MOX–Report No. 43/2012

A Case Study on Spatially Dependent Functional Data: the Analysis of Mobile Network Data for the Metropolitan Area of Milan

Secchi, P.; Vantini, S.; Vitelli, V.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

A Case Study on Spatially Dependent Functional Data: the Analysis of Mobile Network Data for the Metropolitan Area of Milan

P. Secchi [a], S. Vantini [a], V. Vitelli [b]

[a] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

[b] Chair on Systems Science and the Energetic challenge, European Foundation for New Energy – Électricité de France, at École Centrale Paris, Grande Voie des Vignes, F-92295 Châtenay-Malabry, France, and Supélec, Plateau de Moulon, 3 Rue Joliot-Curie, F-91192 Gif-sur-Yvette, France

piercesare.secchi@polimi.it
simone.vantini@polimi.it
valeria.vitelli@supelec.fr

Keywords: Spatial statistics, Functional data analysis, Treelet analysis, Voronoi tessellation, Bagging, Erlang data.

AMS Subject Classification: 62H11, 62H25, 62P30.

Abstract

We analyze geo-referenced high-dimensional data describing the use over time of the mobile-phone network in the urban area of Milan, Italy. Aim of the analysis is segmenting the metropolitan area of Milan into subregions sharing a similar pattern along time, possibly related to activities taking place in specific locations and/or times within the city. To tackle this problem, we develop a non-parametric method for the analysis of spatially dependent functional data, named Bagging Voronoi Treelet Analysis. Indeed, this novel approach integrates the treelet decomposition with a proper treatment of spatial dependence, obtained through a Bagging Voronoi strategy. The latter relies on the aggregation of different replicates of the analysis, each involving a set of functional local representatives associated to random Voronoi-based neighborhoods covering the investigated area. In the presence of spatial dependence the method appears to be both computationally and statistically efficient. Indeed results clearly point out some interesting

temporal patterns interpretable both in terms of population density and mobility (e.g., daily work activities in the tertiary district, leisure activities in residential areas in the evenings and in the weekend, commuters movements along the highways during rush hours, and localized mob concentrations related to occasional events). Moreover we perform two simulation studies, aimed at investigating the properties and performances of the method.

1 Introduction

The metropolitan area of Milan is located in Northern-Italy and, with its 7.4 million inhabitants, is the fifth largest metropolitan area in Europe after the Ruhr, Moscow, Paris, and London. It includes the nine provinces of Milan, Bergamo, Como, Lecco, Lodi, Monza-e-Brianza, Novara, Pavia, and Varese. It is characterized by a very high concentration of working activities (nearly 10% of the entire national Italian gross domestic product comes from this area, providing a per-capita GDP 50% higher than the national average) but also of residential activities (the province of Milan is the most populated in Italy, with more than 1000 inhabitants per km^2, more than five times the national average). The municipality of Milan, located in the center of the metropolitan area, and quite identifiable with the area included within the highway ring-road, is of course the main attractor of the region. Nearly 1.3 million persons live there but every working day its population increases nearly 50% since 600 thousand persons commute from the metropolitan area. This large number of commuters is mainly due to the lack of housing within the municipality; this, together with lack of important investments in improving the transport system in the last decades, has generated a critical situation in terms of mobility. Indeed most roads connecting the municipality of Milan with the metropolitan area have reached their saturation level with peaks of the traffic/capacity ratio up to 150% during rush hours (OECD 2006a,b).

Indeed the OECD identifies housing, transport, and congestion as the bottlenecks for the future growth of the Milan metropolitan area. These factors seem to badly affect the well-being of the city from many perspectives: (*i*) pollution (Milan is the second most air-polluted city in Europe after Moscow), (*ii*) economy (the difficulty in mobility of people and goods is estimated to damp the output of the area of more than 4%), and, of course, (*iii*) demography (while the population of the metropolitan area is growing the population of the municipality of Milan is decreasing; in 1971 nearly 1/2 of the population of the province of Milan lived within the municipality, in 2001 only 1/3).

In recent years a congestion charge has been introduced, the regional railway network has been fully integrated, three new highways and two new subway lines are under construction, and a few bike- and car-sharing initiatives have been promoted. The Green Move project, which the present research is part of, is among these initiatives. Green Move is an interdisciplinary research project financed by Regione Lombardia involving different research groups at the Politecnico di Milano and focused on the development of a vehicle sharing system based on the concept of "little, electric and shared vehicles". This work is a first attempt to gather information about population density and mobility in the metropolitan area of Milan from mobile network data belonging to the Telecom

Italia database. In a long term perspective this information will be used to optimally locate vehicles and docking stations of the car-sharing network. The possible use of this information is of course much wider. Large scale quantitative information on human mobility is of extreme interest to the urban planner, the traffic flow being functionally related to the porosity and the permeability of the urban texture. The same information may, of course, help the city manager, for instance to locate ambulances and police patrols at a certain time in the most suited locations, or to maximize the public exposition of alerts relevant to the well-being of the community.

In the Telecom Italia database, the metropolitan area of Milan is partitioned into a uniform lattice S_0 of 97×109 sites. In each site, the average number of mobile phones simultaneously using the network for calling is provided every 15 minutes for 14 days. This quantity is called Erlang and, at a first approximation, can be considered proportional to the number of active people in that site at that time, thus providing information about people density and mobility. Technically the Erlang $E_{\mathbf{x}j}$ relevant to the site $\mathbf{x} \in S_0$ and to the jth quarter of an hour is computed as

$$E_{\mathbf{x}j} = \frac{1}{15}\Sigma_{q=1}^{Q}|T_{\mathbf{x}j}^{q}| \ , \tag{1}$$

where $T_{\mathbf{x}j}^{q}$ indicates the time interval (or union of intervals) in which the qth mobile phone is using the network for calling while moving within site $\mathbf{x}$ and during the jth quarter of an hour; $|T_{\mathbf{x}j}^{q}|$ indicates its length in minutes. The number of potential phones using the network is indicated with Q. Equation (1) represents the formula actually in use by the mobile company for computing $E_{\mathbf{x}j}$, but its meaning is better captured by the equivalent representation

$$E_{\mathbf{x}j} = \frac{1}{15}\int_{15(j-1)}^{15j} N_{\mathbf{x}}(t)dt \ , \tag{2}$$

which shows that $E_{\mathbf{x}j}$ is the mean over the jth quarter of an hour of the number $N_{\mathbf{x}}(t)$ of mobile phones using the network within site $\mathbf{x}$ at time t, measured in minutes. The equivalence of representations (1) and (2) is easily proved through the following identities

$$\frac{1}{15}\Sigma_{q=1}^{Q}|T_{\mathbf{x}j}^{q}| = \frac{1}{15}\Sigma_{q=1}^{Q}\int_{15(j-1)}^{15j}\mathbf{1}_{\{T_{\mathbf{x}j}^{q}\}}(t)dt =$$
$$\frac{1}{15}\int_{15(j-1)}^{15j}\Sigma_{q=1}^{Q}\mathbf{1}_{\{T_{\mathbf{x}j}^{q}\}}(t)dt = \frac{1}{15}\int_{15(j-1)}^{15j}N_{\mathbf{x}}(t)dt.$$

The Erlang data we deal with are recorded every quarter of an hour, from March 18th, 2009, 00:15, till March 31st, 2009, 23:45. Indeed, in some sites of the lattice the entire temporal profile of the Erlang values is missing, while in other sites only some values are missing or non-admissible since they are negative (and they are treated as missing values). Hence, we restrict the analysis to a non-uniform time grid with $p = 1308$ elements, each element of the time grid being relative to a quarter of an hour for which an Erlang measurement has been observed in at least one site of the lattice.

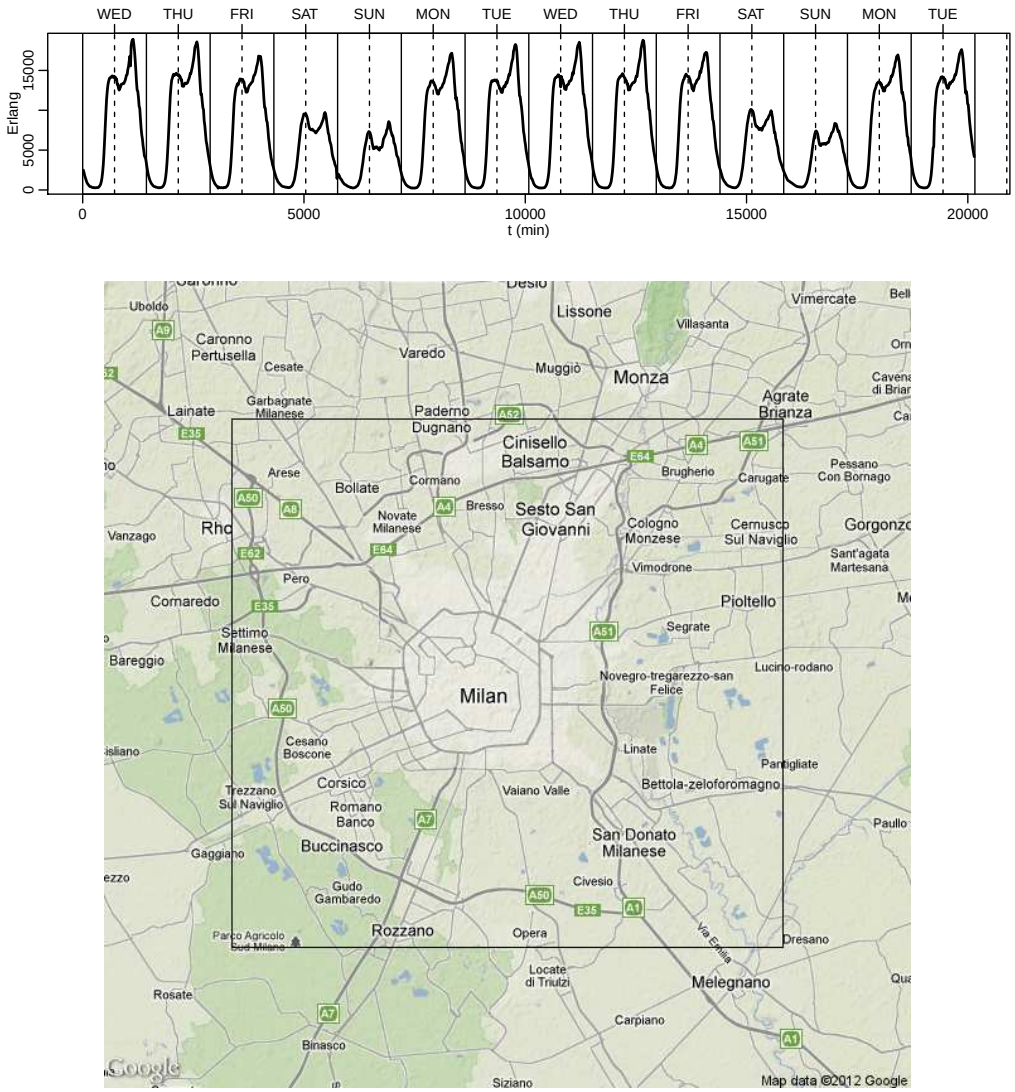

Figure 1: In the top panel, the aggregated Erlang of the investigated area as a function of time. The solid vertical lines are drawn at midnight of each day, and the dotted vertical lines at noon. The first day is Wednesday March 18, 2009. In the bottom panel, a map of the region covered by the lattice of the Telecom dataset.

The lattice S_0 covers an area of 757 km^2, reported in the bottom panel of Figure 1, and included between latitude 45.37° and 45.57° North and longitude 9.05° and 9.35° East. It is divided in $|S_0| = N = 10573$ approximately rectangular sites of size 232m $\times$ 309m. Overall, 13·829·484 records are available, among which 110·475 are missing. The data set at hand can be genuinely considered an instance of spatially-dependent functional data, because of the high within-unit sample size and the very high signal-to-noise ratio. To have a first idea of these data, in the top panel of Figure 1 the aggregated

Erlang for the investigated area, $\sum_{\mathbf{x}\in S_0} E_{\mathbf{x}j}$, is reported as a function of time, measured in minutes. A first inspection shows some global features such as the day/night effect and working/weekend day effect. The aim of the analysis is indeed to identify these global features together with the local ones, more subtle to detect and possibly associated to particular subregions of the investigated area.

The Erlang data are progressively arousing the urban planner community to enthusiasm (Becker et al. 2011; Calabrese et al. 2011). In this work we aim at using Erlang data for segmenting the metropolitan area of Milan into subregions that share the same activity pattern along time in terms of population density and mobility. To our knowledge, this is a first attempt at the exploration of Erlang data with the methods provided by Functional Data Analysis (Ramsay and Silverman 2005), and their extension to the case of spatially dependent data sets.

The rest of the paper is structured in five sections. In Section 2 the methodology used to perform dimensional reduction of spatially dependent functional data is presented: in particular we propose to integrate a treelet analysis (Lee et al. 2008) with a Bagging Voronoi strategy for the exploration of spatial dependence (Secchi et al. 2012). In Section 3 the site-wise temporal smoothing of time-varying Erlang data through a suitable Fourier expansion is described. In Section 4 the results of the analysis of the Telecom Italia database are shown. In Section 5 we describe a simulation study conducted to address some specific methodological issues. Finally, in Section 6, we draw some conclusions and we trace possible directions for future research.

2 Data Analysis: Methodology

The prerogative of Erlang data is they can give insight on different aspects of the urban area they are referred to, and their analysis can be developed with various scopes: the segmentation of the area into districts characterized by homogeneous telephonic patterns; the identification of a set of "reference signals" able to describe the different temporal patterns of use of the mobile phone network; the description of the influence of each detected telephonic pattern in each site of the lattice.

