Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-024-02735-x

ORIGINAL PAPER

Traffic flows time series in a flood-prone area: modeling and clustering on extreme values with a spatial constraint

Maurizio Carpita[1] · Giovanni De Luca[2] · Rodolfo Metulini[3] · Paola Zuccolotto[1]

Accepted: 22 April 2024

Abstract
Time series of traffic flows, extracted from mobile phone origin–destination data, are employed for monitoring people crowding and mobility in areas subject to flooding risk. By applying a vector autoregressive model with exogenous covariates combined with dynamic harmonic regression to such time series, we detected the presence of many extreme events in the residuals, which exhibit heavy-tailed distribution. For this reason, we propose a time series clustering procedure based on tail dependence which is suitable for data characterized by a spatial dimension, since objects' geographical proximity is taken into account. The final aim is to obtain clusters of areas characterized by the common tendency to the manifestation of extreme events, that in this case study are represented by extremely high incoming traffic flows. The proposed method is applied to the Mandolossa, a strongly urbanized area located on the western outskirts of Brescia (northern Italy) which is subject to frequent flooding.

Keywords Traffic flows modelling · Spatial time series clustering · Copula functions · Tail dependence · Spatial proximity · Mobile phone data

1 Introduction

It is widely acknowledged that extreme weather events often entail significant societal implications for communities and individuals. Immediate repercussions include the loss of human lives, destruction of crops, damage to property, and a decline in overall health and economic well-being. Given their substantial social and economic impact, the statistical analysis of extreme weather phenomena can also be approached from a managerial viewpoint. In fact, natural Disaster Management (Mishra et al. 2019) recommends the development of a framework of exposure risk that can exploited in an early warning perspective. In this study, we concentrate on floods. The creation of exposure maps for flooding risk is of paramount importance to effectively address such events. Exposure maps cannot ignore the temporal dynamic of human presence and people mobility. However, traditionally, such maps assume constant crowding over time. This assumption deviates considerably from reality, particularly within metropolitan regions. Therefore, providing a more comprehensive depiction of human presence and mobility is of paramount significance when we aim at assessing the possible consequences of flooding. So, in this paper we model traffic flows, as their in-depth understanding constitutes a fundamental element for building flooding risk maps (and, ultimately, for risk maps of any natural disaster). On the other hand, the use of our model in the construction of risk exposure maps is beyond the aims of this work.

To address this, modern sources of mobile phone data are increasingly integrated with satellite and sensor technologies, as exemplified by Pucci et al. (2022), to assess both crowding (Metulini and Carpita 2021) and dynamic

✉ Rodolfo Metulini
rodolfo.metulini@unibg.it

Maurizio Carpita
maurizio.carpita@unibs.it

Giovanni De Luca
giovanni.deluca@uniparthenope.it

Paola Zuccolotto
paola.zuccolotto@unibs.it

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

[2] Department of Management and Quantitative Studies, University of Naples Parthenope, Via G. Parisi, 13, 80132 Naples, Italy

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

Published online: 25 June 2024

movements (Tettamanti and Varga 2014) in urban areas. Data derived from mobile phone networks have demonstrated their pivotal role in the examination of subjects of considerable importance, such as the social and cultural events' surveillance (Carpita and Simonetto 2014) and the variability in the spatial distribution of human presence in the neighborhoods of large cities (Mariotti et al. 2022). Specifically in the context of flood risk, mobile phone data are employed to achieve dynamic monitoring of population density in regions susceptible to hydrogeological vulnerabilities (Balistrocchi et al. 2020).

Another application of mobile phone data in the realm of traffic flow involves the utilization of mobile phone origin–destination data to extract information regarding traffic patterns. This information is then used to construct statistical models capable of providing accurate predictions of human mobility. Metulini and Carpita (2023) proposed a model that combines vector autoregressive techniques with exogenous covariates and dynamic harmonic regression for this purpose. The application of this method was demonstrated in the case study of Mandolossa, an urbanized area prone to flooding situated on the western outskirts of Brescia (See Balistrocchi et al. (2020) for more details about the area). The study utilized hourly data spanning from September 2020 to August 2021, focusing on traffic flows to and from the municipality of Cellatica. While the model performed well, it is noteworthy that residuals displayed a leptokurtic distribution characterized by heavy tails, primarily determined by a series of extreme events (i.e., days with exceptionally high or low traffic flows). Based on these observations, we recognize the need to propose an analytical approach that takes into account extreme events, which appear as a structural characteristic of the analyzed phenomenon. So, in this work, we resort to the idea of time series clustering based on tail dependence coefficients estimated by copula functions, proposed by De Luca and Zuccolotto (2011). Specifically, we cluster the residuals' time series with respect to their upper tail dependence, because in this context we are interested in the extremely high traffic flows. To take into account the spatial structure of the analyzed setting, we propose a modification of the basic algorithm with the introduction of a spatial proximity coefficient, whose impact is tuned thanks to an iterative procedure. The analysis aims to identify clusters of regions in which extreme events (i.e., extremely high traffic flows) demonstrate a tendency to co-occur.

From a methodological point of view, the novelty of this paper lies in two main points: (*i*) the nontrivial combination of techniques for multivariate time series modeling (both the VARX model and the tail dependence estimation with copula functions) and clustering, and (*ii*) the proposal of a clustering procedure able to account for tail dependence and spatial proximity. From an empirical point of view, a valuable issue is the use of mobile phone data for a Natural Disaster Management purpose.

In Sect. 2 we introduce the mobile phone data and the data processing strategy adopted to obtain the traffic flows in the flood-prone area. Section 3 describes the model adopted to estimate the residuals, which are then further analyzed with the proposed time series clustering procedure, based on upper tail dependence with spatial proximity, which is presented in Sect. 4. Section 5 is dedicated to the application of the methodological strategy to the case of the Mandolossa region. Section 6 concludes the paper.

2 Data and data processing

2.1 Mobile phone data

In our study, we integrate two distinct categories of mobile phone data, specifically the Origin–Destination (OD) data and the Minimization Drive Test (MDT) technology data, which have been generously provided by TIM, the largest telecommunications operator in Italy.

Regarding the OD data, we possess a comprehensive dataset spanning one year from September 1st, 2020, to August 31st, 2021, which is associated with the Aree di Censimento (ACEs)[1] within the province of Brescia. This dataset captures the dynamics of traffic flows, denoted as $flow_{ijt}$, originating from ACE i and arriving at ACE j during the t-th time interval, where each time interval corresponds to a one-hour duration. To elucidate further, the OD data quantifies the count of mobile phone Subscriber Identity Module (SIM) cards that were initially located within a given ACE i during the t-th one-hour interval and subsequently, after a delay of five minutes or more, were identified within ACE j.

The positions of these SIM cards are recorded at five-minute intervals, and only the location of the first arrival during each five-minute interval, referred to as the sampling frequency, is taken into consideration. For instance, let's consider the 1-hour interval t corresponding to 7:00-7:59 AM on February 1st, 2021. If a SIM card is detected in ACE i between 7:00 and 7:04 AM and subsequently arrives in ACE j within the same 5-minute interval, and then reaches a third ACE between 7:05-7:09 AM (labeled as z), the flow data is attributed to $flow_{izt}$. However, it is not attributed to $flow_{ijt}$ or $flow_{jzt}$. This scenario has the potential to result in

[1] According to the Italian National Institute of STATistics (ISTAT), the second highest level of geographical disaggregation is represented by the "Aree di CEnsimento" (ACE), that roughly corresponds to a municipality (or to a portion of a municipality, in case of large cities). Some useful information about the geography and demography of the ACEs, such as the surface area and the number of residents according to the last census, is freely available on the ISTAT website (https://www.istat.it/it/archivio/104317).

an underestimation of the actual traffic flows, particularly in ACEs with limited geographical dimensions that can be traversed in less than 5 min.

It is worth clarifying with an example the distinction between the 1-hour and 5-minute time frequency: suppose the OD data referring to the flows from a specific origin "A" in a specific hour (e.g. 7:00 - 7:59 AM) to a specific destination "B". Each retrieved SIM counts as one flow in that 1-hour interval only if that SIM is retrieved in "A" in a specific 5-minute interval (e.g. 7:00 - 7:04 AM) and in "B" in a subsequent 5-minute interval (e.g. 7:05 - 7:09 AM).

In total, the available data, for each time interval t, is represented as a non-symmetric square matrix with dimensions $N \times N$, where $N = 235$ represents the number of ACEs in the province of Brescia. Rows in this matrix correspond to the ACE of departure, while columns represent the ACE of arrival. Three distinct categories of flows can be identified: flows arriving in ACE i (referred to as "inflows"), flows departing from ACE i (termed as "outflows"), and internal flows from ACE i to ACE i (referred to as "internal flows"). The diagonal elements within each matrix represent the flows departing from and arriving in ACE i, which are recognized as internal flows.

It is noteworthy that the data encompasses both domestic and foreign SIM cards (roaming) connected to the TIM network. The data is derived from two types of SIM cards: human SIM cards, constituting approximately 85% of the total SIM cards, and M2M technology machine SIM cards, which account for about 15% of the total. To prevent the double-counting of users who may possess both a human SIM and devices equipped with an M2M machine SIM, our analysis focuses solely on the count of human SIM cards.

Additionally, the "Minimization of Drive Test" (MDT) technology, a recent innovation offering highly accurate user geolocation data, with an approximate precision of 10 ms, is employed. MDT data has seen limited utilization in academic literature, primarily for technical network control in the field of network engineering. To our knowledge, our studies (Perazzini et al. (2023) and Perazzini et al. (2023b)) mark the initial attempts to employ MDT data for the statistical analysis of traffic flows.

