MOX–Report No. 26/2011

Bagging Voronoi classifiers for clustering spatial functional data

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

Bagging Voronoi classifiers for clustering spatial functional data

Piercesare Secchi, Simone Vantini, Valeria Vitelli

July 1, 2011

MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
piercesare.secchi@polimi.it, simone.vantini@polimi.it,
valeria.vitelli@mail.polimi.it

Keywords: Spatial statistics, bagging, functional data analysis, irradiance data, geostatistics.

AMS Subject Classification: 62H11, 62H30, 62G09, 62P12.

Abstract

We propose a bagging strategy based on random Voronoi tessellations for the exploration of high dimensional spatial data, suitable for different purposes (e.g., classification, regression,...). In particular, we consider the problem of clustering functional data indexed by the sites of a spatial finite lattice. The analysis is based on local representatives from neighboring data, i.e., belonging to the same element of a tessellation: the proposed algorithm accounts for spatial dependence by repeatedly clustering functional local representatives with respect to a random system of neighborhoods. Due to the resampling of tessellations, classification result is a cluster assignment frequency map, which can be used to define an a-posteriori criterion to choose the most suitable grouping structure. Thanks to spatial dependence, local representatives are expected to be less noisy and less correlated than original data, providing better performances. Moreover, this reduction in the dimension of the dataset permits the handling of high dimensional sets of data otherwise intractable without an explicit model for spatial dependence. The performance of the proposed approach is tested on simulated data. An application to environmental data contained in *Surface Solar Energy* database is also illustrated.

1 Introduction

Many methods for the analysis of high–dimensional data have been proposed in the recent years: some of them fall into the framework of functional data

analysis (see Ferraty and Vieu, 2006 [6]; Ramsay and Silverman, 2005 [13]), and only few among these consider spatially dependent functional data (see the interesting review by Delicado *et. al*, 2010 [5]). We aim at considering the problem of unsupervised classification of spatially dependent functional data in a non parametric framework, where each curve is indexed by the sites of a spatial finite lattice $S_0 \subset S$, where S defines the region of interest for the analysis. In particular, our motivating application consist in analyzing a global dataset concerning irradiance data along time: we examine the annual patterns of the maximum amount of energy needed to backup a photovoltaic system in 47880 worldwide non-polar districts (a non uniform lattice S_0 covering the whole earth surface) along the years 1983-2005; the analysis of these patterns is closely related to the sizing of power emergency generators needed in case of consecutive no-sun days (see Richter *et. al*, 2009 [14]). The problem thus consists in associating to each site $\mathbf{x} \in S_0$ a label $l \in \{1, \dots, L\}$, such that sites omogeneous with respect to the distribution generating the functional data are labelled the same: the aim of the analysis is the reconstruction of the latent field of labels. In our motivating application, we aim at identifying different homogeneous macro-areas, interpretable in terms of the observed phenomenon and not captured by customary unsupervised classification procedures that do not take into proper account the spatial dependence among data.

A tentative approach to this problem consist in the use of standard functional clustering procedures, such as functional k-mean (see Cuesta *et al.*, 2007 [4]; Tarpey and Kinadeter, 2003 [15]), that do not properly account for spatial dependence; indeed, while assigning a site to a cluster, information carried out by the neighboring sites is not considered. We thus expect these standard non-spatial approaches to provide good results only when the true groups are associated to very different distributions; in less trivial situations, where the distributions associated to different labels may be very similar, exploiting information carried by neighbors can lead to more accurate results.

