causal predictions of the theory at a very high level of significance. The reported magnitudes are as follows: urbanisation, fertility, and migration explains ~20% of the variation of the clustering coefficient among individuals, which, in turn, explains 48% of the variation in crime among administrative units. These results validate the theory, and offer immediate policy implications.

Multichannel Social Signatures and Persistent Features of Ego Networks

Sara Heydari[1], Sam G.B. Roberts[2], R.I.M. Dunbar[3], J. Saramäki[1,4]
sara.heydari@aalto.fi, s.g.roberts1@ljmu.ac.uk, robin.dunbar@psy.ox.ac.uk, jari.saramaki@aalto.fi
[1]) Department of Computer Science, Aalto University, Espoo, Finland
[2]) School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, UK
[3]) Department of Experimental Psychology, University of Oxford, UK
[4])Helsinki Institute of Information Technology HIIT, Aalto University, Espoo, Finland

The structure of egocentric networks reflects the way people balance their need for strong, emotionally intense relationships and a diversity of weaker ties. Egocentric network structure can be quantified with 'social signatures', which describe how people distribute their communication effort across the members (alters) of their personal networks. Social signatures based on call data have indicated that people mostly communicate with a few close alters; they also have persistent, distinct signatures [1].

However, social relationships are shaped and maintained through a diversity of communication channels. People do not use these channels uniformly–rather, the choice of channel depends on many factors. To examine if the properties of social signatures are generalizable and genuine features of egocentric networks, it is therefore important to look at data from multiple channels of communication, both separately and together.

Combining information on different channels can, however, be problematic because of their intrinsic differences. For example, the number of calls or their total duration is typically used as a proxy for tie strength in mobile telephone call data. But text messages, another common form of communication via mobile devices, have no duration, and the number of text messages between an ego-alter pair is not directly comparable to the number of calls between that pair. In this study [2], we develop a way of constructing comparable call and text-message ego networks as well as mixed ego networks using both channels (see Fig. 1). We observe that the social signatures made from all these types of ego networks display persistent individual differences that remain stable despite the turnover in individual alters. We also show that call, text, and mixed signatures resemble one another both at the population level and at the level of individuals. The consistency of social signatures across individuals for different channels of communication is surprising because the choice of channel appears to be alter-specific with no clear overall pattern, and ego networks constructed from calls and texts overlap only partially in terms of alters. These results demonstrate individuals vary in how they allocate their communication effort across their personal networks and this variation is persistent over time and across different channels of communication.

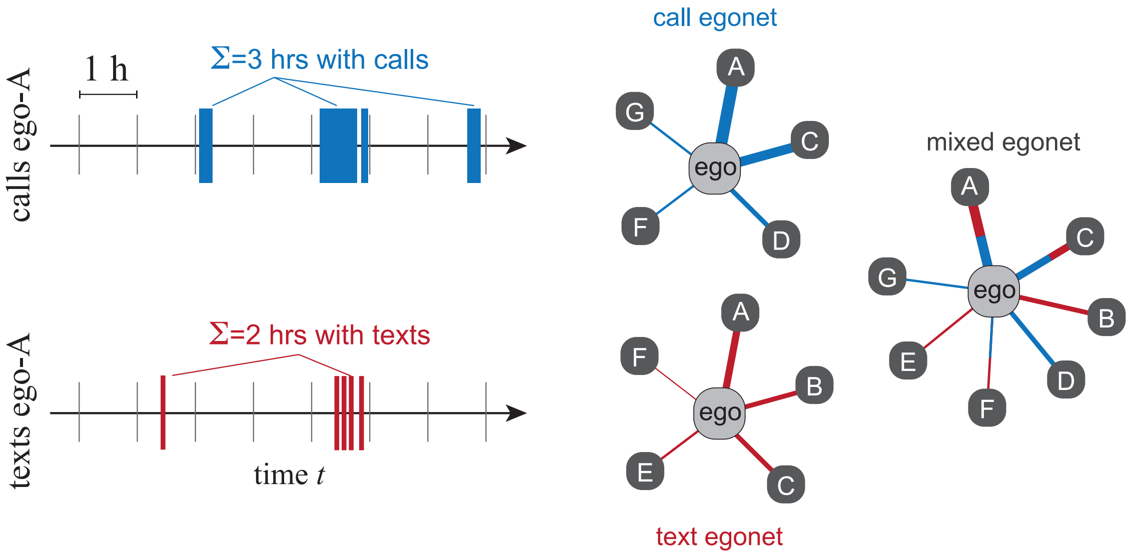


Figure 1: To construct egocentric networks from call and text data, we define for each ego-alter pair the link weight as the number of 1-hour time-bins that contain at least a communication event of the desired type. Weights defined this way are better comparable than numbers of calls and texts whose "units" do not match, as one conversation which requires dozens of text messages can be carried out in one call. For mixed ego-nets, we simply count the number of bins that have at least one event of either type.

References

[1] Jari Saramäki, Elizabeth A Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[2] Sara Heydari, Sam G Roberts, Robin IM Dunbar, and Jari Saramäki. Multichannel social signatures and persistent features of ego networks. *Applied Network Science*, 3(1):8, 2018.

Computable reachability of limited waiting-time processes in large temporal networks

Arash Badie Modiri[1,*], Márton Karsai[2,†], Mikko Kivelä[1,‡]
[1]Department of Computer Science, School of Science, Aalto University, Finland
[2]Univ. Lyon, ENS de Lyon, Inria, CNRS, Université Claude Bernard Lyon 1, LIP UMR 5668, F-69342, Lyon, France.
[*]arash.badiemodiri@aalto.fi [†]marton.karsai@ens-lyon.fr [‡]mikko.kivela@aalto.fi

In many spreading processes a spreading agent may have a limited lifetime δt: like in case of transportation networks with a maximum acceptable transfer time; in social networks where information may become outdated or forgotten; or in case of diseases where the infectious period ends after a certain amount of time. These problems, concerning limited ($\delta t < \infty$) waiting time processes, have been previously studied in temporal networks by simulating the process from a sample of initial nodes and time instances. This approach limits the analysis to either very small networks, or to average statistics (as opposed to event-level statistics or to statistics of the tails of distributions) [1]. To alleviate this problem, recently the *event graph* representation has been proposed [2, 3], with weakly connected components (WCC) giving an upper bound on the number of events (resp. nodes) what a spreading process can follow (resp. reach) [2]. However, as WCCs of event graphs cannot determine the exact reachable set from a node at a given time, the detection of out-components appeared as an open challenge so far.

In this contribution, we present a set of algorithms based on probabilistic counting that allow us to simultaneously measure the number of nodes and events that can be reached from all different starting points and times in a temporal network. Our method works accurately for very large networks, which we demonstrate via the estimation of reachable set of nodes and events from all possible initial conditions in a large mobile phone call network with $\sim$ 320M events (Fig. 1) [2] and in a Twitter mention network of $\sim$ 270M interactions [4]. Further, our method can find the event, which reaches the largest fraction of the network (largest out-component in the event graph) with high adjustable probability. Note that the reachability without limited waiting time ($\delta t = \infty$) appears as a special case here and can be solved as well with our algorithms.

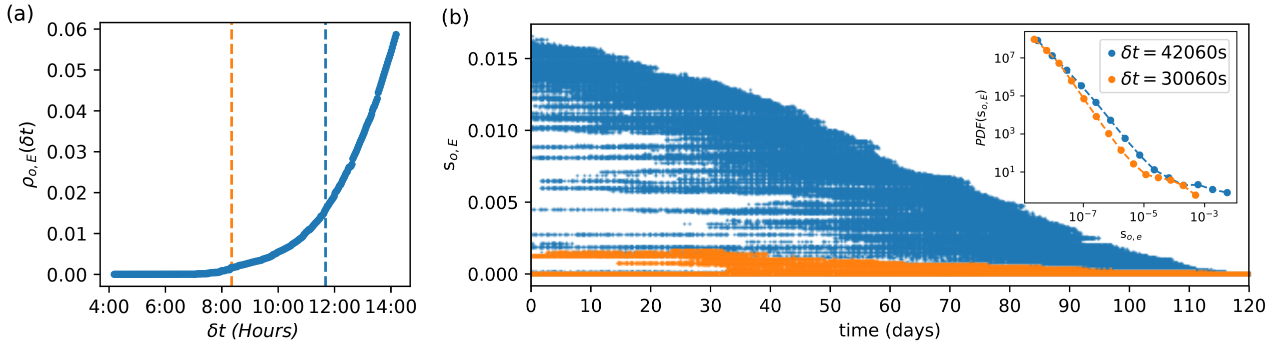


Figure 1: Statistics of paths of δt-connected events calculated for all $\sim$ 320M possible source events in the mobile phone communication network. The sizes of reachable sets are shown as numbers of events. (a) The maximum fraction of reachable events from any starting event $\rho_{o,E}$ (maximum out-component size in event graphs) with varying limited waiting time δt. The probability of missed detection of the correct initial condition is at most 10^{-2}. (b) Fraction of reachable network, $s_{o,E}$, from (all) any event in different times. One point assigns one source event. A sample of 10% of points are plotted due to limitations of plotting software. (b inset) Distribution of number of δt-reachable events from all starting events (i.e. the out-component size distribution in event graphs).

Our work has several advantages as compared to the conventional initial condition sampling approach. It can be used to accurately calculate the tails of the reachability and spreading distributions and it can answer completely new questions on temporal network data, such as, what is the exact maximum number of nodes that can be infected via a spreading process. It can also be used to calculate node/event level statistics, which may lead to new kinds of importance and centrality measures. Further, it opens up a way to analyse percolation phenomena in temporal networks. For example, instead of resorting to upper-bounds via WCC calculations (and lower bounds via sampling), we can now exactly measure the critical parameters of the temporal network unfolding as a directed percolation, or a spreading process evolving on the top of it.

[1] Holme, P. (2005). Network reachability of real-world contact sequences. *Phys. Rev. E*, 71(4), 046119.

[2] Kivelä, M., Cambe, J., Saramäki, J., & Karsai, M. (2018). Mapping temporal-network percolation to weighted, static event graphs. *Sci. Reps.*, 8(1), 12357.

[3] Mellor, A. (2017). The temporal event graph. *J. Complex Networks*, 6(4), 639-659.

[4] Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 177-186). ACM.

Clean up or mess up: the effect of sampling biases on measurements of degree distributions in mobile phone datasets

Adeline Decuyper[1], Arnaud Browet[1], Vincent Traag[2], Alexey Medvedev[*1,3], Vincent D. Blondel[1], and Jean-Charles Delvenne[1]

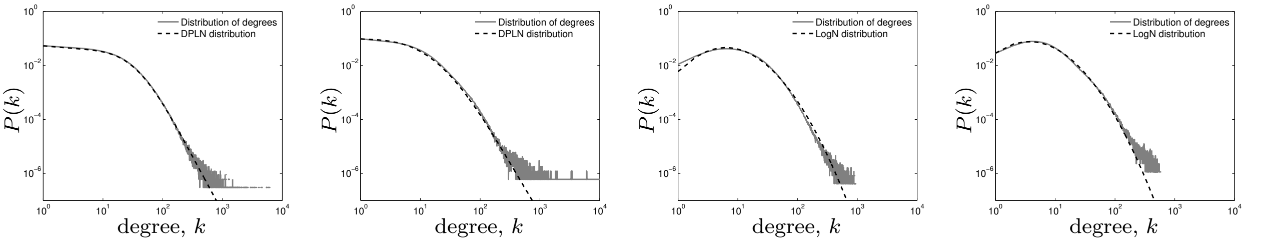
[1] *Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain, Louvain-La-Neuve, Belgium*
[2] *CWTS, Leiden University, Leiden, The Netherlands*
[3] *Université de Namur, Namur, Belgium*

Mobile phone data have been extensively used in the recent years to study social behavior. However, those studies are inherently based on partial data, often covering a subset of the population, a specific time frame or a single country, as mostly provided by a single telecommunication company. Reconstruction of a communication network from mobile phone data is one of the primary tools for analysis of the client audience. The results does not show to be universal across all datasets. For example, in one of the first studies on mobile phone data[*], looking at one day of data, it was observed that the reconstructed network corresponded to a random graph model with a *Power Law* degree distribution. However, in a later study [†], this time using a longer period of one month of data, the authors observed that the mobile call graph had a degree distribution corresponding to a *Double Pareto LogNormal* (DPLN)[‡]. In our work, we point to an observation that the bias due to the limited coverage in time may have an important influence on the results of the analyses performed.

Additionally, we observe significant differences, both qualitatively and quantitatively, in the degree distribution of the network, depending on the way the dataset is pre-processed. Many researchers start by preprocessing the data, often called "cleaning", removing links or nodes that are not active enough. For example, it is common to remove all links that are non-reciprocated[§] or to impose a threshold on activity between a pair of users (e.g. at least n communications in a Δt time interval)[¶]. These apparently innocuous filtering methods applied before the analysis of the network generate additional biases that are often overlooked.

In our work, we study two large databases of mobile phone calls in Belgium (6 months) and in Portugal (15 months). We find that by drawing a link between two users as soon as they have communicated at least once, we obtain a network with a DPLN degree distribution. On the contrary, if we only take into account the users that remain in the network during the whole observation period, thus removing all users that joined or left the network during that time frame, the degree distribution changes from a DPLN to a LogNormal. We validate this observation both in empirical and theoretical scenarios for a wide range of observation periods and present a possible stochastic model of degree evolution based on geometric brownian motion with stopping time T, which corresponds to the time between the first and last observations of activity of the user, representing the time during which the degree of the user could grow and be observed in our dataset. If the observation of this process is stopped at a random time T (which may come from a wide range of distributions), then $X(T)$ follows a DPLN distribution. On the contrary, when T is constant and we observe users over a time window of fixed length, then we observe $X(T)$ distributed as LogNormal [3].

Our observations raise a point about the effects of sampling the datasets and may help in the future to create better synthetic datasets offering a closer correspondence with empirical data, as these results reveal characteristics of mobile phone datasets that may have been overlooked in the past.



Belgium, no filter | Portugal, no filter | Belgium, filtered | Portugal, filtered

Figure 1: Degree distributions in empirical networks of mobile phone communications in Belgium and Portugal for all users and filtered for only active users in the observation window.

[*] corresponding author, email: an_medvedev@yahoo.com

[*] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In Proceedings of the thirty-second annual ACM symposium on Theory of computing, pp. 171–180. ACM, 2000.

[†] M. Seshadri et al. Mobile call graphs: beyond power-law and lognormal distributions. In Proceeding of the 14th ACM SIGKDD, pp. 596–604. ACM, 2008.

[‡] W. J. Reed and M. Jogensen. The double pareto-lognormal distribution - a new parametric model for size distributions. Communications in Statistics-Theory and Methods, 33(8) pp. 1733–1753, 2004.

[§] J.P. Onnela et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 104 (18):7332, 2007

[¶] R. Lambiotte et al. Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications, 387(21):5317– 5325, 2008.

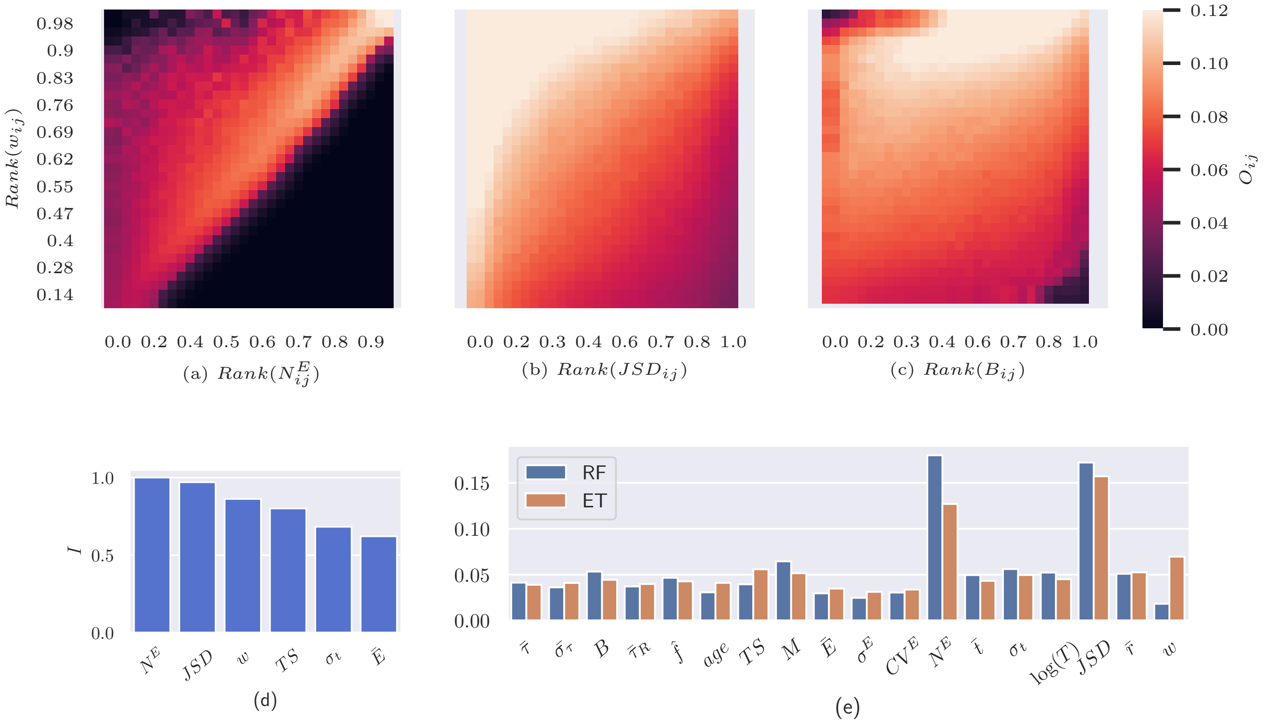
Going beyond communication intensity for estimating tie strengths in social networks

Javier Ureña-Carrion*, Jari Saramäki, Mikko Kivelä
Department of Computer Science, School of Science, Aalto University, Finland
*javier.urenacarrion@aalto.fi

Construction of social networks from auto-recorded communication data, such as mobile phone calls, has led to a revolution in our understanding of social systems. In such studies, the strength of a social tie between two people is almost exclusively calculated as the number or the total time of the contacts. This is due to convention and convenience rather than a justified choice, and discards all the information on timings of the interactions.

Here, we think of *the tie strength as a latent variable which we try to predict based on properties of the time sequences of the dyadic interactions.* Our assumption is that the latent tie strength is expressed as the structural overlap in the social network (independently of interaction times), as formulated by the Granovetter's hypothesis [1, 2]. With this assumption, we can use the time sequences to predict the overlap instead of the latent tie strength—the better the predictor is for overlap, the better it is for the latent tie strength.

We analyze a mobile phone call dataset of ~ 6.5 million people during a period of 4 months, and obtain the topological overlap with an extended network of ~ 77 million users. We compute temporal measures that go beyond intensity features, including variables derived from the inter-event time distribution, daily and weekly patterns, and temporal stability during the observation window. We find that *the topological overlap, measured both in a static and dynamic manner, is correlated with a myriad of tie-level dynamics.* For example, the number of bursty cascades [3] between two nodes (N^E_{ij}) seems to be more suitable than the total number of calls (w_{ij}) for overlap prediction. In addition, we found that differences in the daily outgoing call distribution are also highly correlated to overlap, suggesting behavioral homophily in tie-level dynamics. Our results strongly suggest that there are better proxies for tie strength than the widely used communication intensity. As some of these measures are simple and easy to interpret, they are ready to be adopted for future research on social networks based on communication data.



(a) $Rank(N^E_{ij})$ (b) $Rank(JSD_{ij})$ (c) $Rank(B_{ij})$

(d) (e)

Figure 1 – (a, b, c) The average topological overlap for links which have a given value of communication intensity (number of calls, w_{ij}) and a candidate temporal feature F, $\langle O_{ij} | w_{ij}, F \rangle$. The three temporal features are: (a) Number of bursty cascades $F = N^E_{ij}$. (b) Jensen-Shannon Divergence for node-level daily activity distributions, $F = JSD_{ij}$, where we compare the fraction of outgoing calls each individual makes at each hour of the day. (c) The burstiness coefficient $F = B_{ij} = \frac{\sigma^{\tau}_{ij} - \bar{\tau}_{ij}}{\sigma^{\tau}_{ij} + \bar{\tau}_{ij}}$, a measure of uncorrelated bursty behaviour [4]. In all three cases the inclusion of a new variable adds information not available by using only w_{ij}. (d) Scaled mutual information between overlap and top five variables, where w_{ij} ranks third. (e) Feature importance for overlap prediction problem using Random Forests (RF) and Extra Trees (ET). In both cases, N^E_{ij} and JSD_{ij} outweight most other variables, including communication intensity w_{ij}, as its effect is diminished when N^E_{ij} is in the model.

[1] M S Granovetter, *The strength of weak ties*, Social networks 347-367. (1977)
[2] J-P Onnela et al. *Structure and tie strengths in mobile communication networks*, PNAS 104.18 7332-7336 (2007)
[3] M Karsai, el tal. *Universal features of correlated bursty behaviour*, Scientific Reports, 397 (2012)
[4] K-I Goh, A-L Barabási. *Burstiness and memory in complex systems*, EPL 81.4 48002 (2008)

Session Epidemics

HUMAN MOBILITY SHAPES THE HIV EPIDEMIC IN NAMIBIA

Eugenio Valdano[1], Justin Okano[1], Vittoria Colizza[2], Sally Blower[1]
[1]Center for Biomedical Modeling, Semel Institute of Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles, CA 90024, USA. [2]INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique IPLESP, F75012 Paris, France.

Namibia has a severe HIV epidemic. The infection is endemic in the general population, and there is significant geographic variation in prevalence (6%-40% for women, 0%-24% for men). Along with other countries in sub-Saharan Africa (SSA), Namibia has implemented extensive interventions aimed at preventing infections. However, all of them are based on the implicit assumption that all risk is localized: people get infected by members of their own communities. We argue that, in order to eliminate HIV in SSA (a goal set by the World Health Organization), a key ingredient is still missing: human mobility. The Namibian population is highly mobile. Consequently, individuals may be at risk of acquiring HIV both in their home community (from other residents or from visitors from other communities) or when visiting other communities (from residents of these communities). We quantify the importance of these different types of risk in the different geographic areas, and compute the various flows of risk importation and exportation among them. We use the network of mobility among the 106 Namibian constituencies (administrative areas), inferred from mobile phone CDRs. These data were collected between October 2010 and September 2011, and comprise 9 billion communications from 1.19 million unique SIM cards. These data were used to calculate a mobility network based on the average proportion of time residents spent in each of the 96 constituencies in Namibia that include cell towers. We couple the mobility network to demographic data from the 2011 census and HIV-testing data from 7,600 participants of the NDHS (Namibia Demographic & Health Survey). We use the coupled epidemic-mobility network to calculate, for each constituency, the risk of acquiring HIV that was due to sexual contact with other residents (i.e., localized risk), with visitors, or when traveling. Notably, our results show that mobility does not substantially change (in any area of the country), the number of new infections that would occur in the absence of mobility, but it substantially "redistributes" risk. Hence, mobility changes which individuals become infected and where they acquire infection. Our results demonstrate that it is essential to design specialized interventions that, informed by mobility patterns, reach the groups who are at highest risk of getting infected, and increase adherence to treatment of the infected individuals.

Comparing Sources of Mobility Information for Modelling the Spread of Zika in Colombia

Daniela Perrotta[1*], Enrique Frias-Martinez[2], Miguel Luengo-Oroz[3], Daniela Paolotti[1], Michele Tizzoni[1], Alessandro Vespignani[1,4]

(1) ISI Foundation, Turin, Italy (2) Telefonica Research, Madrid, Spain (3) United Nations Global Pulse, New York, USA (4) Northeastern University, Boston, USA

*Corresponding author for this abstract, daniela.perrotta@isi.it

Human mobility plays a central role in the spatial spread of human infectious diseases. Accurate data on human mobility is therefore key to properly design epidemic models that allow to timely assess the spatial propagation of infectious diseases and to evaluate appropriate control measures and intervention strategies. In this context, mobile phone data, in the form of Call Detail Records (CDRs), implicitly brings a large ensemble of details on human activity, including human movements, which can be identified for example whenever the same phone number is handled by two different mobile phone towers in two consecutive phone calls. Appropriately harnessing such digital traces left by mobile phone users' activity thus represents a relatively low-cost resource to draw a high-level picture of human mobility patterns at an unprecedented scale.

In this study, we focus on the human mobility patterns relevant to the epidemic spread of Zika during the recent outbreak occurred in Colombia in 2015-2016. The aim of the study is to investigate the potential predictive power gained by integrating the human mobility derived from CDRs data into an epidemic modelling approach, as well as more traditional methods, such census data and mobility models. Specifically, more than two billion encrypted and anonymized calls made by around seven million mobile phone users in Colombia have been used to reconstruct the aggregated mobility network describing the flows of people travelling daily among different locations. Moreover, we apply both the gravity model and the radiation model to synthetically infer the population movements in Colombia and generate the corresponding mobility networks.

From the networks' point of view and by considering the census data as a reference, our results show that the gravity model strongly underestimates the mobility flows, whereas the mobility determined by the radiation model and the CDRs data show comparable performances in terms of high correlations and good similarities metrics with census data, thus reproducing well the actual mobility between the 33 administrative units (i.e. departments) in Colombia.

To model the spread of Zika, we employ a metapopulation modelling approach to explicitly simulate the spatial spread of Zika as governed by the transmission dynamics of the virus through human-mosquito interactions and as promoted by population movements across the country. Notably, our approach integrates detailed data on the population, the spatial heterogeneity of the mosquito abundance and the consequent exposure of the population to the virus due to socio-economic factors in order to help provide a more realistic representation of the epidemic progression.

Given the same modelling settings (i.e. initial conditions and parameters), our approach allows to perform numerical simulations of the spatio-temporal progression of the disease by integrating one mobility network at a time to ultimately assess their predictive power by comparing the simulated epidemic profiles with the Zika cases officially reported by the National Institute of Health in Colombia.

Telecom data in public health: building generic prediction models for transmission risk

Jérôme Urbain[1], Kristýna Tomšů[1], Jonathan Frisch[1,2], Astrid Van Lierde[1], Rositsa Zaimova[1], Christophe Bocquet[1]

[1]Dalberg Data Insights, Brussels, Belgium, jerome.urbain@dalberg.com

[2]Apptweak, Brussels, Belgium

Introduction. Over that last decade, a number of studies have shown the relevance of human movement data extracted from mobile phone records to estimate the transmission risk of infectious diseases [1, 2, 3]. With the support of the United States Agency for International Development (USAID), Dalberg Data Insights has conducted two projects aiming at 1) making such approaches available for health authorities combatting zika and 2) generalizing the proposed methods to various diseases and locations, with a single, parameterizable process. We have also explored machine learning modelling to predict the disease incidence over a few weeks in the future. In both cases the outputs of the models were made easily accessible to the end-users through data visualization dashboards.

Generalization of the risk estimation. The generalization part was primarily developed in West Africa, through discussions with local health experts and USAID staff. Taking Tatem's model as a starting point, which estimates the risks of transmission from a place A to a place B by looking at visitors from A to B and visitors from B to A [2], we have generalized it to also consider any location C where residents from A and B could meet. Our risk transmission model is:

$$RiskExport^{t}_{\mathrm{A\to B}} = \sum_{C=1}^{R} \beta^{t}_{\mathrm{C}} \alpha_{BC} \mathrm{N_B} \frac{\mathrm{\alpha_{AC}} I_A}{\sum_{k=1}^{R} \mathrm{\alpha_{kC}} N_k}, \text{ where:}$$

- R is the number of areas,
- $\mathrm{N_k}$ is the population living in area k,
- $\mathrm{I_k}$ is the incidence of the disease in area k,
- β^{t}_{k} is the transmission rate of the disease in area k at time t,
- α_{ij} is the percentage of time residents of area i spend in area j.

The proportion of time spent at each location is estimated using Call Detail Records. Depending on the considered disease, the relevant mobility metrics vary. For instance, for malaria we are interested in the number of nights spent at each place, as malaria is transmitted over night. Dengue, on the other hand, is transmitted by mosquitoes that are active during day time. For other diseases that are transmitted directly from human to human, such as measles or Ebola, all movements should be considered.

Predicting future incidence. For the predictive part, we worked with a ministry of health in Brazil on zika and dengue. In addition to monitoring the evolution of the diseases and the prevention activities conducted by the ministry, we i) estimated the import and export risks as explained above; ii) predicted the incidence over the next 4 weeks. To do so, we trained different models (random forest, linear regression with Lasso, linear regression with Ridge). The models were fed with the past incidence values, population, altitude and surface area, for the target area, the 10 areas that have the biggest mobility flows with the target area and the 10 areas showing the biggest importation risk to the target area (using the import/export risk estimation already explained). Models were trained, validated and evaluated on disjoint data sets of 14, 3 and 4 months, respectively. These models were combined using simple linear regression, allowing to significantly reduce the mean squared error and achieve reliable predictions, as shown on Figure 1.

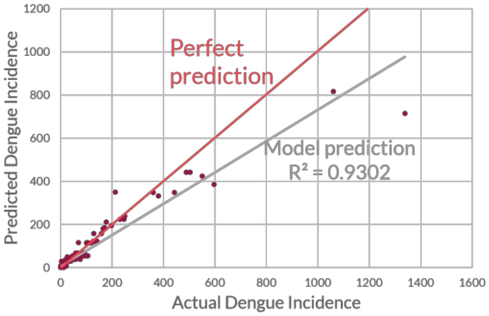


Figure 1: Prediction of dengue incidence

The approach was well received by the ministry of health as i) the prediction part enables to better target their future interventions; and ii) the model remains relatively simple to explain. In addition, the model implicitly integrates time- and location-based factors (e.g., altitude, season), reducing the need for model parameters (β^{t}_{k}) that are hard to accurately estimate, even by experts, beyond trends. The next step is to extend the training data to 10 years in order to capture more complex patterns (e.g., “vaccination” effect when an area was recently hit by the epidemy).

BIBLIOGRAPHY

[1] A. J. Tatem, Y. Qiu, D. L. Smith, O. Sabot, A. S. Ali and B. Moonen, "The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents," *Malaria journal*, vol. 8, p. 287, 2009.

[2] 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 others, "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning," *Malaria journal*, vol. 13, p. 52, 2014.

[3] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsoy, M. A. Johansson, S. B. Rasheed, K. Engo-Monsen and C. O. Buckee, "Impact of human mobility on the emergence of dengue epidemics in Pakistan," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 11887-11892, 2015.

Exploiting Population Activity Dynamics to Predict Urban Epidemiological Incidence

Gergana Todorova, Lancaster University and Anastasios Noulas, New York University

Changes in lifestyle as well as intensifying urbanisation and travel not only pose an increased health risk, but add extra pressure to emergency services that operate day and night in the world's largest cities. In the United Kingdom, the number of health related incidents that require urgent response has been steadily increasing, resulting to an unprecedented number of calls that require an ambulance response [2]. As a result, during 2015 only 5 out of 11 ambulance trusts operating in England have met their operational target to respond to 75% of most urgent calls within eight minutes [1]. Incidents such as *cardiac arrest* require immediate attention by paramedic specialists, with ambulance response times being critical to patient survival rates [5]. Consequently, intelligence on optimally deploying paramedic crews across geographic areas and also with appropriate timetabling becomes a significant challenge for ambulance services worldwide.

To meet the goal of timely responding to urgent health incidents ambulance organisations over the past decade have been moving towards integrating computational methods and data science practices in their planning operations [3]. Such practices include the recruitment of analysts and data scientists, the development of data warehousing infrastructure, as well as the use of software and cartography tools that support planning allowing for the identification of optimal paramedic crew scheduling [4]. However the functionality of the class of optimisation tools that simulate emergency event activity is often opaque. They are being frequently packaged in proprietry software becoming effectively a black box for end users [2]. In the meantime, scholarly work in environmental epidemiology has been focusing more on accurate statistical modelling of ambulance calls, rather than providing plausible interpretations on the socio-economic and urban activity factors that may drive those. Models developed in this setting have used historic frequency information about ambulance calls for instance [1, 6] ignoring socio-economic or real time population activity as well as mobility indicators that could also determine the risk for emergency incidents.

In this work we seek to establish a stronger link between the prevalence of health incidents and spatio-temporal population activities, aiming to bridge the aforementioned gap in the literature. Critically for its novelty, we model activities of local populations using the digital traces emitted by mobile users that engage with location intelligence services like Foursquare. We demonstrate how mobile web indicators of geographic user activity become a useful proxy to understanding the times of the day and days of the week certain urban activities become prominent across city neighborhoods. Moreover, we show how those align across space and time with certain types of ambulance inci-

[1] https://www.bbc.com/news/health-33166204

[2] https://www.intermedix.com/solutions/response-planning-0

Special Session
Data for Refugees

Optimizing the Access to Healthcare Services in Dense Refugee Hosting Urban Areas: A Case for Istanbul

M. Tarik Altuncu m.altuncu@imperial.ac.uk, Ayse Seyyide Kaptaner ayseyyide@gmail.com, and Nur Sevencan nur.sevencan@trtworld.com

Abstract

With over 3.5 million refugees, Turkey continues to host the world's largest refugee population. This introduced several challenges in many areas including access to healthcare system. Refugees have legal rights to free healthcare services in Turkey's public hospitals. With the aim of increasing healthcare access for refugees, we looked at where the lack of infrastructure is felt the most. Our study attempts to address these problems by assessing whether Migrant Health Centers' (MHC) locations are optimal. The aim of this study is to improve refugees' access to healthcare services in Istanbul by improving the locations of health facilities available to them. We use the cellular network usage data provided by Turk Telekom [3] to figure out where optimal locations for MHCs are. Our location optimization cuts the average distance and travel duration for refugees need to travel.

We limited the study to Istanbul as it is one of the densest refugee hosting urban area in Turkey. We then followed the following procedure:

- Estimation of the residential location of refugees in the data set based on their calling activities during nighttime
- Calculating the number of refugee residents per Voronoi cell region
- Clustering Voronoi cell regions to come up more representative scale of regions [1]
- Aggregating the number of refugee residents to the new scale
- Optimizing the MHC locations using the multi-facility location optimization linear programming model [2] with cost function below:

$$\min_x \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \cdot d_{ij} \cdot x_{ij}$$

where n is the number of residential regions, a_i is the number of refugee residents in $i^(th)$ region, d_{ij} is public transit travel distance or duration between $i^(th)$ and $j^(th)$ regions , and x_{ij} is binary decision variable of MHC assignment between regions.

MHC Locations	Travel Distance	Travel Duration
Current	5.9	26
Optimized (Distance)	3.6	20
Optimized (Duration)	4.4	18

Table 1. Cost of travel distances and durations for current MHC locations and optimized MHC locations (on the basis of distance and duration). All distances are provided in kilometers and durations in minutes.

References

1. Arthur, D., Vassilvitskii, S.: K-means++: The Advantages of Careful Seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 1027–1035. SODA '07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007)
2. ReVelle, C.S., Swain, R.W.: Central Facilities Location. Geographical Analysis. **2**(1), 30–42 (1970)
3. Salah, A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y., Dong, X., Dağdelen, .: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey (7 2018)

Acknowledgements: We thank the rest of the members of TRT World's team for the Data for Refugees challenge.

Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey

Michiel A. Bakker[1], Daoud A. Piracha[1], Patricia J. Lu[1], Keis Bejgo[1], Mohsen Bahrami[1,2], Yan Leng[1], Jose Balsa-Barreiro[1], Julie Ricard[3], Alfredo J. Morales[1], Vivek K. Singh[4], Burcin Bozkaya[2], Selim Balcisoy[2], and Alex 'Sandy' Pentland[1]

[1] MIT Media Lab, Cambridge, Massachusetts, USA {bakker,pentland}@mit.edu
[2] Sabanci University, Istanbul, Turkey
[3] DataPop Alliance, New York, USA
[4] Rutgers University, New Jersey, USA

The Syrian civil war that began in 2011 has had an enormous human cost and impact on the region. UNHCR has estimated that over 12 million people fled their homes; 6.6 million Syrians are internally displaced while 5.6 million people fled Syria. As the Syrian conflict has intensified and lengthened, many refugees have faced challenges integrating into their host societies. Here we introduce and evaluate different measures of integration extracted from mobile phone metadata in Turkey, the country that hosts the largest number of Syrian refugees. Ground truth labels revealing the refugee status of a phone's user allow us to study integration of refugees along three dimensions: (1) social integration (2) spatial integration and (3) economic integration. A strong social network has proven to be instrumental in finding housing, employment and healthcare. Nonetheless many prior studies, relying on sparse self-reported data, have failed to accurately measure social interactions. This study measures social integration directly from the data as the refugee to local call volume. Spatial integration, traditionally measured through census data, is measured for each area using the spread of refugees across close-by cell towers. For individual refugees, we measure spatial integration by tracking movement throughout the city while computing likelihood of encountering a local citizen. Finally, even though we do not have access to ground truth economical integration, we infer signs of formal employment through a user's regularity in commuting patterns.

Where previous studies rely on survey data and interviews, we show that integration can be measured and understood directly and at scale through mobile phone data. For example, measurements of spatial integration show that refugees in Istanbul live in more integrated neighborhoods than in Southeastern Anatolia, close to the Syrian border. Additionally, our toolset allows us to observe the important interactions between different dimensions of integration. This leads to interesting insights such as a contrast between strongly correlated social and spatial integration in Istanbul while these same measures are uncorrelated elsewhere in Turkey. The relationships between measures are an important first step towards a more fine-grained understanding of the dynamics behind integration, needed to inform policy decisions and interventions. Finally, leveraging the results from two general elections in Turkey in 2015 and 2018, we confirm earlier findings concerning the impact of refugee presence on voting behavior, and demonstrate that we can better explain voting behavior by incorporating integration metrics. On the whole, this work marks one of the first systematic attempts at employing fine grained mobility to understand refugee integration and impact of integration at scale.

Improve Education Opportunities for Better Integration of Syrian Refugees in Turkey

Marco Mamei[1], Seyit Mümin Cilasun[2], Marco Lippi[1], Francesca Pancotto[1], and Semih Tümen[3]

[1]University of Modena and Reggio Emilia, Italy
[2]Central Bank of the Republic of Turkey , Turkey
[3]Department of Economics, TED University, Turkey
marco.mamei@unimore.it, marco.lippi@unimore.it, seyit.cilasun@tcmb.gov.tr,
francesca.pancotto@unimore.it, semihtumen@gmail.com

Turkey is the largest refugee hosting country worldwide with 3.5M registered Syrian refugees. Considering that Syrians will continue to live in Turkey for many years, and that a significant portion will not turn back to their country, their integration in the Turkish society is crucial.

As more than one third of the Syrian population residing in Turkey is composed of school-age children, schooling and education have a major role in the integration process.

However, from the perspective of any hosting society, mass refugees inflow represents a huge challenge. Refugees generate competition for several resources and many public services get congested. Host countries have limited education resources (schools and teachers), which makes policy making and resource re-allocation a major challenge.

One of the main assets to address these complex challenges is to have up-to-date fine-grained information about refugees and their activities. Call Detail Records (CDRs) are mobile phone data that allow to track activities of refugees and natives at a fine grained scale, thus representing a natural response to this need.

In this project we use mobile phone data to analyze some of the challenges to integrate Syrian refugees in the Turkish education system and the impact on the Turkish society. Educational institutions do not simply transmit human capital, they also pass on social capital in the form of social rules and norms. So, investments in education will not only work toward increasing Syrians' school enrolment rate, they will also impact their integration in the society.

We analyze the distribution of refugees across the country and their possible impact on education facilities. We study logistic obstructions to schooling of Syrian children, and propose and evaluate an optimization mechanism to identify areas where new education resources are required considering the needs of both refugees and natives. Figure 1(a) illustrates provinces where education services are more congested. Figure 1(b) shows the results of an optimization process trying to plan the construction of new schools where most needed according to natives and refugees locations as measured from CDRs.

We also analyze the relationship between education and social integration, and the impact of Syrian refugees' schooling on social integration (via numbers of calls between natives and Syrians). Figure 1(c) shows the correlation between children and calls between natives and refugees. We also consider the impact of Syrian refugees on the education choices of natives and their impact on Turkish economic development

The full version of this report can be found on D4R Challenge web site. *http://d4r.turktelekom.com.tr*

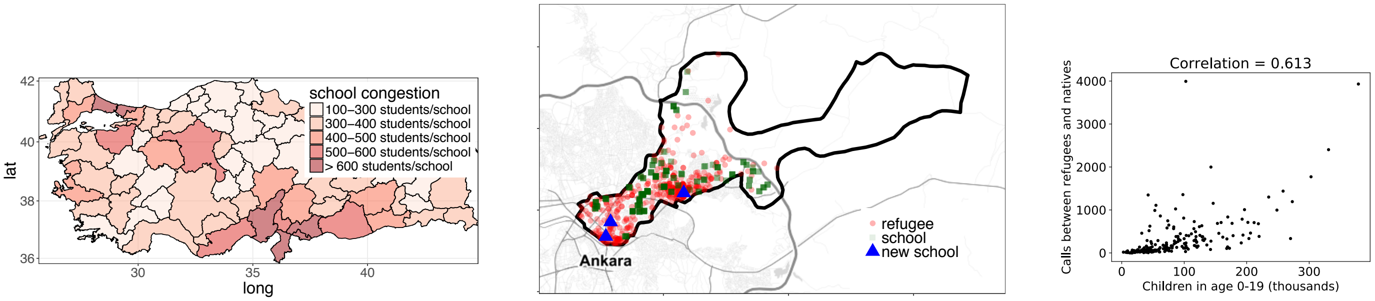


Figure 1: a) School congestion at province level; b) Refugees, existing schools, and where to build new schools in the district of Altindag. c) Correlation between number of children and number of calls between natives and refugees.

Abstract for the netmob19 conference (D4R Challenge Session)

Title:

AROMA_CoDa: Assessing Refugees' Onward Mobility through the Analysis of Communication Data

Authors:

Harald Sterly, University of Vienna, harald.sterly@univie.ac.at

Benjamin Etzold, Bonn International Center for Conversion, benjamin.etzold@bicc.de

Lars Wirkus, Bonn International Center for Conversion, wirkus@bicc.de

Patrick Sakdapolrak, University of Vienna, e-mail: patrick.sakdapolrak@univie.ac.at

Jacob Schewe, Potsdam Institute for Climate Impact Research, jacob.schewe@pik-potsdam.de

Carl-Friedrich Schleussner, Climate Analytics gGmbH, carl.schleussner@climateanalytics.org

Benjamin Hennig, University of Iceland, ben@hi.is

Abstract:

Data availability remains one of the greatest challenges in migration and mobility research, especially for internal migration, and in countries with and contexts of data scarcity. We present a project in which the analysis of mobile phone data (Call Detail Records) has provided insights into the secondary mobility of refugees in Turkey.

Spatial and temporal trajectories at different scale levels have been derived from an anonymous data set that maps the talk time, location and duration of more than 50,000 mobile phone subscribers in Turkey over the course of a whole year, and that was provided in the course of the Data for Refugees (D4R) Challenge. The major migration trends in Turkey are clearly visible: from South to North, from East to West and from rural to urban areas. However, the analysis also provides insights that have not yet been clearly established from existing data, such as specific routes and central nodes in mobility networks; seasonal, cyclical and event-dependent mobility; or the return migration of refugees to the Syrian border. Cultural mobility motives (e.g. visits during the Ramadan feast) are more important, and seasonal mobility to regional agricultural labour markets seems less important than originally expected.

The analysis of CDR - even with a high degree of anonymisation - enables the description of mobility and migration patterns in an unprecedented richness of detail, both temporally, spatially and across scales. However, the complexity of the mobility patterns that emerges in the process requires even more in-depth analyses. The results do also show that the boundaries of established concepts, such as commuting, seasonal or permanent migration can be much more fluid than is often assumed.

Keywords: Call data record, mobility, social integration, Turkey, refugees' movements

Posters

Extensive and Detailed Nationwide Information on Human Mobility from Mobile Data in the Netherlands

Peter van der Mede, Joost de Bruijn, , Klaas Friso, Stefan de Graaf, Jakob Henckel: DAT.Mobility, Deventer, Netherlands; pvdmede@dat.nl
John van de Kooij, Mezuro, Weesp, Netherlands;
Marcel Bijlsma, Frank Ophuis: Mobidot, Enschede, Netherlands
Matthijs de Gier, Frans van der Horst, Vincent Kuijpers, , Kantar, Amsterdam, Netherlands

In 2013 we presented a paper as part of the D4D challenge using Ivory Coast CDR data from Orange. In a proof of concept we showed that it is possible to build a gravity based transport model for a whole country and major cities based on CDR data and the open streetmap (OSM) network.(Nanni et all, 2013).

In the present paper we will show two new developments in creating extensive and detailed (Dutch) nationwide information on human mobility, which is based on CDR/EDR and A-GPS mobile phone data.

First, in co-operation with Mezuro and by using Vodafone CDR/EDR data we have created a national origin-destination (OD) matrix. This matrix comprises of movements between 1,250 zones, based on CDR/EDR from 4-5 million mobile phones . After weighing these data to get national figures, there still are a number of limitations in such a CDR/EDR based OD matrix. The main problem is that the granularity of this OD matrix lacks detail for transport modelling on lower levels, e.g. regional and local roads. Another limitation is that intrazonal movements are lacking in this matrix. Last, due to overlapping gsm cells movements over distances < 7 kilometers are underrepresented in the matrix. To overcome these limitations we developed a number of procedures in which we use socio-economic census data (e.g. inhabitants and work places per zone, and national travel survey data). Using these procedures we have created a 30,000 zone matrix, which includes intrazonal movements, estimates the missing short trips, and which does have enough detail for transport modelling purposes. Next, we assigned this improved OD-matrix to the complete Dutch transport network (1 million links), and we compared the road volumes resulting from this traffic assignment to the volumes in current state-of-the-art regional and national transport models. The results of these comparisons show that the developed methodology leads to high quality a priory OD-matrices, and thus traffic volumes, which make it much easier to create up-to-date transport models.

Though a distinction between train and ‘other modes’ can be made from the raw CDR/EDR data, more detailed insights into travel behaviors, including routes, multimodal trips, trip modality and purpose, travel times and speeds, dwelling time etc. need other measuring methodologies. Such information is highly wanted for all kinds of monitoring and evaluation purposes. Next to actual information on traffic volumes on all roads we would need better insight in the composition of these flows, modal splits, trip-chains etc.
So, secondly, in co-operation with Mobidot and Kantar we initiated the Dutch Mobility Panel. This longitudinal panel is drawn from the 100,000 members panel of whom 20-25% are willing to share data on their location for research purposes. Panel members can download the Kantar app (by invite only). Kantar uses this app for surveying their panel on a multitude of subjects. If a panel member consents in sharing location data, all outdoor movements are collected from their smart phone (iOS and android types) and stored in a database. Both the demographic profile data and the travel data of participants are merged into a real-time updated database.

Both the EDR/CDR based traffic volumes and the A-GPS based outdoor movements can be used in unison to create unprecedented detail of human mobility in the Netherlands.

Title: Distribution of technology (Mobile and Wifi) for total population of cities in India based on the rank size

Author: Jyothi Gupta
Affiliation: UCL, the Bartlett
Email address: ms.gupta.18@ucl.ac.uk

Abstract:

Urban System is explained as a horizon, large city- regions, global city economic relationship with international city. It is the connection between larger system, relationship between space and society. Gidden's (1974) explained us the procedure of nature sciences that is Unity of method between natural and social sciences. Cites are market place as an assumption with system of cities to minimize the transportation cost. The structure of cities could be concentric zones or pattern of pigeon holes in different orientations. India have measurable economic structure and self-sustain process and shares idealism with its religion, pattern of urban formulation, culture diversification.

	Social	Spatial	Relation
Global	Economic	Distance location	Development growth
Country	Culture	Physical Boundaries	State-wise religious planning
City	Social interaction	Scale &Planning	Division of Labour/Money Economy
local	structure	Infrastructure	Nature- sustainability

Urban problem could be stated as enquiry of a problem, social science concepts. Based on the population growth, the world is urbanizing. M.Batty et,(2012) explains Social Media and texting is the kind of information that can be transmitted through particular network being Twitter. Movement on public transport in large cities is increasingly and been collected for analysis and visualisation.

Algorithms are created for mapping the movements between modes and creating a origin and destination., its framework to master the complexity. Batty have identified seven types of initiative for smart city movement which is a summary of digitalization.

Defining the Zipf's law as inversely proportional to its rank in the frequency table, we have considered a first 20 cities in India with its population growth in year 2011. As per the data, we understand that Delhi being the highest populated city in the country compared to 20 other cities in the country. It is ranked at the higher scale and the lowest being Hugli. Basically, after providing a relation ratio of coefficient, data is gathered for analysis:

- Mobile connectivity in rural area
- Access to Internet growth
- Wireless devices and broadband connections with speed

Based on the data breakdown, Rural area growth for about 20 villages have provided the coefficient value as 0.51 whereas in the Urban population for highest growth of 20 cities have coefficient value as 0.88 with the regression line being more persistence. These reality gaps is addressed, and a real understanding of future needs could be prioritised as per the needs and better growth insights to create a technology accessible city.ISRO (2018) published the high speed connectivity in Rural India ranking Asia to have the fastest growth with statistics predicted in technology and connectivity.

Monitoring mobile money networks across providers and countries

Carolina Mattsson
mattsson.c@husky.neu
Northeastern University

Shafique Jamal
sjamal@ifc.org
IFC

Soren Heitmann
sheitmann@ifc.org
IFC

Guy Stuart
guystuart@mfopps.org
Microfinance Opportunities

Introduction Mobile money is an innovative digital financial service that has seen rapid expansion across Africa, South Asia, and Southeast Asia since 2009.[1] Providers support a digital channel for so-called e-money, providing payment processing and transfer services to users over the cellular infrastructure (ie. via SMS, USSD, or a mobile application) and servicing conversion with cash via a large cadre of on-the-ground agents who work on commission.[2]

While mobile money is growing in popularity and impact, providers have struggled to develop strong and sustainable networks among users of the service. This is due in part to high levels of inactivity, and to the widespread tendency for customers to withdraw their e-money into cash straight away rather than to keep it in their accounts or send it onward.[3] This is costly for providers. Servicing a transfer or payment uses scalable digital infrastructure, whereas servicing a cash-in or a cash-out entails compensating agents and rebalancing stocks of cash among them. Analysis tools that monitor trends in network use are of immediate interest.

Methods We introduce two measures of mobile money network strength that are directly comparable across countries and operators, while remaining interpretable to a non-technical audience: *transactions until exit (TUE)* and *proportion remaining in network after 72 hours (PRIN-72)*. These measures capture complimentary aspects of the extent to which e-money continues to circulate once it is in the digital system; we demonstrate using records from a provider in East Africa covering over 300 million transactions. We calculate TUE and PRIN directly from transaction records by “following” e-money according to last-in-first-out.[4]

Results We find that TUE and PRIN-72 discern important network changes. For example, we see TUE jumped in response to an intervention by the provider in September 2016 to address person-to-person fee evasion, which suppresses a particular kind of re-transaction. The intervention targeted large-value deposits, and appears to have been successful in raising the strength of the network surrounding such transactions. But, the effect was temporary elsewhere in the network. This intervention showed no effect on PRIN-72, confirming that it is a complimentary measure to TUE.

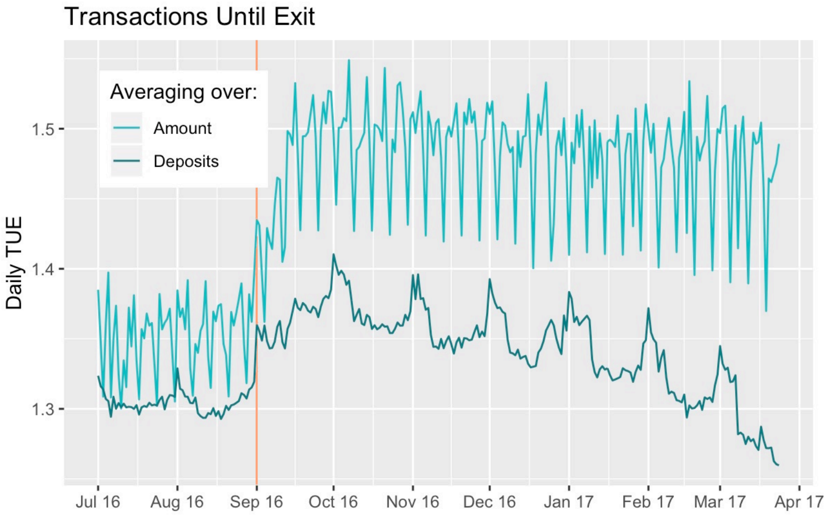


Figure 1: The average TUE for all deposits made on a given day. The darker line normalizes by the size of the deposit, giving all an equal weight. The lighter un-normalized line will reflect especially large-value deposits.

[1] GSMA Mobile Money. State of the Industry 2017: Mobile Money. Tech. Rep., GSMA (2017).

[2] Cull, R., Gine, X., Harten, S., Heitmann, S. & Rusu, A. B. Agent banking in a highly under-developed financial sector: Evidence from Democratic Republic of Congo. World Dev. 107, 54–74, DOI: 10.1016/j.worlddev.2018.02.001 (2018).

[3] Stuart, G. & Cohen, M. Cash In, Cash Out Kenya: The Role of M-PESA in the Lives of Low-Income People. The Financial Services Assesment project (Microfinance Opportunities, 2011).

[4] Mattsson, C. Follow the money: making sense of payment systems. *Under review.* (2019)

Inferring Context of Mobile Data Crowdsensed in the Wild

Rachit Agarwal, Shaan Chopra, Vassilis Christophides, Nikolaos Georgantas, Valérie Issarny
Inria-Paris, France
Email: {rachit.agarwal, shaan.chopra, vassilis.christophides, nikolaos.georgantas, valerie.issarny}@inria.fr

Understanding the sensing context of raw data is crucial for assessing the quality of large crowdsourced spatio-temporal datasets. Accelerometer's precision can vary considerably depending on whether the phone is in-pocket or out-pocket, i.e., held in hand [1]. GPS accuracy can be very low in places like under-ground metro stations [2]. Further, jump-lengths are shorter and have higher frequency when a person is in-door. Hence, we focus on contexts such as in/out-pocket, under/over-ground, and in/out-door that can be essential for reliably inferring human mobility attributes and properties (e.g., location, jump-length, and mobility activity like walking or driving) from crowdsensed data. Our work is motivated by the fact that most of the publicly available crowdsensing datasets (e.g. PRIVA'MOV [3] and Beijing taxi dataset [4]) do not include data from specialized sensors such as light, barometer, etc. considered by state-of-the-art algorithms for detecting the above mentioned contexts. Therefore, we focus on mining context from the limited features available in the publicly available mobility related crowdsensing datasets. Moreover, as ground truth is typically not available in these datasets, we pay special attention to minimizing the training or tuning efforts of the introduced algorithms. Our algorithms are unsupervised binary classifiers with a small memory footprint and execution time. As the lack of certain features prohibits us to consider state-of-the-art algorithms as baselines, we compare the performance of our heuristic algorithms against Machine Learning (ML) models built by an AutoML tool [5] using the same set of features. Our experimental evaluation with a segment of the Ambiciti [6] dataset demonstrates that when compared to the best baseline ML model w.r.t. balanced accuracy (see Table I), our algorithm for in/out-pocket performs equally well, while for under/over-ground and in/out-door contexts, for a specific hyper-parameter, our corresponding algorithms are within 4.3% and 1%, respectively. Concerning memory, our algorithms require 0kB, 4kB, and 0kB, respectively, while they take 0.08sec, 0.17sec and 0.003sec, respectively, for execution. Our algorithms are lightweight enough to be integrated into smartphone applications. Context information mined onboard thus remains private and can be used to annotate users' personal trajectories and incentivize them to participate in crowd-measurement campaigns.

TABLE I
ACCURACY, PRECISION, RECALL AND F1 SCORE REPORTED BY DIFFERENT METHODS.

			Accuracy in %						
			balanced	Precision		Recall		F1 score	
	Method	τ^{uo} in m	80-20 split	in	out	in	out	in	out
In/Out-Pocket	Gaussian NB†	-	54	0.33	0.78	0.19	0.89	0.24	0.83
	Heuristics	-	54	0.19	0.89	0.19	0.89	0.19	0.89
Under/Over-ground	Bernoulli NB†	313	74.5	0.33	0.97	0.81	0.68	0.46	0.79
	Heuristics	313	70.2	0.62	0.78	0.62	0.78	0.62	0.78
In/Out-door	Bernoulli NB†	313	66	0.42	0.84	0.70	0.62	0.53	0.71
	Heuristics	313	65	0.62	0.68	0.62	0.68	0.62	0.68

† TPOT reported method

REFERENCES

[1] E. Miluzzo, M. Pap, N. D. Lane, H. Lu, and A. T. Campbell, "Pocket, Bag, Hand, etc.- Automatically Detecting Phone Context through Discovery," in *First International Workshop on Sensing for App Phones (PhoneSense) at SenSys'10*, p. 21–25, 2010.

[2] K. v. Erum and J. Schöning, "SubwayAPPS: Using smartphone barometers for positioning in underground transportation environments," in *Progress in Location-Based Services 2016*, pp. 69–85, Springer International Publishing, oct 2016.

[3] S. Ben Mokhtar, A. Boutet, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D'Alu, V. Primault, P. Raveneau, H. Rivano, and R. Stanica, "PRIVA'MOV: Analysing Human Mobility through Multi-Sensor Datasets," in *NetMob: Book of Abstracts - Posters*, pp. 19–21, 2017.

[4] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving Directions Based on Taxi Trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, (New York, NY, USA), pp. 99–108, ACM, 2010.

[5] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, "Evaluation of a tree-based pipeline optimization tool for automating data science," in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, (New York, NY, USA), pp. 485–492, ACM, 2016.

[6] V. Issarny, V. Mallet, K. Nguyen, P. Raverdy, F. Rebhi, and R. Ventura, "Do's and Don'ts in Mobile Phone Sensing Middleware: Learning from a Large-Scale Experiment," in *Proceedings of the 17th International Middleware Conference*, (Trento), pp. 17:1–17:13, December 2016.

Characterization of cell activity in mobile networks

Angel Cuevas Muñoz
Universidad Carlos III
Madrid
acuevas@protonmail.com

Rubén Cuevas Rumín
Universidad Carlos III
Madrid
rcuevas@it.uc3m.es

Angel Cuevas Rumín
Universidad Carlos III
Madrid
acrumin@it.uc3m.es

This work aims to characterize the activity of the users in the cells of a large-scale mobile network. To this end, we analyze a dataset containing data from an operator offering nation-wide mobile service in a large country.

The questions addressed are simple: How many users are usually connected to each cell? Is there any temporal pattern in the usage of the cell? If so, can these patterns be used to predict the activity in the cell? Do all the cells have the same temporal usage pattern? Are there well-defined classes of cells in terms of the activity?

These questions are not just driven by curiosity but respond to basic needs in network planning and operation. Responsibles for planning need to know where the users of the network tend to be in order to suitably distribute budget. From the point of view of the operation, knowing the expected level of activity in each cell is crucial in prioritizing repairing tasks. There are many other issues for a mobile operator related to the need to know (and see) where the users are, to have a feel of where they tend to be, and to have a forecast of where they will be later on.

The starting point of our study is the strong spatial and temporal regularity that previous research has shown regarding the mobility of individuals throughout the mobile network.

However, our approach is somewhat different. While most previous research focus on the activity of individual users in the network or in aggregations at a much lower granularity (e.g. a whole city or even a country), we consider a different granularity scale and analyze activity at the cell level (i.e the operator's view).

We are also able to use a different and more detailed dataset recording each time a user has a transaction while being in any cell. We define a transaction here as any interaction between the user and the cell, including calls, but also all the automatic updates that are continuously requested by mobile applications (e.g. automatic mail check, location updates, chat messaging, etc). Since not all the users connected to a cell are calling all the time, our dataset is much more detailed than those based exclusively in call activity (e.g. CDR logs).

The dataset contains information for more than 75.000 cells spreading over a whole country in a period of three months between the spring and the summer of 2017. The construction of the dataset involves extracting, transforming, and synchronizing information coming from several systems which in turn collect data from probes distributed all over the network.

Considering a cell as a fixed spatial area watching users come and by, we conjecture that the regularity observed in individual mobility should translate in regularity at the cell. If users spend most of the time between the same *home* and *work* cells, we should see at each cell the aggregate activity of users being at home and users being at work nearby the cell. If the trajectories of individual users are regular and periodic, the composition of passing trajectories at each cell should also be regular and periodic. If the patterns of movement of individuals are potentially predictable, we should be able to predict the aggregate patterns utilization of cells.

In trying to verify those hypothesis, we first conducted and exploratory analysis using three cells that network engineers regarded as representative. All three showed strong periodicity as expected. Then, clustering algorithms were applied that consistently yielded a clear classification in three well defined groups. Every group presented strong regularity, providing the desired confirmation. In the last step, we tested several time series forecasting algorithms over all the cells of the network, in order to see if the periodicity patterns are predictable at the cell level.

The main findings of our study are:

- The strong regularity observed by previous research at the individual mobility level translates well at the cell level. We show that the composition of passing trajectories at each cell is also strongly regular and periodic, and devise a specific kind of diagram to view weekly evolution of activity in the cells.
- There are well-defined types of cells. We find that 40% of them are *home* cells, another 40% are *work* cells, and the remaining 20% form a class of *low activity* cells.
- Further analysis of this group of *low activity* cell shows that most of them are 2G network cells, that have implicit limitations to achieve higher levels of activity.
- The patterns of movement of individuals are potentially predictable, and we show that it is also feasible to predict the aggregate patterns at the cells, with good precision in the short term. The mean RMSE (Root Mean Squared Error) was about 17 over the whole cell population, and the ME (Mean Error) was about 2, in both cases with a relatively sharp distribution.
- The possibility of predicting activity at the cell level could help operators in addressing critical issues such as network planning and field operation prioritizing.

Understanding Post-Disaster Population Recovery Patterns

Takahiro Yabe[1], Kota Tsubouchi[2], Naoya Fujiwara[3], Yoshihide Sekimoto[4] and Satish V. Ukkusuri[1*]
[1]Lyles School of Civil Engineering, Purdue University, USA
[2]Yahoo Japan Corporation, Tokyo, Japan
[3]Graduate School of Information Sciences, Tohoku University, Japan
[4]Institute of Industrial Science, University of Tokyo, Japan

I. INTRODUCTION

Recent large scale disasters have shown the existence of significant variance in recovery trajectories across communities despite similar damage levels [1]. With the increase in the availability of large mobility datasets (e.g. mobile phone call detail records), longitudinal observations of mobility patterns have become possible [3], and many studies have analyzed the human dynamics during and after various anomalous events [4], [5], [2]. Despite such progress, such studies are fragmented and there is no general understanding on the population displacement and recovery patterns of communities across heterogeneous disasters and countries. In order to bridge this gap in the current literature, we analyzed large scale mobile phone GPS dataset from multiple disasters to 1) unravel the general patterns of recovery after different disasters across countries and 2) to identify key factors that explain such recovery patterns.

II. DATA

We collaborated with 3 different companies across US and Japan and collected GPS location data from mobile phones, and studied the post-disaster movements of more than 2.5 million affected users over a six-month period. We study five disasters that in total destroyed more than 1.5 million households, caused power outages in more than 8 million households, and caused more than $350 billion in economic loss. The disasters studied are Hurricane Maria (Puerto Rico, USA, 2017), Hurricane Irma (Florida, USA, 2017), Tohoku Tsunami (Tohoku area, Japan, 2011), Kumamoto Earthquake (Kyushu area, 2016), and Kinugawa Flood (Ibaraki area, Japan, 2015). The collection of disasters in this study are heterogeneous in various aspects, including the type of disaster, location of occurrence, and socio-economic characteristics of communities in the affected regions.

III. RESULTS

We observe that, despite the differences in the type of disaster and distinct socio-economic characteristics of the affected regions (i.e. Puerto Rico, Florida and Tohoku area), the population recovery patterns after the five disaster cases all follow a similar recovery pattern. Moreover, through a cross comparative analysis of population displacement and recovery patterns of over 250 communities in these five disasters, we find that the variance in initial and long-term displacement rates can be well explained by a set of key common factors, which generalize and further extend the insights obtained from small scale case studies. This work lays out a foundation for studies on understanding the population recovery dynamics of disaster affected cities, which contribute to urban planning efforts for enhancing the resilience of urban systems.

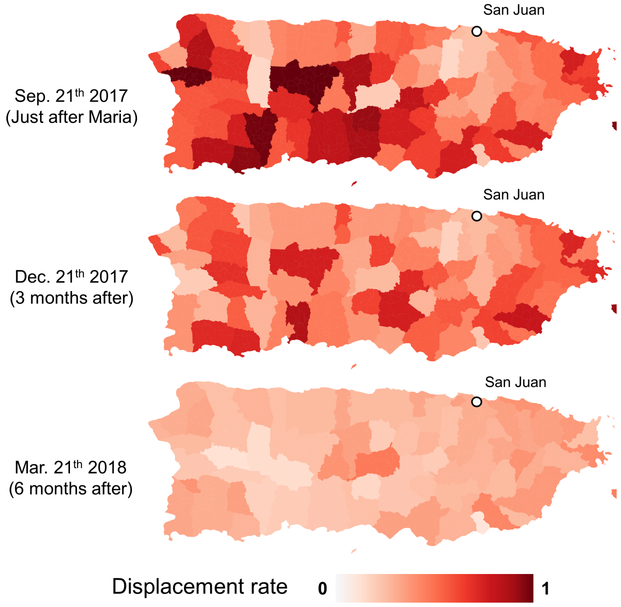


Fig. 1. **Displacement rate in Puerto Rico after Hurricane Maria.** Population recovery rates were estimated from mobility patterns of over 50,000 mobile phone users. We observe large spatial heterogeneity within the island, which can be explained by a set of key socio-economic factors.

References

[1] Daniel P Aldrich. *Building resilience: Social capital in post-disaster recovery*. University of Chicago Press, 2012.

[2] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[3] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

[4] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[5] Takahiro Yabe, Yoshihide Sekimoto, Kota Tsubouchi, and Satoshi Ikemoto. Cross-comparative analysis of evacuation behavior after earthquakes using mobile phone data. *PLoS one*, 14(2):e0211375, 2019.

[*] sukkusur@purdue.edu

Utilizing CDR Data in Migration Studies: Seasonal Labor Migration Among Syrian Refugees and Urban Deep Map for Integration in Turkey

Sedef Turper Alışık[*] sturper@ku.edu.tr

Damla Bayraktar Aksel[*] dabayraktar@ku.edu.tr

Asım Evren Yantaç[*] aeyantac@ku.edu.tr

İlker Kayı[*] ikayi@ku.edu.tr

Sibel Salman[*] ssalman@ku.edu.tr

Ahmet İçduygu[*] aicduygu@ku.edu.tr

Damla Çay[*] dcay13@@ku.edu.

Lemi Baruh[*] lbaruh@ku.edu.tr

Ivon Bensason[†] ibensason@yahoo.com

[*] Koç University, 34450 İstanbul, Turkey
[†] Attorney at Law, İstanbul, Turkey

The ongoing civil war in Syria in its eighth year has displaced more than 10 million people, including 5.6 million people seeking safety in third countries. Since the onset of the Syrian conflict, Turkey adopted a generous open-door policy towards those Syrians fleeing conflict and currently hosts more than 3.6 million registered Syrian refugees under the extended status of temporary protection. However, the official statistics on Syrian refugees in Turkey are compiled mainly through registration data, and therefore, these statistics can offer only a limited capacity to inform well-tuned integration policies that would meet the needs of a dynamically changing target population. Addressing the need to identify spatio-temporal service needs of the highly mobile refugee population living under temporary protection in Turkey, the current study utilizes the mobile call data records of Syrian refugees provided by Turk Telekom D4R (Data for Refugees) Big Data Challenge.

This study provides an overview of the data analysis and visualization steered under "An Urban Deep Map for Integration in Turkey" (UDMIT) project, which uses mobile call data records of Syrian refugees under temporary protection in Turkey. First, to examine Syrian refugees' temporal and spatial dimensions of mobility, the study concentrates on refugees' interprovincial migration patterns within Turkey. Based on an analysis of these patterns, the study offers insights on the potential motivations for regular and seasonal interprovincial mobility, especially regarding access to services and employment opportunities in the formal and informal labor market. The findings are also complemented by policy recommendations on how the D4R data can be of use to central and local authorities on providing occupational health and safety services and on improving refugees' access to information. Second, the current study delivers a web-based deep mapping platform that allows generating and reporting a visual representation of refugee population densities and mobility across Turkey on a real-time basis. The interface enables examining the spatio-temporal D4R data at three scales (country, province and district level) together with other layers of data, including (a) demographic information at the province and district levels, (b) service providers (non-governmental organizations, schools and healthcare services), (c) media analytics and (d) public discussion. Within the scope of this limited study, the deep mapping platform has been developed as an early-version prototype to demonstrate the potential of opening the data to the use of experts and public with a multilayered, visual and interactive tool.

Beyond Traffic Congestion Detection

Jungmin Yoo, Dayea Yim, Eunmi Lyu, and Eunsook Lee
SK Telecom, Repulic of Korea
{jungminyoo00,dyyim,eunmi.lyu,jen_lee}@sk.com

***Abstract*—As mobile traffic increases rapidly, it becomes important for network operators to detect and to optimize traffic congestion by monitoring network traffic in a real-time manner. Traffic congestion can be estimated with Charging Data Record (CDR) data from Offline Charging System (OFCS). However, since CDR data does not include the usage of resource allocation of the wireless channel, there is an inaccuracy in predicting traffic congestion. To improve the accuracy of congestion prediction, this paper suggests a new method that predict congestion status accurately with the data collected from base-station.**

***Index Terms*—Congestion Detection, Intelligent Traffic Management, Physical Resource Block, Quality of Experience**

I. Introduction

In a large-scale event such as a festival or a sport, where massive people want to download and upload an internet traffic, radio cell of the mobile network is overloaded usually. If there exists heavy users who monopolize radio resources, the Quality of Experience (QoE) of other users will be degraded. Therefore, it is important for network operators to monitor mobile traffic in real-time, to detect congestion in a timely manner, and to manage QoE so that all users experience good service quality.

Traffic congestion occurs when the radio cell is overloaded and the capacity is insufficient. Various studies have been conducted to predict the congestion and to perform load balancing. The conventional method of congestion detection infers the cell load with CDR data which reflects core-side network traffic.

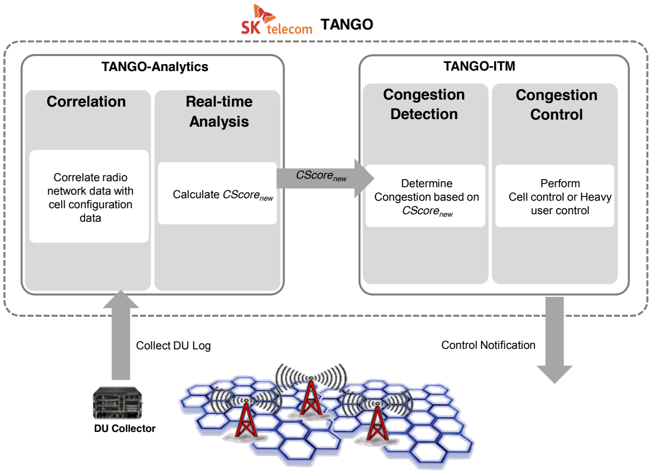


Fig. 1. Overall architecture

In this paper, we suggest a new method to detect actual traffic congestion of a radio cell by analyzing the Physical Resource Block (PRB) usage rate and the number of User Equipments (UEs). This method is applied into SK Telecom's Operation Support System (OSS) called *TANGO (T-Advanced Next Generation OSS)* [1] which collects big data from the whole network, analyzes and optimizes the network. Overall architecture is depicted in Fig. 1 with two parts of *Analytics* and *ITM* (Intelligent Traffic Management). We define *Congestion Score (CScore)* which is the indicator of cell congestion.