MDT data constitutes a collection of radio measurements of signals transmitted over the 3G/4G mobile network with geographic reference to and from terminal devices equipped with GPS functionality and with a firmware suitable for the MDT (the market share of mobile phone with this technology is rapidly increasing). Each signal corresponds to various activities such as phone calls, text messages, internet browsing, or technical network operations. MDT data pertains to devices with SIM cards associated with TIM, retrieved within a specific rectangular area encompassing approximately 150 square kilometers, roughly corresponding to the flood-prone Mandolossa region. Because of their high precision, we can interpret MDT data related to the area covered by streets as a valid proxy for street traffic.

Due to the specialized technology required for MDT signal detection and the time and cost involved in data collection, only five days in November 2021 (Wednesday 10, Friday 19, Saturday 20, Sunday 21, Monday 22) are available for analysis. These days were thoughtfully selected to represent a typical week. The data is collected at a 15-minute time resolution, enabling observations in time intervals such as 00-14, 15-29, 30-44, and 45-59 for each hour of each day. In few cases data is missing (i.e., for ten out of the 480 time intervals). These replacements include the intervals 04:30-04:44 on all five days, 23:30-23:44, 23:45-23:59 for Monday and Wednesday, and 00:00-00:14 on Friday. These missing data have been filled by replacing them with the average value of other intervals within the same hour of the day (despite we found that this replacement strategy is not determinant to our scope, since night traffic is very limited and regular over different weekdays).

The data is presented in the form of a grid of pixels, each measuring 10 ms on each side and identified by their longitude and latitude coordinates. Overall, the database reports signals from about 274 thousand cells.

It is important to note that since a single device can send or receive multiple MDT signals simultaneously, and only about 10% of current electronic devices produce MDT signals, the number of MDT signals in a pixel in a given time interval may be affected by multiple sources of measurement error that might affect the quality of the data. Therefore, we consider the number of cells in the pixel grid from which signals originated in an area, to mitigate this issue.

2.2 Data manipulation

2.2.1 Weighting strategy

In this study, we use Minimization Drive Test (MDT) data to weight the Origin–Destination (OD) traffic data concerning the traffic on the streets on the flood-prone region of Mandolossa. It is worth recalling that OD data represents the traffic from a given whole ACE to another given whole ACE, while this work aims to analyze the traffic in areas at risk. Since MDT data are provided at a fine-grained scale, the joint use of OD and MDT data allows us to filter out the traffic not pertaining to streets in flood-prone areas. To achieve this, we employ the weighting methodology introduced in Sect. 3 of Perazzini et al. (2023) in conjunction with the following resources:

- The administrative boundary map provided by ISTAT, accessible at https://www.istat.it/it/archivio/104317.

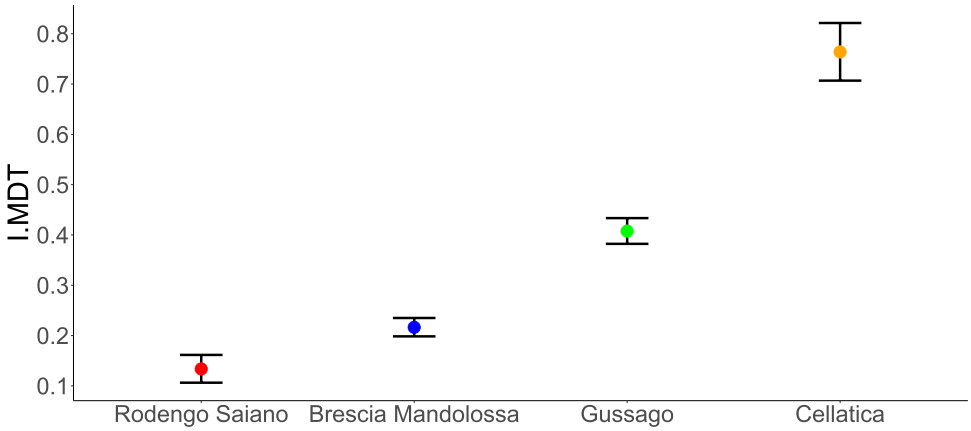


Fig. 1 Mean (bullets) and standard deviation (bars) of the $\overline{I.MDT}_i$ ratio, for $i \in$ {Rodengo Saiano, Brescia Mandolossa, Gussago, Cellatica}. Considered sample: all the four-hour intervals available for all five days, excluding the 0:00-3:59 periods

- The street map, as defined in Perazzini et al. (2023), which is the result of combining two Lombardy Region's released maps related to the Province of Brescia, the first being the "DataBase Topografico Regionale" (version updated in 2021) and the second the "Uso e copertura del suolo della Regione Lombardia 2018".
- The flood risk map with a 20-year time-to-return, as detailed in Balistrocchi et al. (2020).

The strategy proposed (and better detailed) by Perazzini et al. (2023) aims at computing the ratio of phone users on streets that pass by the flood-prone area for each of the 4 ACEs of interest (and, afterward, for the aggregated area constituted by all of them) and to use them as weights. The process that led to the construction of the weights is summarized as follows:

1. Constrain the MDT data to the geographic area overlapping with the four ACEs under investigation.
2. Identify and select the 104 "Sezioni di CEnsimento" (SCEs)[2] that lie within a 500-meter radius from the flood-prone area, as delineated by a flood hazard map computed with a 20-year return period.
3. Align the MDT data with the street map and subsequently narrow down the dataset to MDT data originating from street locations. Subsequently, ascertain the streets traversing the SCEs identified as critical for flood risk management (as determined in step 2), which connect the flood-prone area to the 38 neighboring ACEs.
4. For each ACE and time instance, count the number of grid cells emitting MDT signals related to streets.
5. Among the grid cells identified in step 4, quantify the number corresponding to streets associated with the flood-prone area for each ACE and time instance.
6. Partition the observation period into six intervals of four hours each, 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. Compute the weights for each of the six time intervals using the formula:

$$\text{I\_MDT}_{iT} = \frac{\text{MDT signals from streets connecting the flood prone area}_{iT}}{\text{MDT signals from streets}_{iT}} \tag{1}$$

where i denotes one of the following areas: Brescia Mandolossa, Cellatica, Gussago, Rodengo Saiano, and T represents a given four-hour interval.

As shown in Fig. 1, the computed ratios appear fairly constant at varying time partitions and the observed days. This evidence indicates that the percentage of traffic from an ACE that traverses through the flood-prone area is reasonably stable. So, the weight for each ACE i $\overline{I.MDT}_i$ are computed as the average value of the ratios of the 5 time intervals in the 5 days. The ratio $\overline{I.MDT}_i$ represents the percentage of phone users on the streets subject to flood, and can therefore be interpreted as the portion of traffic of the ACE potentially exposed to floods. This weight is equal to 20% for Brescia Mandolossa, 75% for Cellatica, 40% for Gussago, and 10% for Rodengo Saiano. In addition, the ratio has been calculated for the combined area comprising the 4 ACEs, yielding a value of $\overline{I.MDT}_{agg} = 30\%$.

[2] SCEs, as defined by the Italian National Institute of Statistics (ISTAT), represent the smallest administrative units and can be conceptualized as subdivisions of the ACEs.

2.2.2 Application of the weights to the response variable

Because we are interested in the flows from/to the area of the Mandolossa, our focus is directed towards two distinct groups of ACEs, that we treat separately: the four ACEs intersecting the flood risk map of the Mandolossa, indexed with an i, and specific neighboring ACEs, indexed with a j. 38 ACEs in the vicinity of the Mandolossa region were identified as neighbors, accounting for 84% of the total flows to and from the four aforementioned ACEs of interest (more details can be found in Metulini and Carpita (2023)).

To determine the traffic flows in the flood-prone region at time t between a given ACE i and a given ACE j, that enter the model in Sect. 3 as response variable, we apply the weights as follows:

$$\begin{aligned} Inflow_{jt} &= \sum_i \left(\overline{I.MDT_i} \times flow_{ijt} \right), \\ Outflow_{jt} &= \sum_i \left(\overline{I.MDT_i} \times flow_{jit} \right). \end{aligned} \tag{2}$$

Note that index t represents 1-hour intervals and it should not be confused with the previously adopted index T, representing 4-hour intervals.

For the internal flow, the MDT ratio is applied to the sum of $flow_{iit}$ and $flow_{ii't}$, where both i and i' are ACEs intersecting the flood risk map of Mandolossa:

$$Internalflow_t = \overline{I.MDT}_{agg} \times \sum_i \left(flow_{iit} + \sum_{i' \neq i} flow_{ii't} \right). \tag{3}$$

3 A VARX model for the traffic flows time series

In this research, our focus is directed towards four ACEs located within the flood-prone region of Mandolossa and an additional 38 carefully selected neighboring ACEs.

Our primary objective is to estimate, for each of the 38 neighboring ACEs, a Vector Autoregressive Model with eXogenous variables (VARX), as defined in Tsay (2013). This VARX model enables us to capture the interdependence within each flow and the dependence among the three types of flows, coupled with a Dynamic Harmonic Regression model (DHR). The DHR component is employed to effectively account, through a combination of Fourier bases, for the intricate seasonality patterns.

With the aim to obtain uncorrelated estimated residuals that will subsequently be used for clustering purposes, we have made modifications to the original VARX DHR model introduced and applied by Metulini and Carpita (2023). In particular, we allow lags of order smaller than 24 to be utilized in our analysis.

