Statistical indicators based on mobile phone and street maps data for risk management in small urban areas

Selene Perazzini[1*†], Rodolfo Metulini[2] and Maurizio Carpita[1]

[1*]DMS StatLab, Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, Brescia, 25122, Italy – ORCID ID: 0000-0003-0292-9975 (SP); 0000-0001-7998-5102 (MC).

[2]Department of Economics, University of Bergamo, Via Caniana, 2, Bergamo, 24127, Italy – ORCID ID: 0000-0002-9575-5136.

*Corresponding author(s). E-mail(s): selene.perazzini@unibs.it;
Contributing authors: rodolfo.metulini@unibg.it.;
maurizio.carpita@unibs.it;
†These authors contributed equally to this work.

Abstract

The use of new sources of big data collected at a high-frequency rate in conjunction with administrative data is critical to developing indicators of the exposure to risks of small urban areas. Correctly accounting for the crowding of people and for their movements is crucial to mitigate the effect of natural disasters, while guaranteeing the quality of life in a "smart city" approach. We use two different types of mobile phone data to estimate people crowding and traffic intensity. We analyze the temporal dynamics of crowding and traffic using a Model-Based Functional Cluster Analysis, and their spatial dynamics using the T-mode Principal Component Analysis. Then, we propose five indicators useful for risk management in small urban areas: two composite indicators based on cutting-edge mobile phone dynamic data and three indicators based on open-source street map static data. A case study for the flood-prone area of the Mandolossa (the western outskirts of the city of Brescia, Italy) is presented. We present a multidimensional description of the territory based on the proposed indicators at the level of small areas defined by the Italian National Statistical Institute as "Sezioni di Censimento" and "Aree di Censimento".

Keywords: Minimization of Drive Test Data, T-mode Principal Component Analysis, Crowding Indicator, Traffic Indicator.

1 Introduction

Risk is generally considered as a function of hazard, vulnerability, and exposure [1]. Among these components, exposure represents the people and the tangible human assets located in hazard-prone areas. If the number of people in a certain area increases, the risk increases accordingly. Correctly quantifying the exposure is, therefore, crucial to understand the risk associated with natural (e.g., floods and earthquakes [2]) or non-natural disasters (e.g., car accidents [3], bridge closure [4]) in urban areas. To this aim, information on people crowding, the traffic they generate, and the road network (whose structure shapes the dynamic of traffic) at a high level of disaggregation (i.e., at "small area" level [5]) are needed.

According to the Italian National Institute of STATistics (ISTAT), the highest level of disaggregation is represented by the "Sezione di CEnsimento" (SCE), that roughly corresponds to a portion of a municipality. Some useful information about the geography and demography of the SCEs is freely available on the ISTAT website, such as the surface area and the number of residents according to the last census. However, the sole use of such static information does not allow producing dynamic maps of risk exposure as proposed by Balistrocchi et al. [6] or by Kong et al. [7] in the context of floods studies.

Some risk indicators for small areas can be found in the literature. For example, Lin & Shen [8] and Saghapour et al. [9] develop composite indicators in support of preparedness for pandemic diseases spreads; Wang et al. [10] presents ecological indicators for correctly measuring health in small areas. Nowadays, modern sources of mobile phone data are increasingly combined with satellite and sensor technologies (e.g., [11]) with the aim of producing dynamic information related to the density of people's presences and movements. This approach allows investigating issues of great relevance, such as the monitoring of the impact of social and cultural events [12], the variability in the distribution of presences in the neighborhoods of a large city [13], the seasonality of the second homes in a tourist area [14], the increase of remote working in sparsely populated areas [15] or forecasting traffic in flooding risk areas to support data-driven decision-making [16].

In this paper, we characterize SCEs in terms of the crowding of people (i.e., "city users" according to the definition by Metulini & Carpita [17]) and the dynamic of traffic in different times of the day and on the characteristics of the road network. We combine different kinds of geo-referenced sources of data and construct two composite indicators and three simple indicators of risk: a

"Crowding Indicator" (that we call CRO), a "Traffic Indicator" (that we call TRA), the portion of territory occupied by the road network (that we call STR), the vehicular traffic area dedicated to roadways (that we call ROA), and the portion of the street area occupied by the urban and local roads (that we call ULR).

The case study presented in this work relates to the Mandolossa region, a flood-prone area on the western outskirt of Brescia (Italy). This choice is motivated by our interest in proposing statistical approaches for mapping the exposure in flood-prone areas for the MoSoRe@UniBS Project.

In previous works of ours (e.g., [6] and [17]), we used mobile phone data of people crowding to construct maps of flood exposure. This paper extends the work in Metulini & Carpita [17] by combining a crowding indicator based on city users with traffic indicators constructed using a mixture of two types of mobile phone data and administrative data coming from different sources. In particular, two sources of georeferenced mobile phone data are used to estimate crowding and traffic: the Mobile Phone Density (MPD) data, that were also used in the previous work, and the recently released Minimization Drive Test (MDT) technology data [18]. The two sets of data present different desirable characteristics. The former represents users subscribed to the Telecom Italia Mobile (TIM) company in a squared pixel of 150 meters per side in a 15-minutes time interval. This type of data well captures the crowding of people in small areas and, since available for several years, has been used in various statistical applications in the literature in the last decade (see Section 3.1 for a review). A disadvantage in using this kind of data is that just users from a single company (TIM) are counted, disregarding users of other phone companies. We solve this issue by following the approach in Metulini and Carpita [17] and use the MPD data to define a crowding indicator.

The other type of mobile phone data is very recently released, and, to our knowledge, it has not yet been used in any field other than network engineering, where its use is related to, e.g., smart traffic load maintenance of the network [19]. The MDT data strongly outperforms the MPD in terms of georeferencing. Indeed, the MDT technology captures signals from devices subscribed to TIM and connected to the mobile phone network in a 15-minute time interval with very high accuracy. Similarly to the MPD data, the MDT data are reported on a pixel grid, but the pixels measure 10 meters per side. Both the MPD and MDT data sets represent a spatiotemporal variable observed on a regular pixel grid at subsequent times. However, as we will discuss in more detail in Section 3, the two types of mobile phone data present very different characteristics. MPD data typically offer a detailed representation of the temporal dimension of crowding in (small) areas at a lower geographic resolution. By contrast, being their data collection expensive and time-consuming, MDT data are typically available for short periods of time, but provide a more detailed representation of the spatial dimension. Regarding the MPD data, the data provider recommends considering small aggregations of 2 or 3 pixels, which are quite large. More precisely, a pixel of the MPD grid coincides with the smallest SCEs in the city of Brescia. For this reason, we decided to aggregate the MPD and MDT data in such a way as to represent crowding and traffic in the SCEs. This choice is particularly useful for local administrations because it allows the indicators to be compared with other relevant official statistics published by ISTAT. Furthermore, this choice allows for preserving the small level of geographical detail in the representation of the analyzed phenomena, being the SCE the smallest "nomenclature of territorial units for statistics" in Italy.

In this work, we exploit the different characteristics of the two datasets to analyze two distinct phenomena. Specifically, we use the MPD data to study the crowding and the MDT data to capture the traffic on the roads in a selection of SCEs in the Province of Brescia. In particular, we propose an innovative approach for constructing traffic indicators in small areas. Indeed, the MDT data allows us to estimate the traffic intensity by identifying the devices that are located on the streets. To this aim, the MDT data has been compared to a street map in such a way as to distinguish the signals that originated on the streets and those that did not. For this scope, an accurate street map representing the width of the roadway is needed to identify the cells of the grid that correspond to streets. While retrieving the width of the roads, we realized that available street maps from open-source administrative data are not complete (i.e., some streets are missing). We, therefore, constructed a new comprehensive street map by merging the available ones. This step represents a further contribution of our work to the literature.

We obtain two composite (synthetic) indicators of the exposure of the SCEs, CRO and TRA, that respectively represent the dynamics in the crowding of people and in the traffic. By applying a model-based functional data clustering to the two sets of data we show the presence of different temporal dynamics among SCEs. This evidence motivates us to use a T-mode Principal Component Analysis (PCA) [20] to capture the spatiotemporal patterns in the data. In addition, we define three simple indicators of the static characteristics of the viability of the SCEs - STR, ROA, and ULR - based on the information from the official street maps. We use the five indicators to characterize the exposure of the Mandolossa flood-prone area as well as its complexity, which we express as the variability in the composite indicators among the SCEs of the same municipality (or "Aree di CEnsimento", ACE, according to ISTAT).

The paper organizes as follows: Section 2 discusses the available street maps, shows the construction of the new street map, and defines the indicators STR, ROA, and ULR. Section 3 describes the mobile phone data. In particular, Section 3.1 is dedicated to the MPD data and Section 3.2 to the MDT data, while Section 3.3 discusses the data harmonization process. Section 4 investigates the database by means of model-based functional data clustering. Section 5 illustrates the T-mode PCAs and defines the indicators CRO and TRA. Section 6 presents and discusses results in the flood-prone area of Mandolossa. Section 7 concludes the paper.

2 The street map construction and indicators

The MDT data was used to estimate the traffic since, as we will discuss in detail in subsection 3.2, it captures signals produced by a subset of electronic devices and assigns them to georeferenced cells of a pixel grid with good accuracy. To identify the MDT signals coming from the streets, the pixel grid should be compared with a proper map of the road network. In this Section, we present the construction of the street map, which is a necessary preliminary step for our analysis. Moreover, some data from the street maps are extracted and three variables that represent the main characteristics of the road network in the SCEs are defined. The following analysis and computations were carried out by means of the `R` software. In particular, we referred to the packages **`raster`**, **`rgdal`**, **`rgeos`**, and **`sp`** for geographic data analysis and modeling. The **`mapview`** package was used to produce the presented maps.

