Annals of Operations Research (2024) 342:1629–1654
https://doi.org/10.1007/s10479-023-05195-8

ORIGINAL RESEARCH

Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making

Rodolfo Metulini[1] · **Maurizio Carpita**[2]

Accepted: 17 January 2023 / Published online: 31 January 2023

Abstract
Floods are one of the natural disasters which cause the worst human, social and economic impacts to the detriment of both public and private sectors. Today, public decision-makers can take advantage of the availability of data-driven systems that allow to monitor hydrogeological risk areas and that can be used for predictive purposes to deal with future emergency situations. Flooding risk exposure maps traditionally assume amount of presences constant over time, although crowding is a highly dynamic process in metropolitan areas. Real-time monitoring and forecasting of people's presences and mobility is thus a relevant aspect for metropolitan areas subjected to flooding risk. In this respect, mobile phone network data have been used with the aim of obtaining dynamic measure for the exposure risk in areas with hydrogeological criticality. In this work, we use mobile phone origin-destination signals on traffic flows by Telecom Italia Mobile (TIM) users with the aim of forecasting the exposure risk and thus to help decision-makers in warning to who is transiting through that area. To model the complex seasonality of traffic flows data, we adopt a novel methodological strategy based on introducing in a Vector AutoRegressive with eXogenous variable (VARX) model a Dynamic Harmonic Regression (DHR) component. We apply the method to the case study of the "Mandolossa", an urbanized area subject to flooding located on the western outskirt of Brescia, using hourly-basis data from September 2020 to August 2021. A cross validation based on the hit-rate and the mean absolute percentage error measures show a good forecasting accuracy.

Keywords Mobility · Forecasting · VARX models · Exposure risk · Dynamic harmonic regression

Both authors contributed equally to the manuscript.

✉ Rodolfo Metulini
rodolfo.metulini@unibg.it

Maurizio Carpita
maurizio.carpita@unibs.it

[1] Department of Economics, University of Bergamo, Via Caniana, 2, 24127 Bergamo, Italy

[2] Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 25122 Brescia, Italy

1 Introduction

Information and communication technologies (ICT) with big sources of relevant data has been massively used in the field of modern and smart cities and the study of urban systems (Albino et al., 2015; Bibri & Krogstie, 2017). According to Caragliu et al. (2011), a city can be defined as *"smart"* when *"investments in human and social capital and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources"*. With respect to this definition, the methodological approach proposed in this work goes in the direction of ensuring high quality of life, by developing a framework that can be useful for an early warning detection of flood exposure risk associated to human presence and people mobility in support of the broad topic of Disaster Management (see, e.g., Mishra et al. 2019).

Flood events are one of the natural disasters which cause significant social and economic impacts on human life as, when an area is flooded, many people need to evacuate to a safer place. This phenomenon may be fast and intense and can develop rapidly. Its predictability is the subject on ongoing research and the development of early warning systems is a determinant task. The bulletin provided by the italian national weather vigilance collects signals of relevant weather phenomena for civil protection purposes expected up to the midnight of that day, for the 24 hours of the following day, and also the tendency of the day after. The document is published each day at 3 pm at https://www.protezionecivile.gov.it/en/approfondimento/national-weather-vigilance-bulletins.

A fast response in evacuating people is one of the main flood risk management issues. Maps of flood risk and exposure generally assume people density constant over time, despite this is not the case in the "real-life" world, as crowding is a highly dynamic process in urban areas.

People mobility is thus a relevant aspect for metropolitan areas (Benevolo et al., 2016), where ICT may be used in support of the optimization of traffic flows or in tracking real-time citizens' position. A crucial aspect refers to the monitoring and the forecasting of the dynamic of people's presences and movements. Interpretative tools for tracking and modelling the presences in the metropolitan areas are needed because traditional data sources, such as census and surveys, have some known limitations like high costs and static nature of the data. Moreover, it is a matter of fact that typical smart cities present emerging forms of mobility and time-variations in the use of urban spaces, by both residents and temporary populations.

The new technology of mobile phone network data suits with the aim of producing dynamic information on people's presences (Metulini & Carpita, 2021) and movements (Tettamanti & Varga, 2014). Specific applications also regard the monitoring of the impact of social and cultural events (Carpita & Simonetto, 2014) and the development of dynamic exposure maps to flooding risk for areas with hydrogeological criticality. According to the latter strand of research, Balistrocchi et al. (2020) developed dynamic exposure maps using mobile phone data, while Kong et al. (2022) developed flooding risk maps based on taxi GPS traffic data. From a prevention perspective, the use of mobile phone technologies applied to flooding risk could make the identification of preferential traffic flows possible, thus evidencing potential risks during inundation onsets or emergency situations.

In this work, we use mobile phone data of Telecom Italia Mobile (TIM) retrieving the Origin-Destination (OD) traffic flows from/to different "census areas" (ACEs) of the "Italian National Statistical Institute" (ISTAT) on hourly basis for twelve months, from September 2020 to August 2021. The aim is to develop a statistical modelling framework for the fore-

casting of traffic flows transiting through specific flooding risk areas. It is worth mentioning that flood risk is a function of the hazard (which is related to the probability of occurrence of the flooding disaster), the exposure (which is related to the number of people that are present in the location involved) and the vulnerability (in turn, the lack of resistance to damaging forces) components (Kron, 2002). Here we work on exposure, in the sense that we are interested in forecasting the number of people that are present in flooding risk areas of interest. Once knowing if the amount of people in specific areas (in the near future) is higher than a pre-specified warning threshold, to alert people transiting through that area may be possible. In doing so, it is also worth highlighting that TIM does not provide data at "real-time". It follows the constraint that data of immediately previous hours are not allowed to be used for predicting actual data. For this reason, for the dynamic multivariate model proposed in this study we use lags of 24 hours, so that the model could be used with the actual lags of the TIM data availability.

Empirical evidence (that will be presented in Sect. 2.3) show a strong similarity among the different flows considered. This evidence proves that such flows belong to dependent processes. So, in this work we allow considered flows to be contemporaneously correlated to each other at the same time, but also correlated to each other's past value. Empirical evidence on the temporal dynamic of the OD traffic flows (Sect. 2.3) also highlight the presence of a regular pattern in which a daily and a weekly seasonality are dominant. These evidences, which will be extensively showed in the next section, motivate the choice of using a model with a Dynamic Harmonic Regression (DHR) component (Hyndman & Athanasopoulos, 2018) to deal with a complex seasonality structure, adopting the logic of approximating the temporal periodicity with the use of some Fourier bases.

To properly take into account for the dependence among the considered processes and for the seasonal patterns of the processes themselves, a Vector AutoRegressive with eXogenous variables (VARX) model is proposed in this work, where some DHR components are included among the X part. Similar methodological approaches, based on autoregressive models for the analysis of time series data, have been developed for the forecasting of mobile phone traffic. Guo et al. (2009) proposed the use of a multiplicative seasonal autoregressive integrated moving average modeling strategy. Tran et al. (2015, 2016), to the same goal, adopted a generalized autoregressive conditional heteroscedasticity strategy. However, to the best of our knowledge, this is the first attempt to incorporate a dynamic harmonic term into a vector autoregressive structure in the context of mobile phone traffic.

The performance of our VARX model with a DHR component in terms of forecasting accuracy has been tested by means of a blocked k-folds cross validation (CV) strategy for time series (Snijders, 1988; Hyndman & Athanasopoulos, 2018) which considers, as validation set, folds of 1 day (24 hours) and, as training set, blocks of similar length. We use the Mean Absolute Percentage Error (MAPE) for the performance evaluation using the validation sample. Inspired by the famous algorithm for the early detection of outbreaks, proposed by Farrington et al. (1996) (and developed by Yoneoka et al. (2021)) and based on defining a cut-off for the intensity of a specific event, we then turn observed and predicted values into categories using quintiles representing different levels of exposure risk, and we evaluate predicting performance by means of the Hit-Rate (HR) measure for confusion matrices.

We show the effectiveness of our approach by applying it to "inflows", "outflows" and "internal" flows of the ACE of "Cellatica" in the Mandolossa area, that is a critical zone with flood episodes in the north-west outskirt of the city of Brescia, Italy. The Mandolossa area subjected to flooding, actually, is highly urbanized, with the presence of roads and bridges, as documented in the report "Adeguamento della componente geologica, idrogeologica e sismica del PGT al Piano di Gestione del Rischio Alluvioni (P.G.R.A.)"

https://www.comune.brescia.it/servizi/urbanistica/PGT/Documents/Variante%20Idrogeologica%202017/VI-ALall04l-01c-Relazione%20idrologica%20e%20idraulica%20(Solda-Canale-M.pdf of the Municipality di Brescia. This structure highlights the link between flooding risk and traffic, as also demonstrated by the many flood events that cause traffic flows that happened in the past, such as the one in May 2010 in the area of S. Eufemia. Nevertheless, with appropriate generalizations, the proposed approach may be extended to any flooding risk area of interest.

The paper is structured as it follows. Sect. 2 describes the available mobile phone data and shows preliminary evidences that motivate the adopted methodological approach; Sect. 3 introduces the modeling framework; Sect. 4 is dedicated to the application of the method to real data, and Sect. 5 concludes the paper.

2 The origin-destination mobile phone data

2.1 Data description

Data configures as square OD (Origin-Destination) matrices of dimension $N \times N$, with $N = 235$ being the number of ACEs in the province of Brescia - Italy, according to ISTAT classification (see, e.g., https://www.istat.it/it/archivio/104317). Data are available at intervals of one hour for a total of one year (from September, 1st 2020 to August, 31th 2021) summing to a total of about $24 \times 365 = 8760$ different OD matrices.[1] Data have been provided by Olivetti (www.olivetti.com/en/iot-big-data) and FasterNet (www.fasternet.it) for the MoSoRe Project 2020-2022 (https://ricerca2.unibs.it/?page_id=8548).