We propose a new bagging algorithm for unsupervised classification that exploits spatial dependence by repeatedly generating random connectivity maps and by clustering, at each replication, local representatives of neighboring functional data. The performances of our algorithm are tested in various situations, and compared with standard clustering techniques. The new algorithm is completely non-parametric, since no explicit assumption is made neither on the distribution generating the latent field of labels, nor on the conditional distribution generating functional data. A great advantage of this approach is its flexibility in the exploitation of further information on the considered region, which is not paid off by an excessive increment of the computational cost. The proposed spatial clustering procedure, given the number K of clusters, produces and analizes bootstrap samples in three basic steps: generation of a spatial Voronoi tessellation, identification of a representative for each of the n elements of the tessellation, p-dimensional reduction and clustering of the representatives. For each site of the lattice S_0, the final output is the frequency distribution of cluster assignment to each of the K clusters; the frequency distribution can be summarized in a classification map by means of its mode via a majority vote on the cluster assignment. Moreover, an a-posteriori criterion based on spatial entropy

is proposed to select the optimal number K of clusters. The fact that our data is functional is not irrelevant to the computational cost of standard procedures for the analysis of lattice data; hence another motivation for our method, which implicitly performs a reduction both in the dimension of the sample (by clustering a small number n of representatives) and in the infinite dimension of functional data (through the p–dimensional reduction of the representatives). Moreover, most algorithms for image classification based on Hidden Markov Random Field models, which perform classification via a maximum a posteriori – MAP – criterion (such as *simulated annealing*, or *iterated conditional modes*, see Besag, 1986 [2]; Geman and Geman, 1984 [7] for details on these procedures) heavily depend on hypothesis on the distribution of the observed signal, typically assumed to be Gaussian.

The paper is structured as follows. In Section 2, the bagging Voronoi classifiers algorithm for clustering spatially dependent functional data is introduced and described. In Section 3, some technical issues concerning the algorithm are detailed; in particular, the adopted strategies to capture spatial dependence and to reduce the dimension of the data set, and the chosen approaches to dimensional reduction of functional data and clustering. In Section 4 the properties of the algorithm are explored through a simulation study. Finally, in Section 5 our motivating application is fully described, and results of application of the bagging Voronoi classifiers algorithm to irradiance data are shown. All simulations and analysis of real datasets are performed in R ([12]).

2 Bagging Voronoi classifiers for clustering spatially dependent functional data

Suppose a latent field of labels $\Lambda_0 : S_0 \to \{1, \ldots, L\}$ is associated to each site of the lattice S_0, i.e. $\Lambda_0(\mathbf{x})$ is the true unknown label associated to the site $\mathbf{x} \in S_0$, where $S_0 \subset S$ and S is a measurable subset of $\mathbb{R}^d$; the label sums up some characteristics of the considered area which are interesting for the scopes of the analysis. Moreover, suppose a functional data is observed in each site $\mathbf{x} \in S_0$: given Λ_0, the functional data are independently generated in each site $\mathbf{x} \in S_0$ from a distribution indexed by $\Lambda_0(\mathbf{x})$. Aim of the classification procedure is to reconstruct the unknown field Λ_0 of labels from the analysis of functional data indexed on the considered lattice S_0. Hence, the final result of the procedure is a label assignment for each site of the lattice, according to the observed functional data.

We will now sketch the algorithm via a pseudocode scheme. Specifications for the implementations of the algorithm will be detailed in Section 3. The procedure is a bagging-inspired algorithm, since it is composed by a *bootstrap* sampling phase, articulated in three basic steps, and by an *aggregation* phase (see Breiman, 1996 [3] for details on bagging procedures): at each replication of the three steps, a single weak classifier is found, which exploits a specific structure of spatial dependence, thus obtaining a coarse estimate of the unknown latent field of true labels Λ_0. A more accurate global classifier is obtained after B replications, by *bagging* together all single classifiers. Higher values of B, imply

a higher accuracy of the final estimate (the reconstruction of the latent field of labels Λ_0), which includes all the estimates of the B single classifiers. Moreover, the advantage of such an approach stands in the embarassingly parallel nature of the bootstrap phase, whose computational cost can be dramatically reduced by parallel programming.

Algorithm. Bagging Voronoi classifiers.