We define the vector $\mathbf{Flow}_{jt} = [Inflow_{jt}, Outflow_{jt}, Internalflow_t]'$ of values in the $\mathbb{R}_+$ domain, representing the flows between the flood-prone area and the j-th neighboring ACE at time t (where t represents 1-hour intervals). $Inflow_{jt}$, $Outflow_{jt}$, and $Internalflow_t$ are defined as in Eqs. (2) and (3). We model $\mathbf{Flow}_{jt}$ for each neighboring ACE j as a VARX(p), following the equation:

$$\mathbf{Flow}_{jt} = \boldsymbol{\nu}_j + \sum_{h=1}^{p} \mathbf{A}_{jh} \mathbf{Flow}_{jt-h} + \mathbf{B}_j \mathbf{x}_{jt} + \boldsymbol{\epsilon}_{jt}, \; j = 1, ..., 38. \tag{4}$$

Here, $\boldsymbol{\nu}_j$ represents a constant vector of length 3 defined in the $\mathbb{R}$ domain, p is a positive integer scalar autoregressive parameter and $\mathbf{A}_{jh}$ is a 3×3 time-invariant matrix of coefficients in the $\mathbb{R}$ domain to be estimated. For the stationarity issue, a desirable property is that the eigenvalues of $\mathbf{A}$ lie inside the unit circle in the complex plane. $\boldsymbol{\epsilon}_{jt}$ is the 3×1 vector of error terms for the j-th ACE at time t, each one distributed as a zero mean white noise process, and where each other contemporaneous correlation is allowed. The $l \times 1$ vector $\mathbf{x}_{jt}$ in the $\mathbb{R}_+$ domain denotes the exogenous variables at time t, and $\mathbf{B}_j$ is the $3 \times l$ matrix of coefficients in the $\mathbb{R}$ domain associated to the l exogenous variables, such that $\mathbf{B}_j \mathbf{x}_{jt}$ results in a 3×1 vector.

It is important to note that the model assumes a contemporaneous correlation between inflows, outflows, and internal flows. Furthermore, it should be highlighted that the parameter p differs from parameters p_d and p_w used in the work by Metulini and Carpita (2023). In this context, the lag corresponding to $p_d = 1$ represents a 24-hour delay (i.e., the same hour, the previous day). In contrast, the lag associated with $p_w = 1$ corresponds to a 168-hour delay (i.e., considering the previous week, the same hour of the same weekday). Since the OD data are not "real-time" data and are provided with a delay, we can relax the constraint and consider lags of less than 24 h in this study. Hence, $p = 1$ effectively represents a 1-hour lag.

To capture the seasonality of traffic flows, we model $\mathbf{B}_j \mathbf{x}_{jt}$ as a DHR(K_d,K_w) (Hyndman and Athanasopoulos (2018)). Specifically, each element of the vector $\mathbf{B}_j \mathbf{x}_{jt}$ is a combination of daily (d) and weekly (w) periodic functions, as expressed in Equation (5). To capture the seasonality of traffic flows, we model $\mathbf{B}_j \mathbf{x}_{jt}$ as a Dynamic Harmonic Regression (DHR) with

positive integer scalar parameters K_d and K_w, following the approach described in Perazzini et al. (2023):

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

The parameters β_0, α_k, and γ_k are scalar regression coefficients to be estimated. K_d and K_w represent, respectively, the optimal number of Fourier bases for the daily and the weekly pattern, and the positive integer scalars $m_w = 24 \times 7 = 168$ and $m_d = 24$ represent the weekly and daily seasonal periods, respectively.

Note that β_0, α_k and γ_k are allowed to assume different values in the three elements of $\mathbf{B}_j \mathbf{x}_{jt}$. It is noteworthy that the DHR model entails estimating $2 \times K_d$ parameters for the daily pattern and $2 \times K_w$ for the weekly pattern, for each equation in the VARX model. Additionally, for the sake of simplicity, we maintain constant values for parameters p, K_d, and K_w across all ACEs, as these parameters are calibrated to suit each j in the case study.

We employ the ordinary least squares method, which has been shown to be asymptotically equivalent to the maximum likelihood method in VAR models, as outlined in Tsay (2005). The model's parameters are estimated in `R` using `VARX` function in `MTS` package.

4 Time series clustering on upper tail dependence and spatial proximity

In this section, we outline the clustering procedure we propose to determine groups of time series for which extreme events (extremely high traffic flows, in this case) tend to co-occur. To achieve that, we rely on the method originally introduced by De Luca and Zuccolotto (2011), in which time series clustering is performed on a dissimilarity matrix based on bivariate tail dependence coefficients (that will be detailed in the next sections) estimated using copula functions. The concept of exploiting a copula approach to estimate tail dependence for the purpose of time series clustering has been further explored by Durante and Foscolo (2013), Durante et al. (2014), Durante et al. (2014), Ji et al. (2018), who use conditional Spearman's correlation coefficients to measure dissimilarity, and Durante et al. (2015) who propose to estimate tail dependence coefficients with a non-parametric approach. In another direction, Lafuente-Rego and Vilar (2016) and Vilar et al. (2017) exploit quantile autocovariances, while Liu et al. (2018), Yang et al. (2018) and Yang et al. (2020) resort to the coefficient of weak lower-tail maximal dependence, to the jump tail dependence coefficient and to the α-tail distance, respectively. Additional instances of clustering that rely on copula functions for estimating tail dependence coefficients encompass Jun and Ziping (2013), De Luca and Zuccolotto (2017b), De Luca and Zuccolotto (2017a), Lohre et al. (2020), De Luca and Zuccolotto (2021) and D'Urso et al. (2023), where new clustering algorithms are proposed based on the original idea.

In this paper, we propose a new algorithm, to take into account spatial proximity between ACEs. The proposed procedure is based on the dissimilarity matrix of De Luca and Zuccolotto (2011) that is opportunely modified, as will be clarified below. In addition, with respect to the originally proposed algorithm, this procedure (1) skips the step of expressing time series in a high-dimensional space through MDS and, more importantly, (2) introduces a method that exploits a novel procedure called cutMOB, whose functioning will be explained in the following, to optimally compute the dissimilarity values, based on a combination of tail dependence and proximity coefficients.

The idea of designing a clustering algorithm suited to time series generated in a context with a spatial structure has already been pursued by Coppi et al. (2010), Disegna et al. (2017) and Benevento et al. (2023), but not from an extreme events perspective.

In the following, we will begin by providing a brief overview of copula functions and their application in estimating tail dependence coefficients. Then, we will give details about the proposed clustering procedure.

4.1 Copula functions and tail dependence coefficients

A 2-dimensional copula (Sklar 1959) is a function denoted by

$$C : [0, 1]^2 \to [0, 1].$$

Given the continuous random variables X_j, X_h, and their cumulative distribution functions $U_j = F_j(X_j)$, $U_h = F_h(X_h)$, the 2-dimensional copula function applied to u_j, u_h, is equivalent to the joint distribution function,

$$C(u_j, u_h) = P\big(F_j(X_j) \le u_j, F_h(X_h) \le u_h\big)$$

that is

$$C(u_j, u_h) = F_X\left(F_j^{-1}(u_j), F_h^{-1}(u_h)\right).$$

Then

$$F_X(x_j, x_h) = C\big(F_j(x_j), F_h(x_h)\big).$$

Copula functions provide a highly flexible description of the joint distribution, accomplished through the use of a copula function that joins the univariate marginal distributions of the random variables. When a copula function is employed to describe a joint distribution, various notable characteristics of the multivariate distribution can be readily extracted. Examples of such characteristics include the tail dependence coefficients (TDCs): given two random variables X_j and X_j, the lower and upper TDCs are given, respectively, by

$$\lambda^L_{j|h} = \lim_{v\to 0^+} P(U_j \le v \mid U_h \le v)$$

and

$$\lambda^U_{j|h} = \lim_{v\to 1^-} P(U_j > v \mid U_h > v).$$

In case of upper (lower) tail independence, we have $\lambda^L = 0$ ($\lambda^U = 0$), while non-null values indicate that a dependence exists between the extremely high (low) values of the two random variables, exhibiting stronger dependence as the coefficient value increases.

We distinguish elliptical copulas (the Gaussian copula and the Student's t copula) and Archimedean copulas (see Joe (1997) for a comprehensive review). The main feature of Archimedean copulas is greater flexibility in modelling tail dependencies because the two coefficients can be different, while Gaussian copula does not admit tail dependence and the use of Student's t copula implies assuming equal tail dependence for the two tails.

4.2 Upper tail dependence clustering with spatial proximities

In this study, our focus lies on upper tail dependence, given that the events under scrutiny pertain to exceptionally high traffic flows. To cluster times series based on upper TDCs, the methodology proposed by De Luca and Zuccolotto (2011) requires obtaining the $N \times N$ dissimilarity matrix Δ, whose generic element δ_{jh} is the dissimilarity between the ith and the jth time series, with

$$\delta_{jh} = -\log(\lambda^U_{j|h}). \tag{6}$$

The dissimilarity matrix Δ is then used as a basis for the employed clustering algorithm. In this study, we present a clustering algorithm designed to consider, beyond dissimilarities, spatial proximity between areas. Therefore, we introduce a novel dissimilarity measure. So, we introduce a new dissimilarity measure δ^θ_{jh} as a modification of (6),

$$\delta^\theta_{jh} = -\log(\lambda^U_{j|h}) + \theta c_{jh}, \tag{7}$$

where c_{jh} is a proximity coefficient, similar to that proposed by Coppi et al. (2010). Different choices are possible for the proximity coefficient: a binary indicator assuming value 0 when the $j-$th and $h-$th time series denote traffic flows coming from neighbouring areas, and 1 otherwise, as well as more refined proposals, such as a distance (actual distance, as the crow flies, based on travel time,...) between area i and area j. In this work we opt for the binary indicator. The dissimilarity matrix obtained by (7) is denoted by Δ^θ.

The parameter $\theta > 0$ is aimed to adjust the impact of the coefficient of proximity in the dissimilarity between the time series, and its optimal value can be obtained via an iterative procedure, which is detailed in Algorithm 1.