A concordance table has also been provided so that to make the association of OD data with ISTAT cartography (e.g., the shape files provided at https://www.istat.it/it/archivio/104317) and demographic official data (such as the residential population) possible.

Data can be interpreted as the traffic flow from a specific ACE to another specific ACE in an hour interval. Flows from a specific ACE to itself (also called "internal" flows) are also available and displayed in the main diagonal of the above introduced OD matrices. To make an example, data related to the flow from an ACE (let's say, ACE1) to another ACE (let's say, ACE2) configures as the number of phone SIM that, in that specific hour interval, were retrieved by the antenna in ACE1 and, after five or more minutes, was retrieved by the antenna in ACE2. It is possible to separately count the number of human SIM (that, overall, sums to about 85% of the total SIM) and the number of M2M technology machine SIM (that counts for about 15% of the total SIM). In this work, we just consider the count of human SIM in order to avoid duplicates (e.g. a person who has both the mobile phone and the black box in his/her car). It is worth noting that both Italian and foreigner (in roaming) SIM are counted.

Since proprietary, the data are provided to us without a precise information about the error of measurement. We know that the retrieving system in use allows us a good precision in terms of geolocalization ($\sim$ 100 meters). Since spatial units of aggregations (the ACEs) are quite large, the geolocalization error is attenuated (Cellatica, the ACE that will be analyzed in the case study, measures 6.55 km^2). Moreover, by the design of the data retrievement, the position of the car is collected each 5 min. This leads to two considerations: (i) the counts of traffic flow might be underestimated, because the flow of a person who was recorded in,

[1] Actually, for a limited number of intervals, data are not available, due to technical issue. This aspect will be discussed later in this work.

let say, ACE1, and after 3 minpassed through ACE2, is not counted; (ii) the count of flows between two distant ACEs must be zero. Actually, flows between two distant ACEs are, in rare cases, different from zero. We attribute this fact to an issue related to the bad functioning of the retrieving system.

2.2 Area of interest

For the sake of our goal, which is related to quantify the traffic flows in flooding risk area, we need delimiting the area of interest. The method presented in Sect. 3 will be tested in Sect. 4 over the case study of the Mandolossa region, whose flooding map (with 10 return years, Balistrocchi et al. 2020) is represented in Fig. 1. The ACEs of interest are those which intersect with the identified flooding-risk area. Namely, those ACEs are "Gussago", "Cellatica", "Rodengo Saiano" and "Brescia Mandolossa" (which is, actually, an aggregation of two distinct ACEs in the city of Brescia).

We have identified other 38 neighboring ACEs (in turn aggregated, for a sake of representation, in four macro areas: "Bassa Bresciana", "Brescia", "Franciacorta" and "Valtrompia", located as in Fig. 1), which fulfil the criteria of counting a minimum flow from or flow to the four above defined ACEs. The total flows counted between the four ACEs intersecting the Mandolossa flood risk area and the 38 selected neighboring ACEs counts for about the 84% of the total flows from and to the four ACEs of the Mandolossa. Overall, we believe worthy to be considered i) the sixteen flows between the four ACEs of the Mandolossa (actually, four of them, that we will call "internal" flows, counts the number of flows from an ACE to the same ACE); ii) the sixteen flows from the four ACEs of the Mandolossa to the four neighboring macro areas; iii) the sixteen flows from the four neighboring macro areas to the four ACEs of the Mandolossa.

2.3 Preliminary evidences

Through the use of the above described data, we have studied the characterization of the dynamic of traffic flows over the time by jointly analyzing (as an example, but with the aim of generalizing the evidences to other flows inside the Mandolossa) the three following time series:

1. the flows to Cellatica (purple colored polygon in Fig. 1) from all other 38 neighboring ACEs (herein after, "inflows");
2. the flows from Cellatica to all other 38 neighboring ACEs ("outflows");
3. the flows from Cellatica to Cellatica itself ("internal" flows).

The choice of considering the ACE of Cellatica in this first study is motivated by looking at Fig. 1: the intersecting portion of the flooding risk area with the area of Cellatica is higher compared to the intersection with the other ACEs of the Mandolossa.

The choice of considering both inflows and outflows and internal flows is motivated by the fact that to correctly quantifying street crowding, it is also necessary to consider those moving within the borders of the considered area.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test (Kwiatkowski et al., 1992) for both i) level stationarity and ii) trend stationarity has been performed. The first test does reject the null hypothesis of stationarity while the second test does not. In other words, the three original time series are not stationary, but, when filtered by their complex seasonality, the

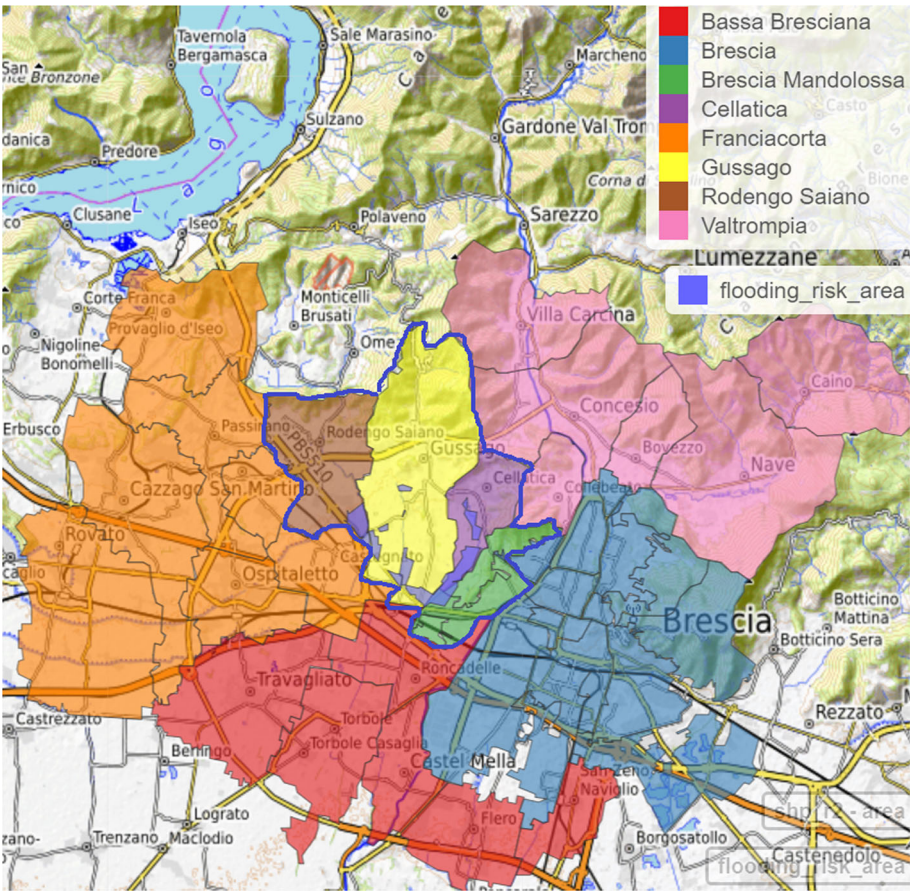


Fig. 1 Map of flooding risk area, ACEs of the Mandolossa and neighboring macro-areas

series turn out to be stationary and so they are allowed to be modeled within a traditional autoregressive setup.

Figure 2 shows a weekly excerpt (i.e., from 22th to 28th February, 2021) of the time series of inflows, outflows and internal flows in Cellatica. The figure highlights a strong similarity among the different time series.[2] This motivates us to model such flows as belonging to dependent processes. A so strong similarity pattern may be due to the topographical characterization of a typical urbanized area of interest, where the same ACE presents both residential and industrial zones nearby located (Fig. 3). This land use structure is well suited for a dynamics of traffic flows in which, for example, to a certain increase in the amount of flows entering a certain area in the morning (e.g., to go to work), an associated increase in the amount of outflow in the same area and at the same time (e.g., to leave the house) is likely.

We analyze the AutoCorrelation Function (ACF) and the Partial ACF (PACF), displayed in Fig. 4 for time lags up to one week (168 hours). ACF and PACF highlights a strong daily pattern. The left charts clearly show a daily periodicity, where the pick of positive

[2] This similarity emerges even if we consider different excerpt of the yearly time series. We have decided to show only one week for greater clarity.

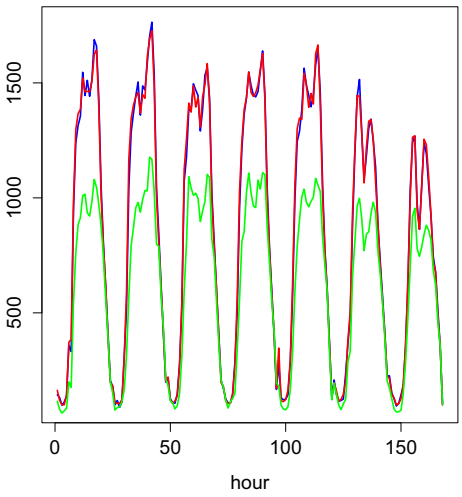


Fig. 2 Outflows from Cellatica (in red), inflows to Cellatica (blue), internal flows within Cellatica (green). Hourly data from 22th to 28th of February, 2021

autocorrelation is in correspondence of 24, 48, 72, ... hours of delay, the pick of negative autocorrelation is in correspondence of 12, 36, 60, ... hours of delay. Right charts show that partial autocorrelation is strong only at exactly 24, 48, 72, ... hours of delay. Moreover, autocorrelation patterns are very similar when considering inflows (Fig. 4a), outflows (Fig. 4b) and internal flows (Fig. 4c).

