Socio-Economic Planning Sciences 90 (2023) 101747

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

Integration of flows and signals data from mobile phone network for statistical analyses of traffic in a flooding risk area

Selene Perazzini [a,*], Rodolfo Metulini [b], Maurizio Carpita [a]

[a] DMS StatLab, Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, Brescia, 25122, Italy
[b] Department of Economics, University of Bergamo, Via Caniana, 2, Bergamo, 24127, Italy

ARTICLE INFO

Keywords:
Vector autoregressive model
Dynamic harmonic regression
Origin–destination data
Minimization of drive test data
United Nations SDGs

ABSTRACT

In this paper, we present a robust spatiotemporal statistical methodology that is capable of accurately forecasting traffic in the flood-prone area of the Mandolossa in the Province of Brescia (Italy). An innovative combination of two sources of mobile phone data is proposed to obtain an extremely accurate representation of the flows of people passing by the streets directly linked to the risky area. Three types of flows have been considered: outflows (from the flood-prone area to the neighborhood), inflows (from the neighborhood to the flood-prone area), and internal flows (within the flood-prone area). The three flows are assumed to be dependent on each other and are modeled using a vector autoregressive approach. We found evidence of both weekly and daily seasonal components in the time series. To capture the seasonality, a dynamic harmonic regression component has been included, where the optimal number of Fourier bases in the periodic functions has been chosen according to a criterion based on the Akaike Information Criteria. On the other side, the set of autoregressive parameters has been defined in such a way as to represent the time period necessary for the mobile phone company to observe, process, and release the data. The forecasting ability of the model has been assessed using blocked k-folds cross-validation along with the mean absolute percentage error and the hit rate. Though the model performs better for non-summer days, we found that it satisfactorily forecasts both the number and the level of people moving.

1. Introduction

Monitoring and forecasting people's mobility in metropolitan areas is crucial for several aspects of life in smart cities and is therefore gaining increasing interest in literature [1]. In this paper, we propose a robust spatiotemporal statistical methodology that is capable of accurately estimating and forecasting traffic in the flood-prone area on the western outskirt of the city of Brescia (Italy). In particular, we contribute to the literature by proposing an innovative approach of combining different types of mobile phone data to obtain small-area estimates of the flows of people. This study considers themes that are very important today, such as the monitoring-optimization of traffic networks and the smart infrastructures for sustainable mobility (Mission 3 of the Italian NRRP, National Recovery and Resilience Plan), which is part of the Next Generation EU (NGEU) Programme. These themes are also very connected to three United Nations SDGs (Sustainable development goals): 9—Industry, innovation, and infrastructure; 11—Sustainable cities and communities; 13—Climate action. This work aims at supporting regulators by providing information useful to manage potentially harmful situations and avoid human losses. Indeed, natural disasters have huge socio-economic consequences and their immediate effects on fragile environments include loss of human lives, as well as severe impacts on health and wealth.

Traditional data sources, such as censuses, are often inadequate for the study of people's dynamics due to their static nature, slow update, and excessive costs. Moreover, smart cities present emerging forms of mobility and time variations in the use of urban spaces. In contrast, mobile phone data allow for a dynamic and fine-grained representation of human activities and are becoming increasingly essential to the analysis of social, economic, and environmental phenomena in urban areas. Information and Communication Technologies (ICT) have been widely adopted for the analysis of smart cities and urban systems [2] to improve the well-being and quality of life. The analytical processing of ICT data can be used, for example, in supporting the optimization of traffic flows or tracking real-time citizens' positions. Reinolsmann et al. [3] use big data collected by ICT and multivariate statistical methods useful to implement real-time and strategic traffic management solutions. They support Advanced Traveller Information Systems that can be integrated with Advanced Driving Assistance Systems at

* Corresponding author.
E-mail addresses: selene.perazzini@unibs.it (S. Perazzini), rodolfo.metulini@unibg.it (R. Metulini), maurizio.carpita@unibs.it (M. Carpita).

https://doi.org/10.1016/j.seps.2023.101747
Received 18 January 2023; Received in revised form 12 September 2023; Accepted 24 October 2023
Available online 27 October 2023

the local level. Nowadays, modern sources of mobile phone data are increasingly combined with satellite and sensor technologies (e.g., [4]) to produce dynamic information related to the density of people's presences [5] and movements [6]. This approach allows investigating issues of great relevance, such as the monitoring of the impact of social and cultural events [7], the variability in the distribution of presences in the neighborhoods of a large city [8], the seasonality of the second homes in a tourist area [9], or the increase of remote working in sparsely populated areas [10]. The use of mobile phone data is not new to the natural risk literature as well. For example, Balistrocchi et al. [11] developed spatiotemporal dynamic maps of flood exposure using mobile phone data about crowding.

Here, we focus our attention on the estimation of traffic intended as the number of people moving and propose an innovative combination of two sources of mobile phone data to obtain an accurate representation of people's movements, namely the Origin–Destination (OD hereafter) and the "Minimization of Drive Test" (MDT hereafter) data. The former type of data has already been used in various statistical applications in the literature, while the MDT data are collected using new technology. It is worth noting that, at the current moment, MDT data have not yet found many applications beyond network engineering, and their statistical application in natural risk management constitutes a further novelty of our approach. Specifically, OD data represent the flows of people from an "Area di CEnsimento" (ACE hereafter) to another from September 2020 to August 2021 at hourly intervals, while the MDT data at our disposal represents the location of phone users in a 15-minute interval with high accuracy (approximately 10 m) and refer to 5 days of November 2021. As we will discuss in detail in Section 2.2, the MDT data only represents a sample of TIM users, but the accurate geolocation allows us to identify the devices that are located on the streets by putting the MDT signals on a street map. In this respect, we refer to the methodology developed by Perazzini et al. [12], where the MDT data has been used to infer the traffic intensity in small areas, and we extend it to the estimation of the proportion of traffic flows related to the flood-prone area in an ACE. Then, the resulting ratios are applied to the OD data as weights, in such a way as to obtain the traffic flows at risk. The combination of these two pieces of information allows us to produce statistical estimation and forecast of traffic flows for short time intervals at the "small area" level.

In the proposed analysis, we apply the modeling strategy presented by Metulini and Carpita [13], who developed a vector autoregressive model with exogenous variables to analyze OD data and forecast traffic flows between specific ACEs subjected to flood risk. In that work, the authors investigated Cellatica (one of the ACEs located in the Mandolossa flood risk area), which is the ACE with the largest intersection with the flood-prone area of the Province of Brescia, and the flows of people that link it to the neighboring ACEs. Here, we consider all the ACEs among which the flood-risk area spans, and we limit our attention to the flows passing by the streets directly linked to the risky area. This allows us to considerably improve the representation of traffic flows at risk. It is worth noticing that, without the methodological strategy proposed in this work, it would not be possible to restrict attention to the flood-prone area and estimate or forecast the flows at risk. The OD data are available for fairly large areas, i.e. the ACEs, and do not allow for estimating traffic flows in a small area exposed to flood risk that expands over limited portions of 4 ACEs. Therefore, the advantage of the proposed methodology consists of a considerable improvement in the degree of geographical resolution of the estimates.

Particular attention has been dedicated to modeling the seasonal component of the traffic flows time series that show a consistent daily trend with less pronounced daily peaks during holidays and weekends. Many applications related to the modeling of traffic with complex seasonality using mobile data can be found in the literature (e.g., 14–17). Here, we refer to [13] and model seasonality through a dynamic harmonic regression model [18] with daily and weekly periodic functions modeled using a proper combination of Fourier bases. We identify the optimal number of Fourier bases in the periodic functions by minimizing the Akaike Information Criteria. Moreover, the autoregressive components of the model are chosen in such a way as to represent the time necessary for the mobile phone company to observe, process, and release the data. This choice allows the model to be adopted by regulators in the Province of Brescia for monitoring traffic in the flood-prone area and to promptly activate actions for traffic control in case of danger. Notably, the chosen methodology is inherently data-driven, and none of the existing alternative methods for analyzing traffic time series correspond directly to it. The most analogous approaches pertain to a subset of methodological strategies that are rooted in autoregressive models tailored for the scrutiny of time series data. For instance, Guo et al. [19] introduced a strategy employing multiplicative seasonal autoregressive integrated moving averages. Pursuing the same objective, Tran et al. [20,21] embraced a strategy constructed upon the underpinnings of generalized autoregressive conditional heteroscedasticity. These studies lean on ARIMA models, which find limited applicability to our specific case study due to the constraints imposed by the availability of OD data. Additionally, the diagnostic assessments of individual time series lend support to the adoption of an ARMA model with a negligible MA component, consequently resulting in its exclusion from our model. Similarly, the incorporation of the DHR arises from the need to capture complex seasonality patterns that prominently manifest within our dataset.

Since the utility of our model pertains to its adoption by the local authorities under the circumstances of critical meteorological forecasts, we focus on short-term forecasting. Indeed, meteorological forecasts, particularly those pertaining to significant occurrences such as flooding, exhibit a notable limitation in terms of both their availability and reliability, extending merely over a temporal horizon ranging from a few days to a matter of hours. Consequently, our research is oriented towards an investigation within this temporal constraint. We assess the forecasting ability of the model using a blocked k-folds cross-validation strategy, which preserves the temporal structure of our dataset. The forecasting accuracy is evaluated using the symmetric mean absolute percentage error and the hit rate. Though the model performs better for non-summer days, we find that it satisfactorily forecasts the number of people moving and achieves from good to excellent performance (depending on the period of the year) in forecasting the level of people moving (i.e., high, moderate, low), to which we refer as traffic intensity for simplicity.

The paper organizes as follows. Section 2 presents the data used. Section 3 describes the weighting strategy. Section 4 applies the weights and shows some preliminary evidence on the traffic flow time series. Section 5 describes the model. Section 6 presents results and discusses the model's forecasting performance. Section 7 concludes.

2. Data

In this work, different sources of data have been combined. In particular, we used two types of mobile phone data: OD and MDT data, provided by the TIM company, which is currently the largest operator in Italy. The OD data represent the flows of people from one ACE to another between September 2020 and August 2021 on an hourly basis. The MDT data provides the accurate location of users in 15 min during 5 days of November 2021. As one can notice, the two databases cover slightly different periods of observation. Indeed, MDT data require particular technologies to be activated and tested in loco before data collection. For this reason, the data collection process is costly and takes time, and the produced MDT datasets typically cover short periods. To overcome this issue and preserve the representativeness of the analysis, days for MDT data collection have been carefully chosen in such a way as to represent a typical week. Moreover, they have been used to infer the proportions of users on streets linked to the flood-prone area and applied to the OD data as weights.

Table 1
Summary statistics of the OD data. Inflows from the other 3 + 38 ACEs (in), outflows to the other 3 + 38 ACES (out), and internal flows (int) are reported for the 4 ACEs exposed to flood risk (Cellatica, Gussago, Rodengo Saiano, Brescia North-West). Reported values are from first to last row: minimum, 1st quartile, median, mean, 3rd quartile, and maximum.

	Cellatica			Gussago			Rodengo Saiano			Brescia NW		
	out	in	int	out	in	int	out	in	int	out	in	int
Min.	0	0	29	0	0	157	0	0	73	0	0	63
1st Qu.	1	1	194	3	3	672	1	1	324	4	4	458
Median	3	3	625	10	11	2218	4	5	1134	11	11	1435
Mean	34	34	578	94	94	2036	51	51	1012	74	74	1359
3rd Qu.	9	9	894	34	34	3205	20	20	1575	36	36	2169
Max.	1731	1731	1731	6180	6180	6180	2812	2812	2812	3837	3837	3837

Three additional sources of information were needed for the analysis, especially for the definition of the weighting system based on MDT data: the map of the administrative boundaries of the Province of Brescia, the flood map, and a street map. The following subsections describe, in order, the OD, MDT, and other open-source data.

2.1. Origin–destination flow data

The main source of data used for this analysis is the mobile phone OD flows. The database reports one year of observations (from September 1st, 2020 to August 31st, 2021) in the ACEs of the province of Brescia. The data can be used to represent the traffic flows $flow_{ij,t}$ from the ACE i to the ACE j in the tth hour of a day. Specifically, the OD data refers to the number of phone SIM cards that were retrieved during a 1-hour interval by the antenna in a given ACE i and, after five or more minutes, by the antenna in ACE j. It is worth noticing that the position of the SIM cards is retrieved at regular intervals of 5 min - i.e., at minutes [00–05),[05–10), … [55,60) - and only the first arrival location registered during the 5-minutes interval is considered. Let us consider, for instance, the one-hour time interval t corresponding to 8:00–8:59 AM on January 1st, 2021. Suppose that a SIM card arrives in i and then moves to j between 8:00 and 8:04 AM of that day, and then arrives at z between 8:05–8:09 AM of the same day. The SIM card is counted in $flow_{iz,t}$, but is not taken into account in $flow_{ij,t}$ nor in $flow_{jz,t}$. This can lead to a potential underestimation of the flows in small ACEs requiring less than 5 min to be crossed.