Algorithm 1 Upper tail dependence clustering with spatial proximities

Require: Two dissimilarity matrices Δ and Δ^θ, obtained as in (6) and (7).
1: Define a sequence Θ of values, starting from 0, that could be plausible values for θ (*e.g.* $0.005, 0.01, 0.015, \ldots, 4$)
2: **for** θ assuming all the values in Θ **do**
3: perform cluster analysis with a hierarchical agglomerative algorithm, using Δ^θ as dissimilarity matrix
4: identify the optimal number of clusters k, by cutting the dendrogram with cutMOB
5: for the clusterization into k groups, compute internal clustering validation indices (*e.g.* Average silhouette width, Dunn index, Calinski and Harabasz index, ...) on the dissimilarity matrix Δ
6: **end for**
7: plot the graphics of the values of the internal clustering validation indices versus θ, and decide its optimal value

Step 3. of Algorithm 1 requires to carry out a hierarchical agglomerative clustering and obtain the corresponding dendrogram, a tree-like diagram that represents the arrangement of clusters produced during the clustering process. The tree structure starts with individual data points at the leaves and progressively merges them into larger clusters according to their similarity (and following a specific criterion called linkage), as moving toward the root of the tree. Dendrograms allow to understand the relationships and hierarchy between different clusters, helping to identify the optimal number of clusters or subgroups in the data, but, to do that, decisions have to be taken about where to cut the tree to obtain a specific number of clusters for further analysis. This decision is not always simple, especially when the analysis requires cutting a large number of dendrograms, as it is the case of analyses that track the changing composition of clusters over time (De Luca and Zuccolotto 2023), or when the adopted algorithm requires performing a big number of clusterizations, as it is the case of the procedure we propose in this paper. In such cases, an automatic cutting procedure helps carry out the analysis in a fast and efficient way.

This automatic procedure is adopted in Step 4. of Algorithm 1, where the optimal number of clusters is obtained through cutMOB (cut MOdel-Based partitioning), a novel method that has proven effective in determining the optimal height at which to cut a dendrogram, proposed in a paper by De Luca and Zuccolotto (2023), where a review of other possible methods for dendrogram cutting is also presented. In brief, the idea on which cutMOB is based starts by drawing a plot of the branches' lengths, in non-decreasing order, from the first to the last iteration of the hierarchical clustering agglomeration path. Recalling that, in general, the dendrogram is cut where, at a visual inspection, the branches begin to be 'too long', the graph aims to detect an elbow, suggesting where the dendrogram should be cut. The elbow is detected thanks to an unusual application of the MOB algorithm proposed by Zeileis et al. (2008). The functioning of cutMOB is explained in detail in the Appendix.

Finally, in Step 5. of Algorithm 1, the efficacy of the clusterization is assessed by the adopted internal clustering validation indices, referencing to the dissimilarity matrix Δ. The rationale behind this selection lies in the fact that proximity between areas is employed to define a set of optimal clusterizations at given values of θ. However, the final choice among them is determined by selecting the one that guarantees the most effective separation among clusters, solely in terms of upper tail dependence.

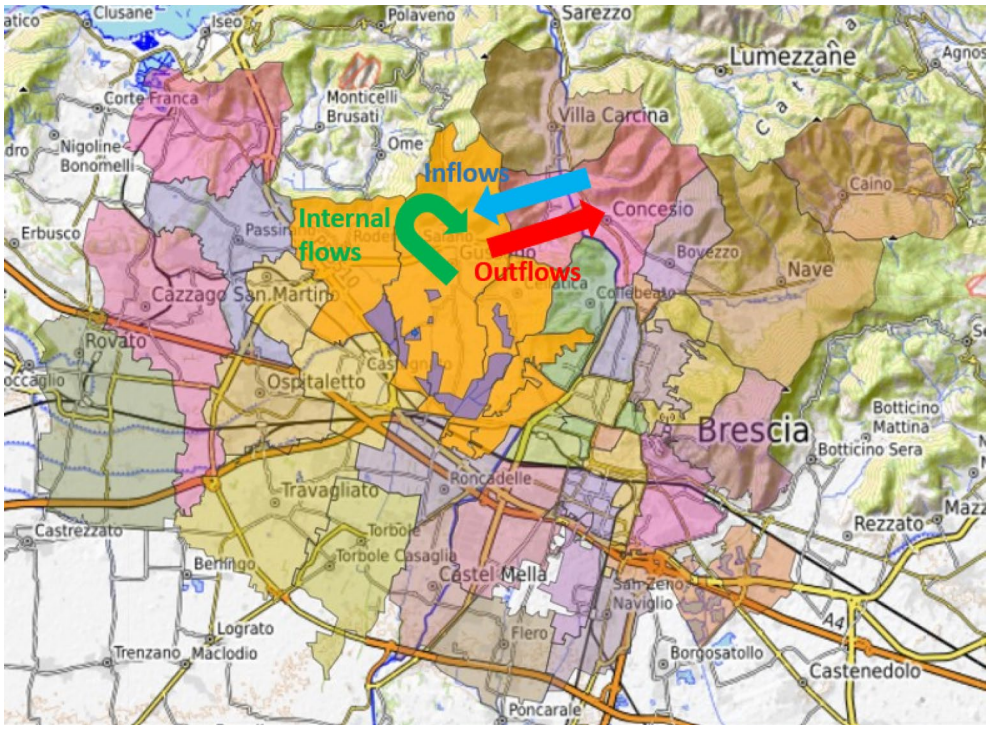
5 Case study: traffic flows in Mandolossa

Figure 2a visually illustrates the analyzed flows of people in our study by taking as an example the flows between Mandolossa and the ACE of Concesio. The Mandolossa region, which intersects the flood-prone area, is delineated by the orange polygon. The figure also depicts various flows, encompassing outflows (indicated by the red arrow), inflows (represented by the light blue arrow), and internal flows (designated by the green arrow). Notably, internal flows encompass traffic between ACEs within Mandolossa and traffic within each single ACE. Moreover, the area prone to flood risk (with a 20-year time-to-return) is marked in violet.

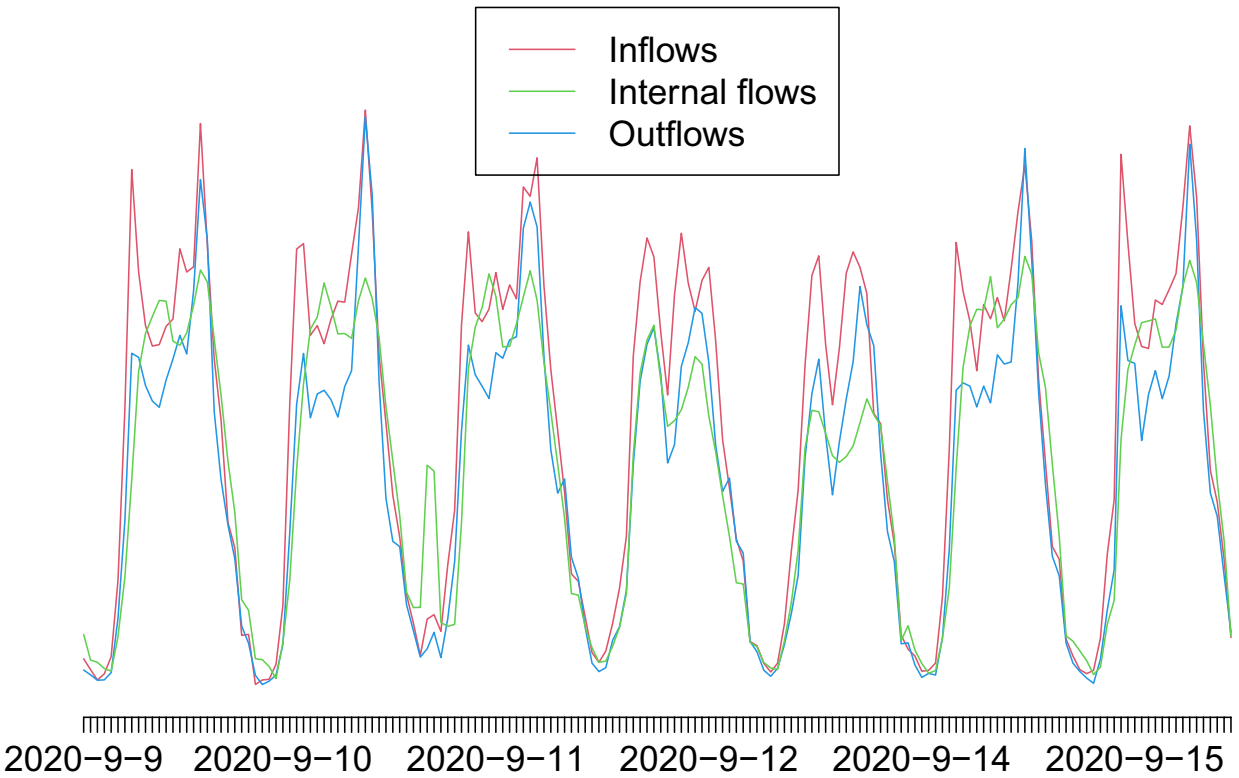
Figure 2b shows one week (September 9^{th}, 2020 – September 15^{th}, 2020) of the time series of the three traffic flows. A strong daily pattern is evident in all the time series. Moreover, a strong contemporaneous correlation is observed among the three flows, which stands at about 0.9 depending on the cases. These findings, common to the other 37 cases, demonstrate the validity of utilizing a VARX model with the DHR component to capture the seasonality.

We carry out the data processing procedure described in Sect. 2.2 on the data described in Sect. 2.1, then we obtain estimated residuals from the model presented in Sect. 3. The R codes can be provided upon request. According to the modeling strategy, here we are allowed to relax the constraint about the impossibility of using previous hours in the AR term that was in place in Metulini and Carpita (2023), because the aim of this work is not about traffic flows' nowcasting. Therefore, all AR orders are allowed to be used here. It is worth highlighting that we estimate 38 different models representing the flow from (to) Mandolossa (as an aggregation of 4 ACEs) to (from) single neighbor ACEs.

