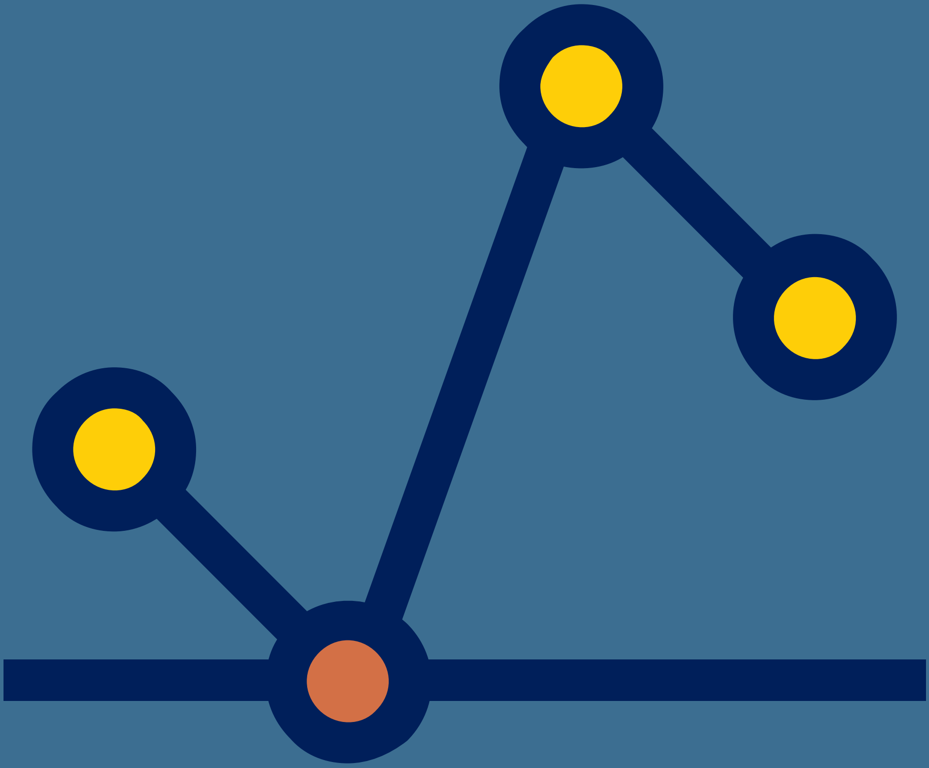
Edited by

Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



Pavia University Press

Copertina: Cristina Bernasconi, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Date le caratteristiche di Internet, l'Editore non è responsabile per eventuali variazioni di indirizzi e contenuti dei siti Internet menzionati.

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

Spatio-temporal statistical analyses for risk evaluation using big data from mobile phone network

Analisi statistiche spazio-temporali per la valutazione del rischio utilizzando big data dalla rete di telefonia mobile

Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

Abstract In this short paper we summarize the ongoing work pertaining the use and combination of mobile phone data to characterize the spatio-temporal dynamic of the presence and the movements of people in the context of smart cities. We develop ad-hoc statistical approaches with the aim of developing small area indicators and forecasting traffic flows. The application of these strategies is related to the evaluation of flood risk in urban areas.

Key words: Vector autoregressive model; model-based functional cluster analysis; T-mode principal component analysis.

1 Introduction

Mobile phone data allow for a dynamic and fine-grained representation of human activities in urban systems. They are particularly suitable for the analysis and the mapping of the density of people's presences [1, 11] and movements [16, 12]. For this reason, they are increasingly adopted for the analysis of smart cities [2]. Some recent statistical applications can be found in [5, 4, 6, 9, 10].

Different sources of mobile phone data exist and might be used for different purposes. Here, we overview our ongoing work related to the development of statistical modeling and classification methods aimed at characterizing the spatiotemporal dynamic of people's presences and movements. The presented works concern the region of the Mandolossa (in the northwest outskirt of Brescia, Italy), which is an

DMS Statlab, Department of Economics and Management
University of Brescia, Contrada Santa Chiara, 50, Brescia. e-mail: selene.perazzini@unibs.it

Department of Economics
University of Bergamo, Via Caniana, 2, Bergamo. e-mail: rodolfo.metulini@unibg.it