For each time interval t, the database collects the flows in a square (non-symmetric) matrix of dimension $N \times N$, where N is the number of ACEs in the province of Brescia and is equal to 235, rows represent the ACE of departure and columns the ACE of arrival. Note that the diagonal represents the internal flows, which are the flows departing from and arriving in ACE i. For each time t and each ACE i, three types of flows can therefore be distinguished: flows arriving in i, flows departing from i, and internal flows from i to i. Overall, the database is constituted by $24 \times 365 = 8,760$ square matrices of dimension 235 x 235, each representing a time interval t.

The database refers to SIM cards connected to the TIM network, including the foreigner SIM cards connected to the roaming. Two types of cards can be distinguished: human SIM (about 85% of the total SIM) and M2M technology machine SIM (about 15%). Since a user might have both a human SIM and some devices with an M2M machine SIM, we restricted our attention to human SIMs to avoid double counting of users. As we will discuss in Section 2.3, we focus our attention on the flows from/to 4 ACEs exposed to flood risk to/from 38 selected neighboring ACEs that display a limited amount of zero flows. Table 1 shows the summary statistics of these flow data.

2.2. MDT signal data

MDT signal data have been collected using a recent technology called the "Minimization of Drive Test". This new technology allows for considerably high accuracy (approximately 10 m) in users' geolocation. The MDT data has been only recently made available by TIM and for the moment has found only a few applications in the academic literature, mostly for technical control of telephone networks in the field of network engineering. In [12], we proposed an innovative application of MDT data for the definition of traffic indicators. In this paper, we extend our previous work by estimating the portion of phone users on streets that are located in or that are directly connected to the flood-prone area.

The MDT technology registers geo-referenced radio measurements of signals transmitted over the 3G/4G mobile network from/to terminal devices with GPS enabled. Each signal corresponds to either a phone call, a text message, internet browsing, or a technical operation on the network (e.g., location update). The MDT data tracking is currently not part of the standard data collections of mobile phone operators. Therefore, the signals are collected ad hoc during specific campaigns organized over a pre-selected area and period of time. The MDT data at our disposal refers to devices with the SIM card associated with the TIM company in a rectangular area of approximately 150 km^2 around the Mandolossa region during 5 days of November 2021—namely Wednesday 10, Friday 19, Saturday 20, Sunday 21, Monday 22. The days of observation were carefully chosen to represent a typical week. The detection of MDT signals requires particular technologies to be activated, and the collection of such data is time-consuming and expensive. So, the first day – Wednesday 10 – was sampled, and, once assessed the adequacy of the data to the analysis, due to budget limit, other 4 days of detection were chosen to characterize the temporal dynamics in the whole typical week. The choice, therefore, fell primarily on the two days of the weekend due to the particular dynamics associated with the closure of work and school activities. Then, the remaining 3 days refer to the beginning of the week, the central part of the working days, and the pre-weekend (i.e., Monday, Wednesday, Friday) in such a way as to characterize the dynamic during the working days.

For each day, the database reports 96 times of observation corresponding 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). Ten times of observation are missing and have been replaced by the average value of the other intervals of the same hour of the day, namely: 04:30–04:44 in the five days; 23:30–23:44, 23:45–23:59 for Monday and Wednesday; 00:00–00:14 on Friday. For each time interval, the MDT technology registered signals and assigned them to the cells of a grid of pixels measuring 10 meters on each side and identified by a pair of longitude-latitude coordinates. The database reports the total number of signals sent or received during a 15 min-interval in a cell of the grid where at least one signal has been generated. As an example, a time interval of the database is shown in Fig. 1. Overall, the database reports signals for 274,005 cells of the grid and 470 time intervals.

A few aspects should be carefully evaluated when analyzing MDT data. First, a device can receive or send multiple MDT signals at a time. As shown in the first column of Table 2, the number of MDT signals registered in 15 min in a pixel (i.e., a 100 m^2 area) varies considerably and may also take extremely high values. Therefore, the number of MDT signals cannot be used to represent the number of individuals in the considered pixel. This is particularly evident if we consider that only about 10% of current electronic devices produce MDT signals and

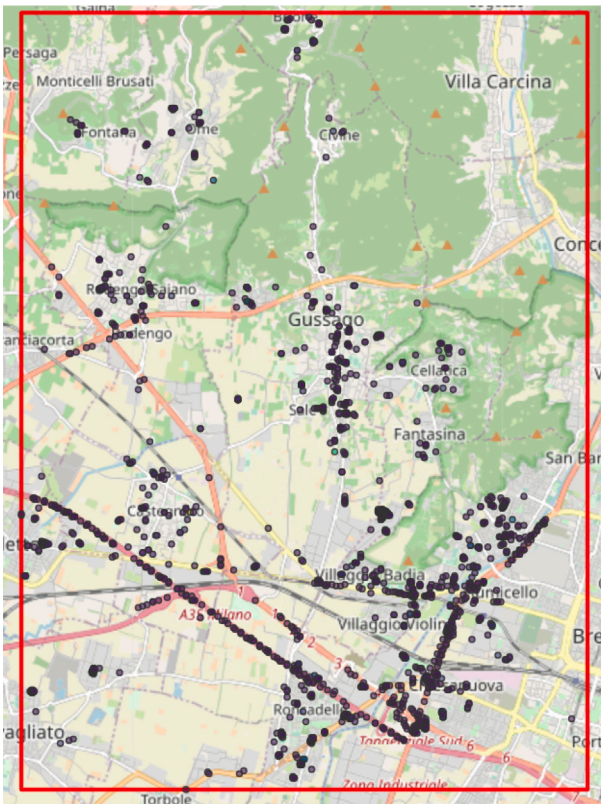

Fig. 1. MDT data signals for the time interval 00:00-00:14 AM of November 10th, 2021. The red line represents the border of the rectangular area captured by TIM. The points represent the cells of the pixel grid for which at least one MDT data signal was detected in the time interval.

therefore the database represents a sample of a much wider population. To overcome this issue, we instead consider the number of cells of the pixel grid from which signals originated in an area. As shown in the second column of Table 2, we find that these values better represent the dimension of the observed sample of individuals in the ACEs intersecting the flood-prone area. This hypothesis is particularly reasonable in our framework of analysis because, as we will show in Section 3, MDT data signals are used as a proxy for street users to capture the flows that are critical for flood emergency management. For this purpose, we restrict our attention to people located on the streets who are likely to move and not stay in a cell for a long time. The limited sample size makes it very unlikely for two devices to simultaneously be in the same small street cell of the grid. Therefore, though a signal cannot be traced back to the device from which it has been generated, we conclude that cells approximate the number of users in the 15-minute interval better than the number of MDT signals.

2.3. Administrative boundaries, flood maps, and street maps

A necessary preliminary step for the analysis is the selection of the area of interest. We considered the flood risk map with time to return equal to 20 years and identified the ACEs in the flood-prone area by overlaying the risk map to the map of administrative boundaries released by ISTAT and available at https://www.istat.it/it/archivio/104317. We found that the flood-prone area is located in the Mandolossa region and extends through 4 ACEs: Gussago, Cellatica, Rodengo Saiano, Brescia Mandolossa (which has been obtained by merging two ACEs in the north-west of Brescia). In addition, Metulini and Carpita [13] identifies other 38 ACEs in the neighborhood that account for 84% of the total flows from/to the four ACEs of interest. We restrict our attention to these two groups of ACEs, which we consider

Table 2
Summary statistics of the MDT data. The first column refers to the number of signals in a 15-minute interval in a pixel. Note that only pixels from which at least one MDT signal originated in the 15-minute interval. The second column refers to the number of pixels from which at least one MDT signal originated during a 15-minute interval in an ACE that intersects the flood-prone area. Reported values are from first to last row: minimum, 1st quartile, median, mean, 3rd quartile, and maximum.

	Number of signals in a pixel	Number of pixels originating MDT signals in an ACE
Min.	1	1
1st Qu.	2	63
Median	4	170
Mean	16	233
3rd Qu.	10	392
Max.	4430	895

separately. The two groups of selected ACEs are shown in Fig. 2: the one intersecting the flood-prone area is represented by the green polygon, and the one composed of the 38 neighboring ACEs is reported in light blue. The Figure also illustrates the different flows of people analyzed in the present work. Indeed, three types of flows can be distinguished between the two macro-areas: outflows (from the flood-prone ACEs to the neighboring ACEs), inflows (from the neighborhood to the flood-prone ACEs), and internal flows (between the ACEs in the flood-prone area). It is worth noticing that the latter type of flow is constituted by both flows of people moving from one ACE of the Mandolossa to another (e.g., from Gussago to Cellatica) and flows of people moving within an ACE (e.g., from Gussago to Gussago). Lastly, the identified area subjected to flood risk (map with time to return equal to 20 years) is depicted in blue.

The last data source needed for the proposed analysis is a detailed road network map. The MDT data, as we will discuss in Section 3, was used to estimate the portion of traffic flows in the flood-risk area per each of the four ACEs. To do so, we first have to identify the MDT data signals that originated from streets and this can be done by comparing the MDT pixel grid with the street map of the area. We refer to the street map defined in [12], which is obtained by merging two maps released by the Lombardy Region for the Province of Brescia: the "DataBase Topografico Regionale"[1] and the "Uso e copertura del suolo della Regione Lombardia 2018". This map has been chosen for two main reasons. First, it does not represent roads as lines but as polygons and shows the width of the roadway. Indeed, this allows for identifying the cells of the MDT grid that correspond to streets. Second, it provides a comprehensive representation of the road network that none of the other official polygon-based maps available for the Province achieves.

3. Definition of the weighting strategy

The OD data capture the flows of people in the four ACEs of the Mandolossa region, but only a portion of the flows concerns the flood-prone area. To capture the flows that are critical for flood emergency management, we computed the ratios of users on the streets of an ACE that are located in the flood risk area. The ratios have then been applied to the OD database as weights and flows $flow_{ij,t}$ have been adjusted accordingly. So, to restrict the analysis to traffic flows potentially exposed to floods, we compute the ratio of phone users on streets that pass by the flood-prone area for each of the 4 ACEs of interest and for the aggregated area constituted by them. To this scope, we consider the MDT data signals for the high accuracy in users' geolocation. The process that led to the construction of the weights is extensively described in Appendix, and can be briefly summarized as follows:

[1] We refer to the version updated in 2021.

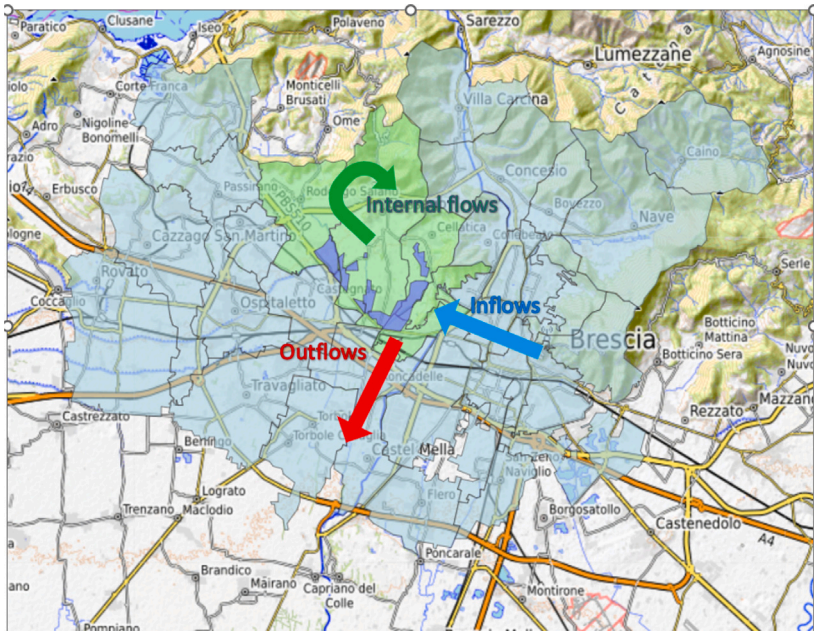

Fig. 2. Map of the ACEs of the Mandolossa (light green) and the neighboring ACEs (light blue). The arrows represent the three flows investigated, the blue polygon reports the flood risk map at 20 years' time to return.