Let $\{E_{\mathbf{x}}(t)\}_{\mathbf{x}\in S_0}$ be the collection of time profiles of the Erlang data, along the sites of the lattice S_0. Details on the smoothing preprocessing of these functional data are described in Section 3. We claim that a few time-varying functions – let us say K – coupled with surfaces are sufficient to capture all information carried by the Erlang profiles. Moreover, they are apt to the segmentation of the area under investigation in terms of subregions sharing the same pattern along time with respect to mobile phone activity, which in turn depends on population density and mobility. The K surfaces express the impact in the area of the coupled time-varying functions, each describing a time profile for mobile phone activity. The other way round, the K time-varying functions express the evolution in time of the coupled surfaces, which provide a segmentation of the area into homogeneous subregions with respect to mobile phone activity. More precisely, these considerations can be formalized in the following model for the exploration of

Erlang data

$$E_{\mathbf{x}}(t) = \sum_{k=1}^{K} d_k(\mathbf{x})\psi_k(t) + \varepsilon_{\mathbf{x}}(t), \tag{3}$$

for $\mathbf{x} \in S_0$, $t \in [0,T]$, and where $\varepsilon_{\mathbf{x}}(t)$ is a noise term such that $\mathbb{E}[\varepsilon_{\mathbf{x}}(t)] = 0$ and $Var[\varepsilon_{\mathbf{x}}(t)] = \sigma^2$ independently from spatial or time coordinate. The set $\{\psi_1(t), \ldots, \psi_K(t)\}$ is that of the time-varying functions while $\{d_1(\mathbf{x}), \ldots, d_K(\mathbf{x})\}$ represent their coupled surfaces. Estimation and interpretation of these two sets of functions is of interest to the urban planner, since they jointly describe both people behavior in time and people mobility in space. In the following subsections we will describe a possible strategy for the identification of these two sets of functions. It is worth stressing that model (3) inevitably implies that the functions ψ and the surfaces d are identifiable only up to a multiplicative factor, exactly like loadings and scores in functional principal component analysis. Indeed this fact is not relevant to our analysis, whose aim is exclusively focused on the segmentation of the area into homogeneous subregions, and on the identification of their activation profiles along time.

2.1 Dimensional reduction: a short tutorial on Treelet Analysis (TA)

We first consider the problem of estimating the set of functions $\{\psi_1(t), \ldots, \psi_K(t)\}$. This means finding a parsimonious description of the sample of Erlang profiles $\{E_{\mathbf{x}}(t)\ \}_{\mathbf{x}\in S_0}$ via a finite set of reference profiles. For the moment we are not considering the spatial dependence which is intrinsic in our data set: we will deal with it in the next subsection.

A possible approach to dimensional reduction of functional data is the use of a *treelet basis*, introduced in Lee et al. (2008). This data-driven basis seems the most suited to the analysis of Erlang data, which present extremely localized functional features. Treelets have been originally designed and developed for treating sparse unordered data. Their property is to have a hierarchical structure, since they are a multiscale orthonormal basis indexed on a hierarchical tree. Indeed, as in multi-resolution analysis, treelets provide a set of "scaling functions" defined on the nested subspaces $\mathbb{R}^J = V_0 \supset V_1 \supset \cdots \supset V_J$, and a set of orthogonal "detail functions" defined on residual spaces $\{W_j\}_{j=1}^J$, where $V_j \oplus W_j = V_{j-1}$ for all $j = 1, \ldots, J$. We remark that treleets are very close to wavelets, even though they are not a wavelet basis. Indeed, in treelet computation, the wavelet approach is mixed with principal component analysis, which is hierarchically performed on the couple of most correlated variables at any given level. At each level of the tree, these are identified and replaced by a coarse-grained *sum variable*, and by a residual *difference variable*: the new variables are computed by a local principal component analysis in two dimensions. Difference variables are then stored, and only sum variables are processed at higher levels of the tree.

More precisely, consider a generic functional sample $\chi_1, \ldots, \chi_N$ and J time instances $t_1, \ldots, t_J$. The algorithm described in Lee et al. (2008) is initialized with the sample design matrix $\mathbb{X} \in \mathbb{R}^{N\times J}$. In our functional specification, $\mathbb{X}$ is the evaluation matrix obtained by setting $\mathbb{X}_{ij} = \chi_i(t_j)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. In the language of treelet analysis, for each $j = 1, \ldots, J$, we interpret $\chi_1(t_j), \ldots, \chi_N(t_j)$ as a sample from

the variable $\chi(t_j)$. Note that in most functional data analyses, each function χ is observed only at discrete time points, often with error. It is thus common to identify the set of functions $\chi_1, \ldots, \chi_N$ by properly smoothing these discrete data and then evaluating each function on the same suitable time grid $t_1, \ldots, t_J$. We will give more details on this preprocessing of our Erlang data in Section 3.

After initializing the set of sum variables with the original variables $\chi(t_1), \ldots, \chi(t_J)$, the algorithm proceeds in the construction of the tree by removing at each iteration the two most correlated variables from the set of sum variables, and by replacing them with the associated first principal component. The second principal component, i.e., the difference variable, is stored along iterations. The algorithm is stopped when the set of sum variables is empty, thus returning the set of difference variables $\{\varphi_1, \ldots, \varphi_J\}$, each represented by a vector in $\mathbb{R}^J$. In our functional specification, the vectors of this set are interpreted as the evaluation of a set of functions $\{\varphi_1(t), \ldots, \varphi_J(t)\}$ at time instances $t_1, \ldots, t_J$. For further details on treelet decomposition see Lee et al. (2008).

The output of the algorithm allows for the choice of any subset of difference variables, which in turn will generate a proper linear subspace of $\mathbb{R}^J$. For instance, the first $L \leq J$ difference variables generate the space $W_1 \oplus \ldots, \oplus W_L$. In our application, the estimation of the set of functions $\{\psi_1(t), \ldots, \psi_K(t)\}$ is indeed accomplished by using the complete treelet basis $\{\varphi_1(t), \ldots, \varphi_J(t)\}$, according to a criterion detailed in the next subsections.

2.2 Bagging Voronoi Treelet Analysis (BVTA)

The model expressed in equation (3) relates the observed functional signal to a linear combination of a set of time-varying functions, each time-varying function contributing to the signal observed in a specific site of the lattice according to the value assumed in that site by a coupled surface. We can exploit the strategy described in the previous subsection and based on treelet decomposition for decoupling observed functional data into their constitutive parts. Indeed, we can directly apply the treelet basis decomposition to the N-dimensional sample of Erlang data $\{E_{\mathbf{x}}(t)\}_{\mathbf{x} \in S_0}$, and then select a K-dimensional subset of the complete treelet basis as an estimate for $\{\psi_1(t), \ldots, \psi_K(t)\}$. The coupled surfaces $\{d_1(\mathbf{x}), \ldots, d_K(\mathbf{x})\}$ will then be obtained by site-wise projection of the Erlang data on the estimates of $\{\psi_1(t), \ldots, \psi_K(t)\}$. In the rest of the paper, we will refer to this strategy as *Treelet Analysis* (TA).

The drawback of this approach is that it does not take into account spatial dependence, neither in the estimation of $\{\psi_1(t), \ldots, \psi_K(t)\}$, nor in that of $\{d_1(\mathbf{x}), \ldots, d_K(\mathbf{x})\}$. Due to the continuity in space of the phenomenon they capture, spatial dependence is intrinsic to our Erlang data. Hence, we develop a novel approach for the identification of the functions ψ and the coupled surfaces d by integrating the treelet decomposition with a proper treatment of spatial dependence. A comparison between this novel approach and TA for dimensional reduction of functional data indexed by a lattice, will be discussed in Section 5 in the light of the results of simulation studies.

We will take into account spatial dependence by following a Bagging Voronoi strategy along the lines depicted by Secchi et al. (2012). The rationale beyond this strategy

is simple, but effective: (*i*) replace the original data set with a reduced one, composed by local representatives of subsets of data belonging to neighborhoods covering the entire investigated area; (*ii*) analyze the local representatives; (*iii*) repeat the previous analysis many times for different reduced data sets associated to different randomly generated systems of neighborhoods, thus obtaining many different weak formulations of the analysis; (*iv*) finally, bag together the weak analyses to obtain a conclusive strong analysis.

At each iteration of the first part of the algorithm, called Bootstrap Step, we generate a partition of the considered region in random neighborhoods, which are used to compute local representatives. Each representative is a summary of the data belonging to the same element of the partition, and it is computed as a weighted mean with Gaussian isotropic weights (Secchi et al. 2012), even though other strategies are conceivable. The sample of functional local representatives exploits a specific structure of spatial dependence, and it is usually less noisy and less spatially dependent. By applying the TA strategy to the sample of local representatives, one obtains a coarse estimate of a reference basis. The coarse estimate of the coupled surfaces is then obtained by projecting each local representative on the estimated basis, and by assigning the corresponding scores to all sites of the lattice belonging to the element of the partition associated to the considered representative. After B replicates of this weak analysis, the intermediate output of the algorithm consists of:

- a collection of reference bases $\{\varphi_1^b(t),\dots,\varphi_J^b(t)\}_{b=1}^B$;
- a collection of sets of surfaces $\{d_1^b(\mathbf{x}),\dots,d_J^b(\mathbf{x})\}_{b=1}^B$.

The second part of the algorithm, the Aggregation Step detailed in the next subsection, *bags* together this intermediate output obtaining a final reference basis, estimate of the time-varying functions $\{\psi_1(t),\dots,\psi_K(t)\}$, and an estimate of the coupled surfaces $\{d_1(\mathbf{x}),\dots,d_K(\mathbf{x})\}$. Larger values of B imply a higher accuracy of the final estimate.

The proposed procedure is sketched in the pseudocode scheme in Figure 2. Note that one has to fix some parameters in advance: n, the dimension of the random partition and of the sample of functional local representatives; B, the number of bootstrap replicates; $d(\cdot,\cdot)$, the most proper metric to measure distances in the considered region. We named this procedure *Bagging Voronoi Treelet Analysis* (BVTA), since it is based on bagging, it uses Voronoi tessellations to compute random partitions of the considered area, and it uses treelets to perform dimensional reduction.

2.3 Aggregation step: 1-median alignment for bases matching

We will here give the details of the Aggregation Step in the BVTA algorithm sketched in Figure 2, whose aim is to bag together the B coarse results obtained in the Bootstrap Step. In the context of the present analysis, this means aggregating sets of treelet basis functions and of their coupled surfaces. The aggregation strategy illustrated in the following lines is a discrete variation of the Procrustes alignment procedures described for instance in Ramsay and Li (1998); James (2007); Kaziska and Srivastava (2007); Sangalli et al. (2009).

Algorithm. Bagging Voronoi Treelet Analysis (BVTA).

Initialize B, n. Choose a metric $d(\cdot,\cdot)$ for computing the random partition and a functional distance $\tilde{d}(\cdot,\cdot)$ for evaluating the similarity between bases.

`Bootstrap Step:`

`for` $b := 1$ `to` B `do`

`step 1.` randomly generate a set of nuclei $\Phi_n^b = \{\mathbf{Z}_1^b, \ldots, \mathbf{Z}_n^b\}$ among the sites in S_0: for $i = 1, \ldots, n$, $\mathbf{Z}_i^b \overset{i.i.d.}{\sim} \mathcal{U}(S_0)$, where $\mathcal{U}$ is the uniform distribution on the lattice. Obtain a random Voronoi tessellation of S_0, $\{V(\mathbf{Z}_i^b|\Phi_n^b)\}_{i=1}^n$, by assigning each site $\mathbf{x} \in S_0$ to the nearest nucleus $\mathbf{Z}_i^b$, according to the specified distance $d(\cdot,\cdot)$;

`step 2.` for $i = 1, \ldots, n$, compute the function g_i^b, acting as local representative, by summarizing information carried by the functional data associated to sites belonging to the i-th element of the tessellation $V_i^b := V(\mathbf{Z}_i^b|\Phi_n^b)$;

`step 3.` obtain an orthogonal basis $\{\varphi_1^b, \ldots, \varphi_J^b\}$ by applying TA to the set of local representatives $\{g_1^b, \ldots, g_n^b\}$; for all $\mathbf{x} \in V_i^b$, let $d_j^b(\mathbf{x})$ be the score of the orthogonal projection of g_i^b on φ_j^b, for $j = 1, \ldots, J$.

`end for`

`Aggregation Step:`

- perform `bases matching`, via the 1-median bases alignment algorithm: obtain the collection $\{\pi_1^b, \ldots, \pi_J^b\}_{b=1}^B$ of permutations defining a reordering of the elements of the bases $\{\varphi_1^b, \ldots, \varphi_J^b\}_{b=1}^B$ and minimizing the distance $\tilde{d}(\cdot,\cdot)$ to the final reference basis $\{\tilde{\varphi}_1, \ldots, \tilde{\varphi}_J\}$;
- identify the collection of surfaces

$$\{\tilde{d}_1^b(\mathbf{x}), \ldots, \tilde{d}_J^b(\mathbf{x})\}_{b=1}^B \equiv \{d_{\pi_1^b}^b(\mathbf{x}), \ldots, d_{\pi_J^b}^b(\mathbf{x})\}_{b=1}^B,$$

  coupled to the B reordered bases. Obtain the estimates $\{\hat{\psi}_1, \ldots, \hat{\psi}_K\}$ of $\{\psi_1, \ldots, \psi_K\}$ as the K elements of the final reference basis associated to the K largest variances computed on the collection of surfaces $\tilde{d}$;
- average, along $b = 1, \ldots, B$, the surfaces $\{\tilde{d}_1^b(\mathbf{x}), \ldots, \tilde{d}_J^b(\mathbf{x})\}_{b=1}^B$. Set the estimates $\{\hat{d}_1(\mathbf{x}), \ldots, \hat{d}_K(\mathbf{x})\}_{\mathbf{x}\in S_0}$ of the surfaces $\{d_1(\mathbf{x}), \ldots, d_K(\mathbf{x})\}_{\mathbf{x}\in S_0}$ to be the mean surfaces whose indexes correspond to those selecting the estimates $\{\hat{\psi}_1, \ldots, \hat{\psi}_K\}$ among the elements of the reference basis.

