https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

Natural Hazards and Earth System Sciences
Discussions

Dynamic maps of people exposure to floods based on mobile phone data

Matteo Balistrocchi[1], Rodolfo Metulini[2], Maurizio Carpita[3], Roberto Ranzi[1]

[1]Department of Civil, Environmental, Architectural Engineering and Mathematics, University of Brescia, Brescia (BS), 25123, Italy
[2]Department of Economics and Statistics, University of Salerno, Fisciano (SA), 84084, Italy
[3]Department of Economics and Management, University of Brescia, Brescia (BS), 25122, Italy

Correspondence to: Matteo Balistrocchi (matteo.balistrocchi@unibs.it)

Abstract. Floods are acknowledged as one of the most serious threats to human lives and properties worldwide. To mitigate the flood risk, it is possible to act separately on its components: hazard, vulnerability, exposure. Emergency management plans can actually provide effective non-structural practices to decrease both people exposure and vulnerability. Crowding maps depending on characteristic time patterns, herein referred to as dynamic exposure maps, provide a valuable tool to enhance the flood risk management plans. In this paper, the suitability of mobile phone data to derive crowding maps is discussed. A test case is provided by a strongly urbanized area subject to frequent floodings located in the western outskirt of Brescia town (northern Italy). Characteristic exposure spatio-temporal patterns and their uncertainties were detected, with regard to land cover and calendar period. This novel methodology appears to be more reliable than crowdsourcing strategies, and has potentials to better address real-time rescues and reliefs supply.

1 Introduction

Floods are natural phenomena whose hazards afflict nearly 20 million people worldwide (Kellens et al., 2013), posing a serious challenge to the protection of human lives and the liveability of urban settlements. Both low-income and high-income countries are strongly impacted by extreme weather events and a high-confidence increase trend in the resulting economic damages and social costs has globally been documented (Kreibich et al., 2019). As reported by Munich RE (2020) over the period 1980-2019, flooding accounts for some 40% of all loss-related natural catastrophes, with losses worldwide totalling more than US$ 1tn. For instance, floods and landslides in Thailand in 2011 resulted in 43 US$ bn, that is the highest flood losses of all time.

Two major factors can be advocated for justifying such a trend: climate change and increased urbanization and people exposure (Hartmann et al., 2013). These factors involve different components of the risk concept (UN ISDR, 2009), which is given by the combination of hazard, exposure and vulnerability. Climate change was popularly acknowledged as a leading cause for the increases in the frequency and intensity of heavy storms and, consequently, of flood episodes (Solomon et al., 2007). This factor is therefore related to the hazard component of the flood risk. However, according to the IPCC (Hartmann et al., 2013), and as confirmed also by up-to-date analyses of flood intensity in Europe (Blöschl et al., 2019), the absence of a global likely trend in the incidence of floods arises. This can be due to various reasons: the high regional variability of heavy storm trends, in terms of type, magnitude and significance, as well as the strong influence played by the watershed hydrologic characteristics and the local flood management practices on the flood generation processes.

On the contrary, the population urbanization represents a non-climatic global trend. In 2008, for the first time in human history, more than half the world population was living in urban settlements and the percentage continues to augment (UN DESA Population Division, 2012). Population urbanization determines dramatic increases in people exposure and vulnerability to floods, since most of recent urbanizations are developed disregarding the natural extension of floodplains. Therefore, urbanizations often lie in flood prone areas and local communities are not able to put in place effective flood defence practices. Moreover, urbanization usually leads to the impairment of the conveyance capacity of the stream network, so that flooding areas are basically larger than in the undeveloped condition. Thus, the urbanization sprawl results in increased damage to communities, private properties and public infrastructures, which is defined as the product of exposure and vulnerability for a given hazard. Indeed, this second factor must be regarded as the main cause for the likely increasing trend of the flood risk (Barredo et al., 2009). A number of researches on flood risk changes under economic and population growth scenarios indicate that this contribution is at least equal to, but commonly larger than, the climate change one (Feyen et al., 2009; Maaskant et al., 2009; Bouwer et al., 2010; Te Linde et al., 2011; Rojas et al., 2013).

This flood risk trend forced an unavoidable shift in the paradigms of flood defence, by recognizing that not all events can be completely controlled and that structural practices have limits (Johnson and Priest, 2008). For instance, the European Flood Risk Management Directive (European Union, 2007) acknowledges that floods cannot be stopped from occurring, and that the focus must be placed on how to mitigate the damages to flood prone communities. Hence, over last decades, flood risk management has evolved from a structural-based defence approach, aiming at decreasing the hazard component, towards a more holistic perspective (Merz et al., 2010; Arrighi et al., 2019), taking into consideration drivers and impacts of flood risk. Vulnerability and exposure were thus investigated in a deeper manner than in the past, whilst novel concepts were introduced, such as residual risk (UN ISDR, 2009), accounting for the potential structural failure of the defence system (Vorogushin et al., 2009; Schumann, 2017; Balistrocchi et al., 2019), and resilience, that is the ability to recover from a damage or to absorb an impact (Liao, 2012).

Actually, vulnerability reduction plays an essential role for successful adaptation to flood risk (Kreibich et al., 2017). Differently from the hazard, the mitigation of the expected damage’s components can be pursued by means of non-structural practices. Among them, a prime role is played by emergency management plans, which allow authorities responsible for the protection of local communities to dispatch timely and appropriate mitigation measures during the occurrence of flood episodes. The development of effective emergency management plans is closely related to the concept of authorities’ preparedness (European Union, 2007). Such plans are actually intended to provide people with early warnings, reliable real-

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

time information and to better address relief supplies and rescue efforts. In this regard, a detailed and reliable picture of the real-time spatiotemporal variability of the flood risk would be highly beneficial.

Presently, large amounts of geospatial data can be obtained from a number of sources, namely remote sensing or aircraft platforms. However, these sources yield situation snapshots, but do not provide information at the spatiotemporal resolution needed for managing urban floodings and are hardly validated in the field. To overcome these problems, the possibility of taking advantage of crowdsourcing techniques has recently attracted much attention (Mazumdar et al., 2017; Rosser et al., 2017; Hirata et al., 2018; Mazzoleni et al., 2018). These techniques have been made available by the widespread proliferation of smartphones and tablets, along with the success of social media. During emergency phases, the advantages in the real-time implementation of the emergency plan are twofold: on the one hand, a large amount of volunteered geographical information suitably georeferenced can be collected (Goodchild, 2007), guiding authorities in developing collaborative flooding maps or in estimating the number and location of exposed people; on the other hand, crucial information can be communicated to exposed people, making them aware of the actual risk magnitude and supporting their capability to face the situation. After the emergency phase, authorities can also exploit the collected data to enhance their preparedness and to better match the emergency management plans to specific real-world needs of the flood prone community.

Potentials and drawbacks of various crowdsourcing techniques have been long debated, even though crowdsourcing has already found successful applications in some weather related disasters (Poser and Dransch, 2010; Hung et al., 2016; Guntha et al., 2018). Such researches have underlined that in many cases, during the emergency phase, crowdsourced data have at least the same quality of the authoritative ones (Goodchild et al., 2017). Nevertheless, several concerns have been pointed out through analyses of crowdsourcing technique applications to real-world disasters: i) the raw data quality is generally poor because of malicious intentions of rough elements of the community or incompetence of stakeholders, so that spurious, erroneous, malformed, redundant or incomplete data must be purged out of the database to make them suitable, ii) the sample significance is basically low, owing to the limited number of exposed people actively participating in crowdsourcing, so that information of general interest must be extrapolated from an exiguous fraction of the whole population, iii) the communication network is not completely reliable, as it frequently fails or malfunctions during disaster occurrences.

Researches have therefore been addressed to filter such rumours from crowdsourced data (Han and Ciravegna, 2019). However, other approaches can be followed to develop effective dynamic information tools through the exploitation of mobile phone data collected by providers. Such data make it possible to geo-localize mobile phone users over the time, that is to derive time-dependent crowding maps. When such maps are intersected with hazard maps, showing the flooding area extensions corresponding to a selected frequency of occurrence, dynamic exposure maps are obtained. As demonstrated by Carpita and Simonetto (2014) with reference to episodic crowd concentrations due to social events, recurrent spatiotemporal patterns can be derived from mobile phone data by means of geostatistic analyses. Herein, the application of this methodology to periodic spatiotemporal variability of resident population related to home-work mobility is investigated. To do so, additional tools are developed to extrapolate the real-world population from the crowding maps of provider's clients.

A suitable case study has been identified in the western outskirt of Brescia town (northern Italy), for which a detailed knowledge of the flooding dynamics and a sizeable set of mobile phone data are available. The suggested approach made it possible to derive reliable dynamic exposure maps with respect to the land coverage and the calendar time periods, obtaining estimates of the expected number of people affected by flood hazards along with its uncertainty.

Hence, the paper is organized according to the following sections: (*i*) firstly, the innovative aspects of the geostatistic analysis methodology herein utilized are illustrated, (*ii*) secondly, the main hydraulic-hydrologic features of the analysed study area are described along with the available mobile phone data, (*iii*) the methodology application and the results are finally discussed.

2 Analysis methodology

The proposed geo-statistical approach relies on Erlang mobile phone measures, which consist in the average number of mobile phone users (MPU) bearing a connected SIM. These data are collected by mobile phone providers and recorded at constant time steps with reference to a georeferenced grid of square cells. The availability of such a kind of data is progressively arousing the enthusiasm of the urban planners' community (Becker et al. 2011; Calabrese et al., 2015), as they offer a variety of potential applications. In this study, MPU spatiotemporal variability was summarized by means of daily density profiles (*DDP*), that provides the variability within a day of MPU referred to a spatial region of interest. Such regions are inundation areas, thus expressing the spatiotemporal variability of people exposed to the flood risk. To define *DDP*, let e_{it} be the number of MPU in the *i*-th grid cell in a generic time interval *t*. Let $I_r = \{i_1, \ldots, i_m\}$ be the set of grid cells related to region *r* of interest. Furthermore, let define $T_d = \{t_1, \ldots, t_o\}$ be the set of time intervals in a day *d*. The daily density profile (DDP_{rd}) can be defined according to Eq. (1), as a vector of length *o* of values describing the sum of MPU in region *r* and day *d:*

$$DDP_{rd} = \{\sum_{l=1}^{m} e_{il,t1}, \sum_{l=1}^{m} e_{il,t2}, \ldots, \sum_{l=1}^{m} e_{il,to}\}'. \quad (1)$$

Herein, the interest lies in analyzing and classifying the occurrences in a time series of DDP_{rd}, related to a set $d = \{d_1, \ldots, d_n\}$ of *n* analyzed days. More precisely, the proposed approach firstly involves the clustering of similar DDP_{rd}, as discussed in detail in the following Section 2.1. The clustering procedure consists of two steps. In the first one, MPU spatial variability inside region *r* is considered by changing the index *i* in a R^2 *x-y* coordinate space; to do so a data reduction strategy is applied. In the second one, DDP_{rd} temporal variability is evaluated by changing index *t* in a R^1 space.

