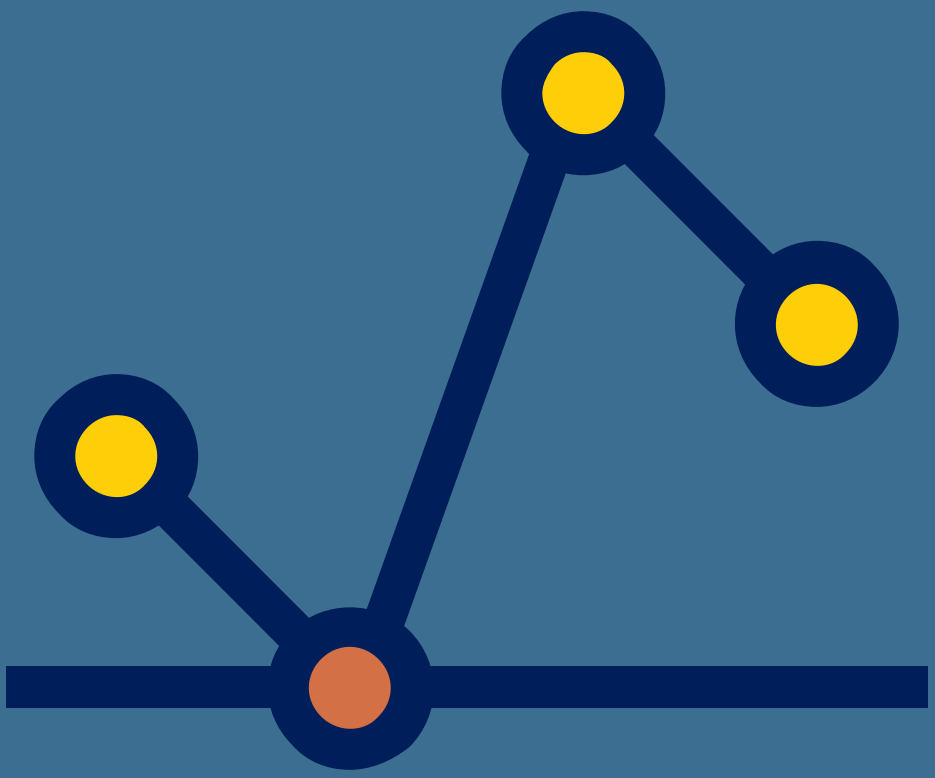
Edited by
Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



PaviaUniversityPress

Copertina: Cristina Bernasconi, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

Modeling and clustering of traffic flows time series in a flood prone area

Modelli statistici e clustering per serie temporali dei flussi di traffico in un'area soggetta a inondazioni

Paola Zuccolotto, Giovanni De Luca, Rodolfo Metulini and Maurizio Carpita

Abstract Time series of traffic flows, recovered by mobile phone origin-destination signals, are used to monitor mobility and crowding in an area subject to flooding risk. We propose a time series model based on vector autoregressive with exogenous covariates combined to dynamic harmonic regression and a subsequent clustering procedure, aimed at obtaining groups of areas characterized by the common tendency to the occurrence of extreme events, that in this case study are extremely high incoming traffic flows.

Key words: Flooding risk, multivariate time series modelling, copula functions, tail dependence, time series clustering

1 Introduction

It is well known that extreme weather events often have huge social consequences for communities and individuals. Their immediate effects consist of loss of human life, devastation of crops, damage to goods, and deterioration of overall health and wealth conditions. Considered their social and economic impact, the statistical study of extreme weather phenomena can also be approached from a management perspective. In fact, natural Disaster Management ([7]) recommends the development

Maurizio Carpita, Paola Zuccolotto
Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 25122, Brescia, Italy e-mail: maurizio.carpita@unibs.it, paola.zuccolotto@unibs.it

Giovanni De Luca
Department of Management and Quantitative Sciences, University of Naples Parthenope, Via G. Parisi, 13, 80132 Naples, Italy e-mail: giovanni.deluca@uniparthenope.it