DMS Statlab, Department of Economics and Management
University of Brescia, Contrada Santa Chiara, 50, Brescia. e-mail: maurizio.carpita@unibs.it

interesting case study because, being exposed to the risk of flooding, local authorities need several accurate information about people in the area. We restrict our attention to applications connected to three United Nations Sustainable Development Goals: 9 - Industry, innovation, and infrastructure; 11 - Sustainable cities and communities; 13 - Climate action. In particular, we focus on human exposure to flood risk, which is constituted by both the individuals staying in and those passing by the area. An accurate representation of human exposure should therefore account for the phenomena' static and dynamic characteristics. We show that the two aspects can be captured using different sources of mobile phone data. Their adoption allows for a multifaceted representation of human exposure at the small-area level, therefore providing a level of accuracy so high that it is rarely achieved by other types of data. We dedicate particular attention to the monitoring of traffic networks, which is a goal of Mission 3 of the Italian National Recovery and Resilience Plan (part of the Next Generation EU Programme). Furthermore, we show some examples of how different mobile phone data sources can be combined with each other's or with other types of data for the analysis of traffic.

2 Mobile phone data

We consider three types of mobile phone data:

- Crowding data (MPD), representing the Mobile Phone Density - i.e., the average number of mobile phone SIM cards in a squared cell of a pixel grid of dimension 150×150 meters in a 15-minute interval;
- Flows data (OD), representing the Origin-Destination flows - i.e., the number of SIM cards moving from one of the "Aree di CEnsimento" (ACE) of the Province of Brescia shown in the left map of Figure 1 to another in a 1-hour interval;
- Signals data (MDT), representing the data collected using the "Minimization of Drive Tests" technology that registers radio measurements of signals (i.e., phone calls, text messages, internet browsing, or technical operations on the network) transmitted over the 3G/4G mobile network from/to terminal devices with GPS enabled. The signals are collected in 15-minute intervals and geo-referenced on a grid of pixels measuring 10 meters per side. Though the MDT data only represents a sample of users, the accurate geolocation allows for small-area estimation. Since a mobile phone can produce multiple signals in 15 minutes, we refer to the number of grid cells from which MDT signals originated in an area.

The three types of data concern users subscribed to TIM and cover different periods and partially different (but overlapping) areas (see the left map of Figure 1). The MPD data was provided by the Municipality of Brescia in the context of a territorial monitoring project between 2014 and 2016. For this reason, they collect phone signals from April 1st, 2014, to August 11th, 2016. The OD and MDT data have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it) for the MoSoRe Project 2020-2022. The OD data cover the period between September 2020 and August 2021, while the MDT data refer to 5 days of November 2021 (namely Wednesday 10^{th}, Friday 19^{th}, Saturday 20^{th}, Sunday 21^{st}, and Monday 22^{nd}). Indeed, MDT data require particular technologies to be activated and tested before data collection. For this reason, the data collection process is costly and takes time, and the produced MDT datasets typically cover short periods and small areas (see the right map of Figure 1). To overcome this issue, days for MDT data collection have been chosen in such a way as to represent a typical week.

It is worth noticing that, while the MPD and the OD data have already found applications in statistics, the MDT data have been very recently released and have been almost exclusively adopted in network engineering.

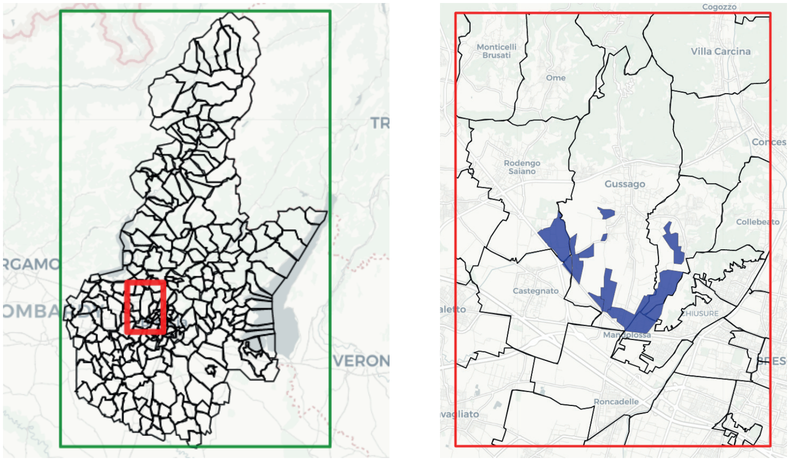


Fig. 1 Left: map of the ACEs of the Province of Brescia (black), and the areas captured by the MPD data (green) and by the MDT (red) datasets. Right: map of the ACEs (black) in the MDT dataset (red) and of the flood risk map with time to return equal to 20 years (blue).