II. Congestion Detection: Past and Now

To manage the flow of internet data traffic efficiently, SK Telecom develops congestion detection algorithms with the data gathered from the whole network. We have defined congestion as a state where too many UEs are connected to base-station and the PRB is insufficient. It is not assumed as congestion when small number of UEs are occupying the whole PRB. Congestion occurs, in general, when PRB is not available due to a large number of UEs. The data used in congestion detection can be summarized below:

- *Past:* CDR data collected from OFCS (Offline Charging System) was used in the congestion detection algorithm. Key indicators of CDR are the number of subscribers connected to the base-station and data throughput per subscriber. CDR data could be used to estimate cell congestion though it does not reflect the cell congestion exactly.
- *Now:* Real-time call log generated from the base-station which includes PRB usage rate and the number of connected UEs is collected and analyzed for congestion detection. Congestion can be measured directly from the base-station and analyzed with cell configuration information.

III. Expectation and Result

We can improve customer experience by allowing users to share resources fairly based on the outcome of congestion detection. Load-balancing or changing network policy of heavy users can be triggered for the better experience. Also, the tracing of congestion area can be used for marketing data, public safety and energy saving. Detailed result will be included in the full paper.

References

[1] https://www.telecomasia.net/content/skt-expanding-use-tango-ai-platform, ”SKT expanding use of TANGO AI platform”

Dynamic approach of spatial segregation: a framework with mobile phone data

Lino Galiana, Insee, Département des Etudes Economiques

Benjamin Sakarovitch, Insee, SSP-lab

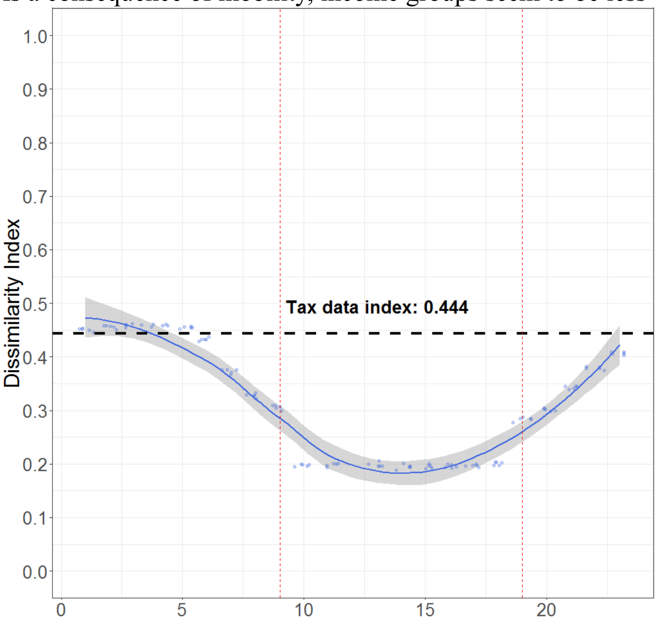
François Sémécurbe, Insee, SSP-lab

Zbigniew Smoreda, Orange Labs

We propose a framework to bring together mobile phone and geocoded tax data to shed a new light on segregation in urban areas. Urban segregation is generally measured through the glance of residential segregation. However, mobility brings together people from different areas and continuously shapes the spatial distribution of income. The effect of mobility on socioeconomic classes copresence in public spaces should be accounted to produce a complete vision of segregation. We propose an infra-day segregation approach that takes into account the effect of individual mobility on within-day segregation dynamics. We use Call Details Record (CDR) andfor their precise spatio-temporal dimensions that enable to estimate the composition of a city neighborhood at a given time.

That paper proposes an innovative methodology to study segregation dynamics at fine spatial and temporal granularities for both low- and high-income groups. We build infra-day segregation indexes using individual estimated position from pseudonymized mobile phone data. We adopt a Monte-Carlo procedure to estimate phone users' likelihood of belonging to low- or high-income groups based on their likely residence. We estimate infra-day segregation indexes by taking into account co-presence at 500x500 meters cells. We propose robustness checks and compare results with residential segregation indexes derived from tax data.

We use September 2007 CDR from Orange mobile network operator in France with information for more than 18,5 millions phone users. It is combined with income information at residence level for every French households. Our methodology is illustrated with Marseilles and Lyon, French second and third largest cities. Results suggest that residential segregation represents the acme of segregation. As a consequence of mobility, income groups seem to be less concentrated during daytime than during nighttime. After correcting for the discrepancy in estimated levels between phone and tax data, we find that segregation is 26% higher at its pinnacle than during the evening. The difference between nighttime and daytime segregation is even higher. We do not find a different pattern between low- and high-income infra-day segregation. An example of low-income infra-day segregation dynamics in Marseilles is presented. It shows significant within-day variation during the day. Nighttime segregation is at a level comparable with tax data measures.



Complete Trajectory Reconstruction from Sparse Mobile Phone Data

Guangshuo Chen (Inria), Aline Carneiro Viana (Inria), and Marco Fiore (CNR-IEIIT)
aline.viana@inria.fr

Mobile phone data are a popular source of information in many recent studies across multiple disciplines, and have largely contributed to improving our understanding of human mobility, interactions and habits [1]. Mobile phone data consist of time-stamped and geo-referenced communication events recorded by network operators, on a per-subscriber basis. They offer an unprecedented possibility of monitoring populations of millions of individuals over long periods that span months.

However, due to the uneven processes that govern mobile communications, the sampling of user locations in mobile phone data tends to be sparse and irregular in time, leading to substantial gaps in the individual trajectory information. As a result, mobile phone data offer a quite partial view of the overall mobility of each user, with substantial continuous periods where the positioning information is entirely absent [2]. In short, such data is usually temporarily or spatially sparse.

We address the problem of sparsity in mobile phone data, and investigate how to solve it through *trajectory reconstruction*, *i.e.*, the completion of individual movement information in the data by recovering the missing positions of each individual. To achieve our objective, we take the following steps.

First, we mine a large-scale mobile phone dataset collected in an operational nationwide network over three months, and provide ***a characterization of the severe sparsity that affects mobile phone data*** of 1.8 million users. We quantify the phenomenon by means of relevant metrics, including the sampling frequency and a dedicated *completeness* measure of individual trajectories in the data. Our results show that sparsity can be dramatic: for instance, only 5% of the actual positions that can be inferred, on average, for each user when working at a temporal resolution of 15 minutes. An important corollary of this result is that pre-processing techniques commonly adopted in the literature –which perform user filtering via (even mild) data sparsity thresholds– remove the vast majority of the available user population from subsequent analyses, hence introducing hard-to-estimate potential biases.

Second, we introduce an ***original approach for the reconstruction of trajectories from sparse mobile phone data***. Our solution leverages well-known features of human mobility (*e.g.*, the temporal prevalence of static phases, the regularity of movements, or the stability of overnight locations) to tailor and solve a *tensor factorization* problem that befits our goal. The proposed methodology allows transforming the highly incomplete positioning information inferred from mobile phone data into seamless individual trajectories that cover locations of all users throughout the full duration of the original dataset. We validate this strategy by using ground-truth mobile phone data collected at high frequency for a subset of 1,450 users: we accurately downsample the ground truth so as to mimic the irregular sampling of typical mobile phone data, and then reconstruct the user trajectories from the downsampled data via our dedicated tensor factorization. Our evaluation demonstrates that the approach achieves full reconstruction of individual trajectories with good accuracy, as exemplified in Figure 1.

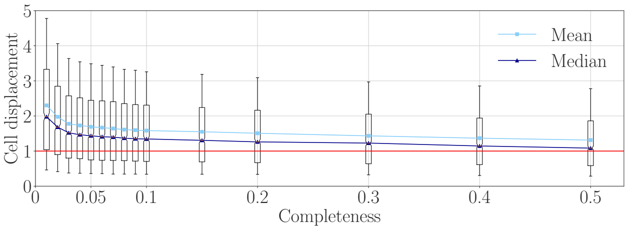


Fig. 1. Cell displacement of locations inferred via tailored tensor factorization with respect to ground truth, versus the level of completeness of the original trajectory (which are sparser at lower completeness). Candlesticks highlight the mean (light blue), median (dark blue), 25th and 75th percentiles (box), and 10th and 90th percentiles (errorbars). The horizontal line (red) highlights a one-cell displacement. This effectively means that, in the reconstructed data, a user is typically placed in the correct cell or in one that is very close to it. Such a level of accuracy is acceptable for metropolitan-scale analyses (where the urban surface is typically covered by hundreds of cells) and is excellent for national-scale studies (as inter-city mobility is perfectly captured).

Third, we give a ***demonstration of the importance of trajectory reconstruction for human mobility analyses***. We revisit three seminal studies by using the complete mobility of 1.7 million users, instead of the incomplete trajectories of a small fraction of especially active users as in the original works. Our investigation yields the following insights: (i) travel distances still follow a truncated power law [3] in the seamless trajectory data of a million-strong user population, yet the cut-off point is shifted away from the radius of gyration identified in the original work; (ii) the high uniqueness of individual trajectories [4] is substantially reduced in complete data, which lets us argue that the uniqueness identified in the original work was largely caused by the the diverse temporal patterns of the mobile communications of each user, rather than by a distinctive mobility; (iii) the theoretical predictability of human movements is even higher than thought [5], as considering the whole user population removes a filtering bias that favored highly mobile individuals with harder-to-anticipate locations.

References

[1] V.D. Blondel *et al.*, "A survey of results on mobile phone datasets analysis," EPJ Data Science 4(1), 2015.

[2] G. Chen *et al.*, "Enriching sparse mobility information in call detail records," Computer Communications 122, 2018.

[3] M.C. Gonzalez *et al.*, "Understanding individual human mobility patterns," Nature 453, 2008.

[4] Y.-A. de Montjoye *et al.*, "Unique in the crowd: The privacy bounds of human mobility," Scientific Reports 3, 2013.

[5] C. Song *et al.*, "Limits of predictability in human mobility," Science 327, 2010.

Mobile Money: Understanding and Predicting its Adoption and Use in a Developing Economy

Simone Centellegher[1,2], Giovanna Miritello[2], Daniel Villatoro[2],
Devyani Parameshwar[2], Bruno Lepri[1] and Nuria Oliver[2]

[1]Fondazione Bruno Kessler
[2]Vodafone Research
{*giovanna.miritello,daniel.villatoro,devyani.parameshwar,nuria.oliver*}*@vodafone.com*
{*centellegher,lepri*}*@fbk.eu*

As of today, there are approximately 2 billion unbanked individuals world wide, adults who are not bank account holders or do not have access to a financial institution. Access to financial institutions is difficult in developing economies and especially for the poor, due to the low penetration of financial services in such countries. The widespread adoption of mobile phones, has enabled the rise of mobile money services. Mobile money bridges the gap between the cash and digital economies, enabling those without access to banks to load cash in a mobile wallet and transact digitally using money transfers, deposits and withdrawals of money, bill payments, etc. through the mobile phone network. Despite the success of mobile money, there is a lack of quantitative studies that unveil which factors contribute to the adoption and sustained usage of such services [2]. In this work, we describe the results of a quantitative study that analyzes data from the world's leading mobile money service, M-Pesa [1], where we analyzed millions of anonymized mobile phone communications and M-Pesa transactions in an African country.

After analyzing the customers' usage of M-Pesa and their patterns of behavior, we tried to understand whether and to which degree past mobile phone usage captures elements of human behavior that are predictive of future mobile money usage.

We build two machine learning-based predictive models (1) "Predicting M-Pesa usage" and "Predicting M-Pesa expenditure", that predict future M-Pesa adoption and intensity of usage, using multiple sources of data, including mobile phone data, M-Pesa agent information, the number of M-Pesa friends in the user's social network, and the type of geographic location where the mobile activity took place. Our models show a competitive performance of AUC=0.691 and AUC=0.619 for each of the previous tasks. We find that the most predictive features (Figure 1) are related to mobile phone activity, to the presence of M-Pesa users in a customer's ego-network and to mobility. Finally we discuss and draw key implications for the design of mobile money services in a developing country.

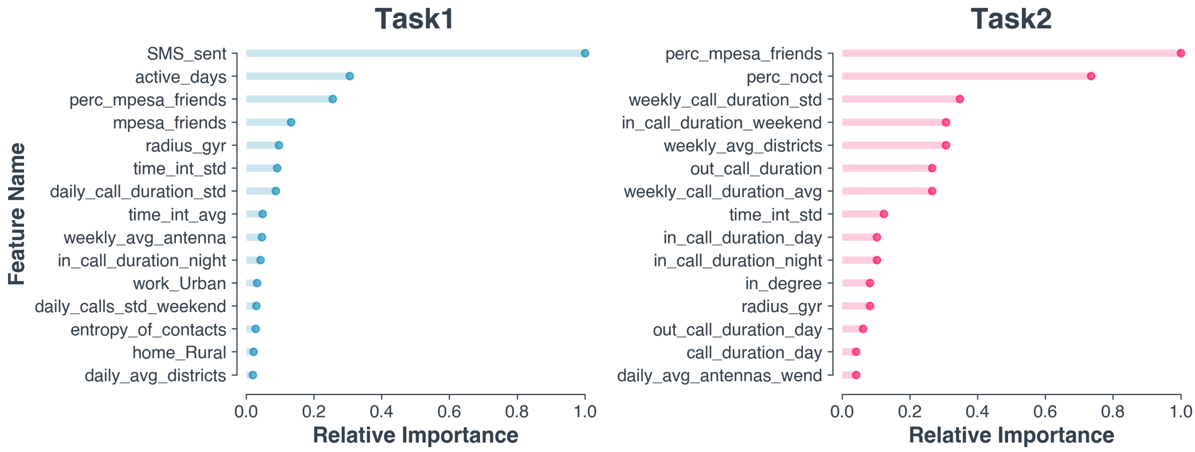


Figure 1: Feature importance in Task 1: "Predicting M-Pesa usage" (left), and feature importance in Task 2: "Predicting M-Pesa expenditure" (right).

References

[1] William Jack and Tavneet Suri. Mobile money: The economics of m-pesa. Technical report, National Bureau of Economic Research, 2011.

[2] Muhammad R Khan and Joshua E Blumenstock. Predictors without borders: behavioral modeling of product adoption in three developing countries. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154. ACM, 2016.

Cities and the Structure of Social Interactions: Evidence from Mobile Phone Data*

Konstantin Büchel and Maximilian v. Ehrlich◇

Frist Draft: December 2016
This Version: January 14, 2019

Abstract

Social interactions are considered pivotal to agglomeration economies. We explore a unique dataset on mobile phone calls to examine how distance and population density shape the structure of social interactions. Exploiting an exogenous change in travel times, we find that distance is highly detrimental to interpersonal exchange. We show that, despite distance-related costs, urban residents do not benefit from larger networks when spatial sorting is accounted for. Higher density rather generates a more efficient network in terms of matching quality, measured via link stability, and information diffusion capacity, measured via clustering.

Keywords: Agglomeration Economies; Social Interactions; Network Analysis; Spatial Sorting.

JEL classification: R10; R23; D83; D85; Z13.

*We benefited from numerous comments by Juan Becutti, Sascha Becker, Aymo Brunetti, Shane Greenstein, Fabian Gunzinger, Christian Hilber, Blaise Melly, Luigi Pascali, Diego Puga, Esteban Rossi-Hansberg, Elisabet Viladecans-Marsal, Alex Whalley, and Yves Zenou. We further received invaluable feedback from participants at the Verein für Socialpolitik Meeting in Tellow, the Meeting of the Urban Economics Association in Minneapolis, the CRED-Workshop in Bern, the SERC Conference in London, the Applecross Conference, the SSES Annual Congress in Lausanne, the NBER Summer Institute 2017, the 6th Lindau Meeting on Economic Sciences, the 2017 SMU Conference on Urban and Regional Economics, the Meeting of the Urban Economics Association in New York and research seminars at the Paris School of Economics, the University of Milan, the IEB Barcelona, and the University of St. Gallen. A very special thanks is due to *Swisscom AG* for providing the facilities and data to conduct this research project; we are particularly indebted to Imad Aad and Holger Müller, who have accompanied the project since its start. We also want to thank *search.ch* and *comparis.ch* for providing data on travel times and usage statistics of messenger apps.

◇University of Bern, Department of Economics and Center for Regional Economic Development, Schanzeneckstrasse 1, 3001 Bern, Switzerland. E-mail: konstantin.buechel@vwi.unibe.ch, maximilian.vonehrlich@vwi.unibe.ch.

Modeling Dynamic Influence of Customers to Predict Churn: An Application with Similarity Forests for Time Series Classification

Laura Calzada-Infante[a*], María Óskarsdóttir[b], Bart Baesens[b,c]
[a] Engineering School, Campus de Gijón, University of Oviedo, 33204 Gijón, Spain
Email: calzadalaura@uniovi.es
[b] Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
[c] Dept. of Decision Analytics and Risk, University of Southampton, United Kingdom

Telecommunication companies are interested in implementing effective retention campaigns for their potential churners, due to the fact that retaining a customer is cheaper than attracting a new one, and that long-term customers are more profitable (1). Multiple research has used binary classification methods to predict potential churners, and there are new methods to consider the influence of the customers' relationships (1,2).

Customers' social relationships influence the decision of changing the operator. Network analysis techniques allow the representation of these relationships and provide a new tool to detect customer behavioral patterns and define accurate models to predict potential churners (3). Call Detail Records (CDR) of telco's customers are the most accessible information that telco companies have about the customer's behavior and influence.

We propose a novel method to extract the dynamic influence of each customer using Social Network Analysis techniques, together with binary classification methods. A temporal network is built based on the CDR to extract behavioral patterns with respect to position of churners in the network. The dynamic influence of each customer is determined by applying different centrality metrics and diffusion propagation methods over a sliding window using aggregated and time-order networks.

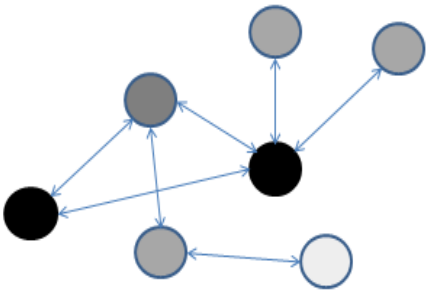


Figure 1. Representation of churners' influence in a customer's social relationships network. Black nodes are churners.

The resulting time series are classified by a recently proposed binary classification method called similarity forests(4). In addition, a comparison to other methods evaluates the accuracy of predicting further in time and the possibility of designing a method that is capable of detecting potential churners both in short and long term.

References

1. Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., Vanthienen, J. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. Expert Syst Appl. 2017;85:204–20.

2. Kim, K., Jun, C., Lee, J. Improved churn prediction in telecommunication industry by analyzing a large network. Expert Syst Appl. 2014;41:6575–84.

3. Óskarsdóttir, M., Calster, T. V., Baesens. B., Lemahieu W, Vanthienen J. Time series for early churn detection: Using similarity based classification for dynamic networks. Expert Syst Appl. 2018;106:55–65.

4. Sathe, S., Aggarwal, C.C. Similarity Forests. Proc 23rd ACM SIGKDD Int Conf Knowl Discov Data Min. 2017;395–403.

Comparison and Analysis of Migrant Identification Methods

V. Frias-Martinez[1], L. Hong[1], J. Wu[1], A. Villareal[3], E. Frias-Martinez[2]
(1) iSchool, UMIACS, University of Maryland, College Park, MD, USA, lzhong@umd.edu ,(2) Telefonica Research, Madrid, Spain. (3) Department of Sociology, University of Maryland, College Park, MD, USA.

Internal migration refers to the migration of individuals from one region to another within the same geopolitical entity, typically within the same country. Considerable attention has been given to the study of migration using spatio-temporal data passively generated, for example cell phone records or social media data. Such type of data enables to carry out large-scale analyses of migration flows as well as the micro-level view necessary to analyze individual behaviors [1]. At its core, research on migration behaviors first requires to identify the internal migrants in the dataset. Methods to identify migrants are based on determining home location changes i.e., a person that was living in a location, changes her home permanently or temporarily within the same country. Several methods have been developed to identify home location using spatio-temporal data [2] and some of these methods have been applied to identify volumes of internal migrants [3]. However, no work has looked into analyzing the impact that the choice of a home location algorithm might has in the identification of migrants. In this paper, we use an 8-month window cell phone dataset from Mexico to identify and compare internal migrations.

We define as internal migrants the individuals who have a consistent home location for at least three months and then move to another place, where they also stay for at least three months. With this definition, the internal migrants we identify can be either long-term or short-term (circular) migrants, depending on whether they go back or not to their original location after our data collection period finishes [4]. The census data we use for validation measures the internal migration flow at the municipality level. As a result, we use four different state-of-the art methods to detect home location at a municipality level: (1) most visited municipality at night, between 6pm to 6am; (2) a shorter period of night time, from 10pm to 6am- the main motivation to do this is that a tighter temporal range might help to reduce the noise in the set of communication towers considered as potential home location; (3) home location as the municipality where the center of gravity across all visited cellular towers is located, weighted by the cell phone activity in each cell and (4) a combination of the temporal window approach (10pm-6am) with the center of gravity approach- the assumption for this approach is that the activities during night hours tend to be closer to one's home location.

We used these methods to identify potential internal migrants and evaluated the accuracy of each method by comparing the computed migration matrix with official census data. Additionally, we looked into the biases introduced by each method, measured as the differences in accuracy when different types of urban and rural flows are approximated via cell phone data. Overall, results show that: (i) methods that use temporal ranges to identify internal migrants perform better than center of gravity-based methods; (ii) longer temporal ranges show better accuracy than shorter ranges; (iv) current methods show biases against rural population; and that (v) those biases decrease when total outbound or inbound flows are considered. Although we focused our study in Mexico, these results could potentially be valid to the great majority of developing economies, as they tend to show similar characteristics, namely: (1) high penetration of cell phones; (2) uneven distribution of the population between urban and rural areas; and (3) reduced cell phone activity when compared to developed economies.

References

[1] J. E. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: evidence from rwanda," Information Technology for Development, vol. 18, no. 2, pp. 107–125, 2012.

[2] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, J. Martonosi, M.and Rowland, and A. Varshavsky, "Identifying important places in peoples lives from cellular network data," in Int. Conf. Pervasive Computing, 2011, pp. 133–151

[3] S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez, "Modeling human migration patterns during drought conditions in la guajira, colombia," in ACM Computing and Sustainable Societies, 2018.

[4] R. D. Bedford et al., New Hebridean Mobility: a study of circular migration. Canberra, ACT: Dept. of Human Geography, Research School of Pacific Studies, The Australian National University., 2017

Analysis of Post-Migration Mobility and Social Ties of Internal Migrants

V. Frias-Martinez[1], L. Hong[1], J. Wu[1], A. Villareal[3], E. Frias-Martinez[2]
(1) iSchool, UMIACS, University of Maryland, College Park, MD, USA ,(2) Telefonica Research, Madrid, Spain. (3) Department of Sociology, University of Maryland, College Park, MD, USA.

Migrations have been studied using macro-level and micro-level analysis. Macro-level studies are carried out using a combination of various survey and census datasets to model large-scale behaviors, however these models fail to provide more nuanced information about the physical or social status of the migrants. Micro approaches, which use interviews and diaries to provide a window into more individual behaviors, could benefit from methods to identify novel or under-studied behaviors that should be addressed in the migration research agenda. In this paper, we present a framework that uses information extracted from cell phone metadata to reveal internal migration behaviors that complement micro-level migration approaches working to understand the physical, social and psychological decision processes behind migration. The proposed framework reveals internal migration behaviors with a focus on immediate post-migration behaviors and the role of pre-migration activities from two perspectives: spatial dynamics and social ties. The main objective is to carry out large-scale analyses of internal migration trends to reveal migrant behaviors that would benefit from further qualitative studies through personal interviews or individual surveys. Ultimately, we expect our analyses to inform migration researchers of pre- and post-migration behaviors that would benefit from further qualitative analysis.

Given cell phone metadata from millions of individuals for a given country, the proposed framework consists of three parts. First, the framework uses features extracted from the cell phone metadata to identify potential migrants in the dataset. We present a method to identify internal migrants and we evaluate its accuracy using real census migration data [1]. Second, the framework uses the aggregated migrants to characterize immediate post-migration behaviors i.e., we analyze the post-migration spatial dynamics and social networks and compare these against behaviors from locals that have not undergone any migration process. Third, we analyze the role that pre-migration spatial dynamics and social networks might play in the same post-migration behaviors shown by internal migrants.

We evaluate the proposed framework to study internal migration behaviors in Mexico, using a dataset with eight months of anonymized cell phone metadata from over 48 million subscribers in combination with open data from the Mexican Statistical Institute (INEGI) [3]. In this context, regarding mobility, we observe among other findings, that in the immediate post-migration period, internal migrants, visit more municipalities and have more irregular behaviors than the local community. In fact, immediately after migrating, migrants have, on average, higher entropy than locals, showing more irregular mobility patterns and having significantly longer trips. These findings could reveal that individuals make an effort to maintain their local connections in their pre-migration municipalities either because of work or personal reasons. Similar findings have been qualitatively reported in other countries [4]. Regarding social ties, migrants communicate with a similar volume of calls than locals, but with a smaller number of contacts, and show lower entropy in their cell phone-based social networks than locals. Also in this case, similar results have been shown in the context of international migrations [5].

References

[1] S. Isaacman, V. Frias-Martinez, E. Frias-Martinez. 2018. Modeling Human Migration patterns during Drought Conditions in La Guajira, Colombia. ACM SIGCAS

[2] L. Hong, J. Wu, E. Frias-Martinez, A. Villareal, V. Frias-Martinez, 2019. Characterization of Internal Migrant Behavior in the Immediate Post-Migration Period using Cell Phone Traces. 10th ACM ICTD

[3] Mexican Statistical Institute INEGI. 2015. National Survey of Demographic Dynamics 2014. http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/ enadid/2014/default.html.

[4] Cati Coe. 2011. What is the impact of transnational migration on family life? Women's comparisons of internal and international migration in Ghana. American Ethnologist 38, 1 (2011), 148–163.

[5] Steven Vertovec. 2004. Cheap calls: the social glue of migrant transnationalism. Global networks 4, 2 (2004).

Understanding Tourist Travel Patterns using Mobile Signaling Data [*]

Qiwei Han
Nova School of Business and Economics
Carcavelos, Portugal
qiwei.han@novasbe.pt

Margarida Abreu Novais
Griffith Business School
Gold Coast, Australia
m.abreunovais@griffith.edu.au

Leid Zejnilovic
Nova School of Business and Economics
Carcavelos, Portugal
leid.zejnilovic@novasbe.pt

1 INTRODUCTION

Understanding how tourists move to reach and travel around within a destination is of paramount importance for tourism decision-makers. Given its significance, the investigation of tourist spatio-temporal patterns represents a critical stream of tourism behavior research. Notwithstanding existing contributions, there are still challenges in fully capturing tourist behavior due to the inefficient data collection approaches currently in use. This paper puts forward high-time-density data from mobile network operators as a powerful data source to investigate tourism spatial behavior in Tuscany, Italy, a region known for its extensive tourism resources, especially when combined with other data sources.

We investigate tourist travel patterns using unsupervised learning at individual level and municipality level, respectively. First, we create tourist *personas* that represent tourist's archetypes, by clustering individual tourists' spatio-temporal behavior. Second, we propose a latent embedding model *Geo2Vec*, an approach originally developed for natural language processing tasks to learn the sequences of tourists' locations and produce a vector space of municipalities. Geo2vec model is able to reveal similarity between municipalities in terms of tourist's travel preferences beyond geographic proximity. Our results demonstrate the advantages of using mobile signaling data to understand tourist mobility in tourism management research, compared to traditional mobile phone data analysis that relies on call detail records.

2 DATA

A large-scale mobile signaling dataset for this study has been provided by a European mobile network operator (referred to hereafter as EURMO). It includes pre-processed logs of anonymized signaling traces of mobile devices with foreign SIM cards connected to the EURMO in Tuscany between May 2017 and February 2018 for over 9.6 million international tourists, including visitor ID, nationality, timestamp and geo-coordinates of the connected cell towers. The data is available at the resolution of minute level when tourist's location is changing (*i.e.* switching to different towers) and at hourly level if the visitor remains at the same location (*i.e.* connecting to the same cell tower). Meanwhile, we retrieve geo-spatial information about terrain categories (*e.g.* forest, water, coast, park and city) and tourism attractions from shapefiles and Tuscany official tourism website, respectively. As such, we can infer tourist's travel preferences towards diverse touristic activities in different seasons. We further clean the dataset and select a subset of 2.95 million tourists from top 6 nationalities (Germany, United States, France, United Kingdom, Netherlands and China).

3 PERSONA CLUSTERING

By aggregating the signaling data at the individual level, a list of 42 features has been generated, to characterize each tourist. These features describe different aspects of tourists' behavior, such as duration and number of locations spent in Tuscany, duration spent in different types of landscape and attractions and summarized spatial coordinates. Then, K-means clustering has been performed across seasons (pre-summer, summer, fall and winter) and nationalities.

The tourists are segmented into 4 groups, where each group exhibit distinct travel patterns: 1) *Florence Visitors* consisting nearly 40% of the tourists that tend to focus their visit on city of Florence within a short time during non-summer season; 2) *Coast Lovers*, representing 28% of the tourists mainly from Germany and France that spend most of their activities along the coast during the summer; 3) *Fast Trippers* consisting 21% of all the tourists from non-European countries such as China and United States that travel beyond Tuscany to visit more places across Italy in a short period; 4) *Explorers*, representing 11% of all tourists that tend to enjoy a long holiday to explore Tuscany in a wide range of activities. Therefore, our findings provide insights to designing more personalized travel marketing campaigns catering to different tourist segments.

4 GEO2VEC

Similar as Word2Vec model that aims to learn the latent embedding of each word from the vocabulary, Geo2Vec performs feature learning through a shallow, two-layer neural network that is trained to reconstruct contexts of locations through tourist's travel sequences. More specifically, we first create an ordered list of visited municipalities for each tourist and train a neural network to learn distributed representations for each municipality. Each municipality is represented as an multidimensional vector in a latent space, known as embeddings, such that municipalities that share common trips are located in close proximity to one another in that space.

Geo2Vec model is able to detect both short and long-range similarity between municipalities directly from the dataset, instead of manually generating features subject to domain context. More interestingly, Geo2Vec can identify adjacent municipalities that are geographically disparate in terms of distances in the latent space, due to their common tourists' travel preferences. Therefore, we can leverage the latent embeddings of municipalities to segment Tuscany into groups of sub-regions beyond geographic proximity. This would allow us to design a data-driven tourism resource distribution strategy according to tourist travel patterns.

[*] We thank Toscana Promozione Turistica and EURMO for their support and domain expertise. We also thank colleagues from Data Science for Social Good Europe who have contributed to this work.

A novel measure of clustering, with application to a mobile money network

Geoffrey Canright*, Rich Ling[§]
*Telenor Research, Snarøyveien 30, 1331 Fornebu, Norway
Email: geoffrey.canright@telenor.com
[§]Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
Email: riseling@gmail.com

Abstract—Inspired by studies of a mobile money transaction network, we have defined a measure Γ of *clustering of closure links* in an egonet. Γ is 0 when the closure links have no nodes in common, is 1 when all closure links form a clique (max clustering), and always obeys $0\leq \Gamma \leq 1$. Our data set is 6 months of p2p transaction data (only), furnished by a mobile money provider. We study recruitment of new users to this network, and—using a fairly strict definition of “recruitment” and “new users”—find that ***(i) good recruiters have (on average) a much higher egonet Γ value than non-recruiters***. This result holds even if one corrects for other measures of activity. In contrast, the standard egonet clustering coefficient *cc* is only weakly, and negatively, correlated with recruitment (Figure 1).

We have also defined a whole-graph (WG) measure of Γ, analogous to the whole-graph *cc*. Γ_{WG} is zero when all triads are isolated, and 1 when all triads form a single clique. We show the evolution of Γ_{WG} and cc_{WG} in Figure 2. We see that cc_{WG} is flat over the 6 months, and small (around 1%), while Γ_{WG} grows monotonically to a rather high value (0.77—cf the average egonet Γ, which is 0.16 in month 6). Furthermore, we can define clusters of triads as sets of triads connected by 0 or one hop. With this definition, we find a largest triad cluster that includes about 200,000 nodes—about 20% of the entire network—in month 3. In short, we find ***(ii) strong growth of Γ-clustering over time***. This cannot be explained in terms of “crowding” of the closure links, since this growth is observed at constant (low) density cc_{WG}.

We offer plausible explanations for these results. ***(i)*** We postulate that new users tend to be recruited into one of their geographically and socially local circles, which has previous adopters, and dense closure links. This is supported by the fact that Γ for the recruiters continues to grow (faster than for non-recruiters) after the recruiting event(s). ***(ii)*** We believe that long-range links (socially and/or geographically) are needed to give the observed large-scale structure. We know however that such links do exist in the network, in the form of transactions between agents—who work for the provider, and participate in a hierarchical network of transactions.

In short—we find that measurement of Γ-clustering gives unique insights into both the recruitment process, and the long-range structure, for a mobile money network. We believe also that the Γ-clustering measure can be useful in a broader range of network studies.

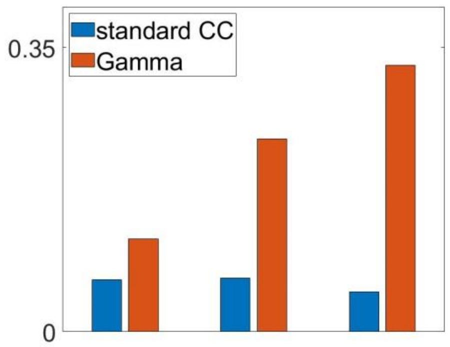


Figure 1. Average *cc* (blue) and Γ (red) for non-recruiters (left), R= 1 recruiters, and R > 1 recruiters (right)

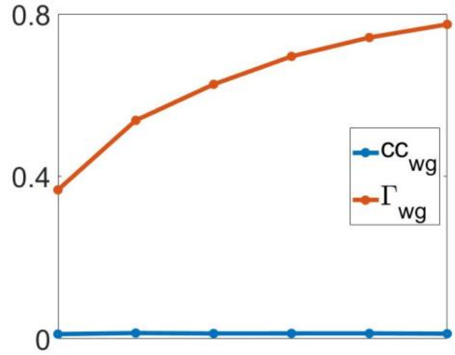


Figure 2. cc_{WG} (lower curve) and Γ_{WG} (upper curve) over 6 months

Refugee Mobility: Evidence from Phone Data in Turkey

Michel Beine[1], Luisito Bertinelli[1], Rana Cömertpay[1], Anastasia Litina[2],
Jean-François Maystadt[3], Benteng Zou[1]

[1]) University of Luxembourg, Avenue de la Faïencerie 162A, L-1511, Luxembourg
michel.beine@uni.lu
luisito.bertinelli@uni.lu
rana.comertpay@uni.lu
benteng.zou@uni.lu
[2]) University of Ioannina, University Campus, 455 00 Ioannina, Greece
alitina@cc.uoi.gr
[3]) University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium
Lancaster University, Baggily, Lancaster LA1 4YW, UK
j.maystadt@lancaster.ac.uk

Our paper employs spatially explicit call detail records provided within the D4R challenge and combines it to several other sources to study one of the multiple aspects of integration of refugees, namely the mobility of refugees across provinces in Turkey.

More specifically, we look at the location of one hundred thousand randomly selected mobile transactions (fifty thousand refugees and fifty thousand non-refugees) recorded by cell towers to define likely decisions of individuals to move between provinces in Turkey. Reconstructing bilateral migration flows at the monthly and at the province level, we employ a gravity model to empirically estimate a series of determinants of refugee movements.

We employ two main sets of determinants of mobility. First, we use the standard gravity model controls, i. e., variables that relate to the attractiveness (resp. repulsiveness) of province d (resp. o) for prospective refugees, the so-called pull (resp. push) factors. These are levels of income at origin and destination provinces, distances across provinces, network effects at destination provinces as well as some refugee-specific determinants such as the presence of refugee camps at origin and destination provinces. The second set of variables that we construct is aimed to capture policy related issues. Our source dataset is the Global Database of Events, Language and Tone, which captures world-wide news media over thirty years, in over one hundred languages and consists of over a quarter billion georeferenced event records in over three hundred categories. The variables we rely upon are as follows: rally for leadership change; boycotts; violent protest; economic aid; humanitarian aid and asylum grants. Considering news as an indicator of policy implemented at the provincial level we gain a better understanding as to how policy can facilitate refugee mobility and thus enhance integration.

We find that standard gravity determinants, such as distance, source income as well as networks apply. Furthermore, policy interventions that are facilitated with political stability, asylum granting and economic aid also matter thus suggesting that there is ample room for policy making.

To benchmark our findings, we estimate the same model for the mobility of individuals with a non-refugee status. The same determinants apply, however the impact of each of these determinants is stronger for non-refugees.

Towards a Reference Methodological Framework for processing Mobile Network Operator data for Official Statistics

Fabio Ricciato*, Albrecht Wirthmann*,
Martijn Tennekes†, Benjamin Sakarovitch‡, Roberta Radini§ and David Salgado¶

**EUROSTAT European Commission.*
†CBS, Netherlands. ‡INSEE, France. §ISTAT, Italy. ¶INE, Spain.

An increasing number of Mobile Network Operators (MNO) are now able to extract signalling data originating by the continuous interaction of mobile stations with the cellular network. Compared to traditional Call Detail Records (CDR), signalling data have better spatial/temporal resolution and represent a more informative data source to analyse human presence and mobility patterns. On the other hand, such data are considerably more costly to extract and more complex to interpret than CDR. Furthermore, they yield a higher degree of heterogeneity across different MNOs, due to proprietary monitoring technologies and tighter dependencies with the underlying network configuration. Their format and semantics change, following the evolution of MNO infrastructure. Consequently, it is difficult for statisticians and researchers without a solid telco engineering background to interpret such data.

In order to overcome these challenges, and thus promote the adoption of MNO signalling data for the production of official statistics, we propose a general Reference Methodological Framework (RMF) intended to facilitate the use of signalling data by statisticians and, in general, by non-telco experts. The RMF is inspired by the principles of modularity, functional layering and the “hourglass model” that lie at the foundation of modern computer network architectures.

Broadly speaking, the RMF encompasses three functional layers: an intermediate convergence layer (C-layer) decouples the complexity of signalling data at the bottom (D-layer) from the statistical definitions on the top (S-layer). Decoupling the D-layer and S-layers allows experts from the two domains, namely MNO engineers and statisticians, to work independently, easing the evolution of both layers. In the proposed vision, MNO engineers implement low-level data processing functions, collectively called “D-to-C” mapping, that transform raw network data into a sequence of geo-located events in a common format, along with additional auxiliary information, that represent C-layer data. The proposed model can accommodate different geo-location methods for individual events, including simplistic Voronoi tessellation (a popular but sub-optimal choice) and more advanced approaches based on cell radio coverage predictions.

From the C-layer data, the statisticians develop estimation methods that are logically placed in the upper S-layer. In so doing, they might fuse geo-located C-layer event data with external geographical data, e.g. transportation maps and/or land use maps.

Adoption of a common RMF enables the reuse of algorithms and processing modules developed by different statistical offices and/or research groups across data from different MNO. In this sense, it will facilitate benchmarking, independent validation and collaborative development of processing algorithms developed by academic and industrial research groups. When coupled with open-source algorithms, it helps to achieve transparency, auditability and public scrutiny onto processing algorithms.

The RMF is being developed within the European Statistical System by a joint effort between EUROSTAT and the National Statistical Institutes of some European countries. In this presentation we will present the general principles underlying the RMF, report on the status of ongoing development and outline future work.

Branching Out the Babytree: The Effects of Dual Peer Group Membership on Social Support During Pregnancy in Online Communities

Lingqing Jiang[1] and Zhen Zhu*[2]

[1]University of Essex, UK
[2]University of Greenwich, UK

Abstract

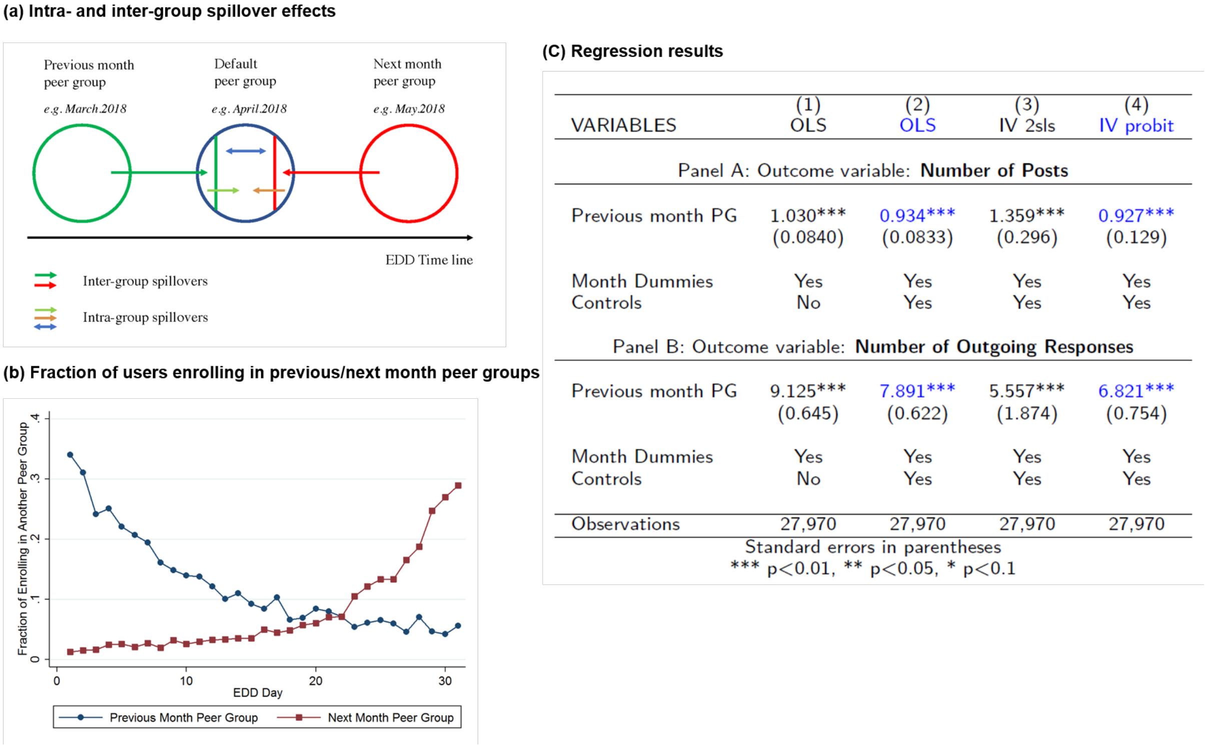
Social support from peers plays a positive and important role in many contexts. Today, the emergence of online social communities provides possibility to seek social support to maternity from other individuals that is complementary to those from the family and government.

Our paper investigates the potential spillovers of social support between peer groups formed by pregnant women (see Figure 1a for the distinction between intra- and inter-group spillovers). We collect a sample dataset from the largest maternity and parenting APP in China, Babytree, which contains over 30,000 pregant women in our sample.

A special feature of Babytree is that users can enroll in peer groups based on their estimated due date (EDD). In principal, users can enroll in any peer groups that are open for enrollment at any given time. A naive comparison between uses enrolling in two peer groups and users enrolling in a single peer group has self-selection problem, e.g. users who are more active in posting and responding might be more likely to enroll in two or more peer groups. If this is the case, the native comparison will simply reflect a positive correlation between enrolling in an additional peer group and users' overall activeness in their default peer group.

We take the advantage of the fact that the enrollment of peer groups is largely determined by the estimated due date (EDD) and use it as an instrument to address the above self-selection issue (see Figure 1b). Figure 1b shows the fraction of users enrolling in an additional peer group of previous or next month for each day of their EDD. Indeed, the fraction of users enrolling in the previous (next) month peer group gradually decreases (increases) over time.

We find that additionally enrolling in the previous month peer group increases users' contribution of social support by 1 post and 6 outgoing responses in their default peer group, respectively (see Figure 1c). No such spillover effects are found when users additionally enroll in the next month peer group. The policy implications of inter-group spillovers are not confined to the Babytree platform but also apply to other fields in which inter-group interaction plays an important role with similar channels, such as the interaction between the freshmen and the sophomores, the juniors and the seniors, as well as the interdisciplinary research.



(a) Intra- and inter-group spillover effects

(b) Fraction of users enrolling in previous/next month peer groups

(C) Regression results

VARIABLES	(1) OLS	(2) OLS	(3) IV 2sls	(4) IV probit
Panel A: Outcome variable: **Number of Posts**				
Previous month PG	1.030*** (0.0840)	0.934*** (0.0833)	1.359*** (0.296)	0.927*** (0.129)
Month Dummies	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes
Panel B: Outcome variable: **Number of Outgoing Responses**				
Previous month PG	9.125*** (0.645)	7.891*** (0.622)	5.557*** (1.874)	6.821*** (0.754)
Month Dummies	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes
Observations	27,970	27,970	27,970	27,970

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 1: (a) Intra- and inter-group spillover effects; (b) Fraction of users additionally enrolling in the peer group of previous/next month; (c) Regression results using OLS and IV.

*Corresponding author. Email: z.zhu@gre.ac.uk.

On the Optimal Marketing Aggressiveness Level of C2C Sellers in Social Media: Evidence from China

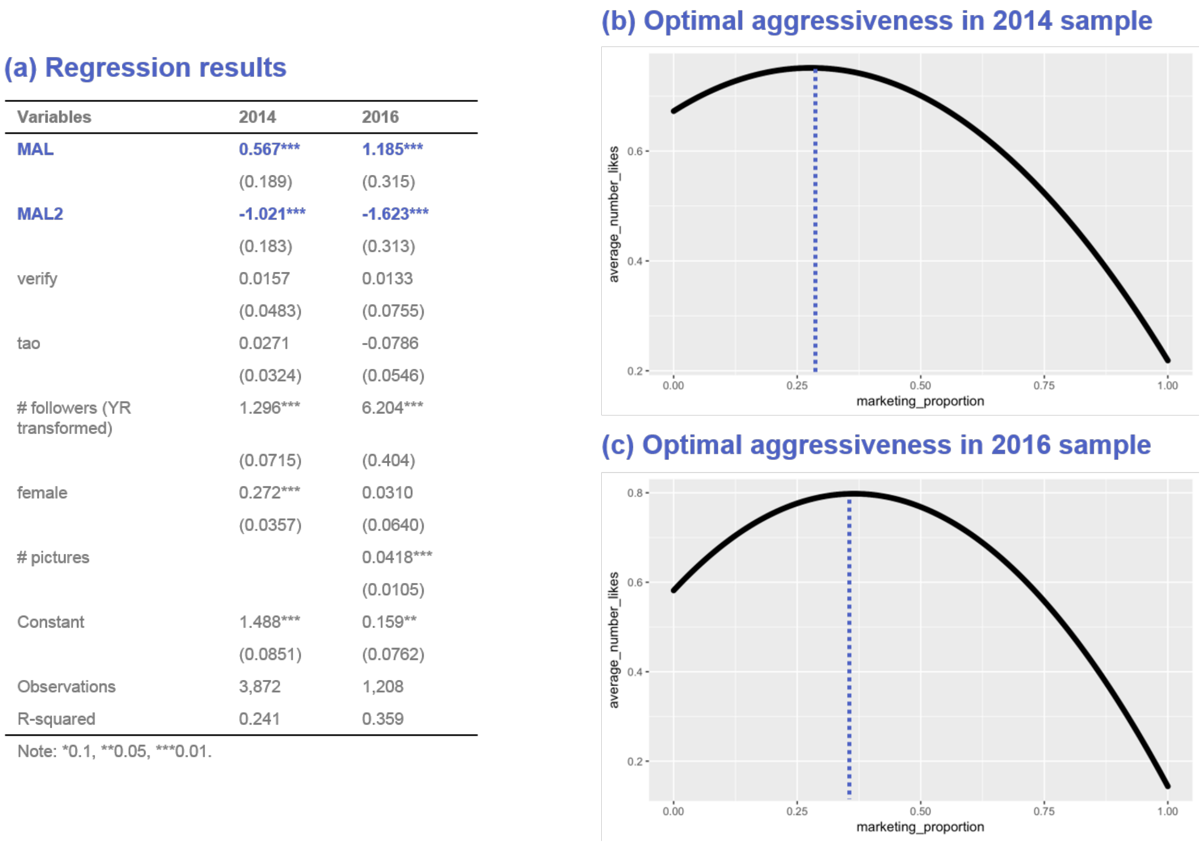
Xu Wang[1], Bart Baesens[2], and Zhen Zhu*[3]

[1]IMT School for Advanced Studies Lucca, Italy
[2]KU Leuven, Belgium
[3]University of Greenwich, UK

Abstract

Social media has become a widely used marketing tool for reaching potential customers. Because of its low cost, social media marketing is especially appealing to customer-to-customer (C2C) sellers. Customers can also benefit from social media marketing by learning about products and by interacting with sellers in real time. However, a seller's marketing microblogs may backfire on her for dominating the social space.

Defining the marketing popularity as the average number of likes each seller receives per marketing-related microblog and defining the marketing aggressiveness level as the proportion of her marketing-related microblogs, this paper empirically explores the relationship between marketing aggressiveness level in social media and the marketing popularity. We gather the data from China's largest microblogging platform, Sina Weibo, and the sellers in our sample are from China's largest C2C online shopping platform, Taobao. We use machine learning classification methods to quantify the marketing aggressiveness level as well as estimate the optimal aggressiveness level to achieve the maximum popularity. We find that the empirical relationship between the marketing aggressiveness level and the marketing popularity follows an inverted U-shape curve (see Figure 1a), where the optimal level is around 30% (see Figure 1b and Figure 1c). In addition, we find a saturation effect of the number of followers on marketing popularity after it reaches around 100,000. Our findings imply that social media marketing should not overlook customers' social needs. Our measure of marketing aggressiveness provides a dynamic business metric for practitioners to monitor so as to improve their marketing and managerial decision making process.



(a) Regression results

Variables	2014	2016
MAL	0.567***	1.185***
	(0.189)	(0.315)
MAL2	-1.021***	-1.623***
	(0.183)	(0.313)
verify	0.0157	0.0133
	(0.0483)	(0.0755)
tao	0.0271	-0.0786
	(0.0324)	(0.0546)
# followers (YR transformed)	1.296***	6.204***
	(0.0715)	(0.404)
female	0.272***	0.0310
	(0.0357)	(0.0640)
# pictures		0.0418***
		(0.0105)
Constant	1.488***	0.159**
	(0.0851)	(0.0762)
Observations	3,872	1,208
R-squared	0.241	0.359

Note: *0.1, **0.05, ***0.01.

(b) Optimal aggressiveness in 2014 sample

(c) Optimal aggressiveness in 2016 sample

Figure 1: (a) Regression results indicating the existence of an optimal aggressiveness level; (b) Visualizing the optimal aggressiveness level in the 2014 sample; (c) Visualizing the optimal aggressiveness level in the 2016 sample.

*Corresponding author. Email: z.zhu@gre.ac.uk.

HOW PHONE CALL DETAIL RECORDS MAY CONTRIBUTE TO IMPROVE DEPRESSION ASSESSMENT IN OLDER POPULATION ? ADDED-VALUE OF THE PHONE CALL DIRECTION PARAMETER

Timothée Aubourg[1,2], Jacques Demongeot[1,3], Félix Renard[1], Hervé Provost[2], Nicolas Vuillerme[1,3]

1 Université Grenoble Alpes, AGEIS, Grenoble, France
2 Orange Labs, Meylan, France
3 Institut Universitaire de France, Paris, France

Abstract

Background

Analysis of phone call detail records (CDRs) is now recognized to represent a promising approach in contributing to the future of mental health research. However, some of its aspects still require deeper investigation. Particularly, no clear consensus exists in the use of the phone call direction parameter, which permits to distinguish outgoing from incoming phone calls in CDRs datasets. It is crucial to address this issue in the field of health monitoring. Indeed, an imbalance in social interactions between the individual and his social network, that could be evidenced by outgoing and incoming phone calls, could be representative of particular disruptions, physical as well as social, occurring throughout his life time. Besides, recent studies that investigate social interactions in *mhealth* by using only outgoing and incoming phone calls separately have reported divergent results. All in all, new consistent tools are required for harnessing CDRs analysis in health on a consistent and relevant way.

Objective

In this train of thought, whether and how the phone call direction parameter could be explicitly used in CDRs analysis and then applied in a mental health context need to be investigated. This paper is specifically designed to address this issue.

Methodology

To this end, we use a 12-successive-month dataset that combines CDRs and geriatric depressive scales (GDS) results of 26 volunteers older than 65 years. We specifically address two issues: (1) the existence of a particular phone call habit in the older adult regarding the direction of his phone calls, and (2) the existence of a relation between such a habit and the depressive health state of the individual.

Results

On the whole, our results show the existence of three clusters of phone call activity named (1) *proactive*, (2) *interactive*, and (3) r*eactive*. Then, by introducing two asymmetry indicators, namely (1) the asymmetry coefficient, and (2) the skewness coefficient, that synthesize these three phone call habits, we find significant correlations between each of these two asymmetry indicators and the GDS scores obtained by the older individual. In particular, we report that older individuals having a r*eactive* phone call habit obtain GDS results significantly higher than the other ones. On the contrary, individuals obtaining a GDS lower than 10 tend to have asymmetry indicators with positive values. Taken together, the present findings suggest the existence of relevant health-related insights contained in CDRs datasets when the phone call direction parameter is harnessed. We believe that taking this parameter into consideration in future studies by using asymmetry indicators could be valuable for improving machine learning of health predictive models in *mHealth*.

Keywords:

Call detail records ; Older population ; Depression ; Geriatric Depressive Scale ; Asymmetry

Address for correspondence

Timothée Aubourg: timothee.aubourg@orange.com

Enabling Socio-Economic Study for Relocating Refugees using Mobile Phone Data and Regional Statistics

M. Saravanan, Researcher, Ericsson Research, Ericsson India Global Services Pvt. Ltd

This study focuses on the necessity of relocating Syrian refugees to affluent prefectures within Turkey based on the inferences from refugee's mobile phone call records and the available regional statistics. Relocation can be expected to improve the socio-economic background of refugees. The Syrian refugee's population in Turkey is approximately 8.5 million which has almost 10% of world level refugee's population. Refugees influx recipient nation require employment and face challenges pertaining to social and economic acceptance. Nations accepting refugees tend to balance the induced burden by localizing refugees at bay. This leads to impoverished and serious survival conditions. In this proposal, we propose a mechanism of redistributing refugees in the recipient nations based on wellbeing data; ensuring efficient employment for satisfactory survival and coordinating seamless socio-cultural interaction thereby improving the life of refugees. We have applied machine learning techniques and statistical inferences to determine the refugee's distribution in different clusters of cities and carefully hypothesize the relocation rudiments. Related to this, we observed the mobility and call pattern of refugees between different identified clusters of prefectures. In this study, our hypotheses are experimentally evaluated with D4R dataset provided by Turk Telecom and suitable discussion were made relevant to this [1]. Data for Refugees (D4R) Turkey is a big data challenge based on refugee tag. Our experimental proofs highlight the substantial rationale behind the suggestion of relocating refugees to progressive prefectures in the Turkey region. Regarding this, we attempted to group the cities of recipient nation based on regional statistics (wellbeing index data) of different prefectures of turkey region [2] and by running different clustering methods [3] on the extracted social parameters. Based on outcomes, we identified the incoming refugee's movements through their calls within relevant cluster and analyzing overlapping community traits among clusters in terms of event (Voice and SMS) trends, refugee/non-refugee preferences and their interactions to propose viable relocation alternatives.

The D4R challenge has provided the CDR data, through which we can extract the different levels of mobility pattern of refugees and non-refugees in Turkey. The extracted data set will be helpful to identify the movement of refugees from one place to another place in search of better livelihood. Our cluster models are trained based on the consideration of providing higher weight to the nation wellbeing parameters: safety and security, health, education, unemployment, social integration and segregation, mobility, and distribution of resources and infrastructure. Also, we compared the refugee and non-refugee call pattern records to understand their present placement and discuss the need for improving social integration. We have trained the data with different combinations of the K-means, PCA, Self-organizing Maps (SOM) and deep clustering algorithm in this process. From the combined effort of SOM and K-Means clustering results, we identified three unique clusters namely *prosperous, less prosperous, and underdeveloped*. With the use of deep clustering, our results are improved by removing outlier samples. To argue our intention of relocating refugees to prosperous cluster, we proposed three null hypotheses and tested through statistical measures.

1. *There will not be any significant difference between refugee placements in different clusters of prefectures in the regions of Turkey*
2. *Refugees movement in regions of Turkey will not represent any patterns relevant to their social upgrade.*
3. *There is no difference in the refugee call patterns across the developed clusters in comparison with non-refugees towards social integration.*

Most of the refugees are not living in the prosperous area. Only 32% of refugees are living in prosperous. Remaining around 68% of refugees are not living in prosperous areas. Out of 68% refugees, we found only 7% are living in less prosperous and 61% are living in the underdeveloped areas. Therefore, by rejecting the first null hypothesis, we proved that there is a significant number of refugees are staying in underdeveloped areas and it clearly necessitates the need for relocation of refugees to the prosperous cluster. The refugees living in underdeveloped should develop more social integration with non-refugees to create a good atmosphere for the country development. Using Chi-square statistics, we compared the three cluster samples and proved the fact that for socially upgrading refugees, there is a necessity to move refugees from underdeveloped to the prosperous or less prosperous clusters. Also, we found evidence that there is a significant difference between the call pattern in refugees and/or non-refugees who are staying in prosperous and underdeveloped clusters. The same thing is applicable for the calls made between less prosperous to underdeveloped and less prosperous to prosperous locations. Hence, we tested using t-test statistics between the calls made by refugees from prosperous locations to others and non-refugees of a similar domain. From the value, it is evident that there is a difference in mean values of their call summary and movement patterns which ensures that the refugees and non-refugees show the same pattern in the prosperous home clusters towards the social integration. Finally, we assure that this relocation refugee study can easily be extended to explore in other countries for improving the socio-economic status of their refugees and by considering the wellbeing status of prefectures attached to that country.

References:

1. Albert Ali Salah, Alex Pentland, Bruno Lepri, Emmanuel Letouz´e, Patrick Vinck, Yves-Alexandre de Montjoye, Xiaowen Dong, Ozge Da˘gdelen. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey, arXiv:1807.00523, July 2018
2. Well-being index for provinces, under the statistical table, http://www.turkstat.gov./PreTabloArama.do.
3. Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Series in Data Management Systems, 4th Edition.

Mobile data to analyse ground access and competition at airports

Amir Brudner[1], Riccardo Gallotti[2,3], Filippo Privitera[4], José J. Ramasco[3], Nicole Adler[1]

[1]School of Business Administration, The Hebrew University of Jerusalem,
Mount Scopus, 91905 Jerusalem, Israel
[2]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy
[3]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB),
07122 Palma de Mallorca, Spain
[4]Cuebiq Inc., 45 W 27th Street, 3rd floor, 10010, New York, NY, USA

Ease of access and egress to airports is likely to impact demand and therefore should be of interest to the aviation supply chain from the perspective of both short-run and long-run investment planning. Data is however often limited in accuracy and based on sporadic local surveys and as a consequence, managers and policy makers must take decisions based on partial information. Recent developments in the use of Information and Communication Technologies (ICT) provide new, alternative data-sources which may lead to more precise detail with respect to individual mobility at different spatial scales [1]. ICT data poses some challenges due to the need to correct potential biases, but equally may overcome some of the traditional limitations.

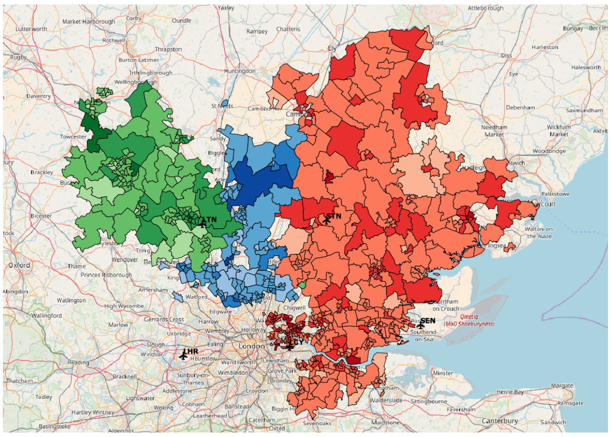
In this research, we investigate mobile data which provides a new, comprehensive perspective on door-to-door airport accessibility in the greater London region. By analysing the data using discrete choice models, we identify airport catchment areas in a large city served by multiple airports. The focus on Greater London allows us to showcase the potential of very large data sources to answer the long-debated question as to whether airports serving the same urban area in fact compete (see Figure).

We analyse the demand of both locals and tourists for each of the six airports drawing from all MSOAs in England and Wales. In total, the dataset includes 104,158 trips by locals and 4,580 trips by tourists in the six month period from January to July 2017. Validating big data is an important element of the analysis [2]. Consequently, we compare the dataset to the UK Civil Aviation Authorities information from 2017, after removing connecting passengers. Although the demand mapped with mobile phone records represents less than 0.1% of the market, the annual market share across airports is very similar, with an error of less than 2% with the notable exception of London City and Gatwick. Indeed, we note a slight bias towards coverage of the more European business-oriented passengers than the general public on average, hence Gatwick is slightly under-represented whilst London City is over represented. Additional validation draws from a comparison of the median measured access times from each MSOA to that of Google maps data.

We utilise the aggregated GPS records generated by the use of mobile apps by anonymous opted-in users to estimate distance and ground access times to airport facilities. Our analysis highlights the role of both ground transportation and the socio-demographic background of passengers in choosing between alternative airports. In support of this data-driven perspective, we use rational choice theory [3] to model the decision behaviour of passengers when facing the option of multiple airports.

The results of the logistic regressions suggest that a reduction of 10% in the relative distance between an MSOA and an airport (for example by improving ground transport access), increases the odds of serving demand by 17%. Furthermore, a 10% increase in the population of an MSOA (for example by increasing density) increases the odds by 12%. Ground transport access is a significant variable both in explaining demand and its magnitude. With respect to car ownership, our results suggest that 10% lower ownership leads to 4.5% higher demand and this is due to the fact that the MSOAs with higher population densities lead to lower car ownership shares.

The quantitative information gathered, integrated with economic data, can be used to inform a two-stage Hotelling style game. This new approach throws light on the opportunity for collusion between airports and between airports and airlines, and may be used by regulators and infrastructure planners to improve decision making at the strategic level.



Luton and Stansted catchment areas. We highlight MSOAs located up to an hour from each airport with access time differences of more than 30 minutes (green and red) and less than 30 minutes (blue). The overlapping region in blue consists of 110 MSOAs and approximately 800,000 people. Of the passengers analysed in these zones, 55% chose Luton and 45% chose Stansted.

[1] Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. Nature physics, 8(1), 32. ISO 690

[2] Khan, N., et al. (2014). Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal, 2014.

[3] McFadden, D. (1974). The measurement of urban travel demand. Journal of public economics, 3(4), 303-328.

Elements for an official statistical production with mobile network data

D. Salgado[1], L. Sanguiao[1], B. Oancea[2], M. Necula[2], T. Tuoto[3], S. Hadam[4], A. Keskyla[5], S. Williams[6], A. Condron[7]

[1] Statistics Spain (National Statistical Office of Spain)
[2] INS (National Statistical Office of Romania)
[3] Istat (National Statistical Office of Italy)
[4] DESTATIS (National Statistical Office of Germany)
[5] Statistics Estonia (National Statistical Office of Estonia)
[6] Office for National Statistics (National Statistical Office of UK)
[7] Central Statistical Office (National Statistical Office of Ireland)

Mobile phone data are a promising data source for producing official statistics. They offer the possibility to reach unprecedented scales of spatial and time disaggregation for relevant statistics in different domains (population, tourism, labour market, etc.) and also, more innovatively, to tackle novel statistical products of social interest related with the newly born network science.

Diverse studies from academia, the research community around Mobile Network Operators (MNOs) and some leading National Statistical Institutes (NSIs) have clearly shown the potential to produce official statistics using mobile phone data. The challenge now is to incorporate them in a standardized production system enabling statistical offices to make use of these data in many diverse domains of social interest. The European Statistical System (Eurostat and European NSIs) is constructing a production framework, facing important scientific and technical challenges, and proposing concrete solutions for each of them:

- Access to mobile phone data, despite the regulatory support in National and European Statistical Acts, is proving to be a major challenge entangling many facets. The goal is to find satisfactory partnership models between MNOs and NSIs. The Reference Methodological Framework (see parallel contribution by Eurostat and others) proposes a layered modular approach in which the notion of *event location* plays a central role in this respect. Furthermore, to comply with regulatory demands, we are advancing in the implementation of privacy-preserving computation methods. All in all, this entails an excellent opportunity to establish long-standing synergic collaborations among academia, MNOs and statistical offices.
- The geolocation of network events is designed to reach functional modularity in the whole process in order to decouple highly technology-dependent stages from the upper-level statistical analysis. Bayesian techniques to deal with overlapping BTS cells avoiding Voronoi tessellations have been implemented in an R package called *mobloc*.
- As with many other new digital data sources, the traditional survey methodology and design-based inference do not apply any longer and a new inference framework for Official Statistics is needed. Bayesian hierarchical models, inspired by the species abundance problem in Ecology and some population estimation methods with administrative data, are under exploration enabling us to integrate data from different sources (population registers, survey/admin data, mobile technology penetration rates…). Another R package called *pestim* is also under construction.
- The quality assurance framework, structured around the European Statistical Code of Practice, needs a revision. Quality dimensions such as timeliness and opportunity are reinforced but others, like accuracy, needs appropriate addressing in the new inferential framework. We investigate quality indicators based on posterior point estimates, credible intervals, and model checking using the posterior predictive distribution. Some of these have been already implemented in the pestim package.

In this contribution we present details of this ongoing work. Mobile phone data for official statistical production offers an excellent opportunity to find synergies among academic researchers, telco professionals, and official statisticians.

Apps, Places and People: strategies, constraints and trade-offs in the physical and digital worlds

Marco De Nadai[1,2], Angelo Cardoso[2], Antonio Lima[2], Bruno Lepri[1] and Nuria Oliver[2]

[1]Fondazione Bruno Kessler
[2]Vodafone Research
{*angelo.cardoso,antonio.lima,nuria.oliver*}*@vodafone.com*
{*denadai,lepri*}*@fbk.eu*

Cognition has been found to constrain several aspects of human activity, such as the number of friends and favorite places a person can maintain stable over time [1, 2]. Do people exhibit similar constraints on digital devices? We address this question through the analysis of a dataset consisting of pseudonymized mobility and mobile apps usage data of 400'000 individuals in a European country for 8 months.

In terms of mobility, we compute the stop locations from GPS data. Regarding applications, we analyze the mobile phone apps used in the foreground. Based on previous definitions in mobility and social interactions, we compute the mobile users' apps and mobility capacity, defined as the set of apps used at least twice in a time window of 20 weeks and the set of stop locations visited at least twice in 20 weeks, respectively.

Despite the enormous heterogeneity of apps usage, we find that individuals exhibit a conserved capacity of applications they regularly use. This capacity steadily decreases with age, while the mobility capacity decreases non-monotonically with age (see Figure 1A). We show that the capacity of applications might be constrained by cognitive features. Next, we identify two profiles of individuals: apps *keepers* and *explorers*, which correspond to those users with stable vs exploratory apps usage behavior. Finally, we show that the capacity of applications predicts mobility capacity and vice-versa (see Figure 1B). By contrast, the profiles do not always match the across domains, such that keepers in the apps domain could be explorers in the physical space and *vice-versa*.

Our empirical findings provide an intriguing picture linking human behavior in physical and digital worlds which is relevant to related research in Computer Science, Social Physics and Computational Social Sciences.

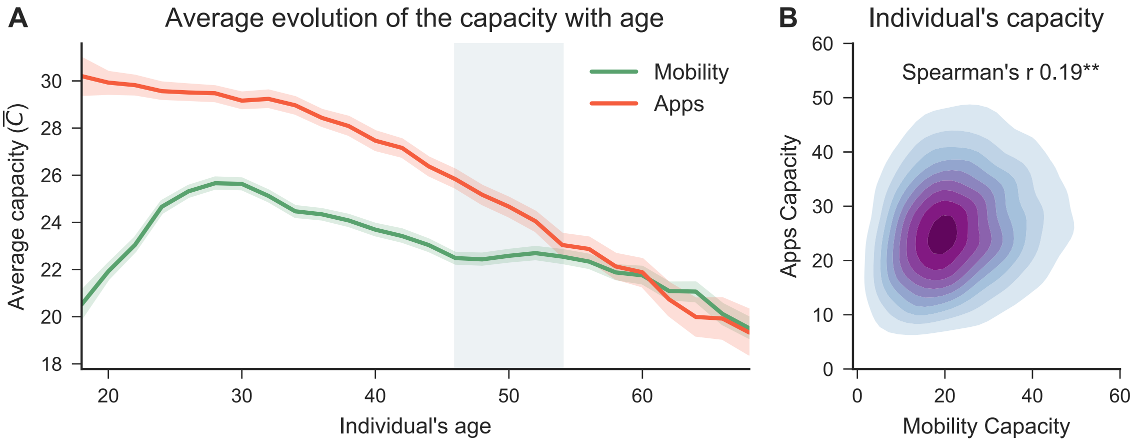


Figure 1: A) The average capacity of applications usage steadily decreases with age, while mobility capacity increases until around 28 years, then it decreases until 46 where it is constant, to decrease again starting from 56 years; B) The correlation between individual's mobility capacity and apps capacity.

References

[1] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491, 2018.

[2] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3:1950, 2013.