Rodolfo Metulini
Department of Economics, University of Bergamo, Via Caniana, 2, 24127, Bergamo, Italy e-mail: rodolfo.metulini@unibg.it

of a framework of exposure risk that can exploited in an early warning perspective. In this work we focus attention on floods. To draw flooding risk exposure maps is of fundamental importance, in order to face, in the best possible way, a flooding event. Such maps cannot ignore human presence and people mobility, but they traditionally assume a constant crowding over time. This assumption is far from reality, especially in metropolitan areas, so, a more detailed description of people presence and mobility is a critical issue to determine an accurate flooding risk. To do that, mobile phone network data have been used to obtain a dynamic monitoring of crowding in areas with hydrogeological criticality ([1]). Another approach consists of using mobile phone origin-destination signals, in order to recover information on traffic flows and then build statistical models, able to give accurate forecasts of people mobility. [6] proposed a model, based on combining vector autoregressive with exogenous covariates and dynamic harmonic regression. They applied the method to the case study of Mandolossa (an urbanized area subject to flooding, located on the western outskirts of Brescia) using hourly data from September 2020 to August 2021. The method worked quite well, but residuals exhibited a leptokurtic distribution with heavy tails determined by a number of extreme events (i.e., days with particularly high or low traffic flows). In this talk we propose to use the method of time series clustering based on copula functions, proposed by [3], in order to cluster the residuals time series with respect to their upper tail dependence. The aim of the analysis is to obtain clusters of areas for which extreme events (in terms of extremely high traffic flows) tend to occur together.

2 Data

Mobile Phone origin-destination (OD) data flows have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it) for the MoSoRe Project 2020-2022 and they refer to one year of hourly observations (from September 1st, 2020 to August 31st, 2021) of traffic among *Aree di CEnsimento* (ACEs) in the province of Brescia. OD data refer to the number of phone SIM cards connected to the TIM network that were retrieved during a 1-hour interval by the antenna in a given ACE i and, after five or more minutes, by the antenna in ACE j. [1] For each time interval t, and for selected ACEs (let say, i and j) in the province of Brescia, three types of flows are available: flows arriving in i (inflows), flows departing from i (outflows), and internal flows from i to i (internal).

[1] Two types of cards can be distinguished: human SIM (about 85% of the total SIM) and M2M technology machine SIM (about 15%). Since a user might have both a human SIM and some devices with an M2M machine SIM, we restricted our attention to human SIMs to avoid double counting of users.

3 Time series modeling

In this work we limit our attention to 4 ACEs inside the flood prone area of the Mandolossa and to 38 properly selected neighbour ACEs. Specifically, with the final aim of obtaining uncorrelated estimated residuals to whom perform the clustering, we estimate, for each single neighbour ACE j the following VAR model with eXogenous variables (**VARX**, [9]):

$$\mathbf{y}_{t,j} = \nu_j + \sum_{h=1}^{p} \mathbf{A}_{h,j}\mathbf{y}_{t-h,j} + \mathbf{B}_j\mathbf{x}_{t,j} + \varepsilon_{t,j}\,, j = 1,\ldots,38\,, \tag{1}$$

where $\mathbf{y}$ is a vector of length 3 made of inflows to i (where i is represented by the union of the 4 ACEs inside the Mandolossa), outflows from i and internal flows in i, and where $\mathbf{B}\mathbf{x}_t$ contains a two-way (i.e., daily and weekly periodicities) Dynamic Harmonic Regression (DHR) component (which is based on a combination of sine and cosine Fourier bases) and proper weekdays and month dummy variables.

The model recalls that used in [6], but lags of order smaller than 24 are here allowed. According to an AIC criterion, we model the DHR component by including 7 daily and 4 weekly Fourier bases. We then calibrate the model by choosing the autoregressive (AR) order based on the AIC, the Auto Correlation Function (ACF), the Partial ACF and the Ljung-Box test. After having tested different AR structures, we opted for a model with the first 25 lags (i.e., $p = 25$), that display a very limited autocorrelation with small values of ACF and with the Ljung-Box test almost always rejected (by varying the AR order). The final model has been used to obtain estimated residuals. Despite the analysis of all estimated residuals might be interesting, in this application we just use the ones related to the inflows, as they allow to cluster ACEs in terms of the dynamic of traffic to the area of the Mandolossa.