Several street maps are currently available, and two types can be distinguished: maps that represent streets as lines (e.g., OpenStreetMap) and those that represent them as polygons. Only the second type is suitable for the purpose of our analyses because it represents the width of the roadway and allows us to identify the phone signals that come from streets.

Two street maps with polygon data are available for the area under analysis: the “DataBase Topografico Regionale” (DBTR) from which we extracted the map of the Province of Brescia (last updated 2021) and the “Uso e copertura del suolo della regione Lombardia 2018” (DUSAF 6.0) released by the Lombardy Region. Both are freely available at the “Geoportale della Lombardia” website (https://www.geoportale.regione.lombardia.it). The DBTR is defined on multiple layers, among which we selected the “vehicular traffic area” and the “street area”. The DUSAF map divides the area into polygons representing land use. Among those, we selected the ones corresponding to streets.

The representation of the streets provided by both maps is rather incomplete. The DBTR is the main street map of the Province and offers a much more comprehensive road network than the DUSAF map, but a few major streets are missing (e.g., the “BreBeMi” highway). Moreover, in each of the two selected layers, many streets interrupt abruptly, and the representation appears patchy. The DUSAF map reports only a few major streets (e.g., highways) and does not capture the road network in residential areas. As one can notice in Figure 1, the two layers of the DBTR and the DUSAF street map only partially overlap. However, the combination of the three (i.e., the two layers of the DBTR and the DUSAF map) provides a detailed representation of the area. Therefore, a new map has been created by overlaying the maps and joining the overlapping polygons. The resulting map (shown in the right map of Figure 1) describes a continuous network and captures the minor roads in the residential areas as well as the fast-paced streets that link the urban areas. Overall, the road network area in the new map is equal to 26 km^2, while the DBTR map, which is the main official street map, covers 22 km^2 and the DUSAF map captures only 9 km^2.

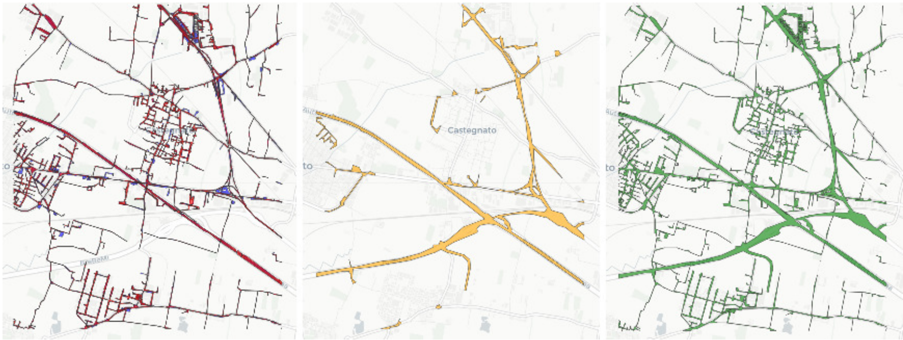

Fig. 1: Example of comparison of the street maps in the DUSAF 6.0 and the DBTR. The maps show the central-west area of the MDT database (latitude 45.53 N - 45.58 N and longitude 10.08095 N - 10.14 N). The left map represents the DBTR and reports the layers of "vehicular traffic area" (blue) and "street area" (red). The central map reports the DUSAF 6.0 map. The right map shows the street map obtained by merging the maps in the DUSAF 6.0 and the DBTR (green).

To describe the main characteristics of the new road network map of the SCEs and provide important insights into the viability of these small areas, we compute three simple indicators. Let | . | be the surface area (computed in m^2) and let N SCE polygons with index j. For each SCE j, we computed the portion of the area occupied by the road network, which we call the STR indicator. We defined STR as the ratio:

$$STR_j = \frac{\mid \text{Streets}_j \mid}{\mid \text{SCE}_j \mid}. \tag{1}$$

The indicator ranges between 0 and 1. We found that, on average, 21% of the area of the SCEs is occupied by the streets. However, the STR indicator largely varies among the SCEs and takes a minimum value equal to 0.002 and a maximum value of 0.99.

A preliminary analysis of the characteristics of the streets reported in the DBTR led to the identification of two additional variables that particularly characterize the SCEs' viability: the portion of the vehicular traffic area dedicated to the roadway and the portion of the street area occupied by the urban and local roads.

The portion of the vehicular traffic area dedicated to the roadway is the part of the road aimed at the flow of vehicles. We computed it from the DBTR but, as previously discussed, a few major roads are not reported in the DBTR. Therefore, information has been integrated with those in the DUSAF maps and the missing streets have been assigned to roadways. We define, for each SCE, the vehicular traffic area dedicated to roadways as the area of roadway

variable	mean	sd	cv	min	max	skew	kurt
STR	0.22	0.14	0.63	0.00	1.00	1.49	7.65
ROA	0.81	0.16	0.20	0.00	1.00	-1.50	6.28
ULR	0.91	0.24	0.26	0.00	1.00	-2.75	9.51

Table 1: Descriptive statistics of the three indicators defined on the street maps (Eq.s 1-3). Reported values are, in order, the variable name, mean, standard deviation, coefficient of variation, minimum, maximum, skewness, and kurtosis.

divided by the area of vehicular traffic:

$$ROA_j = \frac{\mid \text{Roadway}_j \mid}{\mid \text{vehicular traffic}_j \mid}. \tag{2}$$

The indicator ranges between 0 and 1 and we found that the mean value of ROA is 0.8, indicating that on average 80% of the vehicular traffic area of the SCEs is dedicated to roadways.

To compute the portion of the street area occupied by the urban and local roads, we considered the variable "technical-functional classification" in the "street area" layer that divides the roads into 6 categories: highways, primary roads outside urban areas, secondary roads outside urban areas, urban neighborhood roads, urban freeways, and local roads. According to the data, the variable presents some missing values in correspondence to some inhabited centers. A careful graphic inspection revealed that these polygons correspond to urban and local roads. Therefore, the modalities representing urban and local roads have been grouped into a new one and the missing values have been assigned to this class. At last, data have been integrated with the DUSAF maps and, after a graphical inspection, the roads not reported in the DBTR have been assigned to the "highways" modality.

The street categories and their distribution among the SCEs have been analyzed. We found that less than 6% of the SCEs have highways, 14% have primary roads and 11% have secondary roads outside urban areas. At last, we computed the portion of street area per each SCE j occupied by the urban and local roads:

$$ULR_j = \frac{\mid \text{Urban and local roads}_j \mid}{\mid \text{Streets}_j \mid}. \tag{3}$$

The indicator ULR ranges between 0 and 1, with a mean equal to 0.91.

To conclude, Table 1 presents the main characteristics of the indicators STR, ROA, and ULR.

3 Mobile phone data description and processing

Two data sets were used for the analysis: the Mobile Phone Density (MPD) data, and the Minimization of Drive Test (MDT) data. Both datasets refer to users subscribed to the TIM company, which is currently the largest operator in Italy. MPD data was provided by the Municipality of Brescia in the context of a territorial monitoring project developed between 2014 and 2016 by its statistical office in cooperation with the DMS StatLab (Data Methods and Systems Statistical Laboratory) of the University of Brescia. Nowadays, the data is also available on a dashboard released by TIM. The second dataset has 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 MDT data at our disposal refers to 2021 and is not available for previous years. For this reason, we were forced to use data from different periods of time. We address this aspect in Section 3.3, where we also show how the data have been treated in order to overcome this issue. These two sources of data, as well as mobile phone signals in general, suffer from the issue of measurement error, as stressed by many works (see, e.g., [21]). This issue might be mitigated by integrating different mobile phone data sources [22]. Unfortunately, the mobile phone data used in this work, which are proprietary, are provided without the associated measurement error. However, as discussed in the Introduction, the data have been aggregated into SCEs in order to mitigate the measurement error in the MPD data.

The two sets of data do not just differ in the period of observation, they also come with different geographical resolutions. Indeed, both the data are reported on a grid of pixels, but the pixels of the MPD grid measure 150 meters per side, while the MDT ones are 10 meters per side. To clarify, MPD data pixels often cover the smallest SCEs, while MDT data pixels are so small that they can be tied back to bits of roads or buildings. On the other hand, MPD data is much easier to collect and can be provided by the telephone operator within an hour of observing it. MDT data, by contrast, requires longer processing times, as well as the installation and activation of particular technologies, which must be tested on-site before data collection. This makes the collection of MDT signals much more time-consuming and expensive and also poses limits on the extent of the territory observed. As a result, datasets of MDT signals typically cover much shorter time periods and smaller areas than those of MPD data. Specifically, in our setting, the MPD database covers a rectangular area of approximately 80×128 km^2 containing the whole Province of Brescia, while the MDT dataset covers a rectangular area of approximately 10.5×14.4 km^2 within the Province of Brescia.

The following subsections discuss in detail the MPD and the MDT data.

3.1 Crowding data (MPD)

The MPD data come from mobile phone signals retrieved by TIM and have progressively aroused the enthusiasm of the urban planners' community [23, 24]. In our setting, MPD data are the average number of mobile phone SIMs (both calling and not calling) that are assigned to a given cell of the pixel grid in a specific quarter. MPD data have been used in the context of urban planning. For example, Carpita and Simonetto [12] applied statistical methods to analyze the presence of people during big events in the city of Brescia such as the "1000 Miglia" car race. Zanini et al. [25] found, by means of an independent component analysis, a number of spatial components that separate the main areas of the city of Milano. Secchi et al. [26] applied the method of Bagging Voronoi Treelet analysis to identify sub-regions of the metropolitan area of Milan sharing a similar pattern over time.