In A Society of Strangers, Kin Is Still Key: Identified Kin Relations In Large-Scale Mobile Phone Data

Sebastian Diaza[a,b]
Loreto Bravo[b]
Carlos Rodriguez-Sickert[a]
Isabel Behncke[a]
Anna Rotkirch[c]
János Kertész[d,e,f]
Tamás Dávid-Barrett[a,c,g,h]

[a] Universidad del Desarrollo, Facultad de Gobierno, CICS, Av. Plaza 680, San Carlos de Apoquindo, Las Condes, Santiago de Chile, 7610658 Chile
[b] Data Science Institute, Universidad de Desarollo, Av. Plaza 680, Las Condes, 7610658, Santiago de Chile, Chile
[c] Population Research Institute, Väestöliitto, Kalevankatu 16, Helsinki 00101, Finland
[d] Central European University, Center for Network Science, Nador u. 9, Budapest, H-1051, Hungary
[e] Department of Computer Science, Aalto University School of Science, P.O.Box 15500, 00076 Finland
[f] Department of Theoretical Physics, Budapest University of Technology and Economics, H1111, Budapest, Hungary
[g] Trinity College, University of Oxford, OX1 3BH, Oxford, UK
[h] Kiel Institute for the World Economy, Kiellinie 66, D-24105 Kiel, Germany

Abstract

Recent research has identified a way of detecting kin and peer relationship types from anonymous mobile communication patterns, based on a combination of phone call patterns and basic socio-demographic data (age and gender) of the callers. This new methodology exploited the fact that the average mobile phone caller has up to six distinct peaks in the histogram of call frequency, as a function of the alters' age and gender. Although the new methodology allowed a more refined differentiation among close kin networks, it had substantial limitations as these relationships were merely hypothesised. Here, we are able to move beyond this methodological limitation using a unique, large mobile phone data base with information about individual surnames in a population in which people inherit two surnames: one from their father, and one from their mother. Using this mobile phone database, here we focus on the difference between the most frequently called alters depending whether they are family relationship or not. We report tests on two hypotheses: (1) within category kin vs. non-kin relationship drives call frequency and call length; and (2) life-course dependent variation of call pattern is higher in kin than in non-kin.

UNSUPERVISED MODELING OF CHRONOTYPES

IN PHONE CALLS IN OLDER ADULTS.

Timothée Aubourg[1,2], Jacques Demongeot[1,3], Félix Renard[1], Hervé Provost[2], Nicolas Vuillerme[1,3]

1 Université Grenoble Alpes, AGEIS, Grenoble, France
2 Orange Labs, Meylan, France
3 Institut Universitaire de France, Paris, France

Abstract

Background

In the field of complex systems, recent studies have shown relevant results in modeling circadian rhythms of social activity, as well as their chronotype, by means of the analysis of phone call detail records (CDRs) in social networks of young individuals. In the field of health monitoring, this scientific breakthrough offers the opportunity to reinforce current methods of disease surveillance by including real-time personalized social interaction data obtained on a passive and non-invasive way. Interestingly, this from-complex-system-to-health application underlines the potential of a project combining the competence of multi-experts as data scientists and health practitioners for addressing the future of health research. However, key questions still require answers regarding the observed CDRs and their interpretation.

Objective

In particular, whether and to what extent chronotypes could also be observed in phone calls in other types of population remain to be investigated. This paper is specifically designed to address this issue regarding an older population.

Methodology

To this end, we use a 12-successive-month CDRs dataset of 26 volunteers older than 65 years. We specifically implement an unsupervised machine learning algorithm in order to investigate (1) the existence of chronotypes in phone calls in older adults by modeling their phone call activity with Gaussian mixture models (GMM), and (2) the existence of different clusters of chronotypes in phone users based on the GMMs results, by means of a K-means clustering approach.

Results

On the whole, our results do evidence about (1) the ability of GMMs to consistently model circadian rhythms of phone call activity in an older population, and (2) the existence of distinct categories of chronotypes among individuals regarding these rhythms. In particular, it is interesting to note that, by means of GMMs, our unsupervised algorithm permits to model the daytime bimodality of phone calls histograms in our older population, but it can also catch unusual peaks of activity occurring at nocturnal period. By using a K-means clustering approach, we then are able to characterize the morning or evening preference for phone calls time in older adult from the GMMs results. Taken together, these findings suggest that older adults do also present chronotypes in phone call activity that could be evidenced by machine learning approaches. We believe that using these last ones in complement to complex systems methodologies could provide relevant results both in modeling complex human activities, and also in offering innovate health monitoring tools for disease surveillance.

Keywords:

Call detail records ; Older population ; Unsupervised learning ; Clustering ; Gaussian mixture model ; K-means

Address for correspondence

Timothée Aubourg: timothee.aubourg@orange.com

Gender Analytics and Identification Toolkit (GAIT)

Lucio Melito, Kristyna Tomsu, Rositsa Zaimova, Denys Sementsov, Jerome Urbain
Dalberg Data Insights, lucio.melito@dalberg.com

I. Introduction

Telecom operators often lack accurate gender data necessary to proactively work towards connective gender parity, especially in developing countries. Existing research shows that phone usage patterns differ between men and women, confirmed by survey data [1] as well as studies working directly with call detail records to predict subscribers' gender and fill the gender data gap [2, 3, 4, 5]. Commonly reported distinguishing features for women are longer call duration, generally lower phone usage, and lower credit recharges. Different studies report gender prediction accuracy between 70 and 80%.

In partnership with GSMA Connected Women, Dalberg Data Insights has developed a Gender Analytics and Identification Toolkit, first implemented in Bangladesh. This toolkit can provide any telecom operator with gender-disaggregated data on phone usage patterns and machine learning models to predict gender for subscribers with no available reliable gender information.

II. Methodology and data

Data: 3 months of anonymized CDRs from ca 40 million subscribers of Robi Axiata in Bangladesh; for 15,000 subscribers complemented by ground-truth gender labels coming from a phone survey.

Methodology: We computed over 140 indicators summarizing phone behavior of the phone users in Bangladesh. The indicators can be divided in the following categories: generic usage (e.g., number of incoming/outgoing calls per day), social (e.g., number of distinct contacts), top-ups (e.g., an average top-up value), bundles (e.g., bundle value per type), mobility (e.g., radius of gyration).

We used a 15,000 subset of the subscriber base, for which we had ground-truth gender labels, to train and evaluate (80%-20% train-test ratio) several machine learning models to predict the subscribers' gender based on the log-transformed and normalized indicators. Through a combination of Grid, Randomized and Bayesian optimization search we tested hundreds of hyperparameter combinations for several model types (Linear Support Vector classifier, Support Vector Machine, K-Nearest Neighbors, Random Forest and XGBoost), choosing the best set of hyperparameters through 3-fold cross validation, after which the model was re-trained on the whole training set. The model with the highest accuracy was then applied to predict the gender of the rest of the subscriber base.

All the algorithms for data pre-processing, indicator computation, model training and application are packaged in an easy-to-use and easy-to-deploy toolkit, implemented in Python leveraging pySpark API.

III. Results

Survey: According to the survey, 30% of the subscriber base are females. When comparing the survey responses to the gender information available to the operator from the SIM registration process, we found out that 78% of female phone users were registered as men.

Predictive model: The best scoring model was XGBoost reaching an accuracy of **84.5%** (i.e., out of all predicted labels, 84.5% were correct)**,** precision of 79.4% (i.e., out of all predicted female labels, 79.4% were correct) and recall of 61.2% (i.e., out of all females in the dataset, 61.2% were correctly identified by the model).

The top three indicators contributing to the prediction were: i) Average duration of incoming call (women receive on average longer calls), ii) Number of contacts (women have on average less contact numbers), iii) Radius of gyration (typical travel distance) for average active day (women on average travel less).

IV. Discussion

We demonstrated that gender data gaps in telecom data can be filled using a small sample with reliable gender labels and an automated machine learning pipeline. Gender-disaggregated CDRs can help the operators to better understand the differences between men and women in their call behavior and usage of products. This can lead to increased usage of services in the user base as well as better targeted campaigns to acquire more female customers, and thus to increase mobile phone penetration among women, potentially leading to many societal benefits.

Once set up, the packaging of the toolkit allows to run all the computations and predictions in one click, enabling to get gender-disaggregated data even in contexts where the operator doesn't have the time or technical capacities to estimate the subscribers' gender in a different way. The toolkit, including thorough documentation, is now freely available to more than 750 GSMA Member operators across the world, aspiring to fill gender data gaps in several places where gender-disaggregated data is scarce and contributing to the SDG of achieving gender equality.

References:

[1] GSMA, "Connected Women: The Mobile Gender Gap Report 2019," 2019. [Online]. Available: https://www.gsma.com/mobilefordevelopment/resources/mobile-gender-gap-report-2019/.

[2] J. Blumenstock and N. Eagle, "Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda," *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development,* p. 6, 2010.

[3] B. Felbo, P. Sundsøy, A. S. Pentland, S. Lehmann and Y. A. de Montjoye, "Using Deep Learning to Predict Demographics from Mobile Phone Metadata," *arXiv preprint arXiv:1511.06660,* 2015.

[4] V. Frias-Martinez, E. Frias-Martinez and N. Oliver, "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records," *AAAI spring symposium: artificial intelligence for development,* 2010.

[5] J. E, P. R. Sundsoy, J. Bjelland, A. Iqbal, A. Pentland and Y. A. de Montjoye, "Predicting gender from mobile phone metadata," *Netmob 2015 Book of Abstracts: Oral, online resource: http://netmob.org/assets/img/netmob15_book_of_abstracts_oral.pdf,* 2015.

Employment and social capital scores

Lucio Melito, Rositsa Zaimova, Kristyna Tomsu, Denys Sementsov, and Jérôme Urbain

Dalberg Data Insights, Brussels, Belgium, lucio.melito@dalberg.com

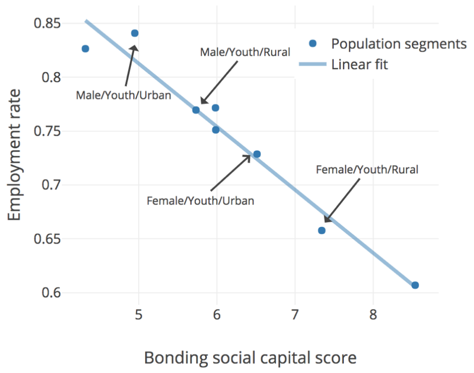


Fig. 1. Bonding social capital. Employment rates of different population segments vs their bonding social capital scores.

Introduction. The concept of social capital has deep historical roots and occupies a prominent place in sociology, where it is defined as “the aggregate of the actual or potential resources which are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition” (1). At the same time, some of the original quantitative research on social networks was kick-started by a number of studies in the 1960s and 1970s (2), which found that a majority of people who had recently changed jobs had learned information about their current position through personal contacts. Perhaps more surprisingly, they described these connections as acquaintances rather than close friends, a finding later known as “the strength of weak ties”. This and other research contributed to academia’s understanding of the subject, with a clear dichotomy widely agreed upon: social capital can be of a bonding or of a bridging kind (3). Bonding capital arises from connections within a tightly-knit group, while bridging capital is from connections between such groups. Given the previous points, it seems natural to hypothesise that bridging social capital would have a sizable positive effect on the employment prospects of individuals, while bonding capital should do the opposite.

We set to investigate this hypothesis in a East African country, where it could be tested on the dataset of Call Detail Records (CDRs) of the major telecom operator in the country, comprising the telephone calls and messages of more than 10 million users. Our research focused on a subset of ten thousand of these subscribers, for whom we had previously gathered ground truth information on age, gender, location and employment status, through a telephone survey that was also used to obtain their explicit consent.

	Bonding score	Bridging score
Employed	0.056 ±0.001	13.5 ±0.48
Unemployed	0.075 ±0.003	7.1 ±0.39

Table 1. Mean bonding and bridging scores for employed and unemployed users, with associated standard errors.

Data and methods. The full dataset of interactions for the month of November 2018 was therefore translated into an undirected graph with nodes representing users and an edge connecting two nodes only if their corresponding users had shared a call or exchanged a text-message during that month. Bonding and bridging social capital scores for the subset of 10000 nodes were computed according to the following definitions:

- **Bonding score**: the probability that any two of that node’s connections would also have a link between them, also known as the clustering coefficient.
- **Bridging score**: the number of shortest paths that pass through that node, i.e. the betweenness centrality of the node; due to the computational impossibility of computing all the shortest paths in a graph with more than 10 million nodes, only a small number of vertices were randomly sampled to compute the score of each node.

Results and discussion. The mean scores and their standard errors, disaggregated by employment status, are shown in Table 1. The differences are stark, with employed people showing significantly higher bridging scores and lower bonding scores, confirming our hypothesis. Figure 1 displays the employment rate of different segments of the population and their average bonding social capital score, with a near perfect (>0.95) negative correlation between the two metrics. These results are also strengthened by the fact that the average numbers of contacts for employed and unemployed persons are extremely similar, when a difference could have partially explained the discrepancy in social capital scores.

The correlation between employment and social capital metrics does not necessarily imply causation, and it is likely that both directions of the arrow of causality are in action: being employed broadens someone’s social circle and put them in touch with a more diverse set of individuals, while having a higher bridging social capital score in the first place will increase the chances of hearing of valuable job opportunities and also increase the probability of landing a job.

Bibliography

1. Pierre Bourdieu. The forms of capital. 1986.
2. Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.
3. Robert D Putnam. Bowling alone: America’s declining social capital. In *Culture and politics*, pages 223–234. Springer, 2000.

Melito *et al.* | bioRχiv | March 15, 2019 | 1–8

Socio-Economic Factors in News Consumption Using Mobile Phone Data

S. Vilella[1,4], L. Ferres[2,3], D. Paolotti[1], and G. Ruffo[4]

[1]ISI Foundation, Turin, Italy
[2]Data Science Institute, Faculty of Engineering, Universidad del Desarrollo
[3]Telefónica R&D, Santiago, Chile
[4]University of Turin, Turin, Italy

Introduction. In previous work, we have been able to show the effects of certain press manipulations by owners of content [BBE+18, EFHB18]. These results, based on hypotheses born out of an operationalization of Herman and Chomsky's Propaganda Model [HC02] and their "filters" give a good idea of how the media behaves. One particular hypothesis is that the media will manipulate content to *target* news to a certain audience. We've been able to show this to a certain extent using Twitter [EFS18]. What we study here is a specialization of that study; namely, how people of different socio-economic backgrounds access particular news outlets of certain characteristics at the finest possible level of granularity using their mobile pphones. The hypothesis is that the more educated the population, the more they will read, but the more specialized they will be (because of audience targetting).

Data and Methods. We obtained a month (between July and August 2016) of anonymized network interaction (deep packet inspection) through the cellphone network of a company that possesses 31% of the marketshare of mobile phones in Chile. Information was aggregated by antenna and by hour, without any single user information, making it virtually impossible to de-anonymize news consumers. The dataset, called `dsUsers`, includes the *number of unique users* that connected to that antenna and to an IP address belonging to one specific news outlet out of a curated list at a certain hour (00, 01, 02... 23), so for example, an entry in our database looks like `ABCD0120,20160706,11,200.12.26.117,1`, means that on July 7, 2016 at 11am, there was one unique user connected to IP 200.12.26.117 at antenna ABCD0120. The exact position of each antenna in the city is known, and we then group together all the antennas within a 1.1km radius, obtaining a lattice of about 700 points that will be our new, "fictitious" antennas. The geo-location of the antennas allows us to cross the data about consumption of news media content with the socio-demographic features of the users living in the different areas of the city. Each antenna is assigned a *census label* that refers to a census district in which it is located, obtained by running a K-means clustering algorithm on the publicly available 2017 Chilean Census data.

Results. By analyzing the access to news media websites in the light of the *census label* of the zone from which they are made, we are able to find patterns and preferences of users that in different socio-demographic contexts, see Figure 1. By looking at the general trends: the consumption of news media content does not correlate with the education level of the user (which, incidentally, is known to correlate very well with income distribution and inequality). Indeed, the most eager consumer of news media content are those living in the lower-middle census areas, with a low education level. By going deeper into this analysis, they seem to prefer the more generic media outlets, like *Biobio* or *Cooperativa*, while the most educated usually prefer to read more specialised news media, like *Diario Financiero*, or more explicitly political oriented media like the conservative *El Mercurio* or the leftist *The Clinic.* Among these evidences, one constant element stands out: the people accessing these websites from the most deprived areas of the city display the lowest activity. Accesses from these areas indeed are always the lowest, in absolute values, regardless of the typology of the news media that we consider.

Conclusions. The intuitive hypothesis that consumption of news media content grows with education level of the user seems to be disproved here. Indeed, it seems that the least educated, those who live in the most deprived areas, are also those who access news media websites the least. This suggests us that public data is, alone, a good reference for basic intervention and policy making, in order to facilitate and promote the access to news media content among particular zones of an urban area, and make a step forward in levelling out urban inequalities; nonetheless, mobile data proves to be an incredibly valuable asset in such an analysis.

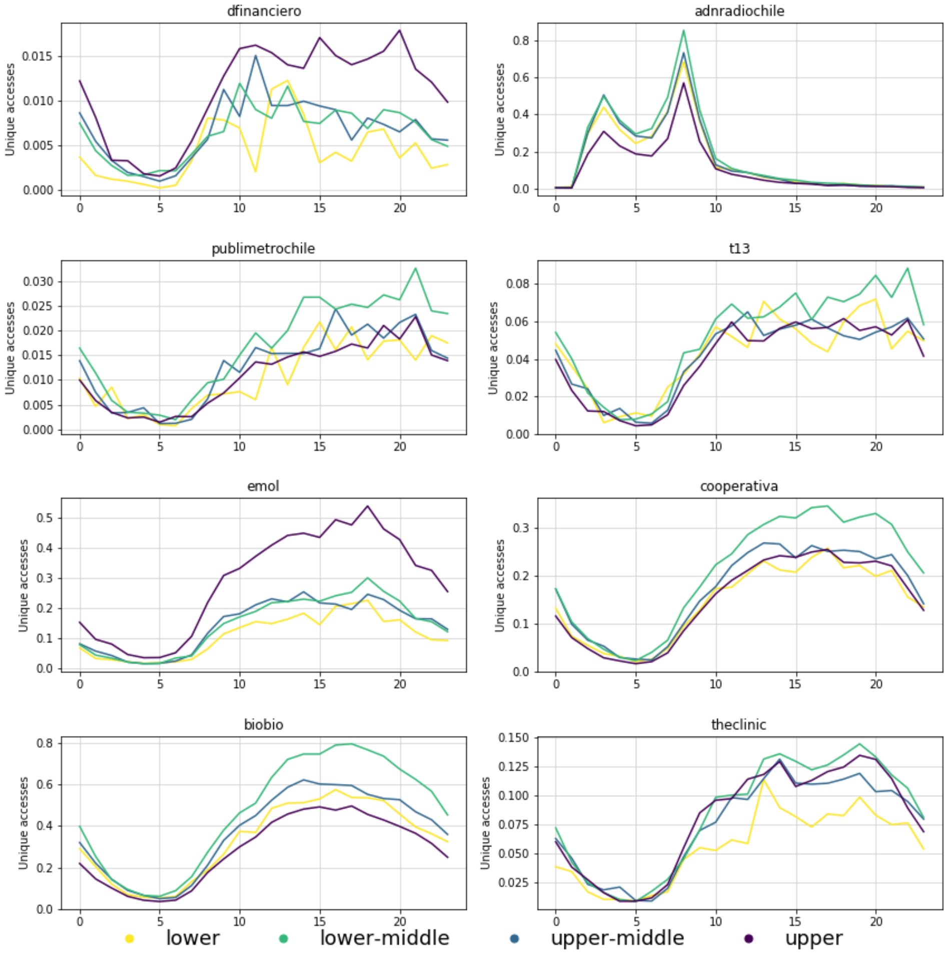


Figure 1

References

[BBE+18] Jorge Bahamonde, Johan Bollen, Erick Elejalde, Leo Ferres, and Barbara Poblete. Power structure in chilean news media. *PLOS ONE*, 13(6):1–18, 06 2018.

[EFHB18] Erick Elejalde, Leo Ferres, Eelco Herder, and Johan Bollen. Quantifying the ecological diversity and health of online news. *Journal of Computational Science*, 27:218 – 226, 2018.

[EFS18] Erick Elejalde, Leo Ferres, and Rossano Schifanella. Understanding news outlets' audience-targeting patterns. *CoRR*, abs/1806.02155, 2018.

[HC02] Edward S. Herman and Noam Chomsky. *Manufacturing Consent: the Political Economy of the Mass Media*. Pantheon Books, New York, NY, 2nd. edition, 2002.

Relations between Mobile Phone Indicators and the Socioeconomic Organisation of Cities

Maarten Vanhoof[a,b], Clementine Cottineau[c]

[a] *Centre for Advanced Spatial Analysis, University College London, London, UK*
[b] *SENSe, Orange Labs France, Chatillon, France*
[c] *CNRS, Centre Maurice Halbwachs, UMR 8097, Paris, France*

Abstract

In this article we explore the convergence between mobile phone data and more traditional socioeconomic data from the national census in French cities. We extract mobile phone indicators from six months worth of Call Detail Records (CDR) data, while census and administrative data are used to characterize the socioeconomic organisation of French cities. We address various definitions of cities and investigate how they impact the statistical relationships between mobile phone indicators, such as the number of calls or the entropy of visited cell towers, and measures of economic organisation based on census data, such as the level of deprivation, inequality and segregation. Our findings show that some mobile phone indicators relate significantly with different socioeconomic organisation of cities. However, we show that relations are sensitive to the way cities are defined and delineated. In several cases, changing the city delineation rule can change the significance and even the sign of the correlation. In general, cities delineated in a restricted way (central cores only) exhibit traces of human activity which are less related to their socioeconomic organisation than cities delineated as metropolitan areas and dispersed urban regions.

Keywords: cities, mobile phone data, mobility, deprivation, segregation, inequality

Email addresses: `m.vanhoof@ucl.ac.uk` (Maarten Vanhoof), `clementine.cottineau@ens.fr` (Clementine Cottineau)

Coverage Anomaly Detection with Mobile Analytics

Daniel H. Chae, Jungmin Moon, Jungmin Yoo, Intaek Lee
SK Telecom, Republic of Korea
{dani75.chae, jmmoon, jungminyoo00, intae31.lee}@sk.com

***Abstract*—As mobile internet connection becomes our daily necessity, it is important for telecommunication company to provide a stable wireless connection. One way of providing stable connection is to operate a mobile network efficiently without a coverage outage. However, since the wireless network equipments such as base-stations are located spreadly over the nation, it is not easy to detect malfunctions of the whole network nodes. This paper introduces how we can improve the network stability by applying mobile data analytics into anomaly detection of service coverage. Big data collected from the whole network is analyzed, some designated network nodes covering a specific area is analyzed with correlation of network equipment (NE) analytics and user equipment (UE) analytics. It shows this method is very efficient in analyzing service quality of massive mobile traffic zones such as subway and train where NE analytics does not represent service quality accurately due to an intermittent traffic pattern.**

***Index Terms*—UE Analytics, Wireless Connection Coverage, Anomaly Detection**

I. Introduction

With streams of network big data analytics, there are needs to improve an operation efficiency in telecom industry. Accurate measurement of customer experience, anomaly detection and fault recovery of NE are highly required for the operation efficiency. Network big data analytics is believed one of solutions for these purposes and it can be applied into network operations supports system (OSS) for maintaining stable wireless service.

Conventionally, NE analytics with performance statistics is widely under practice for network monitoring purpose. As telecom subscribers want more broadband and more uninterrupted service, however, more detailed analysis of customer experience, which is beyond the scope of NE analytics, is required to satisfy them. Automatic measurement of customers experience is important in the era of high demanding quality-of-service (QoS). This is reason why there is a need to extend analytics from NE-level to UE-level by measuring customer's experience accurately with UE analytics.

In this paper, by analyzing individual UE data with less complexity, we introduce a method of coverage anomaly detection in a massive traffic zone where careful network operation should be exercised, and outcomes of accurate tracing of service area and its anomaly detection is shown. This method is useful in providing stable network operation, and it is embedded into SK Telecom's OSS called *TANGO (T-Advanced Next Generation OSS)* [1], [2] as shown in Fig. 1.

II. Correlation of UE and NE Analytics

In perspective of network operation, network status can be assured with NE analytics. NE analytics is performed by analyzing the data provided from NE via northbound interface. NE data includes thousands of key performance indicators (KPIs) and alarm data representing NE status. Additionally, UE data, which contains detailed records of control-plane and data-plane of the subscriber in the network, can be analyzed for assessing subscribers experience though there is a complexity burden due to an amount of UE data size. To detect a service coverage anomaly accurately in a massive mobile traffic zone with less complexity, we introduce a method correlating UE analytics and NE analytics.

III. Expectation and Result

With the suggested method, we can improve customer experience by minimizing anomalies of radio cell coverage in a specific zone efficiently. To find base-stations covering a specific zone, UE analtytics is performed, and then NE analytics is followed in detecting any anomalies of the zone. The method of combining UE and NE analytics can relieve a computation burden in anomaly detection. Detailed result will be included in the full paper.

References

[1] https://www.telecomasia.net/content/skt-expanding-use-tango-ai-platform, "SKT expanding use of TANGO AI platform"

[2] https://ovum.informa.com/resources/product-content/using-ai-in-csp-network-operations-what-operators-have-deployed, "Using AI in CSP Network Operations: What Operators Have Deployed"

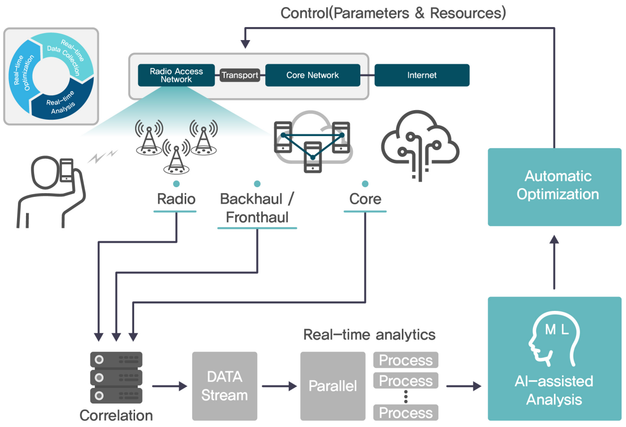


Fig. 1. TANGO Data Analytics

Towards Official Tourism Statistics - Machine Learning for Processing Signalling Data

Marc Ponsen, Yvonne Gootzen, Marco Puts, Martijn Tennekes, Edwin de Jonge, Shan Shah, May Offermans

March 15, 2019

Statistics Netherlands is working on the next step towards official tourism statistics based on anonymised mobile phone signalling data. The privacy preserving process consists of four steps: geolocation, feature extraction, machine learning and aggregation.

Our geolocation algorithm processes information about an antenna network into probabilities that a connecting device is located in a given location.

In the feature extraction step, all observed antennas and their geolocation probabilities are analysed for a given time period. Multi-class particle filters utilise the temporal component of these observations. The output of these models is twofold: trajectory estimation and transportation mode detection. Trajectories describe an estimated location of a device over time. Transportation mode models are used for particle class definitions. The survival rates of each particle class is interpreted as the similarity between the actual movement and a theoretical transportation mode definition.

Extracted features for foreign devices are used as input for an unsupervised machine learning clustering algorithm. The resulting clusters are used to distinguish potentially new tourism developments. This allows for data-driven definitions, rather than analysis influenced by cluster definitions based on foreknowledge.

Detailed pre-aggregation analysis result in more accurate anonymised tourism statistics.

Subway Mobile Analytics for Better Experience

Dayea Yim
KOTRA, SOUTH KOREA
dyyim@kotra.or.kr

Keejoo Bae, Jungmin Yoo, Daniel H. Chae
SK Telecom, SOUTH KOREA
{kjbae, jungminyoo00, dani75.chae}@sk.com

***Abstract*—As consumption of mobile network service s such as video streaming and mobile games is increasing while using public transit, seamless mobile connection service is highly necessary. To satisfy customers with better experience, stable wireless connection should be provided in places where a large number of people gather simultaneously and move quickly. For example, in case of the subway or the high-speed train telecommunications companies manage mobile network service more carefully by continuous monitoring of radio signal quality. Generally, manual measurement of signal quality is performed periodically for this purpose. However, this method does not reflect the latest service quality and incurs labor cost of manual measurement. This paper introduces a new method to measure radio signal quality automatically by analyzing WIFI access point(AP), whose backhaul is LTE connection, in subway which helps wireless internet connection. This method will help analyze service quality accurately by tracing user equipment(UE) in a cost-effective way.**

***Index Terms*—Wireless Network Quality, UE Analytics, Mobility Analytics**

I. Introduction

In places where massive people move simultaneously with high speed, such as the subway, customers will experience network disconnection or degraded connection quality. Since the subway is underground, it is difficult to maintain service quality. In other words, the subway environment presents a challenge of having to remove signal outage and to handle intermittent bursting traffic that causes a bad service quality.

For a seamless mobile connection service, network operators perform manual measurement of radio signal periodically to monitor service quality. The routine-manual measurement, however, incurs labor cost and is not enough to assure service quality of a specific zone.

In this paper, we introduce a new method of analyzing a service quality of subway through UE analytics. SK Telecom deployed LTE hotspot equipment for all cars of the subway to provide WIFI access via LTE connection. SK Telecom collects real-time call-logs of LTE hotspot including service quality data and analyzes each LTE hotspot. This method has been applied to SK Telecom's Operation Support System(OSS) called *TANGO (T-Advanced Next Generation OSS)* [1] to analyze the radio quality of every subway line real-time without manual measurement.

II. Service Management in Subway: Past and Now

The LTE hotspot located in each car of the subway provides WIFI access to the mobile customers via LTE backhaul connection, which helps UE connect to the LTE base-station. By analyzing UE call-trace data of LTE hotspots, network operators can automatically detect service degradation of the network without manual measurement:

- *Past:* Network operators analyze performance data of base-stations that cover subway zones. The performance data of base-stations is aggregate data from all UE, and there is a limitation to estimate the service quality accurately. In addition, manual measurement is required to collect service quality data of each subway line.
- *Now:* By analyzing call logs of LTE hotspots, we can identify to which base-stations the LTE hotspots are connected and can analyze service quality accurately for stable maintenance of subway service. It is useful in detecting anomalies, such as when base-station designated for subway area is covering other area. They happen frequently for unknown reasons.

III. Expectation and Result

UE analytics reduces routine-manual measurement in subway and provides more accurate network status information. Also, analyzing real time logs of LTE hotspots can be used for optimizing the base-stations that cover subway areas efficiently. Collected moving pattern of each UE in subway can be used as marketing data. Detailed result will be included in the full paper.

References

[1] https://www.telecomasia.net/content/skt-expanding-use-tango-ai-platform, "SKT expanding use of TANGO AI platform"

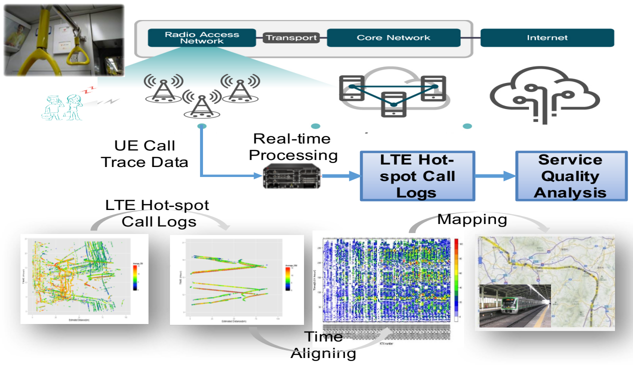


Fig. 1. Data Processing In Analytics

Sensing Population Mobility in Greater Maputo via Mobile Phone Big Data Mining

Mohamed Batran
Data Scientist, Rakuten Inc.
mohamedbatran44@gmail.com

I. Abstract

The rapid growth rate of urban populations may outpace the development of needed urban infrastructure, such as related to transportation, therefore, resulting to inadequacy of public transportation services and traffic congestion. Such as the case of Maputo, the capital and which is considered as an economic and political hub for Mozambique. Accordingly, in order to come up with a solution to this traffic congestion problem and alternative urban planning solutions, initially there is need to acquire a better understanding of the urban dynamics and people flow in the capital.

The mobility patterns and trip behavior of people are usually extracted from data collected by traditional survey methods. However, these methods are generally costly and difficult to implement, especially in developing countries with limited resources. On the other hand, the wide spread of pervasive sensors allowed to capture different levels of human mobility, and to better describe the displacement of people in time and space. Call Details Records (CDR), or mobile phone data, is a good example where timestamp along with the approximate location are recorded with any cellular activity such as a call, a message or a data connection. Such massive amount of data generated by ubiquitous mobile phone usage provide researchers with the opportunity to innovate alternative methods that are inexpensive and easier and faster to implement than traditional methods. However, CDR are spatially and temporally sparse and can only represent a sample of the population who own a mobile phone.

In this work, we leverage hundreds of millions of mobile phone records in Greater Maputo Metropolitan area of Mozambique to infer various aspects of human mobility as an alternative to the expensive traditional methods. This work has three major contribution:

1) We propose a method based on proven techniques to extract the origin-destination (OD) trips from the raw CDR data of mobile phone users and process the data to capture the mobility of those users. The proposed method was applied to 3.4 million mobile phone users over a 12-day period in Mozambique, and the data processed to capture the mobility of people living in the Greater Maputo metropolitan area in different time frames (weekdays and weekends). Subsequently, trip generation maps, attraction maps, and the OD matrix of the study area, which are all practically usable for urban and transportation planning, were generated. Furthermore, spatiotemporal interpolation was applied to all OD trips to reconstruct the population distribution in the study area on an average weekday and weekend. Comparison of the results obtained with actual survey results from the Japan International Cooperation Agency indicate that the proposed method achieves acceptable accuracy.

2) In addition to analyzing inter-city mobility pattern, we propose a method to sense population mobility through city boundary via mobile phone big data. We rely on a validated approach to filter and scale up subscribers and trips for an accurate origin destination (OD) estimation. Furthermore, we slightly modify the algorithm to capture during trip records, en-route CDRs, for more accurate route reconstruction and entry region sensing. The method is applied to Greater Maputo metropolitan area in Mozambique and the output is validated to the most recent traffic count by Japan international cooperation agency (JICA) in 2011.

3) We introduce a method to extract special trip segments from the OD trips generated in (1) that can sufficiently represent and monitor travel time in an urban area from mobile phone data. Our method were able to sufficiently represent and distinguish between travel time in weekday and weekend in Greater Maputo. The result was validated against travel time estimated from an ad-hoc high resolution GPS dataset collected by the study team for that purpose resulting in 87% correlation estimated from both datasets.

This Work was done as part of M.Sc. thesis at The University of Tokyo between Sept. 2016 ~ Aug. 2018 and does not relate to Rakuten Inc.

This work was previously published here:

Batran, M., Mejia, M., Kanasugi, H., Sekimoto, Y., & Shibasaki, R. (2018). Inferencing Human Spatiotemporal Mobility in Greater Maputo via Mobile Phone Big Data Mining. ISPRS International Journal of Geo-Information, 7(7), 259.

Batran, M., Kanasugi, H., Kashiyama, T., Sekimoto, Y., & Shibasaki, R. (2018). Sensing Population Mobility through City Boundary in Greater Maputo via Mobile Phone Big Data Mining. In 2018 IEEE World Congress on Services (SERVICES) (pp. 9–10).

Batran, M., Arai, A., Kanasugi, H., Cumbane, S., Grachane, C., Sekimoto, Y., & Shibasaki, R. (2018). Urban Travel Time Estimation in Greater Maputo Using Mobile Phone Big Data. In 2018 IEEE 20th Conference on Business Informatics (CBI) (Vol. 2, pp. 122–127).

Synthetic Call Detail Records (CDRs) from Authentic CDR Based Subscriber Mobility Profiles

Lasantha Fernando, Viren Dias, Yashothara Shanmugarasa
LIRNE*asia*, 12 Balcombe Place, Colombo 08, Sri Lanka
{lasantha, viren, yashothara}@lirneasia.net

I. Introduction

De-identified CDRs have seen increasing utilization as a valuable secondary data source in studying various aspects of human mobility. However, despite de-identification, a significant portion of subscribers can be re-identified given sufficient matching records[1]. This impairs the shareability of these datasets. As such, novel methods that transform such datasets to safeguard privacy, whilst preserving much of the statistical characteristics of the original dataset, are of significant interest.

Prior research has looked into generating shareable CDR datasets by profiling a region's population as a whole[2]. In this work, we propose an alternative methodology that generates CDRs using the mobility profiles of each individual subscriber, coupled with random noise to safeguard privacy.

II. Methodology

We modeled the subscriber profiles based on de-identified CDR data spanning a single month in 2013 for approximately 6 million subscribers from Sri Lanka. Only location-related attributes were considered since our focus was on eliciting mobility characteristics.

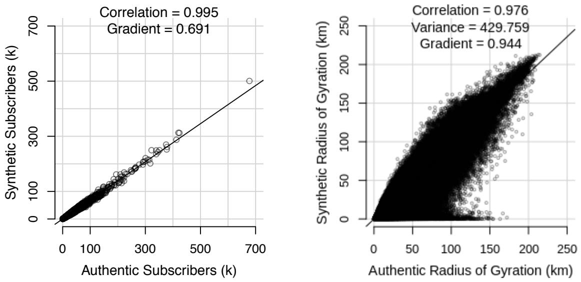
Our first step was to discern homogeneous temporal categories by which to model the profiles. We made the assumption that subscribers' activities were consistent between weekdays, as well as between weekends and national holidays. Within these day classifications, we made the supplementary assumption that subscribers' activities were consistent between corresponding times of the day. Consequently, we segmented each day into octants - 8 contiguous 3-hour segments, resulting in a total of 16 temporal categories. We determined the probability of a record within each temporal category by calculating the portion of total authentic records in that temporal category for each subscriber.

Spatial categories were determined based on the locations of base transceiver stations. Given a temporal category, the probability of a record within each spatial category was determined by calculating the portion of total authentic records for that temporal and spatial category for each subscriber. Finally, the total records for each subscriber were summed up and an error of upto 10% was introduced. This in conjunction with the temporal and spatial probabilities constituted a subscriber's profile.

Given a profile, blank records amounting to the total records were created to initiate the process of generating synthetic records. The records were assigned temporal categories in accordance with the probability distribution, as well as timestamps conforming to that category with equal probability. The records were ordered chronologically, and independently assigned a spatial category in accordance with the probability distribution. The speed between the current and previous records was validated against a predefined constraint, before proceeding to the next. If the validation failed, the spatial category was reassigned and revalidated.

III. Results

The synthetic dataset was subject to rudimentary validations by comparing the number of distinct subscribers aggregated by spatial category with those of the authentic dataset. Despite the high correlation evidenced by Fig. (a), the number of distinct synthetic subscribers was consistently lower by approximately 20%. Almost 3% of the profiles were discarded as they could not generate records that conformed to the established constraints. However, this does not wholly account for the identified discrepancy - we speculate that infrequent, long-distance trips are underproduced as their generation necessitates multiple improbable events in succession, as corroborated by a radius of gyration comparison in Fig. (b).



(a) Distinct subscribers

(b) Radius of gyration

IV. Conclusion

Our methodology generates a synthetic CDR dataset that preserves the core characteristics of the authentic CDR dataset, whilst safeguarding privacy. Extending this work to larger time spans would simply require the addition of a seasonal dimension to the temporal categories. Social characteristics could be preserved with the inclusion of the probability distribution of social contacts in a subscriber's profile.

[1] Yves-Alexandre De Montjoye et al. "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* 3 (2013), p. 1376.

[2] Darakhshan J Mir et al. "Dp-where: Differentially private modeling of human mobility". In: *2013 IEEE international conference on big data*. IEEE. 2013, pp. 580–588.

Confidential sharing of datasets of two mobile network operators: A case study in cross-roaming analysis

Angela Sahk, Reimo Rebane, Jaak Randmets, Dan Bogdanov, Baldur Kubo – Cybernetica [first.last@cyber.ee]
Marko Peterson, Margus Tiru, Siim Esko, Erki Saluveer – Positium [first.last@positium.com]

Introduction

Timely and accurate statistics on cross-border tourism can prove difficult to attain due to various reasons, including privacy and confidentiality barriers, if roaming information from telecom companies is used. Indonesia, led by the vision of the Ministry of Tourism, is one of the first countries in the world to use data from mobile network operators for measuring cross-border tourism activity. Positium has already set up a system for the Ministry based on data from one of the mobile operators. The Ministry wanted to establish a true baseline for roaming market share, which is hard to estimate due to subscribers cross-roaming in the networks of different operators during a single visit. The challenge was how to compare data without sharing it.

Method

A complete answer of the nature of cross-roaming can only be arrived at when unique subscriber information (IMSI) is compared across several operators. Because of privacy reasons, this was a complex task – it requires uniform hashing of IMSIs over at least two operators.

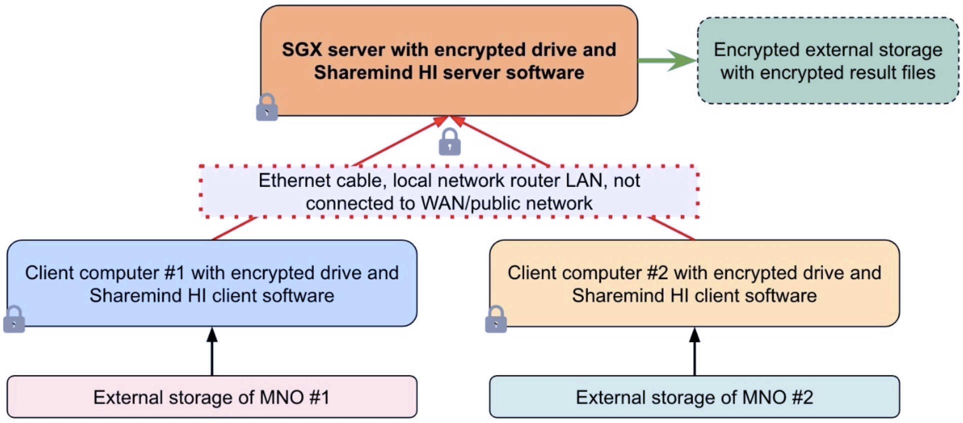
Sharemind is a secure computing platform created to specifically reduce the risk of a privacy breach when processing confidential data. The data is encrypted at the source, by the data owner, and only then sent to the Sharemind service. The host of the service will not have access to the unencrypted data nor the encryption keys. Sharemind traditionally protects data at rest and in transit and surpasses state-of-the-art methods with protecting data in use. It does not remove data protections even while processing, so data will remain protected by cryptographic means during the whole analysis. The Trusted Execution Environment technology used in Sharemind HI to implement privacy-preserving data processing is the Intel® Software Guard Extensions (SGX) available in modern Intel® processors. The three key concepts that SGX provides to protect data are enclaves, attestation and data sealing.

In Indonesia, two mobile network operators prepared a list of IMSIs roaming in certain areas as input. The IMSIs were uniformly hashed from the 7th digit onwards, the file encrypted by operators and uploaded into the Sharemind HI server. The confidential output provided the combinations of 7-digit IMSIs unique for and common between the two operators and correction coefficients to estimates of tourists by country.

Discussion

The project provided the Ministry of Tourism with information on roaming counts and roamer overlap between the two biggest telecom providers in Indonesia, effectively allowing to calculate real roaming market share. Combining this information with other studies provides a more accurate overview of tourism in Indonesia.

The technology is usable in any situation where two sensitive datasets need to be matched against each other.


SGX server with encrypted drive and Sharemind HI server software
Encrypted external storage with encrypted result files
Ethernet cable, local network router LAN, not connected to WAN/public network
Client computer #1 with encrypted drive and Sharemind HI client software
Client computer #2 with encrypted drive and Sharemind HI client software
External storage of MNO #1
External storage of MNO #2

Simple local cloud setup for cross-roaming analysis. Processing takes place in the SGX processor on encrypted input.

Identification of disaster-driven internal displacements from the time series analysis of call detail records

Tracey Li*, Jesper Dejby, Maximilian Albert and Véronique Lefebvre

Flowminder Foundation

*Corresponding author email: tracey.li@flowminder.org

Over 25 million people are forced to leave their homes every year due to environmental disasters. The majority of these people are internally displaced persons (IDPs), meaning that they remain within their home country. IDPs are among the most vulnerable people in the world today, as many States affected by natural disasters are unable to support displaced citizens. New approaches to identifying, understanding, and predicting internal displacements are needed in order to better target assistance and make the humanitarian response more proactive.

We present a method for analysing mobile phone call detail records (CDRs) to identify anonymous individuals that are likely to have been internally displaced as the result of a sudden-onset disaster. In contrast to methods that are typically used to analyse mobility patterns from CDRs, where the regionally aggregated movements of large groups of individuals are analysed, our method offers the advantage that no assumptions regarding the destination, distance, or duration of displacements are necessary. As it enables subscribers that are IDPs to be distinguished from the rest of the population, a detailed study of the mobility behaviour of the IDP subset can then be performed.

We first address the problems of low or irregular spatial and temporal resolution of the data, which are especially relevant in low-income countries. We then detect changes in each individual's 'stay' location over time by computing the distance-to-home time series of each individual, and modelling this as a piecewise-constant signal. We apply a step-detection algorithm to identify change points (level shifts) in the signal and assume that individuals whose stay location changed from their 'normal' location in the days immediately following a disaster are IDPs.

We show the results obtained by the method when analysing datasets pertaining to three natural disasters - Haiti earthquake 2010, Haiti Hurricane Matthew 2016, and Nepal earthquake 2015. The extent to which we observe each location to be affected by displacement is consistent with what would be expected based on the local intensity of the disaster, and matches field observations collected by the International Organization for Migration (IOM). We observe many very short-distance and short-duration displacements, which are typically not included when using conventional CDR analysis methods, in addition to long-distance and long-duration displacements. The inclusion of these movements provides a more complete picture of the scale of a disaster, in terms of the number of people that have been affected and who may require assistance. Mobility information about the individuals who have been displaced, such as the locations of their contacts and frequently visited places, can then be extracted from the CDR data and used to study displacement behaviour. Analyses of disaster-driven displacements will be presented in "Novel mobility and social network metrics to predict disaster-driven displacements from call detail records".

We believe that our IDP identification method can facilitate advances in the analysis and modelling of human mobility in post-disaster scenarios, using CDR and other location data. This will provide crucial information to humanitarian response efforts in contexts where data is often lacking, such as low-income countries. Such information can be used to complement traditional survey methods to assess the scale and characteristics of disaster-induced displacements in a timely manner.

Ncell Digicel IOM·OIM INTERNATIONAL ORGANIZATION FOR MIGRATION THE UNITED NATIONS MIGRATION AGENCY
Funded by European Union Civil Protection and Humanitarian Aid

Explaining urban mobility from urban features and morphology

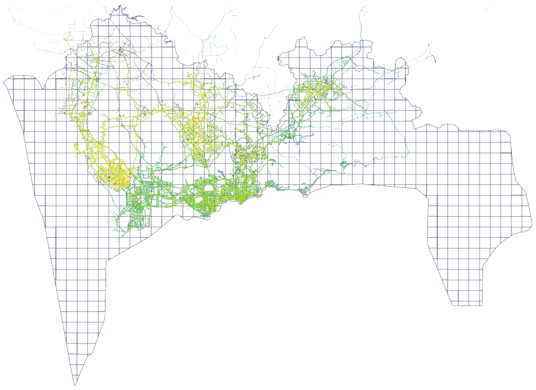
Gevorg Yeghikyan* **Vahan Nanumyan**†
* gevorg.yeghikyan@sns.it, *Scuola Normale Superiore, Pisa, Italy*
† vnanumyan@ethz.ch, *Chair of Systems Design, ETH Zurich, Switzerland*

With the rapidly growing complexity of modern cities, understanding the large scale mechanisms behind intra-urban human mobility and interactions has become of paramount importance for urban planning, management and policy making. The conventional model-driven approaches to human mobility [3, 4] have recently been challenged and augmented by machine learning, specifically deep learning techniques [5, 6, 7]. The latter focus on improving predictive power, but often fail to provide insights into how the observed mobility relates to urban form and socio-economic dynamics. The existing studies on such relationships between urban functions and mobility [8, 9, 10] are narrowly tailored to a specific question, e.g., quantifying flow between certain Point-of-Interest types.

In this work, we explore the dependence of mobility flows on various urban structure features, as well as socio-economic relations between city locations. Our approach is based on principled statistical tools and uses large datasets on fine-grained CDR data and GPS car trajectories spanning several cities.

We follow the complex networks approach by employing a recently developed random graph model [1] and the related multiplex network regression method [2]. We divide the city into a grid of 500m×500m cells. We then represent each cell by a node in the network and the mobility flows between them as directed edges between the nodes. The graph model considers each unit of flow as one unweighted edge. This results in many parallel edges between two nodes.

Then, the random graph model provides a statistical ensemble of all possible flow networks under some constraints: preserving in- and out-flows from each node, as well as respecting heterogeneous pairwise flow propensities of nodes. The multiplex network regression considers these propensities as latent variables, which it infers from observed predictor features. These features include road distances, travel time, demographic and economic statistics of the corresponding urban grid cells (nodes). As opposed to conventional discriminative regression methods, our method intrinsically respects the network constraints. Hence, it allows for stronger interpretation of the modelling results. As with conventional methods, we are able to perform significance tests and model selection, allowing us to judge both the absolute and the relative importance of the urban features for explaining the interactions.



	Exponent	**Std.Err**	p**-value**
time	-1.021	0.0021	<1e-16 ***
speed	-0.716	0.0032	<1e-16 ***
population	-0.252	0.0007	<1e-16 ***
distance	-1.826	0.0023	<1e-16 ***
road distance	-0.222	0.0106	<1e-16 ***
route factor	-0.377	0.0011	<1e-16 ***

Figure 1: (Left) The GPS traces of circa 12,000 taxis over the subdivided area of Shenzhen. (Right) Significant determinants of the mobility flow in Shenzhen.

The figure shows an example result for a limited number of significant predictors for taxi traffic in one city. For instance, we confirm an expected result that the larger the distance, less is the flow between locations. But we also see less trivial results, such as a negative relation between the taxi flow and population density. This is possibly explained by the preference towards public transportation in densely populated areas. In our contribution, we will present the extended analysis for a larger set of mobility predictors, together with a comparison study between multiple cities.

References

[1] Casiraghi, G, Nanumyan, V, Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem. arXiv:1810.06495.

[2] Casiraghi, G, Multiplex Network Regression: How do relations drive interactions? arXiv:1702.02048.

[3] Erlander, S, Stewart, N F, The gravity model in transportation analysis: theory and extensions. Vsp, 1990

[4] Wilson, A G, Urban and regional models in geography and planning. John Wiley & Sons Inc, 1974

[5] Sun, L, Axhausen, K W, Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transportation Research Part B: Methodological, 2016

[6] Jiang, R, Song, X, Fan, Z, Xia, T, Chen, Q, Chen, Q, Miyazawa, S, Shibasaki, R, Deep ROI-Based Modeling for Urban Human Mobility Prediction. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018

[7] Feng, J, Li, Y, Zhang, C, Sun, F, Meng, F, Guo, A, Jin, D, Deepmove: Predicting human mobility with attentional recurrent networks. Proceedings of the 2018 World Wide Web Conference on World Wide Web, 2018

[8] Wang, M, Yang, S, Sun, Y, Gao, J, Human mobility prediction from region functions with taxi trajectories. Public Library of Science, 2017

[9] Zhang, W, Li, S, Pan, G, Mining the semantics of origin-destination flows using taxi traces. UbiComp ’12, 2012

[10] Yao, Z, Fu, Y, Liu, B, Hu, W, Xiong, H, Representing Urban Functions through Zone Embedding with Human Mobility Patterns. IJCAI, 2018

CDR analytics to support the environmental surveillance of infectious diseases

Optimisation of number and placement of surveillance sites

Guilherme A. Zagatti[1], Tracey Li[1], Samantha K. Watson[1], Arend Voorman[2], Vincent Seaman[2], Linus Bengtsson[1] and Véronique Lefebvre[1]*

[1]Flowminder Foundation
[2]*The Bill and Melinda Gates Foundation*

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

Environmental Surveillance (ES) is the collection and testing of sewage samples for the presence of enteric pathogens such as poliovirus. The method allows large numbers of individuals to be regularly tested and has the potential to supplement or replace active clinical surveillance. However, there are uncertainties about the role ES can take in quantifying prevalence and distribution of disease, given unanswered questions about its sensitivity and the interpretation of the results. As ES sites are only sensitive to infections among those defecating in the catchment areas of ES sites, which include both residents and temporary visitors, estimating the sensitivity of a set of ES sites and optimising the placement of future sites partly relies on information on population movements and the number of unique visitors to ES sites over time.

Expansion of environmental surveillance has so far largely been ad-hoc - mostly targeting known high-risk neighbourhoods - and without supporting data on population movements. Because the total number of sites will be limited by laboratory capacity and financial resources, health planners need better data to guide the expansion of ES globally. We developed an analytical framework using Call Detail Records to estimate the fraction of the population covered by ES sites over time, and to optimise the number and placement of new ES sites. We tested our framework in a country with no current cases of targeted diseases (Haiti), where the role of ES sites is to ensure the routine monitoring of the maximum number of people, and not to monitor the spread of an ongoing outbreak.

We first counted the number of unique visitors to each existing and potential site over different time windows (day, week, month, year) to assess the temporal sensitivity of each location, and provide insights on the best times to take samples.
We then took advantage of the movements of people between locations in order to maximise the number of people captured from a given number of locations. This question can be formulated as the classical maximum coverage problem, and we used a greedy algorithm to approximate its solution. We first select the location with the largest number of unique visitors, then select subsequent locations sequentially according to one rule: select the location that was visited by the largest number of visitors that have not visited the previously selected locations (i.e. maximising the number of unique visitors). At each iteration the algorithm provides a set of locations and the maximum fraction of people it can capture, until all possible locations are selected and 100% of the population has been captured. We repeated this optimisation for 3 different time windows (unique visitors to each location in a day, a week, or a month) and found that e.g. 50% of the phone user population can be captured with 45 sites if sampling on 1 day, 12 sites if sampling on 7 days, and 4 sites if sampling every day for a month. We also noted that optimised locations differ from what would be expected from a static population density map, as people's mobility varies across different towns and regions, and because locations with a high number of visitors but low number of residents are selected.

Further work is required in order to make these results operational and considerations pertaining to the fields of hydrology, sewer systems engineering, and chemistry, as well as adjusting for biases present in CDRs need to be included. Then our analytical framework which provides maximum coverage curves of the population as a function of the number of sites, as well as suggestions for site placement, can be used to plan future expansion efforts of the environmental surveillance of infectious diseases.

Quantifying Gender Gaps in Urban Mobility using Mobile Phone Data

Michele Tizzoni[1], Laetitia Gauvin[1], Ciro Cattuto[1], Leo Ferres[2], Simone Piaggesi[1], Andrew Young[3], Natalia Adler[4], and Stefaan Verhulst[3]*

(1) ISI Foundation, Italy, (2) U. Del Desarrollo & Telefónica R&D, Chile (3) NYU The GovLab, New York City, USA (4) UNICEF, New York City, USA

** Corresponding author for this abstract, michele.tizzoni@isi.it*

Mobility is rightfully recognized as a gendered issue. Indeed there is a long tradition of gender and mobility research in the social sciences, urban studies, transportation research and geography concluding that there are differences in the way men and women move about. Nevertheless, most of these studies are mainly based on surveys which might imply relatively few observations over a limited time span, low spatial resolution or can be prone to errors due to self-reporting. More importantly, how the observed differences in mobility can be explained by innate sex-related differences, such as physical differences, or by gendered socially constructed factors, such as household roles, remains a highly debated research issue. Also, long-term trends of gender differences in mobility, driven by global demographic and socioeconomic trends, are hard to capture, and this is particularly true in urban areas, whose population is continuously growing and it is projected to account for almost 70% of the world's population by 2050, according to the United Nations. Indeed, urbanization offers many possibilities to reduce gender gaps through a wealth of new opportunities, but also increases inequalities by, for example, geographical segregation, especially in developing settings. In urban areas, mobility is a critical factor to access the opportunities that cities offer. Thus, investigating the role of gender in urban mobility is key to better understand whether women and young girls can fully benefit from such opportunities and realize their human rights.

In this study, we study urban mobility from a gendered perspective in the greater metropolitan area of Santiago, Chile. Our study has two main objectives: first, to assess and quantify gender disparities in the mobility patterns of Santiago residents in large scale, and, second, to identify socio-demographic factors and the availability of transport options that are associated to such mobility inequalities.

To this aim, we analyze the mobility traces extracted from the CDRs of a large cohort of anonymized mobile phone users disaggregated by sex (male or female) over a period of 3 months. Through estimation statistics, we quantify the effect of gender on different mobility metrics. We find that women visit fewer locations than men and they are more localized, that is, they tend to distribute their time within a few preferred locations. We then map indicators of mobility differences between males and females to 51 municipalities of the Santiago Metropolitan Region and we investigate the association between mobility inequalities and socio-demographic indicators in different areas of the city, as well as their relationship with the Santiago transportation network structure, after controlling for users' activity and gender ratio. Finally, we add a ``semantic layer" to the mobility patterns of Santiago residents by identifying specific points of interests that are more frequently present along women's or men's trajectories in the urban space, thus demonstrating how our approach can identify specific gendered mobility needs.

In conclusion, our study shows how the complex relations between gender, mobility and socio-economic factors can be unveiled by combining telecommunication data with demographic statistics and public transportation data.

Multiview socioeconomic mapping of social networks

Jacobo Levy Abitbol[1], Carlos Sarraute[2], Martin Minnoni[2], Márton Karsai[1]

Univ Lyon, ENS de Lyon, Inria, CNRS, UCBL, F69342, France
Grandata Labs, 550 15th Street, San Francisco, California 94103, USA
Corresponding author: marton.karsai@ens-lyon.fr

Introduction: Despite the pervasive use of mobile devices and the subsequent abundance of data, readily available fine-grained socioeconomic maps are still rare commodity to decision makers, policy experts or researchers. Although these large data streams are regularly being collected and stored, they are generally proprietary. This ensuing data scarcity becomes even more severe when one seeks to blend disparate data sources, usually resulting in aggregated coarse-grained datasets [1] or small-scale fine-grained ones [2]. Earlier socioeconomic status inference methods were based on a variety of data sources, like social media [6, 5], call detailed records (CDRs) [2], or satellite imagery [3, 1]. Although combinations of these datasets [7] have previously been used for similar tasks, large data-driven studies combining high-resolution socioeconomic, social network, and satellite data in large populations are still rare.

In this work we address the inference of the socioeconomic status of a large number of users located in a Latin American country. We build a learning model relying on a large dataset simultaneously recording mobile-phone communications, bank transaction history, and high-resolution satellite imagery of the concerned areas, which enable us to get a fine-grained view of the social structure and living environment of millions of individuals. In doing so, we seek to a) design a model that can coherently aggregate these data sources and learn an interpretable representation of the input and b) detect which high-order correlations yield the most predictive performance of our model in terms of socioeconomic status.

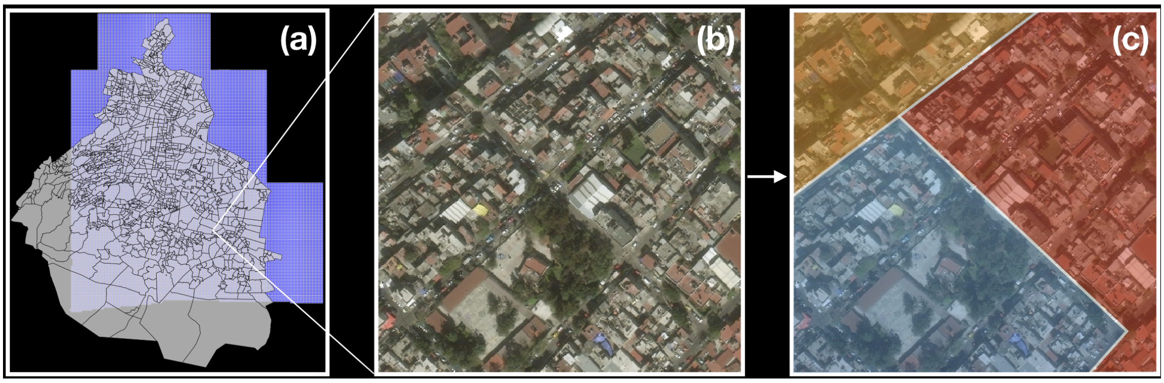


Fig. 1. (a) Available satellite tile coverage of a Latin American city; (b) Display of an individual tile resolution (50cm/pixel); (c) schematic presentation of inferred average socioeconomic status of zip codes.

Data Description: Mobile communication data used in our study records the temporal sequence of 8 billion call and SMS interactions between ~ 112 million anonymised mobile phone customers for a period 21 months. Using the provided CDRs we construct a large social network with users as nodes (whether clients or not of the actual provider), and links drawn between them if they interacted at least once during the observation period. To remove commercial customers, we recursively filter nodes from the network with zero in- our out-degrees. Additionally, due to a combined bank dataset, socioeconomic indicators and demographic information of 6 million bank-mobile customers are also available. The combined bank-mobile dataset contains a single connected component of 1 million people, with communication events and detailed bank records available for all of them. For the purpose of our study, we enrich this dataset with high resolution satellite images collected via the DigitalGlobe Open Data program from 2016-2017. These satellite ties provided a full spatial coverage for several cities within the country.

Results: We build on previously introduced approaches to construct the finest grained socioeconomic inference framework benchmarked on the studied area to date. To do so (*ongoing*), we rely on representation learning tools that enable us to perform a joint embedding of our input where both the satellite tiles and the underlying social network are projected into a lower-dimension space. In doing so, we provide insights on how communication patterns and living environment are intertwined and related to socioeconomic status of people.

References

1. Boris Babenko, et al. NIPS Workshops on Machine Learning for the Developing World, 2017.
2. Joshua Blumenstock, Gabriel Cadamuro, and Robert On. *Science*, 350 (6264):1073-1076, 2015.
3. Ryan Engstrom, Jonathan Samuel Hersh, and David Locke Newhouse. *World Bank*, 2017.
4. Martin Fixman, et al. ASONAM'16 San Francisco, 2016.
5. Jacobo Levy Abitbol, Márton Karsai, and Eric Fleury. ICDMW'18 Singapore, pp. 1192-1199, 2018.
6. Daniel Preotiuc-Pietro, et al. *PLOS ONE*, 10:1–17, 09 2015.
7. Jessica E. Steele, et al. *J. Royal Soc. Interface* 14, 127, 2017.

Understanding Residential Mobility from Long-term Mobile Phone Data

Isabella Loaiza Saa, Samuel Heroy, Esteban Moro, Neave O'Cleary, Alex 'Sandy' Pentland
Human Dynamics Lab, MIT Media Lab, Massachusetts Institute of Technology
Mathematical Institute, University of Oxford

We use nationwide mobile phone data to study the relationship between socioeconomic status (SES) and residential and social mobility in Colombia. Using a 3-years data-set of 15 million lines spanning the entire country, we associate residential moves with permanent changes in users' nighttime locations. We study residential moves in each of the country's five largest cities, validating our results with previous survey-based estimates from Colombia (where approximately 5% of residents move each year).

To quantify residential mobility, we take into account both the proportion of individuals who move in and move out per stratum zone. Additionally (lacking individual-level socioeconomic information), we associate users to a socioeconomic status according to the "socioeconomic stratum" of their pre-move residence. Stratum designations represent governmentally assigned assessments of residential wealth used for income-adjusted pricing of basic services and utilities, taking into account the conditions of a resident's home/street block. For instance, stratum 1 represents very poor conditions, while stratum 6 represents very rich designations, and most residents are assigned 2 or 3. Using this (coarse) designation, we compare the frequency of residential moves for users' of different socioeconomic status (SES), hence we can begin to unpack the link between social mobility (the likelihood to move up or down in SES) and residential mobility. We find that lower SES residents have much hampered residential mobility compared to higher SES residents (in line with survey-based estimates, as explored in Villarage, Sabater, and Módenes, 2014).

In line with previous results, we find that residential mobility in Colombia is low, and there is little difference between cities. We explore the frequency of residential moves between the various strata, showing low social mobility (as measured in this way), as lower (higher) SES residents tend to move among low (high) stratum areas.

While there are numerous studies using call detail records (CDR) to better understand human mobility, our study is novel in its use of CDRs to understand long-term residential moves and social mobility. Because accurate survey-based assessments of residential mobility are relatively infrequent (every 5-10 years), CDRs can provide valuable information in almost real-time to policymakers and other stakeholders regarding the changing makeup of cities and various localities as well as which areas in the city promote low or high social mobility. Given the recent end to a decades long armed conflict, Colombia has in the decade to come unprecedented potential for economic growth and prosperity. We hope to build on this work to be able to parse out what factors influence the rise in social class both for individuals and for whole communities in Colombia.

Comparative Analysis Between Travel Patterns from Cellular Network Data and an Urban Travel Demand Model

Nils Breyer[1] (nils.breyer@liu.se)
David Gundlegård[1] (david.gundlegard@liu.se)
Clas Rydergren[1] (clas.rydergren@liu.se)
Lars Sköld[2] (lars.skold@telenor.se)

[1]Department of Science and Technology, Linköping University
[2]Telenor Sweden

March 15, 2019

Data on travel patterns and travel demand is an important input to traffic planning and forecasting models. Traditionally, travel demand models are based on census data, travel surveys and traffic counts. Problems arise from the fact that the sample sizes are rather limited and that it is expensive to collect and update the data. Cellular network data is seen as a potential large-scale data source to improve the understanding of travel patterns at relatively low cost. In this presentation we will give a better understanding of the potentials and limitations of inferring travel patterns from cellular network data.

The process we propose to infer travel demand from cellular network data consists of a trip extraction step, which identifies trips from the cellular network data and a scaling step, which extends the data to represent the total population (see Figure 1). To find out which types of trips can be extracted from cellular network data, we use a small scale cellular data set collected using 20 mobile phones. In addition to the cellular network data, the phones have collected GPS tracks, which allows a trip-by-trip comparison. Using a large-scale dataset of cellular network data from a Swedish operator for the city of Norrköping, we are able to compare the travel demand inferred from cellular network data to municipality's travel demand model as well as time profiles from public transit tap-ins.

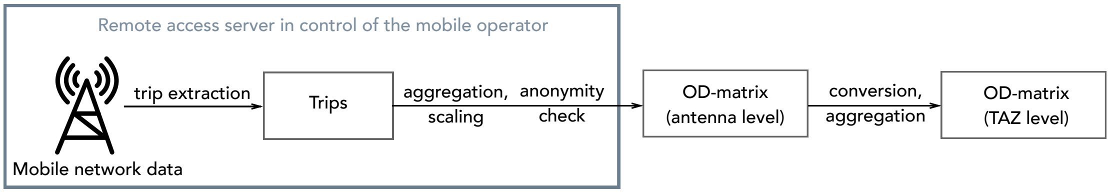


Figure 1: Process to infer travel demand from cellular network data

The results are analysed on several different levels of detail, trip-by-trip and by comparing the results with a traditional travel demand model. Our results show that the recall (trip detection rate) is just about 50% for trips which are only 1-2km long while it is 75-80% for trips of more than 5km length. Similarly, the recall also differs by travel mode with more than 80% for public transit, 74% for car but only 53% for bicycle and walking. After aggregating trips into an origin-destination matrix we find that the correlation to the urban travel demand model used by the municipality depends on the aggregation level of the comparison. While the correlation is weak ($R^2 < 0.2$) using the original zoning used in the traditional model with 189 zones, we find a correlation of $R^2 = 0.82$ when aggregating to 24 zones.

Exploring Temporally Evolving Communities through the Lenses of Location Semantics

Olivera Novović[1,a], Sanja Brdar[1,b], Apostolos N. Papadopoulos[2,c]
[1]BioSense Institute, University of Novi Sad, SERBIA
[2]Department of Informatics, Aristotle University of Thessaloniki, GREECE
[a]novovic@biosense.rs, [b]brdars@uns.ac.rs, [c]papadopo@csd.auth.gr

Introduction

Mobile phone service providers collect large volumes of data related to telecom traffic generated by users. Significant information is recorded in these datasets, and since the data is geo-referenced it has enormous potential for identification of human connectivity patterns in spatio-temporal context. On mobile phone data set covering Milan city [1] we detected communities from connectivity patterns and explored their time evolving characteristics through the lenses of location semantics. In this way we anticipated and explored dynamic change in communities on a city scale.

Location semantic

In urban areas semantics of the location defines its importance to local community and visitors. With increasing number of applications related to location intelligence, there is growing need for data that contain information about spatial semantics of the location. Copernicus Land Monitoring Service provides very detailed land use and land cover data for European urban areas called Urban Atlas. Urban Atlas covering city of Milan with surrounding suburban area contains more than 20 classes defining land use. Along with this data we also extracted points of interest from *OpenStreetMap* database.

Frequent communities

The results of our previous work [2] presented how different areas of the city have different connectivity patterns. Furthermore we performed community detection analysis over telecom data using Apache Spark and Louvain algorithm [3] to identify communities in the city, defined as strongly connected sets of cell ids. We discovered that community structure differs significantly from day to day. Community associated to specific location varies in shape and size. To detect stable patterns in time evolving communities, across 30 days period, we used FP-Growth algorithm for frequent itemset mining. In the next step of the analysis we selected interesting locations that have different spatial semantics. Fig. 1 presents map of Milan city overlaid with frequent communities detected from telecom data around selected points of interest.

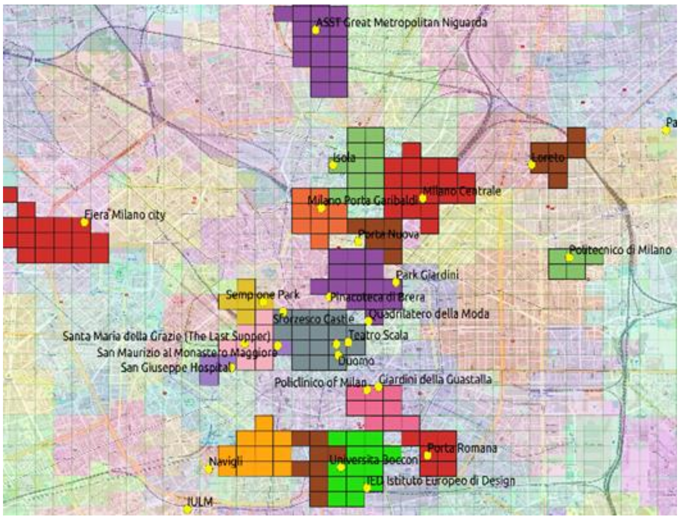


Figure 1. Frequent items generated clusters over selected locations in Milan city

Daily base offset between frequent items area and area of the cluster (i.e. community) is an indicator of cluster stability. As an example, in Fig. 2 we present the dynamic of change in cluster size for two locations with different semantics. The selected areas are *Bocconi University* and *Duomo Cathedral* in the city centre. From Fig. 2 we can notice that *Duomo* has more stable cluster than *Bocconi*. *Bocconi* cluster significantly change in covered area compare to its frequent core.

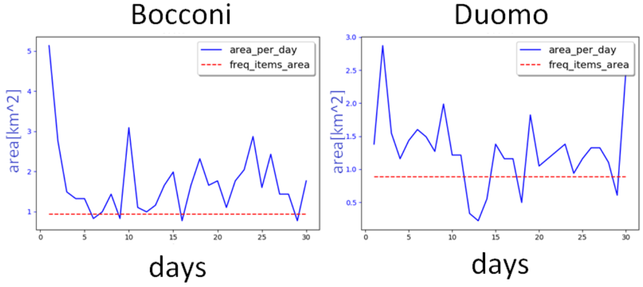


Figure 2. Area of clusters generated daily base compared to area of frequent items cluster

Further systematic analysis of the results across all selected locations provides insights how semantics of the locations impact communities dynamics.

References

[1] Barlacchi, G., et al. (2015). A multi-source dataset of urban life in the city of Milan and the Province of Trentino. Scientific data 2 150055.

[2] Novović, O., Brdar, S., & Crnojević, V. (2017). Evolving connectivity graphs in mobile phone data. In *NetMob, The main conference on the scientific analysis of mobile phone datasets* (pp. 73-75).