Figure 2: Pseudocode scheme of the BVTA algorithm.

If the functional basis used for dimensional reduction of the sample of local representatives were fixed along replicates of the algorithm (e.g. a wavelet basis, or a Fourier basis), one could compute the final reference basis as the output of a proper method for

estimating the centroid of a set of functional data. For instance, one could simply average the B bases component by component, or take their functional median. However, the BVTA algorithm is centered on data-driven bases, i.e., treelets. Hence a matching of the elements of the B different bases generated by the Bootstrap Step of the algorithm is in order before computing the final reference basis.

Different approaches to bases matching are possible. We develop a procedure for 1-median basis alignment, which jointly computes the reference basis from the B coarse bases, while also reordering their elements. This procedure is inspired by the joint clustering and alignment method described in Sangalli et al. (2010), where a Procrustes continuous alignment is integrated in a k-mean clustering strategy, to jointly meet the two tasks of assigning curves to a group, while simultaneously aligning them to the corresponding group prototype. In the context of bases matching, each object is a multivariate functional data (one of the coarse bases), and we look for the unique prototype (the reference basis) which best describes the set of functional objects, while also aligning their components, by permutations in the order of basis functions.

Consider the collection of all bases obtained in the Bootstrap Step, $\left\{\varphi_1^b,\ldots,\varphi_J^b\right\}_{b=1}^B$, and choose a proper measure $\tilde{d}(\cdot,\cdot)$ for the distance (or dissimilarity) between two bases, which in our application will be the $L^1([0,T];\mathbb{R}^J)$ distance.

The 1-median basis alignment is an iterative algorithm, which is initialized by randomly selecting a reference basis $\{\tilde{\varphi}_1^{[0]},\ldots,\tilde{\varphi}_J^{[0]}\}$, among the B coarse bases generated by the Bootstrap Step of the BVTA algorithm. The following two basic steps are then iterated until convergence (consider the l-th iteration, $l>0$):

1. **Alignment step.** For each of the B coarse bases, by permutation of their components, find the best matching to the reference basis $\{\tilde{\varphi}_1^{[l-1]},\ldots,\tilde{\varphi}_J^{[l-1]}\}$ according to the measure $\tilde{d}(\cdot,\cdot)$. For $b=1,\ldots,B$, let $\{k_1^{b,[l]},\ldots,k_J^{b,[l]}\}$ be the permutation of the order of the elements in the basis $\{\varphi_1^b,\ldots,\varphi_J^b\}$ minimizing the distance to the current reference basis;

2. **Estimation step.** Given the B reordered bases, let the new reference basis be that whose j-th element is

$$\tilde{\psi}_j^{[l]}=\underset{\varphi\in L^1(T)}{\operatorname{argmin}}\sum_{b=1}^{B}\tilde{d}(\psi^b_{k_j^{b,[l]}},\varphi),\ j=1,\ldots,J.$$

   Since $\tilde{d}(\cdot,\cdot)$ is the $L^1([0,T];\mathbb{R}^J)$ distance, the reference basis $\{\tilde{\varphi}_1^{[l]},\ldots,\tilde{\varphi}_J^{[l]}\}$ is the functional median of the B reordered bases.

The algorithm is stopped at iteration $\bar{l}$ after two subsequent iterations with no reordering of the basis elements, for all bases. The final reference basis is thus identified as $\{\tilde{\varphi}_1,\ldots,\tilde{\varphi}_J\}\equiv\{\tilde{\varphi}_1^{[\bar{l}]},\ldots,\tilde{\varphi}_J^{[\bar{l}]}\}$.

For $b=1,\ldots,B$, set $\{\pi_1^b,\ldots,\pi_J^b\}$ to be the final permutation $\{k_1^{b,[\bar{l}]},\ldots,k_J^{b,[\bar{l}]}\}$ and let $\{\tilde{d}_1^b(\mathbf{x}),\ldots,\tilde{d}_J^b(\mathbf{x})\}\equiv\{d^b_{\pi_1^b}(\mathbf{x}),\ldots,d^b_{\pi_J^b}(\mathbf{x})\}$, for $\mathbf{x}\in S_0$. For $j=1,\ldots,J$, we now compute the sample variance $\tilde{s}_j^2$ of the dataset $\{\tilde{d}_j^b(\mathbf{x})\}_{\mathbf{x}\in S_0,b=1,\ldots,B}$, whose size is $N\times B$.

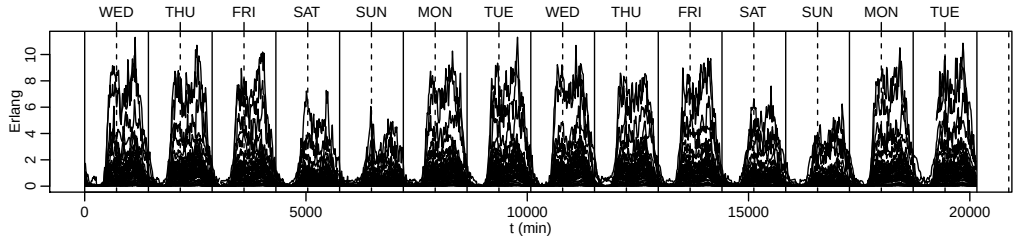

Figure 3: A random selection of 100 Erlang data, drawn at random among the sites of the lattice, as a function of time. The solid vertical lines are drawn at midnight of each day, and the dotted vertical lines at noon. The first day is Wednesday March 18, 2009.

For estimating the time-varying functions $\{\psi_1, \ldots, \psi_K\}$ we take the K elements of the basis $\{\tilde{\varphi}_1, \ldots, \tilde{\varphi}_J\}$ associated to the K largest variances among $\{\tilde{s}_1^2, \ldots, \tilde{s}_J^2\}$, and we call these elements $\{\hat{\psi}_1, \ldots, \hat{\psi}_K\}$.

Indeed, for each given $\mathbf{x} \in S_0$, the same indexes identifying $\{\hat{\psi}_1, \ldots, \hat{\psi}_K\}$ among the elements of the basis $\{\tilde{\varphi}_1, \ldots, \tilde{\varphi}_J\}$, also point to a collection of K data sets among the sequence of data sets $\{\tilde{d}_1^b(\mathbf{x})\}_{b=1}^B, \ldots, \{\tilde{d}_J^b(\mathbf{x})\}_{b=1}^B$; let $\hat{d}_k(\mathbf{x})$ be the mean of the k-th selected data set, for $k = 1, \ldots, K$. We take the surface $\{\hat{d}_k(\mathbf{x})\}_{\mathbf{x} \in S_0}$ to be an estimate of the surface $\{d_k(\mathbf{x})\}_{\mathbf{x} \in S_0}$ coupled with the time-varying function $\psi_k(t)$.

Note that only $K < J$ basis functions in the reference basis are finally selected. Indeed, in practical applications, one follows a Goldilocks approach that finds the "just right" value for K by inspecting the scree-plot generated by the sequence of variances $\tilde{s}_1^2, \ldots, \tilde{s}_J^2$. In our application, only treelets characterized by high values of $\tilde{s}^2$ have a pretty neat interpretation, and are therefore included in the analysis.

3 Data Analysis: Preprocessing

The Erlang data described in Section 1 are an instance of spatially dependent functional data, indexed by the sites of a spatial lattice. In each given site, the discrete sequence of Erlang values can be considered as a sampling of a continuous process in time, describing the average number of mobile phones using the network in that site (see equation 2). An example of the observed Erlang profiles along time is shown in Figure 3: 100 sites have been randomly selected in the lattice, and the Erlang measurements recorded in each selected site have been plotted as a function of time. It can be observed in the picture that, beside a periodic behavior due to night/day alternation in the average use of mobile phone, Erlang data present strongly localized features. Moreover, the average intensity of the Erlang profile can be very different from one site to another.

Indeed, in each site of the lattice we observe a discrete version of the Erlang continuous process, recorded approximately every quarter of an hour: due to discontinuities in the information provided by the network antennas, the Erlang measure is missing at some time instances, and hence the time grid of Erlang measurements is non-uniform. Moreover, some Erlang recordings are negative due to measurement errors, and should

be treated as missing values. We thus need to choose a proper basis expansion to reconstruct the functional form of the time-varying Erlang data, and to evaluate them on a common uniform grid of time values of dimension $J = 200$, before applying the methodology presented in Section 2.

For an extensive description of smoothing procedures for functional data we refer to Ramsay and Silverman (2005). In our application, we perform a site-wise smoothing of the Erlang data via a Fourier basis expansion, due to the evident seasonality in the Erlang profiles. We set the period of the Fourier basis equal to 1 week: hence, the reconstructed functional form of the Erlang profile for site $\mathbf{x} \in S_0$ is a function $E_{\mathbf{x}}(t)$ such that

$$E_{\mathbf{x}}(t) = \frac{c_0^{\mathbf{x}}}{2} + \sum_{h=1}^{H} \left[a_h^{\mathbf{x}} \cos(h\omega t) + b_h^{\mathbf{x}} \sin(h\omega t) \right], \tag{4}$$

where $t \in [0;T]$, $\omega = 2\pi/T$ and $T = 60 \cdot 24 \cdot 7$ is the period expressed in minutes. In the following, the periodic terms in the Fourier basis expansion oscillating at frequency $\omega, 2\omega, 3\omega, \ldots$ will be referred to as first, second, third, ... harmonic, respectively. The coefficients $c_0^{\mathbf{x}}$, $a_h^{\mathbf{x}}$ and $b_h^{\mathbf{x}}$, are estimated by means of least squares.

The basis dimension H should be carefully tuned: it has to be chosen large enough to ensure that the very localized features (sudden peaks, oscillations, ...) which characterize this kind of data (see Figure 3) are properly caught by the smoothing procedure. To select the basis dimension, we analyze the *power spectrum* associated to the site-wise smoothing of the Erlang data with a Fourier basis of large dimension $H = 200$. The power spectrum of the Fourier expansion of a signal represents the amplitude of the signal as a function of the frequency, and at the h-th frequency it is related to the amplitude of the h-th harmonic

$$P_{\mathbf{x}}(h) = \sqrt{(a_h^{\mathbf{x}})^2 + (b_h^{\mathbf{x}})^2}. \tag{5}$$

Hence, the more the h-th harmonic is relevant in the explanation of features occurring in the data, the more $P_{\mathbf{x}}(h)$ will be large. A local maximum in the power spectrum detects the frequency of an harmonic explaining relevant features in the data. When the power spectrum vanishes towards zero, there is no need to include higher frequency harmonics.

In each site we thus obtain a power spectrum from the site-wise smoothing of the Erlang measurements. We choose the most proper value of H by inspecting the shape of the average power spectrum over all the sites of the lattice, i.e. $\overline{P}(h) = \frac{1}{N} \sum_{\mathbf{x} \in S_0} P_{\mathbf{x}}(h)$. The average power spectrum of the Telecom Italia database is reported in Figure 4. A graphical inspection of the plot makes it clear that the frequencies significantly contributing to the Erlang time variation are the smaller ones (all less than 7), capturing differences among days or blocks of days (e.g., the working and weekend days variation), and the multiples of 7, capturing the recurring daily dynamics. Due to the huge dimension of the Telecom Italia database, we choose a basis of very high dimension, in order to be reasonably sure to catch all relevant localized features. The picture indicates that, for frequencies higher than 100ω, the average power spectrum is negligible: we thus set $H = 100$ for subsequent analyses.

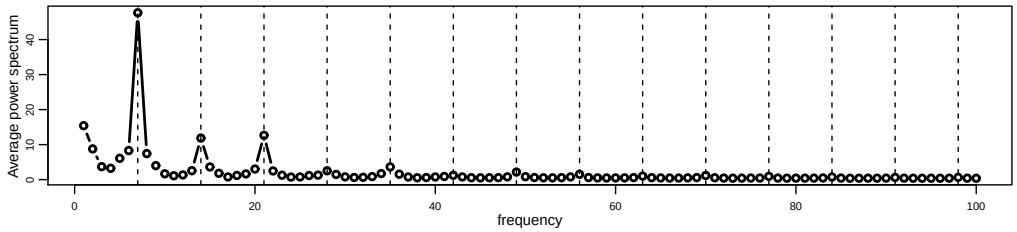

Figure 4: Average power spectrum $\overline{P}(h)$ obtained via site-wise smoothing of the Erlang measures with a Fourier basis of dimension $H = 200$. Only the values of $\overline{P}(h)$ for $h = 1, \ldots, 100$ are shown in the plot. Dotted vertical lines are drawn for multiples of 7.

4 Data Analysis: Results

The smooth functions obtained by the preprocessing of the Erlang data are then analyzed along the method presented in Section 2. In particular, the analysis has been performed for different values of the dimension n of the Voronoi tessellation, ranging from 50 to 2500 elements. For each value of n, $B = 50$ random Voronoi maps have been used in the Bootstrap Step of the BVTA algorithm. The metric $d(\cdot,\cdot)$ for generating the random Voronoi maps is the Euclidean distance on the plane, after having flattened the inspected area using the international WGS84 UTM 32N geographical system map. Local representatives are identified as weighted means with Gaussian isotropic weights.

Choosing the right value for n is a more delicate issue, since the optimal value of this parameter is strongly related to the spatial dependence occurring between data recorded in different sites. Smaller values of n are associated to bigger elements of the Voronoi tessellation, and thus provide a strong reduction of noise together with the aggregation of possibly non-homogeneous data. On the contrary, larger values of n are associated to smaller elements of the Voronoi tessellation: noise is not significantly reduced, but, on the other hand, aggregated data are expected to be more homogeneous. In Secchi et al. (2012), the choice for n is driven by the idea that a good value for n would be the one providing stable results of the performed analysis across bootstrap replicates. In that work a cluster analysis is performed, and thus the concept of bootstrap stability refers to cluster assignment of each site across replicates. To measure stability the authors introduced an entropy criterion, which averages over the entire area a pixel-wise measure of the uncertainty of the cluster assignment distribution along bootstrap replicates.