Our MPD database refers to the TIM mobile phone signals recorded from April 1st, 2014 to August 11th, 2016, in a rectangular region defined by latitude 45.21 N–46.36 N and longitude 9.83 N–10.85 N (province of Brescia). Data were aggregated into 923 × 607 squared cells, and they are available at intervals of 15 minutes, for a total of more than 40,000 million records. For each cell and for each time interval, the corresponding record refers to the number of mobile phones simultaneously connected to the network in that area in that time interval, with phones whose signals were retrieved in more than one area attributed just to the area related to their last signal in that quarter. Data can be seen as a spatial raster, with the grid's color intensity expressing the number of mobile phone signals. Data have been retrieved anonymously. Moreover, the mobility feature of these data is hidden, in the sense that it is not possible to trace a single person over time. It is worth saying that, for MPD data to be reliable, the considered polygon should be at least 400 meters on each side (i.e., 2 or 3 pixels), so, grid cells have been aggregated accordingly.

Since the object of analysis is the SCEs, which are irregular polygons, while raw data are in the form of rectangular grids of a raster, we need an aggregation strategy to obtain the total number of people in each polygon. The weighted scheme adopted by Metulini & Carpita [17] is an overlapping strategy that assigns the number of people in the raster grid to the polygon based on the share of the overlapping area. Given the cells of the grid $Cell_k$ with $k = 1, 2, \ldots, K$ overlapping a specified SCE j, the ratio

$$A_{jk} = \frac{\mid SCE_j \cap Cell_k \mid}{\mid Cell_k \mid} \tag{4}$$

represents the portion of $Cell_k$ covered by the chosen SCE j. Let MPD_k be the MPD in $Cell_k$, the estimated MPD in $SCE\ j$ is computed as

$$MPD_j = \sum_k MPD_k \cdot A_{jk} \tag{5}$$

with A_{jk} defined in Eq. (4). It is worth observing that, by considering MPDs at the SCE level (Eq. 5), we mitigate the impact of the geolocalization inaccuracy since the average area of the considered SCEs is 164559 m^2. Another issue related to this kind of data is that they refer to just one mobile phone company. Although variable rescaling does not affect the results when constructing indicators, we follow [17] and pre-process the data to obtain an estimate of users of other mobile phone companies as well. For a national-level analysis, a convenient solution is represented by using the market share of TIM company (that stands at 30.2% according to "Il Sole 24 Ore" newspaper dated December 2016) and applying it to raw data so to retrieve an estimation for the total number of people. However, for the case of small areas, Metulini & Carpita [17] showed that TIM market share varies among national and municipality dimensions because of differences in incomes and the demographic structure. For this reason, a market ratio for Brescia should be used. The authors also show that TIM market share does not vary among the five districts that constitute the municipality of Brescia, which avoids the need to calculate the market share for smaller areas. According to their strategy, based on comparing the number of residents aged 11-80 at January 1^{st}, 2016 from the ISTAT administrative archive with the number of mobile phone signals in different city districts in three different quarters of the late evening (i.e. 20:00-20:15, 21:00-21:15, 22:00-22:15) and in 42 different weekdays between 2015 and 2016, the market share ratio stands at about 23% in the area of Brescia. We let $ETMS = 23\%$ to be the estimated TIM market share and, as a final step, to obtain an Adjusted measure for MPD (AMPD) which considers all the mobile phone users (not only the TIM users) at quarter t, we multiply MPD by the $1/ETMS$ according to the following formula:

$$AMPD_{jt} = \frac{MPD_{jt}}{ETMS} \tag{6}$$

3.2 Signal data (MDT)

“Minimization of Drive Test” [18, 19, 27] refers to a recent technology that captures signals produced by phone calls, text messages, internet browsing, and technical operations (e.g., location update) of devices with a SIM associated to the TIM company, transmitted over the 3G/4G mobile network from/to terminal devices with GPS enabled. The methodology registers radio measurements of the signals on a geo-referenced grid of pixels measuring 10 meters on each side. MDT data have been only recently made available by TIM and, so far, have found few applications in the engineering literature for purposes related to the technical control of telephone networks (e.g., smart traffic load maintenance). In this work, we propose an innovative application of the MDT data for statistical purposes. In fact, the MDT technology allows for high accuracy in users’ geolocalization (i.e., 10 meters), which cannot be achieved by other types of mobile phone data. For this reason, it is possible to identify the devices that produce MDT signals from streets. In turn, this information can be used to estimate traffic and construct traffic indicators.

Our database collects all the MDT signals registered on 5 different days of the week in November 2021 in a pixel grid representing a rectangular area of 150 km^2 in the Mandolossa region. Since the detection of MDT signals requires particular technologies to be activated, the days of the data collection were carefully chosen. The first day - Wednesday 10 - was sampled and the collected data were analyzed. Once assessed the adequacy of the data for the analysis, the other 4 days of detection were chosen in such a way as to cover a typical week: namely Friday 19, Saturday 20, Sunday 21, and Monday 22. For

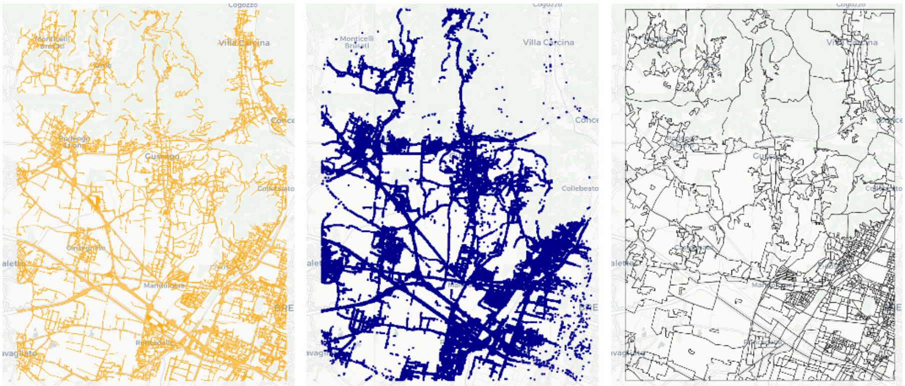

Fig. 2: MDT data pre-processing. Left: street network (orange); center: MDT grid cells on streets (blue points); right: SCE boundaries (black).

each day, 96 times of observations were collected and corresponded to four 15-minute intervals per hour (i.e., for each hour of each day, we have observations at minutes: 00-14, 15-29, 30-44, 45-59). For each time interval, the database reports the total number of signals registered in a cell of the pixel grid.

A few aspects should be carefully evaluated when analyzing MDT data. First, the database represents a sample of a much wider population, as about 10% of the current electronic devices produce MDT signals. Second, a device can produce multiple MDT signals at a time and a signal cannot be traced back to the device from which it has been generated. To overcome this issue, we did not consider the number of signals in a SCE and referred instead to the number of cells of the grid from which signals were generated. Given the limited number of electronic devices captured and the short time intervals observed, this choice nicely represents the traffic intensity in the selected area.

We restricted the database to the cells of the grid located on the streets. To this aim, we compared the MDT database with the street map constructed in Section 2 and reported in the left plot of Figure 2. As previously discussed, our street map represents the width of the roadway. Therefore, we identified the cells of the MDT pixel grid that correspond to streets by overlaying the grid on the street map. Since MDT signals are geolocated with 10 meters accuracy, we considered "on street" all the cells that are at most 10 meters far from the roadway. This step reduced the database to approximately 49% of the grid, as shown in the central map of Figure 2.

We compared the MDT data to the administrative boundaries map (see the right plot of Figure 2) and assigned each street cell to the corresponding SCE. This step led to the construction of an 834 × 480 matrix where each row represents a SCE where at least one MDT signal has been detected on a street during the observed days, the columns refer to the times of observation (i.e., 96 15-minutes intervals collected for the 5 considered days), and entries report the on street-cells counts. Then, the MDT counts have been divided by the area of the road network of the SCE:

$$MDT_{jt} = \frac{Number(\text{Streets Cells}_{jt})}{\mid \text{Streets}_j \mid}. \tag{7}$$

Therefore, we obtained MDT variables representing the presence of individuals on the streets of the SCEs at different times of the day that can be compared independently of the dimension of the streets in each SCE.

3.3 Harmonization of the AMPD and MDT data

In this work, we use AMPD data to estimate the presence of people in the SCEs and construct an indicator of city users' crowding. Moreover, we take advantage of the higher georeferencing accuracy of the MDT data to estimate the traffic intensity and define an indicator of street traffic. To this aim, a few aspects of the data should be carefully evaluated. First, as shown in the previous section, the two datasets refer to different periods of observation: the MPD data were collected between April 2014 and August 2016, while the MDT data refer to five days of November 2021. In order to overcome this issue and guarantee the comparability of the data, we considered the observation of November only. In other words, we restricted the AMPD database to the 30 days of November 2015 and kept the whole MDT dataset. Second, we aim at the definition of small area indicators, but the AMPD and the MDT data are collected on a pixel grid. Therefore, the grid cells have been aggregated into SCEs according to the administrative boundaries map published by ISTAT. Note that the MDT database only captures a fraction of the territory of the SCEs located at the borders of the area under analysis, which refers to the Mandolossa, as clearly shown in Figure 3.