We have also performed an additive decomposition of the time series in trend, daily and weekly seasonality, obtained using Seasonal-Trend decomposition with locally estimated scatter plot smoothing (STL with LOESS, Cleveland et al. 1990). Analyses are performed by R packages for time series (Hyndman, 2022). According to Fig. 5, the original time series (top chart) of inflows (Fig. 5a), outflows (Fig. 5b) and internal flows (Fig. 5c) is additively decomposed in trend (`trend`), daily seasonal pattern (`season_24`), weekly seasonal pattern (`season_168`) and a residual term (`remainder`). The heights of the grey bars help to quantify the importance of each component. Consistently on the three analyzed time series, it emerges that the daily pattern (`season_24`) is more important than the weekly pattern. However, it seems that both daily and weekly seasonalities deserve to be accounted for adequately modelling traffic flows. A confirmation comes from the decomposition of the total variance of the original series in trend, daily pattern, weekly pattern and remainder term (according to the STL additive decomposition). Table 1 shows that the two seasonal components, consistently over the three time series, sum to $\sim 95\%$ of the total variance.

Motivated by empirical evidence, in order to account for the dependence among the three processes and for the daily and weekly seasonalities in the processes themselves, the VARX model with DHR components as described in Sect. 3.1 will be adopted. Here, it is worth mention that possible multicollinearity issues related to jointly modeling such a similar time series through a vector autoregressive strategy are mitigated by the fact that such similarity, as demonstrated above, is mostly attributed to daily and weekly patterns that will not be modeled through the mutual interdependence among the time series themselves but through the Fourier bases of the DHR component. Moreover, similarities in the three time series attributed to the average levels by month and by weekday (see `trend` series in Fig. 5) will be captured by two set of dummies included among the exogenous variables. It follows that the vector autoregressive part will model the remainder component, which is supposed not to be extremely similar among the three time series.

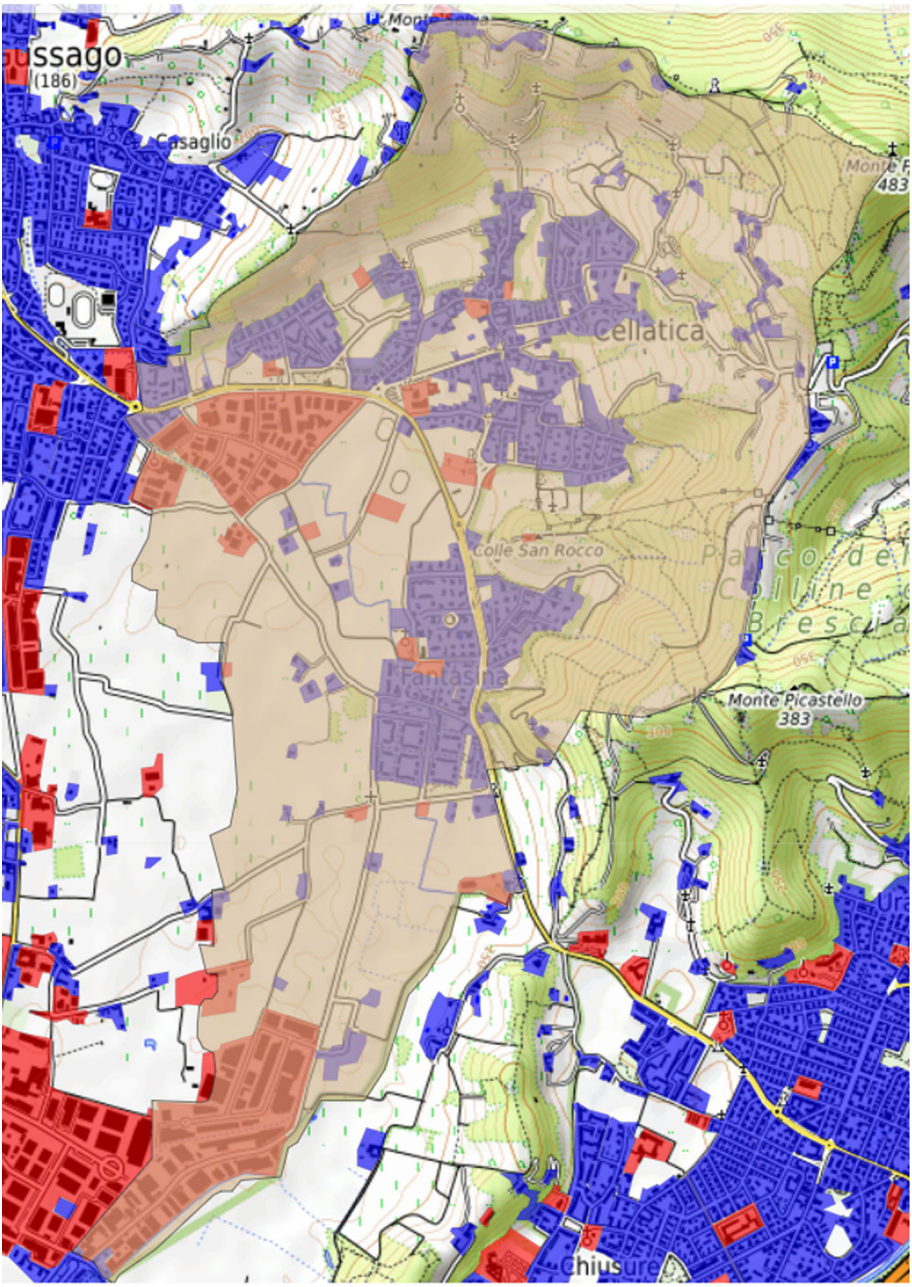
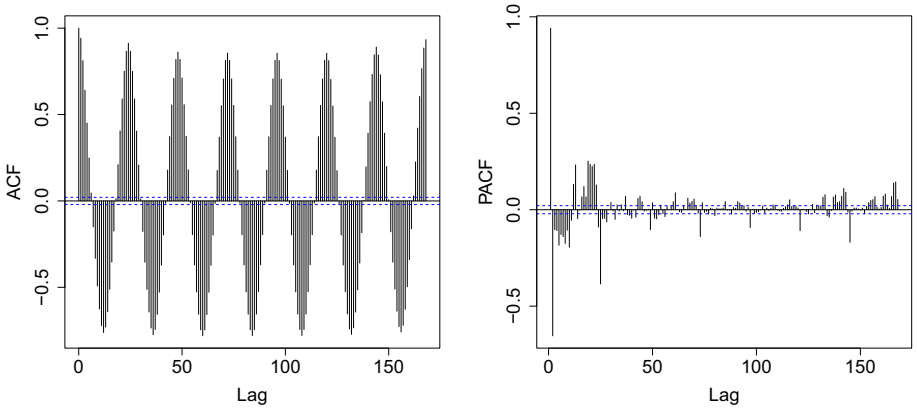
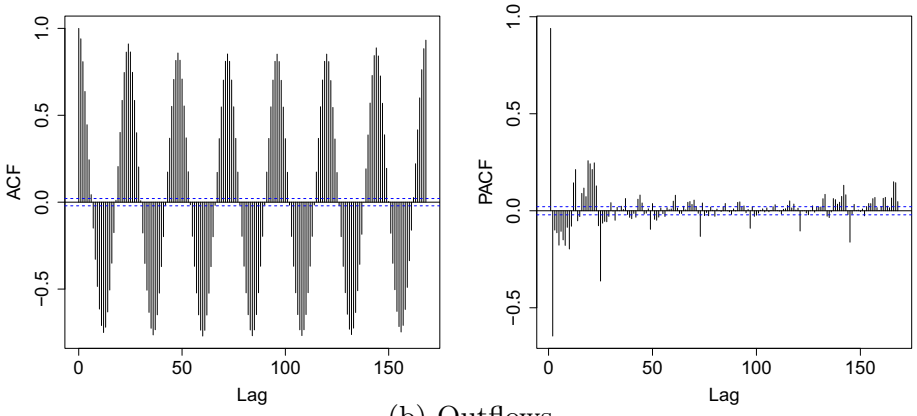


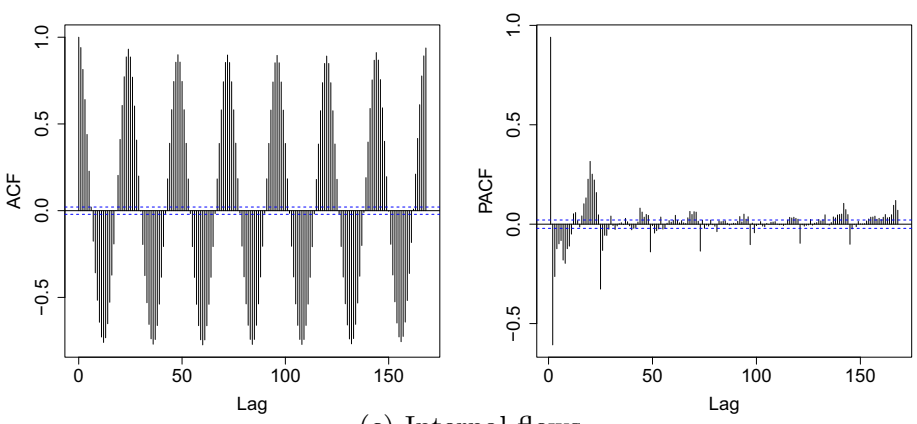
Fig. 3 Map of the ACE of Cellatica, by destination of use. Residential areas are depicted in blue, industrial areas in red. Source: Destinazione d'Uso dei Suoli Agricoli e Forestali (DUSAF) 6.0 2018, last update 14/09/2021. https://www.dati.lombardia.it/Territorio/Dusaf-6-0-Uso-del-suolo-2018/7rae-fng6



(a) Inflows.