According to model calibration, the optimal number of Fourier bases for the DHR component has been inherited from previous works of us (e.g. Metulini and Carpita (2023)), So, the model contains 7 daily and 4 weekly Fourier bases (i.e. $K_d = 7$ and $K_w = 4$). According to the choice of the autoregressive (AR) order, on the model with $K_d = 7$ and $K_w = 4$ Fourier basis for the DHR component, we conducted a comprehensive examination of various AutoRegressive (AR) structures based on the Auto Correlation Function (ACF), the Partial AutoCorrelation Function (PACF), and the Ljung-Box test (Ljung and Box 1978) applied to the estimated residuals. Following this comparison of AR structures, we selected a model characterized by the first 25 lags (i.e., $p = 25$), exhibiting minimal autocorrelation, small values of PACF and ACF, and with the null hypothesis of uncorrelated residuals of the Ljung-Box test always accepted (i.e. for all the 38 residual vectors estimated from the model with $p = 25$). According to a visual inspection



(a) A simplified representation of inflows, outflows, and internal flows on a map.



(b) Time series of inflows, outflows, and internal flows (September 9^{th}, 2020 – September 15^{th}, 2020). Min-max normalization applied to the original time series.

Fig. 2 Flows between Mandolossa and one of the 38 neighbour ACEs (Concesio). Outflows (red), inflows (light blue), and internal flows (green)

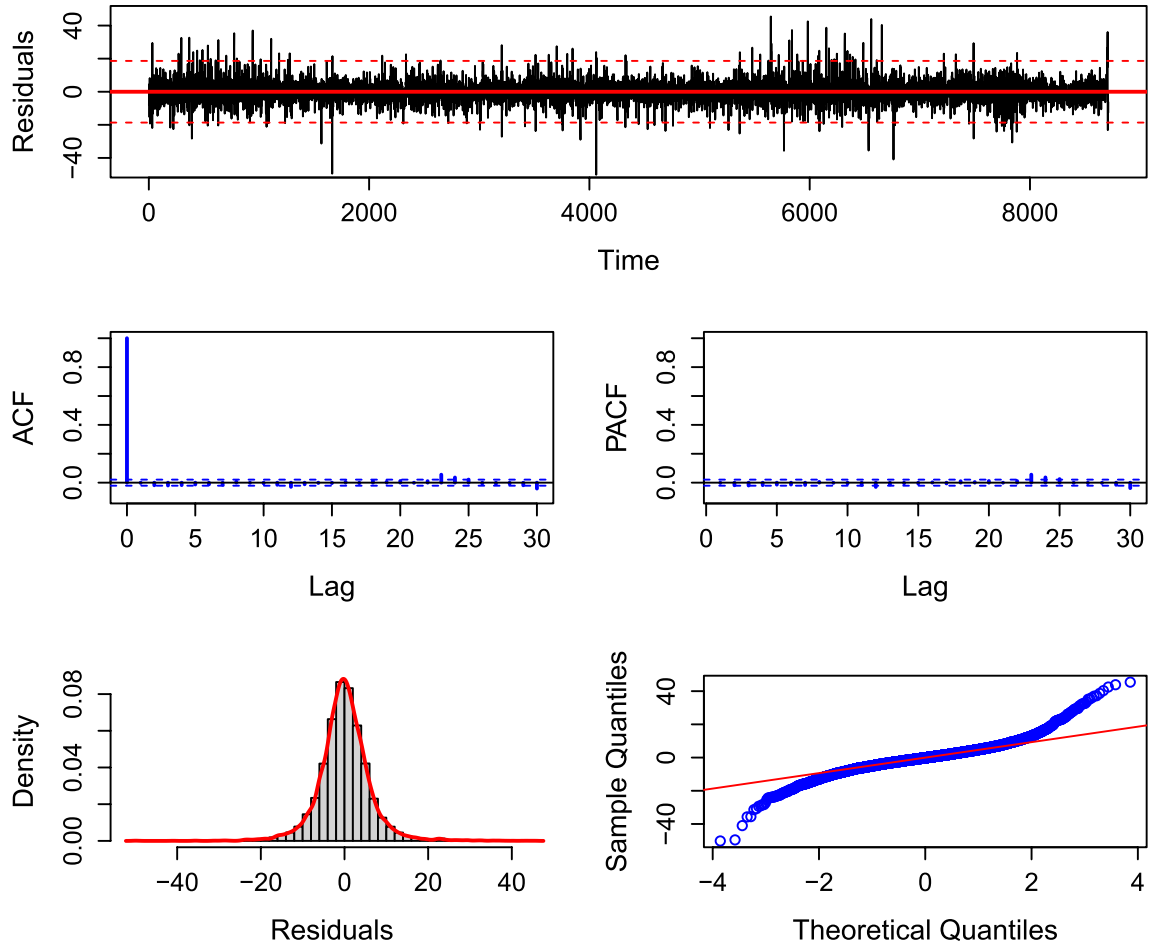


Fig. 3 Residuals' diagnostic: Time series of estimated residuals (top) with 95% confidence bands; ACF (center left) and PACF (center right), with 95% confidence bounds for strict white noise; a histogram depicting the empirical distribution is accompanied by a Normal distribution curve (positioned in the lower-left corner), as well as a QQ-plot assessing the normality of the data. Time series of inflows from the ACE of *Brescia 1*

and the Shapiro-Wilk test, the normality assumption in the estimated residuals is not satisfied. This evidence further justifies the usefulness of the proposed procedure for clustering time series. Figure 3 displays, as an example, the diagnostic of the inflows' estimated residuals using the model in eq. 4 with $p = 25$, $K_d = 7$ and $K_w = 4$ for the ACE called *Brescia 1* (which corresponds to the yellow polygon on the right of the Mandolossa in Fig. 2). The autocorrelation function (center left) displays a strong correlation of the first order, while the partial autocorrelation function (center right) shows some significant correlations at orders 23 to 25 (i.e. the same hours of the previous day). The QQ plot shows a departure from the normality in correspondence of the left and the right tails, which is typical in the case of leptokurtic distribution. This evidence is confirmed by the value of Skewness of 0.299 (slightly skewed distribution) and by the value of Kurtosis of 8.768 (which can be associated with leptocurtic distributions). For this ACE, the Ljung and Box test does not reject the null hypothesis of uncorrelation of the residuals. By considering 1 lag in the test, we obtain a statistic of 0.014 (p-value = 0.906), with 5 lags, the statistic is 1.403 (p-value = 0.924), with 10 lags, 1.678 (p-value = 0.998). It is worth highlighting here that similar diagnostic results are obtained for all 38 considered areas.

While the examination of all estimated residuals could be of interest, in this particular application we only consider those associated with the inflows. This choice is made because inflows' estimated residuals allow to cluster ACEs based on the traffic dynamics from the surrounding neighborhood to the Mandolossa area.

The estimated standardized residuals of inflows of the model (4) with $p = 25$, $K_d = 7$, and $K_w = 4$ applied to all the time series of traffic flows have subsequently been utilized to derive the respective distribution functions $\hat{U}_{jt}$. For each of the $(38 \times 37)/2 = 703$ pairs $(\hat{U}_{jt}, \hat{U}_{ht})$, a set of elliptical (Gaussian and Student's t) and Archimedean (Clayton,

Gumbel, Frank, Joe, BB1, BB6, BB7, BB8 and their rotated versions) copulas are estimated by Maximum Likelihood and the copula which shows the lowest AIC value is selected. Copulas that are most frequently chosen are the Student's t copula (85% of the time) and the 180-degree rotated BB7 copula (13.5%).

After obtaining the estimates of the upper tail dependence coefficients, we executed the clustering procedure outlined in Algorithm 1 with $\Theta = \{0.005, 0.01, 0.015, \ldots, 4\}$ and employing a hierarchical agglomerative algorithm with complete linkage. It is recommended to check the goodness of the choice for Θ, by keeping under control the number of clusters corresponding to different values of θ. Typically, when c_{jh} is defined as a binary variable as in this case study, as θ becomes too large, the procedure generates a huge number of clusters, as a result of a sort of overfitting due to the tendency to cluster together couples of contiguous objects. The number of clusters suggested by cutMOB for each value of θ is displayed in Fig. 4, which shows that the number of selected clusters is quite stable around 3 or 4, and hugely increases as θ approaches the last values of the sequence Θ. This confirms that there is no need to explore higher values for θ. As internal clustering validation indices, we adopted the Average silhouette width, the Calinski and Harabasz index, and the Dunn index, all suggesting an optimal value of around 0.04 for θ (Fig. 5).

With $\theta = 0.04$ the areas are partitioned into four clusters, as depicted in the two panels of Fig. 6. ACEs exhibit a robust spatial neighborhood structure, where extreme events tend to occur simultaneously in geographically adjacent areas. One cluster (colored in blue) consists of ACEs situated in the southern outskirts of Mandolossa, characterized by a significant network of streets leading to Mandolossa. The second cluster (in purple) includes many ACEs not directly contiguous to Mandolossa but connected by extensive roadways. The remaining two clusters comprise only a few ACEs, with Caino forming a distinct group.

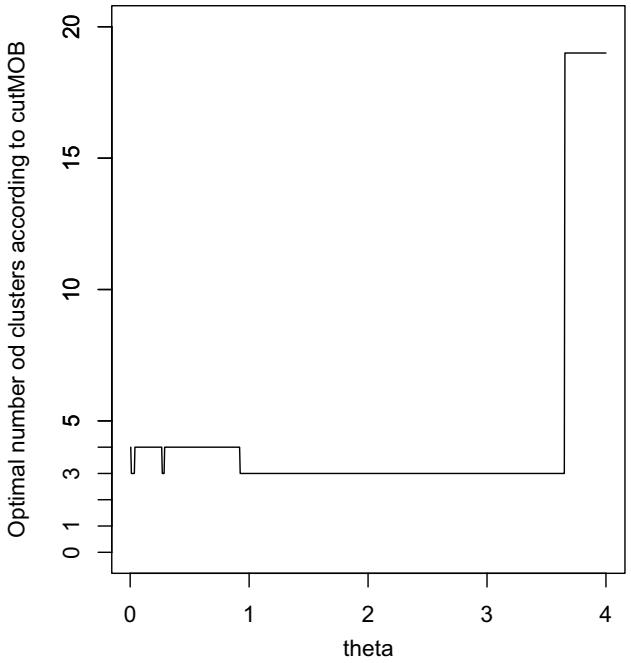


Fig. 4 Graphic of the number of clusters suggested by cutMOB versus θ (point 4 of Algorithm 1)

6 Conclusions

The need to predict traffic flows in a flood-prone area is really crucial. In this work, we have considered 38 areas (ACE) of an Italian region, and estimated 38 dynamic models for the traffic flows from (and to) a specific location. The traffic flows are identified using mobile phone data. After filtering the flows (inflows, outflows, and internal flows) using a vector autoregressive model with exogenous variables, combined with a dynamic harmonic regression model to capture seasonality, the goal has been the clustering the inflow residuals of the 38 areas using a dissimilarity matrix based on the upper tail dependence coefficients and the spatial proximity. The similarity between the two areas is therefore defined in terms of the joint occurrence of extremely high traffic flows, and spatial proximity.