Clustering mobile phone data according to the above described procedure is not straightforward. Nevertheless, traditional techniques (Arabie and De Soete, 1996) cannot be applied. In fact, data amount to *n* observations and $p = m*o$ variables (number of MPU values each day). For instance, if one has one year of available data (i.e. $n = 365$), repeated every 15 minutes ($o = 96$) in a region covered by 500 grid cells ($m = 500$), the variable number is far larger than the number of observations ($p > n$), depicting a high-dimensional context (Donoho, 2000). When analyzing high-dimensional data, several issues, such as the curse of dimensionality (Keogh and Mueen, 2017), need to be faced. With specific regard to data clustering, this issue has been addressed by Bouveyron et al. (2007). However, as suggested by Jovi et al. (2015), high-dimensional data reduction

provides a suitable solution. To do so, an approach based on the Histogram of Oriented Gradients (HOG) is used in this paper. Therefore, data reduction acts on index i, in order to convert the support from R^2 (x-y coordinate space) to R^1.
Once DDP_{rd} are clustered in statistically similar groups, the total number of people in set T_d and in region r can be estimated and associated with descriptive bands (DB), as discussed in Section 2.2. In this regard, a crucial concern is given by the lack of MPU data from all companies providing phone services in northern Italy. To deal with this problem, as firstly suggested by Metulini and Carpita (2020), the approach proposed in this paper adopts a strategy to infer the total number of people by matching available mobile phone data to census data.

2.1 Data reduction and clustering

To cluster similar *DDP* a technique for high-dimensional data reduction is firstly adopted. Then, reduced data are analyzed by a high-dimensional data clustering. Separately for each element of set T_d (i.e. for a given t), let $\varepsilon_{it} = \{e_{1,t}, e_{2,t}, \ldots, e_{im,t}\}$' be the vector of dimension m of MPU in region r in time instant t. The aim is to reduce ε_{it} to a new set of values κ_{it}, of dimension m' $< m$, that preserves the relevant information contained in the set ε_{it}. To do so, set ε_{it}, separately for each t, undergoes a histogram of oriented gradients (HOG) feature extraction (Dalal and Triggs, 2005; Tomasi, 2012). HOG is generally applied to red-green-blue (RGB), or grey scale, images, in order to find similarities among images and to classify them. Since each value is associated with a location in the R^2 x-y coordinate space, vector ε_{it} can be viewed as a raster featuring colors expressing the magnitude of its elements. The application of HOG method, having defined our setting as above, is straightforward.
When clustering in terms of the spatial distribution of users, the interest lies in the relative distribution of MPU in region r, not in the MPU absolute amount. To be consistent with this aim, HOG was applied to a normalized vector of ε_{it}. Vector z_{it} is thus defined as $z_{it} = \{\mathrm{e}_{\mathrm{i,t}} / max_{i \in I_r}(\mathrm{e}_{\mathrm{i,t}}), \forall\ i \in I_r\}$.. In order to generate the vector of features κ_{it}, containing the relevant information in z_{it}, the HOG method firstly divides the grid's cells into a number of S smaller grids $G_1, \ldots, G_S$ ($G_i \cap G_j = \emptyset$, $\forall\ i = 1,\ldots,S$ *and* $\forall\ j = 1,\ldots,S$ *with* $i \neq j$), where $\sqrt{S}$ is a parameter to be chosen. The *direction* and the *magnitude* matrices by using two different *gradient* matrices of each grid G_i (see for details in Dalal and Triggs, 2005) are then obtained. These matrices are used to derive the histogram of gradients with k equal bins, where k is a parameter that need to be chosen. κ_{it} is the vector of features given by all the elements of the obtained histogram of gradients when stacked $\forall$ S. The length of vector κ_{it} is $S*k$. Subsequently, the vector κ_{it} is stacked over the subscript t, obtaining the vector of features κ_d for region r and day d, of dimension $S*k*o$.
κ_d contains the features (variables) undergoing the first clustering step of the method, in which the objects to be clustered, according to how MPU distribute over region r according to index i, are represented by days. For all days in the data set $d = \{d_1, \ldots, d_n\}$, κ_d is computed, and a k-mean cluster analysis (Hartigan and Wong, 1979) is performed, where the n days are the objects to be clustered. κ_d contains the values of the $S*k*o$ (with $S*k*o < m*o$) variables for day d to be attributed to a cluster. According to Hartigan and Wong criterion, the number of clusters H is chosen by analyzing the decreasing trend of the ratio between the total within sum of squares ($Tot\ within\ SS_H$) and the total sum of squares ($Tot\ SS_H$), that needs to be minimized

with respect to the number of groups *H*. For a certain *H*, total within sum of squares is defined as $Tot\ within\ SS_H = \sum_{i=1}^{H} W_{SS}(C_i) = \sum_{i=1}^{H} \sum_{\kappa_d \in C_i} (\kappa_d - \mu_i)^2$, where μ_i is the centroid vector (length *S*k*o*) for cluster *i*; total sum of squares is defined as $Tot\ SS_H = \sum_d (\kappa_d - \mu)^2$, where μ is the centroid vector for the full set of data.. At this point the elements in $d = \{d_1, \ldots, d_n\}$ (the days) have been assigned to a number of clusters $C_1, \ldots, C_H$ ($C_i \cap C_j = \emptyset$, $\forall\ i = 1,\ldots,H$ *and* $\forall\ j = 1,\ldots,H$ *with* $i \neq j$).

In the second step, to account for the MPU variability over time, we consider, for a given region *r*, as objects, the vector DDP_{rd}, considered as the collection of functional observations x_{rd} (T_d), $T_d \subseteq (t_1, \ldots, t_o)$ of length *o*, with *d* varying in $d = \{d_1, \ldots, d_n\}$ (i.e. $\sum_{l=1}^{m} e_{il,t1}$ in t_1). To do so, a model-based functional data clustering method (Bouveyron and Come, 2015) was adopted, and days *d* (cluster's objects) were grouped in terms of the *o* DDP_{rd} values (cluster's variables), separately for each cluster defined in the previous step. The aim is to consider the similarities in the functional form of the DDP_{rd}, seen as a curve of values (*y*-axis) with respect the *x*-axis (time instants). In doing so, each group's curves are modelled by their own set of distributional parameters (see for details Bouveyron and Come, 2015). The method adopted in this work is suitable for high-dimensional dataset, since the clustering process applies the criteria of the sub-space clustering (Agrawal et al., 1998), adopted to consider just the minimum number of variables needed for grouping objects, thus reducing the dimensionality. In detail, it is herein proposed to adopt the following path: *i*) anomalous DDP_{rd} are removed using functional data outlier detection by likelihood ratio test (LRT), as proposed by Febrero-Bande et al. (2008); *ii*) clustering method developed by Bouveyron et al. (2015) is applied, along with *funFEM* package in *R*.

With this strategy we aim at assign the elements in $d = \{d_1, \ldots, d_n\}$ (the days) to a number of final clusters $C_{F,1}, \ldots, C_{F,Z}$ ($C_{F,i} \cap C_{F,j} = \emptyset$, $\forall\ i = 1, \ldots, Z$, $\forall$ j = 1, …, *Z*), with $Z \geq H$. Thus, the adoption of these two steps would permit a representation of the dynamic of MPU's presences, in the form of a DDP for each group of days in region *r*, where, with representative we intend that to each group belongs days that are similar each other and dissimilar in between, in terms of index *i* (spatial distribution of MPU) and index *t* (temporal dynamic of MPU)

2.2 Population assessment

If the clustering strategy makes it possible to represent the dynamic of the presence of mobile phone users in region *r* over a set of time instants for a group of "representative" days, an additional strategy is needed to estimate the total amount of people. Indeed, in most times, data are available just for one mobile phone company. To have a reliable estimated of the total number of people that are actually present in the study area, users of other mobile phone providers must be considered, as well. Collecting all these data is either unfeasible or an unsustainably expensive. To perform a national level analysis, a convenient solution is represented by the use of the market share of the provider company, that can be applied to "correct" the DDP_{rd}. Hence, an estimate of the total number of people (e.g. let s_n to be the national level share than can assume values in the range [0,1], the correct *DDP* may be $DDP_{correct} = \frac{DDP_{rd}}{s_n}$) can be obtained. Country-level estimates are available through *Il Sole 24 Ore* newspaper (Ilsole24ore, 2017). However, market share usually varies significantly among cities, according to different social-economic characteristics of users. For instance, per-capita revenues are on average 19,514 €/year in Italy and 23,418 €/year in Brescia Municipality (data by Ministry of Economy and Finance, Department of Finance, 2016), whereas the percentages of foreigners are 8.5 % and 18.5 %, respectively (data by Italian National Institute of Statistics ISTAT, 2018). Furthermore, families featuring more than 4 people are about 21.0 % in Italy and 16 % in Brescia Municipality, while the percentage of people older than 65 is quite near to the national average of about 22.0 % (ISTAT, 2017).

Thus, to suitably estimate the market share, the smallest level of aggregation, represented by the "Sezioni di Censimento" (SC) (i.e. population census districts), was used in this study (ISTAT, 2017). The following strategy is suggested: data on the number of residents from administrative archives were compared to the number of TIM users on a residential area in late evening hours. Bering in mind the characteristics of the social dynamics of this residential areas, it is reasonable to assume that during these hours residential SC are populated only by residents. Such comparisons were performed separately for each SC using ISTAT data (*Anagrafe Comunale*).