Analogously to Secchi et al. (2012) we define a measure of bootstrap stability coherent with the aims of the present analysis. An intermediate output of the BVTA algorithm consists of the collection of the B surface sets $\{\tilde{d}_1^b(\mathbf{x}), \ldots, \tilde{d}_J^b(\mathbf{x})\}_{b=1}^B$, each surface being coupled with an element of the reference basis $\{\tilde{\varphi}_1(t), \ldots, \tilde{\varphi}_J(t)\}$. Indeed the final estimates of the time-varying functions $\{\psi_1(t), \ldots, \psi_K(t)\}$ and of their coupled surfaces $\{d_1(\mathbf{x}), \ldots, d_K(\mathbf{x})\}$ are exclusively based on this intermediate output, that we therefore require to be stable with respect to the choice of n. To measure stability, for each site $\mathbf{x} \in S_0$ and $j = 1, \ldots, J$, we compute the bootstrap variance of the

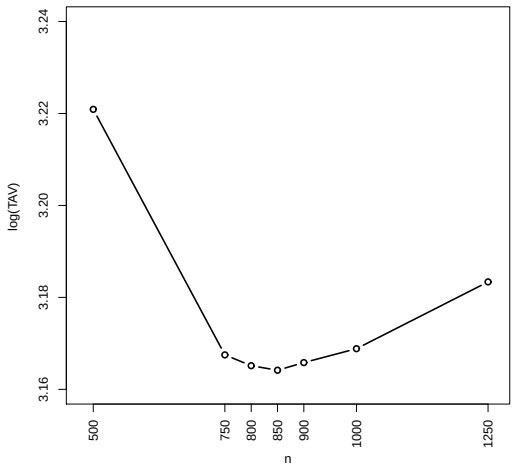

Figure 5: Total average variance (log scale) for the BVTA of the Erlang data as a function of n.

data set $\{\tilde{d}_j^b(\mathbf{x})\}_{b=1}^B$. We then average over $\mathbf{x} \in S_0$, and sum over $j = 1, \ldots, J$. We call this quantity *total average variance* (TAV): its minimization is the criterion driving the choice for n. A small value of TAV implies that for each site and for each element of the reference basis we attain stable scores across bootstrap replicates.

In Figure 5 the logarithm of TAV is reported as a function of n. A minimum is observed for $n = 850$, that is thus the dimension of the Voronoi tessellation used to run the BVTA algorithm for the analysis of the Erlang data. This dimension of the Voronoi tessellation is associated to an average area of the Voronoi elements equal to 0.77 km^2, that corresponds to the area of a circle of diameter nearly equal to 1 km. This indicates that spatial dependence is relevant up to this distance, and thus reveals 1 km to be the practical spatial range of our data.

The Goldilocks approach described at the end of Subsection 2.3 selects $K = 23$ as the "just right" dimension for the reference basis output of the BVTA algorithm. Quite surprisingly, the time-varying functions thus selected, and their coupled surfaces, are also easily interpretable. We here discuss just four of them, that we deem to be particularly interesting for illustrating the peculiar properties of the analysis conducted by means of the BVTA algorithm:

- the population average density function $\hat{\psi}_1$;
- the working/non-working time function $\hat{\psi}_2$;
- the rush-hour function $\hat{\psi}_4$;
- the Milan design week function $\hat{\psi}_9$.

The first two functions correspond to static activities, the third one to mobility-related activities, and the fourth one to a spot event concentrated in space and time.

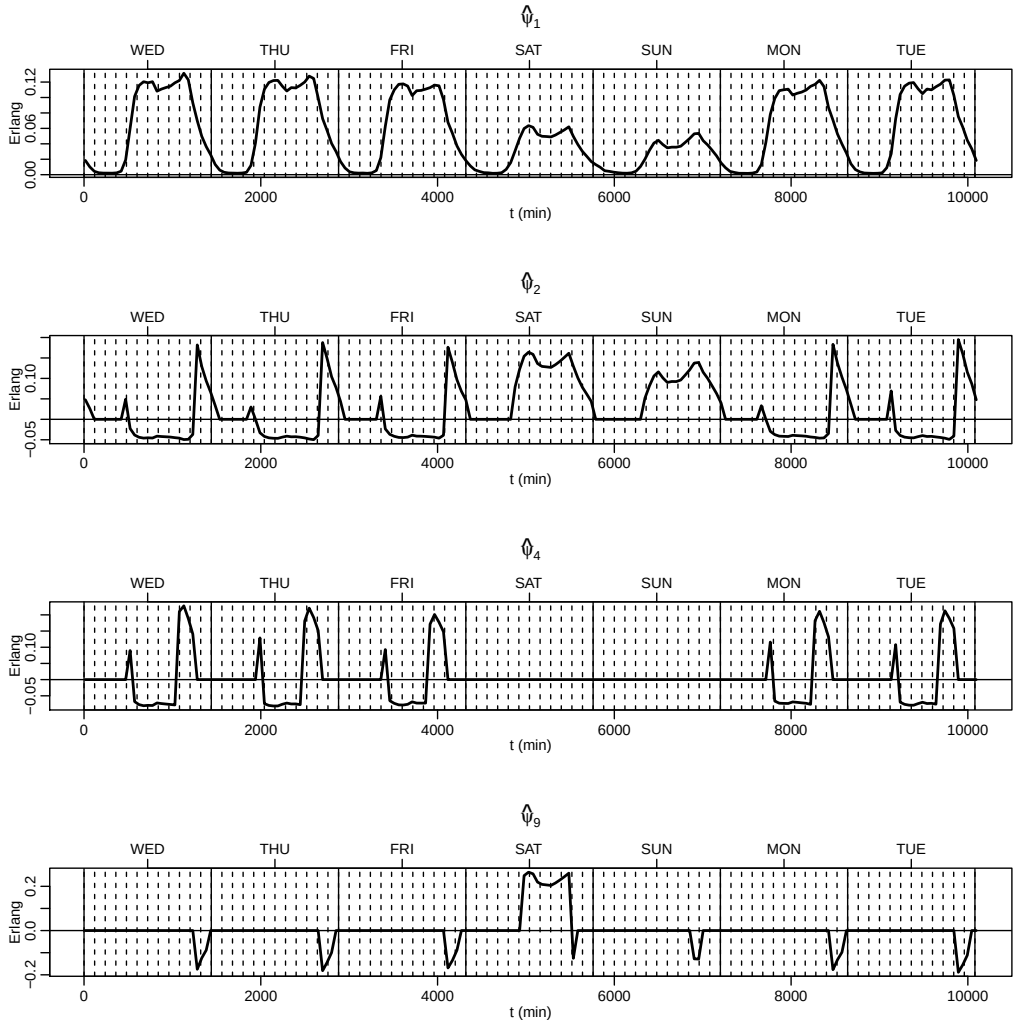

Figure 6: Some selected elements of the reference basis (from top to bottom: $\hat{\psi}_1$, $\hat{\psi}_2$, $\hat{\psi}_4$, and $\hat{\psi}_9$) output of the BVTA algorithm analyzing the Erlang data.

Figure 6 reports the temporal pattern of the basis elements $\{\hat{\psi}_1(t), \hat{\psi}_2(t), \hat{\psi}_4(t), \hat{\psi}_9(t)\}$ over a week-period starting from Wednesday 00:00 and ending with Tuesday 24:00. Full vertical lines separate different days while dotted lines are reported every two hours to help the reader. Their coupled surfaces $\{\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \hat{d}_4(\mathbf{x}), \hat{d}_9(\mathbf{x})\}$ are represented in Figure 7. A value close to 0 in a particular site of the map means that the corresponding reference basis element does not significantly contribute to the Erlang signal measured in that site. On the contrary, a positive/negative large value on the map means that the corresponding reference basis element significantly contributes to the Erlang signal in that site, with sign coherent to the score sign. The 0-level contour lines are traced in bold. Figure 8 zooms on the city center of the previous maps. In the remaining part of the Section, we give more details on the interpretations of $\{\hat{\psi}_1(t), \hat{\psi}_2(t), \hat{\psi}_4(t), \hat{\psi}_9(t)\}$

and of their coupled surfaces, with the aim of illustrating the type of information about the city dynamics that can be drawn from our analysis.

Population Average Density Function $\hat{\psi}_1$. The population average density function is the most important in terms of magnitude, in the sense that it is the one presenting the largest contribution to the Erlang signals of many highly active sites. It can be indeed detected even through much simpler analyses, that do not take into account spatial dependence, or even by simply looking at a random sample of curves (e.g., Figure 3). As it is evident by inspection of Figure 6 (top panel), this reference basis element is always switched on with positive sign with significant values between 7:00 am and 2:00 am in working days and between 8:00 am and 2:00 am in weekend days. It describes daily and weekly periodicity. In particular it points out a larger activity during day-time with respect to night-time, a bi-modal behavior of the daily signal, and confirms Milan to be an attractor during the day-time of working days. This is clear from the lower level observed during the weekend. In the top-left panel of Figure 7 the estimated coupled surface is reported. This map catches the urbanization of the area, clearly pointing out day-time low-density population areas and day-time highly-populated areas. We thus relate this reference basis element to the population average density.

Working/non-working Time Function $\hat{\psi}_2$. Looking at Figure 6 (second panel from the top), we notice that this function contrasts working-time (i.e., from 8:30 am to 8:00 pm of working days) against non-working time (i.e., from 7:00 am to 8:30 am and from 8:00 pm to 2:00 am of working days, and day-time of week-end days). Positive values on the relevant map (top-right panel in Figure 7) indicate high activity during non-working-time and a reduced activity during working-time, and viceversa for negative values. The map clearly spots the historical center connected with a northeast offshoot toward the Central Railway Station, areas mostly devoted to tertiary activities, and where the resident population density is very low. Then, a donut-shaped area around the city center, mostly covering residential or leisure areas, emerges with high positive values. Moving further from the city center, the values of $\hat{d}_2$ tend to vanish except for some non-working hours spots corresponding to satellite towns right outside the city of Milan. Moreover, one can observe a working hours spot in the north direction corresponding to the Bicocca neighborhood, that is a renewed area in the outskirts of Milan mostly devoted to tertiary and to university-related activities. This basis element presents the city center as an attractor during the working hours and the outskirts and the satellite towns as attractors during the non-working hours. This can possibly be explained by the daily mobility of working people from their residence to their working place and backward.

Rush-hour Function $\hat{\psi}_4$. Positive scores with respect to this function point out areas where an high activity is present between 8:00 am to 10:00 am and 5:00 pm to 9:00 pm, which correspond to the morning and evening rush hours (see the second panel from

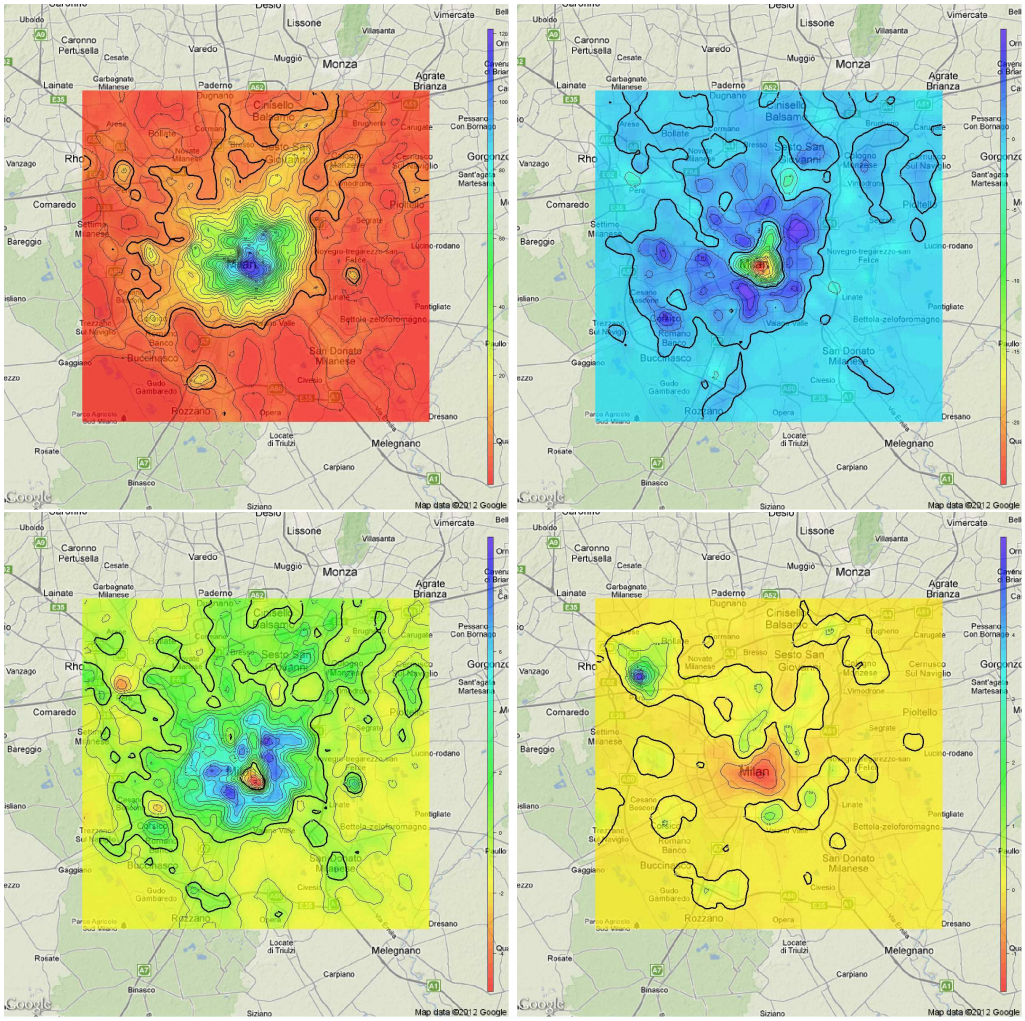

Figure 7: From top to bottom, and then from left to right, maps of the estimated surfaces $\{\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \hat{d}_4(\mathbf{x}), \hat{d}_9(\mathbf{x})\}$ coupled to the reference basis elements reported in Figure 6. The 0-level contour lines are reported in bold.

the bottom in Figure 6). Inspection of the coupled surface $\hat{d}_4$ in the bottom-left panel of Figure 7, shows that areas particularly active during rush hours are concentrated around the third ring-road within the city (Circonvallazione Esterna), at the Central Railway Station, along arteries connecting the city with the satellite towns, along some segments of the highway ring-road, and in Linate Airport (the eastern spot on the map). It is also interesting to note the hole in the very city center corresponding to the congestion charge area, which is restricted only to local traffic during weekdays.