Bootstrap:
Initialize B, n, p, K. Choose a metric $d(\cdot,\cdot)$.
`for` $b := 1$ `to` B `do`
`step 1.` generate a random Voronoi tessellation of the lattice, i.e., isolate neighboring groups of data, to capture potential spatial dependence;
`step 2.` identify a local representative for each element of the tessellation to sum up local information: neighboring data are most likely drawn from the same functional distribution;
`step 3.` perform functional dimensional reduction of local representatives to select relevant functional features in the data, and cluster the projections of local representatives on the space spanned by the obtained basis to finally reconstruct the latent field of labels.
`end for`
Aggregation:
perform `cluster matching`: match the cluster labels along bootstrap replications, to ensure identifiability.
`for` $\mathbf{x} \in S_0$ `do`
• calculate the frequencies of assignment of the site to each one of the K clusters along iterations;
• compute spatial entropy for each site.
`end for`

First we have to fix the number B of replications of the three basic steps, the number n of elements in the Voronoi tessellation and the metric d used to compute distances, the p-dimensional basis for dimensional reduction of the local representatives and the number K of clusters considered in the clustering procedure. Then, for $b = 1, \dots, B$, steps 1-3 are replicated: a set of nuclei $\Phi_n^b = \{\mathbf{Z}_1^b, \dots, \mathbf{Z}_n^b\}$ is randomly generated among the sites in S_0, i.e. $\mathbf{Z}_i^b \overset{i.i.d.}{\sim} \mathcal{U}(S_0)$ for $i = 1, \dots, n$, where $\mathcal{U}$ is the uniform distribution on the lattice. Then, the b-th random Voronoi tessellation of the lattice S_0, $\{V(\mathbf{Z}_i^b|\Phi_n^b)\}_{i=1}^n$, is obtained 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 1`); see Section 3 for details. Given the tessellation, for $i = 1, \dots, n$ the local representative g_i^b, corresponding to the nucleus $\mathbf{Z}_i^b$ of the i-th element of the tessellation $V_i^b := V(\mathbf{Z}_i^b|\Phi_n^b)$, is computed (`step 2`); in this way we construct the b-th bootstrap sample. Then, dimensional reduction of the n local representatives $\{g_1^b, \dots, g_n^b\}$ is performed, by projecting them on the space spanned by a proper p–dimensional functional orthonormal basis, thus

generating the p-dimensional scores vectors $\{\mathbf{g}_1^b, \ldots, \mathbf{g}_n^b\}$ (`step 3`); the scores vectors are then clustered in K groups according to a suitable unsupervised method, depending on the application (e.g. K-mean clustering, PAM, outlier detection, ...). Since we are interested in a final classification map of the lattice S_0, we perform an aggregating phase in which results of each replications are bagged together. In particular, for $k = 1, ..., K$, and $b = 1, ..., B$, indicate with C_k^b the set of $\mathbf{x} \in S_0$ whose label is equal to k: cluster matching is needed to ensure coherence of cluster assignments along replications (see Section 3 for details). Then, the frequency distribution of assignment of each site to each of the K clusters along the B replications is considered. In fact, for each site $\mathbf{x} \in S_0$, one can compute $\pi_{\mathbf{x}}^k = \#\{b \in \{1, ..., B\} : \mathbf{x} \in C_k^b\}/B, \ \forall\ k = 1, \ldots, K$. A final assignment of site $\mathbf{x}$ to *one* of the K clusters can be obtained by selecting that label corresponding to a mode of the distribution $\pi_{\mathbf{x}} = (\pi_{\mathbf{x}}^1, ..., \pi_{\mathbf{x}}^K)$.

The previously described procedure for clustering spatially dependent functional data depends on a number of choices, which define the details of the algorithm, e.g. the parameters B, n, p and K have to be chosen in advance: while B should be chosen big enough to ensure the desired algorithm accuracy, p and n depend on the particular problem at hand, and their choice will be discussed in Section 3.