4 Time series clustering on upper tail dependence

In this Section we describe the clustering procedure we propose to define groups of time series for which extreme events (in this case, extremely high traffic flows) tend to occur together. To do that, we rely on the method originally proposed by [3], where time series clustering is performed on a dissimilarity matrix based on bivariate tail dependence coefficients, estimated by means of copula functions. A 2-dimensional copula ([8]) is a function denoted by

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

Given the 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,

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

that is

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

Then

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

Copula functions describe the joint distribution in a very flexible way, by combining the univariate marginal distributions of the variables and a copula function joining the margins. When a joint distribution is described by means of a copula function, some interesting features of the multivariate distribution can be easily recovered. Examples are the tail dependence coefficients (TDCs): given two variables X_j and X_j, the lower and upper TDCs are given, respectively, by

$$\lambda_{j|h}^L = \lim_{v \to 0^+} P(U_j \leq v \mid U_h \leq v)$$

and

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

In case of tail independence, λ^L (λ^U) is null, while, when λ^L (λ^U) is in the range $(0, 1]$ then the extremely low (high) values of the two variables are dependent, with stronger dependence as the coefficient value increases.

In this work we are interested to upper tail dependence, as events to be monitored are exceptionally high traffic flows. To cluster times series based on upper TDCs, the procedure proposed by [3] requires to obtain the Δ^S dissimilarity matrix Δ, containing the dissimilarities δ_{jh} between all the pairs of the N time series under study, with

$$\delta_{jh} = -\log(\lambda_{j|h}^U). \tag{2}$$

The dissimilarity matrix Δ_N is then used as a basis for the adopted clustering algorithm [3, 4, 2]. In this work we propose a clustering algorithm able to take into account, beyond dissimilarities, the spatial contiguity between areas. So, we introduce a new dissimilarity measure δ_{jh}^θ as a modification of (2),

$$\delta_{jh}^\theta = -\log(\lambda_{j|h}^U) + \theta c_{jh}, \tag{3}$$

where c_{jh} is a contiguity coefficient assuming value 0 when the jth and hth time series denote traffic flows coming from neighbouring areas, and 1 otherwise. The dissimilarity matrix obtained by (3) is denoted by Δ^θ.

The parameter $\theta > 0$ adjusts the impact of the contiguity coefficient in the dissimilarity between two time series, and has to be determined through an iterative procedure, as detailed in Algorithm 1.

Note that the quality of the clusterization is evaluated by the adopted internal clustering validation indices with reference to the dissimilarity matrix Δ. The rationale for this choice is that contiguity between areas is used to define a set of optimal clusterizations at given values of θ, but the final choice among them is done by

Algorithm 1 Upper tail dependence clustering with spatial structure

Require: Two dissimilarity matrices Δ and Δ^{θ}, obtained as in (2) and (3).

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 the method proposed by [5]
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

selecting the one that ensures the best separation among clusters, only in terms of upper tail dependence. We carried out the procedure on the data described in Section 2. The estimated standardized residuals of model (1) applied to the 38 traffic flow time series have been used to derive the corresponding distribution functions $\hat{U}_{jt}$. For each of the $(38 \times 37)/2 = 703$ pairs $(\hat{U}_{jt}, \hat{U}_{ht})$, we estimated by Maximum Likelihood a set of elliptical and Archimedean copulas and selected the best one according to AIC. Once obtained the corresponding estimates of the upper tail dependence coefficients, we carried out the clustering procedure of Algorithm 1 with $\Theta = \{0.005, 0.01, 0.015, \ldots, 4\}$ and using a hierarchical agglomerative algorithm with complete linkage. As internal clustering validation indices we adopted the Average silhouette width, the Dunn index, the Calinski and Harabasz index, that all suggested an optimal value of θ around 0.04 (Figure 1).