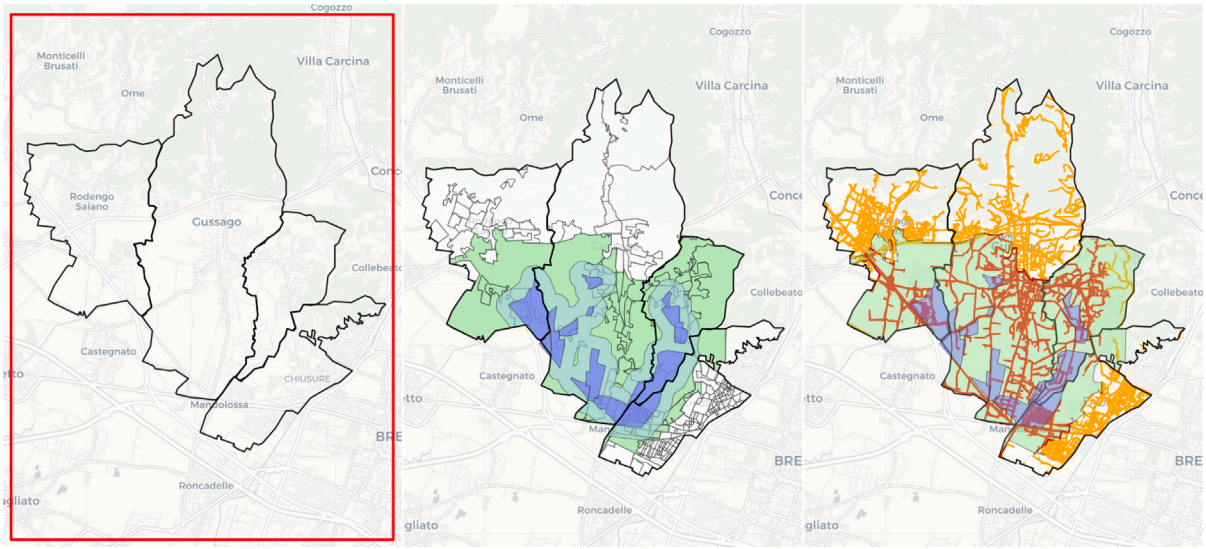

Fig. 3. Weights construction, steps 1–3. **Left**: Step 1—Selection of the MDT data corresponding to the four ACEs of interest. The black lines represent the area of the four ACEs, and the red border indicates the area captured by TIM. **Center**: Step 2—Identification of the area critical for flood emergency management. The map reports the four ACEs (thick black borders), the SCEs that constitute them (fine black borders), the flood risk map (blue), the 500 meters buffer applied to the flood risk map (light blue), and the critical SCEs (green). **Right**: Step 3—Street map and identification of the streets critical for flood emergency management. The map represents the four ACEs (black borders), the flood risk map (blue), the critical area (green), the map of the streets (orange), and the selection of streets that are considered critical for flood risk management (red).

1. The MDT data signals have been restricted to the phone signals in the four ACEs of interest (left map of Fig. 3).
2. 104 "Sezioni di CEnsimento" (SCEs hereafter) – which are subdivisions of the ACEs – that are less than 500 meters far from the flood-prone area were identified as critical for flood management (central map of Fig. 3).
3. We compared the MDT data signals to the street map and further restricted them to the phone signals that originated from streets. Then, we identified the streets passing by the SCEs critical for flood risk management (identified in step 2) that connect the risky area to the 38 neighboring ACEs (right map of Fig. 3).
4. For each ACE and each time of observation we counted the number of grid cells that generated MDT signals corresponding to streets (e.g., left map of Fig. 4).
5. Among the cells identified in step 5, we counted the number of those corresponding to the streets related to the flood-prone area for each ACE and each time of observation (e.g., right map of Fig. 4).

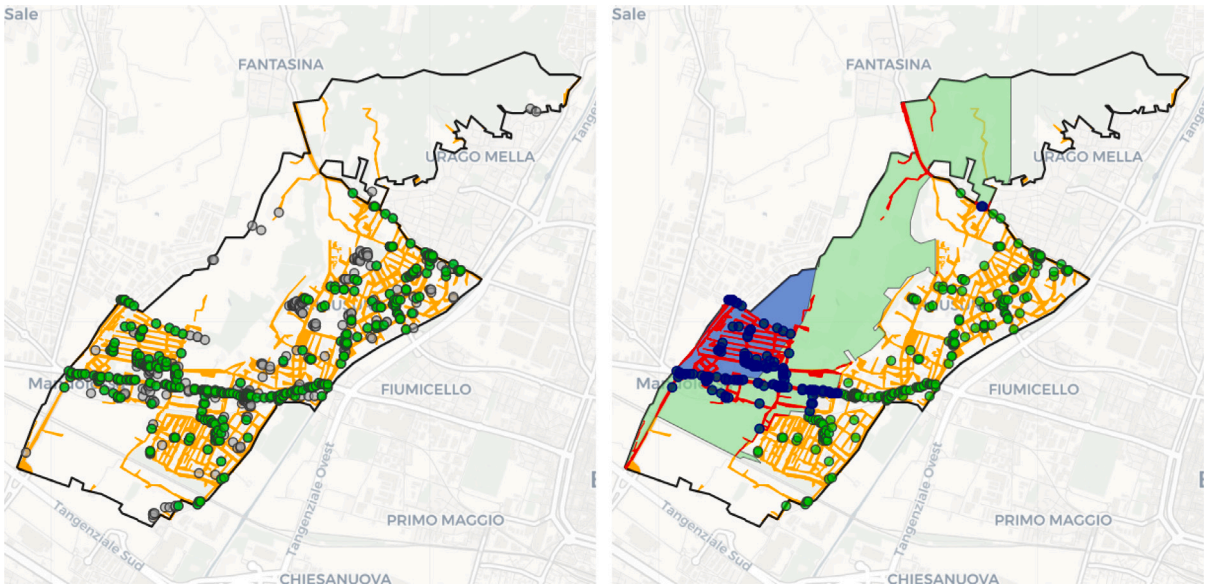

Fig. 4. Weights construction, steps 4–5. The maps represent the ACE "Brescia Mandolossa" (black border) and refer to the time interval corresponding to November 10th, 12:00–12:14 PM. **Left**: Step 4 - Identification of the MDT grid cells corresponding to streets. The map reports the street map of the ACE (orange) and the cells of the pixel grid in the ACE from which MDT signals originated in the time interval (points). The MDT cells are colored green if they correspond to streets, and gray if not. **Right**: Step 5 - Identification of the MDT cells corresponding to critical streets. The map reports the flood risk area (blue), the area critical for flood emergency management (green), the road network of the ACE (red and orange), and the on-streets-MDT cells from which signals originated during the time interval (points). Red polygons represent the streets that are critical for flood emergency management, while orange polygons report the rest of the streets. Blue points identify the cells that correspond to critical streets, and green points represent all the other on-street cells of the grid.

6. We found that the number of MDT cells that originated signals on streets is very low during the night and that the number can vary considerably during the day. Therefore, 6 intervals of 4 h have been considered, namely: 0:00–3:59, 4:00–7:59, 8:00–11:59, 12:00–15:59, 16:00–19:59, and 20:00–23:59.
7. The weights have been computed for each of the 6 time intervals as:

$$I.MDT_{it} = \frac{\text{MDT signals from streets connecting the flood-prone area}_{it}}{\text{MDT signals from streets}_{it}}, \tag{1}$$

where i indicates either Brescia Mandolossa, Cellatica, Gussago, Rodengo Saiano. Similarly, we computed the ratio for the aggregated area.

Fig. 5 reports the resulting ratio separately for the four ACEs of the Mandolossa and over different weekdays and times of the day. The ratios appear fairly constant among the intervals and the 5 days observed. This evidence suggests that the proportion of traffic of an ACE that passes by the flood-prone area is quite stable.[2] For this reason, we compute the weight for each ACE i $\overline{I.MDT}_i$ as the average value of the ratios of the 5 time intervals in the 5 days. The ratios $\overline{I.MDT}_i$ represent the percentage of phone users on the streets that are critical for flood management, and can therefore be interpreted as the portion of traffic of the ACE potentially exposed to floods. The indicator is equal to 20% for Brescia Mandolossa, 75% for Cellatica, 40% for Gussago, and 10% for Rodengo Saiano. In addition, the ratio for the aggregated area constituted by the 4 ACEs has been computed and the obtained value is $\overline{I.MDT}_{agg} = 30\%$.

[2] In this respect, it is worth noticing that the critical areas of the ACEs encompass a diverse array of road classifications (i.e., primary, secondary, local, etc...). Moreover, the flood-prone area exhibits a compositional homogeneity in land use akin to the remaining part of the ACEs, featuring commercial, residential, industrial settlements, and green areas. For this reason, ratios can be assumed constant among the different months of the year.

4. Data pre-processing and preliminary evidence

The MDT ratios computed in Section 3 have been applied to the OD data flows. To obtain the traffic flows in the flood-prone area at time t between i and j, where the ith ACE intersects the flood risk map of the Mandolossa and j is one of the other 38 neighboring ACEs, the weights were applied as follows:

$$\begin{aligned} Inflow_t &= \sum_i \left(\overline{I.MDT}_i \times \sum_j flow_{ij,t} \right), \\ Outflow_t &= \sum_i \left(\overline{I.MDT}_i \times \sum_j flow_{ji,t} \right). \end{aligned} \tag{2}$$

For the internal flow, the MDT ratio was applied to the sum of $flows_{ii,t}$ and $flows_{ij,t}$ where i and j are both ACEs intersecting the flood risk map of the Mandolossa:

$$Internal\,flow_t = \overline{I.MDT}_{agg} \times \sum_i \left(flow_{ii,t} + \sum_{j \neq i} flow_{ij,t} \right). \tag{3}$$

The time series of the obtained inflows, outflows, and internal flows have been analyzed. In general, we observe that all the flows increase up to a certain hour in the morning, remain constant until the mid-afternoon, and decrease in the evening. Some seasonal effects emerged. In particular, exploiting the findings in [5] about the clustering of days, we compared traffic dynamics between summer and non-summer days, and between weekends and weekdays. As shown in Fig. 6, the number of people moving is on average higher during the midweek than on the weekends. It could also be noted that traffic decreases at lunchtime on the weekends. Finally, it emerges that traffic is on average lower in summer. All those pieces of evidence emerge on all the considered flows (inflows, outflows, and internal flows).

Strong daily patterns emerge in the AutoCorrelation Function (ACF) and in the Partial ACF (PACF), which are shown in Fig. 7 for time lags up to 168 h (one week). The ACFs (top charts) clearly show a daily periodicity with picks of positive autocorrelation at lags 24, 48, 72, ..., and picks of negative autocorrelation at lags 12, 36, 60, ... The PACFs (bottom charts) show strong partial autocorrelation at lags 24, 48, 72,

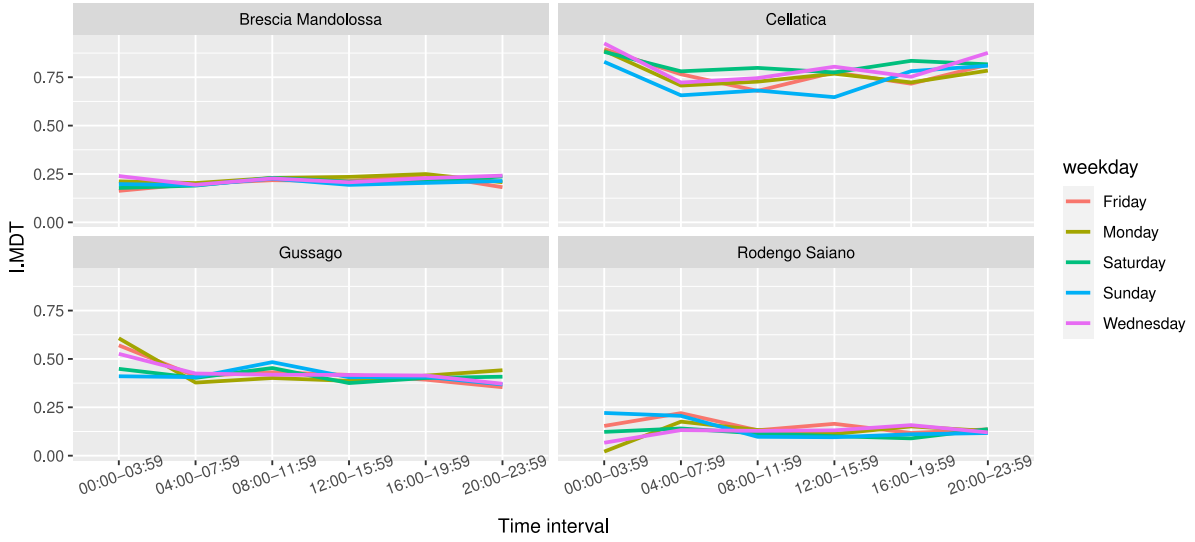

Fig. 5. Ratios $I.MDT_{it}$ in the four selected ACEs and over the six time intervals of 4 h length.