[3] Truică, C. O., Novović, O., Brdar, S., & Papadopoulos, A. N. (2018). Community Detection in Who-calls-Whom Social Networks, In *International Conference on Big Data Analytics and Knowledge Discovery* , (pp.19-33), Springer, Chan

Abstract Submission to NetMob19
Conference on the Analysis of Mobile Phone Datasets

Algorithm and architecture for the generation and mobile high-speed transmission of traffic lights information to connected and automated cars in real time

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 use of communications between vehicles and traffic infrastructure, and more recently, between clouds and vehicles, is a field that is growing and full of challenges. One of them is to be able to take the information from traffic ligths to the vehicle in the shortest possible time. Until recently, radio communications based on the 802.11p [1] [2] protocol were the most effective means. These have the disadvantage of being short-range and have a high cost in their deployment.

Taking advantage of the 4G and 5G microlantency networks that are coming, opens the possibilities of using a cloud-based architecture, which complements short-range communications with large ones. The advantage is a cost reduction in the necessary hardware because the software can be installed in small computers or even mobile phones and the possibility to include more context information from sources integrated into the cloud and no more radio communication hardware is required.

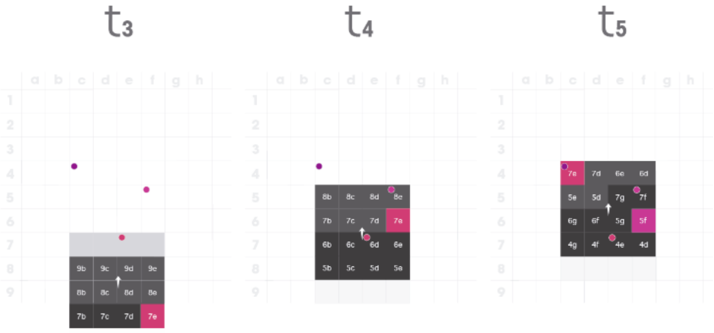
In this work, a methodology has been developed that allows to receive in a device on board a vehicle, information from a set of traffic lights in real time beyond the capacity of vision of vehicle's sensors. This includes the status of each traffic light that affects a vehicle, and the changes that occur in it, in real time. The mobile network has been exploited and a data model and algorithm for the distribution of high-speed notifications have been created. This data model integrate geographical, topological and absolute sequence traffic light information.

The location of the traffic lights is distributed in a grid. In order to optimize communication, an MQTT interface [3] is created into the cloud for a given grid area, taking advantage of the possibility of cascading this interface [4]. Messages from each traffic light are sent to the corresponding topic where the traffic light is located into the grid. The naming mechanism is combined with the identification of the reticule in GeoHash format [5] to achieve a higher calculation speed.

The position reported by each vehicle is integrated in this grid and an information zone of NxN tiles is generated around it. As the vehicle moves, its tiles will include one or more traffic lights. If at that moment there is a topological relationship between the vehicle and the traffic light, being linked by a route of a given length along the streets, the vehicle device subscribes to the MQTT interface of the cloud that includes notifications to it. A direct communication link semaphore-cloud-vehicle is created, based on the spatial distribution of traffic lights and vehicles in real time.

Additionally, the relationship among different traffic light in the network has been implemented on a graph database. This allows to preload the information of the semaphores really related to the one that at that moment is sending information to the vehicle.

Madrid Traffic lights distribution



Vehicle vs traffic light grid

This architecture and transport protocol allows each vehicle to receive the specific information that affects it efficiently and quickly. The data model includes the topology information of the traffic light environment and its status. The protocol conforms to the standard J2735 SPAT and MAP messages [6].

The goal of this study it is the application of protocols and data models to the Madrid traffic light network. The data of 52948 traffic lights from the open data portal of Madrid City Council and the Singularity mobility platform have been used. The tools used have been open source, Python and java languages and PostGIS database. The solution has been implemented into an AWS Amzaon infrastructure.

References

[1] "IEEE 1609 - Family of Standards for Wireless Access in Vehicular Environments (WAVE)". U.S. Department of Transportation. April 13, 2013. Retrieved 2014-11-14.
[2] EN 302 663 Intelligent Transport Systems (ITS); Access layer specification for Intelligent Transport Systems operating in the 5 GHz frequency band
[3] EN 302 663 Intelligent Transport Systems (ITS); Access layer specification for Intelligent Transport Systems operating in the 5 GHz frequency band ISO/IEC 20922:2016(en) Information technology— Message Queuing Telemetry Transport (MQTT) v3.1.1
[4] N. Tantitharanukul, K. Osathanunkul, K. Hantrakul, P. Pramokchon and P. Khoenkaw, "MQTT-Topic Naming Criteria of Open Data for Smart Cities," *2016 International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, 2016, pp. 1-6. doi: 10.1109/ICSEC.2016.7859892
[5] I. S. Suwardi, D. Dharma, D. P. Satya and D. P. Lestari, "Geohash index based spatial data model for corporate," 2015 International Conference on Electrical Engineering and Informatics (ICEEI), Denpasar, 2015, pp. 478-483. doi: 10.1109/ICEEI.2015.7352548
[6] Amsterdam Group, 2015 Signal Phase and Time (SPAT) and Map Data (MAP)

Mobility Analytics for Location Based Advertisement

Jangjun Kim, Hongchan Roh, Jungmin Moon, Jungmin Yoo
SK Telecom, Repulic of Korea
{jjkim81,hongchan.roh,jmmoon,jungminyoo00}@sk.com

***Abstract*—Big data analysis based on mobile communication call-log is essential to manage the flow of large amount of data traffic efficiently, and well-analyzed subscriber-based data is useful in location-based service (LBS). In this paper, we introduce the application method of location-based advertisement (LBA) service with call-logs of non-identified communication subscriber and its advertisement use case. Data is more valuable once it is secured more based on the flow of population. However, in order to generate the data of population flow, pattern analysis of big data collected from the entire communication network should be performed through the correlation analysis between a specific user equipment(UE) and network equipment(NE). High-speed analytic platform, in order to process a large amount of location-based call-log whose data is spatio-temporal, is shown built in a distributed processing platform combined with geometry-specific functions.**

***Index Terms*—Call-log Analytics, Location Based Service/Advertisement**

I. Introduction

As the mobile internet connection becomes common, the daily life of individual exists in the mobile environment. To improve the operational efficiency of the telecommunications industry, it is highly required to manage service quality by customer/service base. An analysis system for real-time tracking and managing the changes of the communication protocol between UE and NE should be built for this purpose. Also, the cause of quality degradation should be analyzed with the system. Recently, LBS services considering the geographical location of mobile terminal devices are expanded. For example, there are messages of push/pull advertisements for point-of-interest (POI), vehicle management, emergency management and asset tracking. For LBS, telco's communication call-logs are most useful in terms of data size and its accuracy.

In this paper, we estimate the location of UE with call-log data, not using location data like GPS tracking, and show its use case in advertising. In order to maximize the effect of indoor/outdoor advertisement shown in the subscriber's path, it is necessary to estimate the movement of people and to characterize the advertising targeted subscribers. For this purpose, we introduce an elaborated location recognition method using individual non-identified UE data. This method works well in a huge network-based environment of the nationwide unit without additional functions for positioning UE, and further advanced tracking will be available through learning patterns of continuously changing signal level between cell-tower and UE. It has been tested based on call-Log generated from SK Telecom's network OSS called TANGO [1].

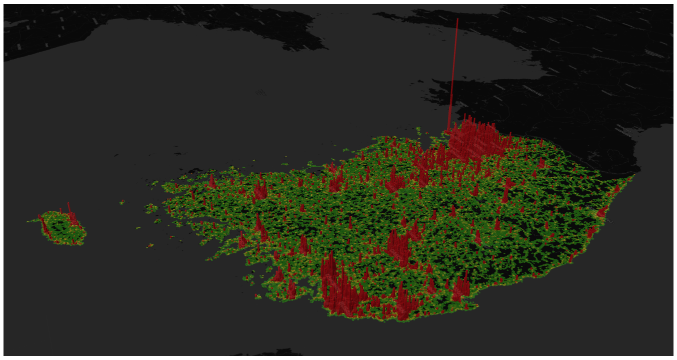


Fig. 1. Location Prediction Example : real-time population density of Incheon

II. End-To-End Call-log based Location Analytics for LBA

The number of advertisements targeted by the effective location requires LBA to be precisely estimated. This can be obtained through analysis of the movement characteristics of the call-log generated by the telco network and the data plane characteristics of the service subscribers. The location can be estimated precisely by performing spatio-temporal analysis on the information transmitted during the periodic or event situation through the interface between UE and the base station. The data plane of the subscriber can finely confirm the target of the advertisement by evaluating the user profile and the user-utilized application information. To do this, we can correlate the access network analysis with core network analysis. We describe the construction of big data analytic platform specialized for high-speed spatial analysis of the correlation and location analysis of access-core network.

III. Expectation and Result

We focused on the possibility of utilizing the nationwide LBA service on a building or a road using access-core end-to-end call-log analyzed by TANGO. Data such as location of cell tower, time-delay between UE and NE, RF-power, and neighbour cells for hand-over are used for this positioning method. We will also discuss how to configure the analytic platform for real-time LBA services.

References

[1] https://www.telecomasia.net/content/skt-expanding-use-tango-ai-platform, ”SKT expanding use of TANGO AI platform”

New Potential Consumers Identification through Telecommunications Network Interaction Analysis[⋆]

Victoria Zevallos[1], Miguel Nunez-del-Prado[1][0000−0001−7997−1739], and Denisse Amanqui[2]

[1] Universidad del Pacífico, Lima - Peru
v.zevallosmunguia@up.edu.pe
m.nunezdelpradoc@up.edu.pe
[2] Entel - Lima, Peru.
denisse.amanqui@entel.pe

Abstract. Competition among telecom operator has radically increased in recent years in Latin America. As a result, operators are making significant investments in developing new strategies allowing them to increase their market share. These strategies have three different approaches. The first is to attract new users to the industry; the second, to attract customers of the competitors; and the third, in retaining the clients. While all fronts are important, the objective of the following study is to attract new clients from other telecom operators. This task presents considerable and interesting challenges due to the lack of information about the subscriber behavior from other telecom operators. Therefore, we rely on the information of the interaction these users maintain with the company's subscribers to determine their future behavior. The main objective is determining which clients belonging to another telecom operator are more likely to become new subscribers based on the analysis performed on the interactions in the telecommunications network.

To attain this objective, we use Call Detail Records (*i.e.,* call traffic and internet consumption) of Small Office Home Office and post-pay Entels subscribers to determine whether a subscriber from another telecom operator will become an Entel subscriber. We model the structure of the mobile social network as a directed graphs G(V,E), where the vertices are weighted by the internet consumption and edges by the volume of incoming and outgoing calls. Then, communities are detected to infer information about the subscribers competitors behavior based on the attributes of the Entel subscribers sharing the same community. Once information is completed, we compute some variables to qualify the changes in subscribers attributes over time. Finally, a classification algorithm is applied to identify the most likely subscribers to migrate to Entel. Our approach achieve an accuracy values around 0.9. However, this classification is not trivial since the dataset is unbalanced. Therefore, we also performed a comparative analysis of resampling techniques to balance the dataset before performing the classification task.

[⋆] Corresponding author: Miguel Nunez-del-Prado. m.nunezdelpradoc@up.edu.pe

Evaluation the wake-up time using mobile phone network data

Gergő Pintér*, Imre Felde*
*Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
pinter.gergo@nik.uni-obuda.hu, felde@uni-obuda.hu

Abstract—**During the last decades, mobile phones have become an integral part of urban life. Mobile devices became an 'extension of human body' and provide ubiquitous possibilities of nearly anytime/anywhere one-to-one communication. Their footprints in the mobile network could help to understand the socioeconomic aspects of urban life. In this paper, we are applying metrics by which mobility and life style characteristics of urban area can be evaluated. Our proposed approach relies on time series analysis of Call Data Records acquisited by a Hungarian Mobile Telecommunication Operator, Vodafone Hungary.**

I. Introduction

The mobile phone data used in this study was provided by Vodafone Hungary, that consists of anonymous Call Detail Records of calls, text messages and data transfer. This data source contains only the so call active records, when a device phone calls, text messages and data transfer, but the type of the record is not specified. Information about cell switching is also not available. As the former would help significantly to identify the wake-up time, the latter would provide much more accurate trajectory of the device. In this paper, we are focusing on two indicators derived from Call Data Records: the Radius of Gyration and the ”wake-up” time.

II. Radius of Gyration

The original radius of gyration [1] is defined as

$$r_g = \sqrt{\frac{1}{N}\sum_{i\in L} n_i(r_i - r_{cm})^2}$$

where L is the list of locations where phone activity is recorded, r_i is the geographical coordinates of the location i, n_i is a weight that can be the visitation frequency or the time spent in a given location. And finally, r_{cm} is the center of mass of the places where the individual visited according to the CDR data.

From the CDR data the ”staying” information is calculated describing how long an individual stays in a given location. As a Call Detail Record represents a momentary information about the location of a device, a 15 minute threshold is applied and it is assumed that the device had stayed in a given location for at least 15 minutes. It has to be noted that when a device is in motion while actively using the mobile phone network, it leaves several traces in different locations in a short time and in this case the threshold does not need to apply as there is newer information about the location of the device. Then the stay information is used to determine the most place where a device present in working hours and out of the working hours to determine the work and the home location of the individual [2]. In this study, the staying time is used as a weight for the n_i parameter and the home location is used instead of the center of mass for the r_cm parameter. The radius of gyration is calculated separately for the weekdays and the holiday in both version (using center of mass and the home cell as the reference point). Using the home location as the center results larger Radius of Gyration because the center of mass is already a mean of the locations. The Pearson correlation coefficient between them is 0.8995 for the weekdays and 0.7454 for the weekends.

III. Wake-up time

As the type of the phone activity is unknown, the wake-up time of the individual is hard to estimate. As making a call requires to be awake (whereas message could be received and data could be transferred autonomously), knowing which activities represent phone calls would provide more accurate information about when people are certainly awake.

The Call Detail Records are counted for every single cell and 15 minute time intervals. This results a time series that is smoothed using the Savitzky-Golay filter with the window size of 31 and the order of polynomial of 5. Using the smoothed curve the minimum and the maximum activity value is selected for every day. The minimum is usually in the middle of the night, and the maximum is in the afternoon. To determine the ”wake-up” time, basically the positive edge of the curve needs to be detected, when the number of the mobile phone activity increases drastically in a short period of time. The time that is considered as the ”wake-up” time is simply where the activity value is the mean of the minimum and the maximum.

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *nature*, vol. 453, no. 7196, p. 779, 2008.

[2] G. Pintér, L. Nadai, G. Bognar, and I. Felde, “Evaluation of mobile phone signals in urban environment during a large social event,” in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, May 2018, pp. 000 247–000 250.

Observing 'devotional togetherness' in Sénégal with mobile phone data

David A. Meyer
Mathematics Department, UC San Diego
dmeyer@math.ucsd.edu

In 2014 the French and Sénégalese telecommunications companies Orange and Sonatel launched a Data for Development challenge, releasing several anonymized sets of mobile phone data from calls within Sénégal during 2013: (1) hourly antenna to antenna calls, (2) call times and antennae for random subsets of 150,000 users in 25 non-overlapping fortnights, and (3) call times and arrondissements for a random subset of 150,000 users for the whole year.[1] Development has as a prerequisite a functional politics. Sénégal is exceptional among West African nations in having a more stable and less repressive political system than others,[2] and since, as O'Brien puts it, "The state in Africa enters the imagination along a religious path.",[3] it is important to understand its relative religious harmony.

Attending the 1963 opening ceremony for the Great Mosque of Touba, the first President of independent Sénégal, Léopold Sédar Senghor, explained:[4]

> *Lundi dernier, le Chef de l'État assistait au Pèlerinage national catholique au Sanctuaire Marial de Popenguine. Aujourd'hui, il est présent à l'inauguration de la grande Mosquée de Touba. Cette double présence n'étonnera que ceux qui persistent à ignorer l'Afrique et ses réalités. Pour nous, Sénégalais, ce sont, là, les fondements de notre politique nationale.*

In this paper we examine this "devotional togetherness"[3] using the mobile phone data from the Orange-Sonatel D4D challenge: Time series analysis of cell tower traffic from dataset (1) reveals variations due to religious rituals, from daily prayers to annual holidays, disaggregated spatially. Comparison of individual call data from datasets (2) and (3) with these baselines indicates participation in, and travel to, these rituals, including, for example, the Grand Magal in Touba.[5] Furthermore, by a novel cross-referencing of individuals between data sets (2) and (3) we can search for those who, like Président Senghor, participate in rituals across religious communities.

[1] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki and V. D. Blondel, "D4D-Senegal: The second mobile phone Data for Development challenge", `arXiv:1407.4885 [cs.CY]`.

[2] L. A. Villalón, *Islamic Society and State Power in Senegal: Disciples and citizens in Fatick* (Cambridge: Cambridge University Press 1995).

[3] D. B. C. O'Brien, *Symbolic Confrontations: Muslims Imagining the State in Africa* (New York: Palgrave 2003).

[4] L. S. Senghor, *Liberté 1 : Négritude et Humanisme* (Paris: Seuil 1964).

[5] C. Coulon, "The Grand Magal in Touba: A religious festival of the Mouride Brotherhood of Senegal", *African Affairs* **98** (1999) 195–210.

Identifying Call Detail Records (CDRs) Affected by Load Sharing

Yashothara Shanmugarasa, Viren Dias, Lasantha Fernando
LIRNE*asia*, 12 Balcombe Place, Colombo 08, Sri Lanka
{yashothara, viren, lasantha}@lirneasia.net

I. Introduction

Pseudonymized CDR datasets are increasingly used as a ubiquitous data source to extract insights into human mobility, transportation, social networks and disaster response behaviors. The utility of these datasets is impaired by the load sharing effect - during peak network traffic, some calls are offloaded to a base transceiver station (BTS) alternative to the closest. Mobile network operators can triangulate locations more accurately using other data sources. However, in the absence of such datasets, the potential implications of load sharing on research must be addressed cautiously.

Existing literature briefly explores this issue; Isaacman et al.[1] handled load sharing by spatially clustering BTSs for each subscriber. In this paper, we discuss several criteria and outline a methodology that can be used to identify records affected by load sharing. We go on to apply this methodology to an existing CDR dataset and also discuss its limitations.

II. Methodology

We used pseudonymized CDR data spanning a single month for approximately 2 million subscribers who had more than 50% of their activity in the Western Province of Sri Lanka, which we selected due to its widespread urbanization.

We clustered BTSs by drawing a grid of 1 km^2 cells across the country based on the methodology described by Maldeniya et al[2].

A trip constituted of consecutive records of a subscriber recorded by different BTSs. We assumed that load sharing happens predominantly between nearby BTSs and is more prevalent than device cloning. Based on these assumptions, we explored different criteria to narrow down possible records affected by load sharing:

- A mean velocity exceeding 110 km/h.
- Between neighboring BTSs.
- A distance less than 5km, 10km, 15km, 20km or 25km

Given the velocity criterion, we conducted analyses based on all possible combinations of the remaining two criteria related to neighbour BTSs and distance.

III. Results

TABLE I
Results of applying different criteria to the dataset

Distance	Only neighbor cells	Pct. of distinct subscribers	Pct. of candidate records
5	False	48.81	0.72
10		51.85	0.89
15		52.32	0.92
20		52.42	0.93
25		52.46	0.94
N/A		52.56	1.00
5	True	43.39	0.53
10		45.26	0.60
15		45.39	0.61
20		45.40	0.61
25		45.40	0.61
N/A		45.41	0.61

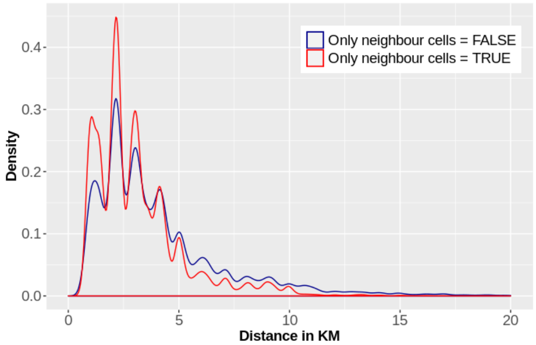


Fig. 1. Density of records affected by load sharing against distance

Density graph in Fig. 1, and Table I suggests that beyond a distance of 10 km, the records affected by load sharing are minimal. This provides a threshold for which the load sharing effect is significant. The marginal increments beyond this distance can be attributed to the presence of cloned devices or infrequent cases of load sharing.

IV. Conclusion

Our method successfully captures a subset of records affected by load sharing. However, some of these records are the result of device cloning. The neighbor BTSs and distance criteria serve to isolate records affected by load sharing from those affected by device cloning, but not wholly. Further experimentation is required to identify a larger portion of

[1] Sibren Isaacman et al. “Identifying important places in people’s lives from cellular network data”. In: *International Conference on Pervasive Computing*. Springer. 2011, pp. 133–151.

[2] Danaja Maldeniya, Sriganesh Lokanathan, and Amal Kumarage. “Origin-Destination Matrix Estimation for Sri Lanka Using 2 . the Four Step Model”. In: *Proceedings of the 13th International Conference on Social Implications of Computers in Developing Countries* May (2015), pp. 785–794.

records affected by load sharing and to segregate them from those affected by device cloning.

DO YOUNG PEOPLE USE THEIR PHONE DIFFERENTLY?

YOUNG ADULT DETECTION THROUGH MOBILE DATA ANALYSIS

Alberto Vázquez (alberto.vazquez@vodafone.com), Rodrigo Buendía (rodrigo.buendia@vodafone.com) and Alberto de Santos (Alberto.santos@vodafone.com).
Vodafone Research, Spain

INTRODUCTION

In a convergent market, it is common practice that several telco products (such as Fiber, Mobile Lines, TV and so forth) are associated to a single paying customer. However, in practice, such products might be used by several people with a diverse range of characteristics (e.g. different members of the same family with varying ages, genders and needs). Unfortunately, the information about the other users of the products and services is missing as the only available information is that of the paying customer. This lack of information is a prevalent and important challenge, which prevents a proper personalization of the customer experience.

A second mobile line is a frequent telco service affected by this ambiguity regarding who is the actual user of the service. A typical use case scenario for this second mobile line are young users whose mobile phone is paid by their parents. Given the different interests and needs of young users versus older adults, we focus on this paper in tackling the problem of automatically differentiating between these two groups through the analysis of their mobile phone behavior. The main hypothesis underlying our work is that young adults exhibit different patterns of mobile usage from older adults and hence may be differentiated.

While several previous works have explored this challenge [1][2][3][4][5], we contribute to the existing literature in three ways: First, we carry out a detailed assessment of the usage of certain social apps, especially those with increasing popularity among young adults, such as Instagram; secondly, we build our models using a large scale dataset of over one million and a half users, much larger than the datasets used in previous works; finally, we study the optimal age threshold to distinguish the behavior between *young* and older adults.

DATA AND METHODOLOGY

We analyze several data sources that capture mobile human behavioral data. All data is fully pseudo-anonymized, provided with user consent and compliant with existing data privacy regulation and our institution’s code of ethics.

In particular, we analyze the data sources below for a random sample of more than a million customers for a period of one month in 2018:

- Volume of data traffic per app and per month, differentiated by hour.
- Use of *Vodafone Pass* special tariffs. These tariffs, hired individually by each user, allows the customer to enjoy unlimited data consumption of certain apps, such as social networks, chats, video, music apps, etc.
- Description of the services purchased by each customer, such as fiber network services, cable TV, prepaid or postpaid mobile services.
- Billing information, including the amount paid by each client per month.
- Geolocation information extracted from Call Detail Records (CDRs).

We apply state-of-the-art supervised machine learning methods in a binary classification task.

In order to train the models, we generate a ground truth dataset composed of the data of over one million and a half customers for whom we have their age and who only have one mobile phone line.

From the available data sources we compute 734 features, as a result of a feature engineering process which includes combining features, aggregating by different time slots, transforming categorical features to numerical features and dealing with null values and outliers.

We split the data into 70% for training/validation and 30% for testing. We then train and test two classification algorithms: Random Forests (RFs) and Gradient Boosted Trees (GBTs).

RESULTS AND DISCUSSION

One of the primary goals of this project is to empirically define the optimal age that would serve as our threshold to define *young* vs *older* adults. Thus, train/test our binary classifiers using a range of ages between 19 and 35, such that users younger than the threshold age would be labelled as *young*.

In all our tests, RFs outperform GBTs. Hence, we only report the age analysis results for RFs. The optimal results have been obtained setting the threshold to distinguish between users older/younger to the age of 21, achieving AUCROC of **0.89,** which is at par with previous work [3][6].

Beyond the classification performance, it is interesting to analyze the most predictive features to shed light on the behavioral patterns that differ between younger and older adults. In our models, the most predictive features are:

- Intensity of usage of social apps, being significantly higher in the younger adult group.
- Volume of data consumed during certain hour slots (especially, at midnight), being also higher in younger adults.
- Usage of special tariffs: related with the use of special Vodafone tariffs (such as the so called *social pass*).

CONCLUSION

In this work we have found empirical evidence of a behavioral difference regarding mobile phone usage between younger and older adults which enables the automatic segmentation of these two groups of customers with high accuracies. We plan to use our models to provide a more personalized and meaningful experience to our customers.

BIBLIOGRAPHY

[1] Nadeem, S. M. S. J. T., & Weigle, M. C. (2012). Demographic prediction of mobile user from phone usage. *Age*, *1*, 16-21.

[2] Sarraute, C., Blanc, P., & Burroni, J. (2014, August). A study of age and gender seen through mobile phone usage patterns in Mexico. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 836-843). IEEE.

[3] Fixman, M., Berenstein, A., Brea, J., Minnoni, M., & Sarraute, C. (2016, November). Inference of Socioeconomic Status in a Communication Graph. In *Simposio Argentino de GRANdes DAtos (AGRANDA 2016)-JAIIO 45 (Tres de Febrero, 2016)*.

[4] Wang, P., Sun, F., Wang, D., Tao, J., Guan, X., & Bifet, A. (2017, April). Inferring Demographics and Social Networks of Mobile Device Users on Campus From AP-Trajectories. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 139-147). International World Wide Web Conferences Steering Committee.

[5] Herrera-Yagüe, C., & Zufiria, P. J. (2012, July). Prediction of telephone user attributes based on network neighborhood information. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 645-659). Springer, Berlin, Heidelberg.

[6] Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., & de Montjoye, Y. A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science*, *6*(1), 3.

Tourism analysis on a large scale using mobile location data

Francesco Tullii, Fabio Pinelli, Francesco Calabrese
Vodafone Research
Email: {Francesco.Tullii, Fabio.Pinelli, Francesco.Calabrese}@vodafone.com

***Abstract*—Mobile phones today represent an important source of information for studying people behaviors, for environmental monitoring, transportation, social networks and tourism. The interest in the use of the data generated by mobile phones is growing quite fast, also thanks to the development and the spread of phones with sophisticated capabilities. In particular, this source of information provided a new and effective way to measure and understand the mobility of the people at different spatial granularity, and it can be used in different scenario by public authorities to improve the quality of service they provide. In this paper, we leverage such kind of data to describe the tourism phenomenon by means of anonymous and aggregated KPIs. In particular we analyzed data of the Italian Sardinia region in August 2018.**

I. Related Work

Despite the presence of several works employing mobile location datasets, they often bring advanced solutions, but working on accurate and selected datasets. The few works that attempt to describe a tourism phenomenon on a large scale in this domain lack of the spatial resolution of the others. For example, [1] describes the spatio-temporal and compositional dimensions of tourists flows in Estonia, but the results are aggregated by counties. Other works present analysis at city level considering a limited temporal window of analysis. In our work we deliver a large-scale analysis on a more refined spatial scale.

II. Data processing

How do tourists travel across the region? What are their profiles? Where do they come from? Do they travel alone or with their families? Do tourists change habits while traveling? The answers to these questions can be used by public tourism authorities to re-design the touristic offers, to improve the quality of services and/or to handle unexpected behaviors. The aim of this work is to build an analytics framework able to leverage mobile network data to provide correct answers to those questions. The proposed framework is designed as an analytics pipeline that ingests mobile network events and extracts mobility indicators (KPIs). These indicators are used, for instance, to describe how and when people move during a tourist visits. In particular, the process pipeline introduced in this work is composed by various steps. Firstly, a cleaning data procedure is performed in order to remove possible erroneous data points on the raw network events with the associated cell id, and to anonymize the data set. On the cleaned data, on a daily basis, dwelling locations (stay locations) for each users are computed using spatial and temporal thresholds. Stay locations represent areas where a user have spent at least a minimum time. Furthermore, a set of anonymous and aggregated KPIs is daily computed on the output of dwelling time process: where people access Sardinia, the number of people visiting an area, how the people move across different area. Moreover, we computed other indicators commonly used in human individual mobility, such as the travel distance distribution and the radius of gyration, to carry out a mobility study. The extracted KPIs are segmented by different profiles based on mobility behaviours. Information related to the roamers are also considered in our analytics framework. Clever combinations of the computed KPIs are finally extracted in order to provide advanced analytics such as frequent movement patters, detect families on trip, and so on.

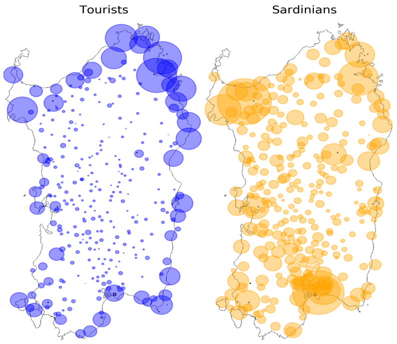


Fig. 1. Tourists vs Local comparison

III. Results

In this section, we describe a series of experiments that we conducted to validate the steps of the framework. In particular, we focus on the results of the dwelling time algorithm, the mobility indicator distributions, and the detected tourist families. In order to validate the results gathered by means of our dwelling time algorithm we carried out two experiments. Firstly, we compared the detected stay locations with the points obtained from another GPS dataset on the same users and the same time window, reaching a level of accuracy of 87.4%. Secondly, we split our set of users, in local and tourists. As an example of a computede KPI, we present in picture 1 a comparison between the two sets in terms of most visited towns, considering the stay locations. The tourists privilege the coasts, in particular the northeastern touristic area. The locals naturally prefer the largest cities. Moreover, we compared the travel distance and radii of gyration distributions between the locals and tourists. We noticed that tourists travel more but constrained in small regions and from a sea town to another, whereas the Sardinians travel less, mainly from and to largest cities.

Furthermore, we performed a comparison between the families extracted among the tourists and the families obtained using mutual calls and text messages. We noticed that we are able to detect a greater number of families which includes the ones obtained with the second method. Moreover, we conducted several experiments to highlight behavioral changes on the smartphone usage while people are travelling. We empirically demonstrate that users on a touristic trip change their usage behaviour w.r.t their typical one. This is true in particular for chat, social networks and maps applications.

References

[1] Janika Raun, Rein Ahas, and Margus Tiru. Measuring tourism destinations using mobile tracking data. *Tourism Management*, 57:202–212, 12 2016.

A General Approach to Detecting Migration Events in Trace Data

Guanghua Chi[1], Fengyang Lin[2], Guangqing Chi[3], Joshua Blumenstock[1]
[1]University of California, Berkeley, [2]Shanghai University of Finance and Economics, [3]Pennsylvania State University
{guanghua, jblumenstock}@berkeley.edu, linfengyang28@gmail.com, gchi@psu.edu

Introduction and Motivation

Migrants play a central role in the global economy. It is estimated that there are 244 million international migrants annually, or roughly 3.3% of the world's population [1]. Internal migrants are even more common, with recent estimates suggesting that as many as 750 million people in the developing world are permanent internal migrants [2]. However, empirical research on migration has historically been hindered by a lack of granular data. Traditional methods — which rely on surveys — are expensive and time-consuming, and are plagued by issues of attrition (since migrants, by definition, do not remain in the same place) [3]. Thus, a growing body of literature has relied on large-scale 'digital trace' data, from mobile phones and online sources, to study the migration of populations [cf. 4, 5].

A critical limitation of prior work is the *ad hoc* approach used to infer migration events from digital trace data. Most existing approaches aggregate all traces in a given time interval (e.g., a month, or a week), assign an individual a location based on the modal or mean location in that interval, and then classify as migrants individuals who are assigned to different locations in adjacent intervals. This approach has two main limitations. First, the modal location, which only focuses on the most common location in a given time interval, does not necessarily represent the home location. For example, we found evidence that the second most common location could have a similar number of appearances to the first one, especially for drivers and commuters who move back and forth between two districts every day. Besides, we also found that people who live close to the border of two districts might receive signals from cell towers in both districts. Second, migration dates are crucial to understand how social networks evolve before and after migration. But migration dates are unknown in prior work because home locations are estimated based on each interval, which can be any day between the two adjacent intervals when the home location changed.

This paper proposes a novel and general approach to detecting migration events in large-scale digital trace data. We develop a segmentation-based algorithm that can flexibly detect both short-term displacement and long-term migration, and isolate the data of migration. We empirically deploy this method on two longitudinal mobile phone datasets (one covering roughly 7 million Afghans over 4 years; the other covering 1.5 million Rwandans over 4.5 years) as well as a randomly selected set of 20,000 Twitter users in the U.S.

Summary of Results

Our approach showed a good performance in identifying migration events in both CDR and geo-tagged tweets. Figure 1 displays the location history, movement segmentation, and the migration date of a random migrant in Rwanda. The detected segments represent home locations in each period of time. Prior modal location based methods will define this person's home as Kigali from 11/15/2008 to 12/01/2008. But our approach can avoid this issue which is usually caused by temporary tourists. Our approach can be applied to other digital trace data sets as long as they contain individual trajectories.

Most importantly, the migration date can be accurately estimated because our method is segmentation based rather than modal location based. Identifying migrants and their migration date would pave the way for many research questions: chain migration, return migration, and dynamic networks, which are impractical to answer using previous methods.

Overview of Methods

Our approach to detecting migrants contains three main steps: detecting movement segments, inferring migrants, and inferring migration date. We borrowed the idea from the DBSCAN algorithm to detect movement segments. In our algorithm, the *radius* is the number of days to search whether a person also appears in the same location. The *minimum number of points per cluster* is the minimum number of days to define a segment. People whose two consecutive segments are located in different locations are migrants. Migration dates are inferred based on the number of error days between home segment and destination segment, which is the number of days when a migrant appears at destination before migration adding the number of days when the migrant appears at home after migration.

[1] G. J. Abel, N. Sander, *Science* **343**, 1520 (2014).

[2] R. E. B. Lucas, *Handbook of the Economics of International Migration* (North-Holland, 2015), vol. 1, pp. 1445–1596.

[3] M. Bell, *et al.*, *Population, Space and Place* **21**, 1 (2015).

[4] J. E. Blumenstock, *Information Technology for Development* **18**, 107 (2012).

[5] E. Zagheni, V. R. K. Garimella, I. Weber, B. State, *WWW '14 Companion* (ACM Press, Seoul, Korea, 2014), pp. 439–444.

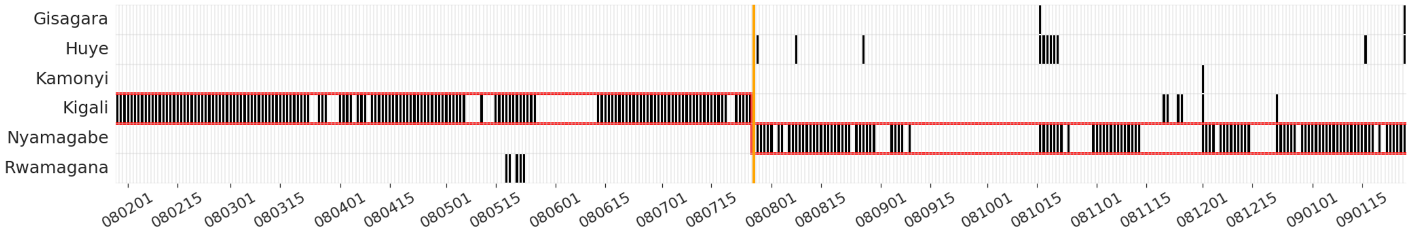


Figure 1: Diagram of a migrant's location history. Note: A black bar represents that this person appears in that district on that day. Red boxes are detected segments. The orange line is the detected migration date. This person migrated from Kigali to Nyamagabe on 7/27/2008.

Preliminary Evaluations on Graph Convolution over an Individual Mobility Network for Estimating Purposes of Stays

Mori Kurokawa,[1] Hao Niu,[1] Kei Yonekawa,[1] Arei Kobayashi[1]
[1]KDDI Research, Inc., Japan
{mo-kurokawa, ha-niu, ke-yonekawa, kobayasi}@kddi-research.jp

Abstract— Call detail records (CDRs) in cellular networks are important data for the inspection of human activities. In this paper, we focus on expressiveness of an individual mobility network (IMN), where nodes and edges represent locations and movements, for estimating purposes of stays. We apply graph convolutional networks (GCN), a recently developed neural network-based graph representation learning, to IMN constructed from CDRs, so as to extract contextual information from IMN. In our experiments, we show the resulting training accuracy to illustrate the potential of IMN and GCN.

Keywords-Call detail records, Individual mobility network, Graph convolution

I. Introduction

Call detail records (CDRs) are an important data source to investigate the diversified mobile phone usage patterns. Offline activities implied by CDRs are beneficial to manage cell towers or provide personalized information. While CDRs have strengths in indoor positioning and saving the batteries of the smartphones as compared with GPS, the difficulty of analyzing CDRs lies in its spatio-temporal sparsity.

In this paper, we try to estimate individual purposes of stays from CDRs. We focus on an individual mobility network (IMN) as a mobility representation. In [1], locations and movements between locations seen in an individual mobility are represented as nodes and edges. Whereas the authors in [1] employed hand-crafted features such as centrality and hubbiness, recently developed graph representation learning [2][3] has achieved to represent structural information as a low-dimensional vector. We exploit semi-supervised graph convolutional networks (GCN) [3] in order to utilize partially available label information of the nodes (purpose of stays).

II. Experiments

A. Data

From Nov. 28 to Dec. 22, 2011, we collected CDRs and activity diary data from 162 examinees with permission. Activity diary data includes ground truth of stay times, locations, and purposes (8 classes; 1: office, 2: school, 3: home, 4: shopping, 5: meal, 6: sightseeing, 7: other private, 8: other business). The mean number of CDRs and registered stay locations was 514.4 and 5.3 per day per examinee, respectively.

B. Methods

For each examinee, we constructed an IMN using CDRs within the above-mentioned period, by placing cell towers as nodes and drawing edges between temporally adjacent nodes. We simplified IMN as an undirected graph with node labels, different from [1] which handles IMN as a directed graph with node and edge attributes. The node labels are defined by comparing access times and locations of the nodes with the ground truth stay times and locations: if the access time is within a stay time (from start to end) and the distance between the node location (i.e. the cell tower) and a stay location is less than 5km, the label of the node is defined as the same as the ground truth purpose of the stay. Finally, we applied a semi-supervised GCN with two hidden layers (16 dim. for each) and one layer perceptron to each IMN.

C. Results and Further Study

The resulting training accuracy was 0.825 in mean (N: 162, min: 0.537, max: 1.0, stdev: 0.097). For inspecting the trained GCN, we selected two examinees (A with relatively high accuracy and B with relatively low accuracy), and illustrated the graph representation in Fig. 1 with different colors of the nodes representing different classes. Fig. 1 shows the hidden layers of the GCN preserve almost the class separability that exists in the IMN (the hidden layer for A seems slightly better).

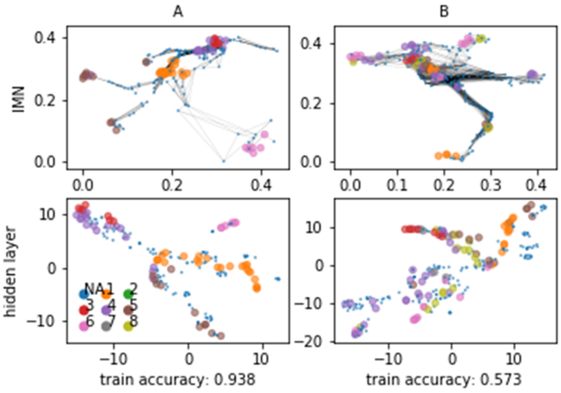


Fig. 1. Graph representation of IMN and 2D visualization (via t-SNE) of the second hidden layer of GCN

For generalizing the GCN trained for each IMN, we have to obtain common representation for all IMNs. In the future, we would like to apply representation alignment techniques such as transfer learning to the GCNs for the generalization.

References

[1] Rinzivillo, Salvatore, et al, "The purpose of motion: Learning activities from individual mobility networks," International Conference on Data Science and Advanced Analytics (DSAA), 2014.

[2] Grover, Aditya, and Jure Leskovec, "node2vec: Scalable feature learning for networks," Proc. of the 22nd ACM SIGKDD, 2016.

[3] Kipf, Thomas N., and Max Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

Study of power-law like distribution in urban scale traffic simulations

Daigo Umemoto[1,2], Nobuyasu Ito[2]

[1) Research Institute for Economics and Business Administration, Kobe University
[2) Center for Computational Science, RIKEN
daigo.umemoto@riken.jp

Power-law like distributions are observed and reported in various complex systems. It was also found in urban scale traffic flow simulations [1]. Because the heavy tail corresponds to congestions, understanding the origin of the behavior is crucial to construct the ideas to design the efficient urban traffic. However, its specific origin is still unknown. Here we show that bottlenecks in networks play key role to produce the heavy tail. Distributions of traffic volumes on road segments (hereafter traffic distribution) robustly obeys power-law like distribution, using digital map of Kobe city in Japan, and randomly generated two dimensional road networks. The traffic demand does not affect the results [1]. Also, As shown in the right panel of Figure 1, traffic distribution obtained from the result of shortest path search without simulation in the network of Kobe city obeys power-law like distribution[3]. Power-law behavior also appears in percolation clusters and Cayley trees. The latter result is shown purely mathematically. These results imply that the origin of the behavior is in road network topology, and specifically, bottlenecks and hieralchy in real networks cause congestions.

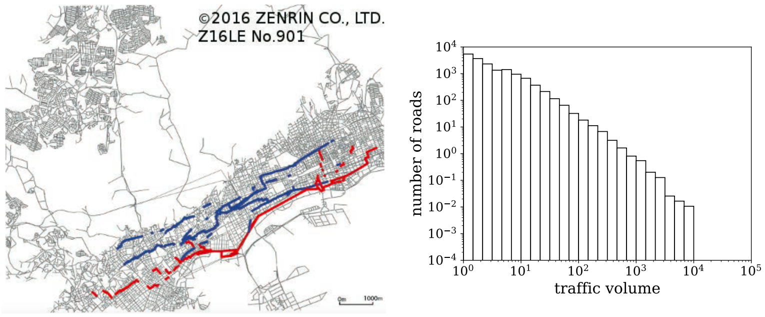


Figure 1: left : Digitized map of Kobe city. Red and blue lines represent most crowded segments, quoted from [2]. right : distribution of road usage on segments, obtained from shortest path searches on randomized OD pairs.

References

[1] Umemoto, D., & Ito, N. (2018). Power-law distribution in an urban traffic flow simulation. *Journal of Computational Social Science*, 1(2), 493-500.

[2] Uchitane, T., & Ito, N. (2016). Applying Factor Analysis to Describe Urban Scale Vehicle Traffic Simulation Results. *Transactions of the Society of Instrument and Control Engineers*, 52, 545-554. (in Japanese)

[3] Umemoto, D., & Ito, N. (2019). Power-law distribution found in city-scale traffic flow simulation. *CSP workshop proceedings* (in prep).

Title

Roaming services in Europe: Estimating the impact of regulatory framework over european consumers surplus

Author

Rafael Rubio-Campillo

Affiliation:

Comisión Nacional de los Mercados y la Competencia (CNMC)

E-mail:

rafael.rubio@cnmc.es

Abstract

Since 2007, intra-EU/EEA roaming services have been addressed through successive EU regulations. The aim of this article is to evaluate the impact this regulation has had on the price level of roaming services and, most importantly, on European consumers' surplus. To reach this objective, this paper uses data on intra-EU/EEA roaming calls from most Member States between 2007 and 2016. We find that the impact of the regulatory framework has critically improved the European consumer surplus over the last 9 years. In addition, the results show that, in spite of the overall improvement in EU countries, there are significant differences among individual countries that can be explained by their own particular characteristics.

Multi-modal Itinerary Reconstruction via Fusion of Cellular Network Signaling, Floating Car and Smart Card Data

Loïc Bonnetain (PhD Student), Angelo Furno (Advisor), Nour-Eddin El Faouzi (Advisor)
Univ Lyon, ENTPE, IFSTTAR, LICIT UMR_T9401, F-69675, Lyon, France
Email: **loic.bonnetain@ifsttar.fr**, angelo.furno@ifsttar.fr, nour-eddin.elfaouzi@ifsttar.fr

I. Introduction

Mobile phone data collected by network operators provide fundamental insights on individual and aggregate human mobility at unprecedented spatio-temporal scales. However, traditional solutions based on Call Detail Records (CDR) are limited due to low accuracy of such data along both the spatial and the temporal dimensions. This reduces their suitability for fine-grained studies on urban mobility.

In our contribution, we focus on mobile phone signaling data and discuss a novel methodology leveraging multi-layer networks and Hidden Markov Model (HMM) to accurately reconstruct multi-modal urban itineraries from such kind of data. In order to accurately grab the dynamics of the transportation system, the parameters of the proposed HMM-based map-matching solution are computed based on network characteristics estimated via fusion of diverse aggregated mobility data sources: floating car data for the road network and smart card data for the public transport network.

II. Methodology

We represent a multi-modal transportation network as a multiplex network composed of four graph layers corresponding to four urban transportation modes: road, bus, tramway and subway. Concerning the cellular network, we propose a method joining traditional Voronoï tessellation to the usage of additional information such as the antenna's directional azimuth, in order to more precisely identify the areas covered by each antenna co-located at the same base station.

The core of the approach is a three-steps map-matching technique. The first step consists in a cleaning phase aiming to reduce noise and oscillation effect in the mobile phone data. Then, by building on best practices from the state of the art [1], [2], we propose an HMM modified solution to compute the k most-likely sequences of nodes of the transportation network from the cellular trajectory (k being an non-zero integer), by using a k-Viterbi algorithm which is a generalization of the original algorithm (corresponding to $k = 1$) [3]. The HMM parameters (initial, transition and emission probabilities) are computed based on the topological properties of the networks and by including dynamic information estimated via fusion of multi-source data (floating car and smart card data). The third step returns the k most-likely itineraries on the multi-layer network with associated probabilities. This step reconstructs complete multi-modal paths (i.e., sequences of adjacent nodes on the multi-layer network) by connecting the possibly disconnected sequences of nodes returned by the HMM step, via (dynamic) shortest path detection.

III. First results

We evaluate our approach in a case-study leveraging real individual anonymized network signalling cellular traces collected by a major french operator in the city of Lyon, France. We propose a validation study at both microscopic and macroscopic levels. At the microscopic level, our approach has been tested using GPS-based traces collected for multiple users via a mobile phone app. Such GPS data, available for a subset of the users of our mobile phone signaling dataset, have been used as ground truth in the evaluation of the reconstructed itineraries. The results show that our approach can properly handle sparse and noisy cell phone trajectories in urban complex environments. Our framework allows associating a likelihood to the different reconstructed alternative itineraries, using the information produced during the k-Viterbi step. Moreover, at the macroscopic level, our approach is compared to the spatial distributions of mobility flows obtained via simulation. The results are promising towards detecting popular paths and performing finer-grained reconstruction of Origin-Destination matrices. Finally, the accuracy of the reconstructed itineraries significantly improves by fusing mobile phone data with floating car and smart card data.

IV. Conclusion and Future Work

The key contributions of our work are the following:

- An unsupervised dynamical HMM-based map-matching approach fusing sparse cellular network trajectories with multi-source data to precisely infer individual itineraries on the urban multi-modal network.
- Instead of inferring a single path, our approach computes k most-likely itineraries on the multi-layer network with associated probabilities to cope with the unavoidable incertitude due to spatio-temporal sparsity of the cellular traces.
- Despite the scarcity of ground truth data, we evaluate our approach in two case studies (i.e., macroscopic and microscopic). In both cases, the potential of mobile phone signaling data for urban mobility analytics is shown.

Future directions should consider the application of the proposed approach on regional scales and in real-time settings towards real-time, large-scale human mobility analytics.

References

[1] F. Asgari, A. Sultan, H. Xiong, V. Gauthier, and M. A. El-Yacoubi, "CT-Mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network," *Computer Communications*, 2016.

[2] E. Algizawy, T. Ogawa, and A. El-Mahdy, "Real-Time Large-Scale Map Matching Using Mobile Phone Data," *ACM Transactions on Knowledge Discovery from Data*, vol. 11, pp. 1–38, 7 2017.

[3] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 4 1967.

Activity-based Land-use Classification with Cellular Data

Peter Widhalm
Austrian Institute of Technology
Giefinggasse 2, A-1210 Vienna
Email: peter.widhalm@ait.ac.at

Norbert Brändle
Austrian Institute of Technology
Giefinggasse 2, A-1210 Vienna
Email: norbert.braendle@ait.ac.at

Klaus Steinnocher
Austrian Institute of Technology
Giefinggasse 6, A-1210 Vienna
Email: klaus.steinnocher@ait.ac.at

The role of massive CDR data as proxy for human movement provides high potential to analyse mobility behaviour for a large number of applications. For example, mobility patterns are a strong indicator for functional land use. Existing work on land use classification can be grouped into unsupervised approaches [1], [2], [3], semi-supervised approaches [4], [5], and supervised techniques [6]. Since the rationale of existing approaches is to observe and analyze aggregate statistics related to *local* network behaviour (e.g. time series of hourly average number of connected mobile phone subscribers or call volumes), they neglect dependencies within sequences of human activities. Here we propose an approach that equates functional land use with the type of activities conducted in a particular region. To extract activities from CDR we first identify stays using the technique described in [7]. Next, we cluster the stays into 18 activity types by training a Hidden Markov Model (HMM) on the daily stay sequences of mobile users, using the start time-of-day and stay duration to describe each activity. We aggregate the extracted activities into spatial raster cells according to a regular 500-by-500 meters grid and compute the total number of stays per day as well as the distribution over the activity clusters to characterize land use.

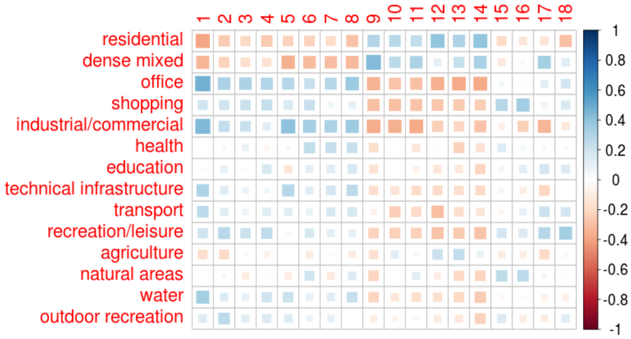
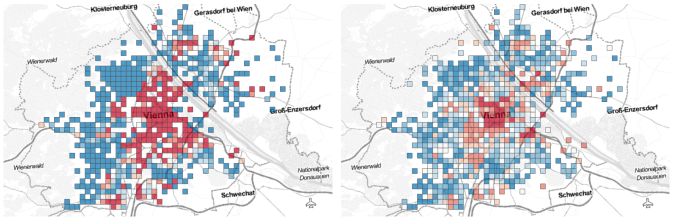


Fig. 1: Correlations between the 18 activity clusters extracted from CDR and zoning labels of Vienna, Austria.

Comparing the aggregated activities to zoning data for Vienna (Austria), we found clear and intuitively comprehensible correlations between activity clusters and land-use types (Fig. 1). However, there is a mix of several different land-use types in most raster cells. Traditional land-use maps describe land-use mix based on fractions of the area covered by each land-use type. In contrast, our approach describes land-use mix as distribution over activity types. Using a Random Forest to learn a regression function between the two representations we obtain results directly comparable to zoning labels (Fig. 2).



(a) zoning labels

(b) CDR

Fig. 2: Residential areas (blue) and office/commercial/industrial land-use (red).

Acknowledgements

This work was funded by the Austrian Federal Ministry for Transport, Innovation and Technology in the Austrian Space Applications Programme, grant No. 865977.

References

[1] K. Madhawa, S. Lokanathan, D. Maldeniya, and R. Samarajiva, "Using mobile network big data for land use classification," in *Communication Policy Research South Conference*, 2015.

[2] J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: analysing cities using the space–time structure of the mobile phone network," *Environment and Planning B: Planning and Design*, vol. 36, no. 5, pp. 824–836, 2009.

[3] G. Engelmann, J. Goulding, and D. Golightly, "Estimating activity-based land-use through unsupervised learning from mobile phone event series in emerging economies," 2017.

[4] T. Pei, S. Sobolevsky, C. Ratti, 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.

[5] V. Soto and E. Frías-Martínez, "Automated land use identification using cell-phone records," in *Proceedings of the 3rd ACM international workshop on MobiArch*. ACM, 2011, pp. 17–22.

[6] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD international workshop on urban computing*. ACM, 2012, pp. 1–8.

[7] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation*, vol. 42, no. 4, pp. 597–623, 2015.

Inference of Mode of Transportation with Mobile Phone Network Data

Eduardo Graells-Garrido
Universidad del Desarrollo
Santiago, Chile

Diego Caro
Universidad del Desarrollo
Santiago, Chile

Denis Parra
Pontificia Universidad Catolica
Santiago, Chile

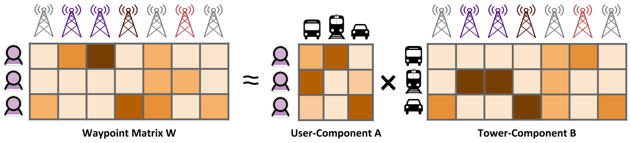


Figure 1: Non-Negative Matrix Factorization Schema.

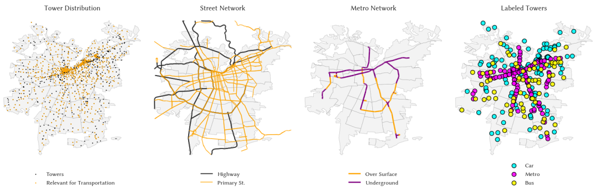


Figure 2: Tower labeling based on Urban Infrastructure.

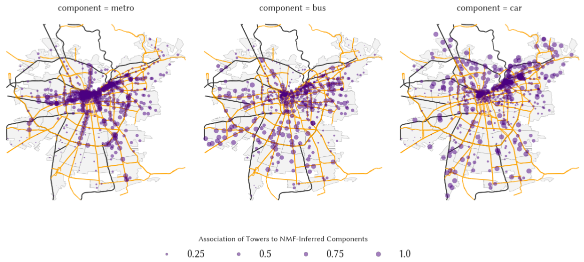


Figure 3: Association between towers and mode of transportation.

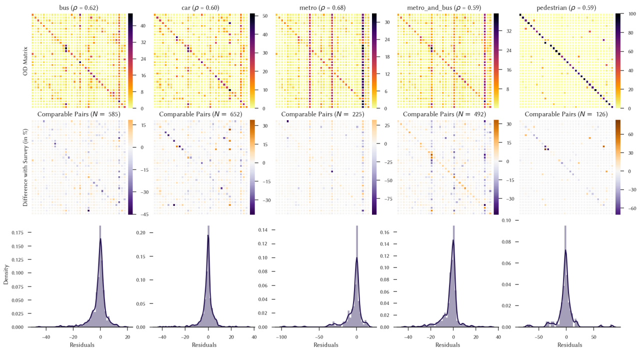


Figure 4: Comparison of O.D. matrices between our method and a travel survey.

ABSTRACT

Traditionally, commuting has been studied with well-known methods such as travel surveys. Even though surveys deliver valuable information, they have drawbacks, including the lack of repeated observations over time and reporting biases and errors. Moreover, surveys (and other traditional methods) are not able to keep the pace of city growth and change, making relevant dynamic phenomena to be invisible for transportation and urban planners.

The availability of large amounts of digital traces has allowed to study urban phenomena at unseen spatio-temporal granularity. One of these data sources is the set of billing records from mobile phone networks, known as *Data Detail Records* (XDR). In this abstract, we overview a recently published work [2] where we sought to answer the following research question: *how to infer the distribution of mode(s) of transportation in commuting within a city using XDR?* An answer would provide insights to manage, plan, and design urban transportation systems, urban infrastructure, and public-policy, among other applications. In this regard, we proposed the following pipeline of analysis to infer mode of transportation usage in the city, with a focus on commuting: 1) identify *trips* and *stays* using computational geometry algorithms over XDR trajectories; 2) infer trip purpose, using known methods [1]; 3) focus on billing records generated *while* commuting, which we represent in a *waypoint matrix*, similar to document-term matrices in Information; 4) decompose the matrix using Topic-Supervised Non-Negative Matrix Factorization [3] (TS-NMF), a semi-supervised matrix factorization algorithm that guides the discovery of latent features toward a desired pattern; and 5) interpret the latent dimensions obtained with TS-NMF and their associations with mode of transportation.

By performing a case study in a big city, Santiago, Chile, we found that the proposed method delivers coherent results, as all modes of transportation under study exhibit similar rank-correlations with a travel survey held in the city (from 0.59 to 0.68, *c.f.* Fig. **??**). Given that the current source of this kind of insight for transportation experts are surveys, that may be outdated, we believe that our work contributes to both disciplines, Data Science and Transportation. By considering our proposed methods and its results, transportation and urban planners will be able to augment their work in a cost-effective way by performing a finer analysis of how people live their cities.

REFERENCES

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies* 58 (2015), 240–250.

[2] Eduardo Graells-Garrido, Diego Caro, and Denis Parra. 2018. Inferring modes of transportation using mobile phone data. *EPJ Data Science* 7, 1 (2018), 49.

[3] Kelsey MacMillan and James D Wilson. 2017. Topic supervised non-negative matrix factorization. *arXiv preprint arXiv:1706.05084* (2017).

CDR mining for emergency situations management

Imed eddine Semassel[a], Marouen Kachroudi[a], Gayo Diallo[b], Sadok Ben Yahia[a]

[a] Faculté des Sciences de Tunis University of Tunis, Tunisa

[b] ERIAS INSERM U897, ISPED, University of Bordeaux, F-33000, France

imededdine.semassel@fst.utm.tn, marouen.kachroudi@fst.rnu.tn, gayo.diallo@u-bordeaux.fr, sadok.benyahia@fst.rnu.tn

Abstract

Anonymous Call Detail Records (CDRs) provide metadata about when and where a person sends or receives a call and/or an SMS message. Using anonymized CDR datasets, researchers has now been enabled to augment spatio-temporal models of both transmitted and non-transmitted diseases. These CDR data can provide information on people's behaviors and relationships with detailed resolution. researchers can track relative population levels in each region of the country, individual movements, seasonal locations, population changes, and migration.

With these datasets, data mining applied to the frequency and timing of calls and/or texts can identify population trends. With these newly, available CDR datasets, population and movement models can now be produced to simulate diseases, which help policies to act with earlier decisions. In this context, data from Orange Senegal Sonatel were used and analyzed to support our proof of concept of our contribution.

Our goal is to analyze this data, to design and develop a recommendation system for the management of critical health cases. The approach analyzes the data by the geolocation of the antennas. Indeed, each resulting model is stratified by a set of complementary data to arrive at two cases.

The first case concerns cardio-vascular attacks and management constraints to save the patient. The second case concerns contagious and communicable diseases.

Our approach is based on simple statistic technique that help to infer information about user's movements considering the differences between day and night, which can show and visualize the time of distances in case of non-transmitted diseases, or the gathering places in case of transmitted diseases.

Our analysis results can be used by the health policies, to defines a vision and plan for future decisions.

Keywords: Call Data Record, CDR, Mobile data, Public Health, emergency situations

Characterizing Transport Perception using Social Media: Differences in Mode and Gender

Paula Vasquez-Henriquez, Eduardo Graells-Garrido, Diego Caro
Data Science Institute, Universidad del Desarrollo, Santiago, Chile.

Transportation plays a key role in people's development in society, greatly affecting the quality of life [1]. Since users base their mobility decisions on factors such as cost, comfort, accessibility, punctuality, quality of service and security [2], transport policies face the growing need to understand their users' perception. However, there is a gap between their perception and the perception of the transportation administrators, since information is collected with an "average" user in mind, usually with little consideration of the opinions of other important user groups, such as women [3].

In this work, we quantify the perception of 300K tweets about transportation in Santiago, Chile. For this, we study the language used by users in their daily experiences about transportation. First, we classify users into the mode(s) of transportation they talk about using a semi-supervised method named Topic-Supervised Non-Negative Matrix Factorization; next, we analyze the lexical associations of the psycho-linguistic lexicon Linguistic Inquiry and Word Count (LIWC) to each mode of transportation and gender. Our work presents the following contributions: a methodology to infer mode of transportation usage from social media content, and a case study of measured differences between modes of transportation, with a gender perspective, using Twitter data from a big city.

This research represents a method of capturing the subjective travel experience in an inexpensive and dynamic way. Our results provide evidence on which aspects of transportation are relevant in the daily experience, working towards defining metrics and indicators of the travel experience as seen from social media. The creation of subjective experience metrics could help to identify pains specific of a segment of users that is otherwise under-represented when transport is designed only with the "average user" in mind.

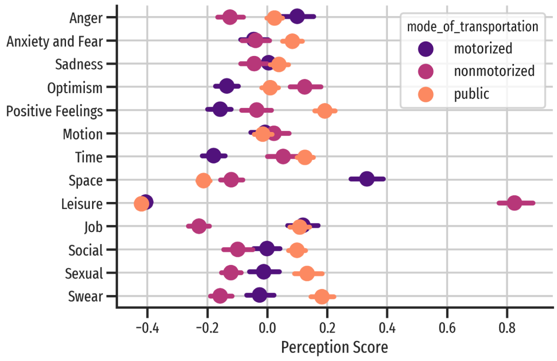


Figure 1: Gross Perception per Mode of Transportation.

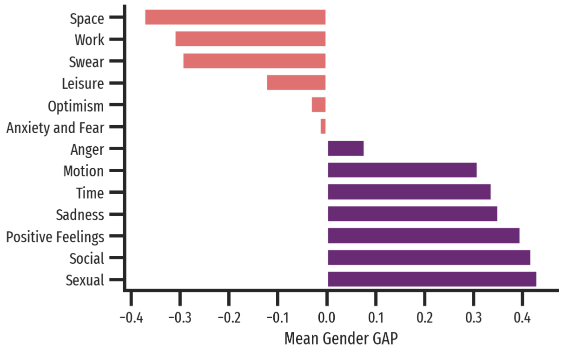


Figure 2: Mean Gender Gaps in Gross-Perception.

References

[1] Alois Stutzer and Bruno S Frey. 2008. Stress that doesn't pay: The commuting paradox. Scandinavian Journal of Economics. 110, 2 (2008), 339–366.
[2] Joe B Hanna and John T Drea. 1998. Understanding and predicting passenger rail travel: an empirical study. Transportation Journal 38, 1 (1998), 38–46.
[3] Daphne Spain. 2000. Run, don't walk: how transportation complicates women's balancing act. In Women's Travel Issues Second National Conference, Drachman Institute of the University of Arizona; Morgan State University; Federal Highway Administration.

Portraits of Swiss Cantons based on Telecom Mobility Insights

Extended abstract

Camelia Elena Ciolac
ciolac_c@yahoo.co.uk

This paper presents a case study of spatial data science applied to open telecom data in the cantons of Switzerland. Combining methods from network science and machine learning with 3D visual analytics, we manage to extract new insights from these multi-faceted datasets.

There exist a multitude of approaches to derive insights into human mobility based on raw telecom data records, many of which were disseminated during the past editions of NetMob Conference. The recent move towards open innovation motivated several telecom operators to open their data in the form of privacy-compliant aggregates related to footfall, OD matrices and network activity at the spatial granularity of official censuses.

This paper benefits from four such datasets [1] and, based on them, we develop several spatial analytics that capture facets of the Swiss cantons' mobility portraits.

The trips between the Swiss cantons provide the basis for building the OD matrix of inter-canton mobility on a daily basis. Using the complex networks science theory, we further detect communities among the cantons based on the flows of people they exchange daily (Fig. 1a). Thus, we build a weighted directed graph whose nodes represent the cantons and the directed edges link the origin to the destination of each mobility flow. The number of travelers on each edge is converted to percentage of the origin's total daily outflow, its median being used as weight in the Louvain method[6].

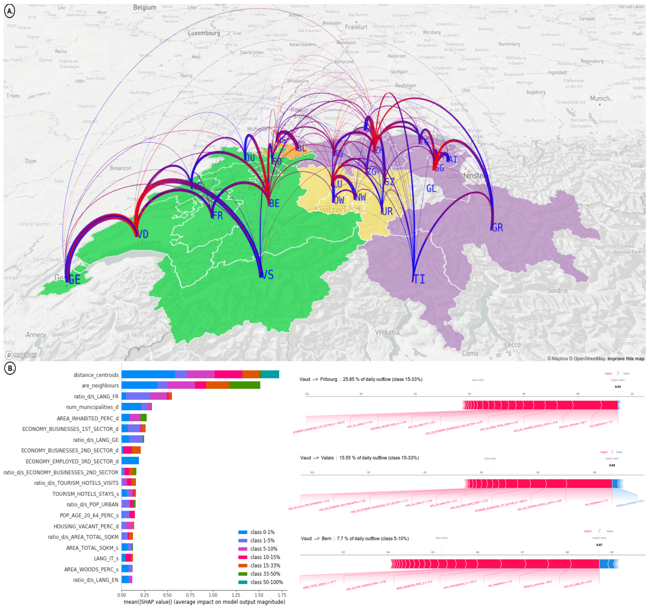


Figure 1: a) OD flows and the corresponding communities of cantons detected for a regular business day; b) XGBoost classifier of inter-canton mobility quanta explained

We train a machine learning model to predict the mobility outflow ratio of a source to other canton based on their geographic and socio-economic attributes[2]. Features' importance and the force plot for individual predictions help interpret our XGBoost classifier(Fig 1b).

Next, we proceed at performing unsupervised machine learning on the CDR derivates (GB of data downloaded, SMS sent and voice calls), which are good proxies for footfall at the canton level.
We compute a spatial clustering among the cantons (Fig. 2), which is a geographically constrained clustering technique aimed at finding similar and geo-spatially contiguous regions in the multivariate timeseries data.

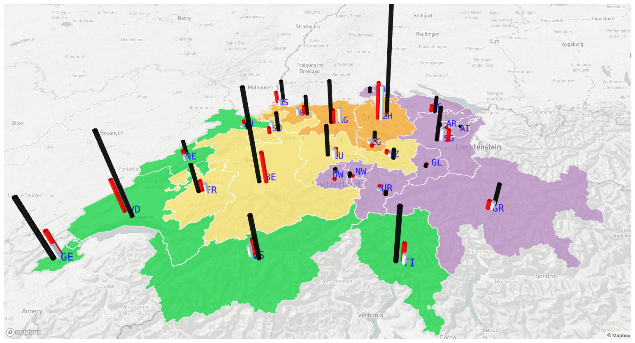


Figure 2: Spatial clustering based on multivariate timeseries of network activity in a regular business day, along with CDR aggregates at canton level during morning time

Thirdly, we investigate the anomalies in the hourly CDR summaries of each canton. For the multivariate outlier detection we use the Isolation Forest, an ensemble model that outperforms the density-based and distance-based methods. The detected anomalies, timestamped (hourly) and geolocated (canton geometry centroid), are then input to Knox space-time interaction tests. Defining the event as the occurrence of such an anomaly, we test whether we can reject the null hypothesis that these events are distributed randomly in space and time and hence conclude that there exist patterns of their spatio-temporal clustering.

References

1. Swisscom Open data, https://opendata.swisscom.com/explore/
2. Federal Statistical Office, Up-to-date regional key figures on the 26 cantons, https://www.bfs.admin.ch/bfsstatic/dam/assets/4662881/master
3. Scikit-learn: Machine Learning in Python, Pedregosa, F. et al,JMLR 12,2825-2830,2011
4. XGBoost: A Scalable Tree Boosting System, Chen, T. and Guestrin, C., 22nd ACM SIGKDD ACM. 785-794, 2016
5. Consistent individualized feature attribution for tree ensembles, Lundberg, S. et al., arXiv:1802.03888
6. Fast unfolding of communities in large networks, Blondel, VD et al., arXiv:0803.0476
7. Exploring network structure, dynamics, and function using NetworkX, Hagberg AA et al., SciPy2008, 2008
8. PySAL: A Python Library of Spatial Analytical Methods, Rey, S.J. and L. Anselin, Review of Regional Studies 37, 5-27, 2007

Special Session
Future Cities Challenge

Disentangling activity-aware human flows reveals the hidden functional organization of urban systems

Riccardo Gallotti[1], Giulia Bertagnolli[1,2], and Manlio De Domenico[1]
[1] *CoMuNe Lab, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy*
[2] *Department of Mathematics, University of Trento, Via Sommarive 14, 38123 Povo (TN), Italy*

Increasing evidence supports the fact that cities are complex systems, with a broad spectrum of structural and dynamical features which lead to unexpected and emerging phenomena. Understanding urban dynamics at individual level – but also as the outcome of collective human behaviour – will open the doors to uncountable applications ranging from enhancing the sustainability and the resilience of the city to improving health and well-being of its inhabitants.

Here, we use a unique data set of longitudinal human flows provided by Foursquare, a leader platform for location intelligence, to characterize the functional organization of a city. First, we build multidimensional network models of human flows corresponding to different types of activities across time. We quantify the efficiency of flow exchange between areas of a city in terms of integration and segregation, respectively. Results reveal unexpected complex spatio-temporal patterns that allow us to gain new insight on the function of 10 megacities worldwide. We discover that large cities tend to be more segregated and less integrated, and that human flows at different hours of the day or between different types of activities enable the identification of different "cities within the city" which indeed show clear dissimilarities in terms of both functional integration and segregation.

Our analysis provides new insights on how human behaviour influences, and is influenced by, the urban environment and, as an interesting byproduct, to characterize functional (dis)similarities of different metropolitan areas, countries, and cultures.

I. INTRODUCTION

Cities are complex systems which process information, evolve and adapt to their environment [1]. To understand how complex systems – and cities more specifically – operate, it is thus important to quantify how information is processed in terms of integration and segregation. To this aim, on the one hand many relevant network descriptors have been introduced, based either on topological features or on dynamical ones, or both. On the other hand, integration has been reflected either in how information flow is accounted for by more complex topological models where multiple relationships co-exist simultaneously [2–5], namely multilayer systems [6, 7], or in causal effects observed in the time course of systems' units [8–17].

Concerning the topological analysis of classical single-layer networks, to date a clear definition of integrated and segregated information flow is still debated and many proxies are used across a broad spectrum of disciplines, ranging from neuroscience to social and urban sciences [18–33], often indicating with the same name very different concepts.

The recent availability of a large amount of human-generated data enables the analysis of urban systems from different perspectives which could not be even considered until a few years ago [34]. In recent times, however, models and analytical tools inspired by complexity science are proliferating. More and more examples are providing convincing evidences of their fruitful application to real cities [35–40]. Applications range from human mobility [41–44] and traffic congestion [45–49], to energy consumption [50], air quality [51, 52] and climate [53], health and wellbeing [54–57], and the associated topic of accessibility to important facilities like hospitals [58]. Indeed, the city can be seen as a growing complex system [59, 60] whose spatial organization [61, 62] dynamically experiences a transition from monocentric to polycentric [63, 64].