The two sets of data were analyzed. We found that the number of cells that originated MDT signals on streets is very low at night and can vary considerably during the day. High variability has been observed in the AMPD quarter time series as well. Therefore, the AMPD and the MDT quarter data have been averaged into hours. Moreover, since Metulini & Carpita [17] observed considerable differences in the temporal dynamics of city users among weekends and weekdays, observations corresponding to the two groups of days have been distinguished and the corresponding hourly intervals have been averaged into two time series of 24 hours. For the AMPD data, the midweek 24-hour time series was obtained by averaging 21 days of November 2015 (5 Mondays and 4 Tuesdays, Wednesdays, Thursdays, and Fridays), and the weekend time series by averaging 9 days (i.e., 4 Saturdays and 5 Sundays)[1]. By contrast, for the MDT data, 3 days of November 2021 (Wednesday 10, Friday 19, and Monday 22) were averaged to compute the midweek time series, while the weekend time series was computed on 2 days (Saturday and Sunday 20-21). The main descriptive statistics of the variables AMPD and MDT are reported in Table 2.

[1] The AMPD time interval corresponding to Sunday, November 1st, 2015 at 00-01 showed anomalous values and has been neglected.

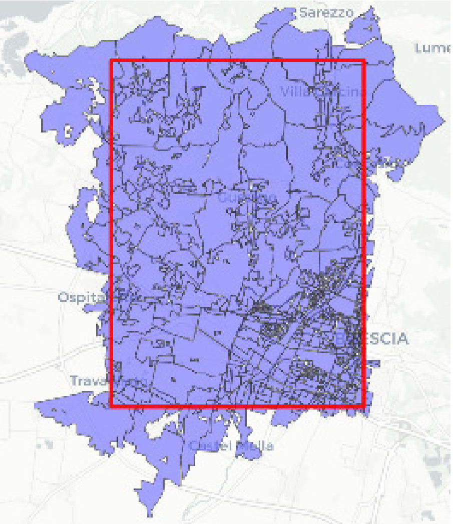

Fig. 3: Map of the SCEs in the MDT database. The map represents the SCEs in the Mandolossa area (blue polygons) and the area described by the MDT database (red rectangle).

variable	mean	sd	cv	min	max	skew	kurt
AMPD	62.11	104.76	1.69	0.01	1329.37	3.85	23.83
MDT	2.71	4.20	1.55	0	72.84	4.72	42.01

Table 2: Descriptive statistics of the variables AMPD and MDT defined in Eq.s (6) and (7) at hourly intervals. Reported values are, in order, the variable name, mean, standard deviation, coefficient of variation, minimum, maximum, skewness, and kurtosis. Values of the MDT data have been multiplied by 10,000. Moreover, statistics of the MDT data have been computed considering SCEs for which at least 12 time intervals with values different from zero were available.

Note that the statistics for the MDT data have been computed considering SCEs for which at least 12 time intervals with values different from zero were available. This choice allows avoiding values to be affected by SCEs with few roads or little traffic flows especially at night.

A clustering of functional curves has been performed on the obtained time series for both the AMPD and the MDT data. The analysis showed the presence of retrieved features in the two sets of data. To capture the spatiotemporal patterns and summarize the information in two synthetic indicators, two T-mode PCAs have been performed. However, for the PCA, the hourly time intervals have been further aggregated into six time intervals of 4 hours: (0-4], (4-8], (8-12], (12-16], (16-20], (20-24]. This choice is motivated by the fact that the MDT hourly data still showed high variability in all the SCEs. Indeed, the hourly variation ranges on average between -40% and +240%. For comparability, the AMPD data have been aggregated into the same 12 time intervals by applying Eq. (6) with t equal to the 12 considered intervals.

4 Model-based functional cluster analysis of MPD and MDT data

In this section, we consider the time series of AMPD and MDT data, at the level of SCE, as functional curves, in order to grasp a possible clustering structure in terms of regularities in the temporal dynamics of city users and traffic. Indeed, the MPD and MDT data can be interpreted as independent realizations of two continuous stochastic processes whose functional expressions are unknown and a realization at discrete points in time is the only known element.

Jacques & Preda (2013) [28] distinguish three classes of clustering methods for functional data: the raw data, the distance-based, and the model-based (also known as filtering) methods. The former ones do not consider the raw data as realizations of a continuous stochastic process and have therefore been excluded. By contrast, the latter two rely on considering the realizations on discrete points as coming from a continuous process and hence estimate proper functional curves. The distance-based methods use clustering algorithms based on specific traditional distances methods (such as the k-means one) adapted to functional data, while the model-based methods approximate the curves into some basis functions and perform clustering using the basis expansion coefficients. Among those, we chose the model-based methods because we are interested in the interpretation of the estimated curves' parameters.

We adopted the model-based functional cluster analysis (M-B FCA) [29, 30] to cluster SCEs, which has been applied by means of the `funFEM` package in `R`. The M-B FCA presented in [30] is a clustering algorithm for functional data that clusters a set of observed curves into K homogenous groups based on a Discriminative Functional Mixture (DFM) model. The algorithm aims to cluster a set of observed curves $\{x_1, \dots x_n\}$ generated by an unknown stochastic process $X(t) = \sum_{r=1}^{p} \gamma_r(X) \upsilon_r(t)$ defined over a random vector $\gamma = (\gamma_1(X), \dots, \gamma_p(X))$ and a set of basis functions $\{\upsilon_1, \dots, \upsilon_p\}$ with p assumed known and fixed. To do so, it estimates the probability that the curve x_i belongs to the k-th cluster by maximizing the expectation of the data log-likelihood conditionally to the $(p \times d)$ orthogonal matrix U of the most discriminative latent subspace for the K groups spanned by d basis functions $\{\varphi_1, \dots \varphi_d\}$. The latter basis functions are obtained as $\varphi_r = \sum_{l=1}^{p} u_{rl} \upsilon_l$ such that $U = (u_{rl})$ is orthogonal for $r = 1, \dots, d$ with $d < K$ and $d < p$. This relationship implies that

$$\Gamma = U\Lambda + \epsilon \tag{8}$$

where Γ is a $(p \times 1)$ random vector, Λ is the random $(d \times 1)$ vector of the latent expansion coefficients of the observed curves $\{x_1, \dots x_n\}$ on the basis $\{\varphi_1, \dots \varphi_d\}$, and ϵ is a vector of length p of independent and random noise terms. We assume that: (i) conditionally to the k-th cluster, Λ is distributed according to a multivariate Gaussian density with $(d \times d)$ covariance matrix Σ_k; (ii) ϵ is distributed according to a multivariate Gaussian density; (iii) the $(p \times p)$ covariance matrix of $W'\Gamma$ conditional to the k-th group is a block diagonal matrix:

$$\begin{bmatrix} \Sigma_k & \mathbf{0} \\ & \\ \mathbf{0} & \begin{bmatrix} \beta & & 0 \\ & \ddots & \\ 0 & & \beta \end{bmatrix} \end{bmatrix} \tag{9}$$

where $W = [U, V]$ and V is the orthogonal complement of U. Eq.(8) and assumptions (i)-(iii) define a family of 12 DFM models capturing the variance of the data of the k-th group through Σ_k and the variance of the noise outside the functional subspace by means of the parameters β. The model's parameters are estimated via the Fisher-EM algorithm [31].

The M-B FCA was performed separately on the two sets of data AMPD and MDT. Indeed, we are not interested to analyse them jointly since the two data represent two distinct phenomena. In our framework, each realization x_i of the stochastic process corresponds to one SCE observed for 48 points in time (i.e., 24 midweek and 24 weekend hours). An inspective analysis of the MDT database showed that only 470 of 1072 analyzed SCEs corresponded to time series with more than 12 hours different from 0. Therefore, for the MDT database only, these 470 SCEs were grouped into a fictitious cluster, and the M-B FCA was performed on the remaining 602 SCEs. Differently, for the AMPD dataset, all the 1072 SCEs have been considered in the M-B FCA. As a preliminary step, the curves have been standardized in such a way that each x_i has 0 mean and standard deviation equal to 1 to compensate for the different geographical extent of the SCEs. In doing so, the amplitude of the curves has been regularized. It is worth noticing that we chose not to regularize curves in terms of their phases as we aim at accounting for the differences in curves' periodicity by means of the clustering analysis.

