Evidences from Cellphone Big Data Applications for Urban Mobility Needs: Case Studies in Italy

Published Version:
Nalin, A., Lantieri, C., Simone, A., Vignali, V., Cappellari, D., Mantegari, G., et al. (2024). Evidences from Cellphone Big Data Applications for Urban Mobility Needs: Case Studies in Italy. Cham : Springer [10.1007/978-3-031-62478-0_24].

Availability:
This version is available at: https://hdl.handle.net/11585/986438 since: 2024-09-22

Published:
DOI: http://doi.org/10.1007/978-3-031-62478-0_24

Evidences from cell phone big data applications for urban mobility needs: case studies in Italy

Abstract. The use of big data has been radically increasing in the last years. In fact, when users' privacy is guaranteed, big data can provide an effective, even customized, answer to the increasing users' needs. A peculiar typology of big data are mobile phone records, which are often recorded automatically by telecommunication operators for billing, management, and network maintenance purposes. Nowadays, they are considered a pertinent tool for understanding how people use city's infrastructures from the point of view of mobility, consumption and environmental impacts, enabling a better comprehension of urban phenomena if compared to the possibilities offered by traditional methods such as travel surveys. This work will focus on some applications, definitions and metrics of cell phone big data by showing a few case studies in Italy. The results of this work will demonstrate some possible solutions in overcoming the traditional drawbacks and limitations of mobile phone big data. Moreover, thanks to the pervasive and ubiquitous presence of connected devices in everyday life, they are a priceless source of information for planning purposes and to enhance smart city and transportation.

Keywords: Big data, Cell phone data, Smart city, Smart transportation.

1 Introduction

The role of big data in everyday life is a challenging matter able to influence urban policies and agendas, and it is strictly connected to the broader theme of 'smart cities' [1]. When dealing with the term 'big data', even if there isn't convergence on a univocal definition [2], most Authors mention the 'five-V-paradigm' (i.e., Value, Variety, Velocity, Veracity, Volume; [3]), with the seldom addition of 'Complexity'. As argued by [3], the key strength of this paradigm lies on the ability to uncover patterns and glean knowledge from data with appropriate and innovative techniques. In parallel to the growth of the size of datasets, which is tightly linked to the increase of storage capacity, computation speed [2], and the number of encompassed fields of application and domains [4], the concept of geospatial big data appeared in debates [5],[6]. This latter notion is referred to all those datasets with specific spatial attributes, such as coordinates. In this context, a peculiar form of big data is the cell phone (often called 'telco') big data. They are collected by telecommunication providers for billing purposes [7], in an uninterrupted data stream that enables companies to handle billions of records on a daily basis. Transportation appears as one of the most appropriate domain for big data application [8], [9], [10], mainly thanks to the pervasive presence of connected devices [11] and the high spatiotemporal resolution of data, which positively affects the accuracy [12], especially when telco big data are coupled to GPS records [13]. One of the more evident strength of telco big data is the sample size, which can account for a considerable share of population [14]. Alongside representativeness, another key strength of big data is the speed in elaborations. While the progression of surveys (i.e., from data collection to data report) with traditional techniques require months or even years, the processing of big data can shorten the needed time, and hence dramatically reduce the overall costs [15]. When dealing with big data, some drawbacks have to be kept in mind aside from the abovementioned strengths. Among them, privacy concerns are probably the most impacting [16], and hence should be always taken into account. Another potential drawback is related to the validation of telco big data with groundtruth data [17]. This paper aims to overcome the mistrust in telco big data as an effective sample of geospatial datasets and as a complementary data source, by showing some novel evidences coming from different applications of telco big datasets, which will demonstrate the capability of unveil mobility patterns even in (near) real time. After a brief description of definitions, metrics and processing methodologies for the different datasets highlighted in Section 2, some case studies in Italy will be presented and discussed in Section 3: 'Padua commuting area' analysis, held in 2018 (Section 3.1); 'Milan city users' analysis, held in 2022 (Section 3.2); 'Ukrainian migrants flows inbounding Italy' after Russian invasion, held in 2022 (Section 3.3).