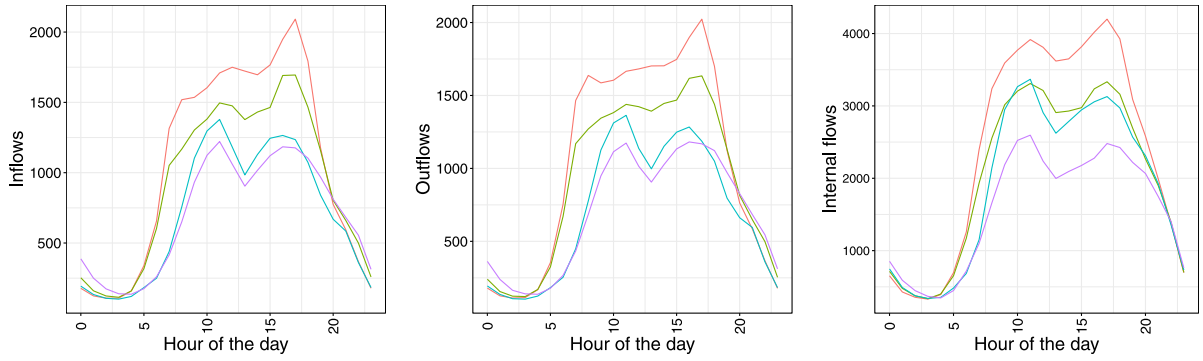

Fig. 6. Curves of daily time series of traffic flows, grouped by periods of the year (median values). Weekdays (in red), weekdays of summer (July and August, green), weekends (blue), and weekends of summer (July and August, purple). Inflows (left), outflows (middle), and internal flows (right).

.... The patterns appear very similar for the inflows, outflows, and internal flows. In addition, an additive decomposition of the time series using the Seasonal-Trend decomposition with LOcally Estimated Scatter plot Smoothing (ST with LOESS, or simply STL 22) has been performed. The original time series of inflows, outflows, and internal flows have been addictively decomposed in trend (`trend`), daily seasonal pattern (`season_24`), weekly seasonal pattern (`season_168`), and a residual term (`remainder`). For this analysis, we used the R packages for time series by Hyndman and Killick [23], and the results are shown in Fig. 8. It emerges that the three flows show consistent daily and weekly seasonality, though the former is stronger.

It is worth stating that anomalous traffic dynamics were observed during holidays, namely: January, 1st and 6th (Epiphany), April, 4th (Easter), April, 5th, April, 25th, May, 1st, June, 2nd, August, 15th (mid-August public holiday), December, 8th (Immaculate Conception), 25th (Christmas), 26th (S. Stefano) and 31th. Therefore, as a preliminary step, holidays have been replaced by the corresponding weekday of the previous week. If the latter corresponds to a holiday itself, the corresponding nearest previous non-holiday weekday is taken. For example, January 1st, 2021 was Friday, and the previous Friday was December, 25th 2020, which was a holiday as well. Therefore, December 18th was taken.

5. The model

To model the daily and weekly seasonality of the inflows, outflows, and internal flows of the Mandolossa flood-prone area and describe the dependence structure among the three, we refer to [13] and adopt a Vector AutoRegressive model with eXogenous variables (VARX hereafter) to capture the dependence within each flow and the interdependence among the three flows, combined with a Dynamic Harmonic Regression (DHR hereafter) model that captures the complex seasonality through a combination of Fourier bases. This approach is particularly suitable to our data since the observed time series show seasonal patterns that do not appear to change over time.

Let us define the vector of flows of the flood-prone area $\mathbf{Flow}_t = [Inflow_t, Outflow_t, Internal flow_t]'$ with $Inflow_t$, $Outflow_t$, and $Internal flow_t$ defined as in Eqs. (2)–(3). We model $\mathbf{Flow}_t$ as a VARX(p_d, p_w):

$$\mathbf{Flow}_t = \boldsymbol{\nu} + \sum_{h_d=1}^{p_d} \mathbf{A}_{h_d} \mathbf{Flow}_{t-24\times h_d} + \sum_{h_w=1}^{p_w} \mathbf{A}_{h_w} \mathbf{Flow}_{t-168\times h_w} + \mathbf{Bx}_t + \boldsymbol{\epsilon}_t \quad (4)$$

where $\boldsymbol{\nu}$ is a constant vector of length 3, p_d and p_w are, respectively, the daily and the weekly autoregressive parameters, $\mathbf{A}_{h_d}$ and $\mathbf{A}_{h_w}$ are two 3×3 matrices of coefficients to be estimated, and $\boldsymbol{\epsilon}_t$ is the 3×1 vector

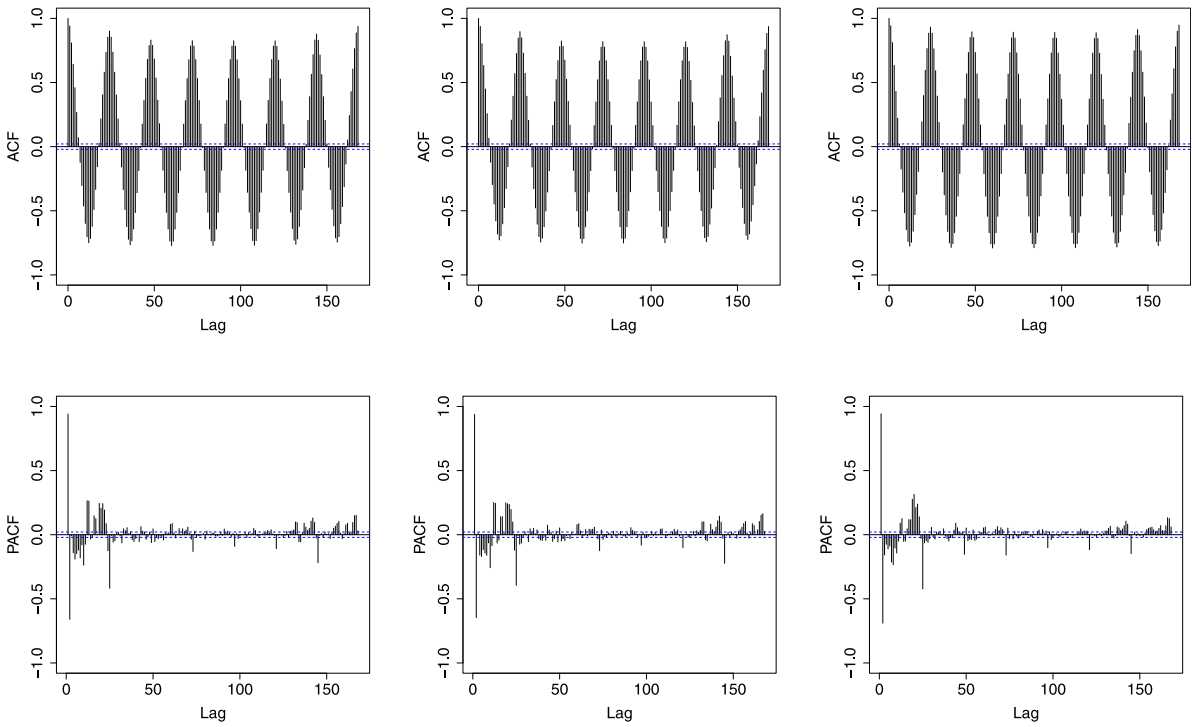

Fig. 7. ACF (top) and PACF (bottom) for time lags up to one week (168 h). From left to right: Inflows, Outflows, Internal flows.

of the error terms at time t; $\mathbf{x}_t$ is the vector of the l exogenous variables at time t, and $\mathbf{B}$ is the $3 \times l$ matrix of coefficients of the exogenous variables, so that $\mathbf{Bx}_t$ is a 3×1 vector. Note that the model assumes that the 3 dependent variables (i.e., inflows, outflows, and internal flows) at time t are related to each other.

Since the OD data are not "real-time" data and TIM is able to provide them at least 24 h after they have been collected, lags of less than 24 h are neglected. In this respect, the lag corresponding to $p_d = 1$ is the lag 24 (same hour of the previous day) and the lag corresponding to $p_w = 1$ is the lag 24×7=168 (same hour on the same weekday of the previous week).

To capture the seasonality of traffic flows, we model $\mathbf{Bx}_t$ as a DHR(K_d, K_w) [18]. More specifically, the rth element of the vector $\mathbf{Bx}_t$ is equal to a combination of daily (d) and weekly (w) periodic functions:

$$\beta_0^{(r)} + \sum_{k_d=1}^{K_d} [\alpha_{k_d}^{(r)} s_{k_d}(t) + \gamma_{k_d}^{(r)} c_{k_d}(t)] + \sum_{k_w=1}^{K_w} [\alpha_{k_w}^{(r)} s_{k_w}(t) + \gamma_{k_w}^{(r)} c_{k_w}(t)], \qquad r = 1, 2, 3,$$
$$s_{k_a}(t) = sin\left(\frac{2\pi k_a t}{m_a}\right), \quad c_{k_a}(t) = cos\left(\frac{2\pi k_a t}{m_a}\right), \qquad a = d, w, \tag{5}$$

where β_0 is a constant term, K_d and K_w are the optimal numbers of Fourier bases for the daily and the weekly pattern respectively, α_k and γ_k are regression coefficients to be estimated, $m_w = 24 \times 7 = 168$ is the weekly seasonal period, and $m_d = 24$ is the daily seasonal period. Note that β_0, α_k and γ_k are allowed to assume different values in the three elements of $\mathbf{Bx}_t$. It is worth noting that the DHR model requires, for each equation of the VARX model, $2 \times K_d$ parameters to estimate for the daily pattern and $2 \times K_w$ for the weekly pattern.

6. Results

In this section, we discuss the application and results of the model described in Eqs. (4)–(5) on the traffic flows of the flood-prone area of Brescia. We refer to the inflows, outflows, and internal flows adjusted by the MDT ratios of users on critical streets as in Eqs. (2)–(3). The following subsections discuss respectively the identification of the number of parameters and the forecasting performance of our model.

6.1. Specification of the model

Four numbers must be chosen in our model: the optimal number of autoregressive parameters p_d and p_w in Eq. (4) and the optimal number of Fourier bases K_d and K_w in Eq. (5). We first identify K_d and K_w following the approach by Metulini and Carpita [24]: first, for each time series, we consider a univariate DHR model with no explanatory variables nor autoregressive or moving average terms and select K_d between 1 and $m_d/2$ using the AIC; then, we include the selected K_d daily Fourier basis in the univariate DHR model and choose K_w between 1 and $m_w/2$ using the AIC. The procedure has been repeated separately on inflows, outflows, and internal flows. Considering a trade-off between model simplicity and performance, the solution with $K_d = 7$ for the daily component and $K_w = 6$ for the weekly component for all flows appears accurate.

To identify the autoregressive parameters, we set $K_d = 7$ and $K_w = 6$ and choose the values p_d and p_w using the approach proposed by [13] which is based on the AIC. As a result, we obtain $p_d = 3$ and $p_w = 4$. So, the chosen model is a VARX($p_d = 3$, $p_w = 4$) with DHR($K_d = 7, K_w = 6$) components as exogenous variables.

Some dummies have been included as exogenous variables in the model: monthly dummies to control for the possible presence of changes in average levels between months (e.g., higher average traffic flows in a specific month); and weekdays dummies to control for the possible presence of changes in average levels between weekdays.

The estimation of the final VARX model with DHR components has been performed using the least squares method[3]. To this aim, we use the functions `VARX` and `VARXpred` in the `R` package `MTS` [25].

[3] For technical reasons due to the provider TIM, data for September, 2nd, 2020 were not available and have been replaced by the data for September, 9th, 2020, i.e. the same day of the following week.