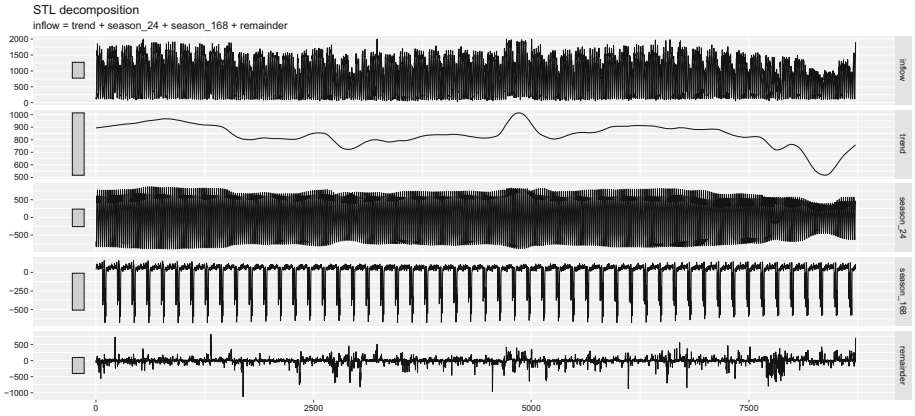


(b) Outflows.

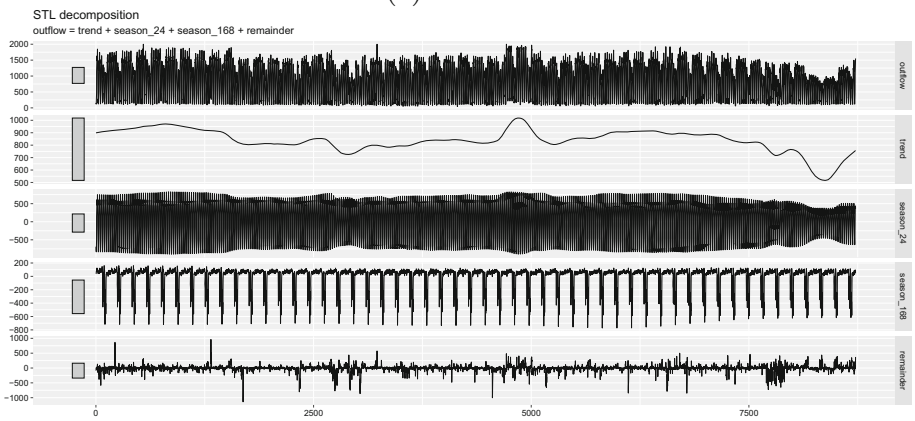


(c) Internal flows.

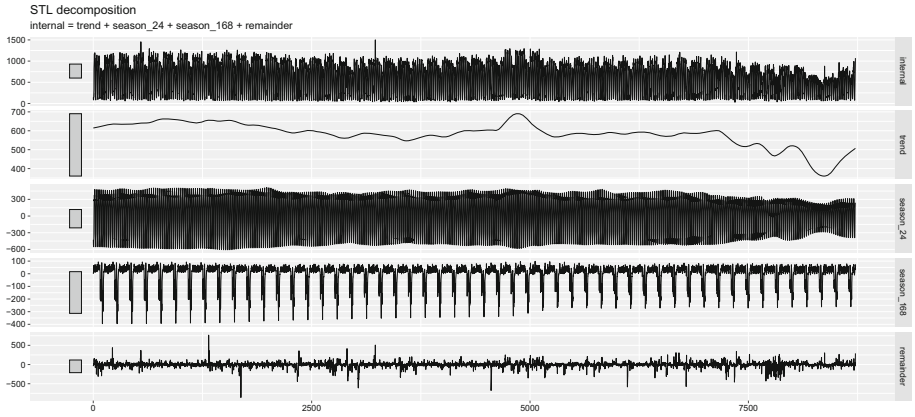
Fig. 4 ACF (left) and PACF (right) for time lags up to one week (168 hours). From top to bottom: Inflows, Outflows, Internal flows



(a) Inflows.



(b) Outflows.



(c) Internal flows.

Fig. 5 STL with LOESS, with 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

Table 1 Total variance decomposition according to STL decomposition (expressed in %)

Series	Trend	Season_24	Season_168	Remainder
Outflows	2.355	87.455	7.296	2.894
Inflows	2.282	87.818	7.107	2.792
Internal flows	2.734	90.016	4.291	2.956

3 Modeling strategy

As explained in the introduction, to model and to forecast traffic flows in areas with hydrogeological criticality, we adopt a new modeling approach that include in a Vector AutoRegressive with eXogenous variables (VARX) model a Dynamic Harmonic Regression (DHR) component, justified by the evidence on the temporal pattern on traffic flows' time series and on the similarity among inflows, outflows and internal flows displayed in Sect. 2.3.

The model is applied to hourly data but, due to the constraint to avoid using recent data that are unavailable, the lags of the AutoRegressive (AR) part cannot be those of the last 23 hours. They are instead either i) the values at the same hour of the previous day, or ii) the value at the same hour and the same weekday of the previous week.

3.1 Model specification

Vector AutoRegressive models (VAR, Hamilton 1994, chapters 11-12), and more in general multivariate time series models have been mainly developed and adopted for the analysis of financial data (Stock & Watson, 2001). In our context, the VAR model with eXogenous variables (VARX, Tsay 2014, chapter 6) seems to be adequate to describe the desired dependence structure where flows display strong correlation patterns. The general VARX model, for time $t = 1, ..., T$, is defined as follows:

$$\mathbf{y}_t = \boldsymbol{\nu} + \sum_{h=1}^{p} \mathbf{A}_h \mathbf{y}_{t-h} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\epsilon}_t \tag{1}$$

where $\mathbf{y}_t$ is the $m \times 1$ vector of dependent variables at time t, and $\mathbf{y}_{t-h}$ is the vector of dependent variables at time $t - h$, $\boldsymbol{\nu}$ is a constant vector of length m, $\mathbf{A}_h$ is an $m \times m$ matrix of coefficients to be estimated; $\mathbf{x}_t$ represents the $l \times 1$ vector of the l exogenous variables, where $\mathbf{B}$ is the related $m \times l$ matrix of coefficients. In our case, we model inflows, outflows and internal flows, so $m = 3$. Furthermore, by adopting a VARX model, we are assuming that the realizations of the three processes in vector $\mathbf{y}_t$ are related to each others.

To model the complex seasonality of traffic flows correlated processes, we employ a DHR structure inside the VARX model by considering in $\mathbf{x}_t$ the vector of selected sin and $cosine$ Fourier bases related to the daily and the weekly patterns. The DHR model (Hyndman & Athanasopoulos, 2018) is based on the principle that a combination of $sine$ and $cosine$ functions can approximate any periodic function related to seasonal patterns. Given y_t the realization of a stochastic process Y evaluated at time t, the DHR model is defined as:

$$y_t = \beta_0 + \sum_{k=1}^{K} [\alpha_k s_k(t) + \gamma_k c_k(t)] + \epsilon_t \tag{2}$$

where β_0 is a constant term, $s_k(t) = sin(\frac{2\pi kt}{m})$ and $c_k(t) = cos(\frac{2\pi kt}{m})$ are, respectively, the *sine* and *cosine* functions, with m being the seasonal period (e.g., if t are hours, the weekly period is $m = 24 \times 7 = 168$), α_k and γ_k, for $k = 1, ..., K$, are regression coefficients to be estimated and ϵ_t is the residual component, generally modeled as a standard $ARIMA$ process. The optimal model is the one with the lowest Akaike Information Criteria (AIC), so the process for choosing the optimal value of K is to start with $K = 1$ and increase K until the AIC is no longer decreasing, by considering the constraint that, by model construction, K cannot be greater than $m/2$.

Actually, DHR model may account for more than one periodic function. According to the preliminary evidence in Sect. 2.3, we want to account for both daily and weekly periodic functions, and we do so by using the following two summation terms:

1. $\sum_{k_d=1}^{K_d}[\alpha_{k_d} s_{k_d}(t) + \gamma_{k_d} c_{k_d}(t)]$, with $s_{k_d}(t) = sin(\frac{2\pi kt}{m_d})$ and $c_{k_d}(t) = cos(\frac{2\pi kt}{m_d})$, where $2 \times K_d$ is the number of parameters to be estimated for the daily pattern and $m_d = 24$ hours.
2. $\sum_{k_w=1}^{K_w}[\alpha_{k_w} s_{k_w}(t) + \gamma_{k_w} c_{k_w}(t)]$, with $s_{k_w}(t) = sin(\frac{2\pi kt}{m_w})$ and $c_{k_w}(t) = cos(\frac{2\pi kt}{m_w})$, where $2 \times K_w$ is the number of parameters to be estimated for the weekly pattern and $m_w = 24 \times 7 = 168$ hours.

Summarizing, we aim at modelling inflows, outflows and internal flows together by combining VARX with DHR approaches and by allowing the realizations of the three processes to affect one each other.

Since in our work the three processes $\mathbf{y}_t$ are represented by directed origin-destination flows evaluated at different hours' intervals, herein after we call its realization at time t from the $i-th$ ACE of origin to the $j-th$ ACE of destination $Flow_{ij,t}$. We furthermore call the realization at time t of the process related to internal flows $Flow_{ii,t}$.

The adopted model is defined as:

$$\mathbf{Flow}_t = \boldsymbol{\nu} + \sum_{h=1}^{p} \mathbf{A}_h \mathbf{Flow}_{t-h} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\epsilon}_t \tag{3}$$

where $\mathbf{Flow}_t = [Flow_{ji,t}, Flow_{ij,t}, Flow_{ii,t}]'$, $\mathbf{x}_t$ is the $2 \times (K_d + K_w)$ vector $[s_{1_d}(t), c_{1_d}(t), ..., s_{K_d}(t), c_{K_d}(t), s_{1_w}(t), c_{1_w}(t), ..., s_{K_w}(t), c_{K_w}(t)]'$ and $\mathbf{B}$ is the $3 \times [2 \times (K_d + K_w)]$ matrix of α and γ parameters in formula 2.

Here it is worth remarking that order p is not referring to hours. Due to data availability constrain, it is instead to be considered either as i) days, or ii) weeks.

Actually, in the application a proper set of dummies related to the month and to the day of the week is also included among the exogenous variables, as will be explained in Sect. 4.