Milan Design Week Function $\hat{\psi}_9$. This function contrasts the Saturday activity carried on between 10:00 am and 8:00 pm and everyday dinner time (see the bottom panel

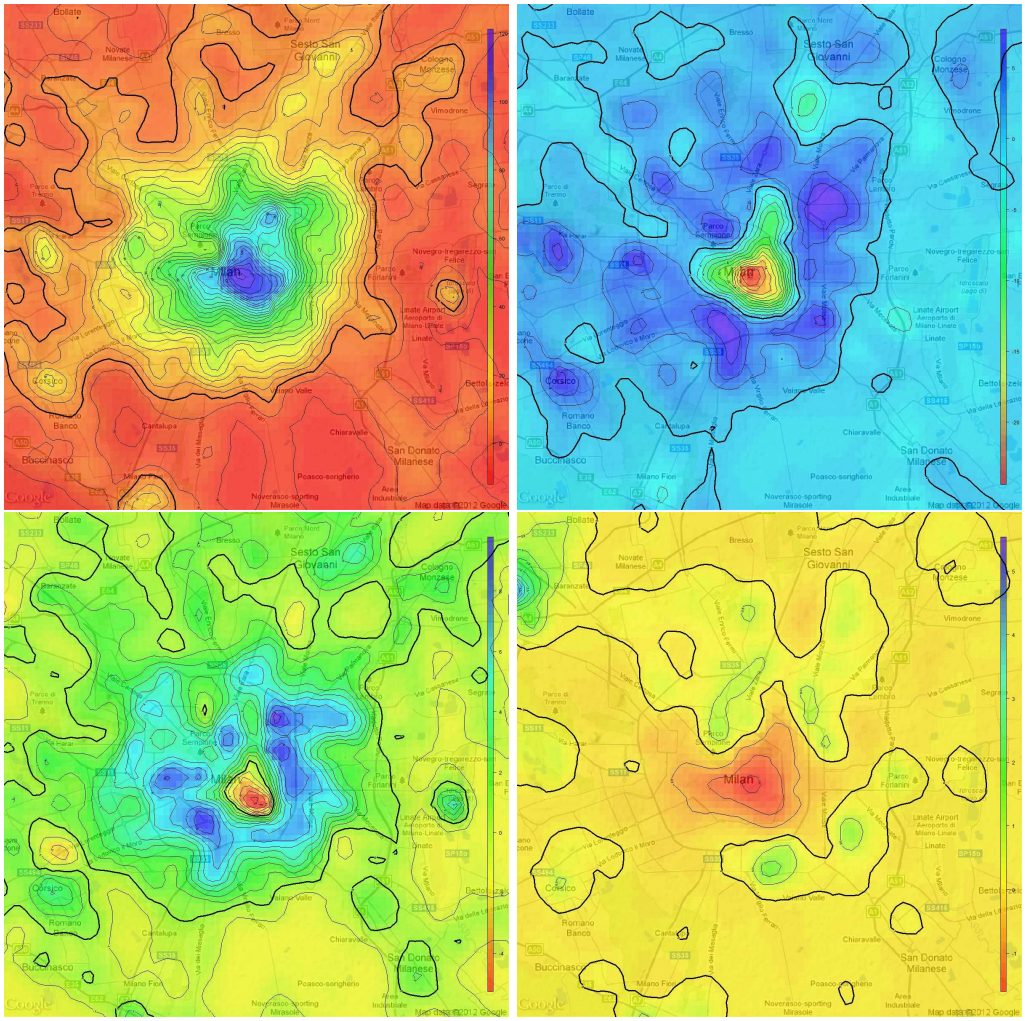

Figure 8: Zooms on the city center of the maps already shown in Figure 7.

in Figure 6), and seems strongly related to the activity, during non-working time, connected to the Milan Design Week. This event has been held between the 18^{th} and the 23^{rd} March, 2009, at the Fiera Milano Exhibition Complex North-West of the city; the activities related to this occasional but highly attractive event clearly affect the Erlang measurements in the time period covered by our data. Indeed, in the bottom-right panel of Figure 7, the Fiera Milano Exhibition Complex can be easily located by the positive peak in $\hat{d}_9$, while a corresponding negative peak is observed in the city center. This is possibly explained in the light of the interpretation given to $\hat{d}_2$, by a flow of people spending Saturday at the Exhibition site and dinner and after-dinner in the city center.

5 Simulation Study

In this Section we describe a simulation study conducted to address some open issues related to the proposed method. The simulations are aimed at supporting the following claims:

(i) minimization of TAV is a good criterion for selecting the optimal value n_{opt} of the Voronoi tessellation dimension n;

(ii) the best estimate $\{\hat{\psi}_1(t),\ldots,\hat{\psi}_K(t)\}$ of the time-varying functions $\{\psi_1(t),\ldots,\psi_K(t)\}$ is obtained for $n = n_{opt}$;

(iii) the estimate $\{\hat{\psi}_1(t),\ldots,\hat{\psi}_K(t)\}$ of the set of time-varying functions $\{\psi_1(t),\ldots,\psi_K(t)\}$ obtained via BVTA algorithm with $n = n_{opt}$ is better than the one obtained via standard TA, which does not take into account spatial dependence.

We analyze a set of functional data $\{Y_{\mathbf{x}}(t)\}_{\mathbf{x}\in S_0}$, indexed by the sites of a square bidimensional lattice S_0 of 50×50 sites. The functional signal observed in each site is generated by the following model

$$Y_{\mathbf{x}}(t) = \sum_{k=1}^{3} d_k(\mathbf{x})\psi_k(t), \quad \mathbf{x}\in S_0,\ t\in[0,T], \tag{6}$$

where $T = 5$, $\{\psi_1(t),\psi_2(t),\psi_3(t)\}$ is a set of three time-varying components, $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ is the set of their coupled surfaces. Model (6) aims at generating smooth data analogous to the data generated by the preprocessing phase of our analysis, as described in Section 3.

The time-varying components are selected to have patterns similar to those in the estimated final reference basis for Erlang data (see Figure 6), in order to support our claims (i)-(iii) in a simulated scenario as close as possible to the case study at hand. Moreover, each time-varying component is orthogonal to the others, to ensure coherence between the set $\{\psi_1(t),\psi_2(t),\psi_3(t)\}$ and the estimates $\{\hat{\psi}_1(t),\hat{\psi}_2(t),\hat{\psi}_3(t)\}$ provided by the TA and the BVTA strategies, which are both based on orthogonal basis decompositions. Finally, we aim at functional patterns generating data $Y_{\mathbf{x}}(t)$ complex enough to be untractable with standard parametric models. The time-varying components $\psi_1(t),\psi_2(t),\psi_3(t)$ are shown in the top panels of Figures 12, 13 and 14, respectively. The function $\psi_1(t)$ (top panel in Figure 12) is represented by the sinusoidal function

$$\psi_1(t) = 2 + \frac{1}{2}\cdot\sin(2\pi t),\ \text{for } t\in[0,T], \tag{7}$$

and it can be interpreted as an average profile across the lattice. The second and third components, $\psi_2(t)$ and $\psi_3(t)$ (top panel in Figure 13 and 14, respectively), are continuous and periodically contrast some selected time intervals, wider in the former function.

Their analytical expressions are the following

$$\psi_2(t) = \begin{cases} 10-50\cdot(t-i), & \text{if } 0.2+i \le t < 0.22+i, \text{ for } i=0,\dots,4, \\ -1, & \text{if } 0.22+i \le t < 0.33+i, \\ & \text{or } 0.67+i \le t < 0.78+i, \text{ for } i=0,\dots,4, \\ 50\cdot(t-i)-17.5, & \text{if } 0.33+i \le t < 0.37+i, \text{ for } i=0,\dots,4, \\ 1, & \text{if } 0.37+i \le t < 0.63+i, \text{ for } i=0,\dots,4, \\ 32.5-50\cdot(t-i), & \text{if } 0.63+i \le t < 0.67+i, \text{ for } i=0,\dots,4, \\ 50\cdot(t-i)-40, & \text{if } 0.78+i \le t < 0.8+i, \text{ for } i=0,\dots,4, \\ 0, & \text{otherwise;} \end{cases} \tag{8}$$

$$\psi_3(t) = \begin{cases} 50\cdot(t-i), & \text{if } i \le t < 0.02+i, \text{ for } i=0,\dots,4, \\ 1, & \text{if } 0.02+i \le t < 0.08+i, \\ & \text{or } 0.92+i \le t < 0.98+i, \text{ for } i=0,\dots,4, \\ 5-50\cdot(t-i), & \text{if } 0.08+i \le t < 0.12+i, \text{ for } i=0,\dots,4, \\ -1, & \text{if } 0.12+i \le t < 0.18+i, \\ & \text{or } 0.82+i \le t < 0.88+i, \text{ for } i=0,\dots,4, \\ 50\cdot(t-i)-10, & \text{if } 0.18+i \le t < 0.2+i, \text{ for } i=0,\dots,4, \\ 40-50\cdot(t-i), & \text{if } 0.8+i \le t < 0.82+i, \text{ for } i=0,\dots,4, \\ 50\cdot(t-i)-45, & \text{if } 0.88+i \le t < 0.92+i, \text{ for } i=0,\dots,4, \\ 50-50\cdot(t-i), & \text{if } 0.98+i \le t < 1+i, \text{ for } i=0,\dots,4, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The coupled surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ are generated according to a Hidden Markov Random Field model (see Kunsch et al. (1995) for details). More precisely, in each site $\mathbf{x} \in S_0$ we generate, independently from one another, three latent labels $\Lambda_1(\mathbf{x}), \Lambda_2(\mathbf{x})$ and $\Lambda_3(\mathbf{x})$ from three different Ising Markov Random fields $\Lambda_1, \Lambda_2, \Lambda_3 : S_0 \to \{-1, 1\}$ with parameters $\beta_1, \beta_2, \beta_3$, respectively. The parameter β of an Ising Markov Random field controls the strength of spatial dependence: higher values of β imply a stronger spatial dependence, and hence generate a field characterized by large macro-areas of the lattice assigned to the same label. We fix $\beta_1 = \beta_2 = 2$ and $\beta_3 = \frac{1}{2}$: in this way, fields generated by Λ_1 and Λ_2 will show a smoother pattern than the patchy one we expect from Λ_3. This choice reflects a situation in which a spiky behavior in time, that described by $\psi_3(t)$, is associated to very localized areas in space. Three realizations of Λ_1, Λ_2 and Λ_3 are shown in Figure 9 (top panels: left, center, and right respectively). Given these three realizations, the surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ are site-wise and independently generated according to the following distributions

$$\begin{aligned} d_1(\mathbf{x}) | (\Lambda_1(\mathbf{x}) = l) &\sim \chi^2(p_l), \\ d_2(\mathbf{x}) | (\Lambda_2(\mathbf{x}) = l) &\sim N(\mu_l, \sigma_2^2), \\ d_3(\mathbf{x}) | (\Lambda_3(\mathbf{x}) = l) &\sim N(\nu_l, \sigma_3^2), \end{aligned}$$

where $p_{-1} = 3$ and $p_1 = 8$, $\mu_{-1} = -1$ and $\mu_1 = 1$, $\nu_{-1} = 0.5$ and $\nu_1 = 1$, $\sigma_2 = 1$ and $\sigma_3 = 0.25$. These choices are related to the desired behavior for functional data associated to

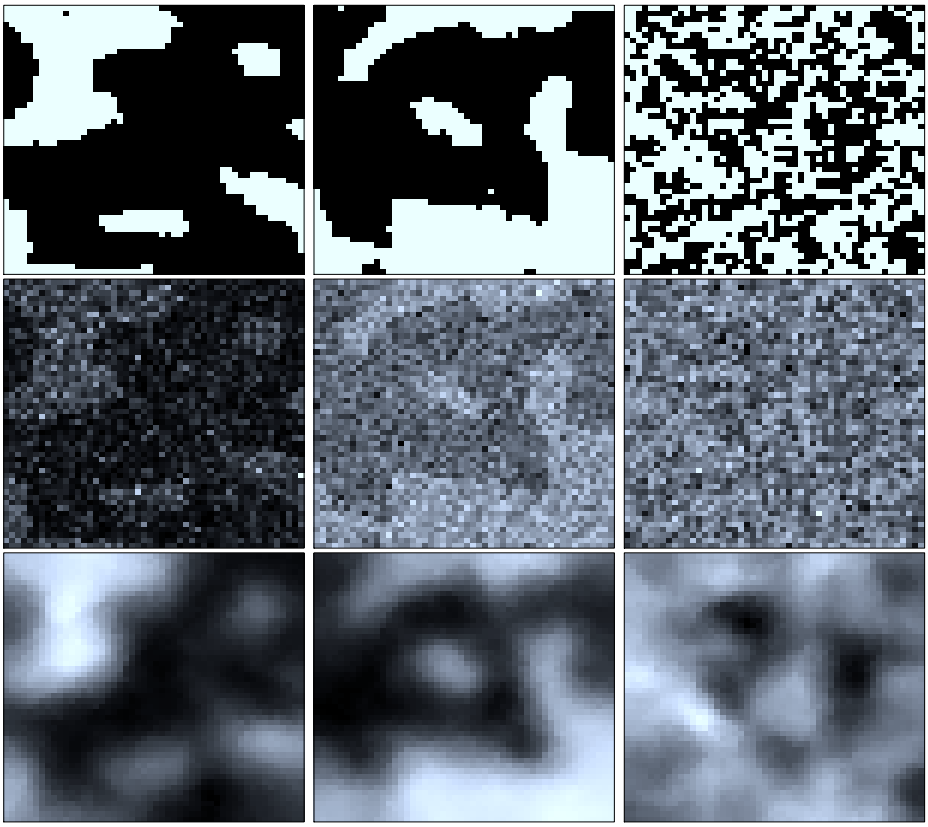

Figure 9: In the top panels, three realizations of the Ising Markov Random fields used to generate $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ for the first simulation study: from left to right, Λ_1, Λ_2 and Λ_3. For each site $\mathbf{x} \in S_0$, $\Lambda_1(\mathbf{x}), \Lambda_2(\mathbf{x})$, and $\Lambda_3(\mathbf{x})$ respectively give the label to determine the distribution of $d_1(\mathbf{x})$, $d_2(\mathbf{x})$, and $d_3(\mathbf{x})$, shown in the central panels (from left to right). In the bottom panels, from left to right, the three estimates $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x})$, and $\hat{d}_3(\mathbf{x})$ obtained via the BVTA algorithm with $n = n_{opt}$, are reported.

different labels: in each site $\mathbf{x} \in S_0$, $d_1(\mathbf{x})$ determines the higher or lower intensity of the average time pattern $\psi_1(t)$, $d_2(\mathbf{x})$ determines the sign and the size of the contrast $\psi_2(t)$, while $d_3(\mathbf{x})$ determines the higher or lower intensity of the peaks in $\psi_3(t)$. The central panels of Figure 9 represent three realizations of the surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}_{\mathbf{x} \in S_0}$ corresponding to the three realizations of the fields Λ_1, Λ_2 and Λ_3 shown in the top panels. Finally, we obtain the synthetic functional data via the generating model (6), and we evaluate the so obtained functional data on $p = 200$ equally spaced time instances in the interval $[0, T]$. An example of the generated synthetic data is shown in Figure 10, where 50 synthetic profiles, randomly selected among the sites of the lattice, have been plotted as a function of time.