A particularly relevant perspective is provided by activity-aware information [65], such as the one provided by users of Foursquare – a leading location intelligence platform – which allows people to investigate human flows at different scales with unprecedented detail [66]. This type of data is of special interest because one can investigate the interplay between the structure of a city and the dynamics of its inhabitants to gain novel insights about the functional organization of the underlying urban ecosystem.

In this work, we stratify human activities in Foursquare to build network models describing the human movements across the urban space – from hours to months – within the different areas (see Methods) of 10 different metropolitan systems worldwide – namely Chicago, Istanbul, Jakarta, London, Los Angeles, New York, Paris, Seoul, Singapore and Tokyo – representing 3 continents.

By classifying existing activities into a few representative macro-categories (see Methods for details) we build a multilayer network [4, 7], where the flows encode how users move between venues of the same macro-category (e.g., from a pub to another one) and between venues of different macro-categories (e.g., from a pub to a cinema). In the following, we will refer to *intra-layer* flow to indicate movements of the first type, and to *inter-layer* flow to indicate movements of the second type.

Our main goal is to better characterize the functional organization of a city through the lens of network science. To this aim we measure to which extent different areas of the city facilitate human flows – i.e., functional inte-

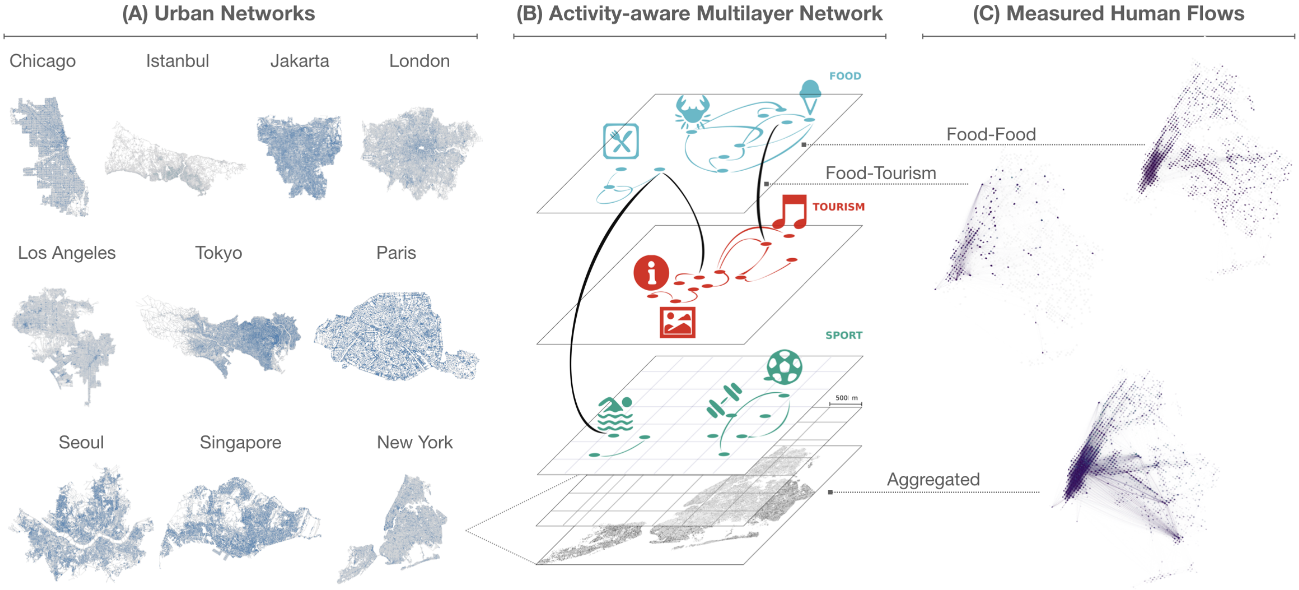


FIG. 1. **Modeling Structure and Function of Urban Systems.** *Left:* Urban structural backbone of the 10 megacities considered here, as described from their street networks (data obtained from Open Street Map [67]). *Middle:* Urban functional networks described by the Foursquare data. The nodes are obtained by dividing the area analysed into cells of 500m × 500m. The edges are subsequent check-ins that might be between activities of the same type (intra-links: e.g. Food-Food, Tourism-Tourism) or different types (inter-links: e.g. Food-Tourism, Food-Sport). The collection of layers and inter-layer flows defines a multilayer network [4, 6, 7], i.e., a multidimensional functional representation of the urban areas. *Right:* The mobility flows between areas are captured as the edges' weights. In the example, describing New York City, we can observe the different spatial distribution of flows between and across different activity layers (see also Fig. 2).

gration – and to which extent there are separate clusters of areas characterized by within-cluster flows larger than between-cluster flows – i.e., functional segregation – (see Methods for details) [68]. By considering those measures simultaneously, it is possible to characterize how well human flows mix through the city according to the existing distribution of venues and the way residents use them. In fact, the dichotomy between integration and segregation – often improperly used as antonyms – is relevant for improving our understanding of the interplay between the urban structure, social relationships and human behavior. To avoid confusion in the reader, it is worth remarking that our measure of integration and segregation is not related to population or cultural mixing [69], but only to how cities are lived by their users.

II. RESULTS

Overview of the data set.— The Foursquare data rendered available for the Future Cities Challenge [70] describe 24 months of check-ins between April 2017 and March 2019 (included). The 10 world mega-cities included in the challenge are Chicago, Istanbul, Jakarta, London, Los Angeles, Tokyo, Paris, Seoul, Singapore and New York City (see Fig. 2 left). The extensive characteristics of the datasets are shown in Tab. I. The flows between different areas are derived by subsequent check-ins to the Foursquare's location-based services and coarse grained with a 500m × 500m granularity (see Fig. 2 middle, and Methods). In the data provided, check-ins are already aggregated by couple of venues (origin and destination), month and hour of the day (morning, midday, afternoon, night, and overnight). The metadata of the venues include a *category* field which describes the type of venue in great detail (e.g.: Knitting Stores, Mini Golf Courses, Rock Clubs, . . .). We defined a set of macro-categories we used to define a limited number of layers (see Methods and Fig. 2 middle). By disentangling the mobility flows into a multilayer network structure, we are able to quantify the differences in the functional organization of the different "cities within a city" that are outlined by movement between different types of activities in a limited number of layers (see Methods and Fig. 2 right).

In Figure 2 we can visually inspect some examples of activity-aware layers. Remarkably, for all the cities considered in this study, the intra-layer connectivity characterizing the transport layer provides a natural link between our functional analysis and the underlying structure of the city. In the data, however, it can be clearly seen in cities where public transport is well developed and largely used, such as Tokyo or Seoul, way more than cities where private transportation is dominant, such as Los Angeles and Istanbul.

Quantifying Functional Integration and Segre-

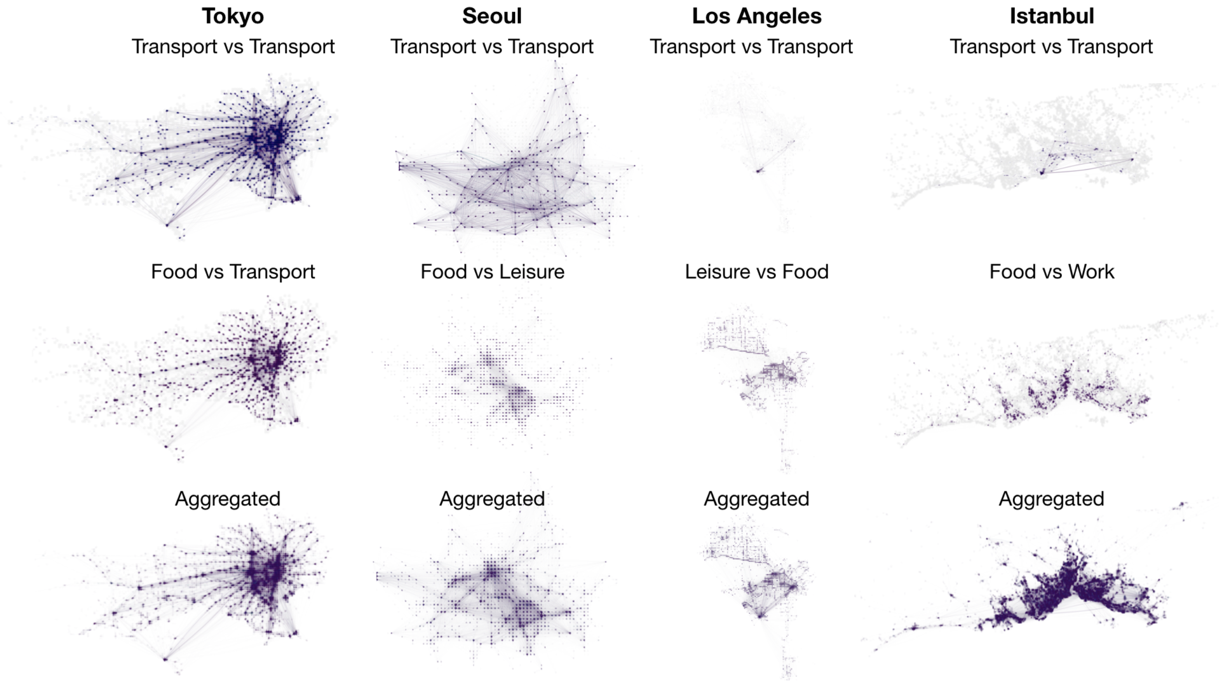


FIG. 2. **Disentangling human flows.** We illustrate here, for four of the ten cities studied in this paper, the strikingly distinct views on the functional organization of a city extracted by isolating intra- or inter-layer flows. These maps outline the different "cities within the city" which we disentangle by decoupling the urban flows into activity-aware multilayer networks.

City	#Venues	#check-ins	L (km)
Chicago	13904	10629110	18.9
Istanbul	113752	13083383	39.3
Jakarta	21813	9281181	27.6
London	22689	10146880	24.0
Los Angeles	15868	10362146	23.6
New York City	32971	11048584	19.3
Paris	13588	9521723	16.6
Seoul	15545	9347489	18.1
Singapore	23324	9691517	23.5
Tokyo	57810	11545155	30.4

TABLE I. **Foursquare data set extensive characteristics.** The figures here are aggregated for all layers, hours, and comprise all 24 months. The linear size L is here estimated as the square root of the total area covered by the data after the aggregation into squares of $500m \times 500m$.

gation.— As previously mentioned, we focus on urban integration and segregation. Integration quantifies, in terms of information exchange efficiency, the ability of a city to favor the flow of people across its areas. Segregation, on the other hand, evaluates the strength of segregated communities, areas of the city with strong flows inside the area and weak inter-areas flows (see Methods for further details).

Synthetic Models.—

Lastly, for the RGGs we also measured the importance of the spatial extension of the network. Fixing the radius below which nodes are connected, we find (see Supplementary Fig.1) that the largest the area ($A = L^2$) covered by a square RGG the more the network is segregated and the less it is, at the same time, integrated. Indeed, here again integration and segregation seem to be very strongly correlated and increasing the radius have a similar effect as reducing the spatial extension.

Functional organization of empirical Urban Systems.— We use the results above as a reference in our analysis of real cities, shown in Fig. 4. More specifically, we stratify human flows by month of the year and by activities, to analyze the corresponding mobility networks. Results are intriguing: the functional organization of some urban systems can dramatically change in different periods of the year (left panel), ranging between small-world and random geometric organization. Comparing with the bottom panel, we notice that the values of segregation for the single months are approximately the same as the aggregated 2 years, and is the value of integration that shows a clear drop. This shows the danger of having under-sampling which could yield to incorrect view on a city functioning by under-estimating its integration. Even if the underlying urban structure changes slowly, or not at all, the functional use of the city, as observed through Foursquare data, may instead display significant variations. When flows are instead aggregated over a long period of time (bottom panel) our view on cities tends to be more compatible with random geometric models, suggesting that the functional organization of large mega-cities is strongly influenced by intermediate-distance movements, captured by random

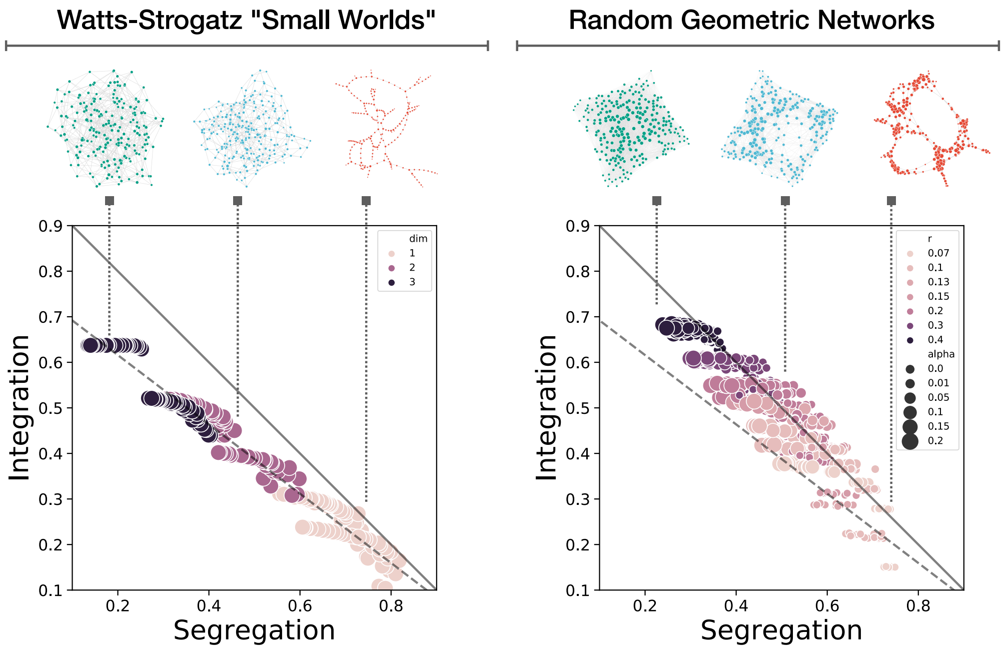


FIG. 3. **Functional organization of synthetic urban models.** *Left:* small-world networks according to the Watts-Strogatz model (see Methods) with different rewiring probabilities (multiple points) and dimensions (from 1D to 3D, encoded by color). The dashed line represents the linear regression relating integration and segregation for this class of models, whereas the solid line is $y = 1 - x$ and it is shown as a reference. *Right:* random geometric networks (see Methods) with different characteristic spatial scale (encoded by color) and different rewiring probabilities (encoded by size). Qualitatively, this class of models clusters along the $y = 1 - x$ line.

geometric models but not by the lattice underlying the small-world model. More in detail, these intermediate distances would here fall beyond the 500m–1000m scale characteristic of lattice first neighbors in the square tiling that we used for describing the urban space (see Methods), but do not represent the long range – across-city – movements, that in our model would be captured by rewiring the existing connectivity.

The fact that some of the cities seem to be even more "integrated and segregated at the same time" than a rewired RGG suggests that the network might be characterized by sparsely connected spatial patches. This result suggests, indeed, that also other complex network models characterized by a mesoscale organization – such as the broad family of stochastic block models – might be suitable candidates for modeling the complexity of megacities.

It is worth checking if this pattern is an intrinsic feature of urban systems or if it is proper of some specific activity layers. To this aim, we perform targeted attacks on each layer of the corresponding multilayer network and measure the response of the systems in terms of changes in segregation and integration. In the bottom panel of Fig. 4 we observe how removing those flows coming from a specific activity type significantly changes urban functional segregation. This is especially true if the activity is Transport, whose removal yields the rightmost outliers in the figure. An even stronger variation would be observed in the integration and segregation restricted to movements between similar layers.

Finally, in Supplementary Figs. 2 and 3 we can observe different "cities within a city" by comparing snapshots of the urban flows at different times of the day or between different types of activities. Indeed, the variability observed for these disaggregated functional networks can be compared with the variability among different cities. Comparing different hours, we find that urban functional integration is systematically higher at daytime and lower in the night, while it is difficult to identify common patterns across cities in its evolution along the day.

When looking at different activities, is even harder to observe common patterns between cities. Segregation and integration of human flows between venues belonging to the same activity layer exhibit high variance with the "food" and "lodging" layers which to be systematically among the more integrated and least segregated. To better understand these differences, in Fig. 5 we link the average values of segregation measured for flows between the same categories across all cities with the corresponding weighted average of geographical distances between nodes. We observe a bulk of correlated points and four clear outliers: two of them represent the typical "second place" of a commute (work or education), one the natural long-range linking layer of transportation, and lastly the locations not associated to a macro category and left as "unknown" (see Methods). Excluding the four outliers,

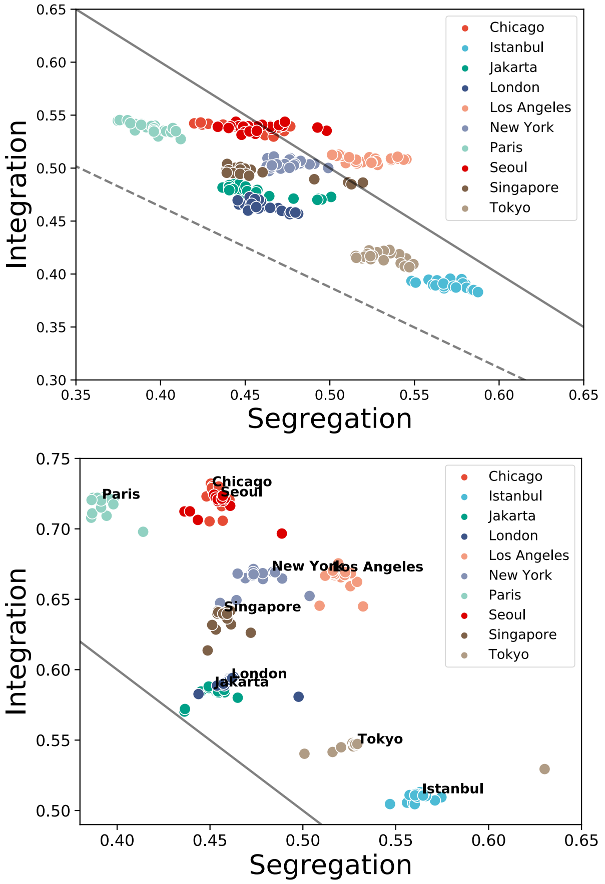


FIG. 4. **Functional (re-)organization of empirical urban systems.** We consider the multilayer networks of human flows for each city (encoded by color). Solid and dashed curves are as in Fig. 3. *Top:* Flows are stratified according to different months (multiple points). Remarkably, some cities exhibit a large variability across time, ranging from a functional organization resembling random geometric networks to one resembling small-world functional organization. Paris, London, Tokyo and Istanbul seem to vary similarly to small worlds, whereas Chicago, Jakarta, Seoul, Singapore, New York and Los Angeles exhibit a broad variability in monthly segregation, associated to higher average values of integration. *Bottom:* Flows are stratified according to different macro-categories used in this work (see Method) and targeted attacks are performed on the layers of each multilayer urban system. Each point corresponds to integration and segregation measured after removing a specific layer of activities, while the first letter of a city name falls in correspondence of the values without any removal. Remarkably, all cities cluster around the solid curve, well describing the behavior observed in random geometric models.

we observe a rather counterintuitive effect: the larger the distance covered by a typical link in that layer, the larger the functional segregation.

This counterintuitive trend – the longer the mobility range the higher the segregation – is also confirmed by comparing the aggregated values for different cities. In Fig. 6 we show how segregation grows and integration drops in cities having larger extension L. We argue that this trend might be a direct effect of geometric constraints, as observed in random geometric networks, leading to an increased amount of segregation when, *ceteris paribus*, we consider networks of larger extension (see Supplementary Fig.1)

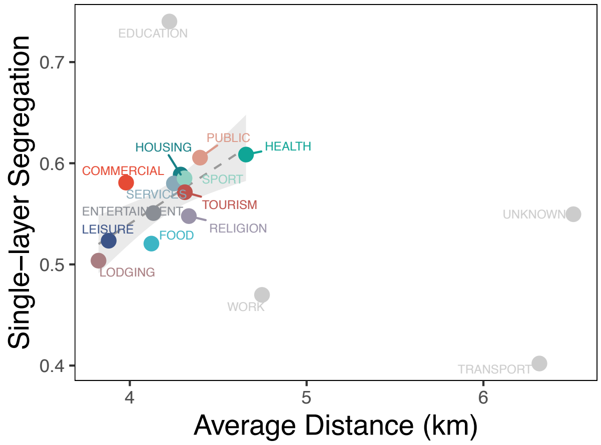


FIG. 5. **Average functional segregation for different activity categories.** The regression, done excluding the 4 clear outliers, highlights a proportionality between the average distance covered in movement inside one layer and the value of segregation. Among the four outlying categories, three are the two "second places" after the home location in individual mobility (education and work), the transport layer, which represents the structural backbone of the city with long range movements and low segregation/high integration (see Supplementary Fig. 4).

A bridge with the urban spatial organization: hotspots' analysis.— Our network analysis of the urban functional organization of megacities leads to a result that, remarkably, create a bridge between network science and quantitative geography. In fact, the organization of cities can be also understood in terms of mobility hotspots [38], i.e., areas characterized by exceptional human flows. The top panel of Fig. 7 shows that the larger is the city the lower is the fraction of nodes which are identified as hotspots using the method introduced in [38]. At the same time, a larger fraction of hotspots is associated to high integration and low segregation (see the bottom panel of the same figure and Supplementary Fig. 5). From a network science perspective, hotspots encode high-degree hubs in the urban functional organization and our results suggest that high-integration and low-segregation features are favored by the abundance of this class of hubs rather than by the presence of a few central nodes.

CONCLUSIONS

Understanding how cities process information, here encoded by human flows, is of paramount importance for designing more efficient and smart urban systems and

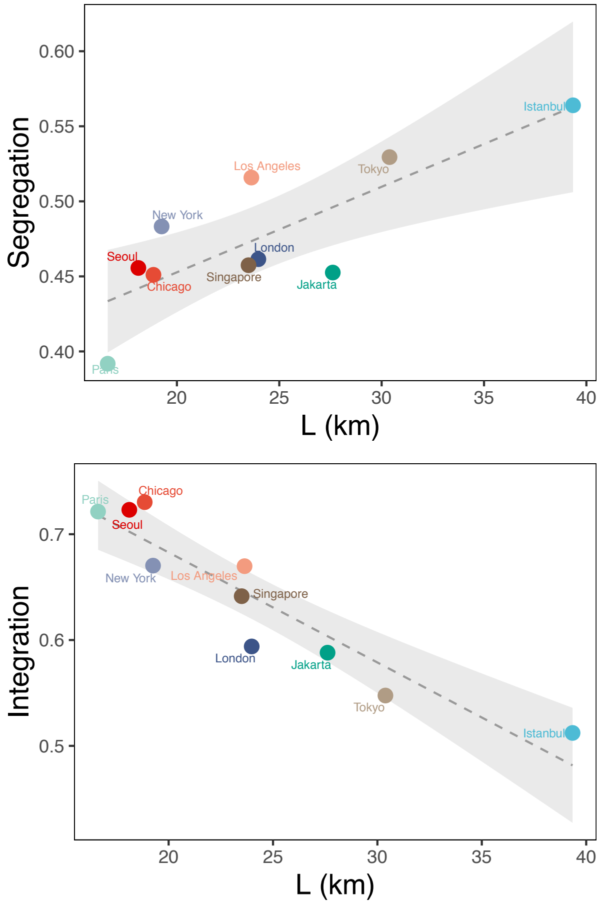


FIG. 6. **Functional segregation and integration *versus* city size.** For each city, we measure $A = Na$, where N is the number of areal units in the city and $a = 500m \times 500m$ their surface. As suggested in [71], the linear extension $L = \sqrt{A}$, represented here, is therefore proportional to the typical commuting distance in a city. *Top:* Functional segregation grows with a city extension. *Bottom:* Functional integration drops with a city extension (Pearson correlation coefficient of -0.93).

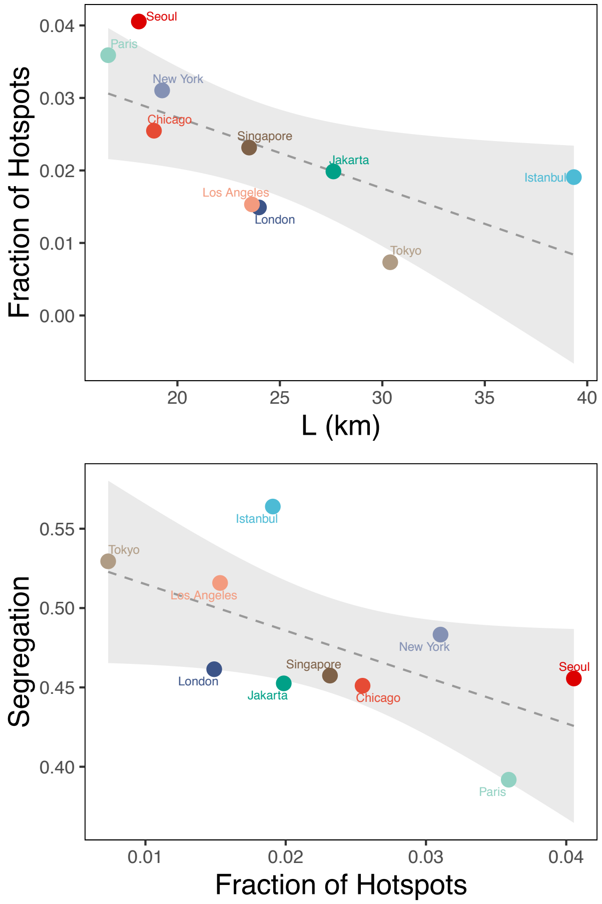


FIG. 7. **City size, fraction of hotspots, and functional segregation.** *Top:* As the size L of the city grow, the fraction of area that is represented by hotspots obtained with the LouBar method [38] decreases (Pearson correlation coefficient of -0.65). *Bottom:* Having in proportion a larger fraction of the urban area covered by hotspots makes the cities less segregated (Pearson correlation coefficient of -0.61), and at the same time the more integrated (see Fig.5).

communities. By considering together multiple activity-aware mobility networks, a functional proxy for the underlying backbone of a city, we have characterized 10 large-scale urban systems in terms of their integrated and segregated flows, revealing interesting spatio-temporal patterns that would otherwise remain hidden without multilayer modeling and analysis.

More specifically, we have shown how network-based analysis can support, and further expand, ongoing discussions about and novel understanding provided by the ICT-data driven quantitative urbanism [38]. For growing cities, it is expected a transition from a monocentric to a polycentric organization, characterized by a sub-linear growth of the number of hotspots with population [63]. Similarly, in the urban functional networks we have extracted from the Foursquare data we observe that in the larger cities hotspots cover, in proportion, a smaller fraction of the total area. However, our analysis of urban segregation and integration suggests a rather different description than that one suggested in [38]: we provide evidence that large polycentric cities ara characterized by a smaller fraction of hotspots (or hubs) and consequently they appear to be more segregated and less integrated than smaller, and monocentric, cities.

From a more methodological perspective, our analysis highlights the importance of data sources for the analysis of the interplay between the city and its main users, i.e., the citizens. Thanks to the unique data set provided by Foursquare we have been able, on the one hand, to quantify the effects of incomplete information: missing human flows related to one activity type dramatically alters the estimated urban segregation. On the other hand, the analysis of longitudinal data at different temporal resolutions, from hours to months, reveals that some cities dynamically change their functional organization even if their backbone does not, whereas other cities do not show the same variability. This allows us to gain novel insights on urban and human behavior in terms of dynamical reorganization of the system. Our analysis of attacks targeted towards specific layers unraveled the importance of different types of human flows for integrating (or segregating) a urban system, revealing the emergence of the phenomenon of *cities within a city*.

Lastly, from a modeling perspective, we discover that many features of complex megacities can be understood from simple mechanisms related to geometric constraints and city's characteristic size, with larger cities tending to be more segregated and less integrated. Random geometric models with long-range connections seems to be a good candidate to reproduce the most salient features measured from empirical data and further research is required in this direction to confirm this finding for a wider spectrum of urban systems. Interestingly, the interplay between heterogeneities in the underlying network connectivity and spatial constraints might be responsible for the emergence of integrated/segregated structures that might be reflected in the functional organization of the city, and future research should point in this direction to gain new insights.

III. METHODS

Geographic coarse-graining.— We reconstruct the flows network by aggregating data over areal units of 500m × 500m, in all 10 cities considered. Flows are reconstructed from subsequent check-ins into Foursquare venues, ignoring the order (undirected network). Flows inside the same area have been integrated into a self-loop link only if the check-ins were between two different locations. Subsequent check-ins in the same location have been excluded from the analysis.

Temporal stratification.— We decouple the functional use of a city i) at different hours of the day (morning, midday, afternoon, night, overnight), and ii) in different months of the year.

Activity stratification.— We use Foursquare's rich system of categories and manually associate them to a reduced number of macro-categories (food, lodging, tourism, work, religion, services, education, health, sport, transport, entertainment, leisure, public, housing and commercial). We do not use [72], except for venue icons in Fig. 1. The few categories that did not fit any macro-category have been labelled as 'unknown'. These categories allow us to build "activity-aware multilayer networks", where activities of different types are associated to different layers of our model. Flows between activities of the same macro-category are encoded by intra-layer links, while flows between different categories are encoded by inter-layer links.

Measuring functional integration.— We measure to which extent a network is integrated in terms of communication, i.e. how efficient nodes are in exchanging information. Given two areal units i and j we can reasonably assume that the communication efficiency between them is inversely proportional to their topological distance d_{ij}; in terms of connections, if information has to travel a long path, the probability that the message is corrupted along the way is high and the communication is inefficient. Globally, the communication efficiency [18] of a city is given by the sum of the pairwise efficiency, adequately normalized:

$$E = \frac{1}{N}\sum_{i\in V}\frac{\sum_{j\in V, i\neq j} d_{ij}^{-1}}{N-1}. \tag{1}$$

Please note that here we did not use distances weighted by the observed flows, to allow for a fair comparison against the values of E obtained from the analysis of synthetic networks, which are not weighted.

Measuring functional segregation.— A usual measure of network segregation, or how strongly the units are organized in into M non-overlapping blocks, is the modularity [73]

$$Q = \sum_{u\in M}\left[e_{uu} - \left(\sum_{v\in M} e_{uv}\right)^2\right] \tag{2}$$

where e_{uu} is the proportion of links inside module u, while e_{uv} accounts for the connectivity between two distinct modules u and v. More in detail, our measure of segregation is the maximum value of the modularity that we find using the Louvain algorithm [74]. We also verify that the observed modularity is significant, by comparison with the values of Q computed over an ensemble of configuration models obtained reshuffling the network. Finally, note that here, instead, we used the weights defined by flows. Values of Q for weighted and unweighted networks are indeed comparable, as opposite to what discussed above for E, and using weights here allowed us to better discern the characteristics of different layers.

Synthetic network models.— We use two standard spatial network models for our analysis.

We first consider a class of networks characterized by small average geodesic distance: the Watts-Strogatz (WS) model. Starting from a regular graph, e.g. a two-dimensional lattice, each link has a probability p of being *rewired*, that is removed and re-placed randomly in the network. If p is large the resulting WS network will look more like an ER random graph than the original lattice. WS networks are also highly clustered, where nodes tend to form closed triangles. WS model are usually referred to as *small-world* networks.

Alternatively to WS, we study also the simplest network model actively involving the spatial dimension model is the random geometric graph (RGG), where nodes randomly distributed in space are connected if they are closer than a fixed threshold distance. The RGGs share many important properties with regular lattices, in particular they are not "small world". For this reason, similarly to the WS case, here also for the RGG we perform a rewiring with probability α.

ACKNOWLEDGMENTS

The authors thank Foursquare for granting access to the data set used in this study and acknowledge Matthew Kamen, Renaud Lambiotte, Jesse Lane, Anastasios Noulas, Cecilia Mascolo, Vsevolod Salnikov, Sarah Spagnolo and Adam Walksman for organizing the Future Cities Challenge. Tha authors asknowledge Giuseppe Lupo for fruitful discussions.

[1] M. Barthelemy, Nature Reviews Physics p. 1 (2019).

[2] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, science **328**, 876 (2010).

[3] M. Szell, R. Lambiotte, and S. Thurner, Proceedings of the National Academy of Sciences **107**, 13636 (2010).

[4] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, Physical Review X **3**, 041022 (2013).

[5] M. De Domenico, Physics of Life Reviews **24**, 149 (2018).

[6] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Journal of complex networks **2**, 203 (2014).

[7] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, Nature Physics **12**, 901 (2016).

[8] T. Schreiber, Physical review letters **85**, 461 (2000).

[9] L. Barnett, A. B. Barrett, and A. K. Seth, Physical review letters **103**, 238701 (2009).

[10] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, Physical review letters **108**, 258701 (2012).

[11] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, science **338**, 496 (2012).

[12] S. Stramaglia, J. M. Cortes, and D. Marinazzo, New Journal of Physics **16**, 105003 (2014).

[13] E. H. Van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara, Nature Climate Change **5**, 445 (2015).

[14] I. Diez, A. Erramuzpe, I. Escudero, B. Mateos, A. Cabrera, D. Marinazzo, E. J. Sanz-Arigita, S. Stramaglia, J. M. Cortes Diaz, and A. D. N. Initiative, Brain connectivity **5**, 554 (2015).

[15] G. Tononi, M. Boly, M. Massimini, and C. Koch, Nature Reviews Neuroscience **17**, 450 (2016).

[16] R. G. James, N. Barnett, and J. P. Crutchfield, Physical review letters **116**, 238701 (2016).

[17] H. Ye and G. Sugihara, Science **353**, 922 (2016).

[18] V. Latora and M. Marchiori, Physical Review Letters **87**, 198701 (2001).

[19] M. E. Newman, Physical review E **70**, 056131 (2004).

[20] R. Guimera and L. A. N. Amaral, nature **433**, 895 (2005).

[21] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, Nature physics **2**, 110 (2006).

[22] D. S. Bassett and E. T. Bullmore, Current opinion in neurology **22**, 340 (2009).

[23] M. Rubinov and O. Sporns, Neuroimage **52**, 1059 (2010).

[24] M. P. Van Den Heuvel and O. Sporns, Journal of Neuroscience **31**, 15775 (2011).

[25] O. Sporns, Current opinion in neurobiology **23**, 162 (2013).

[26] D. Centola, American Journal of Sociology **120**, 1295 (2015).

[27] G. Deco, G. Tononi, M. Boly, and M. L. Kringelbach, Nature Reviews Neuroscience **16**, 430 (2015).

[28] J. R. Cohen and M. D'Esposito, Journal of Neuroscience **36**, 12083 (2016).

[29] H. Aerts, W. Fias, K. Caeyenberghs, and D. Marinazzo, Brain **139**, 3063 (2016).

[30] M. Bertolero, B. Yeo, and M. Desposito, Nature communications **8**, 1277 (2017).

[31] M. A. Bertolero, B. T. Yeo, D. S. Bassett, and M. DEsposito, Nature human behaviour **2**, 765 (2018).

[32] H. Yamamoto, S. Moriya, K. Ide, T. Hayakawa, H. Akima, S. Sato, S. Kubota, T. Tanii, M. Niwano, S. Teller, et al., Science advances **4**, eaau4914 (2018).

[33] M. Stella, M. Cristoforetti, and M. De Domenico, PloS one **14**, e0214210 (2019).

[34] M. Batty, Dialogues in Human Geography **3**, 274 (2013).

[35] Y.-H. Tsai, Urban studies **42**, 141 (2005).

[36] M. Guerois and D. Pumain, Environment and Planning A: Economy and Space **40**, 2186 (2008).

[37] N. Schwarz, Landscape and urban planning **96**, 29 (2010).

[38] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, Scientific reports **4**, 5276 (2014).

[39] C. K. Gately, L. R. Hutyra, and I. S. Wing, Proceedings of the National Academy of Sciences **112**, 4999 (2015), URL `https://doi.org/10.1073/pnas.1421723112`.

[40] R. Ewing and S. Hamidi, Journal of Planning Literature **30**, 413 (2015).

[41] C. Song, T. Koren, P. Wang, and A.-L. Barabási, Nature Physics **6**, 818 (2010), URL `https://doi.org/10.1038/nphys1760`.

[42] T. Louail, M. Lenormand, M. Picornell, O. G. Cantú, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, Nature communications **6**, 6007 (2015).

[43] R. Gallotti, A. Bazzani, S. Rambaldi, and M. Barthelemy, Nature Communications **7** (2016), URL `https://doi.org/10.1038/ncomms12600`.

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

[45] D. Helbing, Reviews of modern physics **73**, 1067 (2001).

[46] D. Li, B. Fu, Y. Wang, G. Lu, Y. Berezin, H. E. Stanley, and S. Havlin, Proceedings of the National Academy of Sciences **112**, 669 (2015).

[47] S. Çolak, A. Lima, and M. C. González, Nature Communications **7** (2016), URL `https://doi.org/10.1038/ncomms10793`.

[48] A. Solé-Ribalta, S. Gómez, and A. Arenas, Networks and Spatial Economics **18**, 33 (2018).

[49] J. Depersin and M. Barthelemy, Proceedings of the National Academy of Sciences **115**, 2317 (2018).

[50] F. Le Néchet, Cybergeo: European Journal of Geography (2012).

[51] B. Stone, Journal of Environmental Management **86**, 688 (2008), URL `https://doi.org/10.1016/j.jenvman.2006.12.034`.

[52] E. Uherek, T. Halenka, J. Borken-Kleefeld, Y. Balkanski, T. Berntsen, C. Borrego, M. Gauss, P. Hoor, K. Juda-Rezler, and J. Lelieveld, Atmospheric Environment **44**, 4772 (2010), URL `https://doi.org/10.1016/j.atmosenv.2010.01.002`.

[53] A. Martilli, Urban Climate **10**, 430 (2014), URL `https://doi.org/10.1016/j.uclim.2014.03.003`.

[54] R. Ewing, G. Meakins, S. Hamidi, and A. C. Nelson, Health & Place **26**, 118 (2014), URL `https://doi.org/10.1016/j.healthplace.2013.12.008`.

[55] D. E. Newby, P. M. Mannucci, G. S. Tell, A. A. Baccarelli, R. D. Brook, K. Donaldson, F. Forastiere, M. Franchini, O. H. Franco, I. Graham, et al., European Heart Journal **36**, 83 (2014), URL `https://doi.org/10.`

1093/eurheartj/ehu458.

[56] M. B. Rice, P. L. Ljungman, E. H. Wilker, K. S. Dorans, D. R. Gold, J. Schwartz, P. Koutrakis, G. R. Washko, G. T. O'Connor, and M. A. Mittleman, American Journal of Respiratory and Critical Care Medicine **191**, 656 (2015), URL https://doi.org/10.1164/rccm.201410-1875oc.

[57] W. Li, K. S. Dorans, E. H. Wilker, M. B. Rice, M. T. Long, J. Schwartz, B. A. Coull, P. Koutrakis, D. R. Gold, C. S. Fox, et al., American Journal of Epidemiology **186**, 857 (2017), URL https://doi.org/10.1093/aje/kwx127.

[58] J. Nicholl, J. West, S. Goodacre, and J. Turner, Emergency Medicine Journal **24**, 665 (2007), URL https://doi.org/10.1136/emj.2007.047654.

[59] L. M. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. B. West, Proceedings of the national academy of sciences **104**, 7301 (2007).

[60] L. M. Bettencourt, Science **340**, 1438 (2013).

[61] A. Bertaud, UC Berkeley IURD Working Paper Series (2004).

[62] V. Volpati and M. Barthelemy, arXiv preprint arXiv:1804.00855 (2018).

[63] R. Louf and M. Barthelemy, Physical review letters **111**, 198702 (2013).

[64] R. Louf and M. Barthelemy, Scientific Reports **4** (2014), URL https://doi.org/10.1038/srep05561.

[65] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, in *International Workshop on Human Behavior Understanding* (Springer, 2010), pp. 14–25.

[66] A. Noulas, C. Mascolo, and E. Frias-Martinez, in *2013 IEEE 14th International Conference on Mobile Data Management* (IEEE, 2013), URL https://doi.org/10.1109/mdm.2013.27.

[67] G. Boeing, Computers, Environment and Urban Systems **65**, 126 (2017).

[68] E. Bullmore and O. Sporns, Nature Reviews Neuroscience **13**, 336 (2012).

[69] R. Louf and M. Barthelemy, PloS one **11**, e0157476 (2016).

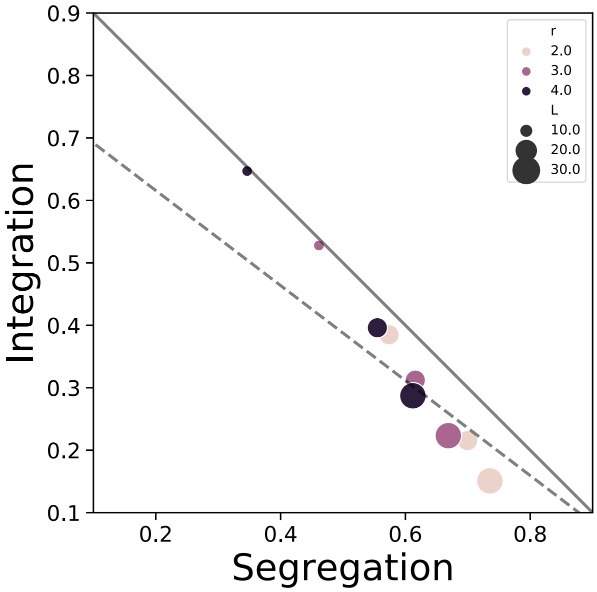
[70] *Future cities challenge*, https://www.futurecitieschallenge.com, accessed: 2019-08-05.

[71] R. Louf and M. Barthelemy, Scientific reports **4**, 5561 (2014).

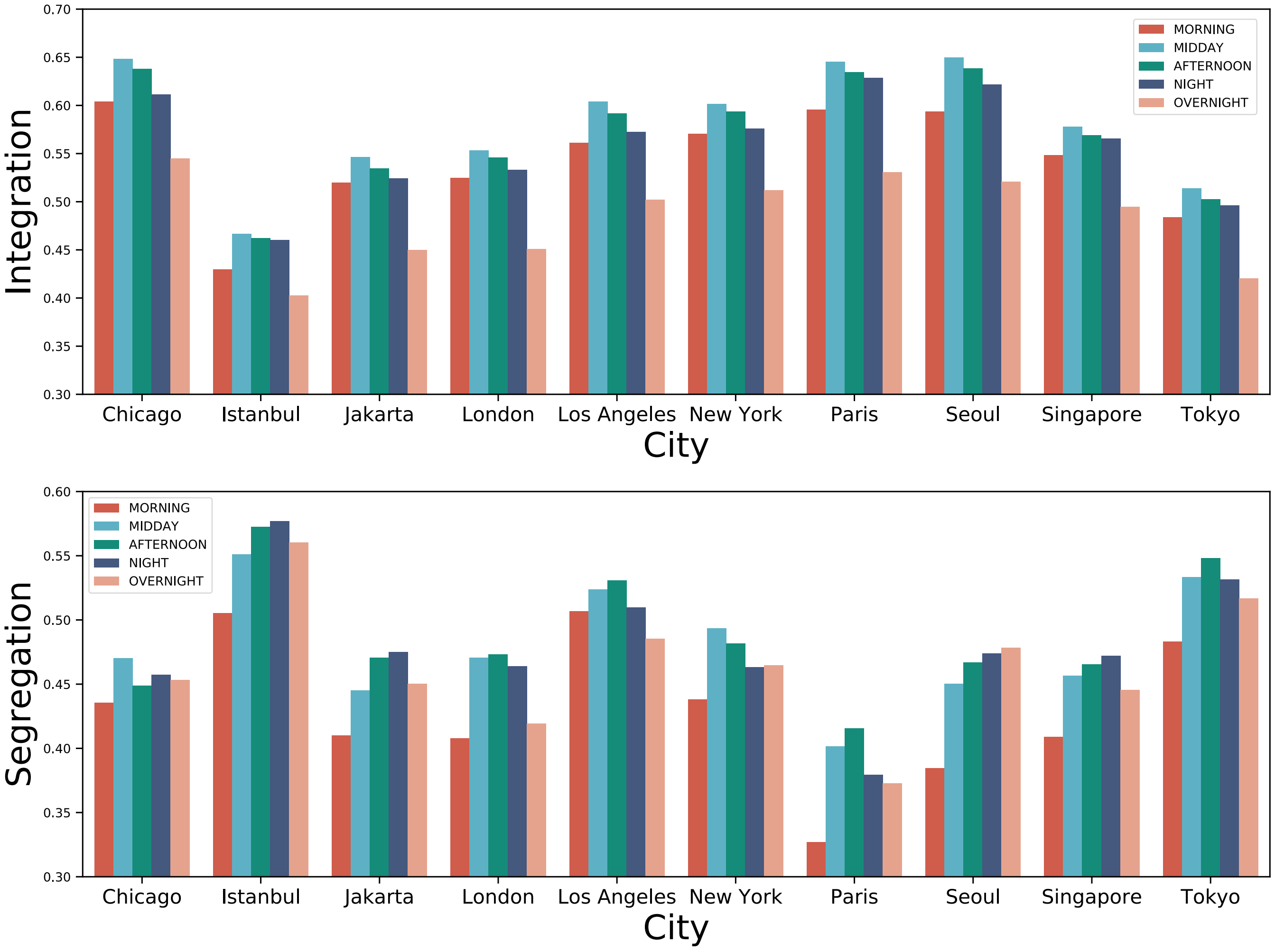
[72] *Foursquare developers venue categories*, https://developer.foursquare.com/docs/api/venues/categories, accessed: 2019-08-02.

[73] M. E. Newman, Physical review E **69**, 066133 (2004).

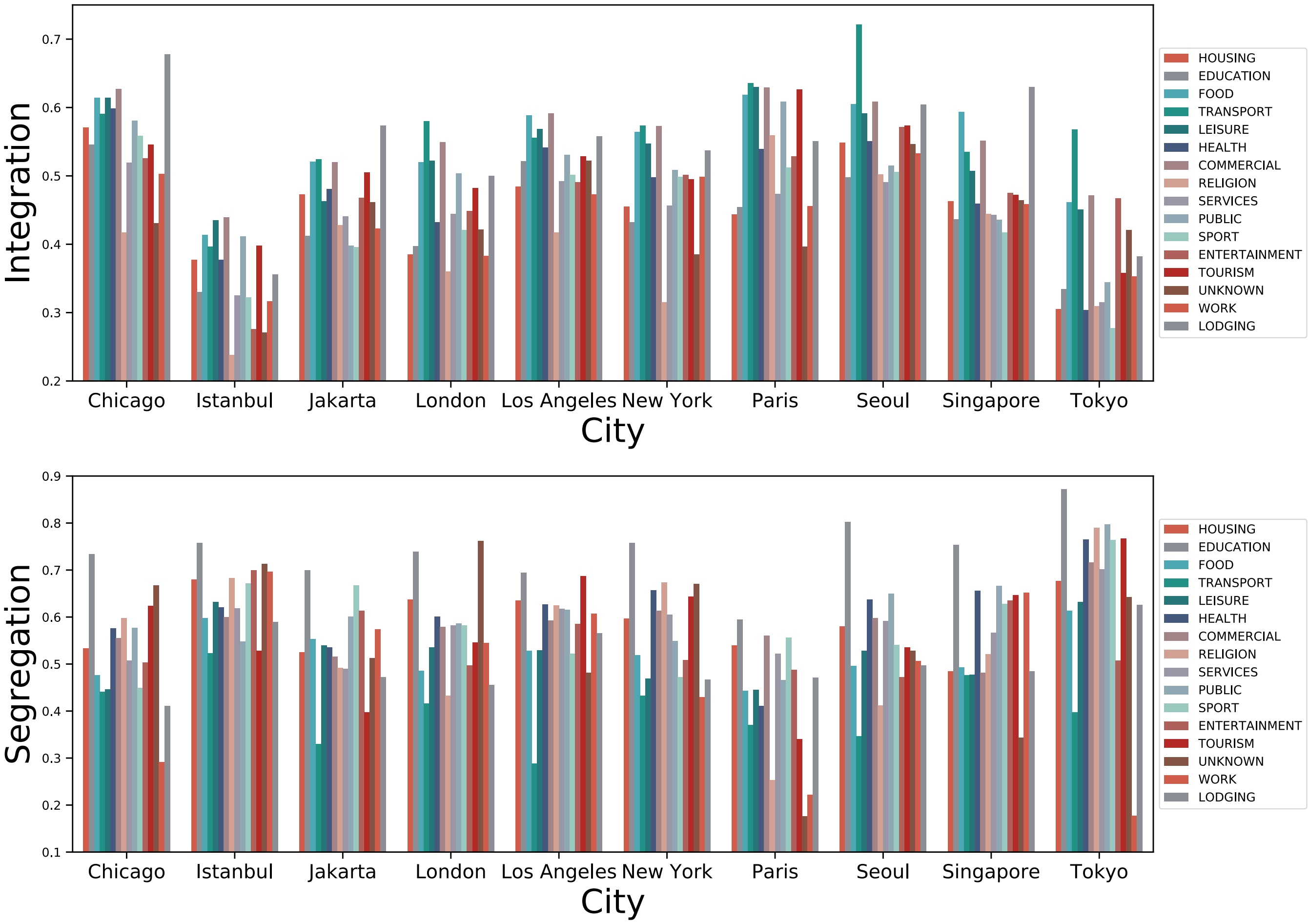
[74] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of statistical mechanics: theory and experiment **2008**, P10008 (2008).



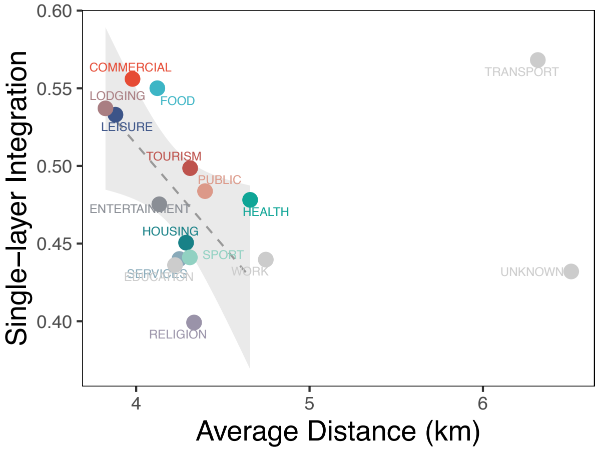
SUPPL. FIG. 1. **Segregation and Integration of Random Geometric Graphs of different sizes.** In this paper, we generate RGGs by i) throwing N nodes in random locations in a square of edge L; ii) connecting all nodes i,j with distance $d(i,j) < r$; iii) rewiring a fraction α of edges. Here, to study the effect of size, we generate networks with identical node density N/L^2 and with no rewiring $\alpha = 0$. For each value of L and r we averaged the values of segregation (modularity Q and integration (Global efficiency E). The result show that, in this scenario, segregation and integration are strongly anti-correlated. High integration is attained for small networks ($L = 10$) with large r, while the opposite yields high segregation.



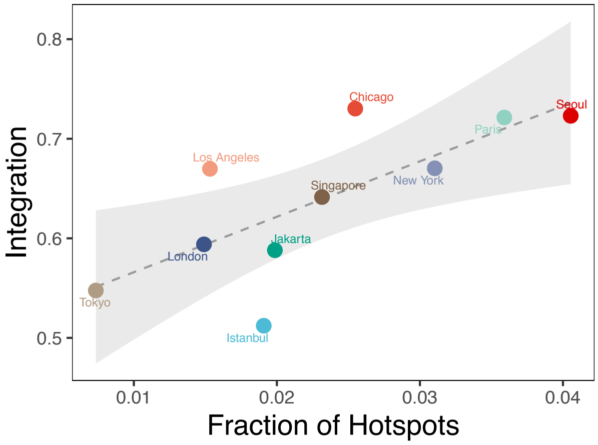
SUPPL. FIG. 2. **Segregation and Integration at different hours of the day.**



SUPPL. FIG. 3. **Single Layer Segregation and Integration.**



SUPPL. FIG. 4. **Average Single Layer Integration *versus* edges characteristic distance.** Similarly to what observed in Fig. 5, if we exclude four outlying layers, the characteristic length of edges on a layer and the layer's Functional integration are correlated (here, more precisely, anti-correlated).



SUPPL. FIG. 5. **Functional integration *versus* fraction of hotspots.** Conversely to what observed in Fig. 7, the Functional integration of a city is larger for cities with a larger number of hotspots per unit area.

hood2vec: Identifying Similar Urban Areas Using Mobility Networks

Xin Liu
University of Pittsburgh
Pittsburgh, Pennsylvania
xil178@pitt.edu

Konstantinos Pelechrinis
University of Pittsburgh
Pittsburgh, Pennsylvania
kpele@pitt.edu

Alexandros Labrinidis
University of Pittsburgh
Pittsburgh, Pennsylvania
labrinid@cs.pitt.edu

ABSTRACT

Which area in NYC is the most *similar* to Lower East Side? What about the NoHo Arts District in Los Angeles? Traditionally this task utilizes information about the type of places located within the areas and some popularity/quality metric. We take a different approach. In particular, urban dwellers' time-variant mobility is a reflection of how they interact with their city over time. Hence, in this paper, we introduce an approach, namely hood2vec, to identify the similarity between urban areas through learning a node embedding of the mobility network captured through Foursquare check-ins. We compare the pairwise similarities obtained from hood2vec with the ones obtained from comparing the types of venues in the different areas. The low correlation between the two indicates that the mobility dynamics and the venue types potentially capture different aspects of similarity between urban areas.

KEYWORDS

Similarity Metric, Urban Area Representation, Venue Category

1 INTRODUCTION

Identifying similar areas in a city can facilitate dwellers and visitors exploring the city better. An intuitive approach for comparing urban areas and providing recommendations is to utilize information about the type of venues within an area. The types of venues within an area (zip code, neighborhood, block etc.) can be thought of as the "signature" of this area. However, there are several assumptions behind this consideration. For example, this implicitly assumes that all venues within the area are *active* through the whole day. For instance, while two areas can appear to have similar venues, they can be significantly different when introducing the temporal dimension (e.g., an area with "lunch restaurants" compared to one with "dinner restaurants"). Furthermore, this static view does not include information on how urban dwellers interact with these areas, as captured through their movements between them.

In this paper, we explore the mobility patterns of Foursquare users in the three US cities included in the Future Cities Challenge (FCC) dataset, namely, New York, Los Angeles and Chicago, borrowing analytical tools from the network science literature. In particular, we build hood2vec that first designs a network between urban areas in a city using the FCC dataset and then obtains a vector representation of each area using a network embedding. We then use these representations to identify similar areas and make comparisons with other similarity metrics based purely on venue types and checkins. *Similar* urban areas will be represented by points (i.e., vector) closer together in the latent space identified by the network embedding. We then calculate the similarity between two urban areas using the (Euclidean) distance between the latent space points for the two areas.

Existing literature has attempted to identify the functionality of urban areas, and consequently, cluster areas based on their functionality. Topic modeling is the dominant techniques in this line of research (e.g., [2, 6]). Other studies have attempted to identify similar areas across cities mainly using the type of activities recorded in the different areas of the different cities (e.g., [3, 5]). In the rest of the paper we formally describe our approach and present the results obtained from the three cities aforementioned.

2 HOOD2VEC: LEARNING AN URBAN AREA VECTOR REPRESENTATION

The FCC dataset provides information about the mobility patterns of Foursquare users. Each data point has the following tuple format: `<start venue, end venue, trip year and month, trip period in a day, number of checkins>`. The *number of checkins* captures the number of times that the specific transitions were observed in the dataset. The dataset also provides information about the name, geographic coordinates and category for each venue.

The majority of the transitions recorded in the dataset are observed only one time. In particular, 95% of the transitions are observed less than 3 times. In order to avoid fitting the noise, we aggregate the transitions (movements) over a wider geographical scale. We also separate the movements according to the time period of movement occurrence according to the data - i.e., overnight (00:00 to 05:59), morning (06:00 to 09:59), midday (10:00 to 14:59), afternoon (15:00 to 18:59), night (19:00 to 23:59). Using MapQuest's Geocoding API[1] we obtain the zip code for each venue and we aggregate the movements at the zip code level (the wider scale). More specifically, we transform the original data to the following format per period: <start zip code, end zip code, trip year and month, number of checkins>. At zip code level, only 10% of the movements have less than 2 observations. However, 20% of the zip codes contain fewer than 10 venues and hence, we filter them out from our analysis. While this might sound a large number to ignore, the checkins within these zip codes cover only 0.5% of the total checkins in the dataset.

Finally, for each city $c \in \mathcal{C}$ = {New York, Los Angeles, Chicago} we define its directed urban flow network $\mathcal{G}_{c,p}$ per period $p \in \mathcal{P}$ = {overnight, morning, midday, afternoon, night} at the zip-code level as follows: $\mathcal{G}_{c,p} = (\mathcal{U}, \mathcal{E})$, where the set of nodes $\mathcal{U}$ is the set of zip code areas in city c. A directed edge $e_{ij} \in \mathcal{E}$ exists between two zip codes $u_i, u_j \in \mathcal{U}$ if there has been observed at least one movement from a venue in u_i to a venue in u_j during period p. We

[1] https://developer.mapquest.com/documentation/geocoding-api/

also annotate every edge e_{ij} with a weight $w(e_{ij})$, which captures the number of checkins of such movements observed.

We would like to note here that while we have chosen the zip codes as our unit, one can define an urban area differently. For instance, one can use the notion of *neighborhoods* that can include several zip codes, or the census tracts, or any other definition of neighborhood [1].

2.1 Vector Representation by node2vec

In order to obtain a vector representation for the nodes of $\mathcal{G}_{c,p}$, i.e., the zip codes at c in period p, we will rely on learning a network embedding. There are several ways to learn a node embedding for a network but in this work we make use of node2vec [4]. Briefly, node2vec utilizes second order random walks to learn a vector representation for the network nodes that optimizes a neighborhood preserving objective function. The framework is flexible enough to accommodate various definitions of network neighborhood and facilitate the projections of the network nodes in the latent space according to different *similarity* definitions. Given that we are interested in the structural equivalence of the urban areas, we pick the parameters of node2vec accordingly ($p = 1$ and $q = 2$ [4]). We also utilize 1,000 random walks for the sampling process, while we set the dimensionality of the latent space to $d = 10$. This is consistent with the dimensionality of another vector representation to be introduced in Section 2.2. node2vec finally provides us with a vector $\mathbf{v}_i \in \mathbb{R}^d, \forall u_i \in \mathcal{U}$, that we can then use to identify the similarity between two urban areas.

2.2 Vector Representation utilizing Venue Categories

As alluded to above, a straightforward way to define the similarity between two areas is to compare the distribution of the type of venues they host. More specifically we can define a vector $\mathbf{z}_i$ for each urban area node, such that its k^{th} element $z_{ik} = \frac{n_{ik}}{N_i}$, where n_{ik} is the number of venues of type k within area i and N_i is the total number of venues within i. For defining vectors $\mathbf{z}_i$ we use the 10 top-level venue categories in Foursquare (thus, $\mathbf{z}_i \in \mathbb{R}^{10}$) : Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. Similar to hood2vec, we can now define the similarity between two urban areas i and j using the distance between vectors $\mathbf{z}_i$ and $\mathbf{z}_j$.

Similar to number of venues, the number of checkins in venues of different types can also be used as the vector representation of an urban area. In particular, we define a vector $\mathbf{z}_i^{\text{check}}$ for each urban area node, such that its k^{th} element $z_{ik}^{\text{check}} = \frac{n_{ik}}{C_i}$, where c_{ik} is the number of checkins of venues of type k within area i and C_i is the total number of checkins of venues within i. We follow the same 10 top-level venue categories for $\mathbf{z}_i^{\text{check}}$ (i.e., $\mathbf{z}_i^{\text{check}} \in \mathbb{R}^{10}$). Then we can also define the similarity between two areas i and j by the distance between vectors $\mathbf{z}_i^{\text{check}}$ and $\mathbf{z}_j^{\text{check}}$.

Period	O	MO	MI	A	N
New York	0.116^{***}	0.152^{***}	0.147^{***}	0.152^{***}	0.144^{***}
Los Angeles	0.184^{***}	0.290^{***}	0.229^{***}	0.219^{***}	0.142^{***}
Chicago	0.284^{***}	0.316^{***}	0.327^{***}	0.336^{***}	0.323^{***}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 1: Correlation between movement and category representations.

3 URBAN AREA SIMILARITY

One of the questions is which urban area representation should we use? Do they even provide us with a different view of the similarity between two areas? In order to explore this we will calculate the pairwise similarities using the network embedding learnt from hood2vec and compare them with the corresponding pairwise similarities obtained from a simple venue-based representation of urban areas (see Section 2.2). Formally, the similarity of two areas i and j, with vector representations $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively, is defined as:

$$\sigma_{ij} = \exp(-\mathtt{dist}_s(\mathbf{x}_i, \mathbf{x}_j)) \qquad (1)$$

where $\mathtt{dist}(\mathbf{x}_i, \mathbf{x}_j)$ is the (Euclidean) distance between the representations of i and j.

We can now examine whether different representations for the urban areas provide different views for their similarity. In particular, if σ_{ij} and σ'_{ij} are the similarities between areas i and j using different vector representations, their Pearson correlation coefficient $\rho_{\sigma,\sigma'}$ will be high if the two representations provide similar information, and low otherwise. We can further compare in the same way the similarity of two areas for the same vector representation over different time periods.

4 EXPERIMENTS AND RESULTS

In this section, we will present the results of our analysis and compare the pairwise similarities obtained from hood2vec and a simple venue category-based representation.

4.1 Movement and Venue Categories

We calculate the correlation between two representations, $\mathbf{v}$ and $\mathbf{z}$, (by the method in Section 3) in three cities: New York City, Los Angeles and Chicago. There is a total of 141 zip codes u_i (9870 pairs) in New York city, 111 zip codes (6105 pairs) in Los Angeles, and, 59 zip codes (1711 pairs) in Chicago. We further extend our comparisons to each time period provided in the data. The results are presented in Table 1. Note that we use the following notation for the five time periods - O: overnight; MO: morning; MI: midday; A: afternoon; N: night. As we can see all the correlations are positive, albeit, small, pointing to the two representations capturing different types of information. We also calculate the correlation between $\mathbf{z}$ and $\mathbf{z}^{\text{check}}$ in three cities. The correlations for these three cities are 0.839, 0.930, 0.936, respectively. I.e., the representations of venue category based on number of venues and checkins are highly correlated. This indicates low correlation of representations between hood2vec and checkin-based venue category.

We further inspect the relationship between the two approaches from the perspective of the top-k neighbors for each zip codes. In

Period	O	MO	MI	A	N
New York City	0.036	0.065	0.062	0.029	0.016
Los Angeles	0.098	0.254	0.150	0.104	0.015
Chicago	0.139	0.136	0.170	0.164	0.129

Table 2: Jaccard index ($k = 5$) for the three cities for the different time periods averaged over the corresponding zip codes.

particular, for each zip code u_i we find the $k = 5$ closest zip codes to i based on their hood2vec representation ($\mathbf{v}$), $\mathcal{N}_{5,i,\text{hood2vec}}$. Similarly, we calculate the top-5 neighbors of zip code u_i based on their venue category representation ($\mathbf{z}$), $\mathcal{N}_{5,i,cat}$. We then calculate the Jaccard index of the two sets:

$$J(\mathcal{N}_{5,i,\text{hood2vec}}, \mathcal{N}_{5,i,cat}) = \frac{|\mathcal{N}_{5,i,\text{hood2vec}} \cap \mathcal{N}_{5,i,cat}|}{|\mathcal{N}_{5,i,\text{hood2vec}} \cup \mathcal{N}_{5,i,cat}|} \quad (2)$$

Table 2 presents the average Jaccard index for every city and time period. Furthermore, Figure 1 presents the Jaccard index as a function of the number of neighbors k considered for every city, averaged over different time periods and zip codes. As one might have expected from the earlier results presented, in general, under different k, there are few shared neighbors when using the two different representations for the zip codes. This strengthens our hypothesis that these two types of representations capture different information for the areas.

Moreover, Figures 2-4 illustrate the Jaccard index for every zip code per city, averaged over the different time periods. As we can see most of the zip codes in all cities have a fairy low Jaccard index. New York City's zip codes exhibit overall lower Jaccard index compared to Chicago and Los Angeles (in accordance to the results in Table 1, 2). Zip codes with high Jaccard index are essentially urban areas for which the two different representations examined identify a high overlap on areas similar to them. This happens to a larger extend in Los Angeles and Chicago compared to New York City. This can potentially be due to (a) the compact nature of NYC that allows people to explore several different areas and hence, geographically remote zip codes are close in the hood2vec latent space, and/or, (b) the different geographic distribution of venues in the three different cities. More specifically, the compact nature may cause venues in New York city more evenly distributed, since they are easily accessible by dwellers. In contrast, scattered nature of Los Angeles may lead to biased venue distribution due to various accessibility of different regions; this could be the reason for slightly high Jaccard indices in some areas. Chicago has fewer zip code areas such that an area can has higher probability of sharing the same closest area(s) in two representations; this can cause slightly high Jaccard indices in some areas. Nevertheless, regardless of the reasons for the differences across the cities examined, in all cases the Jaccard index does not go beyond 0.4. Simply put, there is no zip-code in these three cities, for which the overlap between the top-5 neighbors identified by hood2vec and a simple venue-based vector representation is more than 40%, supporting our hypothesis that these two different approaches capture different information with respect to the similarity of the areas.

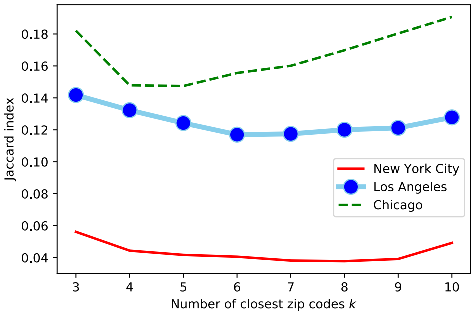


Figure 1: Average Jaccard index as function of number of closest neighbors k.

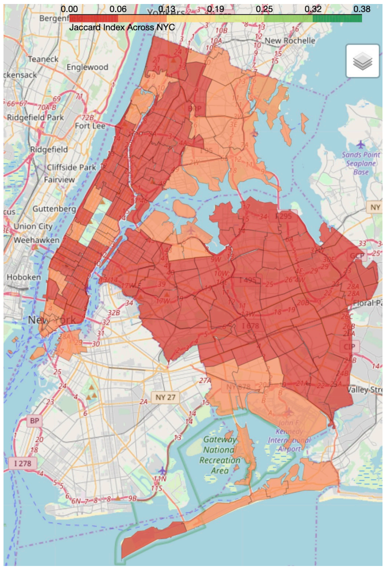


Figure 2: Average Jaccard index over the different time-periods for the zip codes New York City.

4.2 hood2vec representation across time

We further explore how the representation obtained for a zip code through hood2vec changes over time (i.e., over the different time-periods in the dataset). Let us assume the two periods p_1 and p_2, and the corresponding hood2vec representation vectors $\mathbf{v}_{p_1}$ and $\mathbf{v}_{p_2}$ respectively. Then following similar steps as the ones described in Section 3, we can obtain the pairwise correlation of the between periods p_1 and p_2 for the same city. The correlations of each city are shown in Fig. 5, 6, 7.

One can observe that for these three cities, the correlations between any pair of periods are very high, all over 0.9. This means that, the patterns of movements are similar regardless of the time of a day (based on the hood2vec representation). Since New York City and Chicago are more geographically compact, it is easier for dwellers to move within the city for any purpose at any time. This could be the reason that the overall movement patterns within a day are similar. Los Angeles is geographically scattered, which limits the convenience of movements; dwellers tend to move within nearby areas at

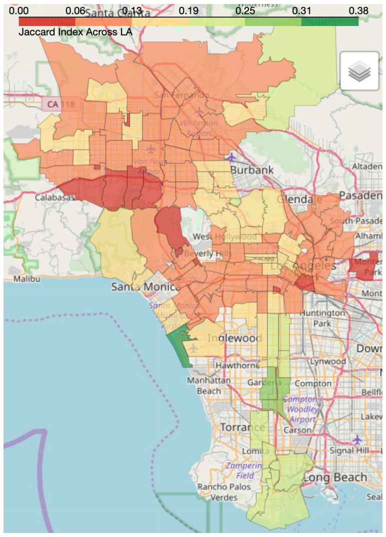


Figure 3: Average Jaccard index over the different time-periods for the zip codes in Los Angeles.

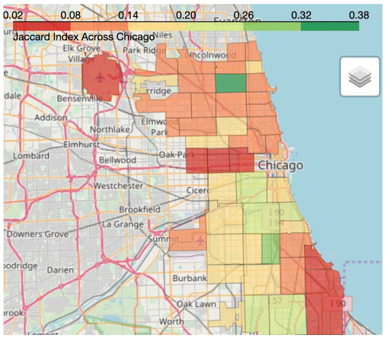
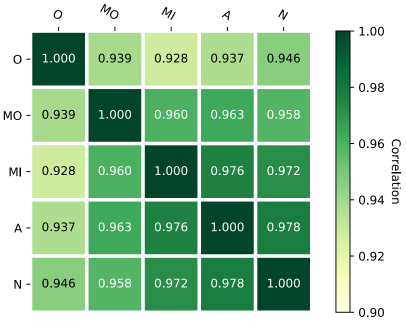


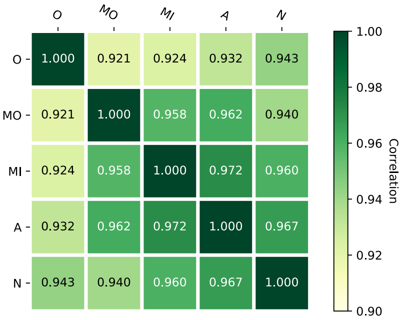
Figure 4: Average Jaccard index over the different time-periods for the zip codes in Chicago.



	O	MO	MI	A	N
O	1.000	0.939	0.928	0.937	0.946
MO	0.939	1.000	0.960	0.963	0.958
MI	0.928	0.960	1.000	0.976	0.972
A	0.937	0.963	0.976	1.000	0.978
N	0.946	0.958	0.972	0.978	1.000

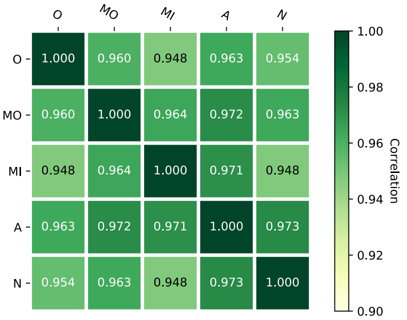
Figure 5: Correlation among representations of different periods in New York City.

any time of the day. This may cause similar movement patterns of all day. The interested reader can explore the different urban area representations at: http://www.pitt.edu/~xil178/hood2vec.html



	O	MO	MI	A	N
O	1.000	0.921	0.924	0.932	0.943
MO	0.921	1.000	0.958	0.962	0.940
MI	0.924	0.958	1.000	0.972	0.960
A	0.932	0.962	0.972	1.000	0.967
N	0.943	0.940	0.960	0.967	1.000

Figure 6: Correlation among representations of different periods in Los Angeles.