3.2 Forecasting and performance evaluation

In this subsection, we describe the methodological strategy adopted for the performance evaluation. Specifically, the performance of the model needs to be evaluated in terms of the ability of forecasting the amount of traffic flows (both inflows and outflows and internal flows) in a specific year's day.

To this aim, we split the data in the training and the validation set by adopting a blocked k-folds cross validation strategy for time series (Hyndman & Athanasopoulos, 2018; Snijders, 1988), where k equals to one day (24 observations) and the training set if 2 months length (60 days). To validate the last 24 hours of the dataset (i.e., the day of August, 31th, 2021) we base on a training set made by the previous two months of data (i.e. from July, 1st, 2021

to August, 30th, 2021). In order to evaluate the performance consistently over all the days of the year, we decide to replicate the performance evaluation using different set of training and validation samples. With a rolling window procedure, to validate the day of August 30th, 2021 the training set refers to the data from June 30th, 2021 to August 29th, 2021, and so on and so forth. We use the validation set for performance evaluation purposes, by comparing predicted versus observed values. (i.e., for each fold). However, since i) the available data are related to exactly 1 year, ii) the training set is 2 month length, and iii) the model is based on lag values up to a certain numbers of weeks, more than 60 days can not be validated.

The RMSE is traditionally adopted for performance evaluation. A disadvantage of the RMSE is that it suffers for being dependent from the unit of measure. In this work, we adopt the MAPE (Hyndman & Koehler, 2006; Tofallis, 2015), which solves that problem by measuring the accuracy as a percentage. The MAPE can be calculated as the average absolute percent error between observed and predicted values using the following formula:

$$MAPE = 100 \times \frac{1}{n_v} \sum_{t=1}^{n_v} \frac{| Y_t - \hat{Y}_t |}{Y_t} \tag{4}$$

where n_v is the number of validation points, Y_t is the observed value at time t of inflows, outflows or internal flows, $\hat{Y}_t$ is the predicted value at time t of inflows, outflows or internal flows.

A second way to evaluate accuracy of prediction is based on HR, which is a metric for categorical data traditionally adopted for the classification of customers in credit scoring (see, e.g., Bensic et al. 2005). Simply speaking, HR is nothing more the ratio of times the values are correctly predicted by such a method. First both observed and predicted values of the validation set are assigned to the categories using distribution percentiles, then this ratio is computed:

$$HR = \frac{1}{n_v} \sum_{t=1}^{n_v} I(\, Y_t \text{ and } \hat{Y}_t \text{ belong to same category}) \tag{5}$$

where $I(c_t)$ is the indicator variable that is 1 if the condition c_t is true and 0 otherwise.

The interest in adopting such a measure arises if we consider as an alternative criteria for the model's performance the ability to predict whether the traffic level is high, moderate or low. In light of this, we will assign observed and predicted values to 5 categories, from "very high" to "very low" flows, using distribution quintiles, whereas a classification in 7 categories has also been tried for comparison purposes.

4 Results of the analysis

In this section, we show the application of our approach to the traffic flows related to the ACE of Cellatica (Fig. 3).

4.1 Modeling

To jointly modeling the traffic flows in both directions of the ACE of Cellatica (inflows and outflows) and the internal traffic flows, it is determinant to find the optimal number of the Fourier bases of the DHR components and then the optimal order for the AR term to be considered in the VARX model.

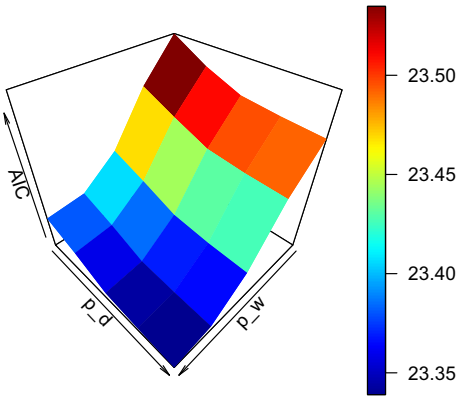


Fig. 6 Surface plot of the AIC of the VARX model with Fourier bases ($K_d = 7$, $K_w = 4$), month, weekday dummies, for $p_d = 1, 2, ..., 5$ and $p_w = 1, 2, ..., 5$

Operationally, first the optimal number of Fourier bases, $K_d = 7$ and $K_w = 4$, have been determined using the method suggested in previous studies for the case of single (Hyndman & Athanasopoulos, 2018) and multiple ([26], Sect. 3) seasonal pattern. We also include monthly dummies among exogenous variables to control for the possible presence of changes in average levels among months (e.g., higher traffic flows in March, as emerged by looking at the `trend` series of the STL decomposition in Fig. 5a–c) and weekdays dummies to control for the possible presence of changes in average levels among weekdays.

For the identification of the AR order, as explained in the introduction and in Sect. 3, we cannot use the latest 23 hours, so that here the lag of order 1, that with hourly data should be the previous hour, has two definitions: i) the value at the same hour of the previous day, ii) the value at the same hour and the same weekday of the previous week. An approach based on AIC and the "elbow" method has been applied to find the optimal orders p_d (the number of lags related to the same hour of previous days) and p_w (the number of lags related to the same hour of previous weeks). More precisely, by means of a surface plot we evaluate the AIC for each combination of p_d and p_w in the range of integer values from 1 to 5, based on a VARX model with Fourier bases $(K_d = 7, K_w = 4) = (7_d, 4_w)$, month and weekday dummies (see Fig. 6).

The surface plot highlights how, for $p_d > 3$ and $p_w > 4$, the AIC does not decrease significantly. So, the chosen model turns out to be a VARX $(p_d = 3, p_w = 4) = (3_d, 4_w)$ model with the DHR$(7_d, 4_w)$ components among exogenous variables.

Estimation results of the final VARX model with HDR components, computed over the full sample of 365 days where data for September, 2nd, 2020 (not available) have been replaced by the data for September, 9th, 2020 and performed in `R` based on `VARX` and `VARXpred` functions in `MTS` package (Tsay, 2014),[3] are reported in Table 2.

The PACF (see bottom left charts in Fig. 7a–c) displays the presence of significant autocorrelation of first order in estimated residuals. This might be due to the fact that first 23 lags are not allowed to be used. Moreover, the histogram of the residuals does not follow a normal distribution. Bottom right charts of Fig. 7a–c highlight that residuals exhibit a leptokurtic distribution with heavy tails. This is the price to pay for does not including traditional lag terms (i.e., the previous hours) in the model.[4] Stationarity on the time series of estimated

[3] We have assumed that traffic flows for these two days are similar.

[4] As a robustness check, the same VARX model has been also estimated replacing the dependent variables with a Box-Cox power transformation of original values via maximum likelihood estimation (Box & Cox, 1964). Residuals do not show a significant improvement in the adherence to normal distribution (results are

Table 2 Results of the VARX($3_d, 4_w$) with a DHR($7_d, 4_w$) component

Endogenous variable	Inflow (s.e.)	Outflow (s.e.)	Internal flow (s.e.)
Inflow_AR(1)_day	0.192 (0.056)	0.063 (0.057)	− 0.009 (0.037)
Inflow_AR(2)_day	− 0.051 (0.057)	− 0.096 (0.058)	− 0.036 (0.036)
Inflow_AR(3)_day	0.028 (0.057)	− 0.003 (0.057)	− 0.039 (0.036)
Outflow_AR(1)_day	0.109 (0.056)	0.230 (0.056)	0.035 (0.036)
Outflow_AR(2)_day	− 0.072 (0.057)	− 0.032 (0.057)	− 0.058 (0.036)
Outflow_AR(3)_day	−0.006 (0.056)	0.024 (0.056)	− 0.006 (0.035)
Internal_flow_AR(1)_day	0.035 (0.032)	0.042 (0.032)	0.258 (0.021)
Internal_flow_AR(2)_day	0.160 (0.033)	0.170 (0.033)	0.165 (0.020)
Internal_flow_AR(3)_day	0.133 (0.032)	0.133 (0.033)	0.176 (0.020)
Inflow_AR(1)_week	0.294 (0.059)	0.136 (0.060)	0.099 (0.037)
Inflow_AR(2)_week	− 0.017 (0.058)	− 0.133 (0.059)	− 0.016 (0.037)
Inflow_AR(3)_week	0.131 (0.059)	0.048 (0.060)	0.039 (0.037)
Inflow_AR(4)_week	0.168 (0.059)	0.084 (0.059)	0.073 (0.037)
Outflow_AR(1)_week	0.012 (0.059)	0.171 (0.059)	− 0.044 (0.037)
Outflow_AR(2)_week	0.070 (0.058)	0.186 (0.059)	0.008 (0.037)
Outflow_AR(3)_week	0.120 (0.058)	0.207 (0.059)	0.050 (0.037)
Outflow_AR(4)_week	0.077 (0.059)	0.160 (0.060)	0.022 (0.037)
Internal_flow_AR(1)_week	− 0.067 (0.032)	− 0.069 (0.033)	0.145 (0.021)
Internal_flow_AR(2)_week	− 0.058 (0.031)	− 0.061 (0.032)	0.046 (0.020)
Internal_flow_AR(3)_week	− 0.217 (0.032)	− 0.221 (0.031)	− 0.044 (0.020)
Internal_flow_AR(4)_week	− 0.174 (0.032)	− 0.168 (0.032)	− 0.041 (0.020)
Exogenous variable	Outflow (s.e.)	Inflow (s.e.)	Internal flow (s.e.)
Sin_day_1	− 45.439 (7.928)	− 37.692 (7.990)	− 47.087 (5.015)
Cos_day_1	− 25.347 (7.432)	− 29.108 (7.490)	− 27.341 (4.701)
Sin_day_2	− 9.508 (2.976)	− 11.809 (3.000)	− 14.312 (1.883)
Cos_day_2	5.004 (2.232)	3.044 (2.249)	5.749 (1.412)
Sin_day_3	4.342 (2.579)	5.511 (2.600)	4.179 (1.632)
Cos_day_3	− 11.551 (3.188)	− 4.073 (3.213)	− 7.625 (2.016)
Sin_day_4	− 0.254 (2.086)	− 2.125 (2.103)	− 3.108 (1.320)
Cos_day_4	1.540 (2.473)	− 2.723 (2.493)	− 3.912 (1.565)
Sin_day_5	0.166 (2.086)	0.348 (2.103)	1.962 (1.320)
Cos_day_5	0.327 (1.956)	− 0.486 (1.971)	− 0.159 (1.237)
Sin_day_6	1.029 (1.941)	0.662 (1.956)	0.839 (1.228)
Cos_day_6	− 1.805 (1.969)	− 0.736 (1.985)	− 1.310 (1.246)
Sin_day_7	− 1.472 (2.023)	0.425 (2.039)	− 0.010 (1.280)
Cos_day_7	− 1.026 (1.969)	− 1.361 (1.985)	− 1.143 (1.246)
Sin_week_1	41.318 (2.358)	41.137 (2.377)	22.975 (1.492)
Cos_week_1	5.778 (2.347)	6.137 (2.366)	5.785 (1.485)
Sin_week_2	− 35.458 (2.324)	− 35.469 (2.342)	− 22.282 (1.470)