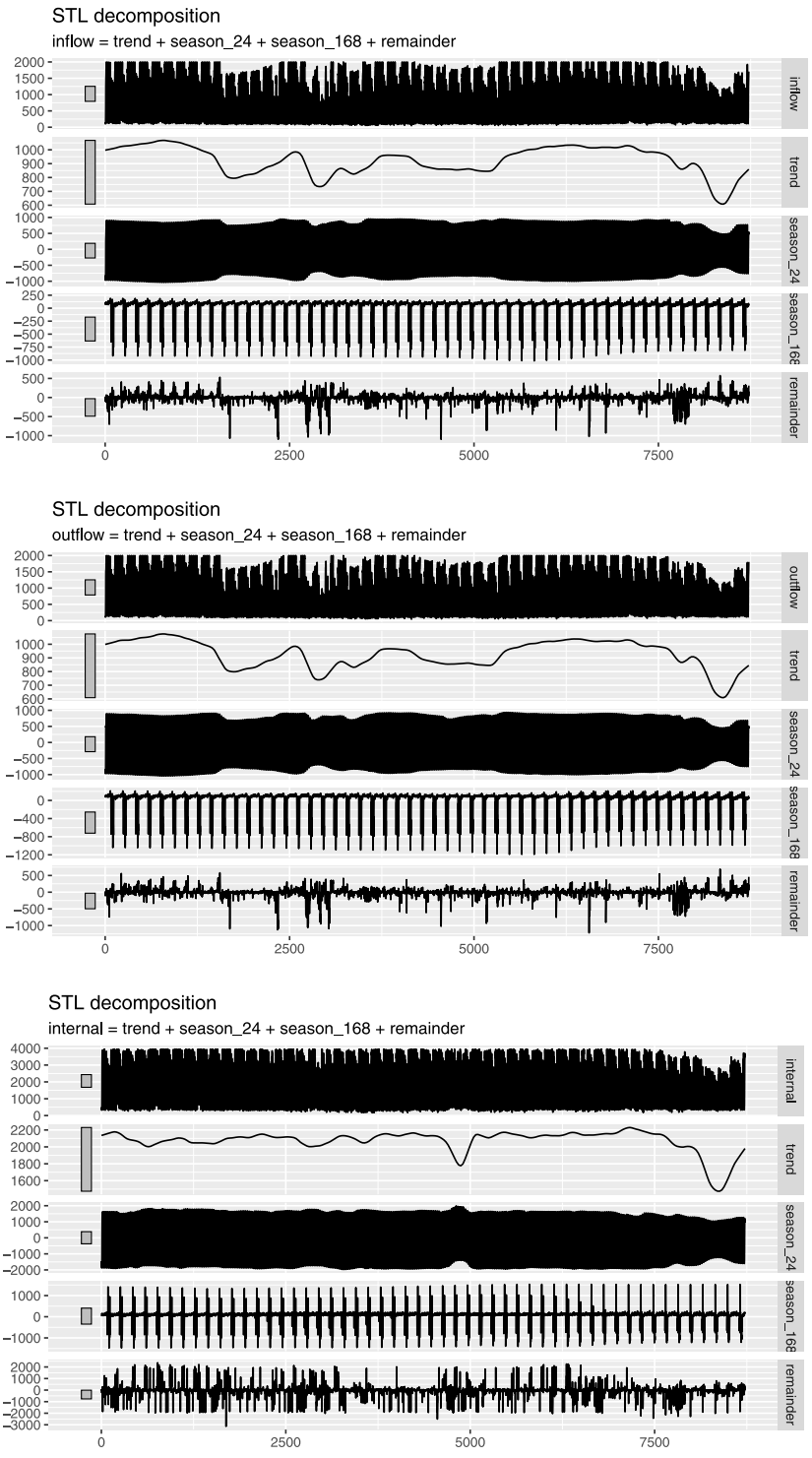

Fig. 8. STL with LOESS, trend, daily and weekly patterns, and remainder component. 1 year of data: from September 1st, 2020 to August 31st, 2021. From top to bottom: Inflows, Outflows, Internal flows.

Results are reported in Table A.1 in Appendix. We do not deepen the analysis and the interpretation of the estimated regression coefficients because we are interested in forecasting accuracy only. However, it is worth observing that, despite some of the estimated coefficients associated with the autoregressive terms are not statistically significant, we decided to keep all of them because a better forecasting performance is achieved in this way. The analysis of estimated residuals is relevant. The PACF of the estimated residuals, reported in the bottom-left charts of Fig. 9, displays significant first-order autocorrelation. Moreover, the residuals are not normally distributed (histograms in Fig. 9), but follow a leptokurtic distribution with heavy tails instead. The possible driver of this unfavorable evidence on the residuals might be the restriction on using lags greater than 24 hours[4] which unfortunately

[4] As a robustness check, the VARX model has been also estimated replacing the dependent variables with a Box–Cox power transformation of original values via maximum likelihood estimation [26]. Residuals do not show a

Table 3
Ljung–Box test on the residuals of the inflows, outflows, and internal flows. The table reports the values of the χ^2 statistic and the associated p-value for the model with lagged variables from order 2 to 40 (first two columns), and for the model with lagged variables from order 1 to 40 (last two columns).

	Lags 2:40		Lags 1:40	
	χ^2	p-value	χ^2	p-value
Inflow	7388.2	<2.2e−16	11 974	<2.2e−16
Outflow	6991.9	<2.2e−16	11 684	<2.2e−16
Internal flow	4496.3	<2.2e−16	7440	<2.2e−16

cannot be relaxed. As clearly shown by the density plot and the QQ-plots in the bottom charts of Fig. 9, the Normality assumption on the estimated residuals is violated. Nevertheless, this does not deny obtaining unbiased estimates of the parameters and the forecasting asymptotic properties are maintained. It is worth noting that non-Normal estimated residuals were also obtained in the analysis of traffic flows in the ACE of Cellatica by Metulini and Carpita [13]. Moreover, the ACF is exponentially decreasing and a notably significant value of the PACF is observed at lag 1. As a result, the null hypothesis of 0 serial correlation in the Ljung–Box test is rejected (see the first two columns of Table 3). However, these outcomes perfectly align with expectations, given the constraints imposed by data availability on the VARX model. Specifically, it is noteworthy that the inclusion of lag 1 in the model is precluded by these constraints. As shown in the last two columns of Table 3, the Ljung–Box test suggests the presence of serial correlation even when the first lag is included. Notwithstanding this, the magnitudes of partial autocorrelations beyond lag 1 remain remarkably low (less than 0.1), rendering them neglectable within the scope of our analysis.

6.2. Forecasting performance of the model

This section presents a detailed discussion of the forecasting ability of our model.

The cross-validation strategy described in [13], which is an adaptation of a blocked k-folds cross-validation for time series [18,27], is applied to assess the forecasting ability of the model. According to this methodology, from the dataset where holidays are replaced, one day of observations is chosen as the validation set and the previous two months of data are taken as the training set. As a result, each validation set has sample size $n_v = 24$ intervals of one hour, and each training set has sample size $n_t = 1440$ (24 observations × 60 days). We forecast the flows in the time intervals of the validation set using the model defined in Eqs. (4) and (5) with parameters $p_d = 3$, $p_w = 4$, $K_d = 7$, and $K_w = 6$ (see Section 6.1) estimated on the training set. To evaluate the performance consistently over all the days of the year, we replicate the analysis on different sets of training and validation samples (see Fig. 10). Note that, being data available for exactly 1 year, 88 days (from September, 1st to November 28th) cannot be validated. In fact, the training set has 60 days and the model trains using autoregressive lags up to 4 weeks (i.e., $p_w = 4$), which we call *lag terms set*.

For each fold, the forecasts of the validation sets are then compared with the corresponding observed values. As an example, plots in Fig. 11 show (top to bottom) the time series of observed (black) versus forecasted (colored) inflows, outflows, and internal flows for four validation days: Wednesday, February 10th; Saturday, February 13th; Tuesday, July 13th; Saturday, July 17th. To assess the accuracy of our forecasts, we refer to two measures: the Symmetric Mean Absolute Percentage Error (SMAPE hereafter) [28] and the Hit Rate (HR hereafter).

The SMAPE is an accuracy measure for quantitative variables and is a variation of the MAPE [29,30] that is particularly suitable for non-symmetrically distributed data, as this is the case. It is not affected by variables' scaling nor by negative or close-to −0 observations and equally penalizes negative and positive errors.

For each validation set, we compute the SMAPE as:

$$SMAPE = \frac{100}{24} \sum_{t=1}^{24} \frac{| f_t - \hat{f}_t |}{(| f_t | + | \hat{f}_t |)/2} \tag{6}$$

where f_t is the observed value of a flow at time t (either $Inflow_t$, $Outflow_t$, or $Internalflow_t$), $\hat{f}_t$ is the corresponding forecasted value, and $|\cdot|$ is the absolute value operator. The SMAPE takes values from 0 to 100. In general, a forecasting performance is considered very good when the SMAPE is lower than 10, and fairly good when the SMAPE takes values between 10 and 20. On the validation days in Fig. 11, we find that the SMAPE varies from 7.5 to 23.5 for outflows, from 7.3 to 23.4 for inflows, and from 6.1 to 31.0 for internal flows. The monthly mean and standard deviation of the SMAPE are reported for the three flows in Table 4. As could be noticed, values do not considerably vary among the different months observed in the validation set. The monthly mean SMAPE ranges from approximately 8 (March) to 15 (January) for inflows and outflows. Even better performance has been observed for internal flows, for which the monthly mean SMAPE ranges from less than 7 (February and April) to about 13 (July)[5]. Moreover, the Table shows that the months associated with a larger standard deviation of the SMAPE are December and January for inflows and outflows, and July for the internal flows. Overall, these findings suggest that the model forecasting performance is slightly weaker for January and summer days.

In addition to the SMAPE, we evaluate the performance of the model in forecasting traffic flows by means of HR. HR is an accuracy measure for categorical variables and is widely adopted in the classification of customers for credit scoring [31]. This measure allows us to assess the ability of our model in forecasting the level of traffic intensity. Indeed, regulators are often more interested in correctly forecasting whether the traffic flow will be higher than a certain threshold rather than the exact amount of people moving. Therefore, for practical applications, our model must have a satisfactory ability to forecast the level of traffic. We consider five categories to represent traffic levels, which we define based on the quintiles of the related distribution, namely: very high, high, moderate, low, and very low[6]. For each validation set and separately for inflows, outflows, and internal flows, each observed value is assigned to the corresponding category according to the distribution of f, and each forecasted value is assigned to the corresponding category according to the distribution of $\hat{f}$. The HR is then computed as:

$$HR = \frac{1}{24} \sum_{t=1}^{24} I\left(f_t \text{ and } \hat{f}_t \text{ belong to the same category}\right) \tag{7}$$

where $I(\cdot)$ is an indicator function that takes the value 1 if the category of f_t equals that of $\hat{f}_t$, and 0 otherwise. The HR takes values from 0 to 1, where 1 indicates that the forecasts perfectly match the observed values. The monthly mean and standard deviation of the HR

significant improvement in the distribution (results are available upon request) and do not justify the use of a Box–Cox transformation on original data. Moreover, we also tried introducing additional lags in the model to improve residuals. Specifically, we considered three alternative sets of additional lags: lags from 25 (the previous hour of the previous day) to 27, from 25 to 29, and from 25 to 30. Unfortunately, we found that neither of these specifications improved residuals. Moreover, we did not even find evidence of improvements in the fitting of the model: the AIC and BIC remain substantially unchanged when adding lags. Therefore, the here proposed model was preferred for parsimony.

[5] The reported ranges of values for inflows, outflows, and internal flows do not account for the month of November, as it was possible to compute the SMAPE for only three days of the month.

[6] The analysis has been repeated considering seven categories of traffic flows. Results are very similar to those obtained with the five-categories setting.