	O	MO	MI	A	N
O	1.000	0.960	0.948	0.963	0.954
MO	0.960	1.000	0.964	0.972	0.963
MI	0.948	0.964	1.000	0.971	0.948
A	0.963	0.972	0.971	1.000	0.973
N	0.954	0.963	0.948	0.973	1.000

Figure 7: Correlation among representations of different periods in Chicago.

5 CONCLUSIONS

In this paper, we propose hood2vec to identify the similarity between urban areas through learning a node embedding of the mobility network captured through Foursquare check-ins. We compare the pairwise similarities obtained from hood2vec with the ones obtained from comparing the types of venues in the different areas. The low correlation between the two indicates that the mobility dynamics and the venue types potentially capture different aspects of similarity between urban areas.

Acknowledgments: This work is part of the PittSmartLiving project (https://pittsmartliving.org/) which is supported in part by National Science Foundation Award CNS-1739413.

REFERENCES

[1] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *AAAI ICWSM.*

[2] Justin Cranshaw and Tae Yano. 2010. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In *NIPS Workshop of Computational Social Science and the Wisdom of the Crowds.*

[3] 4SQ Eng. [n. d.]. What neighborhood is the "East Village" of San Francisco? Retrieved June 9, 2019 from https://medium.com/foursquare-direct/a-hackday-project-what-neighborhood-is-the-east-village-of-san-francisco-229f317f597a

[4] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 855–864.

[5] Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. 2015. Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In *Ninth International AAAI Conference on Web and Social Media.*

[6] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 186–194.

Mining behavioural constraints in urban mobility sequences of Tokyo

Galina Deeva[*] and María Óskarsdóttir[*]
[*]Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium,
Email: {galina.deeva, maria.oskarsdottir }@kuleuven.be

***Abstract*—We analyse the longitudinal mobility data from a sequence mining perspective using a technique that discovers behavioural constraints in sequences of movements between venues. Our goal is to discover distinctive behavioural patterns in the sequences relative to when in the day they were formed. We analyse sequences of venues as well as sequences of subcategories and categories to discover how people move through Tokyo.**

I. Introduction

In this project, we look at the longitudinal mobility data set from a sequence mining perspective. The movements of individuals across time can be viewed as sequential data where each location is an item in the alphabet of venues and a sequence is comprised of an ordered list of locations that describe people's behaviour when moving through a city. In particular, we apply a novel and powerful sequence mining technique called the interesting Behavioural Constraint Miner or iBCM [1]. The technique is capable of discovering expressive and concise patterns using behavioural constraint templates, such as simple occurrence, looping and position in a sequence. In addition, the technique can discover the absence of a particular behaviour, such as the co-existence of two items, which is interesting for the understanding of urban mobility patterns. The output of iBCM is a set of features that represent behavioural constraints in the sequences. Subsequently, supervised analytics techniques can be applied to the feature set to discover meaningful patterns, interesting signals and to classify the sequences.

The goal of this project is to discover distinctive and discriminating behavioural constraints in mobility sequences during different periods of the day. In the context of urban mobility we take a fine-grained look to the venues to see how individuals travel through a city. We apply iBCM to commute sequences and use its window-based approach to discover particular behaviour in certain time periods. We apply the technique in a supervised setting with the goal of classifying the sequences with respect to the part of day. For urban growth and dynamics, we consider, on the one hand, the subcategory and, on the other hand, the category of the venues. Thereby, we analyse sequences of categories, i.e., whether they are for example Residence, Food, or Travel & Transport, or of subcategories. Interesting behavioural constraint templates in this case could be whether after visiting a museum do people go to a restaurant and whether the restaurant is close to the museum or is public transport needed to get there. For this application, our goal is again to classify the sequences with respect to the part of day and to study the feature sets generated by iBCM to discover behavioural constraints that are prominent at each time. This look at urban dynamics allows us to study in which context event interactions happen and could also be applied for location intelligence. Our analyses transition from a fine grained view of movements between venues to an aggregated approach that looks at mobility in the broader sense using categories. This gives alternative perspectives of urban dynamics and behavioural patterns in the context of venue interaction.

Our approach consists of three steps, namely building networks of venues and the movements between them, the generation of pseudo sequences that represent the collective behaviour of the population as it moves through the city, and the subsequent sequence classification and constraint mining. As the data is too aggregated to analyse the mobility sequences of individuals, we will extract pseudo sequences from the networks built using information about individual movements between venues. The networks thus represent the collective movements of the population.

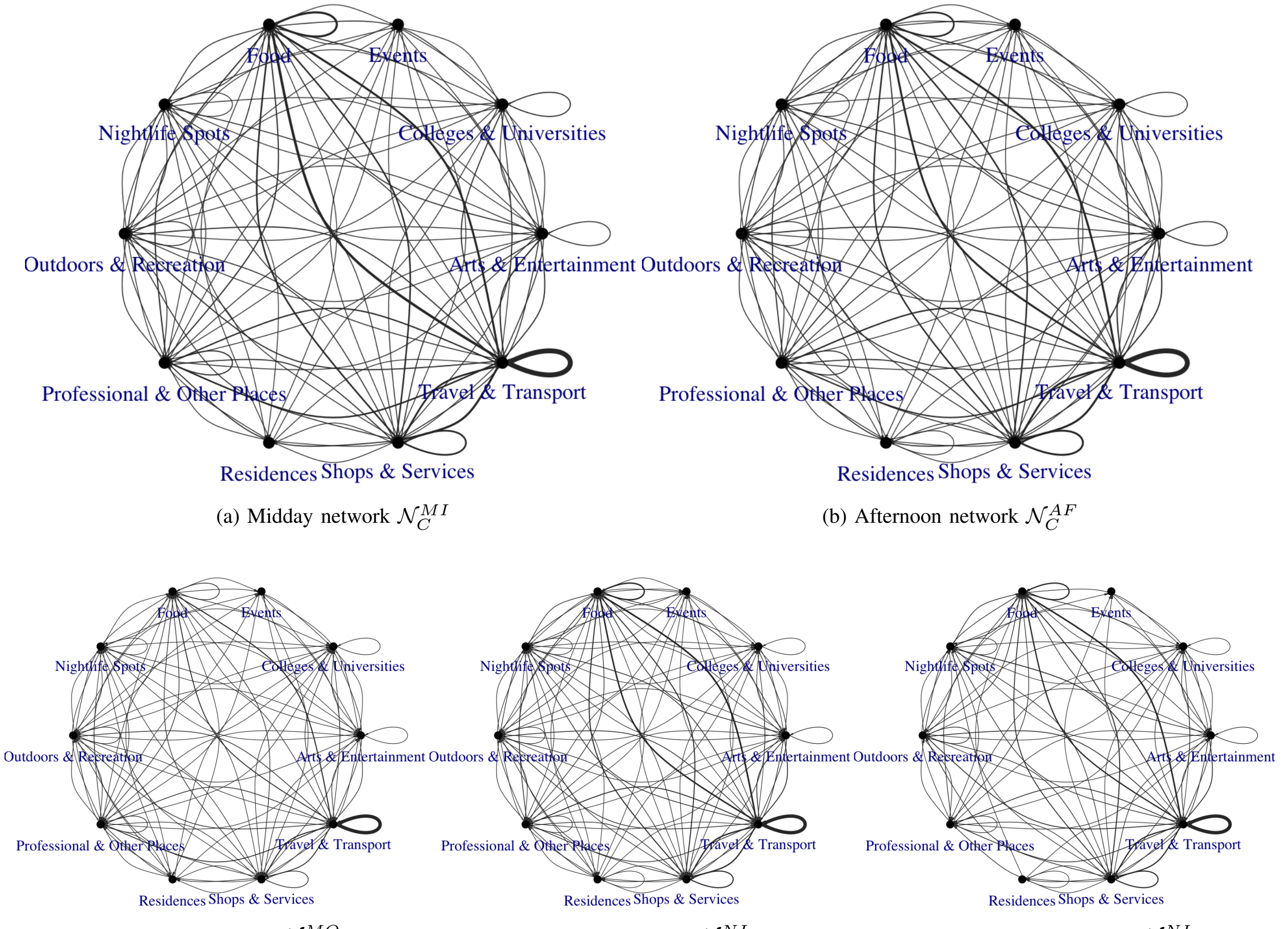
For our analyses we focus on Tokyo and the months April, May and June 2017.

TABLE I
An example of venues, their subcategory and category

Venue	Subcategory	Category
Leuven train station	Train Stations	Travel & Transport
Bart's bar	Bars	Nightlife Spots
Eskimo attire	Clothing Stores	Shops & Services
Moskow Meals	Russian Restaurants	Food

II. Network building

We make networks at three levels of granularity: venue(V), subcategory(S) and category(C). Table I shows an example of a few venues together with their corresponding subcategory and category. The venue networks are created using the aggregated movements between venues. To create the subcategory and category networks, we replace the venues by their corresponding subcategory or category, respectively, and aggregate the edges. To distinguish mobility behaviour at different times of the day, we build separate networks for each time period. The time periods are morning(MO), midday(MI), afternoon(AF), night(NI) and overnight(OV). The definition of the time periods can be seen in Table II.



(a) Midday network $\mathcal{N}_C^{MI}$

(b) Afternoon network $\mathcal{N}_C^{AF}$

(c) Morning network $\mathcal{N}_C^{MO}$

(d) Night network $\mathcal{N}_C^{NI}$

(e) Overnight network $\mathcal{N}_C^{NI}$

Fig. 1. Tokyo category networks.

We denote the networks using the granularity as a subscript and the part of day as a superscript. For example, $\mathcal{N}_V^{MO}$ is the morning venue network and $\mathcal{N}_S^{NI}$ is the night subcategory network. The venue networks have 58 thousand nodes, the subcategory networks have 456 nodes and the category networks have 10 nodes. All the networks are directed, indicating a movement from one location to another, and weighted by the number of movements in the given time period. Figure 1 shows category networks for Tokyo in June 2017 for different parts of the day. The network on the left in the top row shows peoples movements during midday and the network on the right in the top row shows the movements in the afternoon. The networks in the bottom row show, from left to right, movements in the morning, night and overnight. Although the networks appear very similar, there are some noticeable differences. For example, at midday Residences have no self loops and there are no movements from Events to Nightlife Spots. In addition, in the afternoon the weights on the edges from Food to Travel & Transport and from Travel & Transport to Food are almost the same, whereas at midday the weight on the edge from Travel & Transport to Food is 36% higher than on the edge in the opposite direction. The overnight network is also less densely connected than the other networks.

TABLE II
The time periods

Part of day	Start hour	End hour
Morning	6:00	10:00
Midday	10:00	15:00
Afternoon	15:00	19:00
Night	19:00	24:00
Overnight	24:00	6:00

III. Pseudo Sequence generation

As we do not have individual mobility sequences, we generate pseudo sequences that portray the collective mobility of the population. The pseudo sequences are generated by initialising a number of random walks in the networks. These random walks give a representation of the movements between venues and are used in lieu of sequences. In the networks, we know for certain that transitions from A to B and from B to C happen, but we do not know if the order of these transitions is from A to B to C. However, with an assumption that some people have moved more than once around the city, and given that there is a high chance for the location B to be followed by C, we can conclude that some people who moved from A to B continued their way with a movement from B to C. Furthermore, as we generate a considerable number of such random walks, we capture the variability in movements with more common behaviour appearing more frequently in the sequences. Thus, with such pseudo sequences we get an approximation of the average movements. We consider random walks of length 10, 20 and 30.

From the venue networks $\mathcal{N}_V$ we get venue sequences $\mathcal{S}_V$ in which we replace the venues with their corresponding subcategory and category to obtain the sequences $\mathcal{S}_{V\to S}$ and $\mathcal{S}_{V\to C}$. From the subcategory networks $\mathcal{N}_S$ we obtain subcategory sequences $\mathcal{S}_S$ and we replace the subcategories with their corresponding category to obtain the category sequences $\mathcal{S}_{S\to C}$. Finally, from the category networks $\mathcal{N}_S$ we obtain the category sequences $\mathcal{S}_C$.

A pseudo sequence for a venue network with the venues in Table I could for example be

Leuven train station→Eskimo attire→Moskow Meal→Leuven train station $\in \mathcal{S}_V$

with the corresponding subcategory and category sequences

Train Stations→Clothing Stores→Russian Restaurants→ Train Stations $\in \mathcal{S}_{V\to S}$

Travel & Transport→Shops & Services→Food→Travel & Transport $\in \mathcal{S}_{V\to C}$.

The first sequence is very detailed since it shows movements between exact locations. As there are 58 thousand venues there is a lot of variability in the sequences. The venue networks furthermore have multiple small components of only one or two nodes, which means that these venues are never together in a sequence with the majority of venues. The second sequence on the other hand shows a more eminent mobility pattern and the same holds for the third sequence, where the pattern is even more high-level.

The sequences are labelled according to which part of day network they come from.

IV. Sequence classification and constraint discovery

iBCM is a state-of-the-art sequence classification technique that discovers discriminative patterns from sequential data. The behavioural constraints mined by iBCM are based on Declare language. iBCM is a direct sequence classification technique, which means that it generates sequential patterns separately for each class of sequences. Next, each sequence is transformed to a feature vector with binary features indicating whether a certain pattern is present in the sequence. Subsequently, a predictive model is built based on these features.

We look into 5 classes of sequences that represent 5 time periods described above (morning(MO), midday(MI), afternoon(AF), night(NI) and overnight(OV)). iBCM derives binary features, which are subsequently used in a classification algorithm. During the feature generation procedure, we vary the support level (the percentage of sequences that contain a certain sequential pattern) from 0.4 to 0.8 with 0.1 intervals, thus obtaining 5 different feature sets for each dataset. Next, we apply a Random Forest algorithm, chosen because of its fast and efficient performance.

V. Results

Table III presents an overview of the accuracy (*acc*) and lift results produced by iBCM and Random Forest, as well as the number of behavioural constraints generated for different pseudo sequences. The results are calculated using 10-fold cross validation. For each dataset, we only present the results for the support level (*sup*) that gave the best accuracy.

TABLE III
Classification results for pseudo sequences generated from networks with different parameters

Granularity	L	Type	Month	Sup	Con	Acc	Lift
category	10	C	4	0.5	24	1.0	4.9
category	10	S	4	0.7	23	0.99	4.51
category	20	C	4	0.7	20	0.99	4.9
category	20	S	4	0.8	22	0.87	4.54
category	20	V	4	0.4	36	0.96	3.44
subcategory	10	S	4	0.6	14	0.67	2.74
subcategory	10	V	4	0.5	10	0.74	3.09
subcategory	20	S	4	0.5	31	1.0	4.54
subcategory	20	V	4	0.4	15	0.85	2.30
category	10	C	5	0.7	24	0.86	3.35
category	10	S	5	0.7	24	0.87	3.24
category	10	V	5	0.4	28	0.97	3.4
category	20	C	5	0.8	23	0.99	4.99
category	20	S	5	0.8	23	0.87	4.59
category	20	V	5	0.4	41	0.99	3.39
subcategory	10	S	5	0.6	14	0.66	2.67
subcategory	10	V	5	0.5	10	0.71	1.85
subcategory	20	S	5	0.5	41	0.99	4.59
subcategory	20	V	5	0.4	20	0.85	2.3
category	10	C	6	0.5	24	0.85	4.99
category	10	S	6	0.6	19	0.99	4.53
category	10	V	6	0.4	28	0.97	3.44
category	20	C	6	0.8	22	0.84	4.99
category	20	S	6	0.8	23	0.87	4.54
category	20	V	6	0.4	36	0.99	3.44
subcategory	10	S	6	0.6	19	0.83	4.54
subcategory	10	V	6	0.5	12	0.74	3.09
subcategory	20	S	6	0.5	26	0.99	4.54
subcategory	20	V	6	0.4	18	0.87	2.48

It can be observed that the accuracy results obtained by iBCM and Random Forest are relatively high, ranging from 0.66 to 1.0. Such high accuracy obtained in multi-class classification confirms that iBCM was able to find patterns that were discriminating between the classes.

Some examples of the observed patterns for different sequences are listed in Tables IV and V, which illustrate behaviour in April and May 2017. For example, the pattern Existence3(Travel & Transport) in Table IV indicates that Travel & Transport item occurred at least 3 times during a certain window in the morning, which confirms the intuition that people tend to use several types of transportation in a sequence to commute to work in the morning. Similarly, we can conclude that at lunchbreak people tend to use transport to get to a restaurant, that they are less likely to visit Colleges & universities in the afternoon, they go from shops to the train station in the evening, and, finally, that they will to a lesser extent visit restaurants during the night.

Similar patterns occur in Table V. We can observe that people tend to visit some food places in the morning, followed by commuting to work. Then, during lunchbreak, some people might use transportation to visit Shops & Services. Finally, it can be concluded, that the restaurants don't get visits in the afternoon, Colleges & Universities are not visited in the evening, and Shops & Sevices are not visited at night.

TABLE IV
EXAMPLES OF CONSTRAINTS DERIVED BY IBCM (TYPE = VENUE, GRANULARITY = CATEGORY, L = 20, MONTH = MAY)

Label	Constraint
MO	Existence3(Travel & Transport)
MI	CoExistence(Travel & Transport, Food)
AN	Absence(Colleges & universities)
NI	CoExistence(Shops & Services, Travel & Transport)
OV	Absence(Food)_1, Absence(Food)

TABLE V
EXAMPLES OF CONSTRAINTS DERIVED BY IBCM (TYPE = SUBCATEGORY, GRANULARITY = CATEGORY, L = 20, MONTH = APRIL)

Label	Constraint
MO	CoExistence(Food, Travel & Transport)
MI	CoExistence(Travel & Transport, Shops & Services)
AN	Absence(Food)
NI	Absence(Colleges & universities)
OV	Absence(Shops & Services)

In terms of the influence of various parameters on classification results, we make the following observations:

- We generate sequences of 3 possible lengths (L): 10, 20 and 30 items. For sequences with L = 30 no features were obtained, which is in line with intuition that no meaningful sequences of so many movements could be possible within one time period. The results for L=10 and L=20 are similar, however, there is a slight tendency for the accuracy to be higher for longer sequences with the average accuracy of 0.85 and 0.93, respectively. This confirms the expectation that longer sequences could potentially contain more information; however, in the context of this study, there is a clear limitation to a length of pseudo sequences that can still be realistic.
- The sequences with 3 different levels of granularity are analysed: category, subcategory and venues. The latter didn't yield any results, because the alphabet of possible venues was simply too large (58 thousands), making it challenging to search for patterns in short sequences of length 10-20. The average accuracy results for subcategory and category are 0.83 and 0.93, respectively. Thus, it is possible that more interesting patterns can be extracted when looking at the movements between venues from a more general perspective. However, more experiments are needed to confirm this observation.
- Finally, there is no clear tendency for a certain type of network generation to produce better results.

VI. CONCLUSION

In this project we used a sequence classification technique to discover behavioural patterns that are distinctive in mobility patterns during different parts of the day. The high accuracy levels the we obtained show that the pseudo sequences that we generated from the venue, subcategory and category networks contain constraints that are distinct and discriminative for the various parts of the day. The technique is fairly good at detecting these constraints, which in addition are intuitive. They provide an understanding of urban mobility and the specific constraints in sequences from each class could be used to improve transport and accessibility in the city.

For future work, we would like to to carry out the same analysis for the other cities that were provided to see if the mobility patterns are universal. If they are not, we could apply the same technique to classify the cities and discover patterns that are unique for each city. We would have liked to work with real sequences for our analyses, but the data was aggregated for privacy reasons. However, the pseudo sequences provided a representation of movements between venues that were both logical and plausible in the context of urban mobility throughout the day. It would be very interesting to apply iBCM to individual mobility sequences to see how well they are approximated by the pseudo sequences.

REFERENCES

[1] J. De Smedt, G. Deeva, and J. De Weerdt, "Mining behavioral sequence constraints for classification," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

Using Foursquare data to reveal spatial and temporal patterns in London.

Maarten Vanhoof[1*] , **Antonia Godoy-Lorite**[1] , **Roberto Murcio**[1] , **Iacopo Iacopini**[2,1] , **Natalia Zdanowska**[1] , **Juste Raimbault**[3,1] , **Richard Milton**[1] , **Elsa Arcaute**[1] , **Mike Batty**[1]

[1]Centre for Advanced Spatial Analysis, University College London, London, UK
[2]School of Mathematical Sciences, Queen Mary University of London, London, UK
[3]Institut des Systèmes Complexes, National Centre for Scientific Research (CNRS), Paris, France

1 Introduction

Understanding the functional structure of contemporary cities can find many applications ranging from urban planning to local governance regarding inequality issues. This exploration is often constrained by the availability of large-scale data at fine spatio-temporal resolution. Longitudinal data, such as the Foursquare data, allow researchers to explore different aspects of the functional structure of cities, such as the role of places and the human mobility that occurs between them. However, never mind their size, this type of data is often incomplete, capturing only parts of the functional city due to the biases in the data collection that might represent a part of the population only.

In this paper we explore how Foursquare data can help unveil aspects of the functional structure of London. We explore the spatial pattern of venue locations and visit patterns captured in the Foursquare data. We also touch upon the temporal patterns of these data, as well as the interactions between categories of places (as derived by the semantic information that comes with the venues). Our findings reveal some of the characteristics and limitations of the Foursquare data with regard to investigating the functional structure of London and other cities.

2 Brief description of data and data handling

In this report we mostly focus our attention on London. We deployed both movement and venue data. Based on the category tree retrieved from the Foursquare api, we were able to classify the venues to higher level categories. It should be noted that, within this category tree, not all categories have the same depth of branches or number of leaves; and that venues in the data can be classified at different depths of the tree. As such, one can use this category tree only one-directional and it seems wise to evaluate at the highest level only. Those highest categories, and their number of available venues in the London data are in table 1.

To perform our analysis we used open source software, more specifically Python and GIS. If wished for, our codes can be made available on Github to wider audience.

Highest level of category	Number of venues
Food	7340
Nightlife Spot	3169
Shop and Service	3012
NaN	2929
Travel and Transport	2643
Professional and Other Places	1352
Outdoors and Recreation	1114
Arts and Entertainment	828
College and University	256
Event	37
Residence	9

Table 1: Number of venues in London per highest category in the Foursquare category tree.

3 Venue locations

By investigating the location of London venues, one can distinguish a spatial pattern in which more venues are observed in Central London compared to it's surrounding neighbourhoods (figure 1 top). One exception to this pattern is Heathrow Airport in the west, where a high number of Foursquare venues is observed despite its remote location from Central London. The map of the diversity of venues, as given by the entropy of observed categories per LSOA level (a low level administrative boundary in the UK), shows a pattern in line with the number of venues (figure 1 middle). Indeed, the higher diversity of venues in observed in Central London gradually fades out when moving away from it. It is known, however, that entropy values are related to volume and so when normalising entropy by the number of venues, the spatial pattern changes in an interesting way. In fact, after normalising the entropy by the number of venues (figure 1 bottom) one can observe that both London city centre and some remote areas around the City depict very low entropy scores. In Central London, this is due to the high amount of venues observed which brings down the normalised entropy to a large degree. Regarding the remote areas, this is due to a very low number of observations in the first place (as can also be observed in figure 1 top). At the same time, several areas away from Central London display a high normalised entropy, or thus a high diversity of venue categories, despite having a limited number of Foursquare venues.

[*] m.vanhoof@ucl.ac.uk

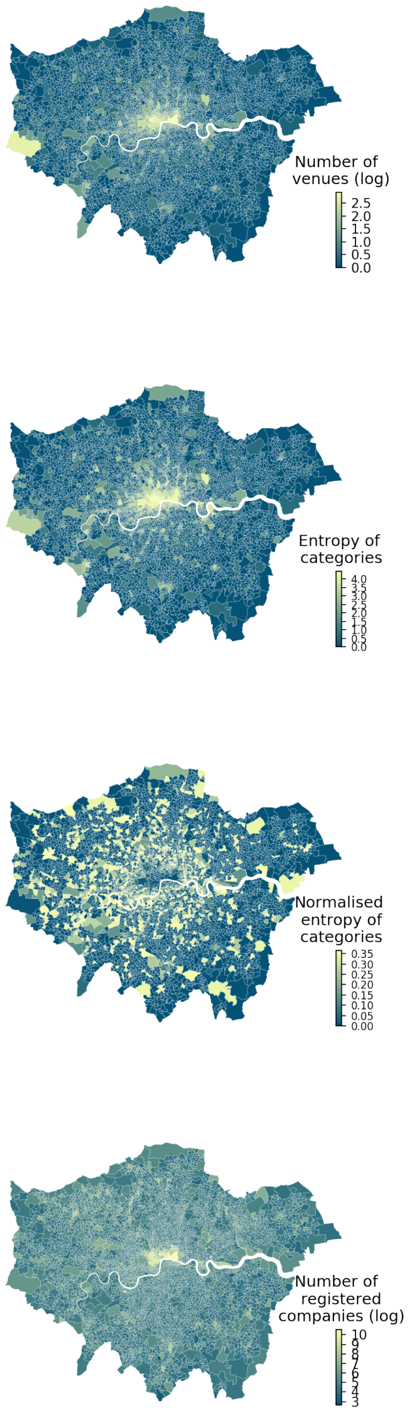


Figure 1: Number of Foursquare venues (top), entropy of Foursquare venue categories (middle top), normalised entropy of Foursquare venue categories (middle bottom), and number of registered companies in the Companies House dataset (bottom) for all LSOA administrative regions in London.

In a way, this finding stresses the importance of diversified functional activity in areas that are away from Central London. One outstanding question is whether limited number of venues in these areas correspond to a diminishing presence of such venues or simply by a diminishing popularity of these areas on the Foursquare platform. A quick comparison between the number of Foursquare venues and the number of registered companies in London -as can be found in United Kingdom's government registrar of companies, Companies House - does not offer a closing answer to this question. The spatial pattern of registered companies appears to be similar to the one of Foursquare venues (figure 1 top vs. bottom), and the distribution of the number of Foursquare venues per registered company for all LSOAs in London is narrow around 2 percent expressing differences between areas when it comes to Foursquare coverage (figure 2). The latter of course does not mean that the Foursquare venues cover companies of all types and sorts, but further investigation is needed in order to better understand this ”popularity bias”. Concluding, one could roughly say that there are three types of areas when it comes to Foursquare venues location in London. Areas that have a lot of venues and diversity (e.g. Central London), areas that have a limited but diversified number of venues (some of the out-of-centre areas) and areas that have little to no venues in the data (e.g. the South-East region of London). Although it is unclear what drives the presence of venues in the Foursquare data, there probably exists a ”popularity bias” which calls for a better understanding.

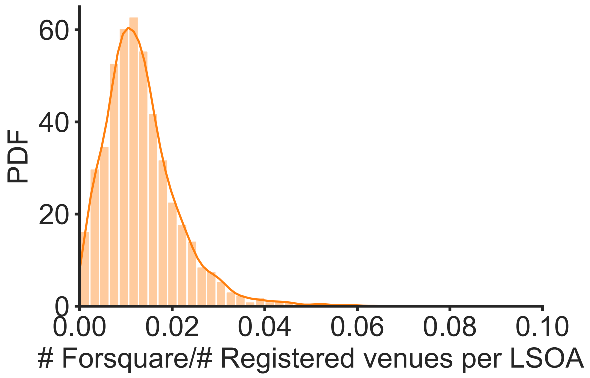


Figure 2: PDF for the number of Foursquare venues over number of registered companies per LSOA.

4 Spatial patterns of categories

Having investigated the diversity of venue categories in areas, the next logical question is whether spatial patterns per category differ much from one another, and if these spatial patterns might form a reflection of the popularity bias. Figure 3 shows the location of different venues according to their highest level category. Clearly, spatial patterns differ by category. For example, the outdoor and recreation pattern depicts a remarkable uniform distribution over the city as the category includes places such as parks or sporting grounds which are typically provided by local authorities to their inhabitants.

Venues in the food, and shops and services categories, are mostly located in Central London and around main roads or local centres. They reflect elements of accessibility and access to a market in order for these venues to function. Another patterns is that of colleges and universities, which are entirely grouped in Central London, expressing their higher-level functional role, their historical character, and their independence from local planning or market demand.

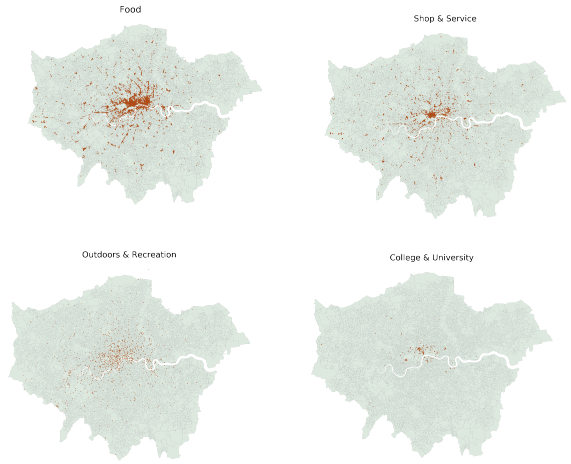


Figure 3: Spatial patterns of venues with different categories in London

5 Hotspots of movements

The movement data provided by Foursquare allows us to add a behavioural component to the location information of venues. To mine this potential, we perform a hot-spot analysis of the venue pairs that were most observed in the provided movement data. For London, we had access to a total of 7,650,994 pairs (Venue_i to Venue_j) that were subsequently visited by Foursquare users. We define our hot-spots on the number of check-ins that happened for each pair during the whole period the data was provided for. Focusing on pairs with more than 1000 check-ins renders a set of 101 pairs of venues (figure 4). As can be observed, most of them reflect micro-movements located in Central London and Heathrow Airport. Nevertheless, some of the pairs depict movements towards train stations or towards outside venues such as Croydon (an extensive night-time and shopping district) or Wimbledon (famous for its tennis competition).

Finally, we defined two types of venues from these hot-spots, Enablers and Receivers. An Enabler venue is drawn from the Venue_i subset and ranks high in frequency (the amount of months-different times combinations this venue was observed in the data) and check-ins. On the other hand, a Receiver venue also ranks high in these characteristics but is drawn form venue_j. The top ten Enables and receivers are shown in (figure 4). They are Heathrow airport, train stations, Harrods, Hyde Park, Marble Arch, and Piccadilly. From these ten, nine are at the same time Enablers/Receivers; St. Pancras is only enabler and Euston is only receiver.

Examining hot-spots for different periods of the day reveals some different patterns too. The morning check-ins, for example, are done mostly at train stations, whereas some areas like Oxford Street (a shopping area) are fundamentally midday locations and others, like Kensington are mostly afternoon/night areas. Some of these patterns can be investigated for an area of central London depicted in figure 5).

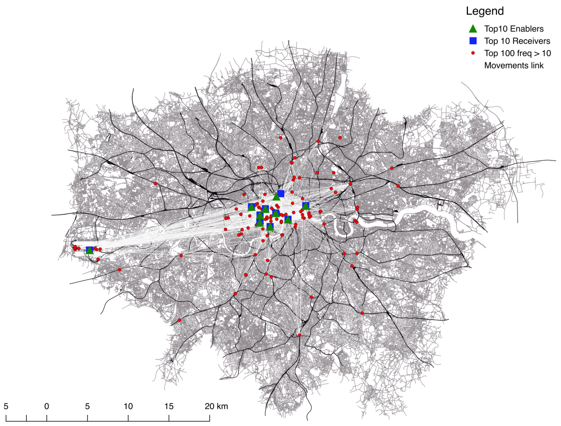


Figure 4: Hot-spots of pairs of venues in London based in the Foursquare movement data. Top 10 Enablers and Receivers are displayed as well

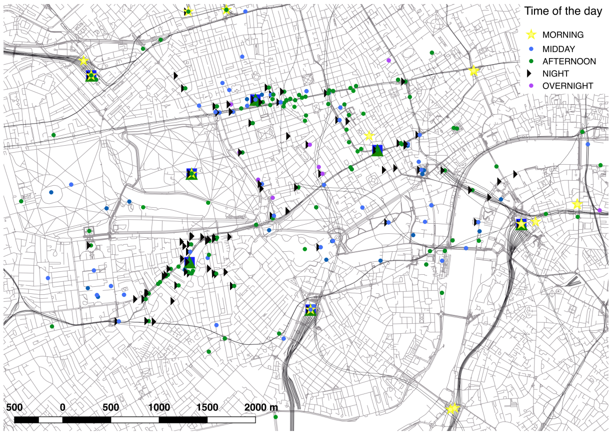


Figure 5: Hot-spots of pairs of venues in London based in the Foursquare movement data. Top 10 Enablers and Receivers are displayed as well

6 Transition matrix between categories

The analysis of movements between pairs of venues and their relation to the time of the days makes one wonder whether, at the very high level, one could detect temporal patterns between categories of venues. Are movements between transport venues and universities more likely to occur during daytime? Is there a seasonal pattern that can be distinguished for the movement between outdoor locations and venues in the category food?

To investigate this, we constructed transition matrices between all high-level categories based on the movement data. In the case of London, most movements occurred between venues in the categories of food, travel & transport, and shop & service. Interestingly, these high-level transition matrices are not influenced by the different seasons in a year (figure 6). And they were only influenced to a limited degree by the time of the day as can be observed in figure 7. In London, for example, interactions with the food category are limited in the morning, but pop up throughout the day and peak in the night period. We found the limited sensitivity to seasons and time of the day to be the case for London, but also for the different other cities available in the Foursquare data such as Istanbul, Tokyo, or Jakarta.

What did reveal though, is that the transition matrices between different cities differ as can be observed in figure 8. In most cities the food, travel & transport, and shop & service remain the major categories, but the degree of their interaction differs. In Istanbul for example, food is by far the main category while travel & transport are almost non existent. The inverse is observed in Tokyo where travel & transport is by far the main category and food is almost non existent. Interestingly, the importance of other categories is different between cities. In Istanbul outdoor& recreation places are rather prominent, while in Jakarta professional places pop up more than in any other city. Without further investigation it is difficult to detect the reasons for these differences, but they could be: a different usage of Foursquare in cities, a different amount of venues from different categories in the cities, and a different functional composition of the cities. Each city thus has its own specific pattern of interactions between venue categories, enabling and limiting the study of their functional structure in different ways.

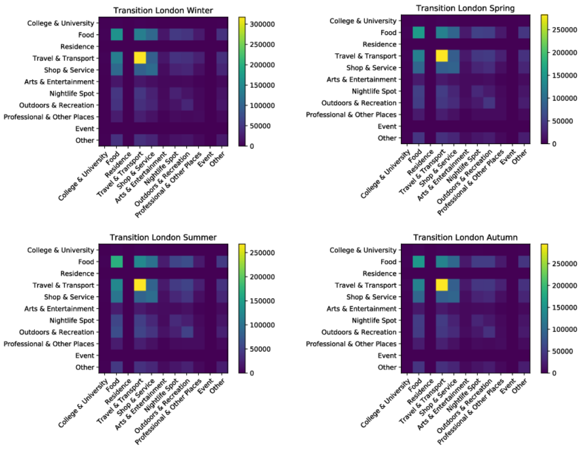


Figure 6: Transition matrices between the different high-level categories of venues for the different seasons in the Foursquare data for London

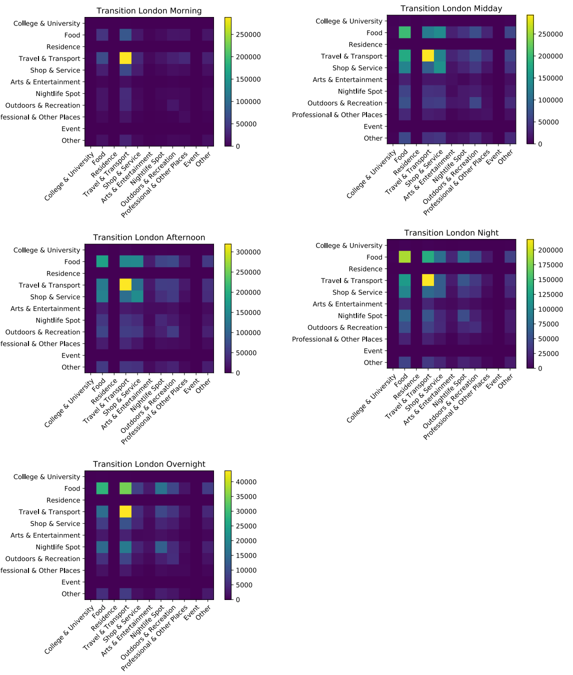


Figure 7: Transition matrices between the different high-level categories of venues for different times of the day in the Foursquare data for London

7 Small conclusion

In this report, we have briefly discussed the foursquare data and their potential to offer a view on the functional structures on contemporary cities like London. Exploring the foursquare venue locations we touched upon the idea of understanding functions and the diversification of function within different city areas to better understand their role in the city as a whole. The spatial patterns of individual categories serve as a good for this example. Nevertheless we reckon that, given the way the foursquare data is collected, some sort of popularity bias exists within the data. Validation with external databases is necessary to better understand this bias, and so we touched upon such validation by comparing the number of foursquare venues in London with the number of firms in London based on the xxxx database on firms in London. Combining the location, functional aspect of the venues with the movement data provided by users interaction with venues we added a behavioural component to our analysis. The analysis of hot-spot of movement pairs allowed us to better understand popular movement patterns in London. In general, we found hot-spots of movement often to be related to tourism movements (airport-center shuttle, Harrods) and places of transport (the many train stations in London). A closer look at places that serve as enablers and receivers, we note that throughout different periods of the day, different places become more important. Functional places, in other words, are depending on the rhythm of the day. This might not be surprising, but when studied in more depth, it will reveal some interesting information on the functioning of places in the city.

Finally, we also investigated the sensitivity of larger patterns of functional interactions to time. More specifically, we wondered whether interactions, as described by the pairs of venues, between high-level categories would differ significantly at city scale over the time periods of season or even different times of the days. For London, we found these high level interactions, as expressed by the transition matrices, to be little to not influenced by time periods. We found a similar insensitivity for other cities in the data, such as New York or Jakarta. We did, however, find differences in the general structure of the high level transition matrices between cities. The reasons for these differences can only be hypothesised, but could relate to different city structures or usage of Foursquare, meaning that the interpretation of functional structures in cities based on foursquare data might be city dependent.

8 Future work

Some ideas for further work we came up with are:

- Elaborate an investigation of the popularity bias by validation foursquare venues with a register of existing companies
- Follow check-ins through day. After aggregating by time of the day the hotspot locations, record Venue$_i$-Venue$_j$ time (let's say Morning). Then look at all Venue$_j$-Venue$_k$ pairs and keep those with time=Midday; then took $Venue_k$-$Venue_m$ pairs and keep pairs with time=Afternoon and so on. This only works if we assume that check-ins happen the same day and only for walkable location, like from Harrods to Hyde Park to Marble Arch or follow locations from train stations.
- Use the patterns of the high-level transition matrices to identify similarities and differences between cities worldwide (unsupervised learning). Investigate whether similarities relate to i) types of Foursquare usage, ii) functional structure of the city and 3) the morphological patterns of the cities.

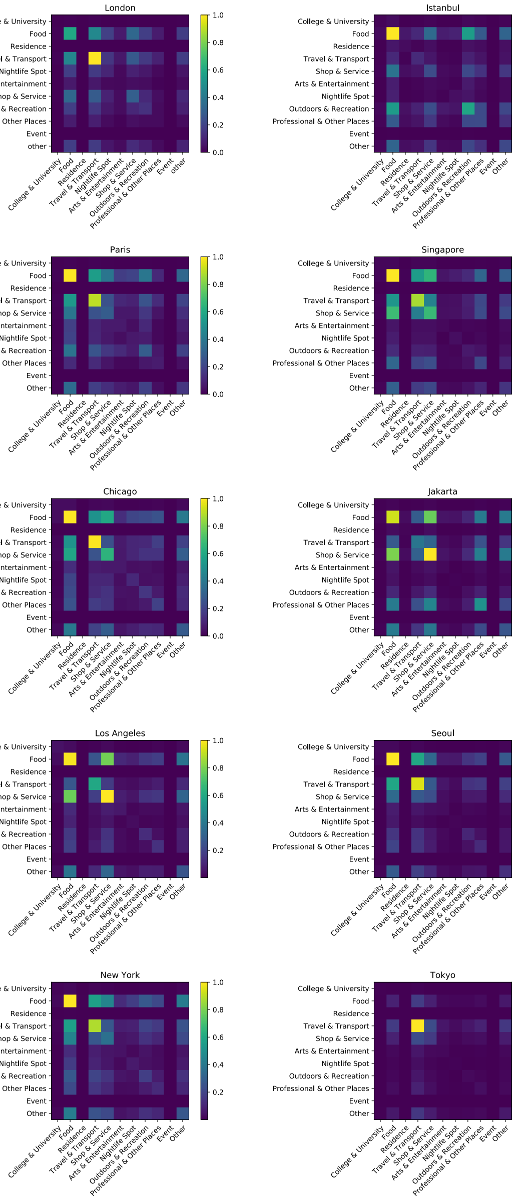


Figure 8: Transition matrices between the different high-level categories of venues for different cities based on the foursquare data

A Geographic Data Science Framework for the Functional and Contextual Analysis of Human Dynamics within Global Cities

Alex Singleton[*] , **Daniel Arribas-Bel** , **Sam Comber** , **Alessia Calafiore** and **Gregory Palmer**

Geographic Data Science Lab, Department of Geography and Planning
University of Liverpool
alex.singleton@liverpool.ac.uk

Abstract

This study implements a Geographic Data Science framework that transforms the Foursquare check-in locations and user behaviour data into a series of global urban networks. Through an innovative spatially weighted community detection algorithm we uncover functional regions for 9 global cities; and through linkage to locally derived catchments for around 330k check-in locations; attribute regions with a range of measures that describe their diversity, morphology and mobility. We then applied k-means clustering to derive a global set of 4 functional region types, enabling the comparison of Human Dynamics and their contexts across the sample of global cities.

1 Introduction

Networks are an increasingly important conceptual and methodological tool in contemporary urban theory to both represent and model various types of interaction, flow or relation (e.g. movement, finance, communication, friendships etc), or to elucidate hidden structure manifest through the agglomeration of human interactions [3, 4]. Many applications of networks focus on human dynamics across a range of temporal scales: from mapping patterns of global migration to daily commuting patterns [2]. However, developing an understanding of contemporary human mobility, behaviour, context and outcome poses great challenge to many existing instruments that urban scholars have traditionally relied upon for the empirical study of cities (e.g. Census, Survey etc) [5].

However, new opportunities to explore the form and function of cities have been enabled by the volunteering of spatio-temporally referenced digital traces left by human use of urban environments [2]. In their raw form, data such as Foursquare check-ins provide insight into both where and when activities are taking place within cities; but as we demonstrate in this short study, through their linkage, augmentation and analysis: also provide great opportunity to model the functional form of cities.

The universality of our implemented Geographic Data Science framework enables transfer across multiple countries to: 1) Define functional regions within cities; 2) Define check-in POI (POI) catchments; 3) Create POI catchment contextual measures related to urban and functional structure; 4) The mapping of diversity of functional and contextual geography.

Through these measures we provide insight into human dynamics within and between sets of defined functional regions across: Chicago, Istanbul, Jakarta, London, Los Angeles, New York, Paris, Seoul, Singapore and Tokyo.

2 Urban Networks & Functional Regions

The objective of this analysis was to provide insight into the emergent functional structure of different cities created through the ordering of human activities into cohesive zones of interconnected activity. In this analysis we create global urban networks of user interactions and implement a geographically weighted community detection technique to identify functional agglomerations of POI.

2.1 Creating and Cleaning the Network

For each city, a network was created using Python's NetworkX library, with nodes associated with each POI and connections between them determined by the agglomeration of all individual's sequences of Foursquare check-in activities. Edges were therefore ascribed the total count of user check-in links. Cities were isolated into separate graphs through the removal of edges that spanned cities; and further edges removed that existed between non-travel related POI categories that exceed what one would consider to be 'a walking distance'. Hypothesizing that these check-ins are a result of the app not registering the use of transportation, we focus only on localized movements, excluding all edges that exceed the median distance value between POIs for each city. Although in this instance we were interested in aggregate functional regions within cities (all activities), this technique could be extended to different temporal bins to ascertain how connectivity between different areas of the city varies by time of day, day of the week or even month of the year.

2.2 Geographically Weighted Community Detection

Although recent work within Urban Geography has demonstrated the utility of community detection techniques in the definition of functional regions [3], these do not explicitly account for geography, which is a primary driver or constraint to interaction.

Broadly speaking, interactions are more likely between POI in "close" proximity; although interactions will vary between urban contexts. As such, we first needed to examine the distances that Foursquare users travelled between check-in loca-

[*] Contact Author

tions. For simplicity, we measured the euclidean distance between each POI for every user; these were then aggregated into distance bins. Figure 1 shows how there is a distance decay in travel between check-ins, but, for some cities (e.g. London and Paris) there is complexity to these general patterns, with lower initial propensity to travel to POI in close proximity.

Community detection was implemented using Python's *community* package's *best_partition* method that searches for network partitions by maximizing the modularity using the Louvain heuristics. However, to better account for geography, we first weighted the network edges by the inverse of geographic distances between their associated nodes; thus prioritizing the importance of more spatially proximate flows. The outcome was a set of functional regions for each of the cities (See Table 1 and Figure 2). Polygon boundaries were created from POI locations by applying Alpha Shapes to the associated POI locations within each identified community.

City	Frequency of Regions	μ POI	μ Distance	σ Distance
Istanbul	81	1344.6	1.497	1.182
Paris	20	643.6	0.627	0.446
Seoul	7	1901.2	0.771	0.670
Singapore	10	2102.6	0.794	0.740
Tokyo	26	2133.5	0.548	0.508
London	20	1069.1	0.659	0.494
Los Angeles	8	1708.8	1.508	1.152
Jakarta	18	1101.6	1.055	0.853
New York	16	1918.5	0.579	0.391
Chicago	7	1656.0	0.850	0.638

Table 1: The frequency of regions identified for each city, mean (μ) POI per community, alongside the mean (μ) and standard deviation (σ) of the edge distance in KM.

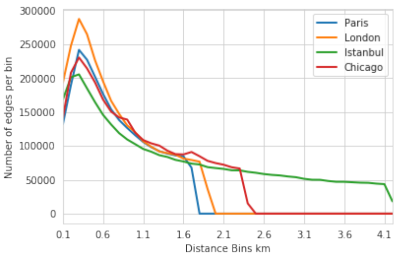
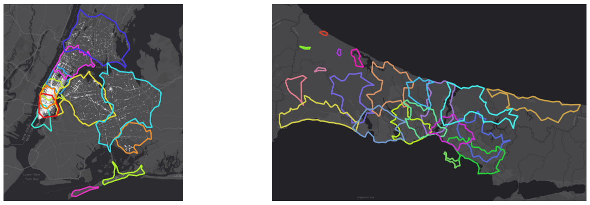


Figure 1: The proportion of check-ins by distance.

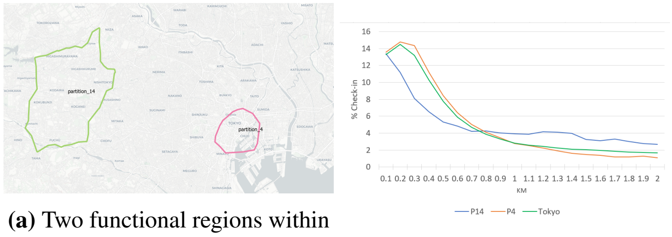


(a) New York, USA.

(b) Istanbul, Turkey.

Figure 2: Functional regions within Istanbul and New York with POI shown as white dots (Note; POI data for Brooklyn was not provided).

Each functional region represents a distinctive cluster of aggregate Foursquare user activity. Prior to the contextualization of these areas, we first examined how the aggregate user check-in trajectory profile varied from the associated city average. To illustrate this we compare two clusters within Tokyo, Japan in Figure 3. The area labelled *partition_4* represents the core of central Tokyo and has high POI density, and more compact structure; whereas *partition_14* is more peripheral, and has lower POI density. Sub-figure 3b illustrates how *partition_14* has a more distinctive set of trip behaviour relative to *partition_4* which is more similar to the city average.



(a) Two functional regions within Tokyo, Japan; highlighting a central area, and one towards the urban periphery.

(b) The proportion of check-ins between two functional regions, and the Tokyo Average.

Figure 3: Functional regions within Tokyo.

3 Contextualizing POI Locations

POI linkage through user interaction and their geographic location are key drivers of those spatially located functional regions demonstrated in the previous section; however, at a local area level patterns of use are both driven by POI type (e.g. travel, food etc); and other measures of the wider built environment. There is much literature within urban planning and architecture that explores how the morphology of the built environment (e.g. street geometry) may influence activity within places and limit or enhance attraction between locations[4]. The objective here was therefore to capture a range of contextual measures for each POI, that would be later used to compare the functional regions.

3.1 Defining a Catchment and Input Variables

The first stage was to define a "catchment" for each of the 330k POI. A polygon delineating the bounding box of all POIs within each city was used to create a NetworkX graph from OpenStreetMap (OSM) data using the OSMnx library [1]. From this city-wide graph of the street network, a sub graph was constructed that was centered at the latitude and longitude of each POI. The radius of each sub graph was extended according to a ten minute walking distance from the POI. A convex hull of the sub graph was then extracted for each POI which are the catchments we use in the subsequent analysis (Figure 4).

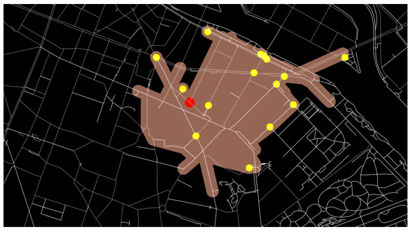


Figure 4: A Catchment showing a ten minute walking catchment along the street network in Paris from a selected POI (red), with other POI within the area highlighted in yellow.

Within each POI catchment area, we defined a series of measures using either OSMnx [1] or spatial analysis of the catchment area characteristics; these are summarized in Table 2.

Measure	Description	Label
Average Circuity	Total edge length divided by sum of great circle distance between nodes incident to each edge.	circuity_avg
k Average	Average degree of catchment.	k_avg
n	Number of nodes in catchment	n
Node Density Km	n divided by area in square kilometers.	node_density_km
Self-loop Proportion	Proportion of edges in catchment that have a single incident node.	self_loop_proportion
Street Length Average	Mean edge length in undirected representation of catchment (meters).	street_length_avg
Street Length Total	Sum of edge lengths in undirected representation of catchment.	street_length_total
Street Per Node Average	Average number of streets per node.	street_per_node_avg
Eigenvector Centrality Ratio	Ratio of eigenvector centrality between POI and other POIs within catchment.	ratio_eig
Percent Same Type	Percentage of POIs of same category inside catchment.	percent_same_type
POI Count	Number of other POIs inside catchment.	n_pois
Average Afternoon	Average Foursquare check-ins of all POIs in catchment between 15:00 and 18:59 (%).	avg_afternoon
Average Midday	Average Foursquare check-ins of all POIs in catchment between 10:00 and 14:59 (%).	avg_midday
Average Morning	Average Foursquare check-ins of all POIs in catchment between 06:00 and 09:59 (%).	avg_morning
Average Night	Average Foursquare check-ins of all POIs in catchment between 19:00 and 23:59 (%).	avg_night
Average Overnight	Average Foursquare check-ins of all POIs in catchment between 00:00 and 05:59 (%).	avg_overnight

Table 2: Initial measures used to compare POI contexts.

Measure	Description
Diversity	Entropy score (Shannon Index) measuring diversity in the categories of POIs
Morphology and Behaviour	The aggregation (mean) of POIs information resulted from the catchment area (CA) morphological profiles and average check-ins by time of day
Mobility	Functional region distance decay function

Table 3: Functional region contextual differentiation.

4 Mapping the Diversity of Functional Regions

To gain an understanding of the functional regions similarities and differences we developed a multidimensional comparison implementing k-means clustering. We posited three drivers of functional region differentiation, which are shown in Table 3.

4.1 Diversity

Entropy was calculated as follows:

$$H = -\sum_{i=1}^{s} p_i ln p_i \tag{1}$$

where s is the number of categories, p_i is the proportion of POIs of each category and ln is the natural log.

4.2 Morphology and Behaviour

The variables calculated for each POI catchment were aggregated into the functional regions. The variables describing the proportion of check-ins by time of day (morning, midday,...) were averaged over those POI within each region. Additionally, the means of the variables(n_pois, percent_same_type, ratio_eig, streets_per_node_avg, street_length_total, street_length_avg, self_loop_proportion, circuity_avg, k_avg, n, node_density_km) characterizing catchment morphology were reduced through application of principal component analysis. Four principal components that explained 77% of the variance derived (Fig 5).

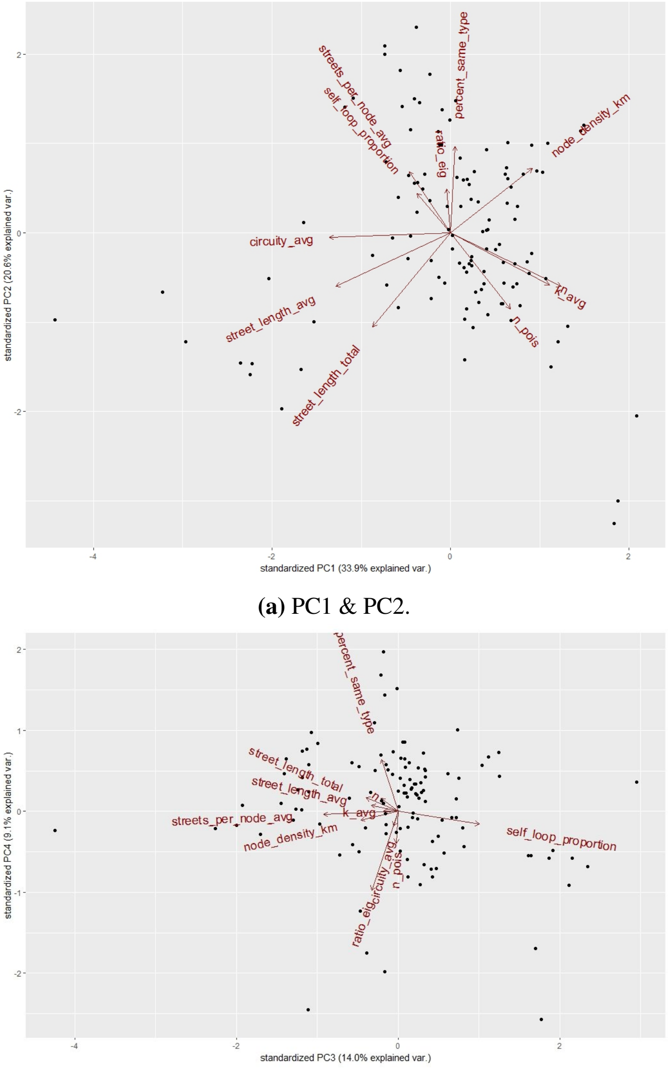
4.3 Mobility

We defined the aggregate regional mobility mathematically through the parameter exponential decay function $f(x)$ in the Equation below, where x is a vector of distances between POI. We then used TensorFlow to obtain the best fit for each of the city's empirical observations y, optimizing parameters c and k by minimizing the $\mathcal{L}_2$ loss between y and $f(x)$. This gave a set of exponential decay parameters that estimated likely relative impedance between POI locations for given regions.

$$f(x) = max_i y_i \times c \times e^{k(max_i x_i - x / max_i x_i)} \tag{2}$$

4.4 Comparing Multidimensional Characteristics

A challenge when implementing k-means clustering is to select an appropriate number of clusters. Here we utilize a clustergram plot which along the x axis plots a range of potential k values; and along the Y, a weighted mean principal component score. Each line relates to a functional region and relative score for each iteration of k; and dots represent the cluster average PCA scores. An indication of the best fit for k relates to a model where these centroids are well separated (See Figure 6). From this figure, we selected 4 clusters as an appropriate k value.



(**a**) PC1 & PC2.

(**b**) PC3 & PC4.

Figure 5: Vars contributing to PC1 & PC2 (a) and PC3 & PC4 (b).

As such, for k=4, we ran k-means 10,000 times (as it is stochastic), extracting the result that had the lowest total within

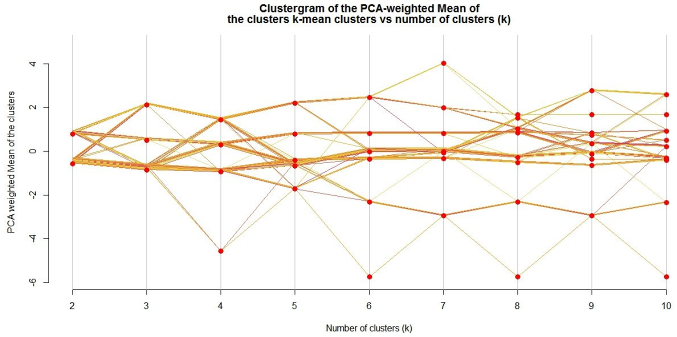


Figure 6: A Clustergram showing how different iterations of K impacted cluster separation.

	1	2	3	4
Istanbul	0.0	27.8	0.8	71.3
Paris	0.0	98.5	0.6	0.8
Seoul	77.3	22.7	0.0	0.0
Singapore	0.0	0.0	0.2	99.8
Tokyo	99.9	0.0	0.1	0.0
London	0.0	91.6	5.4	3.0
Los Angeles	0.0	76.1	12.8	11.1
Jakarta	0.0	47.5	0.0	52.5
New York	0.0	98.1	0.1	1.8
Chicago	0.0	96.3	0.0	3.7

Figure 7: The % of POI within functional region clusters by city.

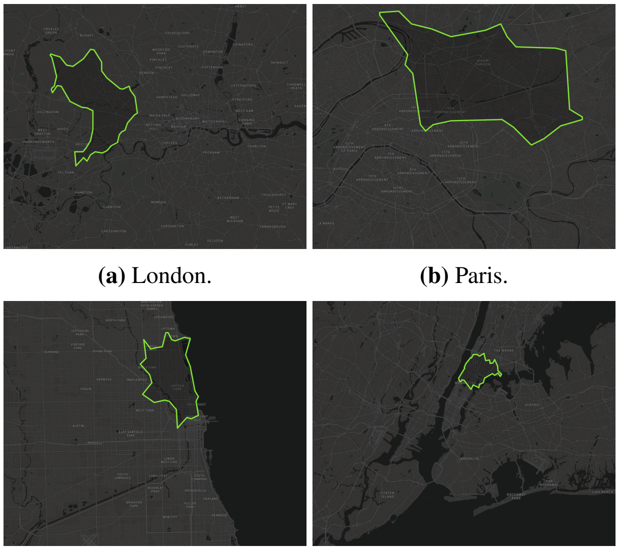
cluster sum of squares; ie, most compact clusters. After clustering, we appended the original data back onto the clusters and examined these distributions relative to the global averages.

- C1: has low diversity (-0.82), the highest value of night check-ins (0.95), and low PC3 (which seems to be mostly driven by the proportions of self loops - see Figure 6)
- C2: has moderately high diversity (0.3), check-ins followed typical peak hours (morning and afternoon), adopts an almost linear decay with regards to the travel distance between venues (high c and low k). This is the most numerous cluster.
- C3: has the lowest diversity and check-ins at night and overnight, while the highest avg of check-ins at midday. PC1 which seems to be driven by node density is very low (these clusters are mostly peripheral areas of the city)
- C4: has the highest diversity (0.65) and overnight check-ins.

4.5 Human Dynamics within Global Cities

The frequency of POI within each of the functional region clusters is shown in Figure 7. No city other than Istanbul (which has a very high POI frequency) are ascribed all clusters, however these distributions do offer insight into the compatibility of these locations in terms of human dynamics and contexts. For example, New York, Chicago, London and Paris share similarly distributed clusters; whereas other cities have more unique distributions such as Singapore and Tokyo.

The utility of these distributions can be illustrated by plotting archetypal functional regions across comparable cities. We do this by examining the input scores for each region, and selecting those that are are closest to the average for a particular cluster (see Figure 8). We would expect these clusters to have similar characteristics in terms of POI diversity, urban morphology and user check-in behaviour.



(a) London. **(b)** Paris.

(c) Chicago. **(d)** New York.

Figure 8: Archetypal functional regions (Cluster 2).

5 Conclusions

This report has outlined how social media data such as Foursquare can be utilized to provide insight into the structure and function of cities. Through our innovative Geographic Data Science methodology we create a framework for identification and description of functional regions across global contexts, linking a range of measures ascribed down to the local level. There are numerous avenues for future research: for example, the exploration of potential modes of transit linking POI sequences, although, this may require more granular time stamps, or the use of modeled distance decay as integral models that may predict future system states.

References

[1] Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65: 126–139, 2017.

[2] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, Ronald A Peterson, Hong Lu, Xiao Zheng, Mirco Musolesi, Gahng-Seop Ahn, et al. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4): 12–21, 2008.

[3] Garrett Dash Nelson and Alasdair Rae. An economic geography of the united states: From commutes to megaregions. *PloS one*, 11(11):1–23, 2016.

[4] Carlo Ratti. Space syntax: some inconsistencies. *Environment and Planning B: Planning and Design*, 31(4):487–499, 2004.

[5] Alex Singleton and Daniel Arribas-Bel. Geographic data science. *Geographical Analysis*, 2019.

An Intelligent Tree Planning Approach Using Location-based Social Networks Data

J. H. van Staalduinen[1]
LIACS – Leiden University
Leiden, The Netherlands
j.h.van.staalduinen@umail.leidenuniv.nl

J. Tetteroo[1]
LIACS – Leiden University
Leiden, The Netherlands
j.tetteroo@umail.leidenuniv.nl

D. Gawehns
LIACS – Leiden University
Leiden, The Netherlands
d.gawehns@liacs.leidenuniv.nl

M. Baratchi
LIACS – Leiden University
Leiden, The Netherlands
m.baratchi@liacs.leidenuniv.nl

***Abstract*—How do we make sure that all citizens in a city can enjoy the necessary amount of green space? While an increasing part of the world's population lives in urban areas, contact with nature remains important for the human well-being. As optional tree planting sites and resources are limited, the best site to plant must be determined. Can we locate these sites based on the popularity of nearby venues? How can we detect groups of people who tend to spend time in tree deprived areas?**

Currently, tree location sites are chosen based on criteria from spatial-visual, physical and biological, and functional categories. As these criteria do not give any insights into the amount of people benefiting from the tree placement, we propose a new criterion taking socio-cultural aspects into account. We combine the Foursquare mobility data set with a tree location data set, both of New York as a case study. Using the Foursquare data set we create a venue interaction network from which we extract venue communities. These communities are then scored based on the amount of trees in the vicinity of their venues. By combining the popularity of venues with the tree density of venue communities we can identify locations where planting a tree can benefit the most number of people and make the largest impact.

***Index Terms*—Urban computing, tree planning, social networks**

I. Introduction

As of 2018, 55% of the world's population lives in urban areas, a number which is projected to grow to 68% by 2050 [4]. The North-American continent stands out in particular, where this number is already at 82%. While it is easy to point out the economical reasons for moving to the city – at least at the first sight [7] – there are certainly downsides attached to urban life. One of them is the inescapable fact that cities, by definition [6], have a higher population density, leading to more built-up areas and thus a scarcer supply of nature than in rural areas. However, as Rohde and Kendle put it, "it is obvious from any casual observation that many human beings do not like to be dissociated from the natural world; as a nation we spend millions of pounds every year on garden and household plants" [15]. Indeed, contact with nature does seem to be linked to human well-being and positive emotional effects and is even said to strengthen urban communities [10, 13]. Apart from socio-cultural benefits, urban greenery can help to mitigate two characteristically urban problems: air pollution due to traffic [11] and (extreme) warmth due to the urban heat island effect [12]. The inclusion of parks and street trees in city landscapes is, therefore, an important aspect of the urban planning process.

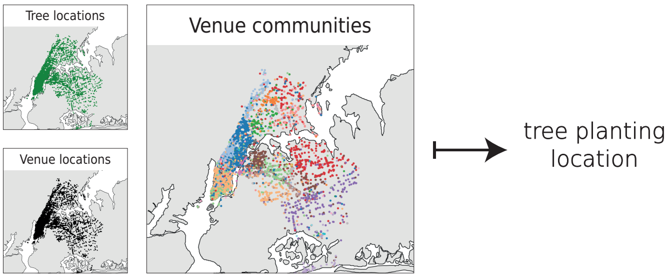


Fig. 1: We combine three types of data (tree locations, venue locations, venue communities) to determine a new criterion which can be used in selecting potential tree planting sites.

To date, socio-cultural arguments play a marginal if not non-existent role in formal frameworks describing criteria for selecting potential tree planting sites. The criteria in these frameworks do not account for the amount of people that are accommodated by the newly planted trees. When following the established criteria, trees may end up in places where they are beneficial to some people, but its effects may not serve the majority of people, or may never reach the people yearning for them most.

To tackle this problem, we propose an additional tree planning criterion. Prioritization should be given to sites visited by many people and specifically people who tend to move between areas lacking trees.

We identify such locations by combining two ways of analyzing the structure of a venue interaction network. By combining the knowledge about venue popularity and venue communities with a low tree density, we can detect popular venues within tree deprived communities and thus provide a prioritization that can be used for site selection in the tree planning process, as schematically shown in Fig. 1. This prioritization can be embedded within the criteria of established tree planning frameworks that currently lack this socio-cultural value and insight.

Our paper has the following contributions:

- We describe a novel criterion for potential tree planting

[1] Both authors contributed equally to this work.

site selection based on network communities within a venue interaction network;
- We combine this criterion with venue popularity, based on network analysis of venue interaction data from the social media platform Foursquare;
- We apply this method to prioritize venues as potential tree planting sites in New York City.

II. Related work

Most of the work in the field of tree planning revolves around selecting appropriate tree species for predetermined planting sites [17, 18]. This reflects the observations by Spellerberg [18] and Pauleit [14] that tree planning is often – or at least has been for some time – an afterthought in the urban design process and characterised by pragmatism. According to an Australian survey, while visual aesthetic of trees and socio-cultural function of green spaces in the city seem to be important motives for planting trees, the first motive only plays a small role in the tree planning process [16] and the second motive is not reflected in the sparse body of site selection criteria that we could find. The work by Amir and Misgav [2], in which they aim to describe a complete tree planning decision framework, does incorporate criteria on site selection. They define three useful criterion categories, which are *spatial-visual*, *physical and biological* and *functional*. Criteria relating to the socio-cultural function of green spaces however, are missing. We observed several works describing site selection criteria [8, 14], but those fall within the category of *physical and biological* criteria that are essential for the survival of the tree. Moriani [11] did use population density in their planting priority index, but as they focused on the air pollution-reducing quality of trees, this still falls within the category of *functional* criteria. We believe then, that the body of site selection criteria is still incomplete and that we can contribute to this framework by introducing a new socio-cultural criterion which takes people movement into account.

III. Methods

A. Venue popularity

A naive approach to maximize the impact of planting a tree, is planting it near a place where many people go. To find this place we compute the degree of all nodes in the undirected network graph $G = (V, E, W)$, where nodes $v \in V$ are venues and edges $e = (v_1, v_2), e \in E$ movements of people between two venues v_1 and v_2, with weight $w_e \in W$ as the number of movements between the pair of venues. The degree of a node v is then defined as the sum of the weights of the edges that are connected to it:

$$\deg(v) = \sum_{e \in \{(u,v) | u \in \text{adj}(v)\}} w_e \tag{1}$$

B. Venue community tree density

Although trees near popular venues may reach many people, they may not reach groups of people who tend to visit other venues. It may be the case that some people never come across arboreal areas. To deal with this shortcoming of the naive approach, we introduce a measure we call the *tree density coefficient*, which is based on communities in the network. A community is a group of nodes which is densely connected with each other, but much less with the rest of the network [5]. By looking at these communities, we use the fact that it is not necessarily bad for a venue not to be covered in trees, if people often move from that venue to a venue that is covered. To detect the communities, we use the Louvain community detection algorithm [3]: a fast algorithm able to find communities with high quality. It is based on the optimization of modularity, a measure that compares the density of connections within a community with the density between communities.

As it is computationally heavy to compute the modularity of a community, the Louvain algorithm uses heuristics to approximate it. Therefore, it does not necessarily return the best community layout. In order to gain confidence in the robustness of our communities we choose to run the algorithm many times to create a large number of community layouts.

To compute the tree density coefficient for a venue, we first count trees in the vicinity of the venues. We approximate this vicinity by creating a grid of the city, where each grid cell is 50 by 50 meters, calculated using Universal Transverse Mercator coordinate system [9]. Each venue v_i is mapped to a cell in the grid and is assigned the number of trees in the cell as its *venue tree density vtd*$_i$.

We compute the *community tree density ctd*$_i$ for a venue v_i by averaging the *vtd*$_i$ with the venue tree densities of all the other venues in its community C_i, over multiple iterations k of the community detection algorithm:

$$\text{ctd}_i^k = \frac{1}{|C_i|} \sum_{v_j \in C_i} \text{vtd}_j, \quad 0 < k \leq k_{\text{max}}. \tag{2}$$

In the end, the *tree density coefficient* c_i for a venue v_i is its average community tree density value over all iterations of the community detection algorithm:

$$c_i = \frac{1}{k_{\text{max}}} \sum_{t=1}^{k_{\text{max}}} \text{ctd}_i^k. \tag{3}$$

C. Combined method

A venue with a low tree density coefficient could have only one visitor, whereas other venues in the same community that have a similarly low coefficient could have many visitors. In this case, the latter venue(s) would be more appropriate as a tree planting site.

We extend the community based density coefficients with venue degrees by combining the two measures and detecting the set of venues that are Pareto efficient, i.e. the venues that are found by minimizing the tree density coefficient and maximizing the influence of the venue: the optimal trade-offs between the two measures. Also called the Pareto frontier, the venues in this set meet our criterion of helping most people needing trees. Tree planners could choose any of the venues along the Pareto frontier, depending on their preference towards either of the two measures.

IV. Case study

A. City of choice: New York

We conducted a case study to investigate the implementation and workings of our criterion using real data. For this we chose to focus on New York City as data on both venue interactions and tree locations was richly available.

We used two data sets to construct our criterion. We used venue interaction data of New York, provided by Foursquare as part of the Future Cities Challenge 2019, to create the venue interaction network. To assign tree density scores, we used a Street Tree Census data set [1]. This section describes the properties of both data sets and how we processed them to implement our methods.

B. Venue interaction data

The Foursquare venue interaction data set comprises of two parts: venues and movements between them. The data set contains information on ten different cities around the world. As we focused on New York in this case study, we used the New York data, but it should be noted this study is applicable to any of the other cities, provided we have access to a corresponding tree location data set.

As not all venues found in the movement data occur in the venue information data, we considered only the venues with known locations for the construction of the network. Additionally, we omitted all 86 venues not connected with the big component as the small components that are not connected never exceed a size of 3 nodes. In the end, we were able to use 15,803 venues in our analysis.