To model the functional curves, we used Fourier basis functions because they are particularly suitable to describe periodic data such as time series [28]. As far as the number of basis functions is concerned, p has been chosen within a range of values between 1 (i.e., a constant function) and 15 (i.e., a function with a constant, 7 cosines, and 7 sines) on the basis of the sum of the root mean squared errors (RMSE) computed between the process' realization x_i and the smoothed estimated curve evaluated at discrete points of time (see the left plots of Figure 4). An illustrative example of the fitting of the smoothing curves is reported in the right plots of Figure 4, where results for a randomly chosen SCE have been reported. As it could be noticed, at least 9 basis functions are necessary for the curves to satisfactorily approximate the observed data. Therefore, $p = 9$ with $\gamma = \{\gamma_0, \gamma_{cos_1}, \gamma_{sin_1}, \gamma_{cos_2}, \gamma_{sin_2}, \gamma_{cos_3}, \gamma_{sin_3}, \gamma_{cos_4}, \gamma_{sin_4}\}$ has been chosen, for both AMPD and MDT data, as the result of a trade-off

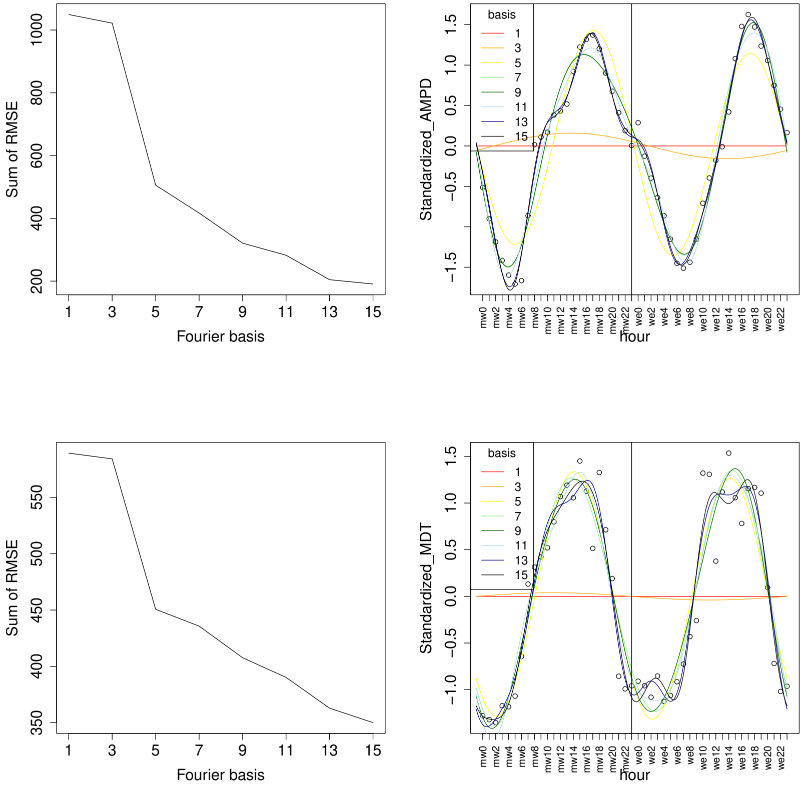

Fig. 4: Selection of the number of Fourier basis functions among $p = 1, 3, 5, 7, 9, 11, 13, 15$. Left: sum of RMSE between the process' realization x_i and the smoothed estimated curve evaluated at discrete points in time. Right: Fitting of the smoothing curves to the observed points in time for a randomly chosen SCE. Top: AMPD data. Bottom: MDT data.

between the need to correctly model the curves and the aim of estimating a parsimonious model.

The best model among the 12 composing the DFM family defined in Eq.(8) and assumptions (i)-(iii) has been chosen along with the optimal number of clusters using the Bayesian Information Criterion (BIC)[2]. Since the BIC tends to overestimate the number of clusters when the M-B FCA is performed on non-Gaussian clusters (see [32], Section 3.3, p. 97), we posed a constraint on the minimum dimension of the clusters (at least 100 SCEs each). This choice also eases the interpretation of the results. We found that the best model for the AMPD data is a DFM where the noise variance is allowed to vary across

[2] Same results are obtained using the Akaike Information Criterion (AIC).

Table 3: Dimension of the clusters for the AMPD and the MDT data.

AMPD		**MDT**	
Cluster	**Dimension**	**Cluster**	**Dimension**
1	223	0	470
2	186	1	216
3	229	2	106
4	229	3	100
5	205	4	180
Total	1072	Total	1072

Table 4: Estimated $\gamma^2_{cos_i} + \gamma^2_{sin_i}$ associated with the Fourier basis functions of the i-th period ($i = 1, \ldots, 4$) for each cluster obtained on the AMPD (top) and MDT (bottom) data.

AMPD

Period (i)	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
1 (24 hours)	1.05	3.75	1.49	0.32	0.02
2 (12 hours)	34.38	1.33	23.63	38.54	34.50
3 (6 hours)	4.99	3.60	1.76	0.36	0.63
4 (3 hours)	0.17	5.83	1.96	1.44	0.68
Total	40.59	14.51	28.84	40.66	35.83

MDT

Period (i)	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
1 (24 hours)	0.04	0.09	0.17	0.13
2 (12 hours)	14.93	10.87	1.75	27.65
3 (6 hours)	0.19	0.17	0.24	0.29
4 (3 hours)	1.13	0.84	0.81	2.02
Total	16.29	11.97	2.97	30.09

groups. For the MDT, the best model resulted in the DFM with a diagonal Σ_k matrix and noise variance not restricted to be common across clusters. The optimal choice resulted in $K = 5$ clusters of curves for the AMPD data. For the MDT data, we obtained $K = 4$, to which we add the previously defined cluster containing SCEs with zero or few MDT signals detected ("cluster 0"). Table 3 reports the dimensions of the obtained clusters.

For each SCE, the probability of belonging to each cluster has been analyzed. We found that most of the probabilities are close to 0 or 1, with very few intermediate values. This evidence suggests that the method performs well in assigning the SCEs to each group.

Let us indicate with γ_{cos_i} and γ_{sin_i} the estimated coefficients associated with, respectively, the i-th cosine and sine for $i = 1, \ldots, 4$. The period (or phase) of the four bases corresponds to, in order, 24, 12, 6, and 3 hours. The adopted M-B FCA method estimates a set of γ coefficients for each cluster, whose associated parameters can be evaluated to characterize the smoothed curves of the clusters' centroids. According to [33], the variability in the values

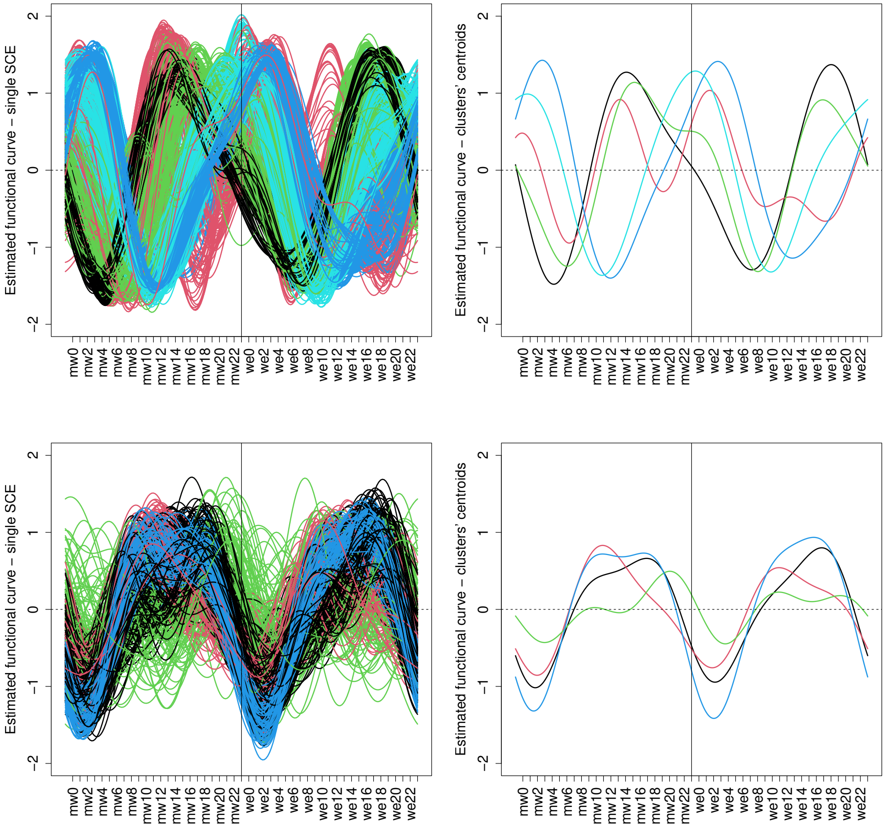

Fig. 5: M-B FCA results for the SCEs in the Mandolossa area. Full set of clustered curves (left) and centroids by cluster (right). Top: AMPD data. Reported curves correspond to clusters 1 (black), 2 (red), 3 (green), 4 (blue), and 5 (light blue). Bottom: MDT data. Reported curves correspond to clusters are 1 (black), 2 (red), 3 (green), and 4 (blue).

of the smoothed curves can be evaluated in terms of the sum of the squares of the estimated coefficients associated with each Fourier basis i, i.e., $\gamma^2_{cos_i} + \gamma^2_{sin_i}$. These values have been reported in Table 4 for each cluster. The corresponding curves for both the two sets of data are shown in Figure 5, where the left charts report all the estimated functional curves and the right ones show the clusters' centroids.

As far as AMPD data are concerned, according to Table 4, the curves of all the clusters but the second are strongly determined by the Fourier basis functions corresponding to the period of 12 hours (estimated values are 34.38, 23.63, 38.54 and 34.50). This evidence clearly emerges in the centroids' curves shown in the top right chart of Figure 5. Indeed, all but the red cluster display one negative and one positive peak per day. By contrast, cluster 2 appears strongly determined by the periods of 6 and 3 hours, and the centroid's curve displays two negative and two positive peaks per day. Moreover, the sum of the estimated parameters per each period, i.e. $\sum_{i=1}^{4} = \gamma_{cos_i}^2 + \gamma_{sin_i}^2$, can be interpreted as a measure of the variability of the estimated curves in the clusters. As shown in the last row of the AMPD Table in 4 (top), clusters 1 and 4 display the highest variability.

For the MDT data, Table 4 (bottom) reports high values for the period of 12 hours for all but the third cluster, and it could be noticed that all but the green curve in the bottom right chart of Figure 5 display one negative and one positive peak per day. According to the sum of the estimated parameters, the largest variation is registered for cluster 4, while cluster 3 shows the lowest.