We shall now tackle the problem relative to the choice of the correct number K of clusters, by means of a *spatial entropy* index evaluating the quality of the final classification. Consider the frequency distribution of assignment $\boldsymbol{\pi}_{\mathbf{x}} = (\pi_{\mathbf{x}}^1, \ldots, \pi_{\mathbf{x}}^K)$ of each site $\mathbf{x} \in S_0$ to each of the K clusters. The entropy associated to the final classification in the site $\mathbf{x} \in S_0$ is obtained as

$$\eta_{\mathbf{x}}^K = -\sum_{k=1}^{K} \pi_{\mathbf{x}}^k \cdot \log(\pi_{\mathbf{x}}^k), \tag{1}$$

which assumes minimum value 0 when $\exists r : \pi_{\mathbf{x}}^r = 1, \pi_{\mathbf{x}}^k = 0 \, \forall k \neq r$, $k, r = 1, \ldots, K$, and maximum value $\log(K)$ when $\pi_{\mathbf{x}}^k = \frac{1}{K} \, \forall k = 1, \ldots, K$. The index in (1) is based on the criterion that the more the frequency distribution in $\mathbf{x}$ is concentrated on one particular label, the more the classification is precise and stable along replications; conversely, if most frequencies are uniformly spread over all labels, the uncertainty associated to the final classification in $\mathbf{x}$ is high. Spatial entropy can be visualized by plotting all the values obtained via equation (1) in each site of the lattice: when the chosen number of clusters K is optimal, we expect a neat plot, mostly equal to zero, and different from zero in sites that are expected to be at the boundaries between regions associated to different clusters. To obtain a global evaluation index, we can compute the *average normalized entropy*

$$\eta^K = \frac{\sum_{\mathbf{x} \in S_0} \eta_{\mathbf{x}}^K}{\log(K) \cdot |S_0|}, \tag{2}$$

including the contribution to the final classification quality of all sites in S_0. For comparisons over different choices of K, the quantity $\eta_{\mathbf{x}}^K$ in equation (2) has been normalized by its maximum value. Indeed, if K is not known, a comparison of the average normalized entropy for different values of K can be performed in

order to choose the optimal value of K: in fact, thanks to the definition given in equation (2) we can choose the optimal number of clusters K^* for the considered region as

$$K^* = \underset{K=1,\dots,K_{max}}{\operatorname{argmin}} \{\eta^K\}.$$

3 Details on the algorithm

We will now expand on the details of each step of the bagging Voronoi classifiers algorithm described through the pseudocode scheme in Section 2. Note that the extreme generality of the proposed algorithm makes each step flexible to different specifications eventually motivated by the application at hand.

3.1 Step 1: Voronoi tessellations

We will here give motivations for the chosen approach to the treatment of spatial dependence in lattice data via Voronoi tessellations. The motivation for using Voronoi tessellations comes from a consistency result proven by Penrose (2007, [11]) in the framework of stochastic geometry.

Consider $S \subset \mathbb{R}^d$. To univocally define a random tessellation it is necessary to select a set of points $\mathbf{x} \in S$ as nuclei for the Voronoi tessellation. Thus, let $\Phi_n = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n\}$ be a set of n points in S sampled from a proper distribution F defined on S: this will be the set of nuclei of the Voronoi tessellation. For each $\mathbf{Z}_i \in \Phi_n$, define the polyhedron

$$V(\mathbf{Z}_i|\Phi_n) = \{\mathbf{x} \in S : d(\mathbf{x}, \mathbf{Z}_i) \leq d(\mathbf{x}, \mathbf{Z}_j), \text{ for all } \mathbf{Z}_j \in \Phi_n,\ i \neq j\}, \tag{3}$$