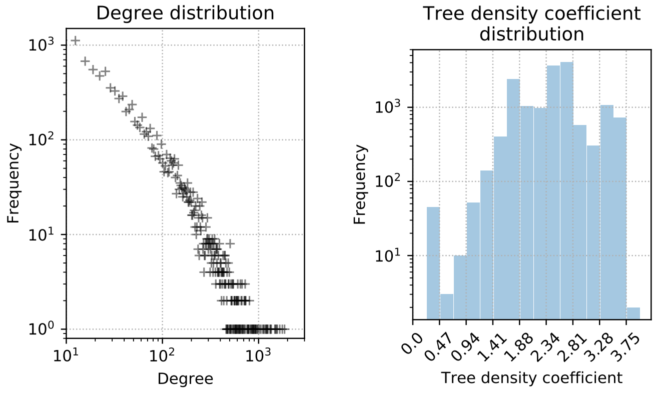
C. Street Tree Census

The Tree Census data set contains information on street trees in New York City and surrounding cities. It contains information on among others the *species*, *health*, as well as *longitude* and *latitude*. Only street trees were counted, which means that trees in parks were not taken into account and are not present in the data set.

V. Results

A. Venue popularity

We computed the venue popularity as the degree of each node and observed that the distribution follows a power law (see Fig. 2a), as is generally the case in scale-free networks modeling natural phenomena. To decide which venues would be interesting as a tree planting site according to this method, one should prioritize venues with higher degrees.



(a)

(b)

Fig. 2: The power law distribution of venue degrees (a) and distribution of tree density coefficients (b).

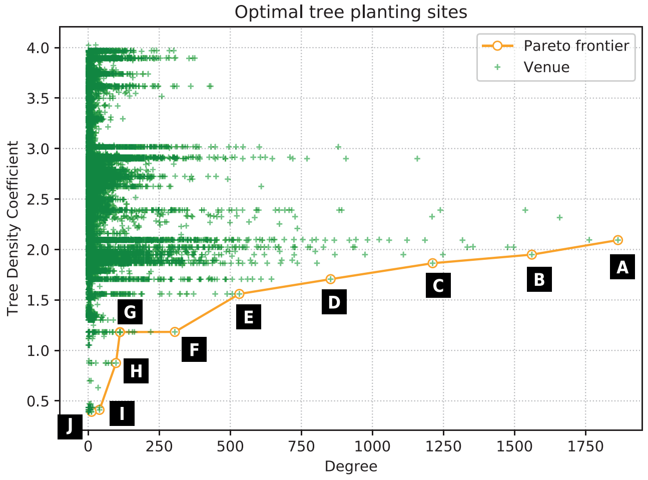
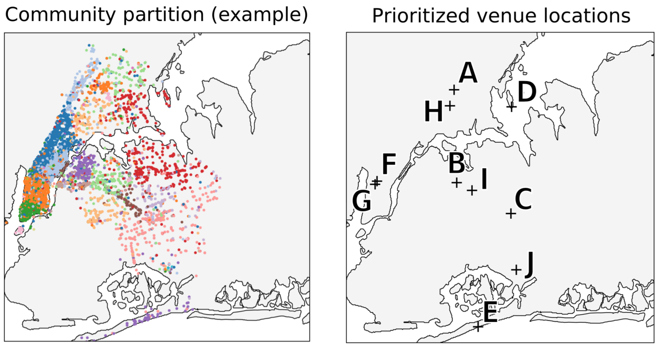


Fig. 3: The distribution of venues according to degree and tree density coefficient. The Pareto frontier shows the venues with the optimal tree planting location according to our criterion. (Venue labels correspond with Fig. 4b and Table Appendix-I)



(a) One of the 1,000 community partitions.

(b) Optimal tree planting locations (see Appendix–Table I).

Fig. 4: Map of New York City showing the optimal tree planting locations based on community structures.

B. Venue community tree density

We used the Louvain community detection algorithm as implemented in the Python `NetworkX` package. We set the resolution to 0.5 to find decently small communities. One such community lay-out is shown in Fig. 4a.

As the communities are detected using the heuristic Louvain algorithm, we averaged the community tree density of the venues over 1,000 runs of the algorithm, each time possibly detecting slightly different communities in the network, to obtain their tree density coefficients.

To find tree-deprived communities, we combined the locations of the venues within the communities with the tree locations in the street tree data set. First, we calculated the tree density for each venue. Then, the average tree density of the venues in the community was computed and returned to each of those venues as its community tree density.

We show the distribution of the tree density coefficient values in Fig. 2b. The distribution is slightly skewed to the right, which means most communities are filled with trees. Some, however, would still benefit from planting more. Prioritization for tree planting sites using this method should be given to the venues with the lowest coefficients.

C. Combined method

In order to select the most impactful planting locations, we combined both methods. This results in the distribution of venues and associated Pareto frontier as shown in Fig. 3. Here we minimize the tree density coefficient of the venues while maximizing their degree. These venues are highlighted by the Pareto frontier and should be prioritized according to our new criterion. To indicate the locations of the venues on the Pareto frontier, we show the venues on a map in Fig. 4b and provide additional insights in the data in Table I in the Appendix.

It is interesting to see that one of the selected venues (venue H) is a rose garden, amidst a park lush with trees. This is explained by the fact that the tree data set contains only street trees, and not trees in parks. Additionally, we found upon inspection using Google Street View that some of them (most notably venues A, B, D, G and H) do seem to be near a number of trees. When inspecting these locations in the tree data base,[2] we see that there are either only a few (venues B and G) or no trees (venues A, D, E and H) recorded in the immediate vicinity of the venues. We see that along with park trees, trees on private grounds are also not recorded.

VI. Conclusion

In this paper, we propose a novel criterion that can be used when selecting potential tree planting sites. The nature of the criterion is socio-cultural, capturing people movement along venues and tree-lacking (social) communities into one measure. Having implemented the measure for a case study on New York City, we show that the measure is applicable in the field and can be used to support decision makers by providing them with optional planting sites along a Pareto frontier.

We do see however, that some venues indicated by our criterion as tree lacking seem to actually be in a green area. We believe that the application of our method can be improved with a more detailed tree location data set. Then, the criterion proposed in this paper can be a meaningful addition to the established site selection criteria.

Acknowledgments

We thank Foursquare and the organisation of Netmob for organising the Future Cities Challenge and providing us with access to Foursquare's venue interaction data set.

References

[1] NYC Parks Recreation - TreesCount! 2015 Street Tree Census. https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh. Accessed: 15 May 2019.

[2] S Amir and A Misgav. A framework for street tree planning in urban areas in israel. *Landscape and Urban Planning*, 19(3):203–212, 1990.

[3] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008, 04 2008.

[4] UN DESA. World urbanization prospects, 2018.

[5] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[6] Susan A Hall, Jay S Kaufman, and Thomas C Ricketts. Defining urban and rural areas in us epidemiologic studies. *Journal of urban health*, 83(2):162–175, 2006.

[7] John R Harris and Michael P Todaro. Migration, unemployment and development: a two-sector analysis. *The American economic review*, pages 126–142, 1970.

[8] Chi Yung Jim. A planning strategy to augment the diversity and biomass of roadside trees in urban hong kong. *Landscape and Urban Planning*, 44(1):13–32, 1999.

[9] Richard B Langley. The utm grid system. *GPS world*, 9(2):46–50, 1998.

[10] Cecily Maller, Mardie Townsend, Anita Pryor, Peter Brown, and Lawrence St Leger. Healthy nature healthy people:contact with natureas an upstream health promotion intervention for populations. *Health promotion international*, 21(1):45–54, 2006.

[11] Arianna Morani, David J Nowak, Satoshi Hirabayashi, and Carlo Calfapietra. How to select the best tree planting locations to enhance air pollution removal in the milliontreesnyc initiative. *Environmental Pollution*, 159(5):1040–1047, 2011.

[12] Briony A Norton, Andrew M Coutts, Stephen J Livesley, Richard J Harris, Annie M Hunter, and Nicholas SG Williams. Planning for cooler cities: A framework to prioritise green infrastructure to mitigate high temperatures in urban landscapes. *Landscape and Urban Planning*, 134:127–138, 2015.

[13] William LI Parry-Jones. Natural landscape, psychological well-being and mental health. *Landscape research*, 15(2):7–11, 1990.

[14] Stephan Pauleit. Urban street tree plantings: identifying the key requirements. In *Proceedings of the Institution of Civil Engineers-Municipal Engineer*, volume 156, pages 43–50. Thomas Telford Ltd, 2003.

[15] CLE Rohde and AD Kendle. Human well-being, natural landscapes and wildlife in urban areas a review. 1994.

[16] Sudipto Roy, Aidan Davison, and Johan Östberg. Pragmatic factors outweigh ecosystem service goals in street tree selection and planting in south-east queensland cities. *Urban Forestry & Urban Greening*, 21:166–174, 2017.

[17] Henrik Sjöman and Anders Busse Nielsen. Selecting trees for urban paved sites in scandinavia. *Urban Forestry & Urban Greening*, 9(4):281–293, 2010.

[18] Ian F Spellerberg and David R Given. Trees in urban and city environments: a review of the selection criteria with particular reference to nature conservation in new zealand cities. *Landscape review*, 12(2):19–31, 2008.

[2] The tree database can be explored on a map at https://tree-map.nycgovparks.org/, last visited 21 May 2019.

Appendix

TABLE I: The venues that are, according to the Pareto analysis, the most efficient to place trees next to.

	Degree	**Tree density coefficient**	**Venue ID**	**Venue Name**	**Latitude**	**Longitude**	**Venue Category**
A	1864	2.09323399	4b637f59f964a5207b7e2ae3	MTA Subway - West Farms Square/E Tremont Av (2/5)	40.8402	-73.8800	Metro Stations
B	1561	1.94936904	4f940fe7e4b059d7da88be53	Junction Blvd	40.7491	-73.8694	Miscellaneous Shops
C	1212	1.86431926	4e7647cffa76059701632021	MTA Subway - 179th St (F)	40.7125	-73.7846	Metro Stations
D	853	1.70625431	4bace08af964a520cf143be3	Sammy’s Fish Box Restaurant	40.8390	-73.7836	Seafood Restaurants
E	532	1.55978734	4cc86db294e1a0933e6c978b	Rockaway Beach - 116th Street	40.5779	-73.8359	Beaches
F	305	1.18353191	4abcfe4bf964a520fa8720e3	Hulu Theater	40.7509	-73.9941	Music Venues
G	112	1.18192331	4c516433d2a7c9b6c4c61911	Bean & Bean Organic Coffee	40.7509	-73.9941	Coffee Shops
H	98	0.87556379	4debdb6b52b11677f060802e	Peggy Rockefeller Rose Garden	40.8592	-73.8735	Gardens
I	40	0.41121454	4d93a4489ef2721e6bffc3d2	I-495 / Grand Central Parkway Interchange	40.7400	-73.8455	Intersections
J	12	0.39198899	4e26fd0f1f6eb1ae139ad929	TSA Security Screening	40.6457	-73.7762	General Travel

Love of food: Do eateries drive London property prices?

Matthias Qian, Charles Rahal and Jan Callies

Abstract

We show that Foursquare venue interaction data can be used to predict out-of-sample changes in property prices. The Foursquare venue interaction data is first used to identify London cafés, restaurants and bars with a surge in popularity. We then show that the neighbourhood around eateries of rising popularity exhibited a 4% higher property price growth rate in 2018 as compared with the rest of London properties. The difference in property growth rates is statistically significant. We conclude that the change in popularity of eateries can forecast property prices.

Introduction

As the Irish playwriter George Bernard Shaw noted, "There is no love sincerer than the love of food." In a similar vein, the World Happiness Report by Helliwell et al. (2012) argues for the essential role of food for the happiness levels of populations. Yet, are epicure homebuyers willing to pay a premium to live close to their favourite eateries? In this paper, we study whether residential property prices are affected by the availability of popular cafés, restaurants or bars, which we collectively call eateries.

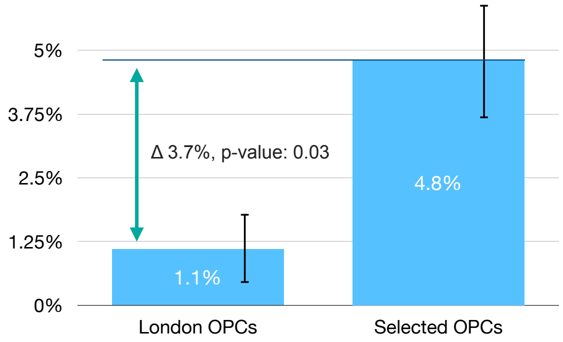
The literature on gentrification reveals that popular eateries are, of course, only one of many sources of abnormal property price returns. Zhou et al. (2017) identify an effect of cultural investment on urban deprivation, while Henneberry (1998) measures the effect of improved infrastructure on property markets. Glaeser et al. (2018) highlight the role of local groceries as an indicator for gentrification.

We introduce the concept of *eateries gentrification* which we define as the price premium of residential real estate induced by the availability of popular cafés, restaurants and bars in its neighbourhood. *Eateries gentrification* does not refer to the location decision of venue managers, who have an incentive to locate their services next to epicure citizens. Instead, we refer with our term only to the causal relationship from popular eateries towards higher property prices and property rents.

Reasons for *eateries gentrification* are threefold. First, the price premium on residential real estate is justified through the productivity gains of having a popular food source in close proximity to the home. This holds particularly true for the 'can't cook, won't cook' households who rely on eateries as their primary food source. Venues in close proximity to their home can reduce the time they spend to procure food and 'shoe-leather' costs. Second, popular eateries in the neighbourhood can improve the relative ranking of a property in the preference relation of prospective buyers or renters. More affluent households can better realise their preferences by outbidding competitors and thereby driving up property prices. Third, *eateries gentrification* can be explained via positive spill-over effects. These popular services can encourage existing property owners to improve the quality of their capital stock which can in turn attract more affluent households. Quality food services can kick-start a virtuous cycle of neighbourhood improvements.

Pundits have long suggested neighbourhoods to blossom around well-received eateries, but their analysis was all too often limited to anecdotical evidence.[1] The contribution of our paper is to put the hypothesis of *eateries gentrification* to a rigorous statistical test. Specifically, we make three contributions. First, we showcase a novel methodology to generate network based predictive features for property prices. Second, we establish the usefulness of a large-scale dataset by Foursquare with venue interactions to predict the returns on housing. Third, our results suggest the use of financial and regulatory incentives to establish new eateries as a powerful tool at the disposal of urban planners.

Figure 1: Comparison of OPC property returns for all of London and the selected OPCs only.



Note: The p-value refers to the Welch test of difference in means with unequal subsample variances. The error bars refer to confidence intervals on the individual mean return estimates. OPC stands for outward postcode.

Data

Our analysis relies on two distinct large-scale datasets that we bring together to put the hypothesis of *eateries gentrification* to a test. Our first dataset comes from Foursquare, an infamous location technology platform. Their longitudinal mobility dataset describes venue interactions in London from April 2017 to March 2019. It lists the ties between pairs of venues (over 7 million of them) which are equally sampled between the months.

To protect the privacy of individuals, the data does not specify the exact day of a venue interaction. Instead, the temporal dimension of the data is aggregated to a monthly frequency.

We match Foursquare's venue interactions dataset with information on the geographical location of venues to

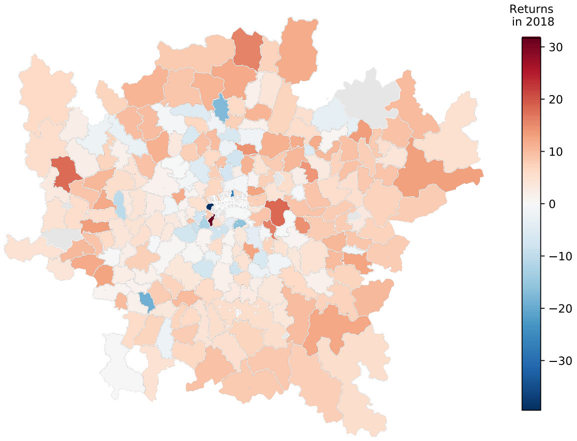
[1] The Telegraph, for example, mentions the restaurants in Brixton Village, as a driver for higher house prices in Brixton: https://www.telegraph.co.uk/property/house-prices/how-the-uks-best-restaurants-are-driving-up-property-prices/

compute distances individuals are willing to travel to visit a venue. We use these distances to compute the degree centrality of venues in a directed and weighted network.

The second large scale dataset is the price paid dataset from Her Majesty's Land Registry. It lists from 1995 onwards the transaction price of residential property sales. In total, the Land Registry details over 24 million property transactions.

We match Foursquare venues with properties in the Land Registry via outwards postcode (OPC) districts. In Britain, every building is assigned a postcode of the form: LLNA NLL where L is a letter, N a number and A alphanumeric. The second L and first A are optional. The first two (optionally one) characters refer to the postcode area, the second two (before the space) the outward postcode district, then the postcode sector, and the final two letters the individual unit\low level clusters of addresses. The outward postcode clusters hundreds or thousands of buildings into neighbourhoods.

Figure 2: The 2018 property returns for London OPCs



Note: OPC stands for outward postcode.

Quantitative Measures

In this section we explain how we compute the key quantitative measures for our analysis which are first, popularity measures for venues, and second, changes in property prices.

Popularity

Our measurement for popularity exploits the venue interactions in the Foursquare dataset. We use the Opsahl et al. (2010) node centrality measure for weighted and directed networks, where the weight is based on the geographical distance between venues. Their central idea is that popularity can be defined as the diversity of connected venues. A higher number of connected nodes should result in higher centrality measure for our focal node.

On the other hand, popularity can be defined as costs that customers are willing to take on to arrive at a venue. The costs grow proportionally with the distance between venues and only popular venues are able to attract a substantial customer base beyond the local neighbourhood. Our node centrality measure allows us to trade-off between the number of relationships between nodes and the strength of the ties. To define our measure of node centrality, we first introduce a measure for the degree of a focal node as

$$k_{it} = \sum_{j=1}^{N} x_{ijt},$$

where i is the focal node, j represents all other nodes, t is the timeframe, N is the total number of nodes, and x is the adjacency matrix, in which the cell x_{ij} is 1 if node i is connected to node j, and 0 otherwise. The node strength, which is the sum of weights, is

$$s_{it} = \sum_{j=1}^{N} w_{ijt}.$$

As we consider a directed network, we distinguish between ties that originate, k^{out}, and that are directed towards a node, k^{in}. The two node centrality measures are defined as:

$$C_{D-out}^{w\alpha}(i,t) = k_{it}^{\text{out}} \times \left(\frac{s_{it}^{out}}{k_{it}^{out}}\right)^{\alpha}$$

$$C_{D-in}^{w\alpha}(i,t) = k_{it}^{\text{in}} \times \left(\frac{s_{it}^{in}}{k_{it}^{in}}\right)^{\alpha}$$

where α is a positive tuning parameter. It balances the node strength and the number of contacts.

As our measure of popularity, we choose to use the latter of the two node centrality measures, $C_{D-in}^{w\alpha}(i,t)$. In our analysis, we choose to use $\alpha = 0.5$. With this choice of the tuning parameter, given a fixed total node strength, a higher number of contacts over which the strength is distributed increases the value of our measure.

Change in popularity

We identify a change in popularity of venues using an unequal variances t-test, also known as Welch's t-test from Welch (1947). Specifically, we compare the mean of $C_{D-in}^{w\alpha}(i,t)$ before and after January 2018. We then select those venues, for which there was the largest increase in centrality, as measured by the p-value of Welch's test. We make the discretionary choice of limiting our attention to ten postcodes that contain the twelve eateries with the largest increase in popularity.

Property prices

To compute OPC level real estate indices, we employ the repeat-sales methodology introduced in Bailey et al. (1963). The idea is to convert the index construction problem to a regression problem, by taking the logarithm of

$$R_{itt'} = \frac{B_{t'}}{B_t} \times U_{itt'}$$

where $R_{itt'}$ is the ratio of the final sales price in period t' to initial sales price in period t for the i-th pair of transactions with initial and final sale in these two periods. $B_t$and $B_{t'}$ are the true but unknown indexes for period t and t', where $t = 0,1,\dots,T-1$ and $t' = 1,\dots,T$. The term $U_{itt'}$ is the transaction specific error term, where we will assume that its log form $u_{itt'}$ has zero mean and the same variance σ^2, and that they are all uncorrelated to each other. The estimation of the unknown B's is treated as a regression problem. Let x_t take the value *-1* if period t is the period of initial sale, *+1* if the period of final sale, and 0 otherwise for each pair of transactions. Let $B_0 = 1$ and $b_0 = 0$. Based on taking the logarithm, the regression problem is then

$$r_{itt'} = \sum_{j=1}^{T} b_j x_j + u_{itt'}$$

To make the computation of the real estate price index more robust, we limit ourselves to consider only those outward postcode districts with more than 100 property transactions per year since 1995.

Residential property returns vary widely across outwards postcode (OPC) districts. As Table 1 shows, the best and worst performing OPCs in London in 2018 have over 30% annual gains and losses, respectively. The standard deviation of OPC property returns in 2018 was 7.6%. The average return was 1.1%. In total there were 118,247 property transactions in London; on average 425 property transaction per OPC. Figure 2 shows a map for the London OPCs with colour-coded 2018 property returns. It indicates that property returns are spatially correlated. Further, we see that the threshold of 100 property transactions per year is particularly binding in the centre of London, which is represented by greyed out OPCs.

Table 1: Best & worst OPC property returns in 2018

Best performers			Worst performers		
OPC	Count	Return	OPC	Count	Return
SW17	877	31.7%	UB9	242	-39.5%
E14	1815	16.5%	W10	229	-39.5%
TW2	422	16.4%	KT17	383	-20.9%
SE3	564	14.0%	KT8	333	-19.5%
SE28	246	12.4%	SE12	305	-18.3%
KT9	261	12.2%	W1H	111	-13.5%
E12	215	11.8%	SM3	336	-11.9%
RM12	485	11.0%	SE22	463	-11.4%
CM23	793	10.9%	N18	235	-11.1%
TW12	339	10.9%	SW12	527	-10.6%

Note: OPC stands for outward postcode.

Analytical strategy

We follow three steps to identify the *eateries gentrification* effect. First, we filter the Foursquare data for cafés, restaurants bars. We screen the filtered list of eateries for venues with large increases in popularity after January 2018. Second, we compute OPC level price changes in residential property prices for the full year of 2018. Third, we perform a hypothesis test about the significance of the restaurant gentrification effect.

The filtering of eateries follows the factor modelling literature for explaining the cross-section of equity returns, as pioneered by Fama and French (1993). We sort postcodes by the popularity of eateries, while they sort equities via size and value factors.

We consider changes in the popularity of venues, instead of the level of popularity, to attempt to identify a causal relationship. The locality of popular venues may not be random, as the manager of the venue may choose to locate his eatery in an affluent neighbourhood. Our identifying assumption is that the change in the popularity of venues is driven by exogenous factors unrelated to property prices.

The interpretation of the *eateries gentrification effect* changes with an analysis of changes in popularity of venues. While we conceptualised *eateries gentrification* as a price premium on residential property due to popular eateries, an analysis in changes in popularity then corresponds to differences in returns on property. We thus find evidence of *eateries gentrification* if we reject the following null hypothesis:

$$H_0\colon \mu_{popular} = \mu_{London}$$

where $\mu_{popular}$ is the expected 2018 property return in postcodes with venues of rising popularities and where μ_{London} is the corresponding expectation for the rest of London.

Results

Table 2 displays the 12 eateries in the ten OPCs that had the largest increase popularity post-2018. On average, customers visit these venues from 20 distinct connected venues per month. They travel on average 3.9 kilometres to visit these eateries. The surge in popularity, as measured by our degree centrality indicator, pre- and post-2018 was 5.89 on average.

Table 2: Characteristics of selected eateries

Name	Category	OPC	Nodes	Dist	Surge
What the Pitta	Vegan Restaurant	E1	19	4.0	4.5
Leon	Café	E14	36	3.7	2.3
Starbucks	Coffee Shop	E16	11	4.0	5.9
The Spread Eagle Bar	Bar	E2	9	1.6	2.3
Costa Coffee	Coffee Shop	E3	21	4.6	2.2
Starbucks	Coffee Shop	NW1	21	3.0	31.0
Goodfare Restaurant	Italian Restaurant	NW1	7	3.1	5.0
Panzer's	Deli	NW8	22	3.2	4.0
Costa Coffee	Coffee Shop	SE18	20	6.4	3.0
Pret A Manger	Sandwich Place	SW5	21	4.1	3.0
Starbucks	Coffee Shop	W6	32	4.7	2.8
Pret A Manger	Sandwich Place	W6	21	4.0	3.0

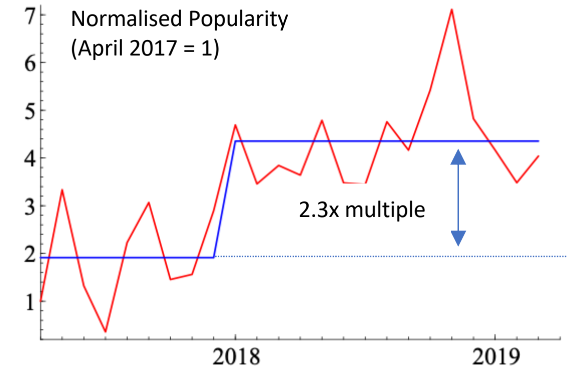
Note: 'Nodes' refer to the number of interactions that the venue has with distinct other venues on average per month. 'Dist' refers to the average distance to the connected venues measured in kilometres. 'Surge' refers to the multiple of Foursquare check-ins before and after January 2018. OPC stands for outward postcode.

Figure 1 showcases an example for the surge venue popularity. We fitted the time-varying mean to the time-series to highlight the structural break in its popularity.

We find high predictive power of the Foursquare venue interaction data for out-of-sample property price movements. On average, OPC districts with one of the selected London venues, exhibit a 4% higher property price growth than postcode districts without these venues. Figure 1 illustrates the differences in 2018 property returns. A

Welch test at a 5% significance level rejects the null hypothesis that the expected return for OPCs of eateries with rising popularity is equal to the return of all London OPCs. Thus we have established *eateries gentrification* in London, which we estimate to be 4% in 2018.

Figure 3: Popularity of "The Spread Eagle Bar"



These results hold in a difficult market environment for London residential property, which has been dominated by concerns over the outcome of the Brexit referendum. This explains why the property price returns in 2018 for the selected OPCs is still modest at an annual growth rate of 5%.

Table 3: The 2018 property returns on selected OPCs

Name	OPC	Return
Leon	E14	16.5%
The Spread Eagle Bar	E2	10.5%
Starbucks	E16	6.2%
Pret A Manger	SW5	6.1%
Costa Coffee	SE18	5.2%
What The Pitta	E1	4.5%
Starbucks & Goodfare	NW1	3.8%
Costa Coffee	E3	1.1%
Starbucks & Pret A Manger	W6	-0.2%
Panzer's	NW8	-0.7%

Average Price Increase in selected OPCs	4.8%
Average prince increase in all London OPCs	1.1%
T-statistic for Welch's test	2.5
P-value of Welch's test	3.2%

Note: OPC stands for outward postcode.

Conclusion

The goal of this paper is to both test the hypothesis of the existence of *eateries gentrification* and to estimate the magnitude of this effect. We construct a measure for the popularity of restaurants and show that postcodes with cafés, restaurants and bars of rising popularity exhibit higher investment returns on residential property. The effect is statistically significant. In London of 2018, the effect size has been 4 percent.

Our results raise further questions. Does the *eateries gentrification* effect change over time? Does it persist in recessions? How does it compare to supermarket gentrification or the London Crossrail gentrification? While open questions remain, one fact is undeniable. The Foursquare dataset provides valuable information to predict residential property markets.

References

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse. "A regression method for real estate price index construction." Journal of the American Statistical Association 58, no. 304 (1963): 933-942.

Fama, Eugene F., and Kenneth R. French. "Common risk factors in the returns on stocks and bonds." *Journal of financial economics* 33, no. 1 (1993): 3-56.

Glaeser, Edward L., Hyunjin Kim, and Michael Luca. "Nowcasting gentrification: using yelp data to quantify neighborhood change." In *AEA Papers and Proceedings*, vol. 108, pp. 77-82. 2018.

Helliwell, John F., Richard Layard, and Jeffrey Sachs. "World happiness report." (2012).

Henneberry, John. "Transport investment and house prices." Journal of Property Valuation and Investment 16, no. 2 (1998): 144-158.

Opsahl, Tore, Filip Agneessens, and John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." Social networks 32, no. 3 (2010): 245-251.

Welch, Bernard L. "The generalization ofstudent's' problem when several different population variances are involved." Biometrika 34, no. 1/2 (1947): 28-35.

Zhou, Xiao, Desislava Hristova, Anastasios Noulas, Cecilia Mascolo, and Max Sklar. "Cultural investment and urban socio-economic development: a geosocial network approach." Royal Society open science 4, no. 9 (2017): 170413.

Metaheuristic macro scale traffic flow optimisation from urban movement data

Laurens Arp[*]
Dyon van Vreumingen[*]
{l.r.arp,d.van.vreumingen}@umail.leidenuniv.nl
Leiden Institute for Advanced Computer Science
Leiden, The Netherlands

Daniela Gawehns
Mitra Baratchi
{d.gawehns,m.baratchi}@liacs.leidenuniv.nl
Leiden Institute for Advanced Computer Science
Leiden, The Netherlands

ABSTRACT

How can urban movement data be exploited in order to improve the flow of traffic within a city? Movement data provides valuable information about routes and specific roads that people are likely to drive on. This allows us to pinpoint roads that occur in many routes and are thus sensitive to congestion. Redistributing some of the traffic to avoid unnecessary use of these roads could be a key factor in improving traffic flow.Many proposed approaches to combat congestion are either static or do not incorporate any movement data. In this work, we present a method to redistribute traffic through the introduction of externally imposed variable costs to each road segment, assuming that all drivers seek to drive the cheapest route. We use a metaheuristic optimisation approach to minimise total travel times by optimising a set of road-specific variable cost parameters, which are used as input for an objective function based on traffic flow theory. The optimisation scenario for the city centre of Tokyo considered in this paper was defined using public spatial road network data, and movement data acquired from Foursquare. Experimental results shows that our proposed scenario has the potential to achieve a 62.6% improvement of total travel time in Tokyo compared to that of a currently operational road network configuration, with no imposed variable costs.

KEYWORDS

traffic flow, urban movements, metaheuristics

1 INTRODUCTION

Even though extensive road networks have been developed to satisfy the high demand for vehicular transportation, overoccupancy of roads still occurs on a daily basis, causing traffic jams which hurt the environment, the economy and the drivers' moods. Finding a solution to traffic congestion is a challenging problem that has occupied many in the past century. After all, traffic dynamics are difficult to predict, due to complex fluctuations in traffic demand, both spatial and temporal. This makes it hard to devise a protocol for traffic flow redistribution that works well in varying conditions.

To date, various approaches have been proposed to alleviate congestion in some way [8]. However, these methods tend to be either static, data-independent protocols, micro-scale solutions (on the level of individual roads) or primarily driven by theoretical models. Our objective instead is to construct a dynamic, data-driven, macro-scale (road network level) approach to address traffic congestion. In this sense, dynamic means that a solution can be adapted to new traffic data with relative ease.

In this work, we propose a method for traffic redistribution fueled by metaheuristic optimisation, which we test on the case of the city centre of Tokyo. We seek to shift the traffic situation away from a state where each driver chooses the fastest, or shortest, route (thus causing congestion on roads that occur in many shortest routes), towards a system optimal equilibrium, as coined by Wardrop [14], where the *total travel time* for all drivers is minimised. By introducing externally imposed variable costs (e.g. tolls, or any other financial or non-financial method a supervisory institution might deploy) on each road, we aim to discourage drivers from all taking the same congested roads. This approach asserts that, on average, each driver is willing to take the cheapest route from their point of departure to the destination, where the total costs to drive a route depend both on the distance travelled, through a spatial cost, and the imposed variable costs encountered along the route.

In order to make predictions of traffic flow and occurrence of congestion, we infer traffic demand from a data set of urban movements. A number of public traffic data sets, such as the Dutch NDW [11], report the traffic flow or density at certain points in time; however, while this gives a detailed picture of a local situation, it provides no information as to what routes drivers are following. Hence, such data is of little use when we wish to redistribute traffic by encouraging sensible alternative routes. For this reason, we use a data set of urban movements, provided by Foursquare as part of the *Future cities challenge*, which allows for the inference of traffic level information needed for this research such as the origin and destination of movements [2].

2 RELATED WORK

The objective of combatting traffic congestion by altering road network setups has been addressed in a large body of work. The use of road pricing as a means to achieve this goal is a prevalent approach [6, 13, 15, 16]. In this context, the marginal cost of congestion is a frequently employed measure to assess the optimal road pricing. A key difference between these papers and our work is that the previously proposed road pricing policies are fixed (e.g. charge a fee within a certain radius of the city centre) instead of dynamic, and do not follow from an optimisation procedure based on actual movement data. Approaches for optimising road networks and traffic flow from a different viewpoint, unrelated to road pricing policies, include metaheuristic optimisation of road improvements [3], development of intelligent traffic light systems [10], optimisation of road graph architectures with evolutionary algorithms [12], and prediction of optimal traffic flow through maximum-entropy methods [9]. A more exhaustive list of methods is provided by Kumar Shukla and Agrawal [8]. In this work, we explore the use of optimisation

[*] Both authors contributed equally to this research.

algorithms in proposing a dynamic pricing mechanism using actual movement data.

3 PROBLEM STATEMENT

The problem of optimising traffic flow through adaptive road pricing is twofold. First, we must estimate traffic flow and congestion in a road network, which is the underlying cause of high total travel time when all drivers follow the cheapest routes from their origins to their destinations. By combining movement data describing the traffic demand in the network and the spatial road network data, traffic flow theory yields these estimations. The demand data is a set D_M consisting of movements between venue locations within the road network. Each element of this set (indexed as k) is a tuple (A_k, B_k, N_k) where A_k and B_k are elements of a venue data set D_V containing spatial information about the venues, and N_k is the recorded frequency of the specific movement from A_k to B_k. Second, having found a method to express the total travel time as a function of the variable cost parameters, we aim to optimise the parameters for minimal total travel time. Our methods for addressing this optimisation problem are set out in detail in the next section.

4 METHODS

4.1 Traffic flow estimation

4.1.1 Road network and routing model. The first step towards prediction and minimisation of congestion is to represent the physical road network as a planar graph G that has road segments for edges, which may be traversed in order to travel from an origin to a destination. Specifically, the graph is a tuple $G = (V, E, S)$ with V the node set, E the edge set and S the set of Haversine lengths of all edges. The node set V contains intersections in the road network, as well as nodes for the origin and destination locations from D_V.

We then introduce, for each road segment $(i, j) \in E$ in the graph, a cost that a driver needs to pay to traverse this segment. The main part of this cost is a *variable cost* p_{ij}. All variable cost parameters collectively form the variable cost vector P which we seek to optimise for minimum congestion. Next, we assign to each segment a *spatial cost*, which is an immutable base cost for travelling from node i to j that is linearly dependent on the length s_{ij} of the segment by a tunable factor β_s. Since the movement data is aggregated into frequency numbers, and is not provided on an individual level for anonymity reasons, we take β_s to be equal for all drivers. Put together, the total cost for a driver to travel via a connected route of segments $R = \{(i, j)\}$ from some origin to a destination, is given by the sum of the individual segment costs occurring on the route:

$$\text{cost}(R) = \sum_{(i,j)\in R} \beta_s s_{ij} + p_{ij}. \tag{1}$$

For the development of traffic flow, we assert that all drivers are selfish and seek to drive the route which incurs the lowest total cost. These routes can be found using a weighted shortest path algorithm. Note that if $p_{ij} = 0$ on all edges, each driver will drive the route of lowest spatial cost, which is exactly the shortest route. From the cheapest routes, which are jointed collections of segments, and the frequency numbers N_k, we can predict the vehicle count n_{ij} on each segment, from which the degree of congestion is computed as set out in the following subsection.

4.1.2 Congestion model. In our congestion model, we assume that the flow of traffic on a road segment is fully described by a Greenshields fundamental traffic flow curve [4], which is a widely used theoretical model for predicting traffic dynamics on a road segment. The used variables are flow f (vehicles passing by per unit time) and density ρ (vehicles present per unit length). In this model, there exists a *maximum density* ρ_m that the road can support, beyond which the total flow is zero. Furthermore, there is a *critical density* ρ_c at which the flow reaches its maximum value: $f(\rho_c) = f_m$. Naturally, $f(0) = 0$, as no flow exists when no cars are present.

A basic curve that fits this description is a concave quadratic function, with zero flow at $\rho > \rho_m$, which we define as

$$f(\rho) = \frac{f_m}{\rho_c^2} \max\left(0, \rho[2\rho_c - \rho]\right), \tag{2}$$

where we note that, since f is quadratic in ρ, $\rho_m = 2\rho_c$. The maximum flow is directly related to the critical density; assuming that the traffic is able to drive at the maximum allowed speed v_m when the density is at its critical point, we set $f_m = \rho_c v_m$.

From this density-flow dependence, we can extract the *space mean speed* $v(\rho)$, the average speed of all vehicles on the road segment, as [4]

$$v(\rho) = \min\left(v_m, \frac{f(\rho)}{\rho}\right), \tag{3}$$

where again the maximum speed enters the relation, this time as a bound on the space mean speed. Finally, the *space mean travel time* $t(\rho)$ on the segment, taken to have length s, is computed as

$$t(\rho) = \frac{s}{v(\rho)}. \tag{4}$$

From the expected number of vehicles n_{ij} on each road segment, we obtain the segment density as $\rho_{ij} = n_{ij}/s_{ij}$. For multi-lane roads, we divide this number by the number of lanes. By inserting the segment density into the density-time relation described above (eq. 4), we find the space mean time t_{ij} spent on the segment. The collection of segment travel times then leads to the definition of the objective function for optimisation by a metaheuristic algorithm.

4.2 Parameter optimisation

4.2.1 Objective function. The objective function computes the measure obj(P), which reflects the extent to which the system optimal equilibrium is reached by a variable cost configuration P. This equilibrium occurs when the total travel times for each driver on their routes are minimal. As such, we define the objective function as the total space mean travel time over all routes driven on the road network. This total travel time can be conveniently expressed using the segment vehicle counts n_{ij}, which are directly dependent on P:

$$\text{obj}(\text{P}) = \sum_{(i,j)\in E} n_{ij}(\text{P})\, t_{ij}(\text{P}). \tag{5}$$

Note that each segment mean travel time t_{ij} is also a function of P since it is dependent on the vehicle count n_{ij}. Algorithm 1 shows, in pseudocode, the routine to compute the objective function.

4.2.2 Optimisation algorithms. The purpose of the optimisation algorithms is to find an optimal variable cost configuration such that the value of the objective function, the total travel time, is minimised. In principle, any black-box metaheuristic optimisation

```
Data: road network graph $G = (V, E, S)$ with node set $V$ inferred
        from venue data $D_V$, $E$ and $S$ inferred from road network data;
        movement data $D_M$;
        spatial cost factor $\beta_s$
Input: parameter vector P
Result: objective function value obj(P)

initialise $n_{ij} \leftarrow 0$, $t_{ij} \leftarrow 0$, $\rho_{ij} \leftarrow 0$ for all $(i, j) \in E$

// Compute predicted vehicle counts $n_{ij}$ on each segment
forall origin-destination-freq. tuples $(A_k, B_k, N_k)$ in $D_M$ do
    Find cheapest route $R_k$ from $A_k$ to $B_k$ according to $\beta_s$ and P
      using weighted shortest-path algorithm
    forall $(i, j) \in R_k$ do
        $n_{ij} \leftarrow n_{ij} + N_k$
    end
end
// Find mean travel times $t_{ij}$ on each segment
forall $(i, j) \in E$ do
    $\rho_{ij} \leftarrow n_{ij}/s_{ij}$
    $t_{ij} \leftarrow t(\rho_{ij})$                (eqs. 2–4)
end
obj(P) $\leftarrow \sum_{(i,j)\in E} n_{ij}\, t_{ij}$        (eq. 5)
return obj(P)
```

Algorithm 1. Routine for computing the objective function for a given parameter vector P.

algorithm could be used to search for local optima of variable cost configurations which might approximate a system optimal equilibrium. That said, for these purposes, algorithms which are robust for high-dimensional problems are preferred, as the number of parameters increases proportionally to the number of edges in the graph.

For our proof-of-concept implementation we use simulated annealing (SA) [7], which is a variation of hill climbing where worse solutions can get accepted depending on the algorithm's decreasing *temperature* value, and a genetic algorithm (GA) adapted for continuous optimisation. For both algorithms each iteration contains 40 objective function evaluations; after one iteration, the GA updates its population, whereas SA resets its temperature value. Both algorithms use mutations generated using a normal distribution with zero mean and unit variance, at a mutation rate of 0.2 per parameter.

5 CASE STUDY: TOKYO CITY CENTRE

In order to test our traffic flow optimisation method, we applied it to movements inside the city centre of Tokyo (i.e. excluding the Greater Tokyo area). We briefly discuss the movements and road network used for the case study, and present experimental results.

5.1 Movement data

The movement data was provided by Foursquare as part of the *Future cities challenge* [2]; we selected only those parts related to the Tokyo city centre. The data contains a list of venues together with their GPS locations, forming the data set D_V introduced in section 3, and a list of movements between the venues. The movement data contains movements of the same form as the tuples in D_M, but with additional indications of the month during which the movement occurred, and the time of the day (periods of 4 to 6 hours). We considered only movements made in the afternoon, and combined the frequencies N_k for the same movement in different months into a single figure. Since the Foursquare dataset does not cover the entirety of vehicular movements inside Tokyo (and, in fact, also includes other types of movements such as subway, biking and walking trips), we viewed the frequencies as ratios rather than absolute numbers, asserting the law of large numbers for sufficient accuracy. We then normalised the frequencies to numbers that the road graph we constructed (see next subsection) could support.

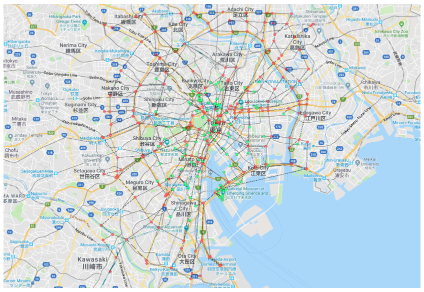


Figure 1. Road network graph example. Green nodes are venues, red nodes are intersections. Map source: Google Maps [5].

As a last modification, we selected the venues occurring in the 100 most popular (i.e. frequent) routes, and clustered all other venues together with their nearest neighbour (in terms of Haversine distance) from the set of most popular venues. The routes were clustered accordingly, going between venue clusters instead of individual venues. This was done in order to substantially reduce the number of routes and therewith the computational complexity of the problem.

5.2 Road network data

The road network data used is based on the Asia shapefile provided by the Earthdata *Global roads open access data set* [1]. It contains information on the road networks of the entirety of Asia with a variable resolution; for the city Tokyo, its resolution is well suited for the algorithm. The road data is translated into a graph representation by finding intersections between lines, and turning these into intersection nodes. The lines themselves are used to create edges between intersections. Venue nodes are created by identifying the location of the venues from the venue data set, which are connected to the nearest intersection node.

5.3 Experimental results

The improvement progress for both optimisation algorithms is shown in figures 3 and 4. For comparison, we ran the objective function once with all variable costs set to 0 to obtain the default flow of the road network. This configuration had an objective value of around 416 million hours spent on the road network. It should be noted that the objective function values are higher than what would be a realistic amount of time spent on roads. As the theoretical traffic flow models do not take relief methods into account for fully congested roads, the speed on those roads in considered to be zero. In these cases, we use an arbitrary low value to represent the speed on the road, resulting in somewhat unrealistic amounts of hours.

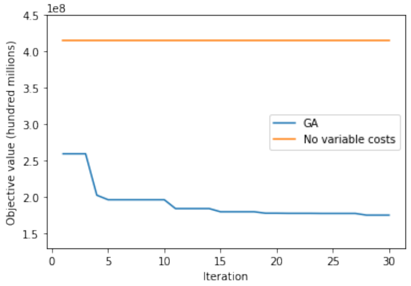


Figure 2. Objective function value plotted to the number of GA iterations. Values are the total time spent on the roads over 6 hours in the afternoon.

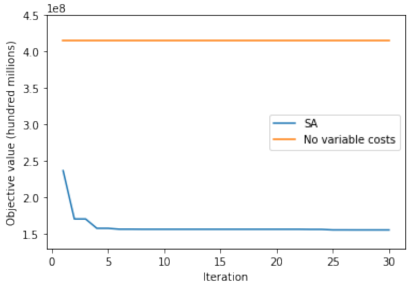


Figure 3. Objective function value plotted with the number of SA iterations. Values are the total time spent on the roads over 6 hours in the afternoon.

Both the GA and SA were able to find solutions which substantially improved traffic flow over the network. After 30 iterations, the lowest fitness value of the GA was around 175 million hours (an improvement of 57.9%), and the lowest fitness value of SA was around 155 million hours (an improvement of 62.6%). Effectively, this means the solution found by SA was able to reduce the total amount of hours spent over a period of 6 hours in the afternoon by 261 million, due to having found a good distribution of variable costs for each road segment. Although the results of the GA were not as good as those of SA, the GA was still successful in improving the unoptimised configuration, which supports the view that any optimisation algorithm could be used for these purposes.
Interestingly, even the first iterations of the algorithms showed values which were an improvement compared to the default traffic flow. This is likely due to the nature of the shortest path algorithm when taking only distance into account, which sends many cars over the same roads unnecessarily. Though this type of behaviour was the basic premise allowing us to optimise, given the large variety of available, very similar roads in the network, any deviation from the over-use of single roads caused by the random initialisation of variable costs would result in better traffic flow. From such a randomly initialised state, both algorithms then improved the solutions such that the more well-suited alternatives were used.

The results show that the algorithms were effective in finding solutions improving traffic flow over the default setting of no variable costs. That said, there is no guarantee the optima the algorithms converged to were global optima, nor that the convergence was as fast as possible. Future work could include more thorough exploration of optimisation algorithms and their parameter settings.

6 CONCLUSION

We have shown that we can successfully address traffic congestion by redistributing traffic through imposing of variable road segment costs, and optimising this cost configuration using metaheuristic algorithms. The best variable costs configuration was found by a simulated annealing routine, improving upon the total travel time corresponding to a configuration with zero variable costs by 62.6%. Both simulated annealing and a genetic algorithm were effective at optimising solutions.

Though the practical implementation of the variable costs may be another non-trivial problem to address first, the positive results show that, at least conceptually, this method could result in improved traffic flow when applied in practice. Consequently, cities may enjoy shorter travel times, better accessibility, cleaner air and, not unimportantly, improved drivers' moods.

ACKNOWLEDGMENTS

We would like to extend special thanks to Foursquare and Netmob for the movement data and to Giorgos Kyrziridis for his support.

REFERENCES

[1] NASA Earthdata. [n. d.]. Global Roads Open Access Data Set. https://earthdata.nasa.gov/. Accessed: 16-05-2019.

[2] Foursquare. 2019. Future Cities Challenge. https://www.futurecitieschallenge.com/. Accessed: 16-05-2019.

[3] Mariano Gallo, Luca D'Acierno, and Bruno Montella. 2010. A meta-heuristic approach for solving the urban network design problem. *European Journal of Operational Research* (2010). https://doi.org/10.1016/j.ejor.2009.02.026

[4] Bruce D. Greenshields, J. T. Thompson, H. C. Dickinson, and R. S. Swinton. 1934. The photographic method of studying traffic behavior.

[5] Google Inc. [n. d.]. Google Maps. https://www.google.nl/maps. Accessed: 21-05-2019.

[6] Theodore E. Keeler and Kenneth A. Small. 2002. Optimal peak-load pricing, investment, and service levels on urban expressways. *Journal of Political Economy* (2002). https://doi.org/10.1086/260543

[7] Scott Kirkpatrick, Charles Daniel Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220 4598 (1983), 671–80.

[8] Sanjiv Kumar Shukla and Anupam Agrawal. 2015. Present and future perspective on optimization of road network management. *International Journal of Computer Trends and Technology* 22, 2 (2015), 64–67. https://doi.org/10.14445/22312803/ijctt-v22p112

[9] L. Liu. 2008. Optimization of traffic flow in a road network. In *2008 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Vol. 1. 1244–1247. https://doi.org/10.1109/ICICTA.2008.21

[10] Pallavi Mandhare, Vilas Kharat, and C.Y. Patil. 2018. Intelligent road traffic control system for traffic congestion: a perspective. *International Journal of Computer Sciences and Engineering* 6 (07 2018), 908–915. https://doi.org/10.26438/ijcse/v6i7.908915

[11] NDW. [n. d.]. Nationale Databank Wegverkeersgegevens. https://www.ndw.nu/. Accessed: 21-05-2019.

[12] Frank Schweitzer, Helge Rosé, Werner Ebeling, and Olaf Weiss. 1997. Optimization of road networks using evolutionary strategies. *Evolutionary Computation* (1997). https://doi.org/10.1162/evco.1997.5.4.419

[13] Alan A. Walters. 1962. The theory and measurement of private and social cost of highway congestion. *Econometrica* (1962). https://doi.org/10.2307/1911814

[14] John G. Wardrop. 1952. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers Part 2*. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[15] Jun Yang, Avralt-Od Purevjav, and Shanjun Li. 2017. The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in Beijing. (2017). https://doi.org/10.2139/ssrn.2948619

[16] Sun Ye. 2012. Research on urban road traffic congestion charging based on sustainable development. *Physics Procedia* (2012). https://doi.org/10.1016/j.phpro.2012.02.231

Love of food: Do eateries drive London property prices?

Matthias Qian, Charles Rahal and Jan Callies

Abstract

We show that Foursquare venue interaction data can be used to predict out-of-sample changes in property prices. The Foursquare venue interaction data is first used to identify London cafés, restaurants and bars with a surge in popularity. We then show that the neighbourhood around eateries of rising popularity exhibited a 4% higher property price growth rate in 2018 as compared with the rest of London properties. The difference in property growth rates is statistically significant. We conclude that the change in popularity of eateries can forecast property prices.

Introduction

As the Irish playwriter George Bernard Shaw noted, "There is no love sincerer than the love of food." In a similar vein, the World Happiness Report by Helliwell et al. (2012) argues for the essential role of food for the happiness levels of populations. Yet, are epicure homebuyers willing to pay a premium to live close to their favourite eateries? In this paper, we study whether residential property prices are affected by the availability of popular cafés, restaurants or bars, which we collectively call eateries.

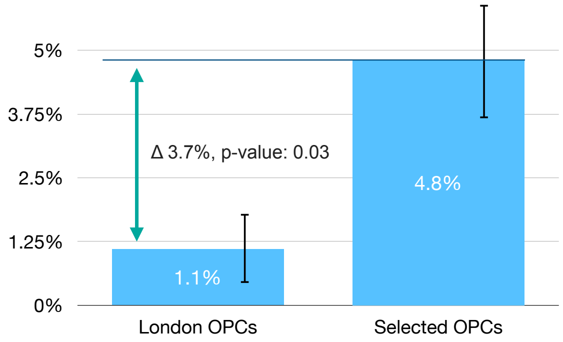
The literature on gentrification reveals that popular eateries are, of course, only one of many sources of abnormal property price returns. Zhou et al. (2017) identify an effect of cultural investment on urban deprivation, while Henneberry (1998) measures the effect of improved infrastructure on property markets. Glaeser et al. (2018) highlight the role of local groceries as an indicator for gentrification.

We introduce the concept of *eateries gentrification* which we define as the price premium of residential real estate induced by the availability of popular cafés, restaurants and bars in its neighbourhood. *Eateries gentrification* does not refer to the location decision of venue managers, who have an incentive to locate their services next to epicure citizens. Instead, we refer with our term only to the causal relationship from popular eateries towards higher property prices and property rents.

Reasons for *eateries gentrification* are threefold. First, the price premium on residential real estate is justified through the productivity gains of having a popular food source in close proximity to the home. This holds particularly true for the 'can't cook, won't cook' households who rely on eateries as their primary food source. Venues in close proximity to their home can reduce the time they spend to procure food and 'shoe-leather' costs. Second, popular eateries in the neighbourhood can improve the relative ranking of a property in the preference relation of prospective buyers or renters. More affluent households can better realise their preferences by outbidding competitors and thereby driving up property prices. Third, *eateries gentrification* can be explained via positive spill-over effects. These popular services can encourage existing property owners to improve the quality of their capital stock which can in turn attract more affluent households. Quality food services can kick-start a virtuous cycle of neighbourhood improvements.

Pundits have long suggested neighbourhoods to blossom around well-received eateries, but their analysis was all too often limited to anecdotical evidence.[1] The contribution of our paper is to put the hypothesis of *eateries gentrification* to a rigorous statistical test. Specifically, we make three contributions. First, we showcase a novel methodology to generate network based predictive features for property prices. Second, we establish the usefulness of a large-scale dataset by Foursquare with venue interactions to predict the returns on housing. Third, our results suggest the use of financial and regulatory incentives to establish new eateries as a powerful tool at the disposal of urban planners.

Figure 1: Comparison of OPC property returns for all of London and the selected OPCs only.



Note: The p-value refers to the Welch test of difference in means with unequal subsample variances. The error bars refer to confidence intervals on the individual mean return estimates. OPC stands for outward postcode.

Data

Our analysis relies on two distinct large-scale datasets that we bring together to put the hypothesis of *eateries gentrification* to a test. Our first dataset comes from Foursquare, an infamous location technology platform. Their longitudinal mobility dataset describes venue interactions in London from April 2017 to March 2019. It lists the ties between pairs of venues (over 7 million of them) which are equally sampled between the months.

To protect the privacy of individuals, the data does not specify the exact day of a venue interaction. Instead, the temporal dimension of the data is aggregated to a monthly frequency.

We match Foursquare's venue interactions dataset with information on the geographical location of venues to

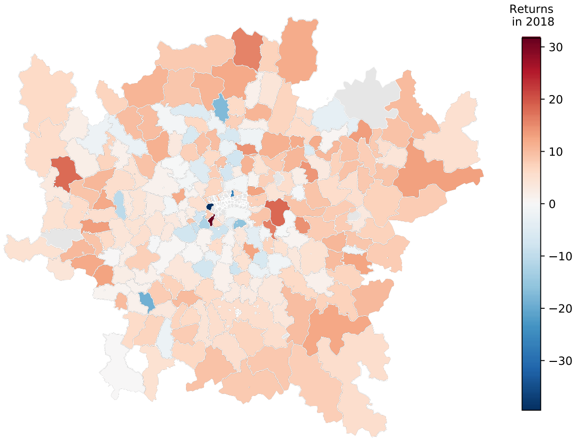
[1] The Telegraph, for example, mentions the restaurants in Brixton Village, as a driver for higher house prices in Brixton: https://www.telegraph.co.uk/property/house-prices/how-the-uks-best-restaurants-are-driving-up-property-prices/

compute distances individuals are willing to travel to visit a venue. We use these distances to compute the degree centrality of venues in a directed and weighted network.

The second large scale dataset is the price paid dataset from Her Majesty's Land Registry. It lists from 1995 onwards the transaction price of residential property sales. In total, the Land Registry details over 24 million property transactions.

We match Foursquare venues with properties in the Land Registry via outwards postcode (OPC) districts. In Britain, every building is assigned a postcode of the form: LLNA NLL where L is a letter, N a number and A alphanumeric. The second L and first A are optional. The first two (optionally one) characters refer to the postcode area, the second two (before the space) the outward postcode district, then the postcode sector, and the final two letters the individual unit\low level clusters of addresses. The outward postcode clusters hundreds or thousands of buildings into neighbourhoods.

Figure 2: The 2018 property returns for London OPCs



Note: OPC stands for outward postcode.

Quantitative Measures

In this section we explain how we compute the key quantitative measures for our analysis which are first, popularity measures for venues, and second, changes in property prices.

Popularity

Our measurement for popularity exploits the venue interactions in the Foursquare dataset. We use the Opsahl et al. (2010) node centrality measure for weighted and directed networks, where the weight is based on the geographical distance between venues. Their central idea is that popularity can be defined as the diversity of connected venues. A higher number of connected nodes should result in higher centrality measure for our focal node.

On the other hand, popularity can be defined as costs that customers are willing to take on to arrive at a venue. The costs grow proportionally with the distance between venues and only popular venues are able to attract a substantial customer base beyond the local neighbourhood. Our node centrality measure allows us to trade-off between the number of relationships between nodes and the strength of the ties. To define our measure of node centrality, we first introduce a measure for the degree of a focal node as

$$k_{it} = \sum_{j=1}^{N} x_{ijt},$$

where i is the focal node, j represents all other nodes, t is the timeframe, N is the total number of nodes, and x is the adjacency matrix, in which the cell x_{ij} is 1 if node i is connected to node j, and 0 otherwise. The node strength, which is the sum of weights, is

$$s_{it} = \sum_{j=1}^{N} w_{ijt}.$$

As we consider a directed network, we distinguish between ties that originate, k^{out}, and that are directed towards a node, k^{in}. The two node centrality measures are defined as:

$$C_{D-out}^{w\alpha}(i,t) = k_{it}^{\text{out}} \times \left(\frac{s_{it}^{out}}{k_{it}^{out}}\right)^{\alpha}$$

$$C_{D-in}^{w\alpha}(i,t) = k_{it}^{\text{in}} \times \left(\frac{s_{it}^{in}}{k_{it}^{in}}\right)^{\alpha}$$

where α is a positive tuning parameter. It balances the node strength and the number of contacts.

As our measure of popularity, we choose to use the latter of the two node centrality measures, $C_{D-in}^{w\alpha}(i,t)$. In our analysis, we choose to use $\alpha = 0.5$. With this choice of the tuning parameter, given a fixed total node strength, a higher number of contacts over which the strength is distributed increases the value of our measure.

Change in popularity

We identify a change in popularity of venues using an unequal variances t-test, also known as Welch's t-test from Welch (1947). Specifically, we compare the mean of $C_{D-in}^{w\alpha}(i,t)$ before and after January 2018. We then select those venues, for which there was the largest increase in centrality, as measured by the p-value of Welch's test. We make the discretionary choice of limiting our attention to ten postcodes that contain the twelve eateries with the largest increase in popularity.

Property prices

To compute OPC level real estate indices, we employ the repeat-sales methodology introduced in Bailey et al. (1963). The idea is to convert the index construction problem to a regression problem, by taking the logarithm of

$$R_{itt'} = \frac{B_{t'}}{B_t} \times U_{itt'}$$

where $R_{itt'}$ is the ratio of the final sales price in period t' to initial sales price in period t for the i-th pair of transactions with initial and final sale in these two periods. $B_t$and $B_{t'}$ are the true but unknown indexes for period t and t', where $t = 0, 1, \dots, T-1$ and $t' = 1, \dots, T$. The term $U_{itt'}$ is the transaction specific error term, where we will assume that its log form $u_{itt'}$ has zero mean and the same variance σ^2, and that they are all uncorrelated to each other. The estimation of the unknown B's is treated as a regression problem. Let x_t take the value *-1* if period t is the period of initial sale, *+1* if the period of final sale, and 0 otherwise for each pair of transactions. Let $B_0 = 1$ and $b_0 = 0$. Based on taking the logarithm, the regression problem is then

$$r_{itt'} = \sum_{j=1}^{T} b_j x_j + u_{itt'}$$

To make the computation of the real estate price index more robust, we limit ourselves to consider only those outward postcode districts with more than 100 property transactions per year since 1995.

Residential property returns vary widely across outwards postcode (OPC) districts. As Table 1 shows, the best and worst performing OPCs in London in 2018 have over 30% annual gains and losses, respectively. The standard deviation of OPC property returns in 2018 was 7.6%. The average return was 1.1%. In total there were 118,247 property transactions in London; on average 425 property transaction per OPC. Figure 2 shows a map for the London OPCs with colour-coded 2018 property returns. It indicates that property returns are spatially correlated. Further, we see that the threshold of 100 property transactions per year is particularly binding in the centre of London, which is represented by greyed out OPCs.

Table 1: Best & worst OPC property returns in 2018

Best performers				Worst performers		
OPC	Count	Return		OPC	Count	Return
SW17	877	31.7%		UB9	242	-39.5%
E14	1815	16.5%		W10	229	-39.5%
TW2	422	16.4%		KT17	383	-20.9%
SE3	564	14.0%		KT8	333	-19.5%
SE28	246	12.4%		SE12	305	-18.3%
KT9	261	12.2%		W1H	111	-13.5%
E12	215	11.8%		SM3	336	-11.9%
RM12	485	11.0%		SE22	463	-11.4%
CM23	793	10.9%		N18	235	-11.1%
TW12	339	10.9%		SW12	527	-10.6%

Note: OPC stands for outward postcode.

Analytical strategy

We follow three steps to identify the *eateries gentrification* effect. First, we filter the Foursquare data for cafés, restaurants bars. We screen the filtered list of eateries for venues with large increases in popularity after January 2018. Second, we compute OPC level price changes in residential property prices for the full year of 2018. Third, we perform a hypothesis test about the significance of the restaurant gentrification effect.

The filtering of eateries follows the factor modelling literature for explaining the cross-section of equity returns, as pioneered by Fama and French (1993). We sort postcodes by the popularity of eateries, while they sort equities via size and value factors.

We consider changes in the popularity of venues, instead of the level of popularity, to attempt to identify a causal relationship. The locality of popular venues may not be random, as the manager of the venue may choose to locate his eatery in an affluent neighbourhood. Our identifying assumption is that the change in the popularity of venues is driven by exogenous factors unrelated to property prices.

The interpretation of the *eateries gentrification effect* changes with an analysis of changes in popularity of venues. While we conceptualised *eateries gentrification* as a price premium on residential property due to popular eateries, an analysis in changes in popularity then corresponds to differences in returns on property. We thus find evidence of *eateries gentrification* if we reject the following null hypothesis:

$$H_0 \colon \mu_{popular} = \mu_{London}$$

where $\mu_{popular}$ is the expected 2018 property return in postcodes with venues of rising popularities and where μ_{London} is the corresponding expectation for the rest of London.

Results

Table 2 displays the 12 eateries in the ten OPCs that had the largest increase popularity post-2018. On average, customers visit these venues from 20 distinct connected venues per month. They travel on average 3.9 kilometres to visit these eateries. The surge in popularity, as measured by our degree centrality indicator, pre- and post-2018 was 5.89 on average.

Table 2: Characteristics of selected eateries

Name	Category	OPC	Nodes	Dist	Surge
What the Pitta	Vegan Restaurant	E1	19	4.0	4.5
Leon	Café	E14	36	3.7	2.3
Starbucks	Coffee Shop	E16	11	4.0	5.9
The Spread Eagle Bar	Bar	E2	9	1.6	2.3
Costa Coffee	Coffee Shop	E3	21	4.6	2.2
Starbucks	Coffee Shop	NW1	21	3.0	31.0
Goodfare Restaurant	Italian Restaurant	NW1	7	3.1	5.0
Panzer's	Deli	NW8	22	3.2	4.0
Costa Coffee	Coffee Shop	SE18	20	6.4	3.0
Pret A Manger	Sandwich Place	SW5	21	4.1	3.0
Starbucks	Coffee Shop	W6	32	4.7	2.8
Pret A Manger	Sandwich Place	W6	21	4.0	3.0

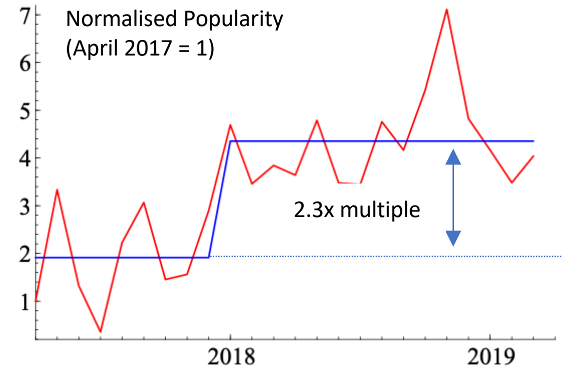
Note: 'Nodes' refer to the number of interactions that the venue has with distinct other venues on average per month. 'Dist' refers to the average distance to the connected venues measured in kilometres. 'Surge' refers to the multiple of Foursquare check-ins before and after January 2018. OPC stands for outward postcode.

Figure 1 showcases an example for the surge venue popularity. We fitted the time-varying mean to the time-series to highlight the structural break in its popularity.

We find high predictive power of the Foursquare venue interaction data for out-of-sample property price movements. On average, OPC districts with one of the selected London venues, exhibit a 4% higher property price growth than postcode districts without these venues. Figure 1 illustrates the differences in 2018 property returns. A

Welch test at a 5% significance level rejects the null hypothesis that the expected return for OPCs of eateries with rising popularity is equal to the return of all London OPCs. Thus we have established *eateries gentrification* in London, which we estimate to be 4% in 2018.

Figure 3: Popularity of "The Spread Eagle Bar"



These results hold in a difficult market environment for London residential property, which has been dominated by concerns over the outcome of the Brexit referendum. This explains why the property price returns in 2018 for the selected OPCs is still modest at an annual growth rate of 5%.

Table 3: The 2018 property returns on selected OPCs

Name	OPC	Return
Leon	E14	16.5%
The Spread Eagle Bar	E2	10.5%
Starbucks	E16	6.2%
Pret A Manger	SW5	6.1%
Costa Coffee	SE18	5.2%
What The Pitta	E1	4.5%
Starbucks & Goodfare	NW1	3.8%
Costa Coffee	E3	1.1%
Starbucks & Pret A Manger	W6	-0.2%
Panzer's	NW8	-0.7%

Average Price Increase in selected OPCs	4.8%
Average prince increase in all London OPCs	1.1%
T-statistic for Welch's test	2.5
P-value of Welch's test	3.2%

Note: OPC stands for outward postcode.

Conclusion

The goal of this paper is to both test the hypothesis of the existence of *eateries gentrification* and to estimate the magnitude of this effect. We construct a measure for the popularity of restaurants and show that postcodes with cafés, restaurants and bars of rising popularity exhibit higher investment returns on residential property. The effect is statistically significant. In London of 2018, the effect size has been 4 percent.

Our results raise further questions. Does the *eateries gentrification* effect change over time? Does it persist in recessions? How does it compare to supermarket gentrification or the London Crossrail gentrification? While open questions remain, one fact is undeniable. The Foursquare dataset provides valuable information to predict residential property markets.

References

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse. "A regression method for real estate price index construction." Journal of the American Statistical Association 58, no. 304 (1963): 933-942.

Fama, Eugene F., and Kenneth R. French. "Common risk factors in the returns on stocks and bonds." *Journal of financial economics* 33, no. 1 (1993): 3-56.

Glaeser, Edward L., Hyunjin Kim, and Michael Luca. "Nowcasting gentrification: using yelp data to quantify neighborhood change." In *AEA Papers and Proceedings*, vol. 108, pp. 77-82. 2018.

Helliwell, John F., Richard Layard, and Jeffrey Sachs. "World happiness report." (2012).

Henneberry, John. "Transport investment and house prices." Journal of Property Valuation and Investment 16, no. 2 (1998): 144-158.

Opsahl, Tore, Filip Agneessens, and John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." Social networks 32, no. 3 (2010): 245-251.

Welch, Bernard L. "The generalization ofstudent's' problem when several different population variances are involved." Biometrika 34, no. 1/2 (1947): 28-35.

Zhou, Xiao, Desislava Hristova, Anastasios Noulas, Cecilia Mascolo, and Max Sklar. "Cultural investment and urban socio-economic development: a geosocial network approach." Royal Society open science 4, no. 9 (2017): 170413.

Metaheuristic macro scale traffic flow optimisation from urban movement data

Laurens Arp[*]
Dyon van Vreumingen[*]
{l.r.arp,d.van.vreumingen}@umail.leidenuniv.nl
Leiden Institute for Advanced Computer Science
Leiden, The Netherlands

Daniela Gawehns
Mitra Baratchi
{d.gawehns,m.baratchi}@liacs.leidenuniv.nl
Leiden Institute for Advanced Computer Science
Leiden, The Netherlands

ABSTRACT

How can urban movement data be exploited in order to improve the flow of traffic within a city? Movement data provides valuable information about routes and specific roads that people are likely to drive on. This allows us to pinpoint roads that occur in many routes and are thus sensitive to congestion. Redistributing some of the traffic to avoid unnecessary use of these roads could be a key factor in improving traffic flow.Many proposed approaches to combat congestion are either static or do not incorporate any movement data. In this work, we present a method to redistribute traffic through the introduction of externally imposed variable costs to each road segment, assuming that all drivers seek to drive the cheapest route. We use a metaheuristic optimisation approach to minimise total travel times by optimising a set of road-specific variable cost parameters, which are used as input for an objective function based on traffic flow theory. The optimisation scenario for the city centre of Tokyo considered in this paper was defined using public spatial road network data, and movement data acquired from Foursquare. Experimental results shows that our proposed scenario has the potential to achieve a 62.6% improvement of total travel time in Tokyo compared to that of a currently operational road network configuration, with no imposed variable costs.

KEYWORDS

traffic flow, urban movements, metaheuristics

1 INTRODUCTION

Even though extensive road networks have been developed to satisfy the high demand for vehicular transportation, overoccupancy of roads still occurs on a daily basis, causing traffic jams which hurt the environment, the economy and the drivers' moods. Finding a solution to traffic congestion is a challenging problem that has occupied many in the past century. After all, traffic dynamics are difficult to predict, due to complex fluctuations in traffic demand, both spatial and temporal. This makes it hard to devise a protocol for traffic flow redistribution that works well in varying conditions.

To date, various approaches have been proposed to alleviate congestion in some way [8]. However, these methods tend to be either static, data-independent protocols, micro-scale solutions (on the level of individual roads) or primarily driven by theoretical models. Our objective instead is to construct a dynamic, data-driven, macro-scale (road network level) approach to address traffic congestion. In this sense, dynamic means that a solution can be adapted to new traffic data with relative ease.

In this work, we propose a method for traffic redistribution fueled by metaheuristic optimisation, which we test on the case of the city centre of Tokyo. We seek to shift the traffic situation away from a state where each driver chooses the fastest, or shortest, route (thus causing congestion on roads that occur in many shortest routes), towards a system optimal equilibrium, as coined by Wardrop [14], where the *total travel time* for all drivers is minimised. By introducing externally imposed variable costs (e.g. tolls, or any other financial or non-financial method a supervisory institution might deploy) on each road, we aim to discourage drivers from all taking the same congested roads. This approach asserts that, on average, each driver is willing to take the cheapest route from their point of departure to the destination, where the total costs to drive a route depend both on the distance travelled, through a spatial cost, and the imposed variable costs encountered along the route.

In order to make predictions of traffic flow and occurrence of congestion, we infer traffic demand from a data set of urban movements. A number of public traffic data sets, such as the Dutch NDW [11], report the traffic flow or density at certain points in time; however, while this gives a detailed picture of a local situation, it provides no information as to what routes drivers are following. Hence, such data is of little use when we wish to redistribute traffic by encouraging sensible alternative routes. For this reason, we use a data set of urban movements, provided by Foursquare as part of the *Future cities challenge*, which allows for the inference of traffic level information needed for this research such as the origin and destination of movements [2].