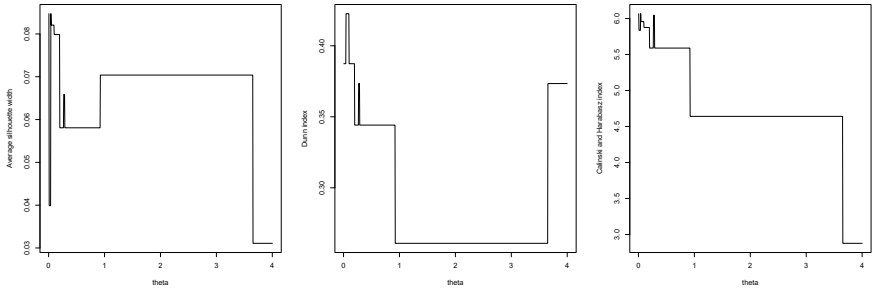


Fig. 1 Graphics of the values of the internal clustering validation indices versus θ (point 7 of Algorithm 1)

With $\theta = 0.04$ the areas turn out to be divided into 4 clusters, as displayed in Figure 2. We can observe that ACEs are grouped in a quite strong spatial neighbourhood structure, with extreme events occurring together in time in geographically contiguous areas. We found a group (coloured in blue) of ACEs located in the south outskirt of the Mandolossa. Those ACEs are characterized by a strong amount of

streets going to the Mandolossa. Belongs to a second cluster (in purple) many of the ACEs that are not contiguous to the Mandolossa but with large streets connecting them with the Mandolossa itself. The other two clusters contain only a few ACEs, with Caino being a group by itself.

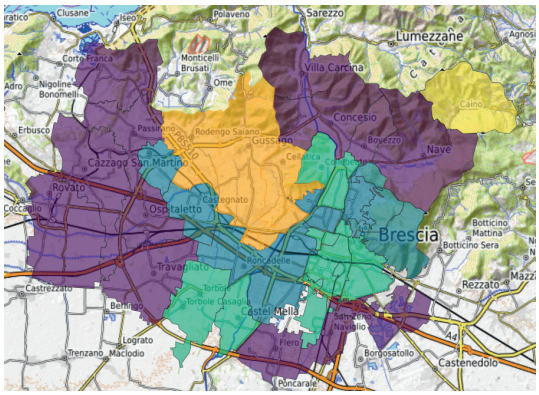


Fig. 2 Map of the 38 *Aree di Censimento* grouped with the upper tail dependence clustering with spatial structure applied to estimated residuals of flows to the Mandolossa (depicted in orange).

Acknowledgements his contribution has been developed for the Spoke 7 "CCAM, Connected Networks and Smart Infrastructure" of the Sustainable Mobility Research Center funded by the Italian NRRP (National Recovery and Resilience Plan) 2021-2026, which is part of the Next Generation EU (NGEU) Programme.

References

1. Balistrocchi M., Metulini R., Carpita M., Ranzi R.: Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, **20(12)**, 3485-3500 (2020)
2. D'Urso P., De Luca G., Vitale V., Zuccolotto P.: Tail dependence-based fuzzy *c*-medoids clustering of financial time series, *forthcoming*.
3. De Luca G., Zuccolotto P.: A tail dependence-based dissimilarity measure for financial time series clustering. Advances in Data Analysis and Classification, **5(4)**, 323-340 (2011)
4. De Luca G., Zuccolotto P.: Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach. International Journal of Approximate Reasoning **139**, 88–103 (2021)
5. De Luca G., Zuccolotto P.: Dynamic time series clustering with multivariate linkage and automatic dendrogram cutting using a recursive partitioning algorithm, *forthcoming*.
6. Metulini, R., Carpita M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research, 1-26 (2023)
7. Mishra D., Kumar S., Hassini E.: Current trends in disaster management simulation modelling research. Annals of Operations Research, **283(1)**, 1387-1411 (2019)
8. Sklar M.: Fonctions de repartition an dimensions et leurs marges. Publications de l'Institut de statistique de l'Université de Paris, **8**, 229–231 (1959)
9. Tsay, R.S.: Multivariate time series analysis: with R and financial applications. John Wiley & Sons (2013)