First, since the MPU grid is made of square cells while SCs are irregular polygons, the number of TIM users belonging to each SC was estimated by intersecting these spatial data. Thus, to count the number of TIM users in each polygon the portion of the cell belonging to the SC polygon were calculated by using the function *extract* in *raster* package, *R*. Let $Cell_j$; j = $1, 2, \ldots, J_{SC}$ be the TIM cells (pixel) overlapping a chosen SC, the ratio A_j in Eq. 2

$$A_j = \frac{area(SC) \cap area(Cell_j)}{area(Cell_j)}, \tag{2}$$

which represents the portion of $Cell_j$ covered by the chosen SC (for each SC, $A_j > 0; j = 1, 2, \ldots, J_{SC}$). If $Cell_j$ is completely included in SC, then $A_j = 1$, otherwise $A_j < 1$. Let TUC_j be the density of TIM Users in $Cell_j$, the estimated number of TIM users in SC ETU_{SC} is computed as shown in Eq. (3).

$$ETU_{SC} = \sum_j TUC_j * A_j \tag{3}$$

The Estimated TIM Market Share in SC $ETMS_{SC}$ is thus given by the ratio in Eq. (4), where P_{SC} is the resident number assessed by the population census the for the SC.

$$ETMS_{SC} = \frac{ETU_{SC}}{P_{SC}} \tag{4}$$

Differently from s_n, $ETMS_{SC}$ range is not necessarily in [0,1], since ETU_{SC} could be larger than P_{SC}. An application example of this procedure can be found in Metulini and Carpita (2019b). In the count of P_{SC}, elderly people (>80 years) and children (< 11 years) were excluded, aiming at taking into consideration only people bearing a smartphone. The distribution of $ETMS_{SC}$ can be used as a proxy for the TIM market share at city level. More specifically, it appears to be convenient to use the median of the distribution of $ETMS_{SC}$, which is preferable to the mean in those cases when the distribution is asymmetric. Let *me(.)* be the median statistics, the estimate $\widehat{DDP}_{rd}$ for a given region *r* for a given day *d* is finally given by Eq. (5).

$$\widehat{DDP}_{rd} = \frac{DDP_{rd}}{me(ETMS_{SC})} \tag{5}$$

2.3 Result representation

For the sake of result interpretation, a graphical representation is herein adopted. Let consider the vector DDP_{rd} to be a curve of functional observations x_{rd} (T_d) representing the sum of MPU in region r and day d (in y-axis) with respect to time instants $T_d \in (t_1, \dots, t_o)$ (in the x-axis). The profile for representative days can be displayed by using functional box plots (FBP), the analogue of the traditional box plot for curves (Sun and Genton, 2011; 2012).

In FBP a group of curves is ordered using the concept of "band depth", a "median" value and an "envelope" is generated, that can be used to define a functional version of traditional descriptive bands. Moreover, it is possible to assign a curve to the outlier group, if it exceeds by 1.5 the envelope margins in at least one time instant.

A FBP strategy is performed separately for each final cluster. Let consider cluster h and let $d_h = \{d_{1h}, \dots, d_{nh}\}$ be the days belonging to cluster h, and let $\widehat{DDP}_{rd,h} = [\widehat{DDP}_{r,d1h}, \dots, \widehat{DDP}_{r,dnh}]$ be the matrix of dimension $o*n_h$ with a $\widehat{DDP}_{rd}$ in each column. Let consider each vector to be a curve. FBP is applied to matrix $\widehat{DDP}_{rd,h}$ to generate the profile plot estimating the dynamic of the total number of people (that we will call "city users", or simply "users") in different hours (with DB) in representative days.

3 Case study description

The study area was selected as an emblematic and widespread situation of the unacceptable high flood risk affecting the foothill zone of the Po River Valley (Lombardy Region, northern Italy). It lies in the western outskirt of Brescia town (Figure 1) and is overall included in the watershed of the Oglio River, a primary left-bank tributary of the Po River. The main drainage is supplied by the Mella River, bounding the eastern side of the study area, that is a left-bank tributary of the Oglio River. As can be seen in Figure 1, the study area features five natural streams originating from the southern boundary of the Alpine chain. From West to East, they are: Laorna, Gandovere, Vaila, La Canale and Solda.

Before the area was anthropized, most of such streams probably had been flooding the alluvial plain swamping into marshes, without a main outlet in the main river network. This was the result of both their almost ephemeral regime and the terrain endorheic morphology. As the agricultural use of the alluvial plain grooved, these streams were connected to the constructed irrigation-drainage network, in order to exploit their low flow for irrigation purposes and to drain the flood flow into the Mella River. These constructed downstream reaches feature two main drainage canals the Gandoverello canal and the Mandolossa canal, depicted in Figure 1. They have become the artificial downstream reaches of the five watersheds, providing a drainage capacity in South direction both for the mountain watersheds and low lands located in the alluvial plain. The total catchment area amounts to 112.3 km^2, featured by an average imperviousness of about 22%; further details on the watershed hydrologic characteristics are available in the supplementary material.

In particular, the Laorna drainage path was strongly manipulated by a constructed straight canal 7 kilometres long, that diverts its streamflow towards the Gandovere stream and intercepts additional surface runoff produced by the northward low land.

Both streams thus confluence into the Gandoverello canal. This was formerly the downstream reach of the Gandovere stream, whose limited conveyance capacity made it necessary to decrease the hydrologic load. Thus, a flow divider was constructed upstream of the Laorna confluence, so that almost half of the Gandovere streamflow is diverted towards the Mandolossa canal inlet by means of a diversion constructed canal that intercepts the Vaila stream. All these flow discharges, along with those coming from the La Canale stream and the Solda stream, converge into the inlet of the Mandolossa canal, which is characterized by the largest conveyance capacity in the study area.

This drainage network is also exploited by the irrigation system of Franciacorta, a vineyard-agricultural district located West of the study area, for the final disposal of the residual flow discharges. In Figure 1 two of the most important irrigation canals belonging to this system, Seriola Castrina and Seriola Nuova, are reported. Their discharges mainly affect the Gandovere and Laorna streamflows. Further contributions come from the West Brescia outskirt, in terms of both urban stormwaters and irrigation excess flows, which directly drains into the Mandolossa canal. Finally, water table resurgences of the high Po River Valley are also present downstream of the Mandolossa canal inlet. Their fresh waters are however intercepted and drained by the downstream reach of the Mandolossa canal.

3.1 Hazard mapping

Since the late 50s of the last century, this area has been subject to a deep urbanization sprawl, yielding the present land cover condition depicted in Figure 2. The dramatic increase in both the urban fabric and the industrial-commercial coverages occurred at the expenses of the croplands and the permanent crops, so that sparse and isolated fabrics have evolved in a continuous and heterogeneous urbanization. In addition, various transport units have been upgraded to speedways and two highways were constructed. The flood risk perception in this this area was historically related to the Mella River inundations, which affected the Brescia outskirt since the late 60s of the last century, until its riverbed underwent severe engineering works. Conversely, with reference to the secondary stream network, the absence of a clear risk perception allowed such an urbanization sprawl to occur regardless of the floodplain extents. As Figure 2 clearly shows, most of the urban fabric areas and the industrial and commercial settlements are adjacent to the stream network.

The increase in the land coverage yielded huge impermeability degrees for the plain watersheds. Except for a combined sewer overflow located in the West Brescia watershed and discharging into the Mella River, all the stormwaters produced by these urbanizations discharge into this secondary stream network, as well as those produced by many settlements in the Franciacorta district, which improperly exploits the irrigation system as a final receipt of combined sewer systems. Moreover, the low risk perception has led to a significant impairment of the functionality of these canals. The stream flow is now constrained into a number of narrow culverts and bridges with low decks and large piers. In addition, urbanized canals are no longer maintained, so that the riparian vegetation groves in an uncontrolled manner. The combination of the increase in the peak flow discharges and in the exposure, along with the decrease in the stream network conveyance capacity has led to a dramatic increase in the flood risk. Flood episodes explained by the secondary hydrographic network insufficiency have been observed since the late 90s, evidencing an empirical frequency of occurrence far less than 20 years. The hazard mapping was thus referred to return periods spanning from 5 years to 20 years, which are significantly less than those conventionally required in Italy for a secondary stream network to be considered verified (20-50 years). The urban areas exposed to floodings are estimated in 160 ha, 231 ha and 330 ha, with respect to 5 years, 10 years and 20 years return periods. About 30% of those areas are residential fabrics whereas the remaining 70% is given by industrial and commercial settlements.

The hazard analysis was herein conducted by using a design event method. As demonstrated by Balistrocchi et al. (2013) in this climatic context, a design event method is capable to provide results comparable to those of more sophisticated continuous approaches, if it is based on the Chicago synthetic hyetograph with duration equal to the double of the catchment time of concentration. A classical leaf hydrologic model was developed in accordance with the sub-catchments subdivision illustrated in Figure 1. Extensive surveys of the stream cross sections and the inline structures were carried out to assess the actual conveyance capacities of the stream reaches. Flooding volumes were hence estimated by limiting the flood hydrographs generated through the hydrologic model to the overflow threshold discharges. Surveys of the historical flooding extensions occurred during the last three decades addressed the delimitation of the flood prone areas. Total exposed urban areas amount to about 160 ha (return period 5 yr), 230 ha (return period 10 yr) and 330 ha (return period 20 yr). Most of such areas are devoted to industrial-commercial settlements (70 %), whereas the remaining part is residential fabrics of medium-low density.

The resulting flood hazard map is reported in Figure 2 along with the land cover, and highlights the large amount of residential fabrics, industrial and commercial settlements potentially affected by storm events featuring low-medium return periods. An unacceptably high flood risk is therefore evidenced for the study area. Most of such areas were too dispersed or small to have suitable intersections with MPU grid cells. Therefore, they were grouped in four macro-areas referred to the river network and distinguished between urban fabrics (red) and commercial-industrial settlements (dark green). As shown in Figure 2 they are: Laorna and Gandovere streams confluence (areas 1 and 5), La Canale and Solda streams (areas 2 and 6), southern Gandovere canal (areas 3 and 7), Mandolossa canal in Roncadelle Municipality (areas 4 and 8).

3.2 Available mobile phone data

This work focuses on mobile phone data provided by Telecom Italia Mobile (TIM), which is currently the largest operator in Italy in this sector. According to national economic newspaper, TIM national share amounted to 30.2 % in December 2016 (*Il Sole 24 ore*). In our analysis, Erlang measure data represent the average number of mobile phone SIMs (both calling and not calling) that are assigned to that grid's cell in that quarter. Erlang measure data have already been used in the context of urban planning along with statistical methods by Carpita and Simonetto (2014) and Metulini and Carpita (2019a), who analyzed the presence of people during big events in the city of Brescia, by Zanini et al. (2016), who find, by mean of an Independent Component Analysis (ICA), a number of spatial components that separate main areas of the city of Milan, and by others (Manfredini et al., 2015, Secchi et al., 2015).