3 An overview of statistical analyses

As a first step, we show that mobile phone data can be used to produce small-area indicators. In this regard, in [13] we use the MPD and the MDT data to represent respectively crowding and traffic intensity in the “Sezioni di CEnsimento” (SCE) (which are subdivisions of the ACEs) within the area captured by the MDT database. To this scope, since the MPD data are defined over a grid of quite large pixels, the number of users in each cell is distributed among the SCEs proportionally to the fraction of the area of the cell overlapping each of them:

$$MPD_{jt} = \sum_k MPD_{kt} \cdot \frac{Area(SCE_j \cap Cell_k)}{Area(Cell_k)} \tag{1}$$

where j indicates a SCE and t a time interval. As far as the traffic intensity is concerned, the MDT database is compared to a street map and is restricted to the phone signals originating from streets. Then, we count the number of cells of the MDT pixel grid corresponding to streets for each SCE j and each time interval t, and the obtained values are divided by the area of the street network in the j-th SCE:

$$MDT_{jt} = \frac{Number(\text{Streets Cells}_{jt})}{Area(\text{Streets}_j)}. \tag{2}$$

The two resulting sets of data (eq. 1 and 2) are analyzed and spatial patterns are investigated. At last, two indicators are defined using two T-mode principal component analyses [3] (see the left map of Figure 2 for an example). Given the strategic role of the road network in flood emergency management, three indicators capturing the main characteristics of the streets in the SCEs are also defined based on the available street maps. The joint analysis of the two mobile-phone-based and the three street-map-based indicators might be used to identify the areas with high concentrations of people or major connecting routes.

Mobile phone data can also be used for estimation and forecasting. In this respect, in [12] we use the OD data to analyze the traffic flows linked to the ACEs overlapping the flood risk map (see the right map of Figure 1). In that paper, a vector autoregressive model with dynamic harmonic components capturing complex seasonality [7] has been defined to forecast traffic flows:

$$\boldsymbol{Flow}_t = \boldsymbol{\nu} + \sum_{h=1}^{p} \boldsymbol{A}_h \boldsymbol{Flow}_{t-24\times h} + \boldsymbol{Bx}_t + \boldsymbol{\varepsilon}_t \tag{3}$$

where $\boldsymbol{Flow}_t$ is a vector of length 3 containing the flows to, from, and within the ACEs exposed to floods. $\boldsymbol{\nu}$ is a constant vector of length 3, p is the autoregressive parameter, $\boldsymbol{A}_h$ is a 3×3 matrix of coefficients to be estimated, $\boldsymbol{\varepsilon}_t$ is the 3×1 vector of the error terms at time t, $\boldsymbol{x}_t$ is the vector of the l exogenous variables at time t, and $\boldsymbol{B}$ is the $3 \times l$ matrix of coefficients of the exogenous variables. The vector $\boldsymbol{Bx}_t$ is modeled using proper dynamic harmonic regression components. Despite the partial autocorrelation function of the estimated residuals showing significant first-order autocorrelation and a leptokurtic distribution with heavy tails, by means of a k-folds cross-validation we find that the model achieves satisfactory performance in forecasting both the number of people moving (see figure 2, that shows true versus fitted values for randomly chosen validation days) and the level of traffic intensity.

Then, in [14] we show that the estimation can be improved by combining the OD data with the MDT. We use the MDT data to estimate the proportion of traffic flows related to the flood-prone area in an ACE. The resulting ratios are then applied to the OD data as weights, in such a way as to obtain the traffic flows at risk. The combination of these two pieces of information allows us to produce statistical estimation and forecast of traffic flows for short time intervals at the small area level.

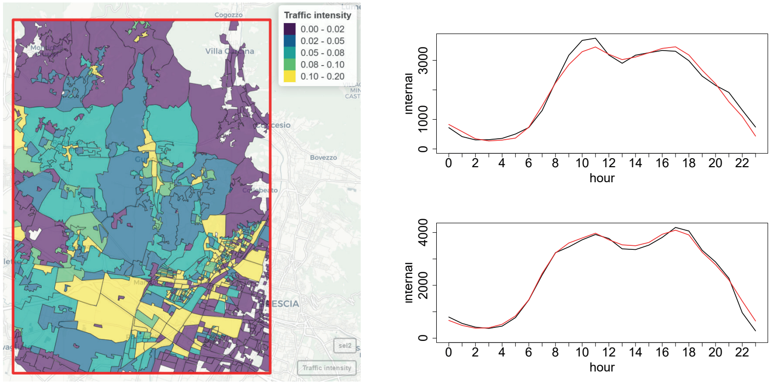


Fig. 2 Results of the analyses. Left: map of the traffic intensity indicator. Right: Observed (black) versus forecasted (colored) internal traffic flows. Validation days: Saturday February, 13^{th}, 2021 (top), Tuesday July, 13^{th}, 2021 (bottom).