2 Materials and methods

The analyses described in Section 3 were based on the Vodafone Italia network, whose coverage ≈ 99% of Italian territory and population [13]. According to AGCOM [18], Vodafone's market penetration in Italy is equal to one third of the resident population (i.e., approximately 20 million human SIMs). In order to extract information from cellphone datasets, raw data were preliminarily processed as follows: recording procedures; anonymization and aggregation; calibration and inferential analyses. After processing procedures, data were analysed with regards to pertinent metrics and analytics. They should be properly set per each project and should be able to detect users' behaviours with regards to specific profiles and mobility patterns.

Analyses are based on high-resolution data, considering both space and time, that are provided by a dense network of telecommunication devices, namely cellphone towers, mobile radio cells and antennas. Raw data, i.e., the location and the time stamp per each user, are usually collected as Circuit-Switched (CS) or Packet-Switched (PS) networks. While the first typically generates Call Detail Records (CDR) which are based on the users' activity (e.g., voice calls, SMS, etc.), the latter is roughly automatic and could account up to 1,000 records per day per SIM [13]. The coverage of each tower depends on both technical reasons and the location where the device is installed. In fact, the coverage area ranges from a few hundred meters in urban environments to kilometres in rural areas [13]. Coverage overlays should be always taken into account. While they are needed in order to keep the cellphone network operative, they could make the location detection more difficult, with a higher analytical effort. Before outlining the analyses, it must be specified that the raw data are irreversibly anonymized and aggregated. In fact, due to normative limitations, analyses should be based on aggregation of people with pre-set grouping thresholds. This means that, during data processing, distinct users or small groups of them cannot be monitored individually. In order to avoid information loss, these data are kept and processed as homogeneous group labelled as ‘masked’ users. With regards to the Italian context and given the reference laws (i.e., [19] and the national laws), the privacy threshold is set in 15 users.

Using appropriate algorithms, data were finally inferred to the universe and not only to Vodafone clients [13]. With regards to users’ provenance, Italians and foreigners were classified in relation to different information. Concerning Italians, a roll-back analysis on the previous 12 months before the analysed event is performed to determine the prevalent night cellphone tower (i.e., where users spent at least 6 hours from 8 PM to 8 AM). With regards to foreign users, the nationality is assigned on the basis of the Mobile Country Code (MCC), a code contained in the SIM. According to Vodafone policies, gender and age are provided only for Italian users. With regards to the analysed case studies, the following **Table 1** provides an overview of the different metrics applied in the analysed case studies.

Table 1. Metrics and users’ behaviours.

Metrics	Definitions	Case study
Unique presences	Analysis of the users’ presence in the study areas. To be detected, users must spend a minimum amount of time in the study area and in the reference time unit. Users are counted only once even if they make multiple visits.	Milan city users
Origin and destination	O/D matrix provided for the entire study area, with different spatial reference units.	Padua commuting area
Trip frequency	Analysis of trip frequency for the entire study area.	Padua commuting area
Users’ provenance	Analysis of the origin of unique users detected at least once in the study area, in terms of usual residence for Italians and of nationality for foreigners.	Padua commuting area
Inbound and outbound	Analysis of first and last Italian province in which a user is registered, with specific reference to Alpine entry and exit points.	Ukrainian migrants
Incidence index	Ratio of foreign users compared to the population officially registered within the study area.	Ukrainian migrants
Prevalent area	Analysis of the area in which users are detected spending their highest share of time. Users are counted in a single area, thus avoiding double counting.	Ukrainian migrants

3 Case study description and results

In the next paragraphs three case studies about applications of telco big data will be presented. They range from the traditional mobility concepts of origin/destination matrices, to innovative metrics, such as inbound/outbound flows, which are detected as main suppliers of other data sources.