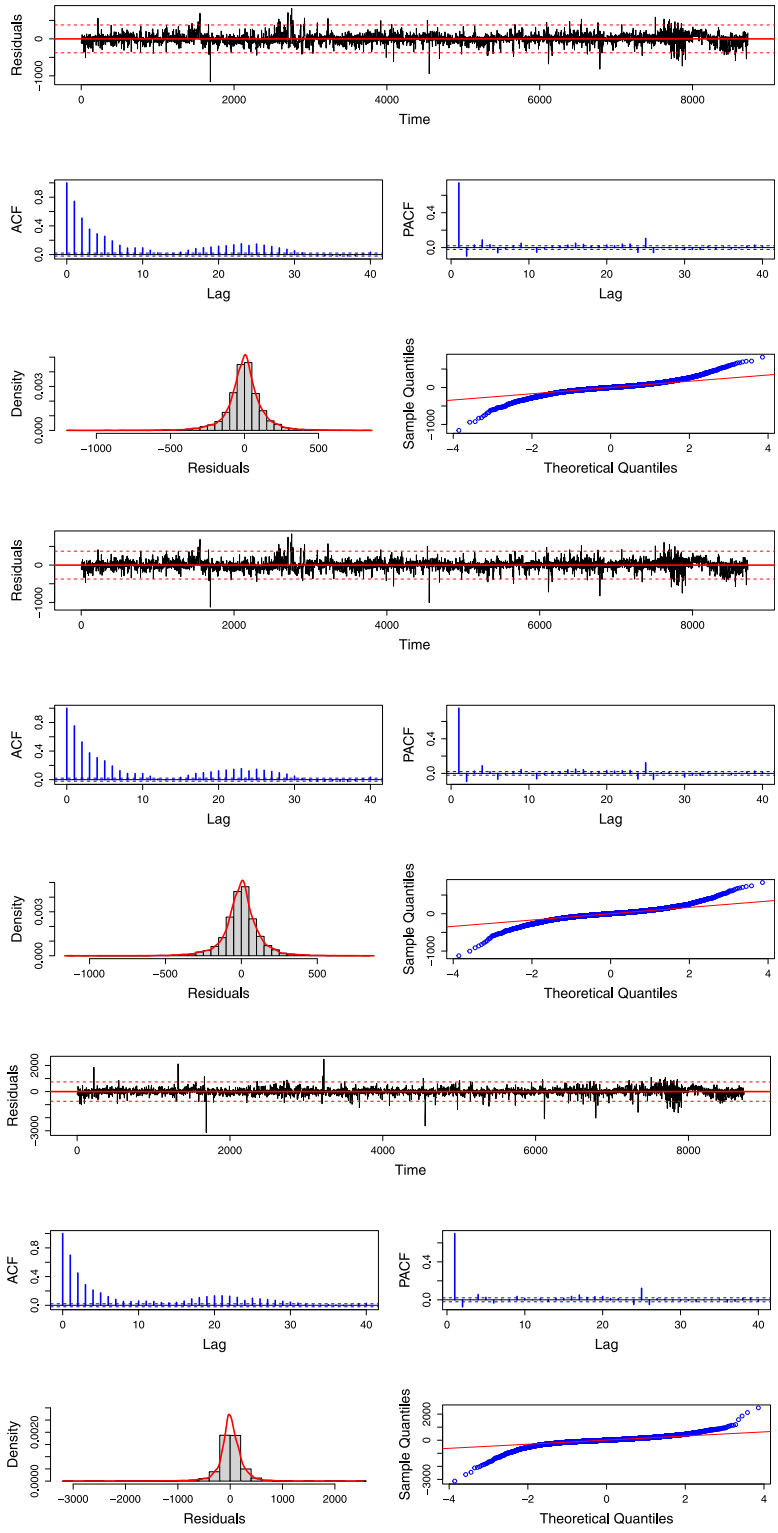

Fig. 9. Residuals' diagnostic: Time series of estimated residuals (top) with 95% confidence bands; ACF (center left) and PACF (center right), with 95% confidence bounds for strict white noise; histogram of the empirical distribution with Normal curve (bottom left) and QQ-plot for normality. Some extremely high values have been replaced with a cut-off value. From top to bottom: Inflows, Outflows, Internal Flows.

have been computed for inflows, outflows, and internal flows and are reported in Table 5. According to the HR, the model performs similarly in different months. Average HR values range from 0.80 to 0.84 for the inflows, from 0.74 to 0.83 for the outflows, and from 0.75 to 0.87 for internal flows.[7] Overall, these results suggest that the model provides good accuracy in forecasting the category of traffic intensity. Similarly

[7] Again, we do not consider the month of November, as only three days of the month are available for the performance evaluation.

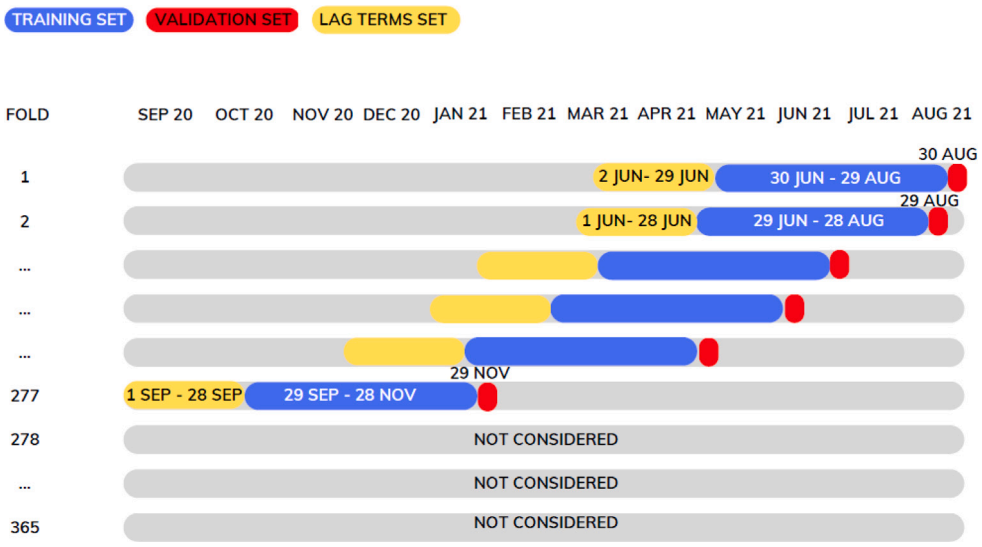

Fig. 10. Illustration of the blocked k-folds cross validation adopted.

Table 4
Mean and standard deviation (inside brackets) of SMAPE from November 2020 to August 2021 by month. Note that SMAPE values for November have been computed on 3 days only.

	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Inflows	8.46	13.37	15.64	11.51	7.89	9.29	8.55	8.76	11.74	12.07
	(4.64)	(7.15)	(7.92)	(6.47)	(4.52)	(4.65)	(3.88)	(3.04)	(5.85)	(4.48)
Outflows	9.23	13.31	15.59	11.47	7.81	9.42	8.39	8.53	11.72	12.38
	(5.78)	(7.24)	(7.96)	(6.50)	(4.74)	(4.91)	(3.88)	(3.17)	(5.94)	(4.57)
Internal flows	4.60	7.32	9.04	6.64	7.05	6.83	7.01	8.12	12.96	9.91
	(1.73)	(2.63)	(3.90)	(3.42)	(4.55)	(3.10)	(3.70)	(3.51)	(8.42)	(3.67)

Table 5
Mean and standard deviation (inside brackets) of HR from November 2020 to August 2021 by month. Note that HR values for November have been computed on 3 days only.

Series	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Inflows	0.74	0.81	0.80	0.80	0.83	0.82	0.84	0.85	0.82	0.84
	(0.12)	(0.09)	(0.09)	(0.10)	(0.11)	(0.08)	(0.07)	(0.10)	(0.10)	(0.07)
Outflows	0.78	0.76	0.76	0.74	0.78	0.80	0.80	0.81	0.78	0.83
	(0.06)	(0.10)	(0.07)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)	(0.07)
Internal flows	0.89	0.83	0.85	0.83	0.84	0.87	0.85	0.81	0.75	0.77
	(0.15)	(0.07)	(0.07)	(0.06)	(0.12)	(0.07)	(0.08)	(0.10)	(0.14)	(0.09)

to the SMAPE (Table 4), the HR indicates that the model performs better for non-summer days, though the indicator appears quite stable over the time period.

Histograms in Fig. 12 show the distribution of HR over the set of available validation days for inflows, outflows, and internal flows. The distributions appear all negatively skewed, with median values larger than 0.8. Inflows and internal flows present a pick for HR equal to 0.90, while outflows at 0.80. The histograms also show that values of HR lower than 0.6 (lower whisker) are obtained for a few validation sets.

At last, the days for which the model achieves poor performance according to the HR have been investigated. Results are shown in the box plot in Fig. 13, which reports results obtained on the validation set. As could be noticed, most of the days corresponding to a low value of HR (i.e., the outliers) refer to July. Nevertheless, the number of days corresponding to a scarce forecasting performance is limited. It is worth underlying that none of these days corresponds to holidays because of the replacement strategy.

7. Discussion and conclusions

In this work, we address the assessment of the exposure to flooding risk in urban areas. We focus on the number of people moving, which we refer to as traffic flows, and present the case study of the Mandolossa region, which is a flood-prone area in the Province of Brescia (Italy). Two aspects are fundamental when analyzing traffic flows in risky areas:

- Traffic flows are dynamic and change over time. Therefore, static representations of human exposure are inadequate.
- Not all the people moving in the proximity of a flood-prone area is passing by the risky areas. It is important to identify who is in danger, and who is not.

In this work, we account for both aspects using an innovative combination of mobile phone data. The OD data flows allow us to capture traffic dynamics over time. In particular, we distinguish between inflows, outflows, and internal flows. The MDT signals data localize users with 10 meters accuracy and allow us to identify the portion of phone users on streets passing by the risky area.

Traffic flows have been analyzed, and weekly and daily seasonal patterns have been detected. To account for the two seasonal components as well as the dependence structure among inflows, outflows, and internal flows, we modeled traffic flows using a VARX model with DHR components. Then, the model forecast accuracy has been assessed using a blocked k-fold validation strategy paired with SMAPE and HR. We found that the model satisfactorily forecasts the number of people

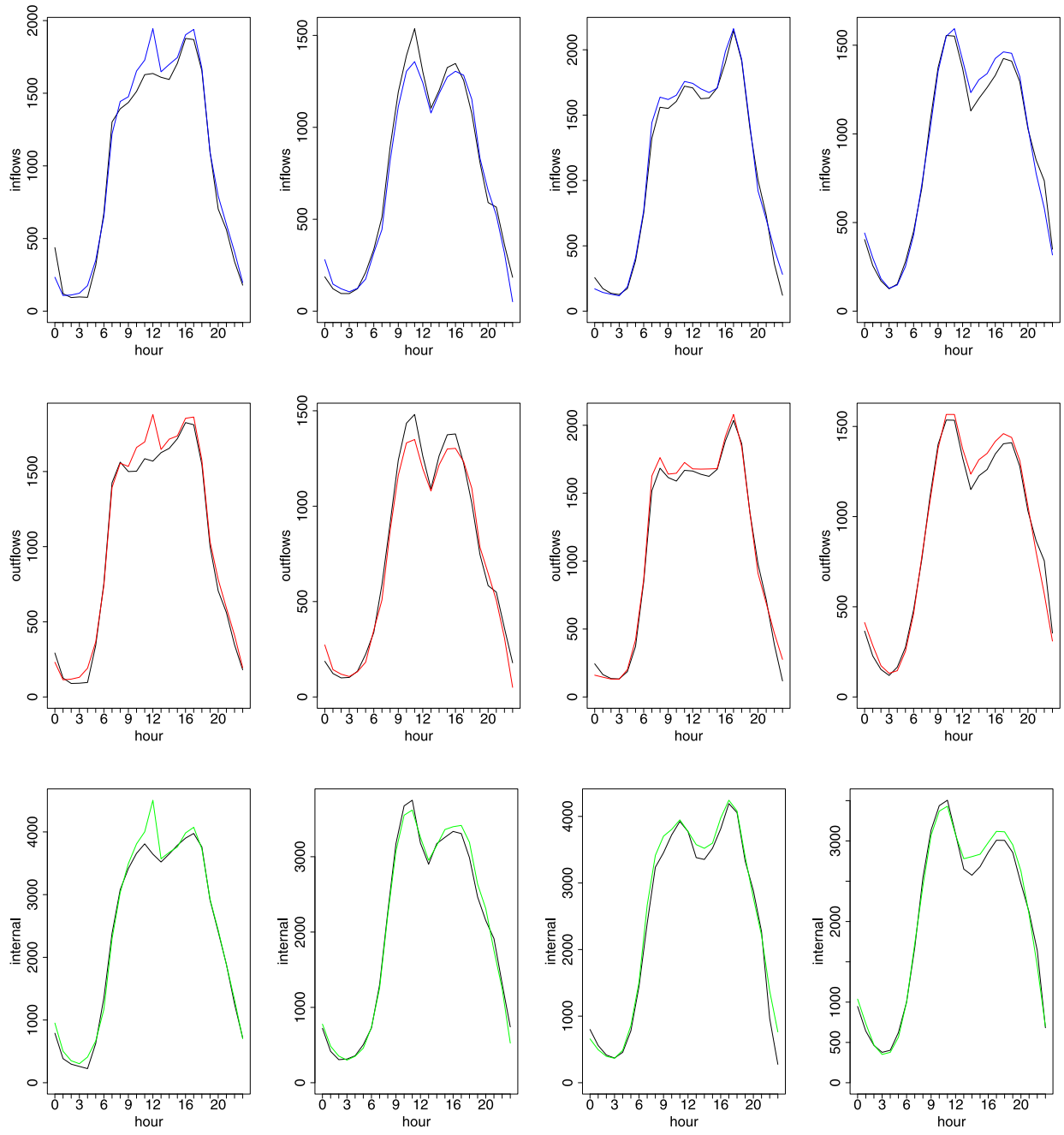

Fig. 11. Plot of observed (black) versus forecasted (colored) traffic flow in the flood risk area of the Mandolossa. Validation days (from left to right): February, 10th (Wednesday), February, 13th (Saturday), July, 13th (Tuesday), July, 17th (Saturday), year: 2021. From top to bottom: Inflows, Outflows, Internal flows.