to be the closed Voronoi cell with nucleus $\mathbf{Z}_i$ for the Voronoi tessellation induced by Φ_n, i.e. the set of $\mathbf{x} \in S$ lying at least as close to $\mathbf{Z}_i$, in the sense of the metric $d(\cdot,\cdot)$, as to any other point of Φ_n: the collection $\{V(\mathbf{Z}_i|\Phi_n)\}_{i=1}^n$ forms a Voronoi tessellation of S. There is no strong restriction on the choice of the metric $d(\cdot,\cdot)$, but the final tessellation will clearly depend both on the choice of the metric, and on the distribution F. Voronoi tessellations have many interesting properties, which make them good tools for partitioning a general domain (see Møller, 1994 [9] for further details on Voronoi tessellations in Euclidean spaces); however, the more interesting property for our purposes is undoubtedly a *coverage* property. Consider the collection of Lebesgue measurable sets $\{A_l\}_{l=1}^L$, $A_l \subset S\ \forall l = 1, \dots, L$, and let $V_i := V(\mathbf{Z}_i|\Phi_n)$; moreover, let

$$A_l^n := \bigcup_{\mathbf{Z}_i \in A_l} V_i,$$

be an approximation of A_l given by the Voronoi cells whose nuclei belong to A_l. The coverage property states that A_l^n represents a consistent estimator of the unknown set $A_l, \forall\ l \in \{1, \dots, L\}$, in the sense that ($\Delta$ denotes symmetric difference of sets)

$$\mu(A_l^n \Delta A_l) \overset{a.s.}{\to} 0, \quad n \to \infty, \tag{4}$$

where μ denotes the Lebesgue measure. The coverage property for Voronoi tessellations expressed in (4) has been proven by Penrose (2007, [11]) under the

reasonable assumption that the support of F includes S, and it represents a strong law of large numbers in the context of Voronoi tessellations.

The coverage property of Voronoi tessellations is fundamental to ensure the validity of our algorithm, since it states that, when the tessellation becomes less and less coarse, subsets of the domain S associated to the same label are well approximated by Voronoi tessellations. Indeed, with a view to our classification problem, we define the collection of sets $\{A_l\}_{l=1}^{L}$ as $A_l := \{\mathbf{x} \in S : \Lambda(\mathbf{x}) = l\}$, for $l = 1, \ldots, L$, where $\Lambda : S \to \{1, \ldots, L\}$ is such that $\Lambda|_{S_0} \equiv \Lambda_0$.

3.2 Step 2: functional local representatives

We consider the situation where we observe a realization $f_{\mathbf{x}} : T \to \mathbb{R}$ of a functional random variable in each site $\mathbf{x} \in S_0$, for $T = [t_{min}, t_{max}]$. Since our approach is completely non-parametric, we do not make assumptions on the distribution of $f_{\mathbf{x}}$. The Voronoi tessellation on S, which induces a tessellation on the lattice S_0, provides a partition of the region in random neighborhoods, and induces a partition in the sample of functional data $\{f_{\mathbf{x}}\}_{\mathbf{x}\in S_0}$. More precisely, for each element V_i of the tessellation, $i = 1, \ldots, n$, we consider the subset of functional data $\{f_{\mathbf{x}}\}_{\mathbf{x}\in V_i}$. To exploit spatial dependence of neighboring data we consider a *local functional representative* of data belonging to the same element of the Voronoi tessellation; in the literature on spatial statistics, this procedure is named spatial smoothing (see Banerjee *et al.*, 2004 [1] for further details). The computation of the functional local representatives here described, which is the one adopted in the application detailed in Section 5, is only one among the possible approaches; all extensions aimed at computing a local centroid (e.g. loess, functional median, ...) are conceivable.