Table 2 continued

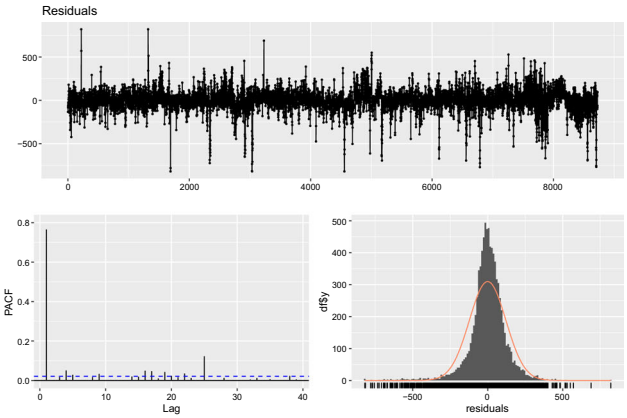
Exogenous variable	Outflow (s.e.)	Inflow (s.e.)	Internal flow (s.e.)
Cos_week_2	10.698 (2.177)	9.993 (2.195)	2.815 (1.377)
Sin_week_3	10.185 (1.978)	10.820 (1.994)	6.417 (1.251)
Cos_week_3	− 23.977 (2.113)	− 23.635 (2.129)	− 10.739 (1.337)
Sin_week_4	16.133 (1.985)	15.494 (2.000)	9.308 (1.256)
Cos_week_4	18.184 (2.069)	18.348 (2.086)	7.757 (1.309)
Month (ref. January): February	43.963 (6.942)	44.745 (6.997)	3.117 (4.392)
March	56.449 (6.775)	56.982 (6.829)	23.425 (4.286)
April	7.840 (6.807)	8.286 (6.861)	− 0.302 (4.306)
May	43.830 (7.116)	44.526 (7.172)	7.260 (4.501)
June	65.148 (7.097)	66.822 (7.153)	21.094 (4.490)
July	68.355 (7.234)	69.637 (7.291)	9.575 (4.576)
August	− 32.542 (7.210)	− 33.130 (7.267)	− 47.680 (4.561)
September	42.524 (7.544)	44.453 (7.604)	11.643 (4.773)
October	102.004 (7.127)	103.378 (7.183)	38.790 (4.508)
November	54.185 (6.947)	54.458 (7.002)	31.033 (4.394)
December	71.030 (6.794)	71.378 (6.847)	27.408 (4.298)
Weekday (ref. Monday): Tuesday	63.740 (26.874)	68.238 (27.087)	31.087 (17.001)
Wednesday	11.398 (26.888)	17.867 (27.101)	− 13.700 (17.009)
Thursday	11.876 (26.920)	23.314 (27.133)	− 10.966 (17.029)
Friday	9.321 (26.875)	27.938 (27.088)	− 27.546 (17.090)
Saturday	− 64.434 (26.842)	− 46.322 (27.055)	− 59.964 (16.980)
Sunday	− 30.698 (26.858)	− 22.364 (27.071)	− 20.947 (16.990)
Intercept	19.612 (13.058)	19.244 (13.161)	61.881 (8.260)
Residual correlation matrix	Outflow	Inflow	Internal flow
Outflow	1	0.983	0.858
Inflow	0.983	1	0.856
Internal flow	0.858	0.856	1
Information criteria	AIC: 23.488	BIC: 23.636	

residuals has been checked by means of the Dickey Fuller (DF) test (Dickey & Fuller, 1979) for unit roots. Since our regression set-up contemplates lags from the order 24, the DF test is employed based on the beta coefficient of the following regression:

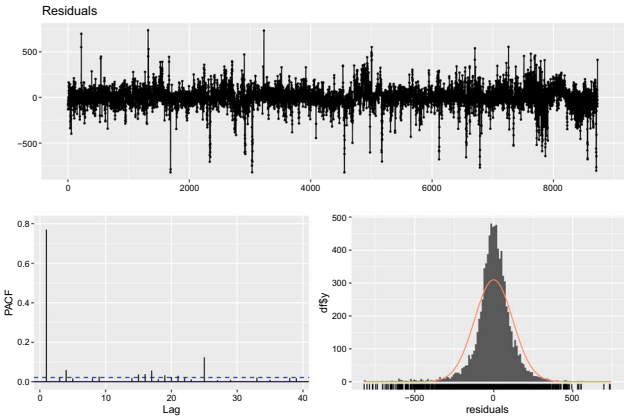
$$\delta Y_t = \alpha + \beta Y_{t-24} + \upsilon_t.$$

The null hypothesis of unit roots is rejected for all the three time-series of estimated residuals. It means the time series of estimated residuals are stationary.

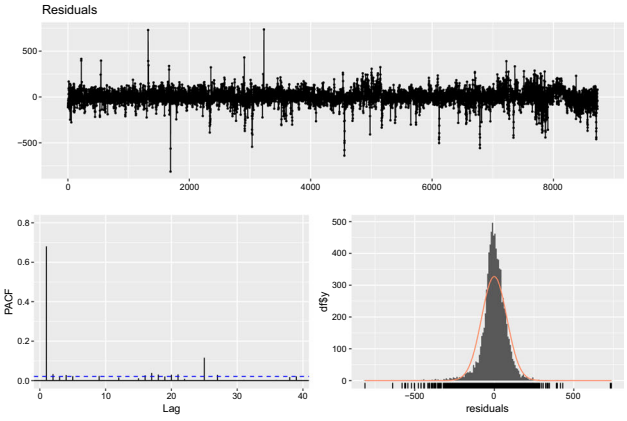
Footnote 4 continued
available upon request) and do not justify the use of a Box-Cox transformation on original data. We have also added the lags 25 (the previous hour of the previous day) and 169 (the previous hour of the same day of the previous week) in the model, but again no significant improvements emerge.



(a) Inflows' VARX estimated residuals.



(b) Outflows' VARX estimated residuals.



(c) Internal flows' VARX estimated residuals.

Fig. 7 Residuals' diagnostic: Time series of estimated residuals (top), PACF with 95% confidence bounds for strict white noise (bottom left), histogram of the empirical distribution with normal curve (bottom right)

4.2 Performance evaluation

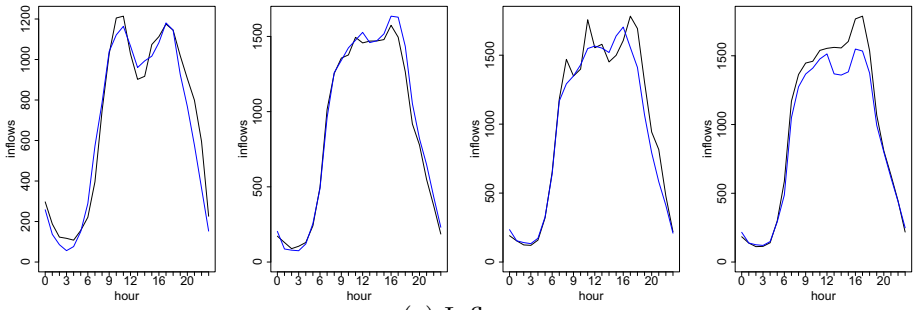
To evaluate prediction performance, we apply the blocked k-folds cross validation strategy for time series described in Sect. 3.2, which is based on splitting the observed time series in a training set and a validation set, by taking as training set the previous two months of data and as validation set the considered day itself. The total number of generated different training and validation set is 277, because for 88 days the validation has not been possible: i) 60 days are excluded because used for the training set, ii) 28 additional days have been used as lag terms in the vector autoregressive model (remember that the order chosen for p_w was 4). As a result, each training set has sample size $n_t = 1440$ (24 daily observations in each of the 60 days) and each validation set has sample size $n_v = 24$.

Schematically, to evaluate the forecasting accuracy we proceed, using the two indices defined in Sect. 3.2, as it follows:

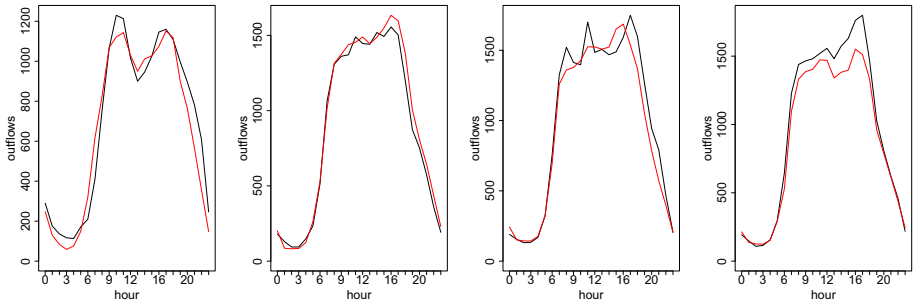
1. first, we replace holidays' values with the values of the same hour of a different day using an appropriate criterion that will be explained below. We do replacement for these days because traffic on holidays likely follows a particular dynamic. We refer to them as anomalous days.
2. in a second step, on the dataset where anomalous days have been appropriately replaced we analyze by means of the MAPE the performance of our method in correctly predicting the amount of traffic in the validation days. In doing so, an analysis by month is shown to demonstrate that the accuracy of the prediction is similar in different time periods.
3. In the third and most relevant step, true and predicted values are classified in 5 categories using distributional quintiles and the HR is computed to evaluate how good is the model to predict whether the traffic is very high, high, moderate, low or very low.