This set of synthetic data is analyzed with two different strategies: the TA strategy described in Subsection 2.1, and the BVTA algorithm described in Section 2.2. We fix

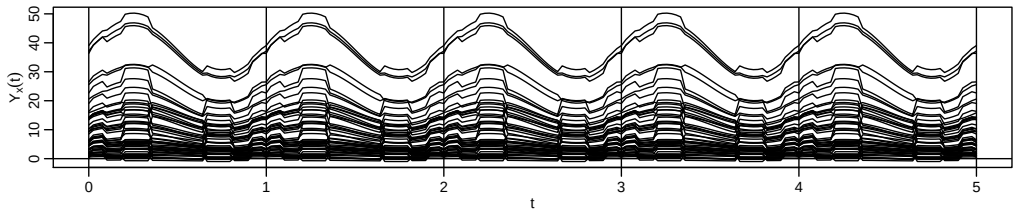

Figure 10: A random selection of 50 synthetic data obtained according to the generating model (6) in the first simulation study. The functions have been drawn at random among the sites of the lattice, and they are shown as a function of time.

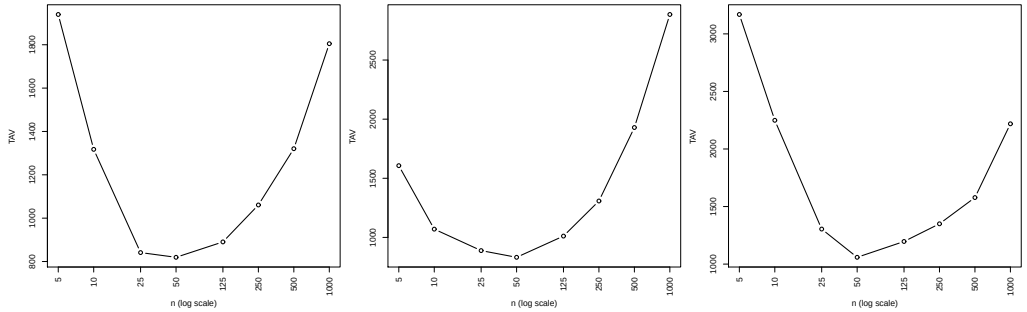

Figure 11: Total average variance (TAV) for different values of n, for three replicates of the BVTA algorithm on three different datasets generated according to model (6) in the first simulation study.

the parameters controlling the BVTA algorithm as follows: $B = 50$, $d(\cdot,\cdot)$ is the Euclidean metric in $\mathbb{R}^2$ and $n \in \{5, 10, 25, 50, 125, 250, 500, 1000\}$. The n representatives are identified as weighted means with Gaussian isotropic weights.

In Figure 11 we plot the TAV for different values of n, and for three replicates of the BVTA analysis conducted on three different synthetic data sets generated according to model (6): the presence of a value of n minimizing TAV can be appreciated in all panels of the picture, thus proving this to be a good criterion to choose n, and addressing our first claim. Moreover, the value of n_{opt} is equal to 50 in all cases, supporting the robustness of our approach.

Figure 12 shows the estimates of $\psi_1(t)$ when the surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ are those represented in the central panels of Figure 9; these estimates are generated by the TA algorithm, and by the BVTA algorithm initialized with different values of n. The estimate $\hat{\psi}_1(t)$ most similar to $\psi_1(t)$ is that obtained by the BVTA with $n = n_{opt}$. In the second panel from the top of Figure 12, the estimate of $\psi_1(t)$ obtained by the TA is pictured; it clearly mixes up some features characterizing $\psi_2(t)$ and $\psi_3(t)$. Furthermore, these confounding coming from different time-varying components is also observed in the third and fifth panels from the top of Figure 12: these are the estimates obtained by the BVTA with a non optimal n. Notice the similarity between the estimate generated by TA and that produced by BVTA for a value of n much larger than n_{opt}. Indeed, at

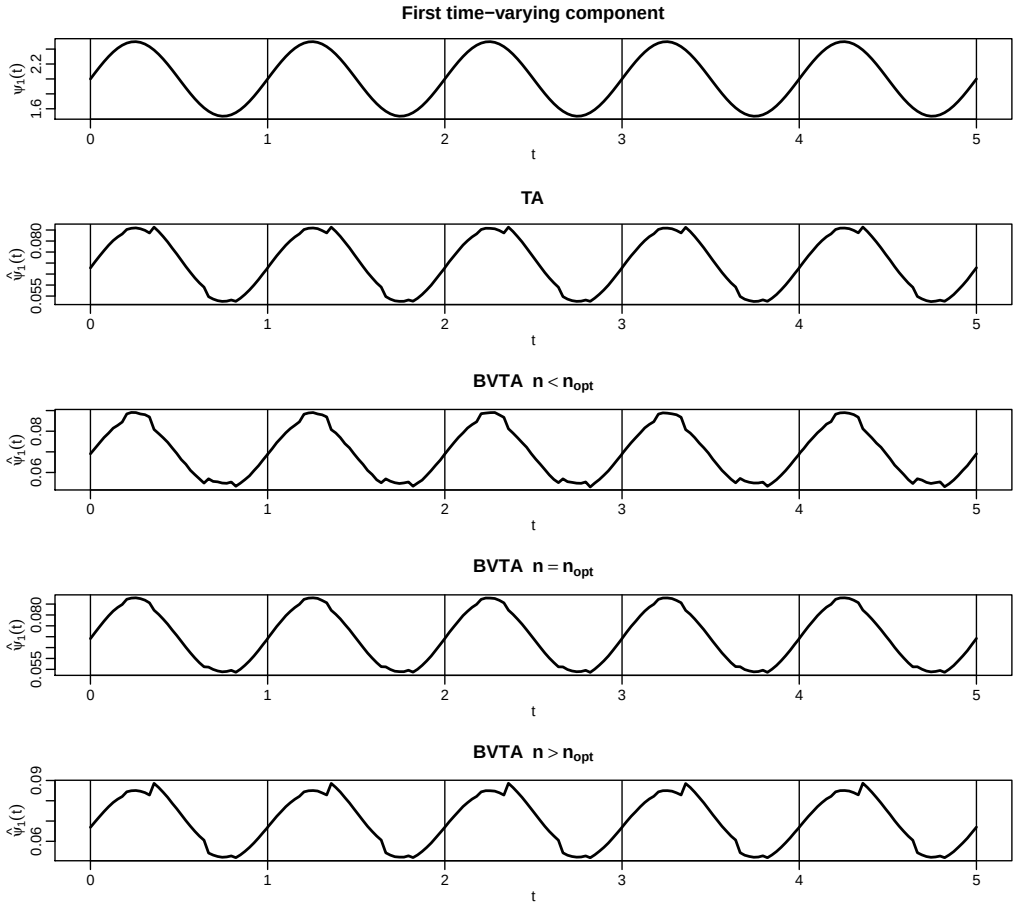

Figure 12: In the top panel, the first time-varying component $\psi_1(t)$ used to generate the synthetic data for the first simulation study according to model (6). From the second panel from top towards the bottom, estimates of the first time-varying component $\psi_1(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$.

least for the first time-varying component, our second and third claim are supported.

Figure 13 reports the estimates $\hat{\psi}_2(t)$ of the second time-varying component $\psi_2(t)$ for the same scenarios considered in Figure 12. It is evident that a pattern quite similar to the true function is obtained by both the TA and the BVTA algorithm with $n \geq n_{opt}$. Indeed, the estimate given by BVTA algorithm with $n = n_{opt}$ is the closest to the true function. BVTA fails to reconstruct the true pattern for values of n smaller than the optimal, mixing up features of both ψ_2 and ψ_3: this can be due to the fact that larger Voronoi elements imply a stronger bias for local representatives, thus making the estimate of $\hat{\psi}_2(t)$ less reliable.

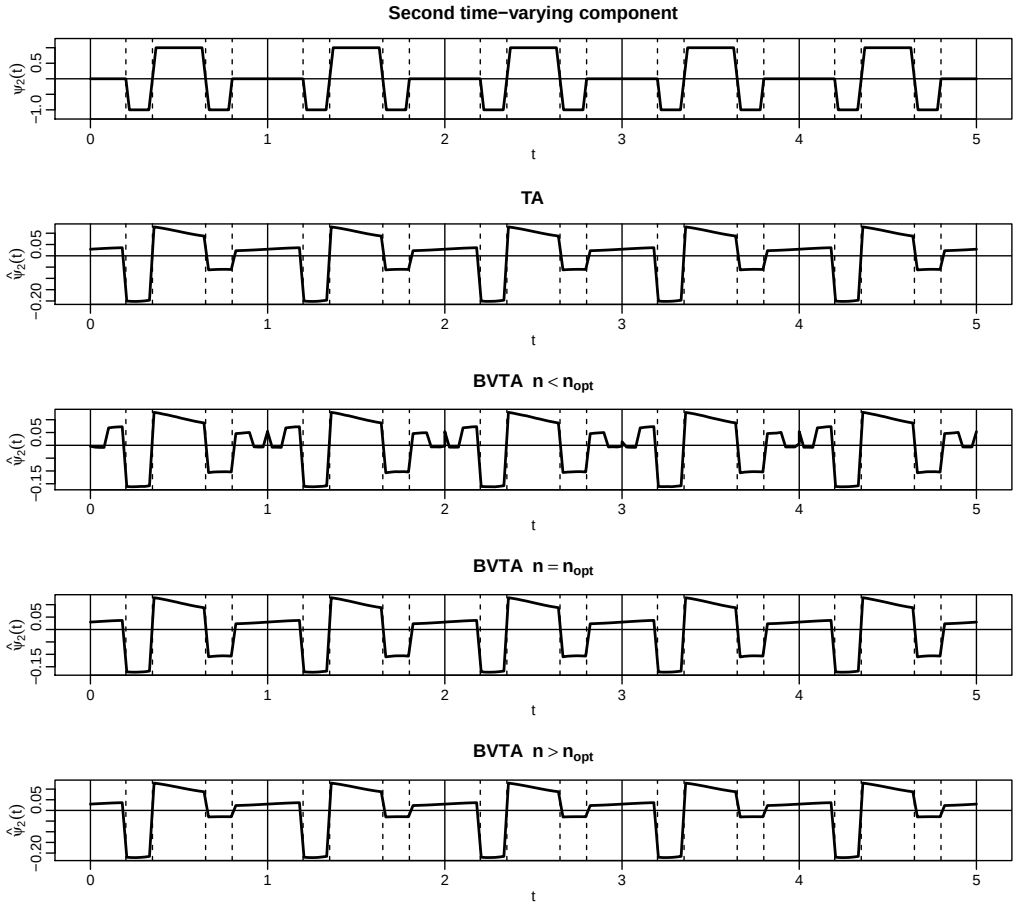

Figure 13: In the top panel, the second time-varying component $\psi_2(t)$ used to generate the synthetic data for the first simulation study according to model (6). From the second panel from top towards the bottom, estimates of the second time-varying component $\psi_2(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$. The dotted vertical lines are drawn at the discontinuities of the true component $\psi_2(t)$, reported in the top panel.

Finally, Figure 14 shows the estimates $\hat{\psi}_3(t)$ obtained in the same scenarios considered in the previous two figures. All estimates seem quite far from the true function $\psi_3(t)$, except for the one generated by BVTA with $n = n_{opt}$, which is very close to the true function $\psi_3(t)$.

Thus, the best estimates of the set of time-varying functions $\{\psi_1(t), \psi_2(t), \psi_3(t)\}$ are obtained by the BVTA strategy with $n = n_{opt}$, which properly exploits spatial dependence to improve the estimation of the model components, supporting claims (i)-(iii).