The local representative g_i of the element V_i, for $i = 1, \ldots, n$, is computed as a local mean weighted with a Gaussian kernel

$$g_i(t) = \frac{\sum_{\mathbf{x}\in V_i} w_{\mathbf{x}}^i f_{\mathbf{x}}(t)}{\sum_{\mathbf{x}\in V_i} w_{\mathbf{x}}^i}, \tag{5}$$

where $w_{\mathbf{x}}^i$ is a Gaussian weight centered in $\mathbf{Z}_i$ and decreasing with respect to $d(\mathbf{x}, \mathbf{Z}_i)$, since we intuitively assume the spatial dependence between two sites to be decreasing with respect to the distance between them. In this sense the local representative already accounts for spatial dependence: a bigger contribution to its calculation, in fact, will be given by functional data associated to sites nearer to the nucleus of the element. The kernel covariance matrix is $\sigma^2\mathbb{I}_2$, where $\sigma = d_{max}/d_{min}$ being d_{max} and d_{min} the maximum and minimum distance between two nuclei of the tessellation, respectively; this choice is motivated by the fact that σ is in this way be related to the mean dimension of the tessellation element via an estimator of the elements mean diameter (see Møller, 1994 [9] for a proof in Euclidean spaces). This setting is general, and can be easily adapted to different situations arising in applications, in presence of proper information: for example, a non–diagonal covariance matrix in the Gaussian kernel could be used to account for anisotropy in the latent random field.

The choice of n, which sets the tessellation dimension and thus the number of local representatives to be computed, has great influence on the algorithm

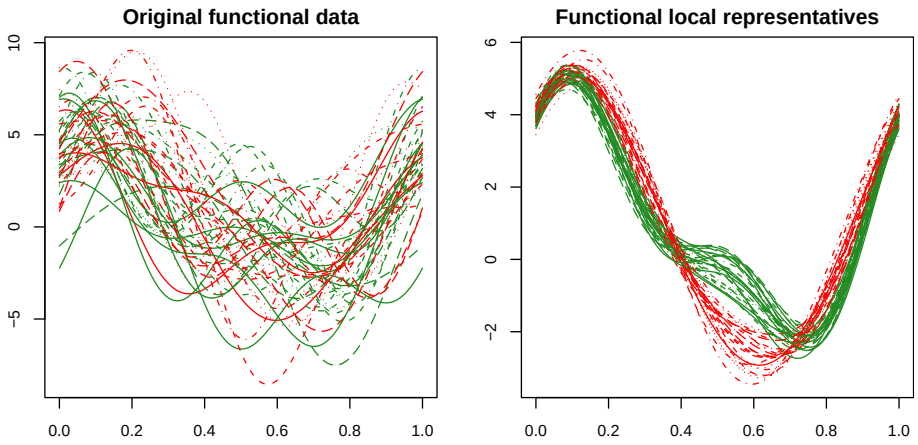

Figure 1: In the left panel, a sample of 50 synthetic data randomly selected from the realization of a Hidden Markov Random Field with Gaussian emission probability function; in the right panel, the sample of functional local representatives obtained using a Voronoi tessellation with $n = 50$. Different colors correspond to different labels.

behavior, since the latent field of labels is unknown: if the labels were known, we would choose a tessellation which perfectly follows the cluster borders, thus independently from the choice of n each local representative would be computed using data drawn from the same distribution. Indeed, labels are always unknown, inducing a bias-variance trade-off which determines the existence of an optimal choice of n. Consider the example in Figure 1. In the left panel a sample of 50 curves randomly selected from a synthetic data set is shown: they are generated according to a Hidden Markov Random Field with Gaussian emission probability function and parameter $\beta = 3$, where the latent field of labels is the realization of a Ising field on a 100×100 lattice of sites, and functional data are obtained using a Fourier basis with fixed ($p = 5$) dimension and coefficients obtained from the emitted random field; the mean vectors of the emission probability function are $\boldsymbol{\mu}_{-1} = (1, 2, 2.25, 0, 0)$ and $\boldsymbol{\mu}_1 = (1, 1.5, 1.25, 0.75, 0.75)$, respectively, and covariance matrices are the identity in $\mathbb{R}^5$ in both cases. Thus the functional distribution of generated data is a mixture, where mixture components are associated to the two labels of the latent field. In the right panel the functional local representatives corresponding to the same data set are depicted, when the tessellation dimension is chosen equal to $n = 50$ and local representatives are computed using equation (5). While in the left panel we can hardly distinguish a grouping structure, since the variability within the two groups is confounding the one between the groups, in the right panel of the picture, instead, we can distinguish curves belonging to two different groups thanks to the evident reduction of the variance in the sample of local representatives. Moreover, the portion of the domain in which the mean curves of the two clusters are the most different is also evident in the right panel; this is due to the fact that bias is reduced in the computation of local representatives.