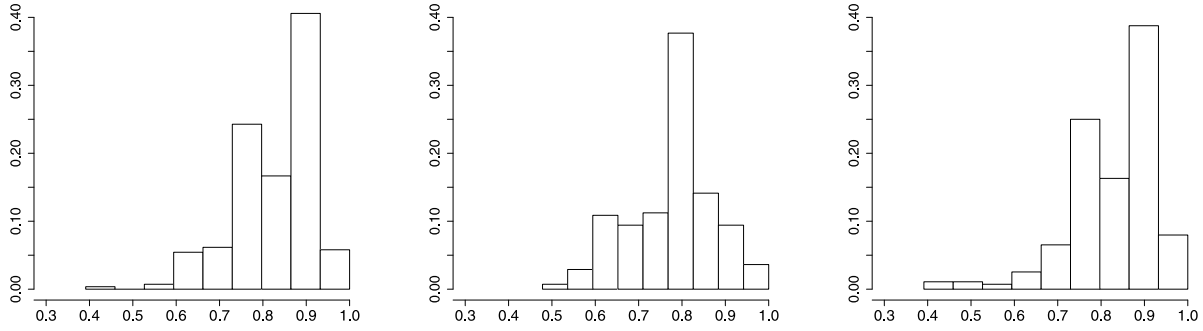

Fig. 12. Histogram of the distribution of the HR computed over the 277 days of the dataset. From left to right: Inflows, Outflows, Internal flows.

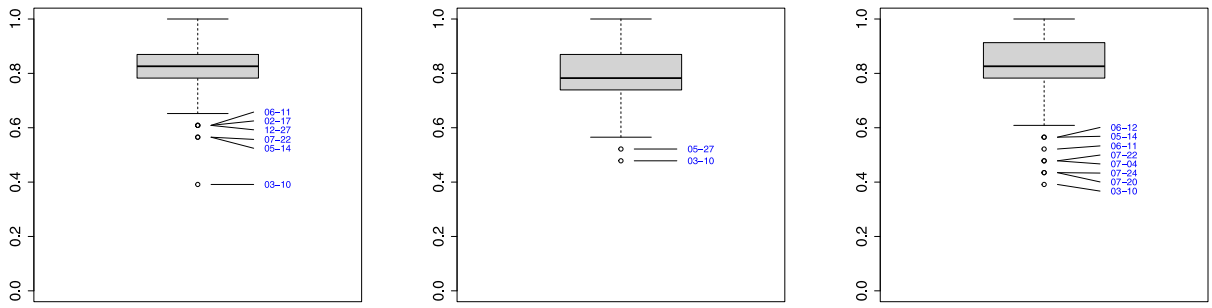

Fig. 13. Box plot of the HR computed over the 277 days of the dataset. The upper (lower) whisker represents the third (first) quartile plus (minus) 1.5 times the interquartile range. Days corresponding to extreme values of the HR are labeled in blue. From left to right: inflows, outflows, internal flows.

moving and achieves a good performance in forecasting the level of traffic intensity (i.e., very high, high, moderate, low, very low), though it performs better on non-summer days.

The proposed model can be used by regulators for monitoring the exposure of people to potentially harmful situations with reasonable advance. This information may support policymakers in evaluating which actions are more appropriate given the hazard, vulnerability, and exposure of the area. In particular, it provides information that local authorities can use to promptly activate crowding control actions aimed at preventing human losses and injuries. If combined with information about residents, it allows authorities to estimate the number of people to be evacuated in an emergency. However, exposure to flood risk is complex and multifaceted. Indeed, the number of people moving over the exposed area is just an important element in a much more complex framework. An interesting advancement that could be evaluated for future developments may concern the estimation of vehicular traffic. People in transit can be spread over a different number of vehicles. Information on vehicular traffic and the number of people on the move can be used jointly to evaluate and implement traffic viability control measures such as the preventive closure of the access roads to the risky area.

The main limitation of the presented work lies in the availability of mobile phone data. Current technology does not allow for real-time data. The data is provided the day after being observed and, for this reason, we were forced to exclude flows' lags of less than 24 h from the model. However, thanks to the fast advancements in technologies (e.g., 5G network), this limit will likely be overcome in the next future. When this happens, it will be possible to use mobile phone data to produce real-time estimates of people on the move. In turn, this information can be used for early warning systems. In this respect, the adoption of mobile data is very promising as it would make it possible to significantly contain the development and operating costs. Current early warning systems are based on very expensive engineering instruments which need to be installed ad-hoc and require adequate maintenance over time. On the contrary, the mobile phone network is already well distributed throughout the territory and would not require specific installation or maintenance interventions. Thus, despite the current limitations, this work has even greater relevance when considering possible future implications.

As a further development, the model might be improved by accounting for the specific characteristics of the non-Normal residuals. For example, skewed-student-t innovations might be assumed to account for skewness and high kurtosis and a GARCH model might be adopted to capture the time-varying variance of the residuals. Alternatively, the information on the heavy tails might be used to cluster areas where intense traffic simultaneously occurs. Furthermore, there exists the opportunity to delve into alternative models aimed at the forecasting of traffic flow. The selection of the VARX model with DHR components was made based on its adeptness in capturing complex seasonality while retaining its suitability within a multivariate framework. Nonetheless, it is worth noting that despite the increased complexity in their application to multivariate analyses, the exploration of VARMA and ETS models could also be contemplated. Moreover, alternative treatments of seasonality might also be considered. Here, we apply a methodology that jointly estimates flows and seasonality. This choice is motivated by the fact that the three types of flow exhibit strong correlations as well as strong autocorrelation. It might be interesting to investigate seasonality outside the model by seasonally adjusting the series and fitting a VAR model. Moreover, additional explanatory variables might be introduced, such as the census population of both the ACEs of origin and destination, or weather variables, such as temperature or level of precipitation, due to the relationship between floods and meteorological measures.

CRediT authorship contribution statement

Selene Perazzini: Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Rodolfo Metulini:** Methodology, Software, Formal analysis, Investigation, Data curation, Validation, Writing – original draft, Writing – review & editing. **Maurizio Carpita:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This study has been developed for the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), and it is partially funded by the project "Sustainable Mobility Center (MOST)" 2022–2026, CUP D83C22000690001, Spoke N° 7 "CCAM, Connected networks and Smart Infrastructures". Thanks to FasterNet S.r.l. (www.fasternet.it) and to the MoSoRe Project 2020–2022 (https://ricerca2.unibs.it/?pageid=8548) for providing the TIM flows and signals Data. Also thanks to the Team of Olivetti-TIM (www.olivetti.com/en) for the technical support to the data provision.

Appendix. Construction of the weights

The proportion of traffic flows passing by the flood risk area has been computed for the 4 ACEs of interest using a 7 steps-procedure:

Table A.1
Results of the VARX($p_d = 3, p_w = 4$) with a DHR($K_d = 7, K_w = 6$) component.

Endogenous variable	**Inflow (s.e.)**	**Outflow (s.e.)**	**Internal flow (s.e.)**
inflow_AR(1)_day	0.570 (0.061)	0.414 (0.061)	0.264 (0.118)
inflow_AR(2)_day	−0.098 (0.062)	−0.103 (0.063)	−0.066 (0.121)
inflow_AR(3)_day	−0.010 (0.062)	−0.049 (0.063)	−0.169 (0.121)
outflow_AR(1)_day	−0.256 (0.060)	−0.095 (0.060)	−0.403 (0.116)
outflow_AR(2)_day	−0.029 (0.060)	−0.031 (0.060)	−0.118 (0.117)
outflow_AR(3)_day	0.082 (0.062)	0.112 (0.063)	0.077 (0.121)
internal_flow_AR(1)_day	0.030 (0.062)	0.029 (0.063)	0.381 (0.121)
internal_flow_AR(2)_day	0.049 (0.059)	0.051 (0.060)	0.124 (0.115)
internal_flow_AR(3)_day	0.015 (0.009)	0.019 (0.009)	0.162 (0.018)
inflow_AR(1)_week	0.246 (0.009)	−0.041 (0.009)	0.149 (0.018)
inflow_AR(2)_week	0.008 (0.009)	−0.145 (0.009)	−0.025 (0.018)
inflow_AR(3)_week	0.180 (0.009)	0.044 (0.009)	0.245 (0.017)
inflow_AR(4)_week	0.132 (0.055)	0.038 (0.055)	0.180 (0.107)
outflow_AR(1)_week	0.074 (0.056)	0.368 (0.056)	−0.057 (0.109)
outflow_AR(2)_week	0.030 (0.053)	0.189 (0.053)	0.015 (0.103)
outflow_AR(3)_week	0.004 (0.053)	0.140 (0.054)	−0.010 (0.104)
outflow_AR(4)_week	−0.052 (0.054)	0.044 (0.054)	−0.061 (0.105)
internal_flow_AR(1)_week	−0.059 (0.051)	−0.062 (0.051)	0.121 (0.099)
internal_flow_AR(2)_week	0.003 (0.009)	0.000 (0.009)	0.054 (0.018)
internal_flow_AR(3)_week	−0.027 (0.010)	−0.026 (0.010)	−0.004 (0.019)
internal_flow_AR(4)_week	−0.003 (0.009)	−0.003 (0.009)	0.017 (0.018)
exogenous variable	**outflow (s.e.)**	**inflow (s.e.)**	**internal flow (s.e.)**
sin_day_1	−50.647 (8.908)	−41.909 (8.990)	−149.059 (17.349)
cos_day_1	−73.328 (8.447)	−72.018 (8.524)	−183.885 (16.450)
sin_day_2	−14.188 (3.580)	−16.070 (3.613)	−64.125 (6.972)
cos_day_2	−9.015 (2.463)	−9.535 (2.485)	1.559 (4.797)
sin_day_3	17.476 (2.775)	16.328 (2.800)	39.824 (5.403)
cos_day_3	7.020 (3.523)	11.478 (3.556)	2.408 (6.862)
sin_day_4	2.570 (2.009)	2.659 (2.027)	−4.066 (3.912)
cos_day_4	−0.693 (2.769)	−3.913 (2.795)	−15.203 (5.393)
sin_day_5	−8.219 (2.091)	−8.222 (2.110)	−9.158 (4.072)
cos_day_5	0.001 (1.923)	−0.069 (1.941)	−1.656 (3.745)
sin_day_6	0.030 (1.932)	−0.262 (1.950)	2.467 (3.762)
cos_day_6	−0.170 (1.967)	0.526 (1.985)	−1.654 (3.831)
sin_day_7	2.654 (2.041)	3.748 (2.060)	6.174 (3.975)
cos_day_7	−3.930 (1.965)	−3.940 (1.983)	−9.272 (3.827)
sin_week_1	163.312 (10.114)	157.363 (10.207)	322.871 (19.697)
cos_week_1	20.020 (10.274)	20.168 (10.368)	33.954 (20.008)
sin_week_2	−109.552 (5.709)	−108.577 (5.761)	−191.979 (11.118)
cos_week_2	37.770 (5.602)	37.824 (5.653)	43.589 (10.909)
sin_week_3	42.107 (3.753)	42.816 (3.787)	76.218 (7.308)
cos_week_3	−54.223 (4.109)	−53.980 (4.147)	−106.939 (8.003)
sin_week_4	43.265 (3.137)	42.930 (3.165)	87.501 (6.108)
cos_week_4	33.244 (3.366)	34.436 (3.397)	69.129 (6.556)
sin_week_5	−69.542 (3.133)	−69.733 (3.161)	−109.762 (6.101)
cos_week_5	3.196 (3.092)	−0.230 (3.121)	14.329 (6.022)
sin_week_6	54.724 (2.802)	55.723 (2.828)	94.478 (5.457)
cos_week_6	−35.847 (3.077)	−31.771 (3.105)	−65.942 (5.992)
month (ref. January): February	64.869 (7.038)	66.002 (7.102)	52.810 (13.706)
March	18.464 (6.886)	20.391 (6.949)	28.904 (13.410)
April	−0.683 (6.732)	−0.783 (6.794)	0.359 (13.111)
May	40.883 (6.887)	39.275 (6.950)	59.165 (13.411)
June	65.458 (7.080)	65.504 (7.145)	92.917 (13.788)
July	68.160 (7.344)	69.405 (7.412)	55.784 (14.303)
August	−34.724 (7.126)	−36.678 (7.191)	−144.595 (13.878)
September	25.884 (7.165)	26.603 (7.231)	6.497 (13.954)
October	108.951 (7.105)	110.412 (7.170)	139.807 (13.838)
November	16.092 (6.819)	17.361 (6.881)	35.251 (13.279)
December	106.026 (6.757)	107.372 (6.818)	13.2837 (13.158)
weekday (ref. Monday): Tuesday	−48.998 (14.966)	−46.302 (15.103)	−61.781 (29.146)
Wednesday	−87.868 (18.447)	−83.178 (18.616)	−179.269 (35.925)
Thursday	−110.401 (20.767)	−102.202 (20.957)	−238.359 (40.443)
Friday	−157.515 (20.982)	−147.637 (21.174)	−332.078 (40.862)
Saturday	−87.272 (19.188)	−81.322 (19.363)	−187.802 (37.368)
Sunday	109.832 (15.186)	111.145 (15.325)	207.293 (29.574)
intercept	82.660 (21.072)	71.637 (21.258)	345.997 (39.537)
residual correlation matrix	outflow	inflow	internal flow
outflow	1	0.971	0.666
inflow	0.971	1	0.664
internal flow	0.666	0.664	1
information criteria	AIC: 25.795; BIC: 25.953		