According to the first step, for the holidays we have considered as anomalous the following days: January, 1st and 6th (Epiphany), April, 4th (Easter) and 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. We have replaced the values of the inflows, the outflows and the internal flows with the values of the same weekday of the previous week (under the assumption that the replaced one is a "normal" day). Actually, some of these days are replaced with the nearest same weekday that is not an anomalous day. That is the case of, e.g., January 1st, whose same weekday of the previous week was another anomalous day (December, 25th).

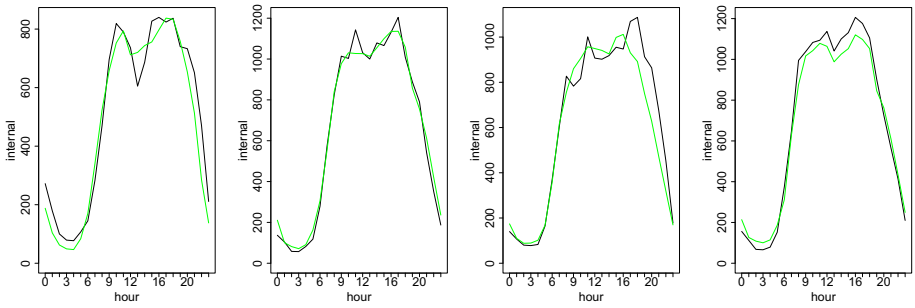
According to the second step, MAPE has been computed for all the available validation days, using the dataset where the holidays have been replaced and using the predicted values determined by the chosen VARX$(3_d, 4_w)$ model with DHR$(7_d, 4_w)$ components. As an example of forecast, Fig. 8a–c show, respectively from top to bottom, the time series of observed (black) versus predicted (colored) inflows, outflows and internal flows - where outflows represent the amount of traffic exiting Cellatica, inflows the amount of traffic entering Cellatica, internal flows the amount of traffic within Cellatica - for the validation of four representative days (from left to right): February, 15th (Monday), February, 17th (Wednesday), February, 19th (Friday), February, 21st (Sunday). According to these validation days, the MAPE (computed according to Eq. 4) varies from a minimum of 4.8 (Wednesday) to a maximum of 24.3 (Monday) for outflows, from a minimum of 3.8 (Wednesday) to a maximum of 23.5 (Monday) for inflows and from a minimum of 7.4 (Wednesday) to a maximum of 23.5 (Monday) for internal flows. Here it is worth considering that the MAPE ranges from 0 to 100, where values lower than 10 are generally associated to a very good performance and values between 10 and 20 to fairly good performance.



(a) Inflows.



(b) Outflows.



(c) Internal flows.

Fig. 8 Plot of observed (black) versus predicted (colored) traffic flow in Cellatica. Validation days (from left to right): February, 15th (Monday), February, 17th (Wednesday), February, 19th (Friday), February, 21st (Sunday) 2021

Mean and standard deviation of the MAPE by month have been computed for both inflows, outflows and internal flows. Table 3 shows that, according to MAPE, the performance of our prediction method is similar between different months. Excluding November, which average values are determined by three days only, we range from a minimum of about 13 for the month of May (better predicting performance) to a maximum of 22/25 in the month of July (worst prediction performance). The month of July also presents the larger standard deviations. Despite it emerge that our model presents a worst prediction performance in summer days and in the days of January, the difference with respect to the other months is not large.

Table 3 Mean and standard deviation (inside brackets) of MAPE, by month (The month of November relies on 3 days only)

Series	January	February	March	April	May	June	July	August	November	December
Inflows	20.89	14.47	16.10	17.83	13.74	14.69	22.89	20.07	9.83	16.12
	(11.74)	(8.49)	(11.33)	(9.62)	(9.42)	(6.50)	(14.76)	(8.67)	(3.68)	(7.50)
Outflows	21.25	14.81	16.27	17.88	13.95	14.77	23.23	20.39	10.14	16.13
	(11.88)	(8.43)	(12.21)	(9.78)	(9.53)	(6.59)	(15.09)	(9.10)	(4.90)	(7.67)
Internal	17.90	14.70	13.64	15.58	13.03	15.31	25.41	17.68	11.16	13.39
	(10.09)	(9.91)	(8.93)	(7.29)	(8.47)	(6.10)	(16.49)	(7.50)	(0.44)	(5.35)

As a third step, we compute the Hit-Rate on the available validation days after having turned true and predicted values in five categories according to quintiles.[5] It is worth recalling that this step is the core of the validation strategy, because, for early warnings purposes, one is not interested to know the exact amount of traffic, but whether traffic is higher than a threshold. The performance must be evaluated in terms of the ability to predict whether the amount of traffic is on the correct category interval.

Mean and standard deviation of the HR by month have been computed for both inflows, outflows and internal flows. Table 4 shows that, according to HR, the performance of our prediction method is similar between different months. Excluding November, which average values are determined by three days only, average HR values range from a minimum of 0.70 for internal flows in July to a maximum of 0.87 for inflows in the month of November. Standard deviations are also quite similar each others. Similar to the results of MAPE in Table 3, we found the worst prediction performance for summer days. We can also notice that prediction performance for internal flows is slightly worst. However, overall the predicting performance in terms of HR is stable over the time period. Recalling that the HR ranges from 0 to 1 (where 1 corresponds to a perfect match between observed and predicted values), overall, results prove a good accuracy in predicting whether the amount of traffic flows in flooding risk area will fall into one of the 5 categories.

Histograms in Fig. 9a–c show the distribution of HR over the set of available validation days, respectively, for inflows, outflows and internal flows. For all cases, distributions show a negative skewness with high frequencies in the range [0.80, 0.95] for the case of inflows and outflows and in the range [0.70, 0.95] for the case of internal flows.

In all cases, the median of the HR is larger than 0.8, while very few are the validation days reporting an HR lower than the lower whisker (which is approximately 0.6).

Box plots in Fig. 10a–c, where the extremes of the box represent the first and the third quartiles (Q_1 and Q_3), the horizontal line is the second quartile (Q_2) and whiskers corresponds to, respectively, $Q_1 - 1.5 \times (Q_3 - Q_1)$ and $Q_3 + 1.5 \times (Q_3 - Q_1)$, display the distribution of the HR computed, respectively, on inflows, outflows and on internal flows. Further anomalous days are labeled in blue and corresponds to those days for which the performance is not good. Among those new anomalous days, no holidays are detected.

An alternative replacement strategy based on Exponential Weighted Moving Average (EWMA) (Hunter, 1986) has been applied. Actually, we used a variation of the EWMA where $t - 1$ is replaced by $t - 24$ (in order to just consider the values of the same hour of the day in the weighting scheme) according to the formula:

$$EWMA_t = \alpha Y_t + (1 - \alpha)(EWMA)_{t-24},$$

where $\alpha = \frac{2}{n-1}$ and n represents the backward window length, that we set to $n = 5$.

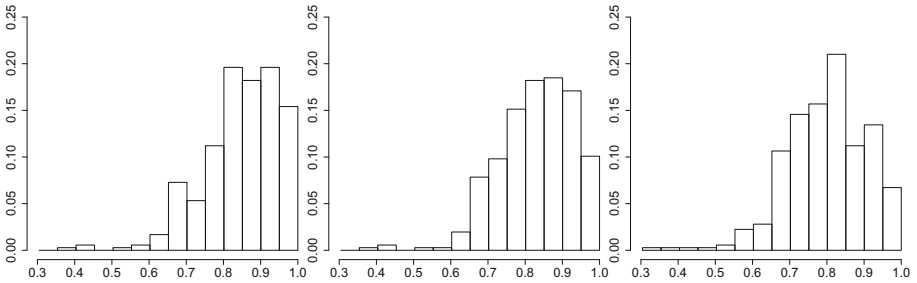
Histograms in Fig. 11a–c display the distribution of the Hit-Rate computed after having replaced holidays with our EWMA strategy. By doing such a way, the accuracy of our model does not change significantly though a larger frequency of HR values (for outflows and inflows) is found for the bin corresponding to the interval [0.9, 0.95]. According to the boxplots in Fig. 12a–c, we can notice the presence of the same few anomalous days (compared to the other replacement method).

It is worth observing that, regardless of the replacing method adopted, in this step of analysis the anomalous days (labeled in blue) are just few.

[5] As a check, a classification into seven categories has also been tried and results appear really similar to the classification in five categories.