In general, a certain number of representatives will be optimal in terms of

misclassification error: the optimal choice of n is the one that finds a good compromise between variance and bias. This is due to the mentioned bias-variance trade-off:

- as n decreases, noise is reduced in the local representatives sample, since local representatives are weighted sample means calculated on an averagely larger dataset (minimal variance); however, at the same time the associated Voronoi tessellation follows less accurately the boundaries in the true latent field of labels, thus including different mixture components in the calculation of local representatives (maximal bias). The limiting case is $n \equiv 1$, when all sites in the finite lattice belong to the same Voronoi element, and are thus used to compute a single representative;

- as n increases, the resulting Voronoi tessellation approximates more accurately the boundaries of the latent field of labels (minimal bias), but at the same time the variability reduction due to spatial smoothing is smaller (maximal variance). The limiting case is $n \equiv |S_0|$, when all sites in the finite lattice are nuclei, and thus no spatial smoothing is performed.

The optimal value of n determined by this trade-off depends both on the strength of spatial dependence, and on the mixture components of the distribution of the functional signal. In Section 4 a simulation study aimed at stating the existence of the optimal value of n in some realistic situations is detailed.

3.3 Step 3: dimensional reduction and clustering

We now describe in details the third step of the spatial clustering procedure, which aims at performing data dimensional reduction and at clustering reduced data, to obtain a classification map. We thus introduce functional data analysis techniques aimed at catching the most relevant features in the functional distribution of the sample, useful for the treatment of local representatives $\{g_1, \ldots, g_n\}$ (see Ramsay and Silverman, 2005 [13]).

In order to obtain a reduction in the infinite dimension of functional data, we need to find the best projection of data onto a proper functional basis. The choice of this basis is extremely open, and heavily depends on the application: we describe one of the possible methods, the one adopted in our motivating application described in Section 5. However, in general we shall distinguish two possible situations. If the functional basis used for p-dimensional reduction of the sample of local representatives is fixed along replications (e.g., a wavelet basis, or a Fourier basis), then the only task to accomplish is the projection of the local representatives on the given basis, but the resulting algorithm is less flexible. If, instead, the functional basis is data–driven, then there is great flexibility with respect to the functional features possibly arising in applications.

Among the latter approaches to p–dimensional reduction, one of the possibilities consists in performing functional principal component analysis, i.e., Karhunen-Loève decomposition; indeed, the basis composed by functional principal components, is a complete orthonormal basis of $L^2(T)$ (see Ferraty and Vieu, 2006 [6] for theoretical details). Note that we have to assume $\{g_1, \ldots, g_n\}$

to be independent realizations of a random function $g = \{g(t), t \in T\}$ with finite second moment, in order to consistently estimate the covariance structure needed to find principal components. However, we expect the spatial dependence in the local representatives sample to be highly reduced, so that we can reasonably consider local representatives as an independent sample from a functional mixture distribution.

Hence, we estimate the covariance operator $\Gamma_g(s,t)$ via

$$\hat{\Gamma}_g(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(g_i(t) - \bar{g}(t))(g_i(s) - \bar{g}(s)), \tag{6}$$

where $\bar{g}$ is the estimated functional mean of the local representative sample, and these functional estimates are obtained via numerical approximations provided the density of the grid of measurements for each functional data is sufficiently large. The orthonormal eigenfunctions $\{\nu_1(t), \nu_2(t), \ldots\}$, and the associated eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$, of the covariance operator are estimated as the solutions $\{\hat{\nu}_k, \hat{\lambda}_k\}_{k\geq 1}$ of the eigenequation