5 T-Mode principal component analysis of MPD and MDT data

The preliminary analysis of the two datasets showed the presence of heterogeneity among the SCEs and retrieved features over time. In order to ease the interpretation of such datasets, the Principal Component Analysis (PCA) is typically applied to reduce the dimensionality in such a way that most of the information in the data is preserved. Several types of PCAs for spatiotemporal data can be found in the literature [34] and, among those, the PCA with T-mode decomposition of the dispersion matrix is particularly suitable to our data. Indeed, this analysis simplifies the observed time series into discrete groups of clustered time intervals and isolates subgroups of observations with similar spatial patterns [20]. The methodology has been widely adopted in meteorology literature with the aim to identify spatial patterns in climate phenomena (e.g., [20, 35–37]), but it is suitable to investigate any other spatiotemporal phenomena.

The T-mode PCA requires transposing the data matrix Z in such a way that each analyzed variable refers to a time of observation [38]. Therefore, the obtained data matrix Z_T has n rows representing small areas and m columns reporting a variable observed at different times. In our case study, the rows of Z_T correspond to the SCEs j, columns represent the time intervals t, and entries are either the $AMPD_{jt}$ or MDT_{jt}. The two variables have been standardized to prevent scaling to affect results. Specifically, our Z_T matrix has dimensions (1043×12). Indeed, some outliers in the AMPD data have been detected and the corresponding 29 small SCEs have been excluded. These outliers correspond to very small SCEs mostly on the borders of the analyzed rectangle of coordinates and their limited area might have affected the measurement of the AMPD data. As a consequence, the analyzed correlation matrix is the 1043×1043 $Z_T' Z_T$. The T-mode PCA decomposes Z_T in a 1043×12 standardized scores matrix F_T and a 12×12 loading matrix A_T:

$$Z_T = F_T A_T'. \tag{10}$$

Per each component, the loadings in A_T can be plotted as a time series, and spatial patterns can be observed by mapping the corresponding scores in F_T. The `R` package `FactoMineR` was used for the PCA computations. Moreover, some tests from the packages `PCAtest` and `EFAtools` were used to assess the adequacy of the data to the analysis and the goodness of results.

As a preliminary step, the KMO sampling statistic [39] has been computed to assess the adequacy of the two sets of data to the PCA. The KMO measures the proportion of variance among a set of variables that might be common variance and represents the degree to which each observed variable is predicted by the others. The higher the KMO, which can take values between 0 and 1, the higher the proportion of (potentially) common variance. We found that the overall KMO value is 0.829 for the AMPD data and 0.932 for the MDT and that the single variable's KMO values are all above 0.75. Therefore, we conclude that the two sets of data as well as all the variables that they contain are suitable for the analysis.

Then, we tested the presence of non-random correlation in the data, which is a necessary condition for the PCA. To this aim, we applied the Vieira's ψ [40] and the Gleason and Staelin's ϕ [41] statistics. Both statistics capture the degree of correlation between the observed variables by computing the area between the ranked eigenvalues e_w line and the horizontal line at 1. Indeed, if variables are uncorrelated, their associated eigenvalues tend to 1 and approximate a horizontal line when plotted in a ranked order. By contrast, the more the variables are correlated, the bigger some eigenvalues and the steeper the curve of the ranked values. Specifically, the Vieira's statistic is defined as $\psi = \sum_w (e_w - 1)^2$ and takes values between 0 and $m(m-1) = 132$; while the Gleason and Staelin's statistic is $\phi = \sqrt{(\sum_w e_w^2 - m)/(m^2 - m)}$ and takes values between 0 and 1. The statistics are then compared with the corresponding ones estimated from over 1000 uncorrelated datasets generated by randomly permuting the measurements on each variable of the original dataset. In this way, we are able to determine the probability of the null hypothesis that the obtained value of the statistics can be observed upon a dataset of uncorrelated variables. For the AMPD data, we found $\psi = 112.62$ and $\phi = 0.92$, while for the MDT data, we found $\psi = 64.31$ and $\phi = 0.70$. The associated p-values are all approximately 0. Therefore, the two tests strongly support the hypothesis that the correlation structure in the two sets of variables is non-random. At last, the variables have been standardized to prevent results from being affected by the scaling.

Three criteria have been used for the selection of the principal components (PC): the Kaiser–Guttman criterion (i.e., "eigenvalues greater than one"), the elbow method, and the rank-of-roots statistic [42], which is defined as the percentage of explained variance. The three methods led to the identification of one component only for the AMPD variables, which explains 92.8% of the variance. The component is highly and positively associated with all the observed time intervals (see the left plot of Figure 6). As a further check, we computed

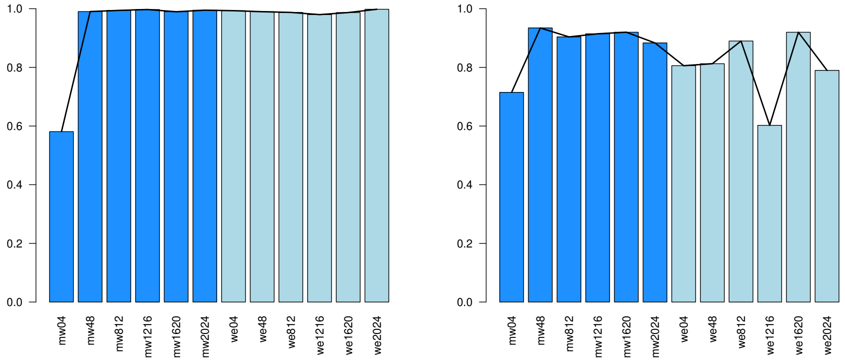

Fig. 6: Loadings for the first principal component obtained from the PCA on AMPD (left) and MDT (right) variables. Midweek intervals are indicated by *mw* and are represented by blue bars, while weekends are indicated as *we* and represented by light blue bars. Time intervals are reported in order (from 0-4 to 20-24) and indicated by the starting and ending hour.

the Vieira's index of the loadings [40], which quantifies how likely is each variable to be associated with a PC. Specifically, the index is here computed as $IL_t = (u_t^2 e^2)/s_t$ where t indicates a variable (i.e., an observed time interval), u_t is the loading of the t-th variable, and s_t is the standard deviation of the t-th variable. The index of the loadings is used to test the significance of the contributions of the original variables to the PC via permutation. Applied to the AMPD data, the index indicates that all the variables' contributions to the first component are significant. We conclude that all the information in the AMPD data is summarized in the first component, which can therefore be interpreted as a measure of crowding, and to which we refer as the CRO indicator.

As far as the MDT variables are concerned, two principal components can be identified according to all three aforementioned criteria. However, the first principal component explains 71.6% of the variance, while the second one 10.3%. Therefore, we restricted our attention to the first one. This choice is supported by the Vieira's index of the loadings, which indicates that all 12 intervals have significant contributions to the first component. The component is positively associated with all the variables and is mostly determined by the daytime hours and the midweek days, as shown by the loadings plot in Figure 6 (right). Therefore, this component is a measure of the street traffic and we refer to it as the TRA Indicator.

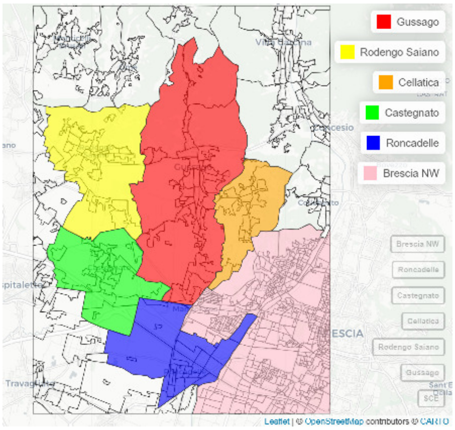

Fig. 7: Map of the six selected ACEs.

6 Results and discussion

In this Section, we discuss the results of the M-B FCA and of the T-mode PCAs. Results are presented both at the SCE level (i.e., by showing the distribution of the indicators on the full sample) and at the ACE ("Area di CEnsimento" ISTAT, i.e. by studying the distribution of the indicators among the SCEs that belong to the same ACE) level. In this respect, 5 major ACEs (that are municipalities) have been identified, namely Castegnato, Cellatica, Gussago, Rodengo Saiano, and Roncadelle. In addition, a new ACE has been created by merging 14 small ACEs of the municipality of Brescia included in the analyzed area. We refer to the latter as "Brescia North-West (NW)". The selected ACEs are shown in Figure 7. As one can notice, Brescia NW includes most of the SCEs (many very small) in the area, namely 67% of the SCEs that constitute the selected ACEs. We focused our attention on these 6 ACEs as they cover the flood-prone area and are therefore crucial for risk management.

In Figure 8 the M-B FCA results on the AMPD and on the MDT data are plotted on a map. In this way, each SCE is represented on the map with a color corresponding to the cluster it belongs to. Some spatial patterns emerge in both maps, especially in the AMPD data (left map). Overall, the two sets of data result in very different geographical distributions of the clusters.