The application of the cutMOB algorithm allows the identification of four clusters. As seen from the map in Fig. 6, the 4 groups obtained with the cutMOB procedure are substantially coherent for road and spatial characteristics. The most evident result concerns the group with the single ACE of Caino (yellow), which is very far from the Mandolossa alluvial area (orange) and whose traffic inflows mainly pass through secondary roads. A second cluster (purple) groups ACEs located to the west (direction Rovato), northeast (direction Cesio), and southeast (direction Flero), all areas being relatively far from the alluvial area and whose traffic inflows mainly pass through provincial and state roads. A third cluster (light blue) groups ACEs being very close to the south of the alluvial area (directions Roncadelle and Brescia) and whose traffic inflows mainly pass through provincial roads, while the last cluster (green) groups ACEs to the south and southeast of the alluvial area whose traffic inflows mainly pass through secondary roads.

The results obtained with this study, in particular the cluster analysis of the extreme traffic flows, have made it possible to acquire further useful information regarding the mobility of the area considered. This information could

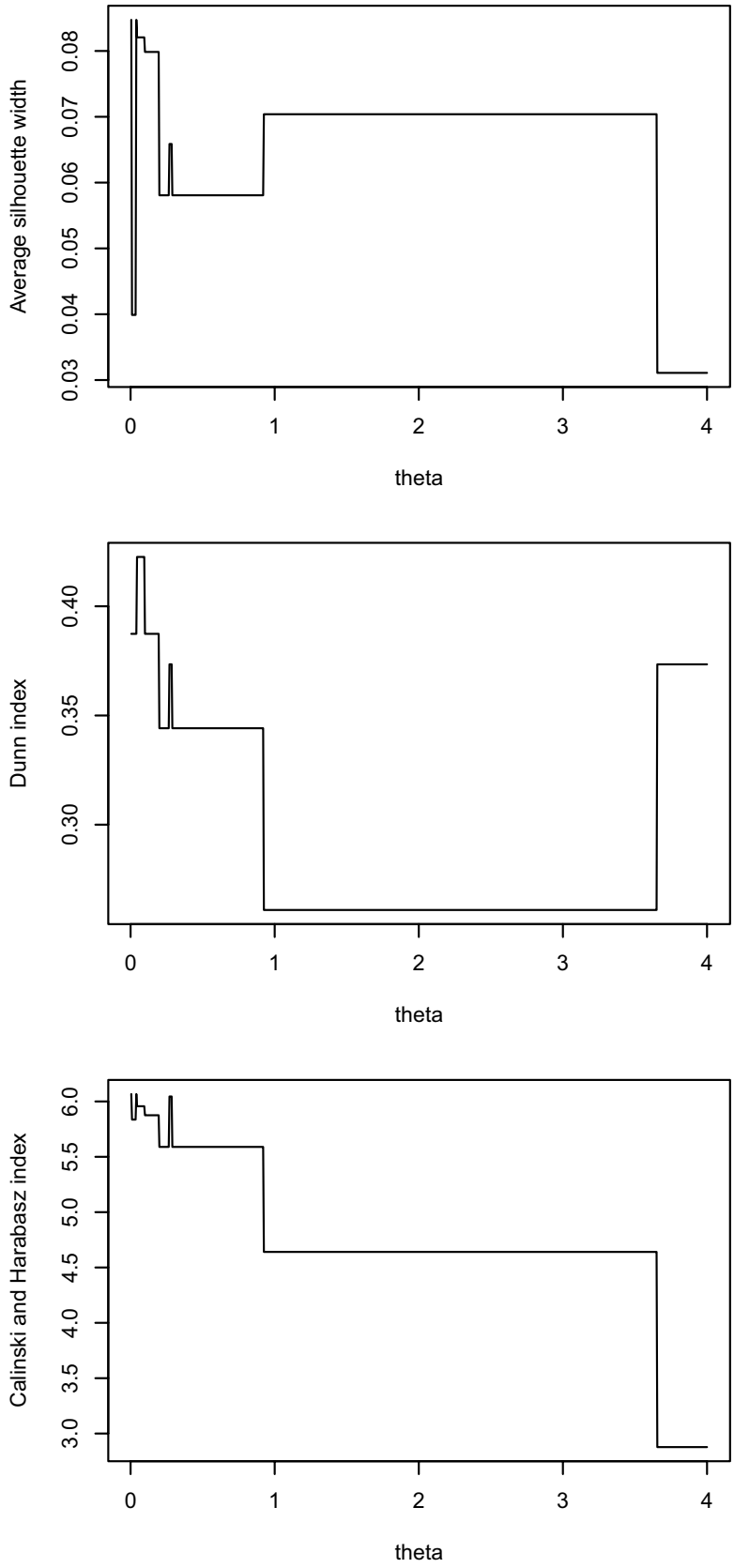
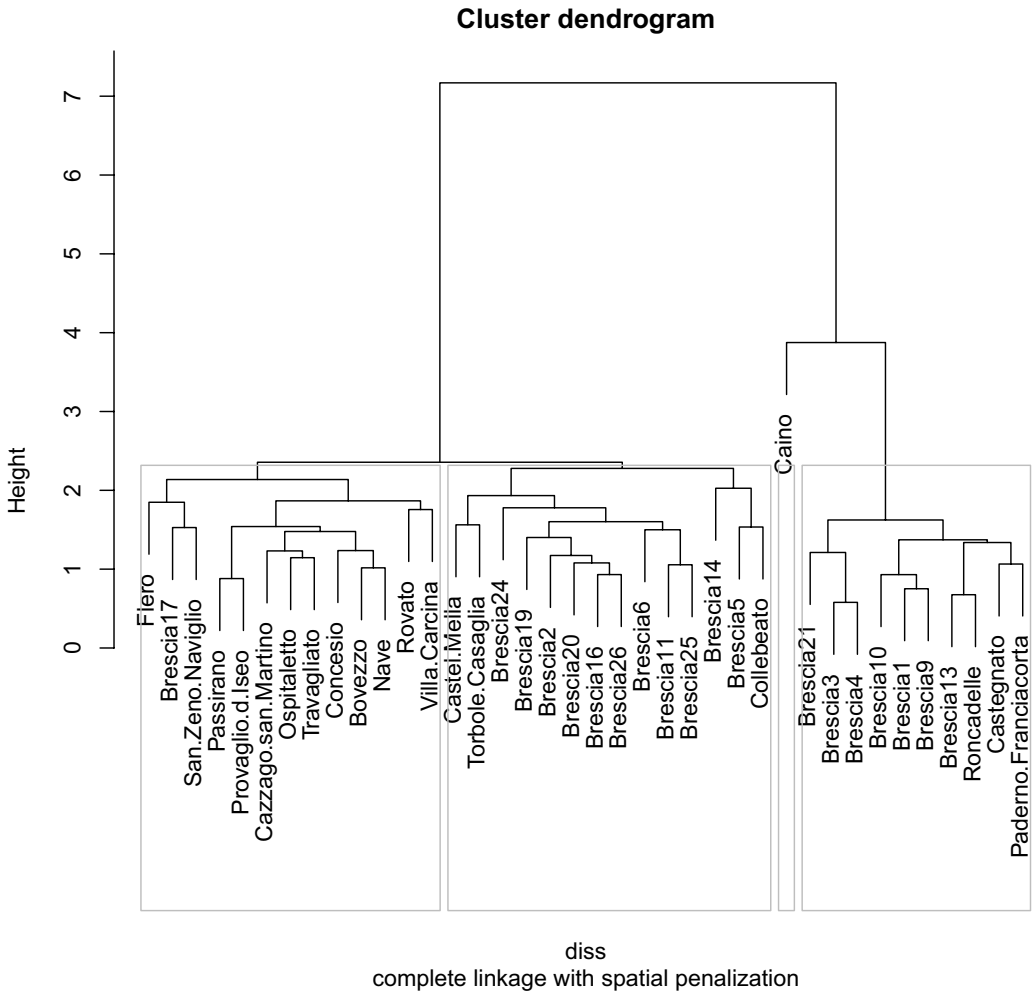


Fig. 5 Values of the indices of the internal clustering validation versus θ (Algorithm 1), point 7



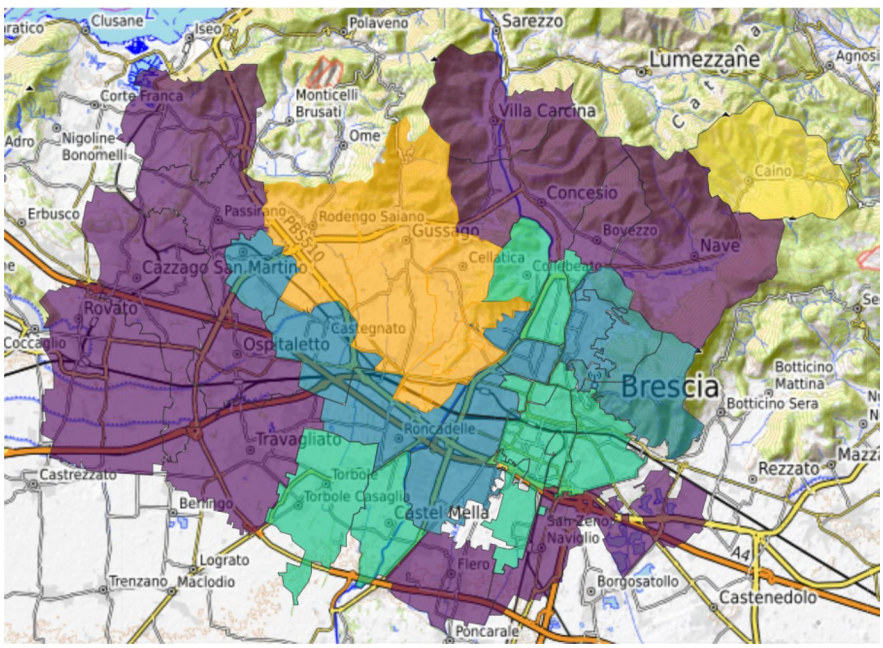


Fig. 6 Dendrogram resulting from the clusterization procedure (top) and map of the 38 clustered *Aree di Censimento* near to Mandolossa (bottom, Mandolossa depicted in orange)

be used to add other explanatory variables to improve the dynamic model used for estimating traffic flows in the Mandolossa alluvial area or other areas of interest due to critical issues related to mobility.