2 RELATED WORK

The objective of combatting traffic congestion by altering road network setups has been addressed in a large body of work. The use of road pricing as a means to achieve this goal is a prevalent approach [6, 13, 15, 16]. In this context, the marginal cost of congestion is a frequently employed measure to assess the optimal road pricing. A key difference between these papers and our work is that the previously proposed road pricing policies are fixed (e.g. charge a fee within a certain radius of the city centre) instead of dynamic, and do not follow from an optimisation procedure based on actual movement data. Approaches for optimising road networks and traffic flow from a different viewpoint, unrelated to road pricing policies, include metaheuristic optimisation of road improvements [3], development of intelligent traffic light systems [10], optimisation of road graph architectures with evolutionary algorithms [12], and prediction of optimal traffic flow through maximum-entropy methods [9]. A more exhaustive list of methods is provided by Kumar Shukla and Agrawal [8]. In this work, we explore the use of optimisation

[*] Both authors contributed equally to this research.

algorithms in proposing a dynamic pricing mechanism using actual movement data.

3 PROBLEM STATEMENT

The problem of optimising traffic flow through adaptive road pricing is twofold. First, we must estimate traffic flow and congestion in a road network, which is the underlying cause of high total travel time when all drivers follow the cheapest routes from their origins to their destinations. By combining movement data describing the traffic demand in the network and the spatial road network data, traffic flow theory yields these estimations. The demand data is a set D_M consisting of movements between venue locations within the road network. Each element of this set (indexed as k) is a tuple (A_k, B_k, N_k) where A_k and B_k are elements of a venue data set D_V containing spatial information about the venues, and N_k is the recorded frequency of the specific movement from A_k to B_k. Second, having found a method to express the total travel time as a function of the variable cost parameters, we aim to optimise the parameters for minimal total travel time. Our methods for addressing this optimisation problem are set out in detail in the next section.

4 METHODS

4.1 Traffic flow estimation

4.1.1 Road network and routing model. The first step towards prediction and minimisation of congestion is to represent the physical road network as a planar graph G that has road segments for edges, which may be traversed in order to travel from an origin to a destination. Specifically, the graph is a tuple $G = (V, E, S)$ with V the node set, E the edge set and S the set of Haversine lengths of all edges. The node set V contains intersections in the road network, as well as nodes for the origin and destination locations from D_V.

We then introduce, for each road segment $(i, j) \in E$ in the graph, a cost that a driver needs to pay to traverse this segment. The main part of this cost is a *variable cost* p_{ij}. All variable cost parameters collectively form the variable cost vector P which we seek to optimise for minimum congestion. Next, we assign to each segment a *spatial cost*, which is an immutable base cost for travelling from node i to j that is linearly dependent on the length s_{ij} of the segment by a tunable factor β_s. Since the movement data is aggregated into frequency numbers, and is not provided on an individual level for anonymity reasons, we take β_s to be equal for all drivers. Put together, the total cost for a driver to travel via a connected route of segments $R = \{(i, j)\}$ from some origin to a destination, is given by the sum of the individual segment costs occurring on the route:

$$\mathrm{cost}(R) = \sum_{(i,j)\in R} \beta_s s_{ij} + p_{ij}. \tag{1}$$

For the development of traffic flow, we assert that all drivers are selfish and seek to drive the route which incurs the lowest total cost. These routes can be found using a weighted shortest path algorithm. Note that if $p_{ij} = 0$ on all edges, each driver will drive the route of lowest spatial cost, which is exactly the shortest route. From the cheapest routes, which are jointed collections of segments, and the frequency numbers N_k, we can predict the vehicle count n_{ij} on each segment, from which the degree of congestion is computed as set out in the following subsection.

4.1.2 Congestion model. In our congestion model, we assume that the flow of traffic on a road segment is fully described by a Greenshields fundamental traffic flow curve [4], which is a widely used theoretical model for predicting traffic dynamics on a road segment. The used variables are flow f (vehicles passing by per unit time) and density ρ (vehicles present per unit length). In this model, there exists a *maximum density* ρ_m that the road can support, beyond which the total flow is zero. Furthermore, there is a *critical density* ρ_c at which the flow reaches its maximum value: $f(\rho_c) = f_m$. Naturally, $f(0) = 0$, as no flow exists when no cars are present.

A basic curve that fits this description is a concave quadratic function, with zero flow at $\rho > \rho_m$, which we define as

$$f(\rho) = \frac{f_m}{\rho_c^2} \max\left(0, \rho[2\rho_c - \rho]\right), \tag{2}$$

where we note that, since f is quadratic in ρ, $\rho_m = 2\rho_c$. The maximum flow is directly related to the critical density; assuming that the traffic is able to drive at the maximum allowed speed v_m when the density is at its critical point, we set $f_m = \rho_c v_m$.

From this density-flow dependence, we can extract the *space mean speed* $v(\rho)$, the average speed of all vehicles on the road segment, as [4]

$$v(\rho) = \min\left(v_m, \frac{f(\rho)}{\rho}\right), \tag{3}$$

where again the maximum speed enters the relation, this time as a bound on the space mean speed. Finally, the *space mean travel time* $t(\rho)$ on the segment, taken to have length s, is computed as

$$t(\rho) = \frac{s}{v(\rho)}. \tag{4}$$

From the expected number of vehicles n_{ij} on each road segment, we obtain the segment density as $\rho_{ij} = n_{ij}/s_{ij}$. For multi-lane roads, we divide this number by the number of lanes. By inserting the segment density into the density-time relation described above (eq. 4), we find the space mean time t_{ij} spent on the segment. The collection of segment travel times then leads to the definition of the objective function for optimisation by a metaheuristic algorithm.

4.2 Parameter optimisation

4.2.1 Objective function. The objective function computes the measure obj(P), which reflects the extent to which the system optimal equilibrium is reached by a variable cost configuration P. This equilibrium occurs when the total travel times for each driver on their routes are minimal. As such, we define the objective function as the total space mean travel time over all routes driven on the road network. This total travel time can be conveniently expressed using the segment vehicle counts n_{ij}, which are directly dependent on P:

$$\mathrm{obj}(\mathrm{P}) = \sum_{(i,j)\in E} n_{ij}(\mathrm{P})\, t_{ij}(\mathrm{P}). \tag{5}$$

Note that each segment mean travel time t_{ij} is also a function of P since it is dependent on the vehicle count n_{ij}. Algorithm 1 shows, in pseudocode, the routine to compute the objective function.

4.2.2 Optimisation algorithms. The purpose of the optimisation algorithms is to find an optimal variable cost configuration such that the value of the objective function, the total travel time, is minimised. In principle, any black-box metaheuristic optimisation

```
Data: road network graph $G = (V, E, S)$ with node set $V$ inferred
        from venue data $D_V$, $E$ and $S$ inferred from road network data;
        movement data $D_M$;
        spatial cost factor $\beta_s$
Input: parameter vector P
Result: objective function value obj(P)

initialise $n_{ij} \leftarrow 0$, $t_{ij} \leftarrow 0$, $\rho_{ij} \leftarrow 0$ for all $(i, j) \in E$

// Compute predicted vehicle counts $n_{ij}$ on each segment
forall origin-destination-freq. tuples $(A_k, B_k, N_k)$ in $D_M$ do
    Find cheapest route $R_k$ from $A_k$ to $B_k$ according to $\beta_s$ and P
      using weighted shortest-path algorithm
    forall $(i, j) \in R_k$ do
        $n_{ij} \leftarrow n_{ij} + N_k$
    end
end
// Find mean travel times $t_{ij}$ on each segment
forall $(i, j) \in E$ do
    $\rho_{ij} \leftarrow n_{ij}/s_{ij}$
    $t_{ij} \leftarrow t(\rho_{ij})$                (eqs. 2–4)
end
obj(P) $\leftarrow \sum_{(i,j)\in E} n_{ij}\, t_{ij}$        (eq. 5)
return obj(P)
```

Algorithm 1. Routine for computing the objective function for a given parameter vector P.

algorithm could be used to search for local optima of variable cost configurations which might approximate a system optimal equilibrium. That said, for these purposes, algorithms which are robust for high-dimensional problems are preferred, as the number of parameters increases proportionally to the number of edges in the graph.

For our proof-of-concept implementation we use simulated annealing (SA) [7], which is a variation of hill climbing where worse solutions can get accepted depending on the algorithm's decreasing *temperature* value, and a genetic algorithm (GA) adapted for continuous optimisation. For both algorithms each iteration contains 40 objective function evaluations; after one iteration, the GA updates its population, whereas SA resets its temperature value. Both algorithms use mutations generated using a normal distribution with zero mean and unit variance, at a mutation rate of 0.2 per parameter.

5 CASE STUDY: TOKYO CITY CENTRE

In order to test our traffic flow optimisation method, we applied it to movements inside the city centre of Tokyo (i.e. excluding the Greater Tokyo area). We briefly discuss the movements and road network used for the case study, and present experimental results.

5.1 Movement data

The movement data was provided by Foursquare as part of the *Future cities challenge* [2]; we selected only those parts related to the Tokyo city centre. The data contains a list of venues together with their GPS locations, forming the data set D_V introduced in section 3, and a list of movements between the venues. The movement data contains movements of the same form as the tuples in D_M, but with additional indications of the month during which the movement occurred, and the time of the day (periods of 4 to 6 hours). We considered only movements made in the afternoon, and combined the frequencies N_k for the same movement in different months into a single figure. Since the Foursquare dataset does not cover the entirety of vehicular movements inside Tokyo (and, in fact, also includes other types of movements such as subway, biking and walking trips), we viewed the frequencies as ratios rather than absolute numbers, asserting the law of large numbers for sufficient accuracy. We then normalised the frequencies to numbers that the road graph we constructed (see next subsection) could support.

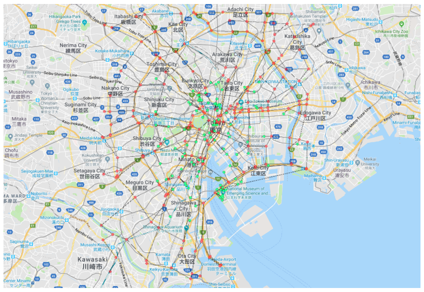


Figure 1. Road network graph example. Green nodes are venues, red nodes are intersections. Map source: Google Maps [5].

As a last modification, we selected the venues occurring in the 100 most popular (i.e. frequent) routes, and clustered all other venues together with their nearest neighbour (in terms of Haversine distance) from the set of most popular venues. The routes were clustered accordingly, going between venue clusters instead of individual venues. This was done in order to substantially reduce the number of routes and therewith the computational complexity of the problem.

5.2 Road network data

The road network data used is based on the Asia shapefile provided by the Earthdata *Global roads open access data set* [1]. It contains information on the road networks of the entirety of Asia with a variable resolution; for the city Tokyo, its resolution is well suited for the algorithm. The road data is translated into a graph representation by finding intersections between lines, and turning these into intersection nodes. The lines themselves are used to create edges between intersections. Venue nodes are created by identifying the location of the venues from the venue data set, which are connected to the nearest intersection node.

5.3 Experimental results

The improvement progress for both optimisation algorithms is shown in figures 3 and 4. For comparison, we ran the objective function once with all variable costs set to 0 to obtain the default flow of the road network. This configuration had an objective value of around 416 million hours spent on the road network. It should be noted that the objective function values are higher than what would be a realistic amount of time spent on roads. As the theoretical traffic flow models do not take relief methods into account for fully congested roads, the speed on those roads in considered to be zero. In these cases, we use an arbitrary low value to represent the speed on the road, resulting in somewhat unrealistic amounts of hours.

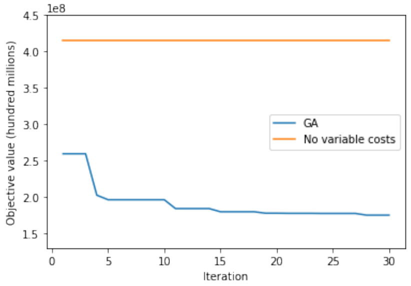


Figure 2. Objective function value plotted to the number of GA iterations. Values are the total time spent on the roads over 6 hours in the afternoon.

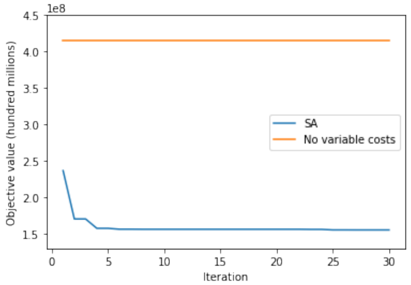


Figure 3. Objective function value plotted with the number of SA iterations. Values are the total time spent on the roads over 6 hours in the afternoon.

Both the GA and SA were able to find solutions which substantially improved traffic flow over the network. After 30 iterations, the lowest fitness value of the GA was around 175 million hours (an improvement of 57.9%), and the lowest fitness value of SA was around 155 million hours (an improvement of 62.6%). Effectively, this means the solution found by SA was able to reduce the total amount of hours spent over a period of 6 hours in the afternoon by 261 million, due to having found a good distribution of variable costs for each road segment. Although the results of the GA were not as good as those of SA, the GA was still successful in improving the unoptimised configuration, which supports the view that any optimisation algorithm could be used for these purposes.

Interestingly, even the first iterations of the algorithms showed values which were an improvement compared to the default traffic flow. This is likely due to the nature of the shortest path algorithm when taking only distance into account, which sends many cars over the same roads unnecessarily. Though this type of behaviour was the basic premise allowing us to optimise, given the large variety of available, very similar roads in the network, any deviation from the over-use of single roads caused by the random initialisation of variable costs would result in better traffic flow. From such a randomly initialised state, both algorithms then improved the solutions such that the more well-suited alternatives were used.

The results show that the algorithms were effective in finding solutions improving traffic flow over the default setting of no variable costs. That said, there is no guarantee the optima the algorithms converged to were global optima, nor that the convergence was as fast as possible. Future work could include more thorough exploration of optimisation algorithms and their parameter settings.

6 CONCLUSION

We have shown that we can successfully address traffic congestion by redistributing traffic through imposing of variable road segment costs, and optimising this cost configuration using metaheuristic algorithms. The best variable costs configuration was found by a simulated annealing routine, improving upon the total travel time corresponding to a configuration with zero variable costs by 62.6%. Both simulated annealing and a genetic algorithm were effective at optimising solutions.

Though the practical implementation of the variable costs may be another non-trivial problem to address first, the positive results show that, at least conceptually, this method could result in improved traffic flow when applied in practice. Consequently, cities may enjoy shorter travel times, better accessibility, cleaner air and, not unimportantly, improved drivers' moods.

ACKNOWLEDGMENTS

We would like to extend special thanks to Foursquare and Netmob for the movement data and to Giorgos Kyrziridis for his support.

REFERENCES

[1] NASA Earthdata. [n. d.]. Global Roads Open Access Data Set. https://earthdata.nasa.gov/. Accessed: 16-05-2019.

[2] Foursquare. 2019. Future Cities Challenge. https://www.futurecitieschallenge.com/. Accessed: 16-05-2019.

[3] Mariano Gallo, Luca D'Acierno, and Bruno Montella. 2010. A meta-heuristic approach for solving the urban network design problem. *European Journal of Operational Research* (2010). https://doi.org/10.1016/j.ejor.2009.02.026

[4] Bruce D. Greenshields, J. T. Thompson, H. C. Dickinson, and R. S. Swinton. 1934. The photographic method of studying traffic behavior.

[5] Google Inc. [n. d.]. Google Maps. https://www.google.nl/maps. Accessed: 21-05-2019.

[6] Theodore E. Keeler and Kenneth A. Small. 2002. Optimal peak-load pricing, investment, and service levels on urban expressways. *Journal of Political Economy* (2002). https://doi.org/10.1086/260543

[7] Scott Kirkpatrick, Charles Daniel Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220 4598 (1983), 671–80.

[8] Sanjiv Kumar Shukla and Anupam Agrawal. 2015. Present and future perspective on optimization of road network management. *International Journal of Computer Trends and Technology* 22, 2 (2015), 64–67. https://doi.org/10.14445/22312803/ijctt-v22p112

[9] L. Liu. 2008. Optimization of traffic flow in a road network. In *2008 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Vol. 1. 1244–1247. https://doi.org/10.1109/ICICTA.2008.21

[10] Pallavi Mandhare, Vilas Kharat, and C.Y. Patil. 2018. Intelligent road traffic control system for traffic congestion: a perspective. *International Journal of Computer Sciences and Engineering* 6 (07 2018), 908–915. https://doi.org/10.26438/ijcse/v6i7.908915

[11] NDW. [n. d.]. Nationale Databank Wegverkeersgegevens. https://www.ndw.nu/. Accessed: 21-05-2019.

[12] Frank Schweitzer, Helge Rosé, Werner Ebeling, and Olaf Weiss. 1997. Optimization of road networks using evolutionary strategies. *Evolutionary Computation* (1997). https://doi.org/10.1162/evco.1997.5.4.419

[13] Alan A. Walters. 1962. The theory and measurement of private and social cost of highway congestion. *Econometrica* (1962). https://doi.org/10.2307/1911814

[14] John G. Wardrop. 1952. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers Part 2.* https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[15] Jun Yang, Avralt-Od Purevjav, and Shanjun Li. 2017. The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in Beijing. (2017). https://doi.org/10.2139/ssrn.2948619

[16] Sun Ye. 2012. Research on urban road traffic congestion charging based on sustainable development. *Physics Procedia* (2012). https://doi.org/10.1016/j.phpro.2012.02.231

PolisNet: Designing Future Cities with Deep Neural Networks

Gianni Barlacchi*
gianni.barlacchi@gmail.com
Fondazione Bruno Kessler
Trento, Italy

Marco De Nadai*
denadai@fbk.eu
Fondazione Bruno Kessler
Trento, Italy

Yahui Liu*
yliu@fbk.eu
Fondazione Bruno Kessler
Trento, Italy

Bruno Lepri*
lepri@fbk.eu
Fondazione Bruno Kessler
Trento, Italy

Luca Pappalardo*
luca.pappalardo@isti.cnr.it
ISTI-CNR
Pisa, Italy

ABSTRACT

Cities are undoubting the protagonists of world migration of the last decades. People have increasingly moved to urban centers attracted by a greater pool of opportunities, jobs and education at eased reach thanks to shorter distances and public transportation. Recently, the advent of new data sources and methods has acted as a real game-changer in urban studies. The digitalization of society has allowed to observe, understand end even predict many aspects of human mobility and social behavior. However, despite the increasing ability to describe human behavior in cities, how to exploit big data to improve the design of urban areas is still unclear. Starting from the data provided by the Future Cities Challenge, we propose PolisNet, an AI-aided approach to design and modify urban areas to make them *successful.* This framework is based on Deep Neural Networks (DNNs) that exploits urban imagery and Point Of Interests (POIs) to propose *what* can be improved in an urban zone, *how*, and *where.*

1 INTRODUCTION

Urban planners, sociologists and policymakers have always been concerned about improving distressed districts and create the conditions to spur urban life. Neighbourhoods, in particular, are the most fundamental unit of city life that shape the crime [16], diversity [8] and even the individual's economic success in life [18]. However, there is still little guidance on designing them and *what* improve on existing neighborhoods, *how*, and *where.*

Here, we propose PolisNet, an AI-aided approach to design and modify neighborhoods to make them successful. This framework is based on a Deep Neural Network (DNN) that exploits urban imagery and Point Of Interests (POIs) data to propose *what* can be improved in a neighborhood, how, and *where.* In particular, we empirically describe each neighborhood in a city through an OpenStreetMap [1] aerial image, which describes the built environment, its POIs and the metrics of urban success, which is based on the well-established vitality of a neighborhood [3, 8, 21]. This metric was previously used to prescribe planning principles and sustainable human activity and urban development [21]. By leveraging aerial image, POIs and success data, PolisNet perturbs the aerial image of each urban zone to propose *where* it might be modified and *what* POI to add or eliminate to make the urban zone more successful. For example, given an aerial image of south-central Los Angeles, PolisNet suggests the type and the geographical position of POIs that, added or deleted, increase the success of south-central Los Angeles.

To the best of our knowledge, this is the first research effort towards the creation of AI-aided design systems that could be a valid support for policymakers and city architects, and help reduce the considerable costs related to urban design and simulation. We anticipate that our network will be a groundbreaking system to improve urban life and design neighborhoods with lower crime rate, cultural and economic distress.

The rest of the paper is organized as follows. Section 2 provides an overview on urban related works, whereas Section 3 presents an in-depth description of system. In Section 4, we characterize the employed datasets and explain our experimental setup, together with the obtained results. Finally, conclusions and future research are drawn in Section 5.

2 RELATED WORK

Our article primary speaks to two streams of literature. The first one is the community of urban computing, which studies human behaviour in cities through massive data about human displacements, socio-economic and geographical data. The second one is the stream of literature that tries to understand, recommend and design cities that work by collaborating with local authorities, communities and policymakers.

In Urban planning, many approaches tried to study neighborhoods and their revitalization. Notably, Jane Jacobs with her *The Death and Life of Great American Cities* [8] influenced many urban plans in the world. In her book she praised diversity as an ingredient to have convenient social interactions, face-to-face encounters, and a spontaneous sense of community that spurs economy, security and urban life. More recently, some guidelines and policies have been shared to revitalize neighborhoods [10, 19] based on high-tech entrepreneurship, walkability, and a sense of community similar to the old-fashioned Jacobs' ideas.

In Urban computing, many approaches tried to model and predict the activities and characteristics of urban areas. Notably, Noulas *et al.* [13] proposed a model for predicting an urban zone's prominent activity using Foursquare and mobile phone data. Zhang *et al.* [22] proposed an online, semi-supervised, and multimodal embedding method for geo-located information with space, time and text.

Several works have also targeted land use classification and functional area detection. For example, Yuan *et al.* [20] propose a framework to classify functionalities of an area for the city of Beijing

*Authors equally contributed to the manuscript.
[1] https://www.openstreetmap.org

using POIs and trajectories of taxis. Barlacchi *et al.* [2] propose a novel machine learning representation based on the encoding of Foursquare POIs to classify the most prominent land use of an urban area.

Another strand of research focus on defining measures of success for a city or urban zone. Yue *et al.* [21] define *urban vitality* as the capacity of an urban environment to boost social activities. They discover that urban vitality showed a positive correlation with phone usage density, and that the urban structure of Shanghai plays a crucial role in its urban vitality. De Nadai *et al.* [3] use the average number of mobile Internet connections throughout a typical business day, divided by the an urban zone's area, as a proxy for urban vitality.

3 POLISNET SYSTEM

Our framework, PolisNet, is an AI-based framework that uses digital data and Deep Neural Networks (DNNs) to design a neighborhood and propose *what* improve, *how*, and *where*. The architecture of PolisNet is illustrated in Figure 3.

3.1 Urban Success Metric (USM)

Defining urban success is not an easy task and many scholars have tried to define it through extensive studies [14]. Although urban success can manifest itself in different ways, we straightforwardly define what is *not*: an area without visitors. Thus, we build upon the sociological theory of Jane Jacobs, which defines vitality and diversity as an essential factors for urban success. Recent work in urban computing empirically defined vitality by measuring the activity of people in a neighborhood from mobile phone data [3, 4]. Similarly, we here define as vital a place where people geo-localize through Foursquare data. Thus, vitality in a neighborhood i is formalized as:

$$\text{vitality} = \sum_{p \in P_i} C(p) \tag{1}$$

where P_i is the set of POIs in the neighborhood and $C(p)$ is the number of check-ins for a particular POI. Similarly, we define the economic vitality counting only the vitality of the Shops and Food POIs. Note that our framework might be used with different metrics such as the number of crimes and the number of cultural events in a neighborhood.

3.2 Urban Success Evaluator (USE)

As shown in Figure 1, we design an Urban Success Evaluator (USE) to distinguish the zone to be successful or not, which can be formulated as a binary classification problem, where "0" and "1" refer to `unsuccessful` and `successful`, respectively. Recently, deep convolutional neural networks (CNNs) have led to a series of breakthroughs for image classification [6, 11]. Thus, we build our model based on two state-of-the-art methods, including Inception-v3 [17] and ResNet-v2 [7], which are widely used as basic networks in various computer vision tasks.

Specially, we define the training set as $\mathcal{S} = \{(\mathbf{I}_n, \mathbf{M}_n, C_n), n = 1, 2, \ldots, N\}$, where the $\mathbf{I}_n$ denotes the original map image, $\mathbf{M}_n$ denotes the mask of POIs, and $C_n \in \{0, 1\}$ denotes the binary classes. To train our model, we calculate the cross-entropy loss:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{n=1}^{N} C_n \log(p(C_n)) + (1 - C_n) \log(1 - p(C_n)), \tag{2}$$

where $p(C_n)$ refers to the probability of zone $(\mathbf{I}_n, \mathbf{M}_n)$ to be correctly classified. For simplicity, we set equal loss weights for both classes.

3.3 Urban Success Designer (USD)

We then design an Urban Success Designer (USD) to model the distribution of POIs in the spatial space in different cities, which shows the potential ability that can be applied to propose ideal new POIs to transform an `unsuccessful` zone in a `successful` one. We re-formulate such kind of problem into a location regression problem. Inspired by fast object detection methods [12, 15], we assume that each POI can be enlarged with a buffer area (formed as bounding box in object detection). After that, our approach discretizes the output space of POI locations into a set of default boxes over different aspect ratios per feature map location. At prediction time, the network generates scores for the presence of each POI category in each default box and produces adjustments to the box to better match the shape of the POI buffer.

Figure 2 shows the pipeline of our USD, in which the training objective is derived from the MultiBox objective [12] but is specified to handle the regression of multiple POI categories. In contrast with previous object detection methods, we try to predict location and category of the removed POI. Thus, we define the training set as $\mathcal{Y} = \{(\mathbf{I}_n, \mathbf{M}_n^{\dagger}, \mathbf{p}^{\dagger}, C^{\dagger}), n = 1, 2, \ldots, N\}$, where the $\mathbf{I}_n$ denotes the original map image, $\mathbf{M}_n^{\dagger}$ denotes the mask of POIs with a randomly removed POI (located in $\mathbf{p}^{\dagger}$ with a category $C^{\dagger} \in \{1, 2, \ldots, K\}$, K refers to number of POI categories). The overall objective loss function is a weighted sum of the localization loss (loc) and the confidence loss (conf):

$$\mathcal{L}_{reg} = \frac{1}{N}(\mathcal{L}_{conf}(x, C^{\dagger}) + \alpha \mathcal{L}_{loc}(x, l, g)) \tag{3}$$

where $x_{ij}^{C} \in \{0, 1\}$ be an indicator for matching the i-th default box to the j-th ground truth box of category C. The localization loss is a Smooth L1 loss [5] between the predicted box (l) and the ground truth box (g) parameters. The confidence loss is the softmax loss over multiple classes confidences (C) [12].

3.4 PolisNet Framework

First, we define a neighborhood as a uniform and non-overlapping regular cell of approximately 600 × 600 meters based on the Slippy Map standard [2] at zoom 15. Then, we associate each cell with the corresponding OpenStreetMap aerial image, the metrics of urban success, and the map mask built over the characteristics of the POIs contained in the cell. For each unsuccessful urban zone, i.e., an urban zone with the success metric lower than 0.5, PolisNet proposes several changes in order to make the zone successful.

The design process of PolisNet is divided into two phases: in the first phase, the USE evaluates whether an urban zone needs to be redesigned or not. In the case the evaluation reveals that the zone is successful, the system outputs the new generated successful area indicating the required changes: (i) *where* it requires changes,

[2] https://wiki.openstreetmap.org/wiki/Slippy_Map

and (ii) *what* POIs should be added in such locations. Differently, if the zone needs refinements, the urban zone representation is forwarded to the second step. In this step the USD, by relying on the location regression model, proposes a new list of possible POIs, with their corresponding geographical locations, that can be added in order to turn the zone into a successful one. The two steps are repeated in a loop for a maximum number of iterations or until when the urban zone does not turn into a successful zone.

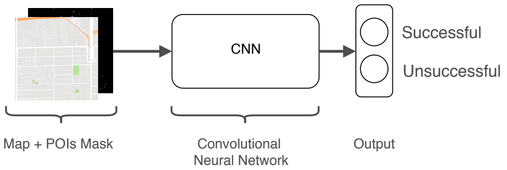


Figure 1: Schema of the USE classifier.

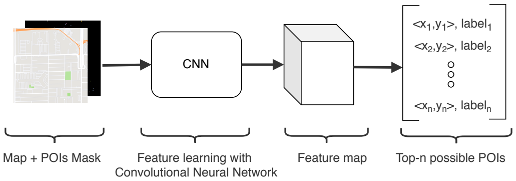


Figure 2: Schema of the USD model: given an input area, it decides *where* and *what* to add to make the input area successful.

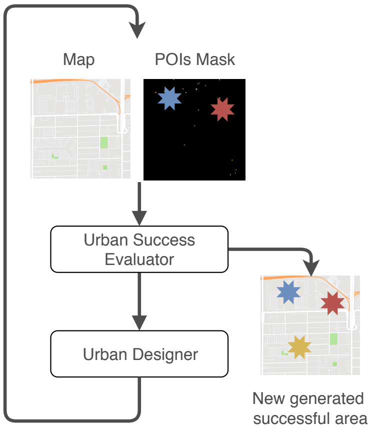


Figure 3: Schema of PolisNet: it proposes *where* an urban zone should be modified, and *what* POI to add or eliminate to make the zone successful.

4 EXPERIMENTS

Our experiments aim at demonstrating the effectiveness of our models on two phases: (i) the urban success evaluation and (ii) the urban success design. We use a database of cities composed by 8 kinds of POIs that belong to 9 different cities around the world.

4.1 Experimental Setup

As aforementioned, we represent each neighborhood through non-overlapping square cells based on Slippy Maps of zoom 15. For each cell we download the aerial map of OpenStreetmap removing any visual information connected to the POIs. This was done through a custom map style of Mapbox [3] and it ensures higher independence between the POI mask and the Map. Then, we use the Future Cities Challenge [1] dataset to obtain the Foursquare *venues* and the *movements* information. From the *venues* we extract the category, latitude and longitude, while from the *movements* we aggregated the number of check-ins in each venue. We aggregate the categories of the venues to the highest hierarchy of the Foursquare category tree [4] and assign a color to each category. We discard the `event` and `residents` categories. For each cell we build a POIs Mask where each venue is represented with a color based on its category, and located in a (x, y) pixel depending on its geographical location in the map. Finally, we represent each neighborhood through a triplet (aerial image, POIs Mask, label), where the label is binarized into `successful` and `unsuccessful`. The label is `successful` when the vitality (Equation (1)) and the economic vitality are above the 20th percentile of the distribution of vitality for that city, `unsuccessful` otherwise. We build our data for nine cities namely Chicago, Istanbul, London, Los Angeles, New York, Paris, Seoul, Singapore, Tokyo. This resulted on 20,060 images of neighborhoods.

We train our model using Adam [9] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and batch size 4 for evaluator model and batch size 1 for designer model. For data augmentation we flip the images horizontally with a probability of 0.5. The initial learning rate is set to 0.0002. We train for 30 epochs and linearly decay the rate to zero over the last 20 epochs. It takes about several hours for each model with a single GeForce GTX 1080 Ti GPU. All these models are implemented using the deep learning framework PyTorch. In particular, the weight term α is set to 1.0 in the designer model. To measure the impact of our models as well as the baselines, we used well known metrics for assessing the accuracy of DNNs.

We use the accuracy score to assess the performances on the binary classifier in the USE. To evaluate the USD we rely on two metrics: (i) the *top-k distance error*, which is the minimal distance among the top-k predicted POIs and the target POI in the image coordinate, and (ii) the *closest accuracy*, which measures whether or not the label of closest predicted POI is equal to the target label.

4.2 Preliminary Results

In this section we present some preliminary results for the two steps that compose our PolisNet system. Table 1 shows the accuracy of the USE classifier. The Table shows that conditioning the model with the city name does not provide significant improvements. The Inception-v3 and Resnet networks with the city information have a 0.4% and -4.8% of absolute points improvement, respectively.

In Table 2 we report preliminary results for the very challenging task of proposing changes to apply to unsuccessful urban zones. The accuracy and distance of the top-k predicted results show that it is very difficult to guess the correct POI category and image

[3] https://www.mapbox.com/gallery/
[4] https://developer.foursquare.com/docs/resources/categories

Model	City-conditioned	Accuracy
Inception-v3	×	0.778
Inception-v3	✓	**0.798**
Resnet-v2-50	×	0.794
Resnet-v2-50	✓	**0.750**

Table 1: Accuracy of the USE in the classification of successful areas.

coordinates of the removed POI. However, when more POIs candidates are considered, the model shows promising results both in terms of accuracy and distance error. This suggests that potential improvements can be applied to further develop this idea. The low performances with $k < 50$ might be due to the very basic input representation we are providing to our network. Indeed, some improvements can consider (i) to design a more expressive dense representation of POIs instead of using the simple one-hot encoding representation and the POI macro-category; (ii) to better balance the dataset both in terms of city map and categories of POIs; (iii) to provide a more informative visual input than a simple map mask with colored pixels.

k	Accuracy	Distance
10	0.074 ± 0.078	155.37 ± 37.51
20	0.112 ± 0.074	140.51 ± 35.48
50	0.149 ± 0.014	125.77 ± 37.50

Table 2: Accuracy and distance error of the USD in proposing changes for a given unsuccessful urban zone.

5 CONCLUSION AND FUTURE IMPACT

In this paper we presented PolisNet, an AI-based system that recommends how to improve the success of urban zones in a city. Since the underlying idea is rather flexible, PolisNet can be improved and extended in several directions.

First, while we use in this paper a fixed threshold for defining an urban zone as successful or unsuccessful (i.e., the median of the urban vitality), we may let the user choose the threshold to use for the discrimination. In this way, a policy maker may decide, according to the available economic resources, how much to change a zone to achieve the desired level of urban success.

Second, with slight modifications, our system can be adapted to generate an optimized street network given the current mobility fluxes observed in an urban zone. In this case, PolisNet would perturb the underlying street network or the position of existing POIs, given certain constraints, in order to reduce indicators of *mobility success*, such as travel times or distances, the emergence of traffic jams, or the number of car accidents.

REFERENCES

[1] [n.d.]. Future Cities Challenge. https://www.futurecitieschallenge.com/

[2] Gianni Barlacchi, Alberto Rossi, Bruno Lepri, and Alessandro Moschitti. 2017. Structural Semantic Models for Automatic Analysis of Urban Areas. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 279–291.

[3] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. 2016. The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 413–423.

[4] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. 2016. Are safer looking neighborhoods more lively?: A multimodal investigation into urban life. In *MM*. ACM, 1127–1135.

[5] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*. Springer, 630–645.

[8] Jane Jacobs. 1961. *The death and life of American cities*. Random House.

[9] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[10] Gabe Klein and David Vega-Barachowitz. 2015. *Start-Up City: Inspiring Private and Public Entrepreneurship, Getting Projects Done, and Having Fun*. Island Press. https://www.amazon.com/Start-Up-City-Inspiring-Entrepreneurship-Projects/dp/1610916905?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1610916905

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*. Springer, 21–37.

[13] Anastasios Noulas, Cecilia Mascolo, and Enrique Frias-Martinez. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, Vol. 1. IEEE, 167–176.

[14] Amos Rapoport. 2002. The role of neighborhoods in the success of cities. *Ekistics* (2002), 145–151.

[15] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7263–7271.

[16] Robert J Sampson and W Byron Groves. 1989. Community structure and crime: Testing social-disorganization theory. *American journal of sociology* 94, 4 (1989), 774–802.

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826.

[18] Laura Tach, Sara Jacoby, Douglas J Wiebe, Terry Guerra, and Therese S Richmond. 2016. The effect of microneighborhood conditions on adult educational attainment in a subsidized housing intervention. *Housing policy debate* 26, 2 (2016), 380–397.

[19] Peter A Tatian, G Thomas Kingsley, Joe Parilla, and Rolf Pendall. 2012. Building successful neighborhoods. *Washington, DC: The Urban Institute* (2012).

[20] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *KDD*. ACM, 186–194.

[21] Wenze Yue, Yang Chen, Qun Zhang, and Yong Liu. 2019. Spatial Explicit Assessment of Urban Vitality Using Multi-Source Data: A Case of Shanghai, China. *Sustainability* 11, 3 (January 2019), 1–20. https://ideas.repec.org/a/gam/jsusta/v11y2019i3p638-d200854.html

[22] Chao Zhang, Keyang Zhang, Quan Yuan, Fangbo Tao, Luming Zhang, Tim Hanratty, and Jiawei Han. 2017. ReAct: Online Multimodal Embedding for Recency-Aware Spatiotemporal Activity Modeling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 245–254.

A Multi-Objective Approach for Optimal Store Placement

Jeroen Rook
LIACS, Leiden University
j.g.rook@umail.leidenuniv.nl

Brent Verpaalen
LIACS, Leiden University
b.a.a.verpaalen@umail.leidenuniv.nl

Daniela Gawehns
LIACS, Leiden University
d.gawehns@liacs.leidenuniv.nl

Mitra Baratchi
LIACS, Leiden University
m.baratchi@liacs.leidenuniv.nl

***Abstract* - When a store owner wants to open a new store he or she desires a location that attracts a large number of customers. Previous work has shown how location based social networks can contribute to this decision process. However, opening a store also has an impact on the surrounding neighbourhood. With the use of urban planning theories we define a score showing the impact of these store placements. We propose a framework, that selects the best venue categories for a given location in a city according to their scores for both perspectives. These scores are computed from metrics extracted using a location based social network from Foursquare. Our experiments, based on the city of New York, show that the number of suitable store categories, for a single location, are often not singular. This indicates that this multi-objective approach is necessary in solving the optimal store placement problem.**

1. INTRODUCTION

Imagine you want to open a new restaurant. Where in the city would you allocate your restaurant? As a store owner you desire a location where the characteristics of the surrounding area give you a large number of customers, resulting in an increased revenue. However, store owners need permissions from local governments, who also consider urban planning objectives of the city. These objectives are set out to maintain a liveable and economically thriving city. As a store owner it is beneficial to take these objectives into account. These objectives provide a sustainable future for the store, preventing bad cityplanning, which could be disastrous for the cities economy [1]. Earlier research used Foursquare data sets in order to predict the expected number of visitors [2]. With the addition of the governmental view on store placement we extended the optimal store placement problem. Approaching the placement problem, to our knowledge, has not been covered in previous research.

In this paper, we rewrite the optimal store placement problem by trying to find the location that is best for both the store owner and the city itself. This is done by mining features from a location based social networks data set, provided by Foursquare. We use known features, and also introduce two new features derived from a relative neighbourhood graph. With these features we predict the popularity of a location and define an objective function representing the impact on the city. These two functions allow us to obtain a Pareto optimal set of the most suitable categories for a given locations, helping potential store owners to obtain popular store locations and local governments in understanding which venue categories should be targeted for a specific location.

In this paper we focus on the following contributions:

- Creating a multi-objective approach on location based store category prediction;
- Constructing new features from a relative neighbourhood graph in order to capture the function of the neighbourhood within the city;
- Applying our proposed method to the problem of identifying optimal store locations in New York City using a Foursquare movements network.

2. RELATED WORK

Traditionally, the optimal store placement problem revolved around the central place theory, the spatial interaction theory and the theory of minimal differentiation [3]. With the rise of location based social networks (LBSNs), which combines data from social networks with real-world objects, other approaches in solving the placement problem arised. Karamshuk et. al. [2] defined this problem as a ranking problem for a set of locations using one particular type of store. Another proposed approach is to make a prediction where a category would be ranked by its popularity at a given location using matrix factorisation [4]. Other socially generated data such as written visitor reviews have shown to be able to contribute in solving this problem [5, 3]. The focus of these works are based on the optimality from a store owners perspective, which is defined as a maximum number of visitors.

Considering this problem from the point of view of urban planners is much wider than only the number of visitors, but cannot be defined on one single property. Proposals of definitions have been made [6] and commonly fall back on traditional theories. One of these theories is mixed-land use [7], which sets out to have residential, commercial, and working locations within the same neighbourhoods. This encourages non-auto commuting, which accounts for a decrease of traffic congestion, and results into less CO2 emissions. Both having a positive effect on the city.

3. METHODS

Before the most suitable venue category for a given location can be found, several steps need to be taken. Our proposed approach is performed in two phases, training and ranking. In the training phase, features for each location in the data set are constructed, and the number of check-ins per venue are calculated. These check-ins are used as our ground truth for a regression model predicting the number of check-ins from the given features. In the ranking phase, for any given location we artificially simulate a venue category placement, computing resulting features. This is done for all possible categories. For these features the predicted popularity is retrieved using the regression model created in the training phase. Also the city impact score for each of these categories is computed. In the

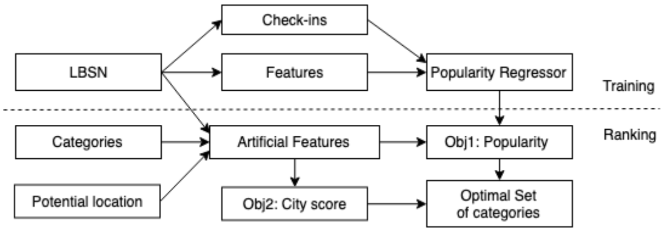


Figure 1: The full framework, for getting multiple objective rankings comparing different categories at a given location. These rankings are used to compute an optimal set of categories.

end we can obtain the Pareto optimal set of categories for that particular location, based on the two objectives. Categories in the Pareto optimal set have the best pairs of scores. All these steps and their underlying dependencies are visualised in Figure 1.

3.1 Feature engineering

In order to make rankings on popularity and city impact we hand-craft features, for each location per category, from the LSBN. We make heavy use of the features used by Karamshuk et al. [2], which are based on the work of Jensen [8]. Karamshuk et. al. extracted a total of eight different features from the location based social network, which are the first eight described features in Table 2. These features are divided in two different groups. The first group are features constructed from the geographical situation and the other group are features constructed using the movements between venues, which are called mobility. We extend this feature set by looking at the geographical behaviour of the movements in the LSBN. These features are described in 3.1.1.

All of the Karamshuk et al. features are based on the interactions occurring in the direct neighbourhood of the venue under investigation. A neighbourhood is defined as the area within a circular radius r, which is set to $200m$ for all features. The Haversine distance [9] is used as a distance metric. Due to lack of space, we refer to their paper for the full details on the construction of the features.

We will now outline the two new features Bypass and Betweenness centrality:

3.1.1 FlowGraph

The existing features do not capture the the function of the neighbourhood within the city as a whole. For example, it does not show what paths people potentially took to get from one venue to another. To get a better understanding of this we create a relative neighbourhood graph (RNG) [10] from all venues. This graph only has edges between neighbours from a geographical perspective. An edge between the venues (a, b) only appears if the following holds:

$$\forall c \in V : dist(a, c) > dist(a, b) \wedge dist(c, b) > dist(a, b) \quad (1)$$

Where V is the set of venues in the RNG and $dist$ is the Haversine distance between two venues. When all eligible edges, according to equation 1, are defined in the graph, the movements from the LSBN are mapped to corresponding edges of the shortest path between the venues in the RNG. The obtained weighted graph is referred to as FlowGraph, which shows dense areas and areas which acting as hubs for the whole graph. As a byproduct it allows for a visualisation of the interaction structure in the city, as can be seen in Figure 3. Features constructed from this graph are grouped under *Geographical Flow*.

Table 1: Overview of the constructed features extracted from the venue movements network

Feature	**Description**
Geographical	
Density	Number of neighbours
Neighbours Entropy	Spatial heterogeneity
Competitiveness	Amount of same category venues in the neighbourhood
Quality by Jensen	Amount of categories which tend to occur together
Mobility	
Area Popularity	The number of check-ins in the neighbourhood
Transition Density	The number of check-ins of venues within the neighbourhood
Flow	The number of check-in from movements outside the neighbourhood
Transition Quality	The expected number of movements from venues in the neighbourhood based on their categories.
Geographical Flow	
Bypass	Indegree of bypassing users
Betweenness centrality	Amount of times the location is on a shortest path

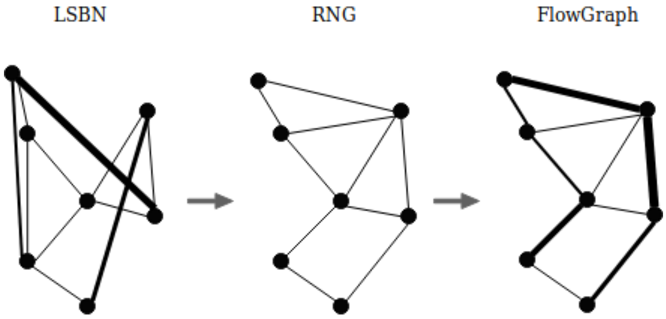


Figure 2: Sequential construction of the FlowGraph, which is build by creating a relative neighbourhood graph based on the venue locations and mapping the weights of the edges from the LSBN to all edges on the shortest paths in the RNG.

The *bypass* feature is the indegree of a location in the FlowGraph. A high value suggests that many people are passing through the selected location. This is especially useful in less occupied neighbourhoods, since people passing by are potential customers.
Another feature we propose is *betweenness centrality*, which was not considered in [2]. This centrality is defined by the frequency a node appears in the shortest paths between all nodes in the graph. This is related to the hub function for each location. For example the area which connects Long Island with Manhattan has the highest betweenness centrality.

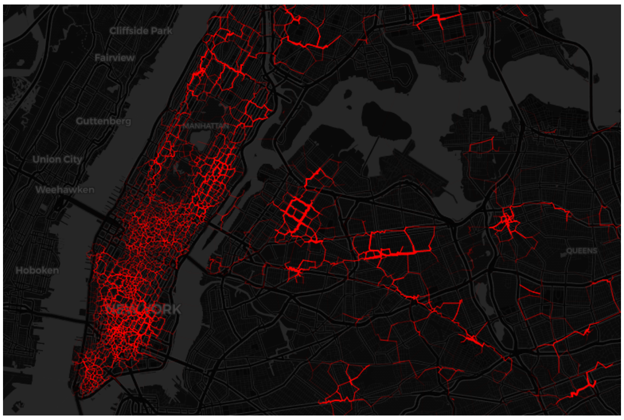


Figure 3: The FlowGraph for the venues of New York City. Busy areas and corridors can quickly be identified.

3.2 Objectives

In order to capture the venue categories that are most suitable considering both the store owners and the city perspectives we define an objective for each of them. Thus, for each category at a given location, two scores are computed. A category is covered if and only if there exists another category for which all scores are higher than it's own. From the set of categories we define the optimal set as the ones that are not covered.

3.2.1 Popularity

The benefit for the store owner is defined in terms of popularity, i.e. check-ins the venue would potentially receive. For each location we only know the popularity for the category, which is already there. For all other categories at a location this is unknown. A regression model is trained on the existing venues with their category in the data set. The model predicts for a given category and corresponding extracted features, as described in the previous section, what the popularity probably will be. With this model we can get a better understanding of how popular the venue would become for all other categories at a location. The popularity is the objective for the store owners and should be maximised.

3.2.2 City impact score

We investigated what objectives could be important from the city perspective. We found one important feature that can be extracted from data, the *Neighbours Entropy* [11, 2]. This feature represents the diversity of categories within a neighbourhood, which is in essence mixed-land use, the criteria which is regularly considered by urban planners. When different categories are given for a location the entropy changes accordingly. A high entropy means a high diversity of the surrounding, so this score should be maximised.

4. EXPERIMENTS

In this section we perform a study on the Foursquare data acquired from the city of New York. We apply our proposed methods for the city of New York in order to see if the extracted features from the FlowGraph significantly contribute towards better popularity predictions. Furthermore, we show the results for the optimal set of categories. Before these experiments can be conducted the data set is pre-processed to a suitable format.

4.1 Data set

The data set consists of two different data structures. The first structure holds information of all the known venues in the city, such as venue name, longitude, latitude, and the venue category. The other structure consists of aggregated movements between two venues. The provided movements are between two venues within the city, and also includes movements going to or leaving a single venue in the city. We defined the number of check-ins. i.e. popularity, as the total number of movements going towards a venue.

In total there are 17,382 venues for the city of New York where at least one check-in occurred. The number of distinct categories is 503, which is quite specific. To be sure this would not result in an insufficient number of training examples per category, we clustered multiple categories to prevent overfitting on a handful of venues. To do this we reduced the categories to their top-level category from Foursquare, resulting in a total of 10 categories: *Arts & Entertainment, Shops & Services, Professional & Other Places, Food, Residences, Travel & Transport, College & Universities, Outdoors & Recreation, Nightlife Spots, and Events.*

4.2 Popularity Prediction

We created venue popularity predictions for locations using different venue types and their corresponding features. For this predictions we choose to use a regressor because of the continuous property of our popularity variable. A Catboost regression model [12] is used for getting popularity predictions. Catboost uses gradient boosting on decision trees. We evaluated the performance of the algorithm using the root mean square error metric. The data set was split into a train and test set with a distribution of respectively 80 and 20 percent. The model was trained on the train set and the test set was used for evaluation, preventing data leakage. When training our algorithm we got the first row of results shown in Table 2

Table 2: Error scores (RMSE) of the popularity regressions models provided with different feature sets

	Train	Test
All features	129.18	128.54
Original features	129.28	128.65

In order to see if our constructed features improved the popularity prediction we performed the test twice using different feature sets. The first set included all described features, the second set contained only the original features excluding ours. The results show that the added features in our situation have low to no impact on the popularity prediction error. A paired t-test between predictions from both sets yielded a p-value of 0.7. This means that our proposed features do not create a significant difference in performance.

4.3 Optimal sets

From all venues in the data set we took a random sample of 1000 venues. We created features for each category available at all locations. Each location we analysed the resulting category sets. Recall that each category is assigned with two scores. As an example the categories are projected on a 2D plot where each score is on one of the axis, as can be seen in Figure 4. In the original dataset this is a Taco Shop in the Upper East Side. In total there are four categories in the optimal set, highlighted in red. These categories are thus to

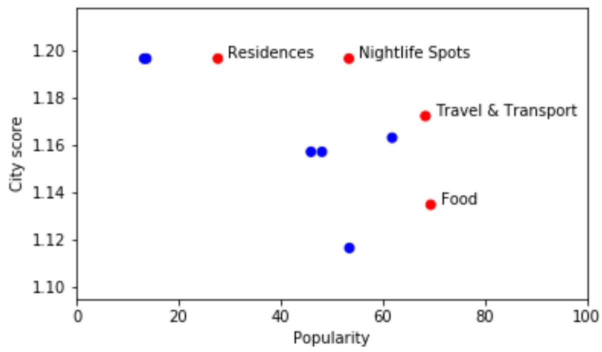


Figure 4: Projection of all categories for a location, currently used as a Taco store in the Upper East Side, with the optimal venues shown in red.

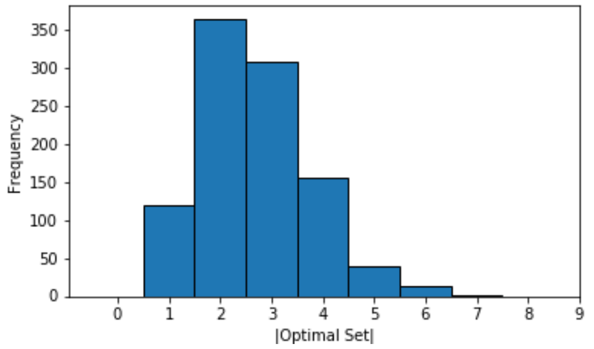


Figure 5: Histogram of the number of categories in the optimal set from a sample of 1000 locations

be considered as most suitable for that particular location. In this set *residential* and *nightlife spots* would give a higher diversity in the area and a food venue is suggested be most popular.

The histogram at Figure 5 shows the distribution of categories in the optimal set for all locations in the sample. A majority of locations have more than one category in their optimal set. Meaning that there often is a conflict between the selection of the best category between each score. This conflict shows us that the venue prediction problem is a multi-objective problem. Because a large majority has multiple optimal venue categories we know that it is hard to find the best category when only looking at one objective.

5. CONCLUSIONS

For future work we want to take an in depth look into other city planning objectives. Currently, we are focused on the shop diversity as a city planning objective. City planning is more than just this feature, but is also often a city specific task. If we want to research this, we should create a closer co-operation with the cities we are analysing, increasing the effectiveness of our city score. Another addition is taking more different venue categories into account. We focused on a generalised set of categories. If we could gather more data from different venues we could look at a lower type in the Foursquare venue type hierarchy. One of the proposed goals could be to analyse different food type venues, such as but not limited to: fastfood, fine dining, coffee shop, etc.

In conclusion, using a Foursquare movements network we created a framework where the optimal store placement problem can be scored using the perspective of the store owner, and the perspective of the city as a system. We introduced a projection of movements on a relative neighbourhood graph, which we called FlowGraph, and used this graph to create two new features. The impact of these features were minimal and not significant for the performance of popularity prediction. Overall we showed that both objectives rank the categories in a different way, resulting in multiple categories considered optimal. This indicates that the multi-objective approach is a necessity in order to work towards a city improvement based on city planning guidelines while having satisfactory store owners.

6. REFERENCES

[1] M. Townsend, J. Surane, E. Orr, and C. Cannon. America's" retail apocalypse" is really just beginning. *URL: https://www.bloomberg .com/graphics/2017-retail-debt/*, 2017.

[2] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM, 2013.

[3] H. Damavandi, N. Abdolvand, and F. Karimipour. The computational techniques for optimal store placement: A review. In *International Conference on Computational Science and Its Applications*, pages 447–460. Springer, 2018.

[4] Z. Yu, M. Tian, Z. Wang, B. Guo, and T. Mei. Shop-type recommendation leveraging the data from social media and location-based services. *ACM Trans. Knowl. Discov. Data*, 11(1):1:1–1:21, July 2016.

[5] Y. Fu, G. Liu, S. Papadimitriou, H. Xiong, Y. Ge, H. Zhu, and C. Zhu. Real estate ranking via mixed land-use latent models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 299–308. ACM, 2015.

[6] A. Gil Solá and B. Vilhelmson. Negotiating proximity in sustainable urban planning: A swedish case. *Sustainability*, 11(1):31, 2019.

[7] R. Cervero. Mixed land-uses and commuting: Evidence from the american housing survey. *Transportation Research Part A: Policy and Practice*, 30(5):361–377, 1996.

[8] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, 74(3):035101, 2006.

[9] C.C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[10] J.W. Jaromczyk and G.T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992.

[11] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6638–6648, 2018.

Crime Rate Prediction with Region Risk and Movement Patterns

Shakila Khan Rumi
RMIT University
Melbourne, VIC
shakilakhan.rumi@rmit.edu.au

Phillip Luong
RMIT University
Melbourne, VIC
phillip.luong@rmit.edu.au

Flora D. Salim
RMIT University
Melbourne, VIC
flora.salim@rmit.edu.au

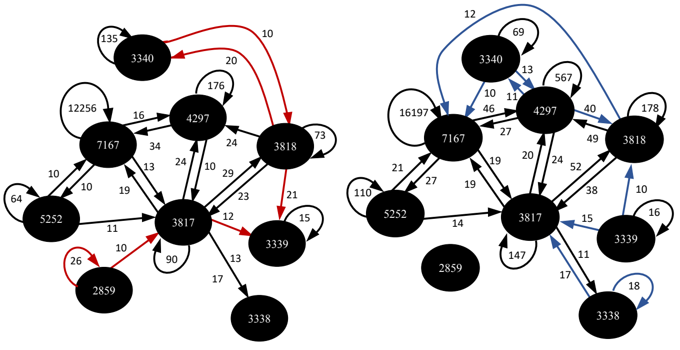
***Abstract*—Location Based Social Network, Foursquare helps us to understand the human movement of a city. It provides data that characterises the volume of movements across regions and Places of Interests (POIs) to explore the crime dynamics of the city. To fully exploit human movement into crime analysis, we propose region risk factor which combines monthly aggregated crime and human movement of a region across different time intervals. Number of features are derived from this risk factor. Extensive experiments with real world data in multiple cities verify the effectiveness of the features.**

I. Introduction

Safe and secure living place is one of the basic demands to every person in society. Therefore, it is important to find ways to control the crime rate. High crime rate hinders the economic development of a city. Understanding the root causes of what increase the likelihood of any particular crime event occurring at any time has great benefits for law enforcement to prevent crime events. According to the criminology theory, the surrounding environment including neighbourhood regions and movement of people play crucial role in crime event prediction. The widespread use of location based social networks such as *FourSquare* open the door of opportunities to analyse crime event occurrences in a timely manner. In this paper, we study the crime rate prediction with the help of urban mobility data.

Recently, there has been research exploring the link between crime events and urban dynamics using FourSquare data [1]. However, this study focused on a region's check-in information to predict the crime events. There was no focus on hyperlinking human movement between two regions. In [2] the authors considered the movement between regions using taxi flow data for crime inference, however, did not consider the variation of movement in different time periods of day. For example, the people who move to work place from home will move to the opposite direction in the afternoon. In this paper, we analyse the crime inference with the movement of people at different period of day. In Figure 1 we can observe the difference in the amount of people moving in morning and afternoon for New York City. The changes of human movements are highlighted with different colour. To fully exploit the human dynamics in crime inference problem, we further introduce region risk which associates crime and people movement in that region in a certain time interval, which is denoted as *Region Risk*. The hypothesis is that if the people are coming to a place from a high crime risk area, such movement implies a high crime risk in the arrival area too. We derive numerous features with assistance of this region risk. The significance of this features alongside other features is verified for three different cities including Chicago, Los Angeles and New York City across different time interval of a day. The contributions of this paper are summarized as follows:

- This is the first work that predicts crime rate based on the dynamic features that associate region risk and movement patterns between regions.
- Different features associated with region risk and the human mobility in different periods of time are crafted.
- The work verifies the effectiveness of different features in crime inference problem using correlation and regression analysis. Real-world crime data and FourSquare movement data are used for evaluation. The experimental results show that the region risk features are highly correlated with crime count of a region.



(a) Morning (b) Afternoon

Fig. 1: Few check-in movements in New York in different time intervals for ”2018-03”

II. Related Work

Many data mining research have been developed recently to verify the impact of human mobility into crime study. In [3], the authors extracted human behavior from mobile network activities and demographic features of people connected to the network over different regions and times. The study showed that the combination of mobile activity data and demographic data can predict crime event in a region with better accuracy.

Ambient population is measured through FourSquare check-in data and is used to understand the long-term crime event occurrences [4], [5]. In [1] the authors proposed several dynamic features using FourSquare data to measure the social diversity of a region and predict short-term crime event occurrences. To understand crime event occurrences, it is important to explore the correlation between places. The mobility data represented by taxi flows and Points of Interest (POI) can lift the performance of crime rate inference [2]. Here, the authors' hypothesis is that the social interaction between two places can be inferred through taxi trips and the crime rate propagate based on the connection between places. In [6], the authors proposed crime-specific dynamic features by analyzing individual risk factor of the users and extracted multiple features based on the risk analysis.

However, none of this work correlate the large scale human movement in different time period of day with crime in a region. Our work attempts to fill this gap.

III. Dataset Description

The datasets are collected for three different cities in USA including Chicago, Los Angeles and New York City. We collect different types of data including check-ins and crime events for each city. We segment each city into 400×400 grid.

A. POI and Check-in Data

The POI and check-in information is collected from FourSquare check-in. The dataset is provided as part of the *Future Cities Challenge* at Netmob [1]. The check-in information describes an aggregated count of all movements from one venue to another separated by month and five time intervals including Morning, Midday, Afternoon, Night and Overnight. In the collected data, we focus on the three cities mentioned above for the year of 2018. The aggregated number of venues and different venue movements are summarised in Table I.

TABLE I: Number of Venues and Venue Movements for each city

City	Venues	Unique Movements
Chicago	13,904	5,396,723
New York	32,971	5,296,809
Los Angeles	15,868	5,683,763

B. Crime Data

We collect crime event records that lie in 2018 for Chicago, Los Angeles and New York from Open Data Portal of the respective city councils [2,3,4]. Each dataset consists of the longitude, latitude, and the time and date of crime event occurrences. The total number of crime event occurrences are 263,515, 226,498 and 452,958 for Chicago, Los Angeles and New York respectively.

IV. Feature Description

This section presents the detailed description of the features. We consider each city as a directed graph, $G = (V, E)$. Each grid, $v \in V$, represents a node in the graph. E represents the set of edges between two nodes, which are weighted and directed. If there is a directed edge between two nodes, i and j (denoted as (i, j)), in a time interval, t, there is at least one check-in from i to j over t. The cost (or weight) of each node is the aggregated number of check-ins in a month. For each node $v \in V$ in a time interval, t, the following nodal and edge features are calculated.

A. Nodal Features

Nodal features describe the characteristics of the focal grid only.

1) Historical Features: To retain the historical knowledge about crime event occurrence, we calculate the following feature:

Crime Event History: We measure the number of crime events in node, v, during interval, t, in the past month. This is represented as:

$$NH(v,t) = \sum_{j \in m} Cr_j(v,t). \tag{1}$$

Here, $Cr_j(v,t)$ denotes the number of crime events that happened on j-th day in node, v, during time interval t. The variable, m, represents the day in the past month.

2) POI Features: The regional information of a node is described using the following features.

POI Density: For each node, $v \in V$ the POI density is calculated as follows:

$$NP(v) = \frac{N(v)}{N(V)}. \tag{2}$$

The total POI of the city is represented as $N(V)$. $N(v)$ denotes the number of POI in focal node, v.

Venue Category Distribution: Each type of venue has different impact on crime. Hence, it is important to extract the distribution of venue types in node, v. It is calculated as follows:

$$ND_i(v) = \frac{N_i(v)}{N(v)} \tag{3}$$

Here, $N_i(v)$ represents the number of i-th category venue in node, v.

Venue Diversity: Shannon's entropy [7] measurement is applied to determine the diversity of venue types, P, in node, v. Thus, Venue Diversity is modelled as:

$$NE(v) = -\sum_{i \in P} \frac{N_i(v)}{N(v)} * log_2\left(\frac{N_i(v)}{N(v)}\right). \tag{4}$$

3) Movement Features: The human dynamics of region in time interval is represented using the following features:

[1] https://www.futurecitieschallenge.com/

[2] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[3] https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

[4] https://data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr

TABLE II: Feature Correlation Analysis for the New York and Chicago

Feature Name	Chicago					New York				
	Morning	Midday	A'noon	Night	O'Night	Morning	Midday	A'noon	Night	O'Night
Check-in Entropy	**0.365**	**0.277**	**0.273**	**0.360**	0.326	**0.387**	**0.264**	**0.304**	**0.300**	0.250
Risk Count	**0.210**	0.088	0.077	0.095	**0.611**	**0.402**	**0.424**	**0.512**	0.282	**0.421**
Risk Ratio	0.113	-0.010	0.002	0.042	0.326	-0.018	-0.036	0.000	0.037	0.249
Risk	0.154	0.095	0.070	0.116	**0.422**	0.113	0.050	0.076	0.225	**0.355**
Number of Crimes	0.001	0.039	-0.026	0.035	0.322	0.357	0.074	0.053	**0.309**	0.040
Historical Density	**0.637**	**0.867**	**0.872**	**0.671**	**0.794**	**0.936**	**0.983**	**0.723**	**0.961**	**0.921**
Food	0.175	**0.116**	**0.119**	**0.228**	0.262	0.034	0.016	0.018	0.031	0.233
Total POI	**0.517**	**0.651**	**0.631**	**0.638**	**0.451**	**0.509**	**0.655**	**0.673**	**0.586**	**0.524**
Venue Diversity	**0.318**	**0.287**	**0.255**	**0.366**	**0.428**	**0.410**	**0.367**	**0.356**	**0.375**	**0.361**

Incoming Movement: The density of incoming movement into node, v, in time interval, t, is measured here. If a check-in is performed from other nodes to node, v, then it is considered as incoming movement. This we obtain the density of incoming movement as,

$$NI(v,t) = \frac{C_i(v,t)}{C(v,t)}, \tag{5}$$

where $C(v,t)$ and $C_i(v,t)$ denote total check-in and incoming check-ins respectively, performed at node, v, during time interval, t, in each month.

Outgoing Movement: If a check-in is performed from node, v, to other nodes then it is considered as outgoing movement. The density of outgoing movement is represented as,

$$NC(v,t) = \frac{|C_o(v,t)|}{|C(v,t)|}, \tag{6}$$

where $C_o(v,t)$ represents outgoing check-ins performed in node, v, during time interval, t, in each month.

Stationary Movement: When the origin and destination of a check-in is the same node, it is denoted as stationary movement. The density of stationary movement for $C_s(v,t)$ stationary check-ins in node, v, in time interval, t, is as follows:

$$NS(v,t) = \frac{|C_s(v,t)|}{|C(v,t)|} \tag{7}$$

Diversity of Movement: The heterogeneity of movement type is measured here:

$$NM(v,t) = -\sum_{i \in M} \frac{|C_i(v,t)|}{|C(v,t)|} * log_2\left(\frac{|C_i(v,t)|}{|C(v,t)|}\right). \tag{8}$$

The set, M, consists of three movement types: incoming, outgoing and stationary.

B. Edge Features

Edge features determine how crime rate of a region is influenced by its connected region. Two types of feature have been crafted. One is based on the crime rate in the neighbourhood region and another one is based on the risk analysis of a region based on movement data.

1) Neighbourhood Crime: The number of crime events for each adjacent node of focal node, v, is computed in each time interval. It reflects the situation of the surroundings.

2) Region Risk: To compute region risk, we analyse the risk associated with each node to support the following intuition. If the incoming check-ins of a node in a time interval are from high risk area, it imposes high risk of crime event occurrence in that node in that time interval. The region risk of node, v, for time interval, t, is as follows:

$$RR(v,t) = \frac{|Cr(v,t)|}{|C(v,t)|}, \tag{9}$$

where $Cr(v,t)$ denotes the crime events that happened in node, v ,in time interval, t. Several edge features are crafted which consider this region risk, $RR(v,t)$.

Risk Distribution: The risk distribution consists of the mean and median of the region risk associates with the regions $r \in R(v,t)$ from where people are moving to focal node, v, in time interval t.

Risk Count: Risk count in node, v in time interval, t, determines the number of regions with high risk than average from where incoming movement occur in node, v, in that time interval. The risk count is denoted by,

$$RC(v,t) = |\{r : r \in R(v,t) \text{ and } RR(r) > \frac{1}{|R|}\sum_{n \in V} RR(n,t)\}|. \tag{10}$$

Here, $R(v,t)$ denotes the regions which are origin of check ins to node v, and $|R|$ is the total number of regions.

Risk Ratio: Risk count determines the absolute number of regions with high risk. We normalise this feature based on total regions with incoming movement. The risk ratio is modelled using,

$$RT(v,t) = \frac{RC(v,t)}{|R(v,t)|}. \tag{11}$$

Region Risk: This feature represents absolute value of region risk that is associated to the focal node, v in time interval,t, $RR(v,t)$

C. Feature Correlation Analysis

We conduct Pearson correlation analysis to see how the proposed features are individually correlated with the monthly crime count of a region in a certain time interval of day. The correlation value between the features and crime count is illustrated in Table II for Chicago and New York City across different periods of a day. Here, we note only the features which have high correlation with crime count due to space limit. We observe that crime count is highly correlated with

Crime Event History in both cities in many intervals. According to the Near Repeat theory [8], crime tends to happen in the vicinity of past crime. The features derived from region risk analysis also have good correlation with crime count for both cities especially overnight. Such positive correlation proves the importance of such features in crime count. Surprisingly, many POI densities are negatively correlated with crime count. Although venue diversity has good correlation with crime which verifies that mixed land use has good impact on crime.

V. Model Description

The main purpose of this paper is to show the effectiveness of the features derived from human movement in crime inference. To serve this purpose, we apply Linear Regression (LR) to count the features of a node in a time interval. We adopt LR as inference model because it is a simple and most straightforward regression technique.

In this study, we only use the region and time interval where check-in movement exists to train the model. For example, in a month, m, for a region, v, during time interval, t, if there is any type of movement, it generates the training and test data based on the features described in previous section. Finally, LR model is trained with different feature settings to know the effectiveness of the feature. The performance metric compares which feature sets are significant for the crime inference.

VI. Experiment

A. Settings

The dataset used in this experiment is introduced in Section 3. Each day is segmented in five intervals and for each interval the are aggregated in monthly level. To prevent extreme sparsity situations, only check-in movement data with 10 or more unique movements in a month. The aim of the experiment is to examine the effectiveness of the proposed features in crime inference model for a month in a certain period of day. We partition the data about 75% as training set and the rest 25% as test set for all three cities. Particularly, the data lies between January 2018 and September 2018 (inclusive) is used as training data and the data between October 2018 to December 2018 (inclusive) as test data. If new regions are found with check-in movement greater than, or equal to 10 in test data for a time interval, the risk value for that region is set 0.

Two performance metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) have been applied to verify the effectiveness of the crime inference model.

B. Performance Study

We evaluate the performance of the proposed features for monthly crime count. We build linear regression based model and apply 'leave-one-out' to measure the performance of a set of features. First, we train a model with all the features. Next, we train another model without a set of features to identify the effectiveness of that set of features. If the MAE and RMSE increase for second model compare to the first one, that set of features is considered important in monthly crime count.

C. Feature Importance

We measure the importance of each group of feature using regression method. If the MAE and RMSE value is higher without a set of features, it verifies the importance of that feature set. The importance of each feature set is illustrated in Figure III for New York City. We observe that historical features are the dominating set of features among all of the features. The edge features based on Region Risk analysis also show the effectiveness across different time interval. The same analysis has been done for the other two cities.

TABLE III: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for different times of the day for New York

		Experiment Settings.[a]					
Time	Error	1	2	3	4	5	6
Morning	MAE	7.76	12.25	7.69	7.78	7.86	7.65
	RMSE	11.1	17.01	11.14	10.53	11.32	10.74
Midday	MAE	12.38	23.08	12.3	12.24	12.35	12.4
	RMSE	16.36	32.28	16.28	15.22	16.36	15.76
A'noon	MAE	21.29	23.45	21.3	21.88	21.29	22.02
	RMSE	42.55	32.11	42.61	44.11	42.54	44.46
Night	MAE	12.62	23.71	12.44	13.47	12.7	13.986
	RMSE	16.56	31.29	16.51	17.82	16.66	17.52
O'night	MAE	13.39	17.34	14.06	14.78	13.35	18.10
	RMSE	17.5	21.3	18.09	18.55	17.46	21.93

[a] List of experiments:
1. All Features Present, 2. Historical Features Omitted
3. Nodal Check-In Omitted, 4. Geographic (POI) Features Omitted,
5. Neighbourhood Features Omitted, 6. Region Risk Features Omitted

VII. Conclusion

This work provides new perspective to understand crime dynamics with the help of human mobility. It captures the relationship between the monthly aggregated crime data and the movement of people in a region across different time period of a day. The experiments verify that a group of people from high crime risk area increase the crime risk of their destination.

References

[1] S. K. Rumi, K. Deng, and F. Salim, "Crime Event Prediction with Dynamic Features," *EPJ Data Science*, 2018, in Press.

[2] H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime rate inference with big data," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 635–644.

[3] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in *Proceedings of the 16th international conference on multimodal interaction*. ACM, 2014, pp. 427–434.

[4] C. Kadar, J. Iria, and I. P. Cvijikj, "Exploring foursquare-derived features for crime prediction in new york city," in *The 5th International Workshop on Urban Computing (UrbComp 2016)*. ACM, 2016.

[5] C. Kadar and I. Pletikosa, "Mining large-scale human mobility data for long-term crime prediction," *EPJ Data Science*, vol. 7, no. 1, p. 26, 2018.

[6] S. K. Rumi, K. Deng, and F. Salim, "Theft prediction with individual risk factor of visitors," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2018.

[7] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[8] J. Q. Wilson and G. L. Kelling, "Broken windows," *Atlantic monthly*, vol. 249, no. 3, pp. 29–38, 1982.