In this study, reliable Erlang measures of MPU recorded by TIM company are available. The investigated area is marked in black solid line in Figure 1 (WGS 84 UTM 32 N coordinates: 5,040,920–5,049,980N, 585,970–592,970E, area about 64 km^2) and is centered on the Mandolossa-Gandovere network. The area is covered by a georeferenced grid of square cells with 150 m sides, which provides the number of TIM users every 15 minutes. In details, for each grid's cell and for each time interval, the corresponding record refers to the average number of mobile phones simultaneously connected to the network. For instance, Figure 3 shows a detail of the spatiotemporal distribution of TIM MPUs occurred on Wednesday, November 18th 2015, in a sample area of 20x20=400 cells, near to the Mandolossa inlet. Therein, exposed areas, obtained by intersecting the urban covers with the flooding areas, were also reported. Thus, Figure 3 provides a sequence of snapshots of a dynamic map of people exposure to floods. As can be seen, the spatial distribution of raw data is realistic, as major densities suitably concentrate along the main street network and in the urban areas. The temporal variability is also reasonable; for instance, lower densities are evidenced during nighttime in industrial sectors and main streets; see, for instance, the industrial settlement near to the confluence of La Canale stream in the Mandolossa canal (flooding area marked with 6 in Figure 2). The mobility feature of these data is hidden, meaning that it is not possible to trace the path followed by a single MPU over time. Measures are available in the period 2014–2016, even though after data inspection a more limited subset was found to be suitable for the analysis (from July 1st 2015 to August 11th 2016), due to data collection issues.

4 Analysis procedure application

4.1 Procedure parameterization

The application of the HOG procedure to reduce the dataset dimensionality was performed for each quarter of day in T_d, by dividing the original grid in 9 smaller grids G_i, $i = 1, \ldots, 9$. The parameter $\sqrt{S}$ was thus set at 3. For each G_i, gradients and direction were then computed and the histogram of gradients with k=4 bins corresponding, respectively, to angles 0-45, 45-90, 90-135 and 135-180 was obtained. In general, the recognition of the analyzed object is improved by increasing the number of bins. This value was chosen in order to maximize k but, at the same time, avoiding the presence of zeros in the vector of HOG features. Extracted features count for 3^2*4 = 36 in each quarter, which correspond to a dimensionality reduction in the order of 400/36 ≈ 11. The final vector κ_d with all quarters of the same day stacked for sample area near to the Mandolossa inlet (sample area evidenced in Figure 3) and for day d amounts to 36*96 = 3456 features.

The hierarchical k-means cluster analysis, where the objects of the cluster are the days d and the variables are represented by the features of κ_d, was performed on a total amount of 360 days (d = 360) from July 1st 2015 to August 11th 2016. After data inspection, only the days of the last available year (from July 1st 2015 to August 11th 2016) were included in the analysis, since the first year (April, 2014 to June 30th 2015) features some collection problems. In effect, a configuration with 3 clusters sharply separating the days of the first year (till June 2015) and the days of the second year (by July 2015) was estimated, by performing the cluster analysis by using the full set of data. In the final sample all holidays were removed, in consideration of their specific characteristics with respect to typical days. More precisely, August, 15th, 1st and 2nd November, 8th December, 24th to 26th December, 31th December, January 1st and January 6th. 27th and 28th March (Easter), 25th April, May 1st, June 2nd were removed. In addition, those days where a large amount of data (>10%) were missing were removed, as well. Conversely, data in those days where missing data were less than 10% were maintained. A test for possible presence of curse of dimensionality, based on the distribution of the distances among pairs of objects, has been performed. A unimodal distribution that suggests the absence of such a problem was derived. On the whole, the amount of suitable data appears to be sufficient to get reliable estimates of people exposed to the flooding risk in the study area.

The number of first-step clusters was chosen according to the relative decreasing trend of the total within sum of squares with the increase in the number of groups. Figure 4 shows this trend, evidencing that a splitting in 4 clusters would decrease by half the total within sum of squares with respect to a 1 cluster splitting. Since this decrease appears to be satisfactory, the sample of days was split in $H = 4$ clusters, where, C_1 mainly corresponds to all days of July, August and September (green spine-plots shown in Figure 5), C_2 corresponds to working days from February to June (blue spine-plots shown in Figure 5), C_3 corresponds to working days from October to January (red spine-plots shown in Figure 5) and C_4 corresponds to the weekends except for those of summer (yellow spine-plots shown in Figure 5).

Hence, SMU variability over time instants was accounted for by considering the Mandolossa's DDP of each day as a functional curve. Firstly, those days that have to be considered outliers were removed by using the curve outlier detection method (Febrero-Bande et al., 2008) separately for each first-step cluster. Secondly, it was evaluated whether days should be further grouped in terms of dissimilarity in *DDP* functional curve dynamic. To do so, the assumption of independence of our functional data was tested by using Portmanteau (Gabrys and Kokoszka, 2007) and distance correlation (Székely and Rizzo, 2013) tests. Model based functional data clustering techniques (Bouveyron et al., 2015) suggests to split the "summer" group in 3 sub-groups, containing, respectively, the days of July, the days of August and the days of September. This second-step splitting leads to Z=6 final clusters, where $C_{F,1}$ includes days of July, $C_{F,2}$ includes days of August, $C_{F,3}$ includes days of September and $C_{F,4}$, $C_{F,5}$ and $C_{F,6}$ match, respectively, C_2, C_3 and C_4.

Illustrations of *DDPs* for representative days by using functional box plots are reported in Figure 6 (residential fabrics) and in Figure 7 (commercial and industrial settlements), for each of the 8 flooding areas shown in Figure 2 for the 10 years return period. Although functional data clustering suggests the splitting in six groups, for sake of clarity summer months (July, August and September) were combined in a single final cluster (Cluster 1, C1), as well as all the working-days from October to June in a single final group (Cluster 2, C2) and the weekends from October to June (Cluster 3, C3), thus leading to 3 clusters.

To extract the number of people in each quarter of each day from the grid's cells to the irregular polygon of each area (i.e. to find $DDP_{ri,d}$ for area r_i , i = 1, .., 8, and day d) the procedure described in Section 2.2 was applied. Hence, MPUs were firstly divided by a constant $c = 0.85$, in order to consider children (>12 years) and old peoples (>80 years), who likely do not have smartphones (i.e. about 85% of people are in the age range [12,80] in Brescia). Then, by estimating the median value of the market share ratio at SC level adopting the strategy in Section 2.2, which amounts to about 20%, the estimated DDP_s for each area and for each day were derived by applying Eq. (5). The estimated market share is also consistent with that found by Carpita and Fabbris (2019). Estimated $DDP_{ri,d}$ underwent the functional box plot strategy, separately for the days d in the 3 clusters (with outliers excluded) and for r_i corresponding to the 8 areas.

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

4.2 Results and discussion

Figure 6 and Figure 7 report, respectively, the resulting functional box plots for residential and for productive areas, reporting the estimated number of city users in different time instants, separately for the three clusters of days. Overall, the number of city users is lower during the first hours of the day and it increases in the morning, reaching a peak during working hours (9 am–1 pm and 2–6 pm), both in residential and in productive areas. In Moie di Sotto residential area located at the confluence of the Laorna stream and the Gandovere stream (flooding area 1 in Figure 2), people number is estimated at about 200, during the first hours of the day and during the night and increases to about 250 in working hours (inhabitant density about 25-30 ha^{-1}). The dynamic, similar in all the three clusters, shows irrelevant differences among different periods of the year.

In Villaggio Badia residential area located North of Mandolossa canal inlet (flooding area 2 in Figure 2), the city user number varies between a minimum of 1200 people and a maximum of 1400 people, during and average day (inhabitant density about 30-35 ha^{-1}). During the working-days of months from October to June (cluster *C2*), the peak reaches 1600 users. Moreover, the descriptive bands appear to be wider in summer (cluster *C1*) and on weekends (cluster *C3*) as compared to cluster 2, where bands are narrower (*i.e.* lower variability between days).

Residential areas along the southern Gandovere canal (flooding area 3 in Figure 2) are little populated. Only 50–70 users are there during an average day of summer (cluster *C1*) or during the week end (cluster *C3*) (inhabitant density about 18-23 ha^{-1}). The number amounts to more than 80 people on working hours of working days (cluster *C2*). Number of city users in Roncadelle's residential area located along the Mandolossa canal (flooding area 7 in Figure 2) is less sensitive to working hours, especially during summer. In summer the number of city users varies from a minimum of about 600 up to a maximum of 700. City users are about 800 during working hours of working days and weekends (days belonging to clusters *C2* and *C3*).

Industrial and commercial settlement of Moie di Sotto (flooding area 5 in Figure 2) feature 1000–1500 people (night, first hours of the day–working hours) in summer. These numbers increase up to about 1200–1800 in working days and to about 1100–1600 in weekends. This high density is mainly due to the presence of a commercial outlet of regional interest in this area. Industrial and commercial settlements along La Canale and Solda stream network (flooding area 6 in Figure 2) are very highly populated by city users and the difference between the number of people in summer and in working days is significant. Daily minimum and maximum values are included between 2000 and 3000 in summer, and between 2500 and 3500 during working days. Weekends follow a stable dynamic (about 2500 people all along the day).

Flooding areas related to southern Gandovere canal (flooding area 8 in Figure 2) presents a productive area with an average number of users varying from 250 to 380 in summer and from 300 to 420 during working days. In the same way of Villaggio Badia (southern part of flooding area 2 in Figure 2), the number of users along the weekends is stable and it stands to about 320/330. The industrial sector of Roncadelle (flooding area 4 in Figure 2) is another highly populated area, featuring more than 2000 people during the day. In some particular working days and weekends this number reaches about 3000 (red dashed lines of outliers in Figure 7). In this area a remarkable difference in the number of users between working days, summer days and weekends was not detected.