The spine plots in Figure 9 represent the proportion of SCEs belonging to each M-B FCA cluster in the analyzed ACEs. As far as AMPD data are concerned (left chart), we note that most of the SCEs in Brescia NW belong to the green, red, or black clusters which correspond to areas where people gather during leisure time; Castegnato is almost equally composed of all the clusters but the red one; Cellatica collects mostly SCEs in the green and light blue cluster, which correspond to areas mostly crowded at night and might therefore be traced back to residential areas; while Gussago, Rodengo Saiano, and Roncadelle are mostly composed of blue clusters' SCEs. By contrast, smaller

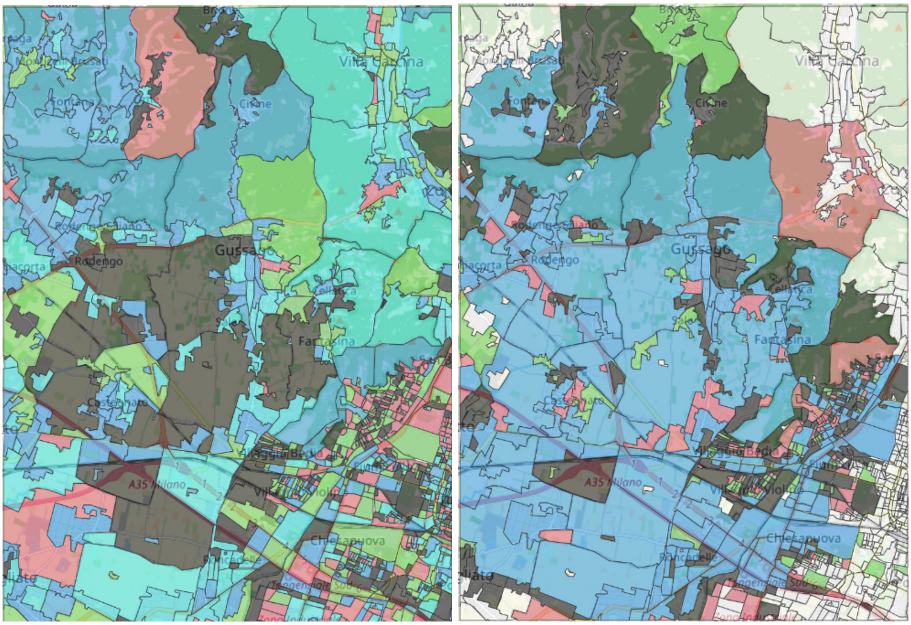

Fig. 8: Map of SCEs, colored by the belonging of the M-B FCA clusters. Clusters are represented with colors as in Figure 5. Left: AMPD data. Right: MDT data. The white SCEs in the MDT map indicate areas where less than 12 hourly intervals are different from 0 (cluster 0).

differences emerge in the MDT clusters (right chart of Figure 9). Brescia NW appears largely composed of the white cluster representing the SCEs for which few or no signals have been detected on the streets. This might be partially due to the very limited area of the SCEs in the city center, which determines a lower probability of detecting MDT signals. Despite this, all the ACEs appear mostly composed of black and blue clusters.

After all, these results highlight the existence of a strong heterogeneity in the temporal dynamic of city users and traffic among SCEs, as well as among the SCEs of the same ACE. This heterogeneity further motivated us to propose indicators of exposure for small areas that are able to account for the temporal dynamics of crowding and traffic on the road network.

The T-mode PCAs presented in Section 5 led to the identification of two composite indicators: the crowding CRO and the traffic TRA. The SCEs' scores for CRO and TRA have been normalized to the $[0, 1]$ range and averaged to represent the ACEs. The average values of the two indicators in the SCEs of the ACEs are shown in Figure 10. The left map reports the CRO indicator, and Brescia NW appears the most crowded area. In the right plot of Figure 10 representing the TRA indicator the ACEs with intense traffic emerge. In particular, the maximum value corresponds to Roncadelle, where there are a few

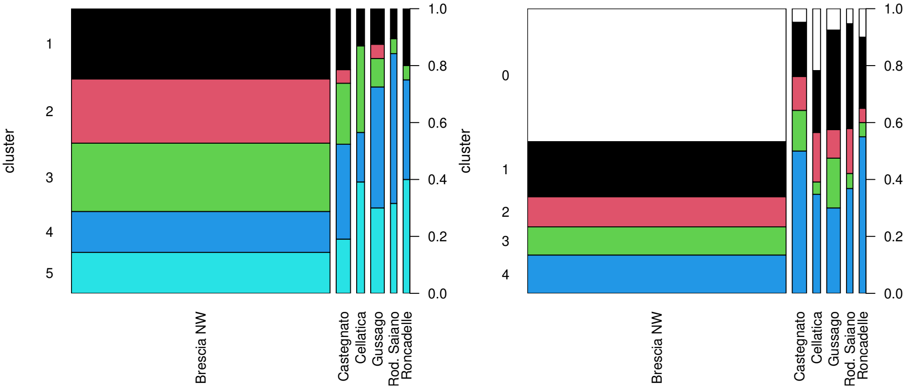

Fig. 9: Spine plot representing the distribution of the SCEs of the 6 chosen ACEs by M-B FCA clusters. The widths of the bars are proportional to the number of SCEs in that ACE. Left: AMPD data. Right: MDT data. The white cluster in the MDT spine plot collects SCEs where less than 12 hourly intervals are different from 0 (cluster 0).

major roads that connect the city center to the northwest Province of Brescia. High values of TRA are also found for the ACEs bordering Roncadelle in correspondence to the route of the "Serenissima" and "BreBeMi" highways and of the city's west ring road.

The main characteristics of the 6 ACEs are plotted in Figure 11, where the indicators CRO and TRA are reported along with the indicators STR, ROA and ULR defined in Eq.s (1), (2), and (3) respectively. The spider plots report the mean, the 10th, and the 90th percentiles of the SCEs' variables in the considered ACEs. Note that the scale of the plot is normalized with respect to the maximum and minimum of each ax. A few differences between the six ACEs emerge. In particular, Brescia NW is the most crowded ACE, the network of roads is dense, and the ACE is very busy. Roncadelle, which is adjacent to Brescia NW, appears the busiest ACE, and it is also considerably crowded. Overall, the road network appears quite similar to Brescia NW, although fewer roadways constitute the streets of Roncadelle. Gussago, Cellatica, and Castegnato show similar characteristics for all the indicators: the road network is dense but, the areas are neither very crowded nor particularly busy. Lastly, Rodengo Saiano is associated with low values of the indicators TRA and CRO and has a moderate road network.

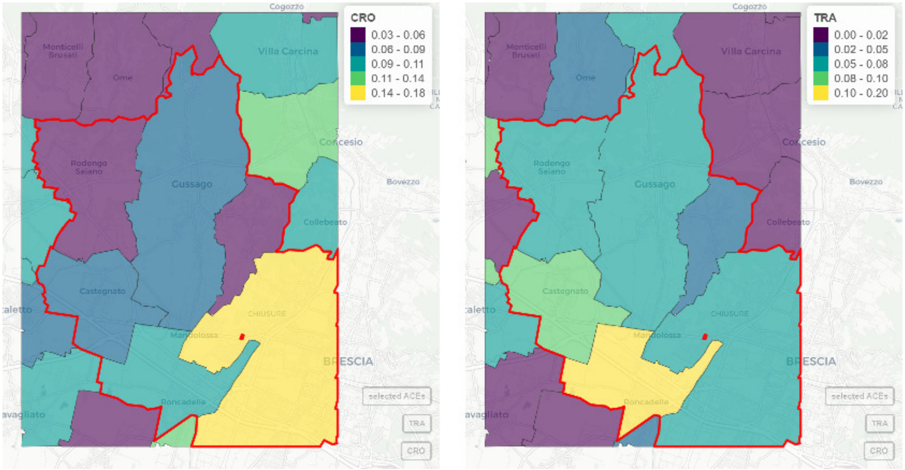

Fig. 10: ACE scores for the CRO (left) and TRA (right) composite indicators from T-mode PCA. Values represent the average normalized score of the SCEs in the ACE. The red line identifies the six main ACEs of the area under analysis.

7 Concluding remarks

The adoption of new sources of big data coming from sensors and tracking systems in conjunction with traditional data is important for the correct quantification of the exposure to risks at the small area level. In this work, we showed how the combination of the information from five different sources (i.e., DBTR, DUSAF, MPD, MDT, and the map of the administrative boundaries) may provide a multifaceted representation of the exposure to risks in small areas, based on the street network, the city user crowding, and traffic on the territory.