In a broader sense, the methodology proposed in this paper has several practical applications in real-life scenarios, such as in:

- Public Safety Alerts, where real-time monitoring of traffic flows can help authorities issue timely warnings and advisories to residents in flood-prone areas, reducing the risk of casualties and property damage.
- Emergency response planning: By analyzing traffic flows, authorities can identify critical evacuation routes and allocate resources more effectively during flooding events, ensuring timely evacuation of vulnerable areas.
- Infrastructure planning: Understanding traffic patterns can inform infrastructure development, such as building or upgrading roads and bridges to improve flood resilience and minimize disruptions during extreme weather events.
- Insurance and risk assessment: Insurance companies can use traffic flow data to assess flood risk for different areas, adjusting premiums accordingly and providing incentives for flood mitigation measures.
- Urban planning: City planners can incorporate insights from traffic flow modeling into land use planning and zoning regulations to minimize flood exposure for residential and commercial areas.
- Climate change adaptation: As climate change intensifies extreme weather events, understanding traffic flows can inform adaptive strategies to mitigate flood risks and enhance community resilience.

Appendix: details about the cutMOB procedure

As mentioned above, the idea underlying cutMOB begins with plotting the lengths of branches in non-decreasing order across the iterations of the hierarchical clustering agglomeration path. Typically, dendrograms are cut where, upon visual inspection, the branches appear excessively long. The objective of the graph is to identify an elbow point, indicating the optimal location for cutting the dendrogram. In this Appendix we show how cutMOB, an automatic procedure for dendrogram cutting, works.

We show the example of a set of $N = 38$ objects, grouped into $k = 4$ clusters, drawn from a 15-dimensional random variable by means of the simulation process proposed in Waller et al. (1999) and then analyzed with the hierarchical clustering algorithm with average linkage. Figure 7 shows the resulting dendrogram and the corresponding plot of the branches' lengths, where - through visual judgment - we detect two possible elbows, at iteration 33 or 34 (corresponding to the two black points).

In order to detect the elbow in an automatic way, we may consider all the possible partitions of the ordered branches heights into two parts and, for each partition, fit the points by two separate regression lines. The idea is that the partition ensuring the highest difference in the slope of the two regression lines informs about the optimal dendrogram's cut. Figure 8 shows how the idea works, for two selected separations in the branch height plot of Fig. 7, at iterations 33 and 34. Splitting the branch heights at iterations 33 and 34 correspond to identify the elbow in those points, which yields, respectively, 5 and 4 clusters.

For the selection of the iteration that induces the best partition, cutMOB relies on the MOB algorithm proposed by Zeileis et al. (2008), a technique that combines the statistical modeling culture and the idea of recursive partitioning. In detail, MOB proceeds through the following steps: (1) a given parametric model $Y = f(\boldsymbol{X}) + \varepsilon$ is selected to fit data, where Y is an outcome variable modelled as a function of a set of covariates $\boldsymbol{X}$; (2) starting from the whole dataset, data are repeatedly split into two parts, according to a set of partitioning variables $\boldsymbol{Z}$ (all the possible splits); for each split the model selected in (1) is fitted to the two sets of data and a test for the model parameters' instability is performed; (3) if absence of parameter instability is rejected, the data are divided into two groups according to the value assumed by partitioning variable (and the cutoff value) inducing with the highest instability, thus generating two separate sub-datasets; and (4) the procedure is reiterated in each of the sub-datasets until the test accepts the null hypothesis of no parameter instability. Note that the variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not necessarily have to be distinct.

In cutMOB, the outcome variable Y is set to be the branch height, modeled as a function of a unique covariate X given by the iteration number, which is also designed as (unique) partitioning variable Z. The statistical model used to fit Y as a function of X is simple linear regression. The procedure requires fitting data with the MOB and considering only the first split, i.e. the one determining with the highest parameter instability. Given that the search for the split that causes the greatest parameter instability in the regression lines leads to the presence of two regression lines with significantly different slopes, this method aligns with the notion of detecting the elbow. Figure 9 displays the results of MOB applied to the above-mentioned simulated data: the first split suggests to assign the first 34 iterations to left node and the last three to the right node, thereby identifying the elbow at iteration 34. As shown in the bottom panel of Fig. 8, this yields a dendrogram's cut that generates $k = 4$ clusters

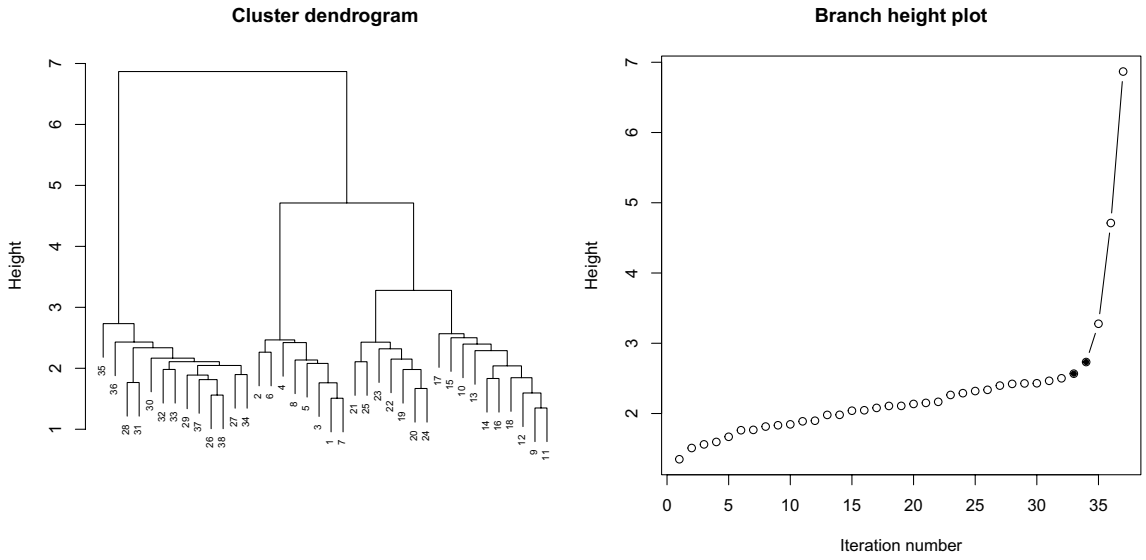


Fig. 7 Simulated data: dendrogram (left) and branch height plot (right)

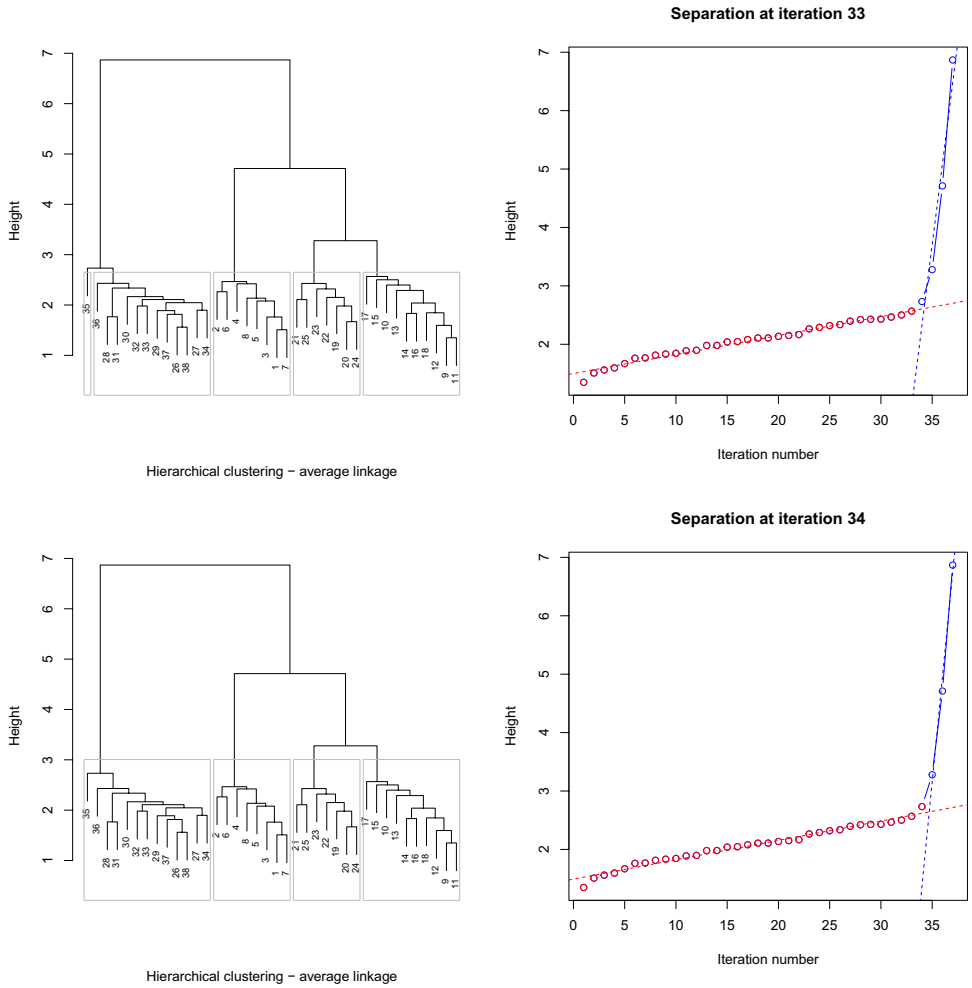


Fig. 8 Split of the branch height plot into two parts, with a regression line fitting the two sets of points (right), with the corresponding dendrogram cuts (left). Examples of split at iterations 33 (top) and 34 (bottom)