5 Conclusions

In this paper a novel approach to the assessment of the risk related to people exposure to floodings based on a geostatistical analysis of Erlang measures was proposed. Such a procedure takes advantage of data reduction (histogram of oriented gradients discussed in Section 2.1), in order to face the dimensionality curse issue. Its suitability and potentials were demonstrated with regard to an urban outskirt area located near to Brescia (Lombardy, Northern Italy), which is affected by widespread and frequent floodings. In Figure 3 the possibility of expressing the spatiotemporal variability of exposed people by using time variable maps was illustrated. These data feature high spatial resolution (150 m) and short time step (15'), thus providing reliable assessments even for the smallest analyzed areas (about 4–5 ha) and a precise evaluation of the temporal dynamic. Indeed, daily density profiles can be derived according to this procedure. Then, these profiles can be clustered yielding groups of similar daily time patterns. Clustering results are definitively meaningful, since working days and weekends are acknowledged to show different temporal dynamics, when they belong to working months (from October to June). Conversely, daily dynamics in summer months (July, August and September, usually exploited for the longest holydays in Italy), must be regarded as different from the others. In addition, working days and weekends feature more similar daily density profiles during such months. As can be seen in Figure 6 and in Figure 7, the daily temporal variability of people exposed to floodings can be assessed with respect to the day cluster and the type of urban areas (residential or industrial–commercial), both in terms of expected value (the median) and uncertainty (confidence band), thus providing a comprehensive information to agencies and authorities devoted to the flood risk management.

The need to assess the entire population would theoretically require the gathering of a huge amount of datasets, from all the providers that operate in the area of interest. This issue would lead to a relevant increase in the data collection cost and would be difficult to overcome. Nevertheless, census data make it possible to infer the total population from the users of a single provider, by means of local estimates of its market share, as discussed in Section 2.2.

It worth underling that this statistical support, along with the high spatiotemporal resolution and the reliability of the raw data, makes the proposed procedure particularly appealing in order to decrease the errors of exposed people estimates. Such a support is not provided by crowdsourcing techniques, which are based on voluntary data supplies and commonly rely on very limited datasets with respect to the amount of the exposed people. A second advantage that must not be disregarded lies in the possibility of exploiting dynamic exposure maps, or alternatively the clustered daily density profiles, off-line that is independently of the potential malfunctioning of the mobile phone connection during the flood episode. Conversely, crowdsourcing could be strongly compromised by the difficulties of connecting to the network during the emergency period. Indeed, dynamic exposure maps derived by mobile phone data have strong potentials to substantially improve emergency plans, so that real-time rescues, relief supplies and traffic management could be better addressed.

6 Acknowledgements

This research was developed in the course of FLOod RIsk MAnagement Policies (FLORIMAP) Project which received funding by the Cariplo Foundation under grant number 2017/0708. The content of this article does not reflect the official opinion of the Cariplo Foundation. Responsibility for the information and views expressed in this paper lies entirely with the authors.

References

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, ACM Press, 94–105, doi:10.1145/276304.276314, 1998.

Arabie, P., and De Soete, G.: Clustering and classification, World Scientific, doi:10.1142/1930, 1996.

Arrighi, C., Pregnolato, M., Dawson, R. J., and Castelli, F.: Preparedness against mobility disruption by floods. Sci. Total Environ., 654, 1010–1022. doi:10.1016/j.scitotenv.2018.11.191, 2019.

Barredo, J. I.: Normalised flood losses in Europe: 1970-2006, Nat. Hazard. Earth Sys., 9/1, 97–104, doi:10.5194/nhess-9-97-2009, 2009.

Balistrocchi, M., Grossi, G., and Bacchi, B.: Deriving a practical analytical-probabilistic method to size flood routing reservoirs, Adv. Water Resour., 62, 37–46, doi:10.1016/j.advwatres.2013.09.018, 2013.

Balistrocchi, M., Moretti, G., Orlandini, S., and Ranzi, R.: Copula-based modelling of earthen levee breach due to overtopping, Adv. Water Resour., 134, 103443, doi:10.1016/j.advwatres.2019, 2019.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., and Volinsky, C.: Route classification using cellular handoff patterns, in: Proceedings of the 13th international conference on Ubiquitous computing, 123–132, ACM press, 2011.

Blöschl, G. et al.: Changing climate both increases and decreases European river floods, Nature, 573, 108–111, doi:10.1038/s41586-019-1495-6, 2019.

Bouveyron, C., Girard, S., and Schmid, C.: High-dimensional data clustering, Comput. Stat. Data An., 52(1), 502–519, doi:10.1016/j.csda.2007.02.009, 2007.

Bouveyron, C., Come, E., and Jacques, J.: The discriminative functional mixture model for a comparative analysis of bike sharing systems, Ann. Appl. Stat., 9(4), 1726–1760, doi:10.1214/15-AOAS861, 2015.

Bouwer, L. M., Bubeck, P., and Aerts, J. C. J. H.: Changes in future flood risk due to climate and development in a Dutch polder area, Global Environ. Chang., 20/3, 463–471, doi:10.1016/j.gloenvcha.2010.04.002, 2010.

Calabrese, F., Ferrari, L., and Blondel, V. D.: Urban sensing using mobile phone network data: A survey of research, ACM Comput. Surv., 47(2), 2655691, doi:10.1145/2655691, 2015.

Carpita, M., and Fabbris, L.: The mobile phone big data tell the story of the impact of Christo's The Floating Piers on the Lake Iseo, in: ASA Conference 2019 Statistics for Health and Well-being Book of Short Papers, 2019.

Carpita, M., and Simonetto, A.: Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city, Electron. J. Appl. Stat. Anal., 5/1, 31–41, doi:10.1285/i2037-3627v5n1p31, 2014.

Dalal, N., and Triggs, B.: Histograms of oriented gradients for human detection, in: Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR ’05), doi:10.1109/CVPR.2005.177, 2005.

Donoho, D. L.: High-dimensional data analysis: The curses and blessings of dimensionality, AMS Math Challenges Lecture, 1–32, 2000.

European Union: Directive 2007/60/EC on the assessment and management of flood risks, Official Journal of European Union, 227, 27–34, 2007.

Febrero–Bande, M., Galeano, P., and Gonzalez–Manteiga, W.: Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels, Environmetrics, 19(4), doi:10.1002/env.878, 331–345, 2008.

Feyen, L., and Dankers, R.: Impact of global warming on streamflow drought in Europe, J. Geophys. Res.-Atmos., 114/D17, D17116, doi:10.1029/2008JD011438, 2009.

Gabrys, R., and Kokoszka, P.: Portmanteau test of independence for functional observations, J. Am. Stat. Assoc., 102(480), 1338–1348, doi:10.1198/016214507000001111, 2007.

Guntha, R., Rao, S., Benndorf, M., and Haenselmann. T.: A comprehensive crowd-sourcing approach to urban flood management, in: Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (LNICST), 218, 13–24, Springer Verlag, 2018.

Goodchild, M. F.: Citizens as sensors: The world of volunteered geography, Geojournal, 69/4, 211–221, doi:10.1002/9780470979587.ch48, 2007.

Goodchild, M. F., Aubrecht, C., and Bhaduri. B.: New questions and a changing focus in advanced VGI research, T. GIS, 21/2, 189–190, doi: 10.1111/tgis.12242, 2017.

Han, S., and Ciravegna, F.: Rumour detection on social media for crisis management. Paper presented at the Proceedings of the International ISCRAM Conference, 660–673, 2019.

Hartigan, J. A., and Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm, J. R. Stat. Soc., Series C (Applied Statistics), 28(1), 100–108, doi:10.2307/2346830, 1979.

Hartmann, D. L., et al.: Observations: Atmosphere and surface, In: Stocker, T. F., et al. “Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change”, Cambridge University Press, Cambridge, UK, 2013.

Hirata, E., Giannotti, M. A., Larocca, A. P. C., and Quintanilha, J. A.: Flooding and inundation collaborative mapping - use of the Crowdmap/Ushahidi platform in the city of Sao Paulo, Brazil, J. Flood Risk Manag., 11 (Supplement 1), S98–S109, doi:10.1111/jfr3.12181, 2018.

Hung, K.-C., Kalantari, M., and Rajabifard., A.: Methods for assessing the credibility of volunteered geographic information in flood response: a case study in Brisbane, Australia, Appl. Geogr., 68, 37–47, doi:10.1016/j.apgeog.2016.01.005, 2016.

Kellens, W., Terpstra, T., and De Maeyer, P.: Perception and communication of flood risks: A systematic review of empirical research, Risk Anal., 33/1, 24–49, doi:10.1111/j.1539-6924.2012.01844.x, 2013.

Keogh, E., and Mueen, A.: Curse of dimensionality, in: Sammut, C., and Webb, G. I. (eds.), Encyclopedia of Machine Learning and Data Mining, 314–315, Springer, doi:10.1007/978-1-4899-7687-1_192, 2017.

Johnson, C. L., and Priest. S. J.: Flood risk management in England: A changing landscape of risk responsibility?, Water Resour. Dev., 24/4, 513–525, doi:10.1080/07900620801923146, 2008.

Kreibich, H. et al.: Adaptation to flood risk: Results of international paired flood event studies, Earth's Future, 5/10, 953–965. doi:10.1002/2017EF000606, 2017.

Kreibich, H., Thaler, T., Glade, T., and Molinari, D.: Preface: Damage of natural hazards: assessment and mitigation, Nat. Hazards Earth Syst. Sci., 19, 551–554, https://doi.org/10.5194/nhess-19-551-2019, 2019.

Jovi, A., Brki, K., and Bogunovi, N.: A review of feature selection methods with applications, in: Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1200–1205, doi:10.1109/MIPRO.2015.7160458, 2015.

Il Sole24ore, Telefoni cellulari: è corsa al ribasso per le tariffe, 23-05-2017. URL: https://www.ilsole24ore.com/art/telefoni-cellulari-e-corsa-ribasso-le-tariffe-AESqTnQB

ISTAT: Basi territoriali e variabili censuarie, 2017. (Last accessed: 17.07.2019)

Liao, K.-H.: A theory on urban resilience to floods–A basis for alternative planning practices, Ecol. Soc., 17/4, doi:10.5751/ES-05231-170448, 2012.

Maaskant, B., Jonkman, S. N., and Bouwer, L. M.: Future risk of flooding: an analysis of changes in potential loss of life in South Holland (the Netherlands), Environ. Sci. Policy, 12/2, 157–169, doi:10.1016/j.envsci.2008.11.004, 2009.

Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., and Vitelli, V.: Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region, in: Paganoni, A. M. and Secchi, P. (eds.), Advances in Complex Data Modeling and Computational Methods in Statistics, Contributions to Statistics, Springer, Cham, 133–147, doi:10.1007/978-3-319-11149-0__9, 2015.

Mazumdar, S., Wrigley, S., and Ciravegna, F.: Citizen science and crowdsourcing for earth observations: An analysis of stakeholder opinions on the present and future, Remote Sens., 9/1, doi:10.3390/rs9010087, 2017.

Mazzoleni, M., et al.: Exploring the influence of citizen involvement on the assimilation of crowdsourced observations: A modelling study based on the 2013 flood event in the Bacchiglione catchment (Italy), Hydrol. Earth Syst. Sc., 22/1, 391–416, doi:10.5194/hess-22-391-2018, 2018.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Assessment of economic flood damage, Nat. Hazard. Earth Sys., 10/8, 1697–1724, doi:10.5194/nhess-10-1697-2010, 2010.

Metulini, R., Carpita, M.: The HOG-FDA Approach with Mobile Phone Data to Modeling the Dynamic of Peoples Presences in the City, in: Bini, M., Amenta, P., D'Ambra, A. and Camminatiello, I. (Eds.), IES 2019 Innovation & Society - Statistical Evaluation Systems at 360: Techniques, Technologies and new Frontiers Book of Abstracts, Cuzzolin Editing, 2019a.

Metulini, R., Carpita, M.: A strategy for the matching of mobile phone signals with census data, in: Arbia, G., Peluso, S., Pini, A., and Rivellini, G. (Eds.), SIS 2019 Smart Statistics for Smart Applications Book of Short Papers, 427–434, Pearson Publishing, 2019b.

Metulini, R., Carpita, M.: A Spatio-Temporal Indicator for City Users based on Mobile Phone Signals and Administrative Data, Social Indicator Research, 2020 (Online First). Doi: 10.1007/s11205-020-02355-2.

Munich RE: Risks from floods, storm surges and flash floods, underestimated natural hazards, https://www.munichre.com/en/risks/natural-disasters-losses-are-trending-upwards/floods-and-flash-floods-underestimated-natural-hazards.html, accessed in June 2020.

Poser, K., and Dransch, D.: Volunteered geographical information for disasters management with application to rapid flood damage estimation, Geomatica, 64/1, 89–98, doi:10.5623/geomat-2010-0008, 2010.

Rojas, R., Feyen, L., and Watkiss, P.: Climate change and river floods in the European Union: socio-economic consequences and the costs and benefits of adaptation, Global Environ. Change 23/6, 1737–1751, doi:10.1016/j.gloenvcha.2013.08.006, 2013.

Rosser, J. F., Leibovici, D. G., and Jackson, M. J.: Rapid flood inundation mapping using social media, remote sensing and topographic data, Nat. Hazards, 87/1, 103–120, doi:10.1007/s11069-017-2755-0, 2017.

Secchi, P., Vantini, S., and Vitelli, V.: Analysis of spatio-temporal mobile phone data: A case study in the metropolitan area of Milan, Stat. Method. Appl., 24(2), 279–300, doi:10.1007/s10260-014-0294-3, 2015.

Solomon, S., et al.: AR4 Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK, 2007.

Schumann, A.: Flood safety versus remaining risks - options and limitations of probabilistic concepts in flood management, Water Resour. Manag., 31(10), 3131–3145, doi:10.1007/s11269-017-1700-z, 2017.

Sun, Y., and Genton, M. G.: Functional boxplots, J. Comput. Graph. Stat., 20(2), 316–334, doi:10.1198/jcgs.2011.09224, 2011.

Sun, Y., and Genton, M. G.: Adjusted functional boxplots for spatiotemporal data visualization and outlier detection, Environmetrics, 23(1), 54–64, doi:10.1002/env.1136, 2012.

Székely, G. J., and Rizzo, M. L.: The distance correlation t-test of independence in high dimension, J. Multivariate Anal., 117, 193–213, doi:10.1016/j.jmva.2013.02.012, 2013.

Te Linde, A. H., Bubeck, P., Dekkers, J. E. C., De Moel, H., and Aerts, J. C. J. H.: Future flood risk estimates along the river Rhine, Nat. Hazard. Earth Sys., 11/2, 459–473, doi:10.5194/nhess-11-459-2011, 2011.

Tomasi, C.: Histograms of oriented gradients, Computer Vision Sampler, 1–6, 2012.

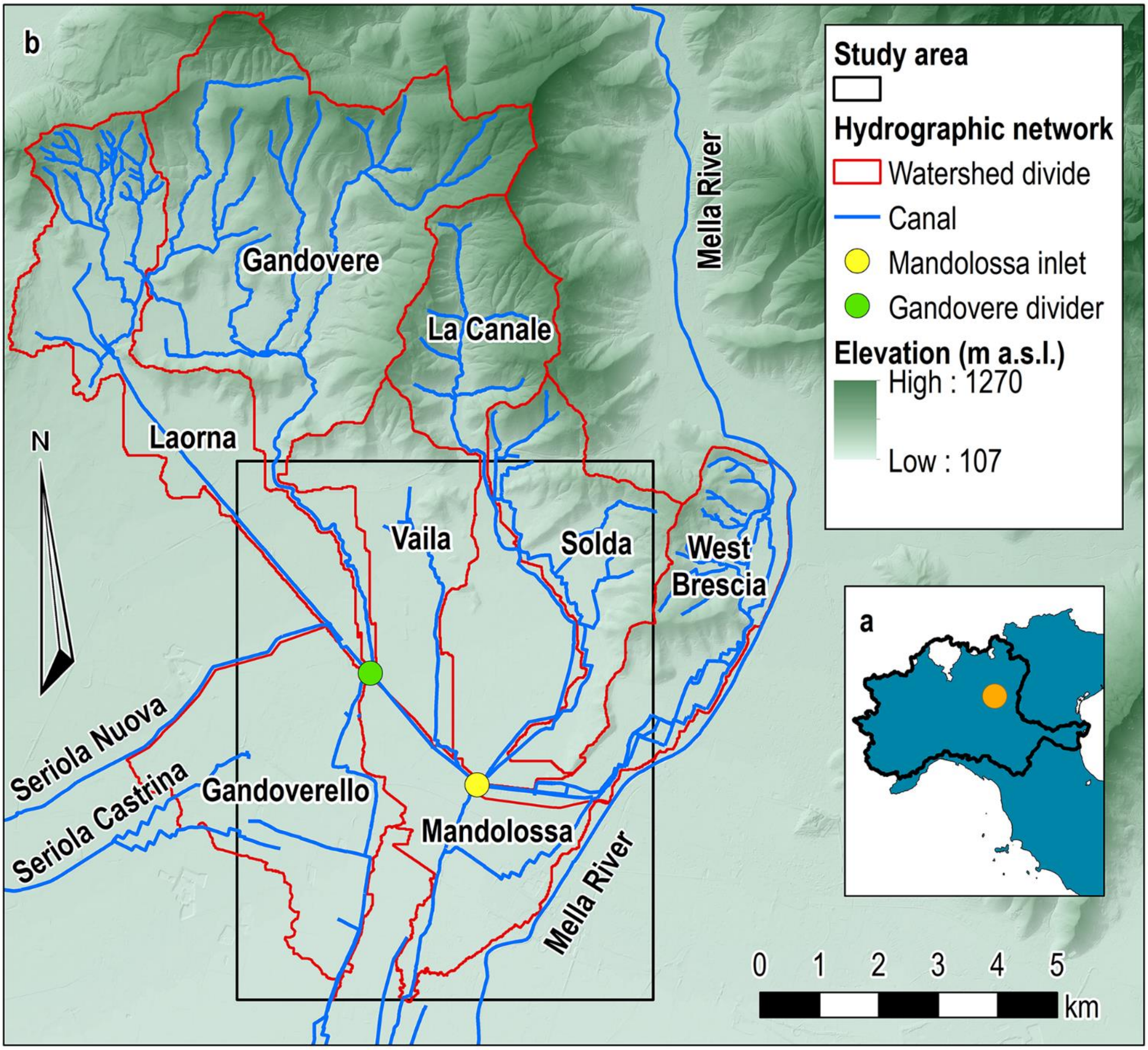
UN DESA Population Division: World urbanization prospects: The 2011 revision, United Nations Department of Social Affairs (UN DESA) Population Division, New York, USA, 2012.

UN ISDR: 2009 UNISDR terminology on disaster risk reduction, United Nations International Strategy for Disaster Reduction (UN ISDR), Geneva, CH, 2009.

Vorogushyn, S., Merz, B., and Apel, H.: Development of dike fragility curves for piping and micro-instability breach mechanisms, Nat. Hazard Earth Syst. Sci., 9/4, 1383–1401, https://doi.org/10.5194/nhess-9-1383-2009, 2009.

Zanini, P., Shen, H., and Truong, Y.: Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data, Ann. Appl. Stat., 10(2), 812–833, doi:10.1214/16-AOAS913, 2016.

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020



585 **Figure 1.** Location of the study area with respect to the Po River Valley in northern Italy *(a)*, and main hydrographic features of the foothill area west of Brescia town *(b)*; base map 5 m Digital Elevation Model provided by Lombardy Region (www.geoportale.regione.lombardia.it).