3.1 ‘Padua commuting area’ analysis

During the SUMP update phase in Padua, Italy, a comprehensive analysis of urban and metropolitan commuting area was carried out. The analysis was implemented from February 2018 to May 2018 and included 20 municipalities (i.e., Padua, which was split into 26 Traffic Analysis Zones – TAZs, and surrounding towns). Data were processed in order to detect mobility patterns of residents and city users such as visitors, students, commuters, tourists, etc. Temporal analyses were carried out detailing the day of the week and the different time slots, and including both daily and nocturnal movements. Among others, O/D matrices, trip frequency and users’ provenience were the most scrutinised metrics (see **Table 1**). It is worth noting that telco data outputs offered results that are comparable to official statistics and other surveys conducted at the same time.

As shown by **Fig. 1**, where flows were reported in accordance with their magnitude, most of the trips were done within Padua. Part of the O/D matrix is blank because flows between other municipalities were not monitored, or flows were below the privacy threshold. As a general reference, coherence should be found by comparing flows inbounding city centre and industrial areas, where most of the workplaces lie, and outbounding to the peripherical neighbourhoods, where most of the population live.

Fig. 1. Origin/Destination matrix.

The trip frequency distribution analysis was set in order to detect the main commuting patterns. When frequency $f_{trips} \geq 15$ in a month for a given O/D, users were labelled as ‘commuter’. **Fig. 2** shows the flows’ origins and the destinations filtered by the abovementioned frequency threshold, so only ‘commuting’ O/Ds are detected. As reported in the previous matrix, each cell is coloured in accordance to the magnitude of the monitored variable, i.e., trip frequency f.

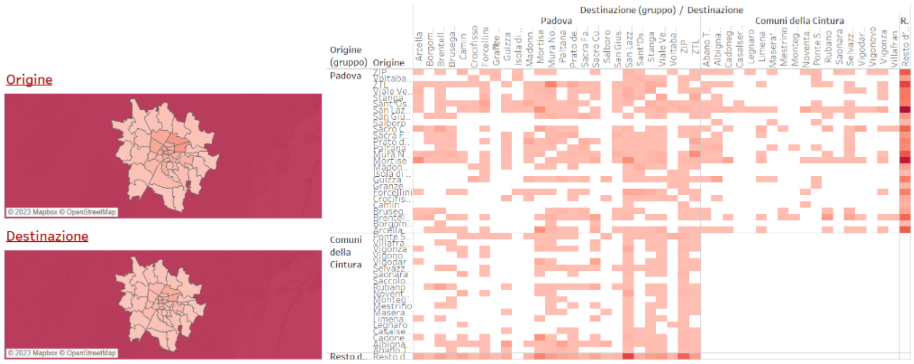

Fig. 2. Trip frequency *f* (filtered visualization, with $f_{trips} \geq 15$).

After data analysis, a machine learning clustering procedure based on the described matrices was set in order to detect homogeneous mobility-based TAZs within Padua municipality. This procedure is based on the grouping of nodes formed by antenna locations that are more densely connected to each other than to the rest of the network. Each 'virtual' node represents travel O/D and are grouped into 'communities' based on the intensity of the number of flows between the two nodes. Clusters hence represent areas that are more interconnected by flows. **Fig. 3** shows the outputs of the algorithms compared to original TAZs. Differences between 'B' and 'C' clusters lie on grouping thresholds.

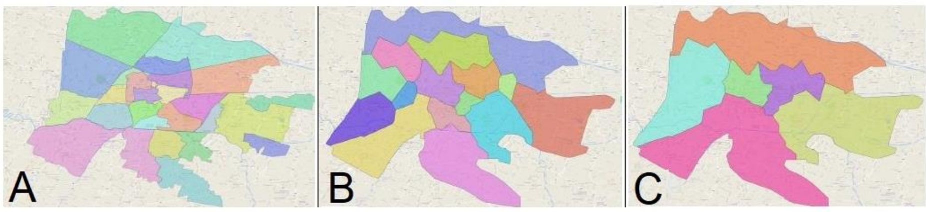

Fig. 3. Machine learning clustering. A: TAZs; B: 15-zones clustering; C: 6-zones clustering.