This information can be used by local authorities to promptly activate traffic control actions aimed at preventing human losses and injuries.

4 Discussion and future developments

In this paper, three modern sources of mobile phone data have been presented, and their potential in the analysis of people's presence and movements has been shown through the discussion of some applications. We showed that the combination of different mobile phone databases can improve the accuracy of estimation. This aspect is still part of our current research. For example, we are working on a VARX model for forecasting traffic flows in risky areas, including traffic intensity and crowding indicators as regressors. Moreover, we presented some examples of how mobile phone data can be combined with other data sources to provide a multifaceted representation of a complex phenomenon. Nowadays, we are extending our research by introducing further data sources in the analysis and exploring the dynamics that link them. For example, in the context of the project "Data Science for Brescia", we are jointly analyzing mobile phone data and expenditure indices defined on Mastercard payment data to monitor the social and economic impacts of cultural events in the city of Brescia.

Acknowledgements This contribution has been developed for the Spoke 7 "CCAM, Connected Networks and Smart Infrastructure" of the Sustainable Mobility Research Center funded by the Italian NRRP (National Recovery and Resilience Plan) 2021-2026, which is part of the Next Generation EU (NGEU) Programme.

References

1. Balistrocchi, M., Metulini, R., Carpita, M., & Ranzi, R. Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, **20(12)**, 3485–3500 (2020)
2. Bibri, S.E., and Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable cities and society. **31**, 183–212 (2017)
3. Compagnucci, R. H., and Richman, M. B.: Can principal component analysis provide atmospheric circulation or teleconnection patterns?. International Journal of Climatology: A Journal of the Royal Meteorological Society, **28(6)**, 703-726 (2008)
4. Carpita, M., Manisera, M., and Zuccolotto, P.: Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia. In: Lombardo R., Camminatiello I., Simonacci V. eds. IES 2022: Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, Book of Short Papers of the 10th Scientific Conference of the SVQS, 575-581. PKE Press, Milano (2022)
5. Carpita, M., and Simonetto, A.: Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation. **5 (1)**, 31–41 (2014)
6. Curci, F., Kërçuku, A., Zanfi, F., Novak, C. et al.: Permanent and seasonal human presence in the coastal settlements of Lecce. An analysis using mobile phone tracking data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 57–71 (2022)
7. Hyndman, R.J., Athanasopoulos, G.. Forecasting: principles and practice. OTexts (2018)
8. Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis **71**, 92—106 (2014)
9. Manfredini, F., Lanza, G., Curci, F., et al.: Mobile phone traffic data for territorial research. Opportunities and challenges for urban sensing and territorial fragilities analysis. TeMA-Journal of Land Use, Mobility and Environment. **2**, 9–23 (2022)
10. Mariotti, I., Giavarini, V., Rossi, F., and Akhavan, M.: Exploring the "15-Minute City" and near working in Milan using mobile phone data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 39–56 (2022)
11. Metulini, R., and Carpita, M.: A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Social Indicators Research. **156 (2)**, 761–781 (2021)
12. Metulini, R., and Carpita, M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research. to be published. (2023)
13. Perazzini, S., Metulini, R., Carpita, M. Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Submitted to journal.
14. Perazzini, S., Metulini, R., Carpita, M. Integration of flows and signals data from mobile phone network for statistical analyses of traffic in a flooding risk area. Submitted to journal.
15. Pucci, P., Gargiulo, C., Manfredini, F., Carpentieri, G., et al.: Mobile phone data for exploring spatio-temporal transformations in contemporary territories. TeMA-Journal of Land Use, Mobility and Environment. **2**, 6–12 (2022)
16. Tettamanti, T., and Varga, I.: Mobile phone location area based traffic flow estimation in urban road traffic. Advances in Civil and Environmental Engineering. **1 (1)**, 1–15 (2014)