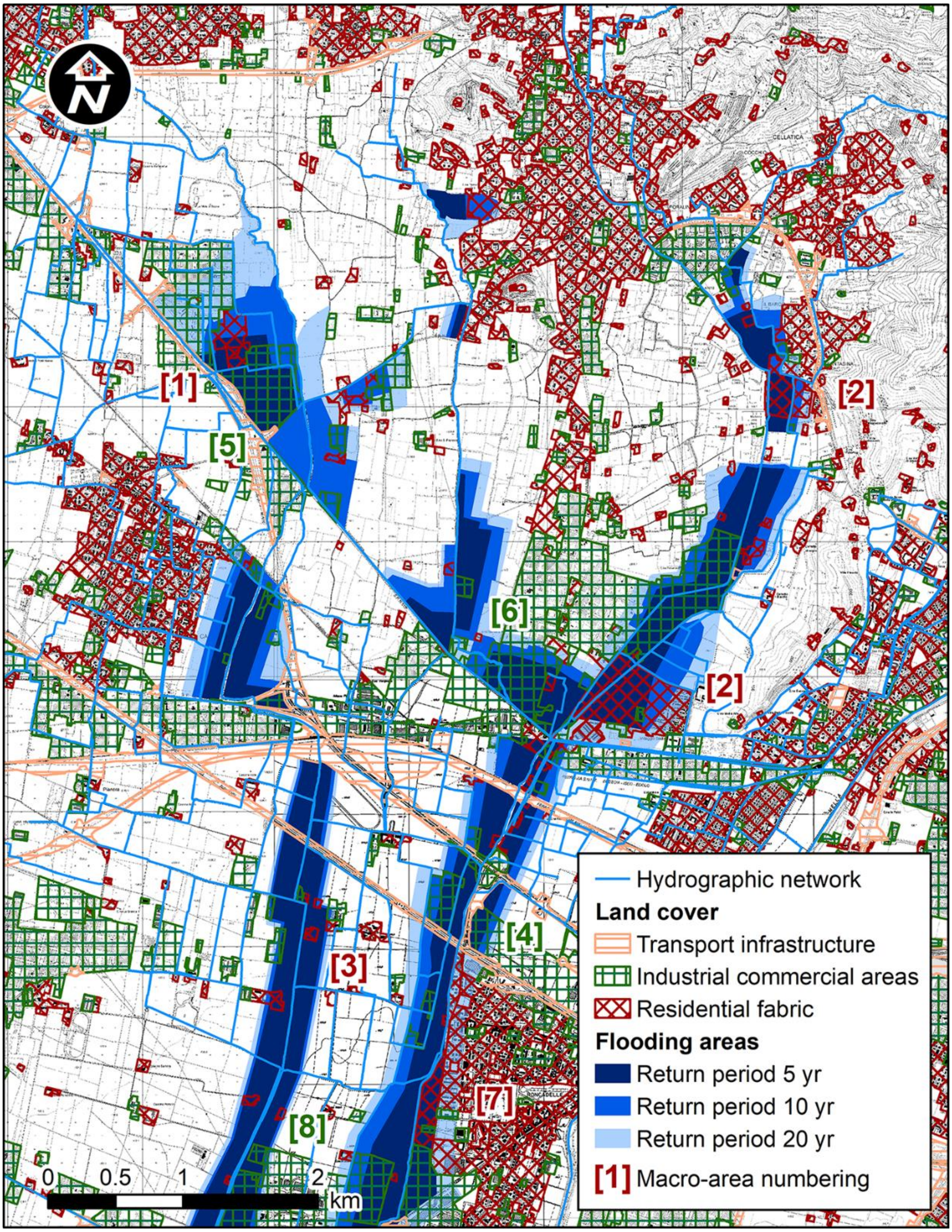


Figure 2. Flooding hazard map of the study area comparing present land cover and flooding area extensions referred to return periods varying between 5 years and 20 years; exposed residential areas are: [1] Laorna and Gandovere streams, [2] La Canale and Solda streams, [3] Southern Gandovere canal, [7] Mandolossa canal, and exposed industrial and commercial settlements are: [4] Mandolossa canal, [5] Laorna and Gandovere streams, [6] La Canale and Solda streams, [8] Southern Gandovere canal; base map Lombardy Regional Technical Map CTR 1:5000 provided by Lombardy Region (www.geoportale.regione.lombardia.it).

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

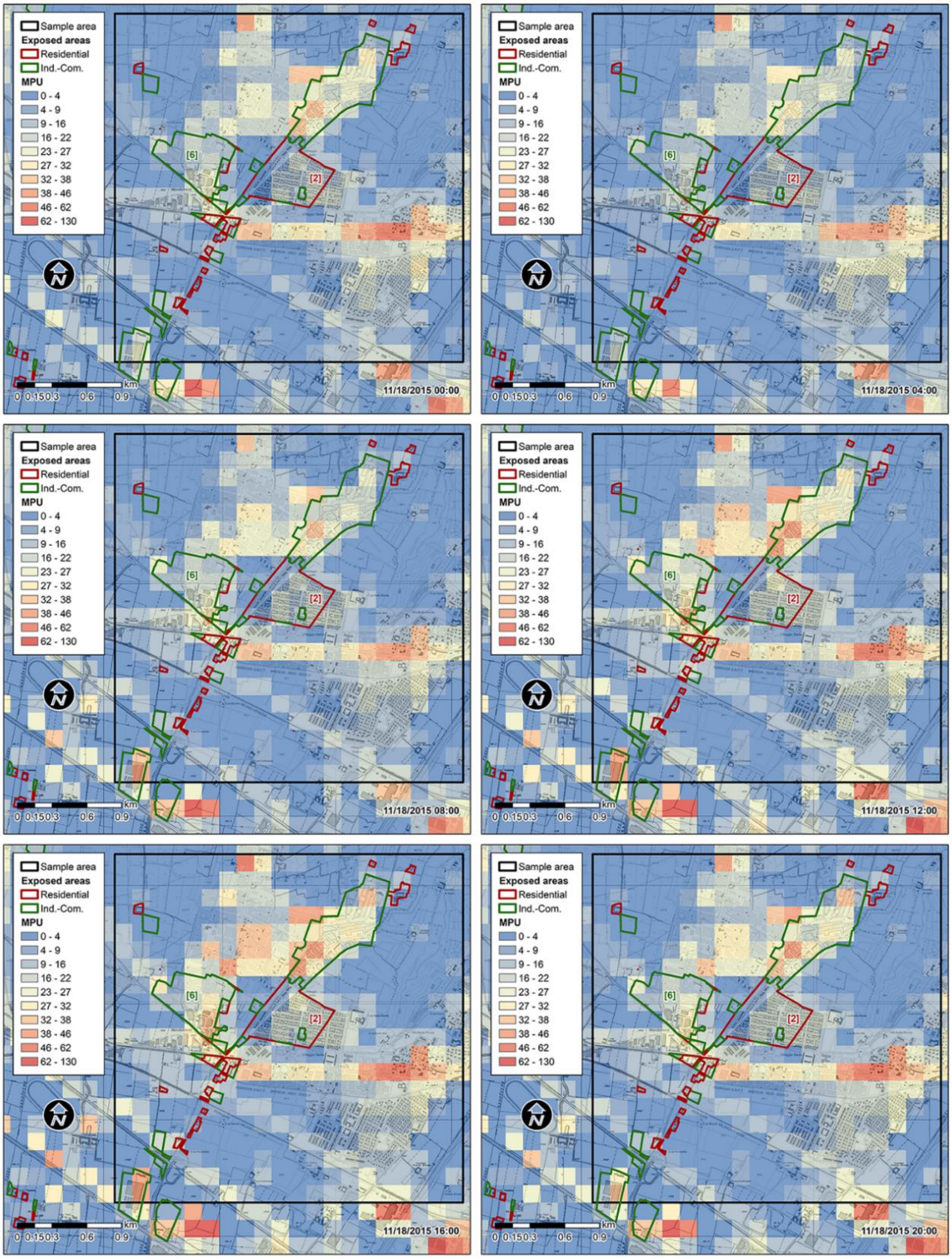


Figure 3. Snapshots of a dynamic map showing the spatiotemporal distribution of mobile phone users (MPU) occurred on 18/11/2015 (Wednesday) in urban areas exposed to 10 year return period floodings; base map Lombardy Regional Technical Map CTR 1:5000 provided by Lombardy Region (www.geoportale.regione.lombardia.it).

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

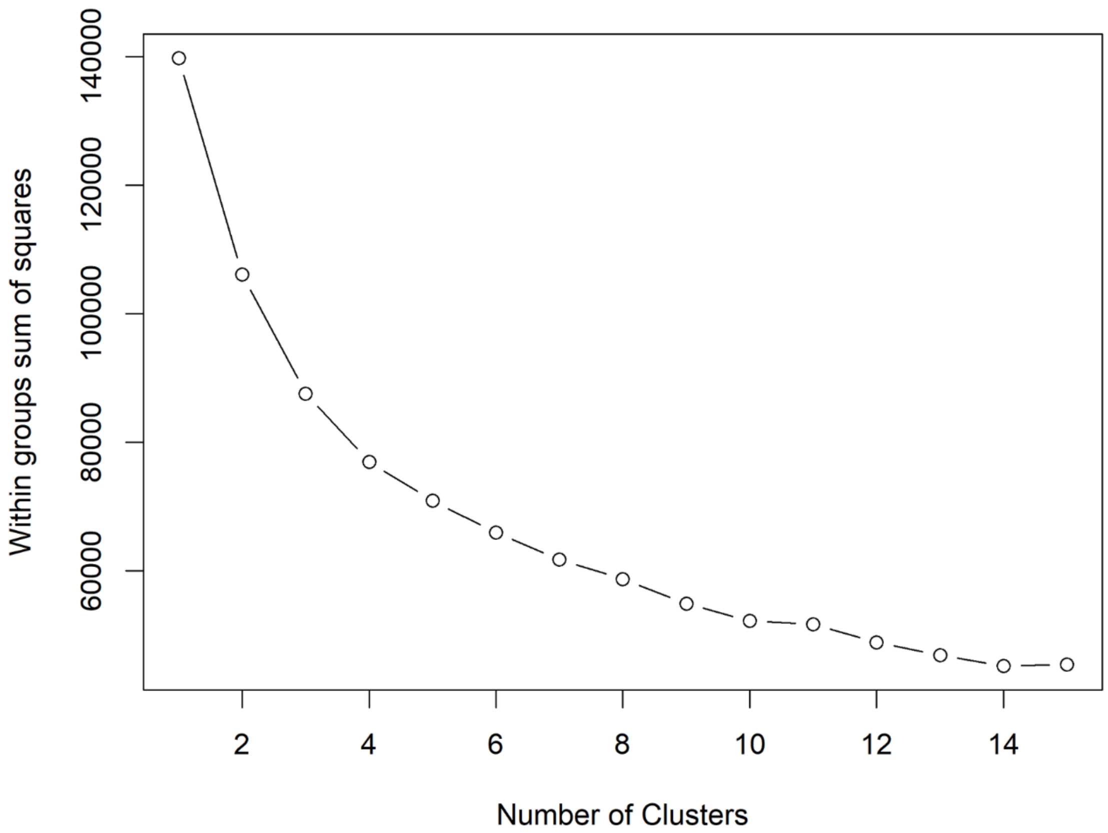


Figure 4. Diagnostic for the choice of the number of first-step clusters based on the within groups sum of squares.