3.2 'Milan city users' analysis

The term 'city users' encompasses different traveller categories, such as leisure tourists, business workers, off-site students, etc. As the second-largest Italian city, and the main economic centre and major transport hub, Milan represents a significant context and a very interesting case study for understanding how to better manage the impacts on urban dynamics of 'vibrancy', crowding and presence of 'non-residents' by using telco big data [20], without any other effective reference datasets. The analyses included 15 months (January 2022 to March 2023), when mobility restrictions against the COVID-19 pandemic were almost ended. As a spatial reference unit, ACEs ('Area di Censimento'; Household survey areas; [21]) provided by ISTAT, the Italian Institute of Statistics, were chosen instead of traditional and smaller survey sections, because they were seen as a balanced compromise between information granularity and privacy concerns (i.e., the smaller the areas, the higher the probability to detect user groups below the privacy aggregation thresholds). City users were labelled as: ‘residents’; ‘inner commuters’, i.e., who lives in a given ACE but works in another ACE; ‘commuters’, with regards to their commuting frequency and their mobility patterns; ‘visitors’, i.e., who was detected occasionally in Milan. This latter category represents the highest share of users depicted with orange pattern in **Fig. 4**. With regards to the spatial distribution of users, as an evident and expectable mark the central ACEs (i.e., city and historical centre; the CBD close to Garibaldi train station; the Centrale train station; the Navigli) are the most crowded areas.

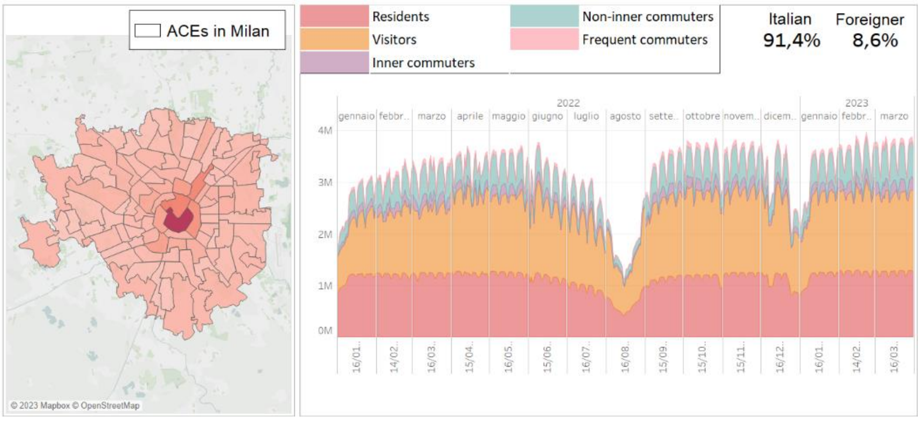

Fig. 4. Unique presences.

3.3 ‘Ukrainian migrants flows inbounding Italy’ analysis

After the Russian invasion of Ukraine, which dates back to February the 24th, 2022, European countries were involved in rescuing and civil defence activities. The dramatic turn of events caused unprecedented flows of migrants evacuating the belligerent homeland towards other nations. As a major immigration country, even for Ukrainians, Italy was supposed to be a main destination of displaced people. In order to monitor the flows, a comprehensive and detailed roll-back analysis was carried out from November 2021 to August 2022. The invasion day was considered as watershed: before it, flows were considered ‘normal’ movements (i.e., workers, people returning to homeland, etc.), while after it flows were labelled as ‘emergencial’. As for spatial references, administrative units, hospital districts and border crossings, both rail and road, were used. No information about maritime and air transport were provided. Among others, the main metrics were inbound and outbound flows; incidence index; and prevalent area (see **Table 1**). This analysis represented a novel effort to quantify the migration phenomenon from a comprehensive point of view. In fact, due to the extreme heterogeneity of war migrants aims and patterns (e.g., people could range from permanent family reunions to temporary evacuates), official statistics or surveys couldn’t be able to detect precisely all the flows’ characteristics, their final destinations and the presence of migrants within urban areas, during such an emergency phase, with the aim of both rescuing and providing assistance. Considerable insights can be seen from **Fig. 5**, which compares inbound (top) to outbound flows (bottom), **Fig. 6** describing the prevalent area of detection and **Fig. 7**, representing the incidence index. While a considerable amount of migrants remained in Italy for months (the incidence index ranged from 0.00069 during the first day of survey to 0.00258 at last day of survey) the highest peaks (≈ 0.004) included who decided to stay in Italy as well as people just transiting. With regards to the former, the provinces of Rome, Milan and Naples counted the highest shares of migrants (**Fig. 6**). With regards to the latter, and detailing border crossings (**Fig. 5**), both inbound and outbound flows were detected mostly in the eastern side of Norther Italy (i.e., Gorizia, Trieste and Udine border crossings). A considerable share of outbounding people was detected at the Imperia border crossing only a few days after the invasion, and they could be interpreted as transiting flows towards France. Although no information about gender and age are provided and confirmations are quite hard to be proved, outbounding flows registered at the eastern border crossings (i.e., towards eastern countries and Ukraine itself) could be interpreted as people eligible for military enlistment backing to homeland.