The three estimates $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x})$ and $\hat{d}_3(\mathbf{x})$ generated by the BVTA with $n = n_{opt}$ are reported in the bottom panels of Figure 9 (from left to right): they refer to the data set

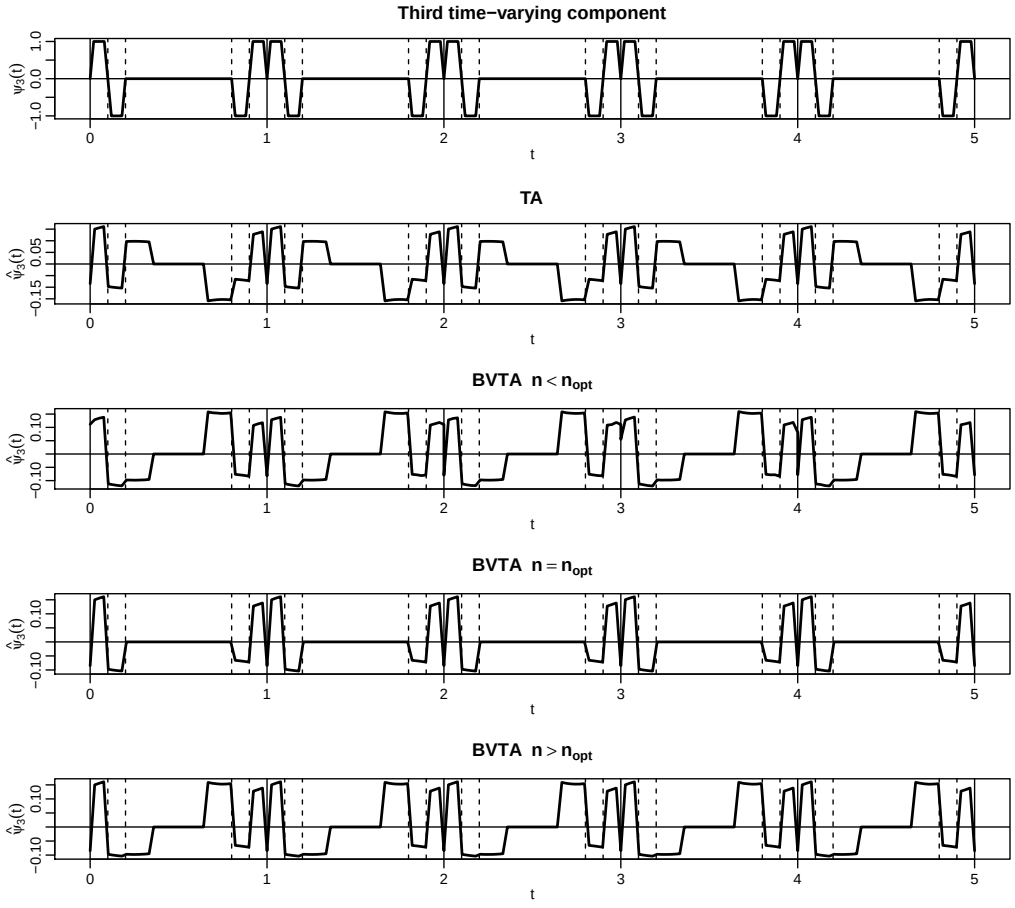

Figure 14: In the top panel, the third time-varying component $\psi_3(t)$ used to generate the synthetic data for the first simulation study according to model (6). From the second panel from top towards the bottom, estimates of the third time-varying component $\psi_3(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$. The dotted vertical lines are drawn at the discontinuities of the true component $\psi_3(t)$, reported in the top panel.

shown in the central panels. The latent fields of labels are clearly detectable from the corresponding estimates of the surfaces: the segmentation of the area into subregions homogeneous with respect to the strength and sign of the signal represented by the corresponding time-varying function ψ appears to be quite effective.

We finally describe the result of a second simulation study, conducted to test the robustness of the BVTA strategy when the set of time-varying components is no longer a set of orthogonal functions. In this second study, we analyze a set of functional data generated using the same model described in equation (6), where the new set of coupled surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}_{\mathbf{x}\in S_0}$ is obtained according to the same HMRF previously

described (a realization of the surfaces is shown in Figure 15, central panels). The only modification to the scenario introduced for the first study affects the new set of time-varying components $\{\psi_1(t), \psi_2(t), \psi_3(t)\}$: while the first two components have been only slightly modified, the third component is instead a function that we do not expect to be completely captured via our BVTA algorithm, since it is not orthogonal to the space spanned by the first two. It is included in model (6) to check for stability of the estimates of the first two components, when a third non–orthogonal component is present.

The time-varying components for the second simulation study are shown in the top panels of Figures 17, 18 and 19, respectively. The new function $\psi_1(t)$ (top panel in Figure 17) is the same function represented in equation (7). The new second component $\psi_2(t)$ (top panel in Figure 18) periodically contrasts some selected time intervals, and is quite similar to the old one in equation (8)

$$: \psi_2(t) = \begin{cases} -1, & \text{if } 0.25+i < t < 0.33+i \text{ or } 0.7+i \leq t < 0.9+i, \text{ for } i = 0,\ldots,4, \\ 1, & \text{if } 0.33+i \leq t < 0.7+i, \text{ for } i = 0,\ldots,4, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Finally, the third component $\psi_3(t)$ (top panel in Figure 19) is not a contrast, and shows a periodic and positive spiky behavior, quite different from the previous one in equation (9)

$$: \psi_3(t) = \begin{cases} 1, & \text{if } 0.33+i \leq t < 0.4+i, \text{ or } 0.65+i \leq t \leq 0.8+i, \text{ for } i = 0,\ldots,4 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Note that, while $\psi_1(t)$ and $\psi_2(t)$ are (almost) orthogonal, $\psi_3(t)$ is not orthogonal to $\psi_1(t)$.

The functional synthetic data obtained in the second scenario are again analyzed using both the TA and the BVTA strategies. We fix the parameters controlling the BVTA algorithm as previously described. In Figure 16 we plot the TAV for different values of n, and for three replicates of the BVTA analysis conducted on three different synthetic data sets generated according to model (6) in the second simulation scenario: the presence of a value of n minimizing TAV can be again appreciated, with the value of n_{opt} being equal to 25 in all cases.

Figure 17 shows the estimates of $\psi_1(t)$ when the surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ are those represented in the central panels of Figure 15; these estimates are generated by the TA algorithm, and by the BVTA algorithm initialized with different values of n. The estimate $\hat{\psi}_1(t)$ most similar to $\psi_1(t)$ is that obtained by the BVTA with $n = n_{opt}$. Both the estimate of $\psi_1(t)$ obtained by the TA, and the ones obtained by BVTA with a non optimal n, clearly mix up some features characterizing $\psi_2(t)$ and $\psi_3(t)$. Indeed, at least for the first time-varying component, the BVTA estimates are robust with respect to the absence of orthogonality.

Figure 18 reports the estimates $\hat{\psi}_2(t)$ of the second time-varying component $\psi_2(t)$ of the second simulation study, for the same scenarios considered in Figure 17. It

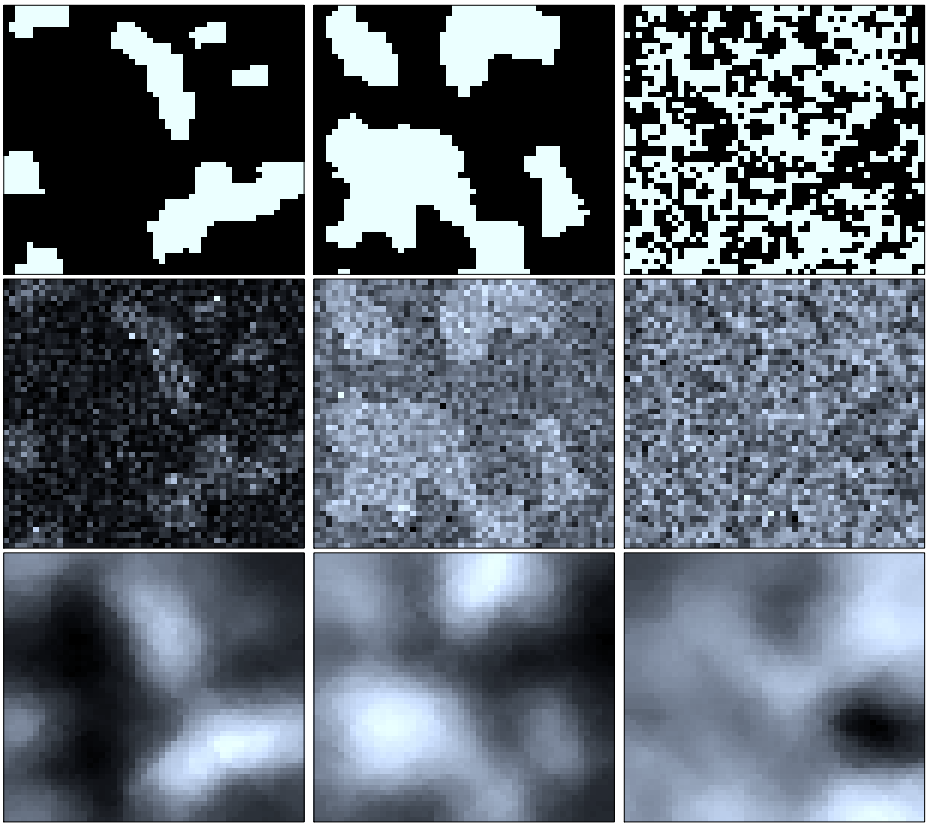

Figure 15: Three realizations of the Ising Markov Random Fields used to generate $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x})\}$ for the second simulation study, and their estimates via BVTA: from left to right, Λ_1, Λ_2, and Λ_3 are shown in the top panels, d_1, d_2, and d_3 are shown in the central panels, and the three estimates $\hat{d}_1, \hat{d}_2$, and $\hat{d}_3$, obtained via the BVTA algorithm with $n = n_{opt}$, are reported in the bottom panels.

is evident that a pattern quite similar to the true function is obtained by the BVTA algorithm with $n \geq n_{opt}$, and also by the TA algorithm.

The third time-varying component $\psi_3(t)$ is not orthogonal to the space spanned by $\psi_1(t)$ and $\psi_2(t)$. Hence, we do not expect to estimate it efficiently neither with the TA strategy, nor with the BVTA strategy. Figure 19 shows the estimates of $\psi_3(t)$ obtained in the same scenarios considered in the previous two figures. Indeed all estimates seem quite far from the true function $\psi_3(t)$, even though the best estimate appears to be the one generated by BVTA with $n = n_{opt}$. This supports our claim that the BVTA strategy, provided an optimal value of n is selected, is quite robust to violations of orthogonality. Note that both TA and BVTA aim at estimating $\psi_3(t)$ with a function almost orthogonal to $\hat{\psi}_1(t)$ and $\hat{\psi}_2(t)$, and are thus not able to capture the positive spike pattern of $\psi_3(t)$, far from being orthogonal to $\psi_1(t)$.

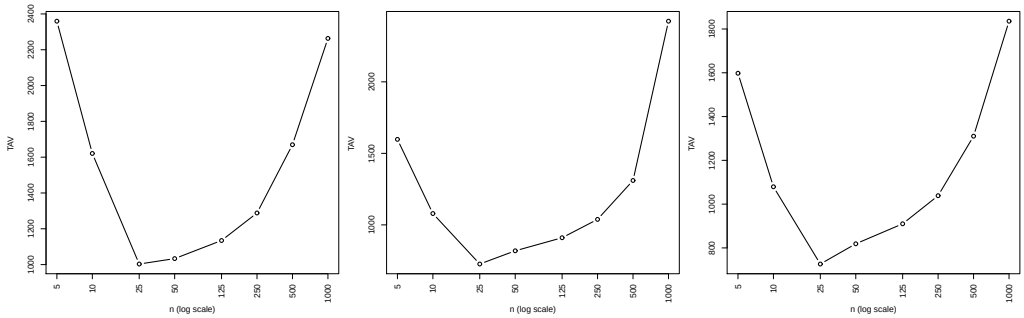

Figure 16: Total average variance (TAV) for different values of n, for three replicates of the BVTA algorithm on three different datasets generated according to model (6) in the second simulation study.

The three estimates $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x})$ and $\hat{d}_3(\mathbf{x})$ generated by the BVTA with $n = n_{opt}$ are reported in the bottom panels of Figure 15 (from left to right): they refer to the data set shown in the central panels. Similarly to the results of the first simulation study, the latent fields of labels are clearly detectable from the corresponding estimates of the surfaces.

We shall remark that we do not expect a perfect matching between the function $\psi_i(t)$ of the second simulation study, and its treelet estimate $\hat{\psi}_i(t)$, for $i = 1,2,3$, due to the fact that treelet decomposition looks for an orthonormal set of basis functions, while the selected set of time-varying functions (shown in the top panels of Figures 17, 18 and 19, respectively) is clearly non-orthogonal. Nevertheless, we decided to test the BVTA strategy also in this synthetic scenario, in order to check its robustness in a situation possibly arising in applications. A possible future research direction goes towards the use of dimensional reduction techniques which remove the orthogonality assumption.

6 Conclusions

In this work, a real case study concerning the dimensional reduction of spatially dependent functional data, describing the average number of mobile phones simultaneously using the Telecom Italia mobile network for calling at a given time, has been described. This work is a first innovative attempt to gather information about population density and mobility in Milan from mobile network data belonging to the Telecom Italia database. We believe that the applicability and impact of the proposed analysis are broad, both to the purposes of the Green Move project and, more generally, for future urban planning and development.