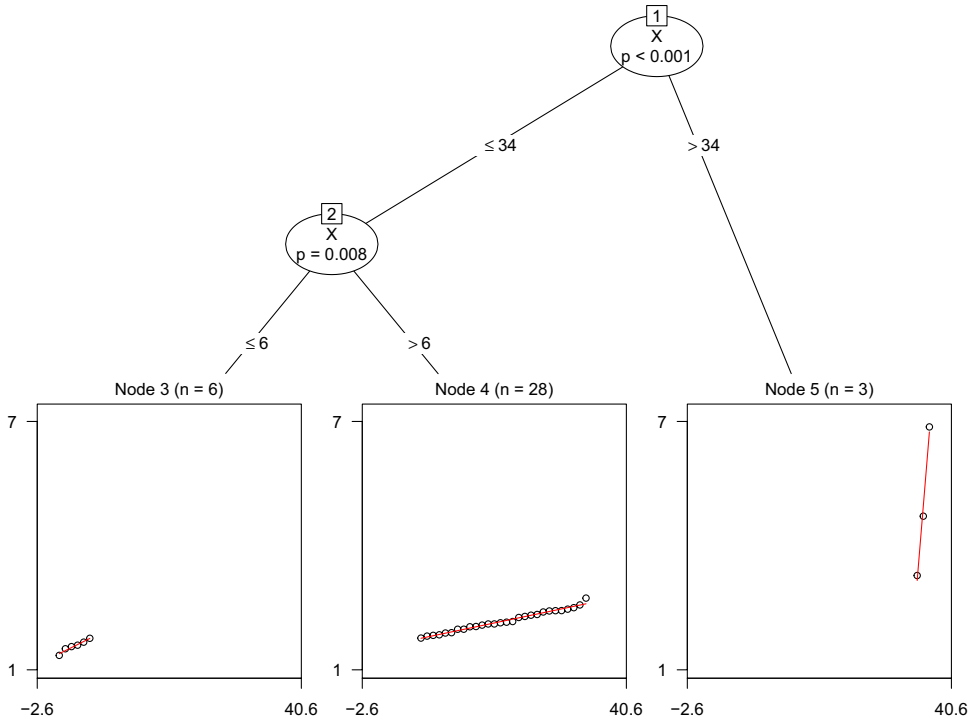


Fig. 9 MOB for the set of simulated data

(as many as implied by the data generation process used to simulate the data).

Further details on cutMOB and a simulations study can be found in the seminal paper (De Luca and Zuccolotto 2023).

Acknowledgements This contribution has been developed for the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project "Sustainable Mobility Center (MOST)" 2022-2026, CUP D83C22000690001, Spoke Number 7 "CCAM, Connected networks and Smart Infrastructures", and for the project "Study of mobile phone siGNals for the evalUation of the interconnections between Mobility and the environment in Lombardia (SIGNUM)" CUP: F53D23010910001- PRIN 2022 PNRR M4C2 - financed by the European Union - Next Generation EU (DD MUR n. 1409 del 14/09/2022).

Author contributions All authors contributed equally to this work.

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

Declarations

Competing interests The authors declare no competing interests.

References

Balistrocchi M, Metulini R, Carpita M, Ranzi R (2020) Dynamic maps of human exposure to floods based on mobile phone data. Nat Hazard 20(12):3485–3500

Benevento A, Durante F, Pappadà R (2023) An approach to cluster time series extremes with spatial constraints. In: Chelli FM, Ciommi S, Ingrassia S, Mariani F, Recchioni MC (eds) Book of the Short Papers SIS 2023 - Statistical Learning. Sustainability and Impact Evaluation, Pearson

Carpita M, Simonetto A (2014) Big data to monitor big social events: analysing the mobile phone signals in the Brescia smart city. Electron J Appl Stat Anal: Decis Support Syst Serv Eval 5(1):31–41

Coppi R, D'Urso P, Giordani P (2010) A fuzzy clustering model for multivariate spatial time series. J Classif 27:54–88

De Luca G, Zuccolotto P (2011) A tail dependence-based dissimilarity measure for financial time series clustering. Adv Data Anal Classif 5(4):323–340

De Luca G, Zuccolotto P (2017a) A double clustering algorithm for financial time series based on extreme events. Stat Risk Model 34:1–12

De Luca G, Zuccolotto P (2017b) Dynamic tail dependence clustering of financial time series. Statistical Papers, pages 1–17

De Luca G, Zuccolotto P (2021) Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach. Int J Approx Reason 139:88–103

De Luca G, Zuccolotto P (2023) Dynamic time series clustering with multivariate linkage and automatic dendrogram cutting using a recursive partitioning algorithm. Inf Sci 649:119605

Disegna M, D'Urso P, Durante F (2017) Copula-based fuzzy clustering of spatial time series. Sp Stat 21:209–225
Durante F, Foscolo E (2013) An analysis of the dependence among financial markets by spatial contagion. Int J Intell Syst 28(4):319–331
Durante F, Foscolo E, Jaworski P, Wang H (2014) A spatial contagion measure for financial time series. Expert Syst Appl 41(8):4023–4034
Durante F, Pappadà R, Torelli N (2014) Clustering of financial time series in risky scenarios. Adv Data Anal Classif 8(4):359–376
Durante F, Pappadà R, Torelli N (2015) Clustering of time series via non-parametric tail dependence estimation. Stat Pap 56(3):701–721
D'Urso P, De Luca G, Vitale V, Zuccolotto P (2023) Tail dependence-based fuzzy clustering of financial time series. Ann Oper Res, pages 1–27
Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts
Ji H, Wang H, Liseo B (2018) Portfolio diversification strategy via tail-dependence clustering and ARMA–GARCH vine copula approach. Aust Econ Pap 57(3):265–283
Joe H (1997) Multivariate Models and Multivariate Dependence Concepts. CRC Press, Boca Raton
Jun Z, Ziping D (2013) Distance measure of financial time series based on the coefficients of temporal tail dependence. Int J Adv Manag Sci 2(4):143–146
Lafuente-Rego B, Vilar JA (2016) Clustering of time series using quantile autocovariances. Adv Data Anal Classif 10(3):391–415
Liu X, Wu J, Yang C, Jiang W (2018) A maximal tail dependence-based clustering procedure for financial time series and its applications in portfolio selection. Risks 6(4):115
Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297–303
Lohre H, Rother C, Schäfer KA (2020) Hierarchical risk parity: Accounting for tail dependencies in multi-asset multi-factor allocations. New Developments and Financial Applications, Machine Learning for Asset Management, pp 329–368
Mariotti I, Giavarini V, Rossi F, Akhavan M (2022) Exploring the" 15-minute city" and near working in Milan using mobile phone data. Territorio Mobilità e Ambiente, TeMA, p 15
Metulini R, Carpita M (2021) A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Soc Indic Res 156(2–3):761–781
Metulini R, Carpita M (2023) Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Ann Oper Res pages 1–26
Mishra D, Kumar S, Hassini E (2019) Current trends in disaster management simulation modelling research. Ann Oper Res 283:1387–1411
Perazzini S, Metulini R, Carpita M (2023) Integration of flows and signals data from mobile phone network for statistical analyses of traffic in a flooding risk area. Socioecon Plann Sci 90:101747
Perazzini S, Metulini R, Carpita M (2023b) Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Stat Methods Appl pages 1–28
Pucci P, Gargiulo C, Manfredini F, Carpentieri G, et al (2022) Mobile phone data for exploring spatio-temporal transformations in contemporary territories. In Tema. J f Land Use Mobil Environ pages 6–12. ITA
Sklar A (1959) Fonctions de répartition án dimensions et leurs marges. Publications de l'Institut Statistique de l'Université de Paris 8:229–231
Tettamanti T, Varga I (2014) Mobile phone location area based traffic flow estimation in urban road traffic. Adv Civ Environ Eng 1(1):1–15
Tsay RS (2005) Analysis of Financial Time Series. Wiley, New York
Tsay RS (2013) Multivariate Time Series Analysis: With R and Financial Applications. Wiley, New York
Vilar JA, Lafuente-Rego B, D'Urso P (2017) Quantile autocovariances: a powerful tool for hard and soft partitional clustering of time series. Fuzzy Sets Syst 340:38–72
Waller NG, Underhill JM, Kaiser HA (1999) A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. Multivar Behav Res 34(2):123–142
Yang C, Jiang W, Wu J, Liu X, Li Z (2018) Clustering of financial instruments using jump tail dependence coefficient. Stat Methods Appl 27(3):491–513
Yang H, Wang M-h, Huang N-j (2020) The α-tail distance with an application to portfolio optimization under different market conditions. Comput Econ, pages 1–30
Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. J Comput Graph Stat 17(2):492–514