600

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

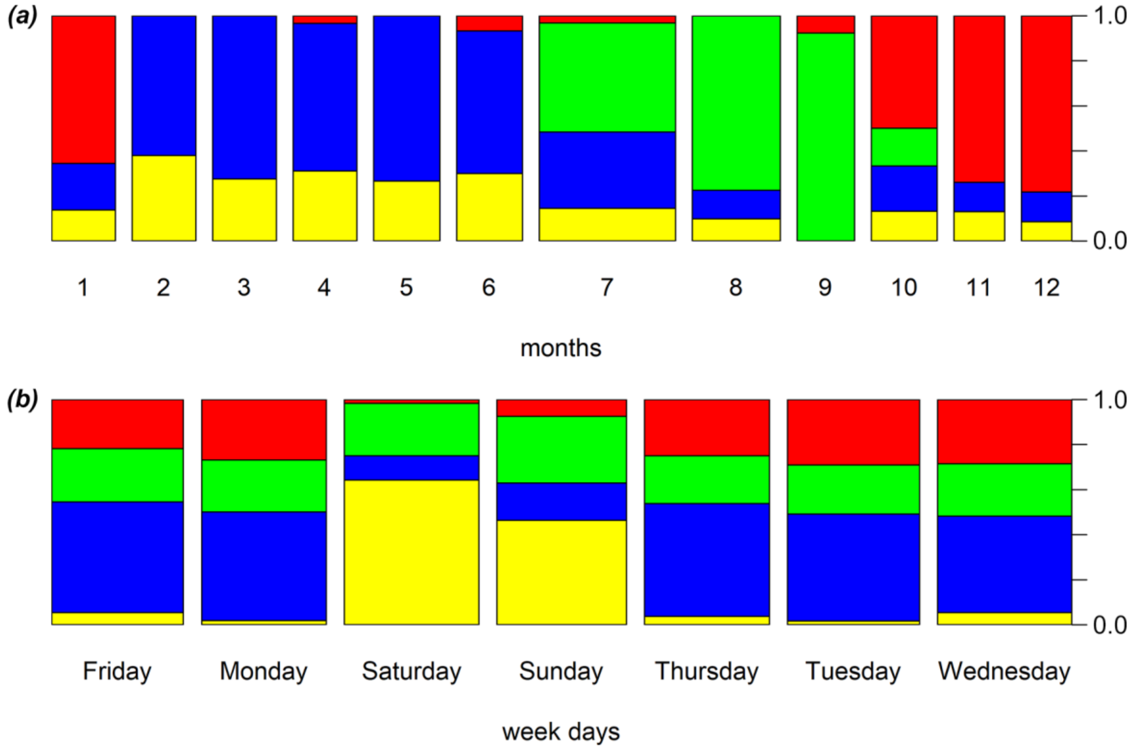


Figure 5. Spine-plots representing the first-step clustering of days along *(a)* days of the week and *(b)* months (green: all days from July to September; blue: working days from February to June; red: working days from October to January; yellow: weekends from October to June).

605

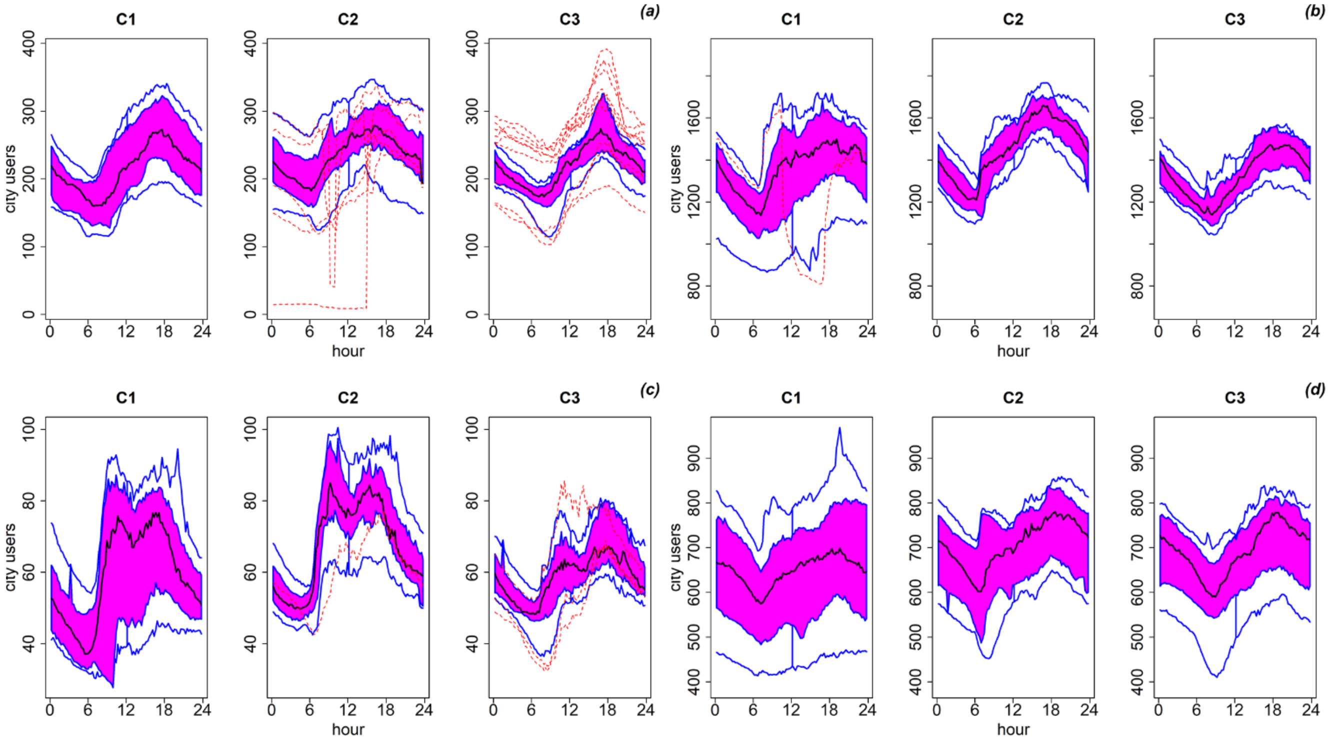


Figure 6. Functional box plots of exposed people ("city users") inside residential areas: *(a)* Moie di Sotto (area 1 in Figure 2), *(b)* Villaggio Badia and Fantasina (area 2 in Figure 2), *(c)* southern Gandovere canal (area 3 in Figure 2), *(d)* Roncadelle (area 7 in Figure 2). Cluster 1 (July, August, September, C1), Cluster 2 (working-days from October to June, C2), Cluster 3 (week-ends from October to June, C3).

610

https://doi.org/10.5194/nhess-2020-201
Preprint. Discussion started: 2 July 2020

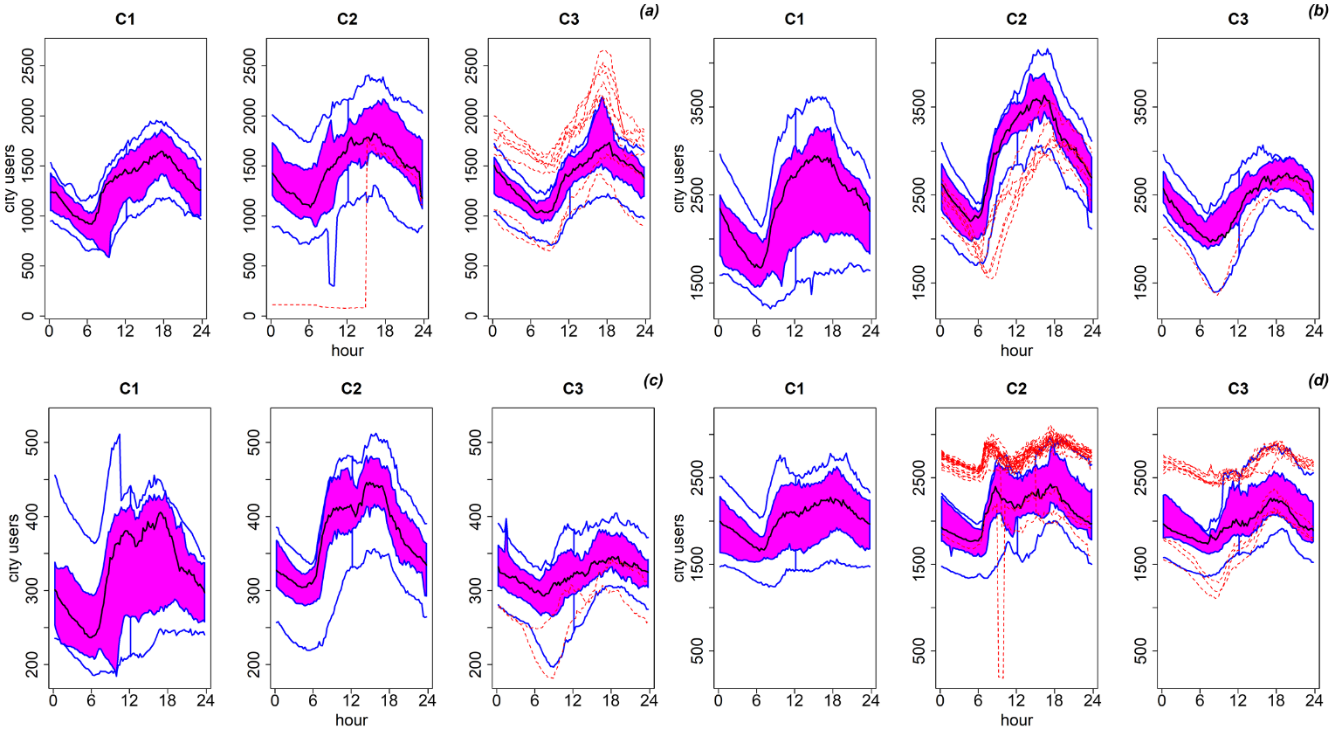


Figure 7. Functional box plots of exposed people ("city users") inside industrial-commercial settlements: *(a)* Moie di Sotto (area 5 in Figure 2), *(b)* Villaggio Badia and Fantasina (area 6 in Figure 2), *(c)* southern Gandovere canal (area 8 in Figure 2), *(d)* Roncadelle (area 4 in Figure 615 2). Cluster 1 (July, August, September, C1), Cluster 2 (working-days from October to June, C2), Cluster 3 (week-ends from October to June, C3).