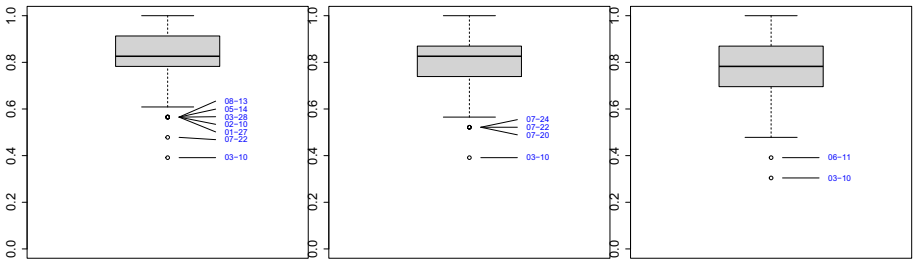
Table 4 Mean and standard deviation (inside brackets) of HR, by month (The month of November relies on 3 days only)

Series	January	February	March	April	May	June	July	August	November	December
Inflows	0.83	0.82	0.83	0.85	0.83	0.85	0.78	0.79	0.87	0.84
	(0.09)	(0.10)	(0.14)	(0.08)	(0.09)	(0.10)	(0.12)	(0.08)	(0.13)	(0.08)
Outflows	0.82	0.81	0.83	0.81	0.82	0.83	0.76	0.79	0.82	0.80
	(0.10)	(0.09)	(0.13)	(0.10)	(0.09)	(0.09)	(0.11)	(0.08)	(0.04)	(0.08)
Internal	0.78	0.79	0.76	0.82	0.78	0.74	0.70	0.77	0.85	0.84
	(0.08)	(0.08)	(0.13)	(0.07)	(0.09)	(0.11)	(0.11)	(0.09)	(0.15)	(0.08)



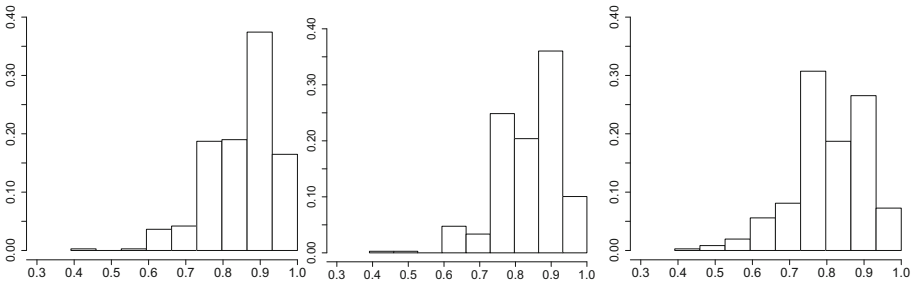
(a) Inflows. (b) Outflows. (c) Internal flows.

Fig. 9 Histogram of the distribution of the HR computed over the 277 days of the dataset, where holidays are replaced



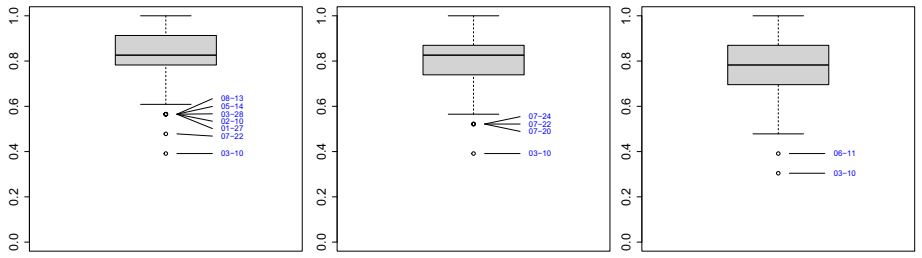
(a) Inflows. (b) Outflows. (c) Internal flows.

Fig. 10 Box plot of the distribution of the HR computed over the 277 days of the dataset, where holidays are replaced. Further anomalous days labeled in blue



(a) Inflows. (b) Outflows. (c) Internal flows.

Fig. 11 Histogram of the distribution of the HR computed over the 277 days of the dataset, where holidays are replaced with the EWMA strategy



(a) Inflows. (b) Outflows. (c) Internal flows.

Fig. 12 Box plot of the distribution of the HR computed over the 277 days of the dataset, where holidays are replaced. Further anomalous days labeled in blue

5 Concluding remarks

Flooding risk exposure maps traditionally use only administrative and census data, that allow static flooding risk exposures mapping, where the amount of presences is constant over time, although crowding is a highly dynamic process in metropolitan areas.

Today real-time monitoring of people and vehicles presences is a relevant aspect for "smart" cities, especially when they could be subjected to flooding risk or other natural disasters. Public decision-makers can take advantage of the availability of data-driven systems based on telecom data that allow to monitor hydrogeological risk areas and that can be used for real-time predictive purposes to deal with future emergency situations.

To account for the complex dynamic of traffic flows, in this study we have made use of the cutting-edge data coming from the mobile phone network related to origin-destination signals on traffic flows and available hourly basis from September 2020 to August 2021.

With the aim of forecasting the exposure risk and thus to make outbreaks' early detection and warning to who is transiting through that area, we have proposed to model the complex seasonality by adopting a novel methodological strategy based on combining Vector AutoRegressive with eXogenous variables (VARX) and Dynamic Harmonic Regression (DHR) models.

By adopting a robust cross validation strategy along with the Mean Absolute Percentage Error (MAPE) and the Hit-Rate (HR) indices on a case study represented by the area of Cellatica in the Mandolossa region, a critical zone with flood episodes located in the north-west outskirt of Brescia:

1. we came to prove that our method presents a good performance in predicting whether traffic in flooding risk areas is "very high", "high", "moderate", "low", "very low".
2. we are able to detect those year's days for which the traffic forecast based on our method is not good enough.

The poor prediction on anomalous days is the price to pay for the technical limitation given by not including the immediately previous hours lag terms in the model. Actually, we would like to highlight that this limitation may turn out to an advantage. In fact, we prove that a fairly good predicting performance on traffic flows may be obtained even if previous hours' information is not accounted for.

The useful results obtained in this study are promising, and will be used to further develop the main Project MoSoRe@UniBS 2020-2022 (see Acknowledgement). Possible future directions may regard: i) applying the method to other flooding risk areas, ii) refining the considered areas interested in flooding risk. According to point ii), a solution may

be to match original traffic flows data with additional sources. For example, Minimization Drive Test (MDT) technology data (Baumann, 1996), which present an excellent accuracy in terms of geolocalization, might be matched with origin-destination data to allow inferring the portion of traffic related to the road network.

Acknowledgements The authors would like to thank Antonio Naimoli (University of Salerno), Roberto Ranzi (University of Brescia), Ilia Negri, Emilio Zanetti Chini and Sergio Ortobelli Lozza (University of Bergamo) for their precious feedback. This work has been made in the framework of the MoSoRe@UniBS (Infrastrutture e servizi per la Mobilità Sostenibile e Resiliente) Project of Lombardy Region, Italy (CallHub ID 1180965; bit.ly/2Xh2Nfr).

Funding Open access funding provided by Università degli studi di Bergamo within the CRUI-CARE Agreement.

Declarations

Conflict of interest No competing interests to declare.

References

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology, 22*(1), 3–21.

Balistrocchi, M., Metulini, R., Carpita, M., & Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data. *Natural Hazards and Earth System Sciences, 20*(12), 3485–3500.

Baumann, D. (1996). Minimization of drive tests (MDT) in mobile communication networks. *Proceeding zum Seminar Future Internet (FI) und Innovative Internet Technologien und Mobilkommunikation (IITM), 9*, 1–7.

Benevolo, C., Dameri, R. P., & D'auria, B. (2016). Smart mobility in smart city. *Empowering organizations* (pp. 13–28). Springer.

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management: International Journal, 13*(3), 133–150.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis, 120*, 70–83.

Bibri, S. E., & Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society, 31*, 183–212.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211–243.

Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology, 18*(2), 65–82.

Carpita, M., & Simonetto, A. (2014). 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.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–33.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431.

Farrington, C. P., Andrews, N. J., Beale, A. D., & Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159*(3), 547–563.

Guo, J., Peng, Y., Peng, X., Chen, Q., Yu, J., & Dai, Y. (2009). Traffic forecasting for mobile networks with multiplicative seasonal arima models. In 2009 9th international conference on electronic measurement and instruments, pp. 3.377-3.380. IEEE.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hunter, J. S. (1986). The exponentially weighted moving average. *Journal of Quality Technology, 18*(4), 203–210.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice OTexts.

Hyndman, R. J. (2022). CRAN task view: Time series analysis.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18) . Springer.

Kong, X., Yang, J., Qiu, J., Zhang, Q., Chen, X., Wang, M., & Jiang, S. (2022). Post-event flood mapping for road networks using taxi GPS data. *Journal of Flood Risk Management, 15*(2), e12799.

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

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1–3), 159–178.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303.

Metulini, R., & Carpita, M. (2021). A spatio-temporal indicator for city users based on mobile phone signals and administrative data. *Social Indicators Research, 156*(2), 761–781.

Metulini R., & Carpita, M. Forecasting traffic flows with complex seasonality using mobile phone data. In R. Lombardo, I. Camminatiello, V. Simonacci (Eds.), *Book of Short Papers IES 2022: Innovation and Society 5.0: Statistical and Economic Metodologies for Quality Assessment* pp. 38–43, ISBN: 978-88-94593-36-5.

Mishra, D., Kumar, S., & Hassini, E. (2019). Current trends in disaster management simulation modelling research. *Annals of Operations Research, 283*(1), 1387–1411.

Snijders, T. (1988). On cross-validation for predictor evaluation in time series. On model uncertainty and its statistical implications, (pp. 56–69), Springer.

Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives, 15*(4), 101–115.

Tettamanti, T., & Varga, I. (2014). Mobile phone location area based traffic flow estimation in urban road traffic. *Columbia International Publishing, Advances in Civil and Environmental Engineering, 1*(1), 1–15.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society, 66*(8), 1352–1362.

Tran, Q. T., Ma, Z., Li, H., Hao, L., & Trinh, Q. K. (2015). A multiplicative seasonal ARIMA/GARCH model in EVN traffic prediction. *International Journal of Communications, Network and System Sciences, 8*, 43–49.

Tran, Q. T., Hao, L., & Trinh, Q. K. (2016). A novel procedure to model and forecast mobile communication traffic by ARIMA/GARCH combination models. *Advances in Computer Science Research, 58*, 29–34.

Tsay, R. S. (2014). *Analysis, multivariate time series, with R and financial applications*. John Wiley and Sons.

Yoneoka, D., Kawashima, T., Makiyama, K., Tanoue, Y., Nomura, S., & Eguchi, A. (2021). Geographically weighted generalized Farrington algorithm for rapid outbreak detection over short data accumulation periods. *Statistics in Medicine, 40*(28), 6277–6294.