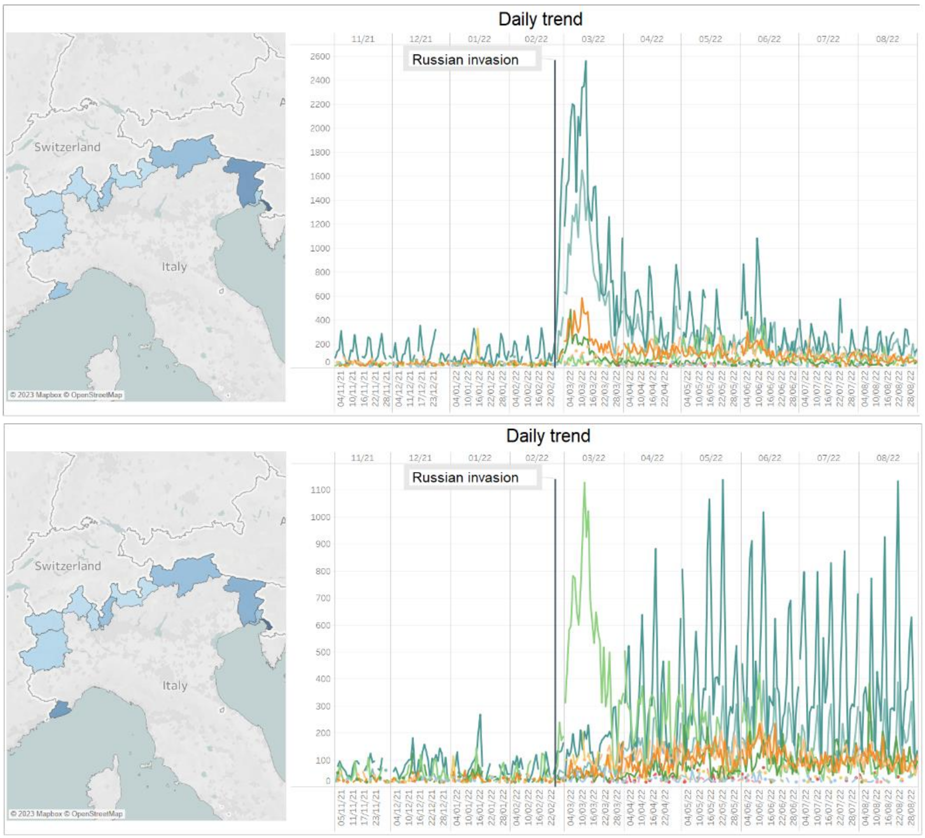

Fig. 5. Inbound (top) and outbound (bottom) flows.