1. **Selection of the MDT data signals corresponding to the four ACEs of interest.** We compared the MDT grid of pixels with the map of the administrative boundaries in order to identify the phone signals related to the four ACEs of interest. The area of the MDT database corresponding to the four ACEs that we used for the analysis is shown in the left map of Fig. 3. This step restricted the MDT database to 113397 cells of the pixel grid.
2. **Identification of the area critical for flood emergency management.** We selected the "Sezioni di CEnsimento" (SCEs hereafter) – which are subdivisions of the ACEs – that are critical for flood management. The procedure for this step is shown in the central map of Fig. 3. First, we considered the map of the 300 SCEs constituting the 4 ACEs and compared them to the flood risk map. In the map, the SCEs and the flood maps correspond to the black borders and the blue area respectively. To capture all the streets passing by or linked to the flood risk area, we applied a 500 meters-buffer to the risk map (light blue area of the map). The selected buffer reflects the geographical accuracy of the OD data indicated by the TIM data operator. We found that the identified area intersects 92 SCEs, which we, therefore, considered critical for flood risk management. Furthermore, a graphical inspection revealed 12 additional SCEs in which there are important routes connecting the flood-prone area to the rest of the province of Brescia. We included them in the selection as well. Overall, we identified 104 SCEs that define the area critical for flood emergency management, which is reported in the map as the green polygon.
3. **Street map and identification of the streets critical for flood emergency management.** Most of the MDT cells of the pixel grid selected in Step 1 correspond to buildings. Since we aim at capturing traffic flows, we further need to restrict the MDT database to the cells representing streets. We identified the streets in the 4 ACEs by means of the street map presented in Section 2.3. Then, we restricted our attention to the SCEs selected in step 2 and identified the streets linked to or passing by the flood-prone area through a graphical inspection. This step is shown by the right map of Fig. 3, where the red polygons represent the streets that have been identified as critical for flood management, and the orange ones the rest of the streets in the ACEs. As it could be noted, not all the streets in the critical area (green polygon) have been included in the selection of critical streets. Indeed, we restricted our attention to those directly connected with the 20-year flood-risk map only. In practice, all the roads in the critical area have been considered critical but those that run along the edges of it without ever approaching the flood-prone zone.[8] This is the case, for example, of the major road on the northwest border of the critical area. This choice is motivated by the fact that this street runs alongside the flood-prone area without ever entering it. The high number of grid cells from which MDT signals originated located in it would have distorted the estimation of traffic flows at risk and we therefore considered it appropriate not to include it in the selection of critical streets.
4. **Identification of the MDT grid cells corresponding to streets.** We compared the MDT grid of pixels with the streets of the four ACEs (step 3) and identified the MDT cells on streets for each ACE and each time of observation (i.e., a 15-minute interval). Since the MDT signals are georeferenced with 10 meters accuracy, we considered "on street" all the cells of the grid that are at most 10 meters far from the roadway. Overall, we found that 54998 cells of the grid from which MDT signals were detected correspond to streets in the four ACEs. For this step, an illustrative example is presented in the left map of Fig. 4. The map refers to November 10th, 12:00–12:14 PM, and the ACE of "Brescia Mandolossa", but the same process has been repeated for all the time intervals and the other 3 ACEs. Then, for each ACE i and each time interval t, we counted the number of street-cells SC_{it}.
5. **Identification of MDT cells corresponding to critical streets.** Among the MDT grid cells on streets found in Step 4, we identified the ones corresponding to the streets critical for flood emergency management selected in Step 3. We repeated this procedure per each time interval and each ACE, and an example is presented in the right map of Fig. 4. For any combination of i and t, we counted the number of on MDT cells that correspond to critical streets, to which we refer as CSC_{it}.
6. **Aggregation of the data in time intervals of 4 h.** The variables SC_{it} and CSC_{it} obtained at steps 4 and 5 have been analyzed. We found that the number of MDT cells that originated signals on streets is very low during the night and that the number can vary considerably during the day. Therefore, the total number of MDT cells on streets and on critical streets have been aggregated in larger time intervals. Various alternatives have been explored, and we found the best solution in 6 intervals of 4 h, namely: 0:00–3:59, 4:00–7:59, 8:00–11:59, 12:00–15:59, 16:00–19:59, 20:00–23:59. As discussed in Section 2.2, ten 15 min-times of observation were missing, and have therefore been replaced with the average values of the time intervals available for the corresponding hour.
7. **Weights computation.** The proportions of street-cells corresponding to critical streets were computed for each ACE and each of the 6 time intervals as:

$$I.MDT_{it} = \frac{CSC_{it}}{SC_{it}} \tag{8}$$

where t indicates one 4-hour interval in one day. Results are reported in Fig. 5. The ratios for the interval 00:00-03:59 appear – again – strongly affected by the low number of signals detected and have been neglected. The remaining ratios are fairly constant among the intervals and the 5 days observed. We, therefore, computed the weight for each ACE i $\overline{I.MDT}_i$ as the average value of the ratios of the 5 time intervals in the 5 days.

[8] The impact of the chosen buffer on the resulting critical area and critical streets has been investigated. We find that, overall, buffers of 200, 300, and 400 meters would lead to substantially the same selection of SCEs and the same street area results. More in detail, we find that the three buffers would reduce the selected area of the critical street by about -8%, -5%, and -3% respectively. Both increasing or decreasing the buffer of 100 meters just slightly modifies the critical area. More in detail, a 400-meter buffer would lead to a decrease in the number of selected SCEs of 3 units, while a 600-meter buffer would increase it by 4 units. In either case, we get to the same selection of critical streets with approximately the same road surface.

References

[1] Benevolo C, Dameri RP, D'auria B. Smart mobility in smart city. In: Empowering organizations. Springer; 2016, p. 13–28.
[2] Bibri SE, Krogstie J. Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustain Cities Soc 2017;31:183–212.
[3] Reinolsmann N, Alhajyaseen W, Brijs T, Pirdavani A, Ross V, Hussain Q, et al. Delay or travel time information? The impact of advanced traveler information systems on drivers' behavior before freeway work zones. Transp Res F 2022;87:454–76.
[4] Pucci P, Gargiulo C, Manfredini F, Carpentieri G, et al. Mobile phone data for exploring spatio-temporal transformations in contemporary territories. TeMA-J Land Use Mob Environ 2022;2:6–12.
[5] Metulini R, Carpita M. A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Soc Indicators Res 2021;156(2):761–81.
[6] Tettamanti T, Varga I. Mobile phone location area based traffic flow estimation in urban road traffic. Adv Civ Environ Eng 2014;1(1):1–15.
[7] Carpita M, Simonetto A. Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. Electron J Appl Stat Anal: Decis Syst Serv Eval 2014;5(1):31–41.

[8] Mariotti I, Giavarini V, Rossi F, Akhavan M. Exploring the "15-Minute City" and near working in milan using mobile phone data. TeMA-J Land Use Mob Environ 2022;2:39–56.
[9] 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 2022;2:57–71.
[10] 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 2022;2:9–23.
[11] Balistrocchi M, Metulini R, Carpita M, Ranzi R. Dynamic maps of human exposure to floods based on mobile phone data. Nat Hazards Earth Syst Sci 2020;20(12):3485–500.
[12] Perazzini S, Metulini R, Carpita M. Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Stat Methods Appl (online first) 2023.
[13] 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. Ann Oper Res (online first) 2023.
[14] Wei Y, Wang J, Wang C. Network traffic prediction based on wavelet transform and season ARIMA model. In: International symposium on neural networks. Springer; 2011, p. 152–9.
[15] Jašek R, Szmit A, Szmit M. Usage of modern exponential-smoothing models in network traffic modelling. In: Nostradamus 2013: prediction, modeling and analysis of complex systems. Springer; 2013, p. 435–44.
[16] Guo X, Songling Z, Wu J. A hybrid seasonal autoregressive integrated moving average and denoising autoencoder model for atmospheric temperature profile prediction. Big Data 2022;10:493–505.
[17] Liu Z, Wang F, Dang A. Creating engines of prosperity: Spatiotemporal patterns and factors driving urban vitality in 36 key Chinese cities. Big Data 2022;10:528–46.
[18] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2018.
[19] Guo J, Peng Y, Peng X, Chen Q, Yu J, Dai Y. Traffic forecasting for mobile networks with multiplicative seasonal ARIMA models. In: 2009 9th international conference on electronic measurement & instruments. 2009, 3–377–3–380. http://dx.doi.org/10.1109/ICEMI.2009.5274287.
[20] Tran T, Ma Z, Li H, Hao L, Quang Khai T. A multiplicative seasonal ARIMA/GARCH model in EVN traffic prediction. Int J Commun Netw Syst Sci 2015;08:43–9. http://dx.doi.org/10.4236/ijcns.2015.84005.
[21] Tran T, Hao L, Quang Khai T. A novel procedure to model and forecast mobile communication traffic by ARIMA/GARCH combination models. 2016, http://dx.doi.org/10.2991/msota-16.2016.8.
[22] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A seasonal-trend decomposition. J Off Stat 1990;6(1):3–73.
[23] Hyndman RJ, Killick R. CRAN task view: Time series analysis. R Foundation for Statistical Computing; 2022, URL https://cran.r-project.org/web/views/TimeSeries.html.
[24] Metulini R, Carpita M. Forecasting traffic flows with complex seasonality using mobile phone data. In: Book of short papers IES 2022 innovation & society 5.0: statistical and economic methodologies for quality assessment. PKE Publisher; 2022, p. 38–43.
[25] Tsay RS. Multivariate time series analysis: with R and financial applications. John Wiley & Sons; 2013.
[26] Box GE, Cox DR. An analysis of transformations. J R Stat Soc Ser B Stat Methodol 1964;26(2):211–43.
[27] Snijders TA. On cross-validation for predictor evaluation in time series. In: On model uncertainty and its statistical implications. Springer; 1988, p. 56–69.
[28] Armstrong J. Long-range forecasting: from crystal ball tocomputer. Wiley: New York; 1978.
[29] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast 2006;22(4):679–88.
[30] Tofallis C. A better measure of relative prediction accuracy for model selection and model estimation. J Oper Res Soc 2015;66(8):1352–62.
[31] Bensic M, Sarlija N, Zekic-Susac M. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intell Syst Account Finance Manag: Int J 2005;13(3):133–50.

Selene Perazzini is a Research Assistant in Statistics at the University of Brescia. She received a Ph.D. degree in Economics, Management and Data Science at IMT School for Advanced Studies Lucca in December 2020, after obtaining a Bachelor's degree in Statistics cum laude, and a Master's degree with merit in Statistical, Financial and Actuarial Science at the University of Bologna. Her research interests include risk analysis and insurance.

Rodolfo Metulini is tenure track Assistant Professor in "Statistics for Experimental and Technological Research" at the University of Bergamo. He is the Principal Investigator of the PRIN/PNRR project "SIGNUM: Study of mobile phone siGNals for the evalUation of the interconnections between Mobility and the environment in Lombardia" (CUP: F53D23010910001). He received his Ph.D. in Statistics at the University of Bologna and he previously served IMT School for Advanced Studies, Sant' Anna Institute, University of Brescia and University of Salerno. Rodolfo published around 30 peer-reviewed international journals and conference proceedings and he presents a long-standing teaching activity on the discipline of Statistics. He served as Guest Editor of a thematic issue on the "Italian Journal of Applied Statistics".

Maurizio Carpita is Full Professor in Statistics and Scientific Director of the DMS StatLab at the University of Brescia (Italy), and Scientific Coordinator of the Research Group SVQS - "Statistics for Evaluation and Quality in Services" of the Italian Statistical Society (www.svqs.it). He is a Co-Editor in chief of the EJASA—Electronic Journal of Applied Statistical Analysis (siba-ese.unisalento.it/index.php/ejasa/index). His methodological research interests are latent variable and psychometric models, statistical methods and data mining tools for business intelligence. His studies are in the fields of the analysis of big data from telecom and Internet, assessment of the services quality and job satisfaction, testing of student performances (www.researchgate.net/profile/Maurizio-Carpita).