We used MPD and MDT data to estimate people crowding and traffic intensity respectively. The temporal dynamics of the MPD and MDT data were explored using the M-B FCA, while their spatial dynamics were analyzed by means of the T-Mode PCA. Then, we proposed five indicators: three on the static characteristics of the viability of the area - STR, ROA, and ULR - and two on the dynamic patterns of people crowding and traffic - CRO and TRA. The indicators might be used to identify the areas with a high concentration of people or major connecting routes. This information can be extremely important in various practical contexts, for example for the definition of traffic risk indicators on which insurers can rely for vehicle liability insurance. Moreover, the indicators can also be used to obtain a multi-dimensional description of the territory at different levels, for example from SCEs to ACEs (municipalities), as shown in Section 6. In this respect, we proposed an application of the indicators to the Mandolossa, a flood-risk area in the northwest of the city of Brescia. Overall, the joint analysis of the different data sources has thus

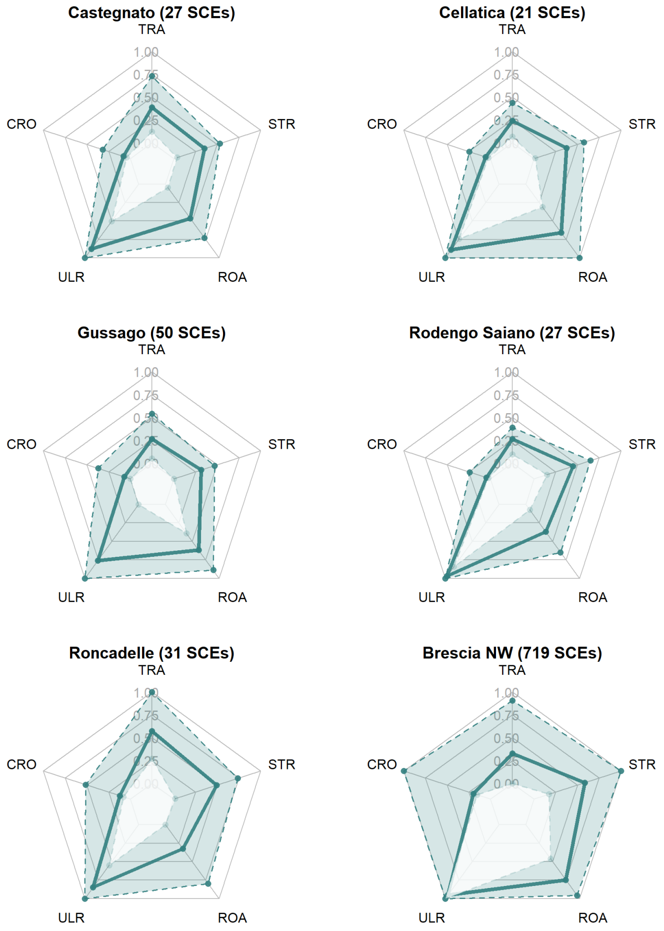

Fig. 11: The five indicators measured on the six major ACEs of the Mandolossa area. The green continuous line represents the indicator's mean value in the ACE SCEs, and the green dotted lines report the 10th and 90th percentiles. The scale of the plot is normalized with respect to the maximum and minimum of each axis.

allowed us to grasp the main characteristics of the small areas that make up the territory. The analysis can potentially be extended to other areas of the Lombardy Region as well as of the national territory. To this aim, a major problem emerges in the availability of MDT data. The MDT data, unlike the MPD one, is highly accurate in georeferencing but is not systematically detected and requires specific technologies and methods for the collection of signals from the mobile phone network.

In this work, MPD and MDT data were analyzed separately. We exploited the different characteristics of these data to capture different phenomena. A possible future development concerns the joint analysis of the two types of data in order to understand the relationship between road traffic and crowding in small urban areas.

Declarations

Funding. This study has been developed for the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project "Sustainable Mobility Center (MOST)" 2022-2026, CUP D83C22000690001, Spoke N 7 "CCAM, Connected networks and Smart Infrastructures".

Conflict of interest/Competing interests. Authors declare no conflict of interest.

Availability of data and materials. Street maps data are available at the "Geoportale della Lombardia" website (https://www.geoportale.regione.lombardia.it). The mobile phone data used in this study are subject to commercial restrictions.

References

[1] Kron, W.: Keynote lecture: Flood risk= hazard× exposure× vulnerability. Flood defence, 82–97 (2002)

[2] Perazzini, S., Gnecco, G.S., Pammolli, F.: A public-private insurance model for disaster risk management: an application to italy. Italian Economic Journal, 1–43 (Online first, 2022)

[3] Borgoni, R., Gilardi, A., Zappa, D.: Assessing the risk of car crashes in road networks. Social Indicators Research **156**(2), 429–447 (2021)

[4] Torti, A., Arena, M., Azzone, G., Secchi, P., Vantini, S.: Bridge closure in the road network of lombardy: a spatio-temporal analysis of the socio-economic impacts. Statistical Methods & Applications, 1–23 (2022)

[5] Ghosh, M., Rao, J.: Small area estimation: an appraisal. Statistical science **9**(1), 55–76 (1994)

[6] 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)

[7] Kong, X., Yang, J., Qiu, J., Zhang, Q., Chen, X., Wang, M., Jiang, S.: Post-event flood mapping for road networks using taxi gps data. Journal of Flood Risk Management **15**(2) (2022)

[8] Lin, Y., Shen, Z.: An innovative index for evaluating urban vulnerability on pandemic using lambdamart algorithm. Sustainability **14**(9), 1–19 (2022)

[9] Saghapour, T., Giles-Corti, B., Jafari, A., Qaisrani, M.A., Turrell, G.: Supporting pandemic disease preparedness: Development of a composite index of area vulnerability. Health & place **70**, 102629 (2021)

[10] Wang, Y., Pirani, M., Hansell, A.L., Richardson, S., Blangiardo, M.: Using ecological propensity score to adjust for missing confounders in small area studies. Biostatistics **20**(1), 1–16 (2019)

[11] 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)

[12] Carpita, M., 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)

[13] Mariotti, I., Giavarini, V., Rossi, F., 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)

[14] 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 **2**, 57–71 (2022)

[15] 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 **2**, 9–23 (2022)

[16] Metulini, R., 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, 1–26 (2023)

[17] Metulini, R., 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)

[18] Micheli, D., Diamanti, R.: Statistical analysis of interference in a real lte access network by massive collection of mdt radio measurement data from smartphones. In: 2019 PhotonIcs & Electromagnetics Research Symposium-Spring (PIERS-Spring), pp. 1906–1916 (2019). IEEE

[19] Baumann, D.: Minimization of drive tests (mdt) in mobile communication networks. In: Proceeding zum Seminar Future Internet (FI) und Innovative Internet Technologien und Mobilkommunikation (IITM), vol. 9, p. 7 (2014)

[20] Compagnucci, R., Richman, M.: Can principal component analysis provide atmospheric circulation or teleconnection patterns? International Journal of Climatology - INT J CLIMATOL **28**, 703–726 (2008)

[21] Caceres, N., Romero, L.M., Benitez, F.G., del Castillo, J.M.: Traffic flow estimation models using cellular phone data. IEEE Transactions on Intelligent Transportation Systems **13**(3), 1430–1441 (2012)

[22] Gilardi, A., Borgoni, R., Mateu, J.: Spatial statistical calibration on linear networks: an application to the analysis of traffic volumes. METMA X, 103 (2022)

[23] Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: Route classification using cellular handoff patterns. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 123–132 (2011)

[24] Calabrese, F., Ferrari, L., Blondel, V.D.: Urban sensing using mobile phone network data: a survey of research. Acm computing surveys (csur) **47**(2), 1–20 (2014)

[25] Zanini, P., Shen, H., Truong, Y.: Understanding resident mobility in milan through independent component analysis of telecom italia mobile usage data. The Annals of Applied Statistics **10**(2), 812–833 (2016)

[26] Secchi, P., Vantini, S., Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of milan. Statistical Methods & Applications **24**(2), 279–300 (2015)

[27] Scaloni, A.: Minimization of drive test (mdt) an innovative methodology for measuring customer performance on mobile network. In: The GeoSynthesis Project"", ITU Workshop On" Benchmarking of Emerging Technologies and Applications. Internet Related Performance Measurements" Geneva, Switzerland (2019)

[28] Jacques, J., Preda, C.: Functional data clustering: A survey. Advances in Data Analysis and Classification **8**, 231–255 (2013). https://doi.org/10.1007/s11634-013-0158-y

[29] Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis **71**, 92–106 (2014)

[30] Bouveyron, C., Côme, E., Jacques, J.: The discriminative functional mixture model for a comparative analysis of bike sharing systems. The Annals of Applied Statistics **9**(4), 1726–1760 (2015)

[31] Bouveyron, C., Brunet, C.: Simultaneous model-based clustering and visualization in the fisher discriminative subspace. arXiv preprint arXiv:1101.2374 (2011)

[32] Bouveyron, C., Celeux, G., Murphy, T., Raftery, A.: Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2019). https://doi.org/10.1017/9781108644181

[33] Pollock, D.S.G.: Topics in time-series analysis the fourier decomposition of a time series. lecture notes. https://www.le.ac.uk/users/dsgp1/COURSES/LEIMETZ/FOURIER.pdf. (2011)

[34] Preisendorfer, R.W., Zwiers, F.W., Barnett, T.P.: Foundations of Principal Component Selections Rules. 81-4, vol. 192. Scripps Institute of Oceanography., ??? (1981)

[35] Barreira, S., Compagnucci, R.: Spatial fields of antarctic sea-ice concentration anomalies for summer - autumn and their relationship to southern hemisphere atmospheric circulation during the period 1979-2009. Annals of Glaciology **52**, 140–150 (2011)

[36] Isaak, D., Luce, C., Chandler, G., Horan, D., Wollrab, S.: Principal components of thermal regimes in mountain river networks. Hydrology and Earth System Sciences **22**, 6225–6240 (2018)

[37] Ibebuchi, C.C.: Patterns of atmospheric circulation in western europe linked to heavy rainfall in germany: preliminary analysis into the 2021 heavy rainfall episode. Theoretical and Applied Climatology **148**, 269–283 (2022)

[38] Richman, M.: Review article, rotation of principal components. J. Climatol. **6**, 293–355 (1986)

[39] Kaiser, H.F., Rice, J.: Little jiffy, mark iv. Educational and psychological measurement **34**(1), 111–117 (1974)

[40] Vieira, V.: Permutation tests to estimate significances on principal components analysis. Computational Ecology and Software **2**, 103–123 (2012)

[41] Gleason, T., Staelin, R.: A proposal for handling missing data. Psychometrika **40**, 229–252 (1975)

[42] ter Braak, C.: CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1), vol. 11. (1987)