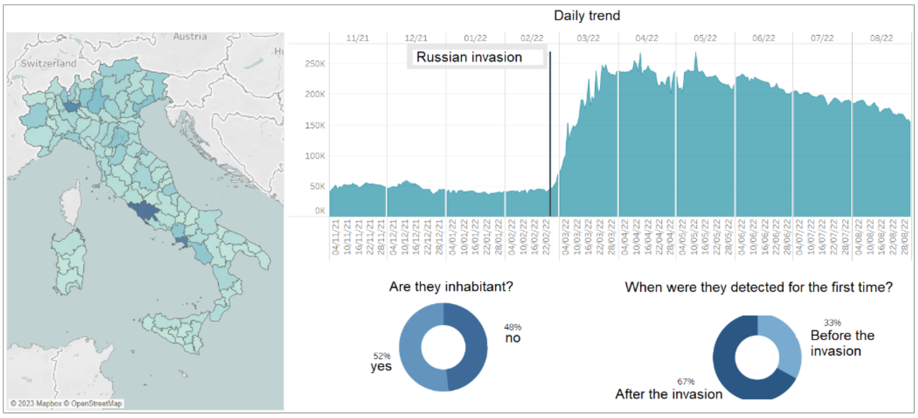

Fig. 6. Prevalent area of Ukrainians.

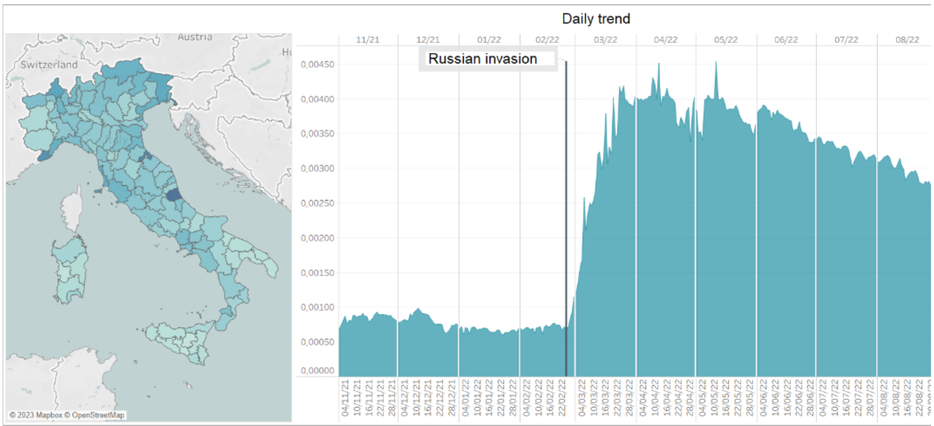

Fig. 7. Incidence index of Ukrainians.

4 Conclusions

The massive spread of connected devices has been played a key role in the enhancement of ‘smartness’ in urban areas. The most evident component of such this ‘connected’ environment is the smartphone. When turned on, are they calling, sending SMSs, online browsing or not, they are automatically connected to the telephone network, in a bidirectional and seamless stream of information. At early stages of cellphone usage, information was collected by telecommunication companies only for billing and technical purposes, while today data are processed as valuable (big) datasets for some cases, especially for transport-related matters, or to monitor urban vibrancy, or to assist people rescuing during emergencies. As showed by the discussed case studies, where different telco big data applications, metrics and contexts have been described, this technology enables to quantify the ‘effective’ status quo of people behaviours (i.e., how people move, how and where they spend their time, where they live, etc.) even in (near) real time, so the elapsing time between events and analysis could be dramatically reduced. Additionally, such this technology is based on the real movements, and not on the official statistics. As general remark of this paper, and recommendation for further works, this latter concept could be considered at the same time a major strength and weakness of telco big data. In fact, they could be considered as a main reference for some fields (such as presence of people in a given area or their spatial patterns) while for others (e.g., comparison between people flows and traditional transportation surveys outputs) a solid validation process should be properly conduced.

References

[1] R. Mortaheb and P. Jankowski, ‘Smart city re-imagined: City planning and GeoAI in the age of big data’, *Journal of Urban Management*, vol. 12, no. 1, pp. 4–15, 2023.