The methodology developed to perform dimensional reduction of spatially-dependent functional data is an innovative integration between a treelet analysis (see (Lee et al. 2008)) and the Bagging Voronoi strategy for the exploration of spatial dependence (see (Secchi et al. 2012)), and it is thus named *Bagging Voronoi Treelet Analysis*. We exploit the potentialities of both techniques, improving on the results that can be obtained using

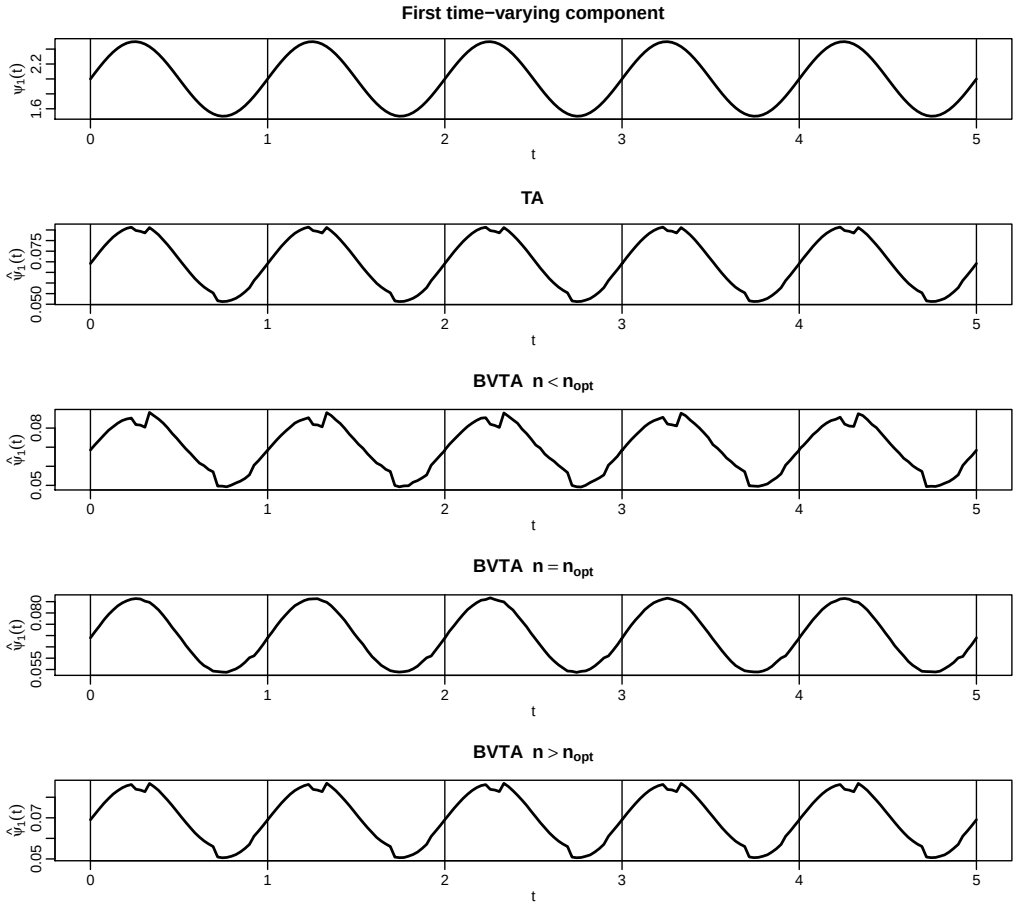

Figure 17: In the top panel, the first time-varying component $\psi_1(t)$ used to generate the synthetic data for the second simulation study according to model (6). From the second panel from top towards the bottom, estimates of the first time-varying component $\psi_1(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$.

the original methods alone. The method is proven useful in the applicative context of interest, and in a simulated scenario close to the real one.

Further research developments concern both the treatment of spatial dependence, and the dimensional reduction technique. The former can be improved by considering, either in the random generation of the set of nuclei for the tessellation, or in the distance $d(\cdot,\cdot)$ used to compute the Voronoi elements, relevant information concerning the area under investigation. For instance, the diffusion tensor describing the traffic mobility could be used to define a city-adapted measure of the distances, thus obtaining Voronoi elements capable of "following" the flow of people. For what concerns the latter aspect, the dimensional reduction strategy can be modified removing the assumption of

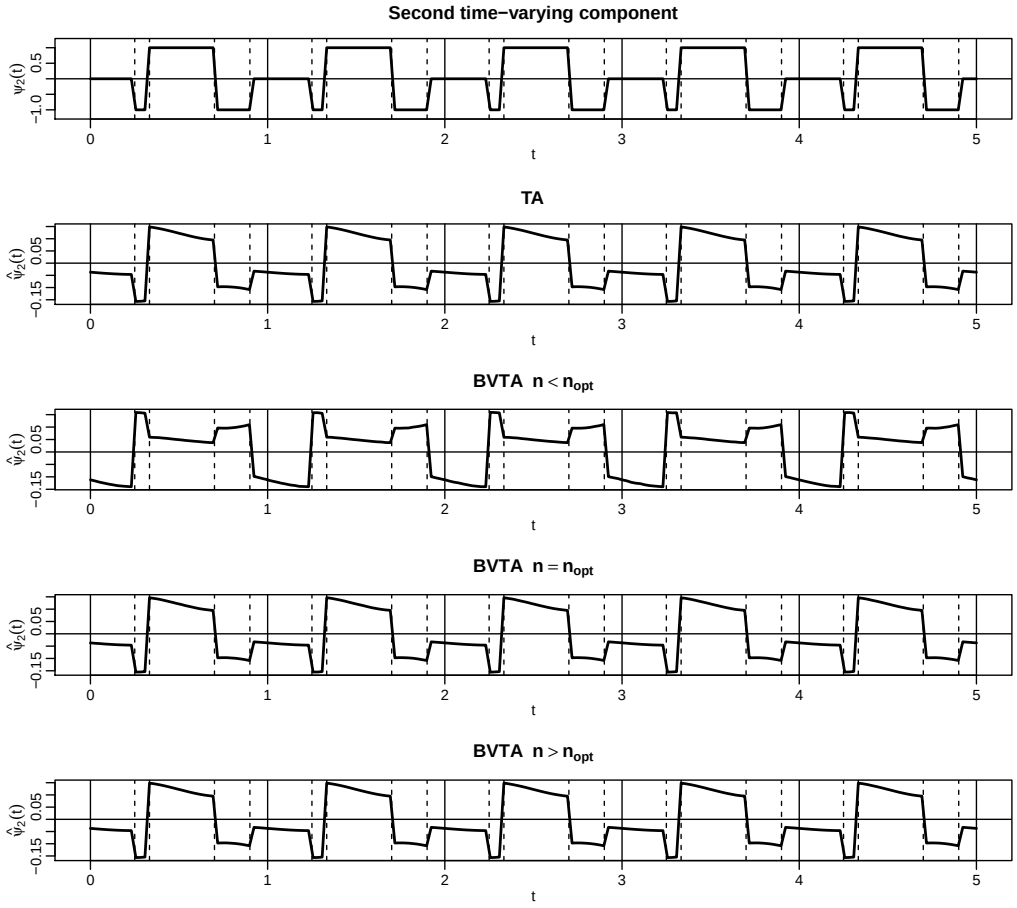

Figure 18: In the top panel, the second time-varying component $\psi_2(t)$ used to generate the synthetic data for the simulation study according to model (6). From the second panel from top towards the bottom, estimates of the second time-varying component $\psi_2(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$. The dotted vertical lines are drawn at the discontinuities of the true component $\psi_2(t)$, reported in the top panel.

orthogonality among the elements of the reference basis. This assumption is indeed non physical, and seems restrictive in the real application at hand.

References

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., and Volinsky, C. (2011), "A Tale of One City: Using Cellular Network Data for Urban Planning," *IEEE Pervasive Computing*, 10, 18–26.

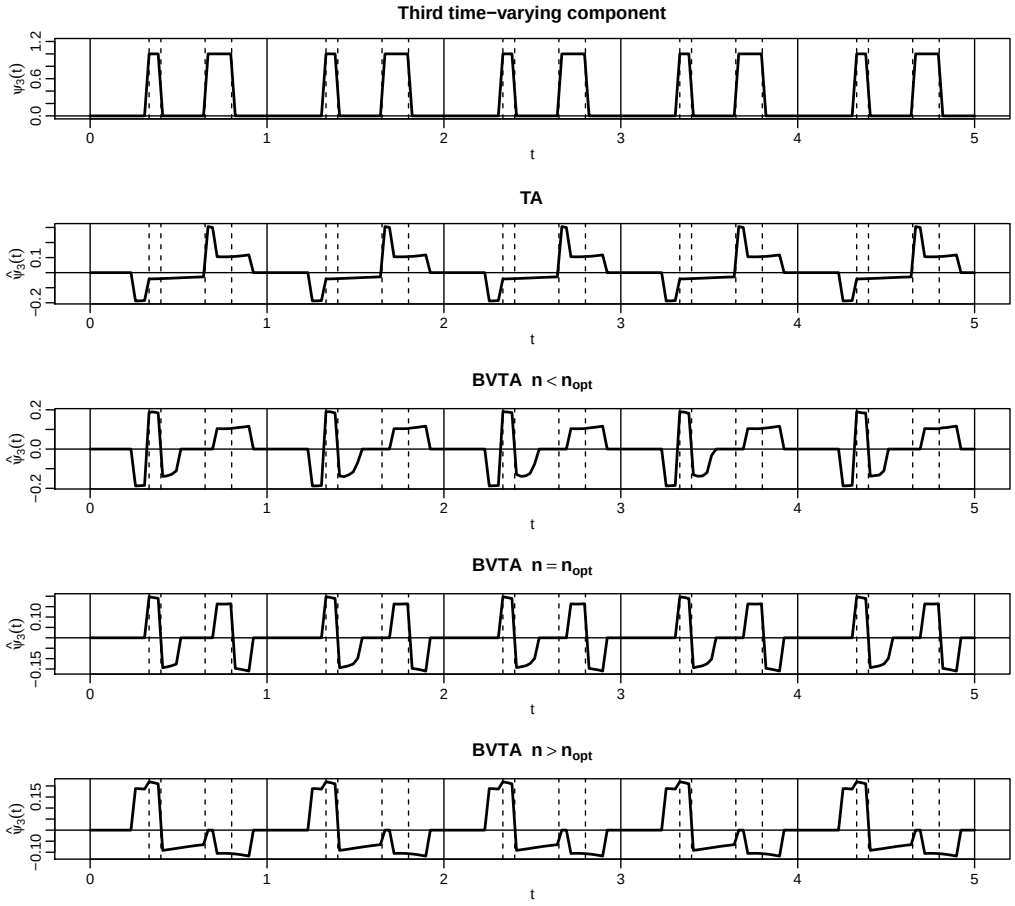

Figure 19: In the top panel, the third time-varying component $\psi_3(t)$ used to generate the synthetic data for the simulation study according to model (6). From the second panel from top towards the bottom, estimates of the third time-varying component $\psi_3(t)$ obtained with the TA strategy, and with the BVTA strategy for different values of n: $n = 5 < n_{opt}$, $n = n_{opt}$, and $n = 1000 > n_{opt}$, respectively. The solid vertical lines are drawn for $t = 0,1,2,3,4,5$. The dotted vertical lines are drawn at the discontinuities of the true component $\psi_3(t)$, reported in the top panel.

Calabrese, F., Lorenzo, G. D., Liu, L., and Ratti, C. (2011), "Estimating Origin-Destination Flows Using Mobile Phone Location Data," *IEEE Pervasive Computing*, 10, 36–44.

James, G. M. (2007), "Curve alignment by moments," *The Annals of Applied Statistics*, 1, 480–501.

Kaziska, D. and Srivastava, A. (2007), "Gait-Based Human Recognition by Classification of Cyclostationary Processes on Nonlinear Shape Manifolds," *Journal of the American Statistical Association*, 102, 1114–1128.

Kunsch, H., Geman, S., and Kehagias, A. (1995), "Hidden Markov Random Fields," *The Annals of Applied Probability*, 5, 577–602.

Lee, A. B., Nadler, B., and Wasserman, L. (2008), "Treelets – An adaptive multi-scale basis for sparse unordered data," *The Annals of Applied Statistics*, 2, 435–471.

OECD (2006a), *OECD Territorial Reviews: Competitive Cities in the Global Economy*, OECD Publishing.

— (2006b), *OECD Territorial Reviews: Milan, Italy*, OECD Publishing.

Ramsay, J. O. and Li, X. (1998), "Curve registration," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60, 351–363.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer.

Sangalli, L. M., Secchi, P., Vantini, S., and A.Veneziani (2009), "A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery," *J. Amer. Statist. Assoc.*, 104, 37–48.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010), "K-mean alignment for curve clustering," *Computational Statistics and Data Analysis*, 54, 1219–1233.

Secchi, P., Vantini, S., and Vitelli, V. (2012), "Bagging Voronoi classifiers for clustering spatial functional data," *International Journal of Applied Earth Observation and Geoinformation*, DOI: http://dx.doi.org/10.1016/j.jag.2012.03.006, in press.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

43/2012 Secchi, P.; Vantini, S.; Vitelli, V.
A Case Study on Spatially Dependent Functional Data: the Analysis of Mobile Network Data for the Metropolitan Area of Milan

42/2012 Lassila, T.; Manzoni, A.; Quarteroni, A.; Rozza, G.
Generalized reduced basis methods and n width estimates for the approximation of the solution manifold of parametric PDEs

41/2012 Chen, P.; Quarteroni, A.; Rozza, G.
Comparison between reduced basis and stochastic collocation methods for elliptic problems

40/2012 Lombardi, M.; Parolini, N.; Quarteroni, A.
Radial basis functions for inter-grid interpolation and mesh motion in FSI problems

39/2012 Ieva, F.; Paganoni, A.M.; Ziller, S.
Operational risk management: a statistical perspective

38/2012 Antonietti, P.F.; Bigoni, N.; Verani, M.
Mimetic finite difference approximation of quasilinear elliptic problems

37/2012 Nobile, F.; Pozzoli, M.; Vergara, C.
Exact and inexact partitioned algorithms for fluid-structure interaction problems with finite elasticity in haemodynamics

36/2012 Canuto, C.; Verani, M.
On the Numerical Analysis of Adaptive Spectral/hp Methods for Elliptic Problems

35/2012 Pigoli, D.; Aston, J.A.D.; Dryden, I.L.; Secchi, P.
Distances and Inference for Covariance Functions

34/2012 Menafoglio, A.; Dalla Rosa, M.; Secchi, P.
A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space