[2] C. M. Dalton and J. Thatcher, ‘Inflated granularity: Spatial “Big Data” and geodemographics’, *Big Data and Society*, vol. 2, no. 2, Dec. 2015, doi: 10.1177/2053951715601144.

[3] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, ‘Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 2, p. e1289, 2019.

[4] S. E. Bibri, ‘The anatomy of the data-driven smart sustainable city: instrumentation, datafication, computerization and related applications’, *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0221-4.

[5] G. Yin *et al.*, ‘How to quantify the travel ratio of urban public transport at a high spatial resolution? A novel computational framework with geospatial big data’, *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103245, 2023.

[6] Z. Huang *et al.*, ‘Modeling real-time human mobility based on mobile phone and transportation data fusion’, *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 251–269, Nov. 2018, doi: 10.1016/j.trc.2018.09.016.

[7] H. Huang, X. A. Yao, J. M. Krisp, and B. Jiang, ‘Analytics of location-based big data for smart cities : opportunities, challenges, and future directions’, vol. 90, 2021.

[8] A. I. Torre-Bastida, J. D. Ser, I. Laña, M. Ilardia, M. N. Bilbao, and S. Campos-Cordobés, ‘Big Data for transportation and mobility: recent advances, trends and challenges’, *IET Intelligent Transport Systems*, vol. 12, no. 8, pp. 742–755, 2018.

[9] M. D. Gennaro, E. Paffumi, and G. Martini, ‘Big data for supporting low-carbon road transport policies in Europe: Applications, challenges and opportunities’, *Big data research*, vol. 6, pp. 11–25, 2016.

[10] R. Battistini, A. Nalin, A. Simone, C. Lantieri, and V. Vignali, ‘How do University Student Cyclists Ride? The Case of University of Bologna’, *Applied Sciences*, vol. 12, no. 22, p. 11569, 2022.

[11] J. Steenbruggen, E. Tranos, and P. Nijkamp, ‘Data from mobile phone operators: A tool for smarter cities?’, *Telecommunications Policy*, vol. 39, no. 3–4, pp. 335–346, May 2015, doi: 10.1016/j.telpol.2014.04.001.

[12] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, ‘Towards real-time road traffic analytics using telco big data’, in *Proceedings of the international workshop on real-time business intelligence and analytics*, 2017, pp. 1–5.

[13] C. Lorenzo *et al.*, ‘Exploring mobile network data for tourism statistics: the collaboration between Istat and Vodafone Business Italia’, *RIVISTA DI STATISTICA UFFICIALE*, vol. 3, pp. 43–76, 2022.

[14] K. Vidović, P. Čolić, S. Vojvodić, and A. Blavicki, 'Methodology for public transport mode detection using telecom big data sets: case study in Croatia', *Transportation Research Procedia*, vol. 64, pp. 76–83, 2022.

[15] N. Caceres, L. M. Romero, and F. G. Benitez, 'Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys', *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 574–601, Jan. 2020, doi: 10.1080/23249935.2020.1720857.

[16] A. Ogulenko, I. Benenson, I. Omer, and B. Alon, 'Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data', *Computers, Environment and Urban Systems*, vol. 85, Jan. 2021, doi: 10.1016/j.compenvurbsys.2020.101550.

[17] F. Calabrese, L. Ferrari, and V. Blondel, 'Urban Sensing Using Mobile Phone Network Data: A Survey of Research', *ACM computing surveys*, vol. 47, no. 2, pp. 1–20, Jan. 2015.

[18] Autorità per le Garanzie nelle Comunicazioni – AGCOM. 2022. and AGCOM, 'Osservatorio sulle Comunicazioni, N. 2/2022'. 2022.

[19] European Parliament, Council of the European Union, *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)*. Accessed: May 02, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32002L0058

[20] R. Metulini and M. Carpita, 'A spatio-temporal indicator for city users based on mobile phone signals and administrative data', *Social Indicators Research*, vol. 156, no. 2–3, pp. 761–781, 2021.

[21] ISTAT, 'Descrizione dei dati geografici e delle variabili censuarie per sezione di censimento Anni 1991, 2001, 2011'. 2015.