$$\int_T \hat{\Gamma}_g(s,t)\hat{\nu}_k(s)ds = \hat{\lambda}_k\hat{\nu}_k(t),$$

where each function $\hat{\nu}_k$ detects an orthonormal direction in the functional space $L^2(T)$, explaining a decreasing portion of variability $\hat{\lambda}_k$ in the sample of local representatives; the projections of local representatives in each direction are called *scores*. Thus, to perform dimensional reduction of the sample of local representatives $\{g_1, \ldots, g_n\}$, only the first p eigenfunctions are used to represent data, those which, according to a graphical inspection, are explaining features associated to a grouping structure. Alternatively, when using Karhunen–Loève decomposition, p could be determined by fixing a given portion of the variability to be explained by selected components.

Once chosen the most relevant p components of the functional basis, only projections of data along relevant components are analyzed, thus performing dimensional reduction. Hence, to meet the final task of the classification procedure, we can perform k-mean clustering in a multivariate setting considering the sample of FPCA scores $\{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$. Note that other clustering methods can be used to obtain a final classification map, e.g., hierarchical methods, PAM or sparse clustering, depending on the scopes of the analysis and on the application.

3.4 Aggregation: cluster matching

Consider the b-th replication of the bagging Voronoi classifiers algorithm, and suppose to have generated a Voronoi tessellation (`Step 1`), and to have obtained a sample of local representatives (`Step 2`), which have then been projected on a proper basis and clustered (`Step 3`). Let $\Gamma_1^b, \ldots, \Gamma_n^b$ denote the labels of the local representatives at the b-th replication, i.e. $\Gamma_i^b \in \{1, 2, \ldots, K\}$ is the final cluster assignment of the function g_i^b, for $i = 1, \ldots, n$, to one of the K clusters; hence, all sites $\mathbf{x}$ in V_i^b get the label Γ_i^b. For $k = 1, ..., K$, indicate with C_k^b the set of $\mathbf{x} \in S_0$ whose label is equal to k. Since our final aim stands in obtaining a

final classification map of the lattice S_0, all we need is the frequency distribution of assignment of each site to each of the K clusters along the B replications. The computation of the frequency distribution of assignment, however, is based on the assumption that cluster labels $\{C_1^b, \ldots, C_K^b\}$ are coherent along replications of the algorithm; more specifically, we want cluster labels $\{C_1^b, \ldots, C_K^b\}$ to be coherent with $\{C_1^m, \ldots, C_K^m\}$, for all $m < b$, and $b \geq 2$.

The coherence of cluster labels is ensured by *cluster matching*. The idea is the following: we consider only subsequent replications. If $b \geq 2$, in the current replication of the procedure the labels identifying the clusters $C_1^b, ..., C_K^b$ are renamed by matching them with the cluster assignments $C_1^{b-1}, ..., C_K^{b-1}$ obtained at the previous replication; indeed the algorithm looks for the label permutation $\{l_1, \ldots, l_K\}$ in the set $\{1, \ldots, K\}$ that minimizes the total sum of the off-diagonal frequencies in the contingency table describing the joint distribution of sites along the two classifications $C_1^{b-1}, ..., C_K^{b-1}$ and $C_{l_1}^b, ..., C_{l_K}^b$. Other different procedures for cluster matching are conceivable.

4 Simulation study on synthetic data: optimal choice of n

We now describe a simulation study to test bagging Voronoi classifiers algorithm on synthetic data. It aims at testing the algorithm performance with respect to the choice of the number of elements composing the Voronoi tessellation, n, under different correlation structures in the latent field of labels: we assume stronger/weaker spatial dependence in the latent field of labels, and we compare results obtained varying the coarseness of the Voronoi tessellation. Since the focus of this simulation is on spatial dependence, and on the existence of a value of n (optimal tessellation) which minimizes the misclassification error, in this simulation study the emitted random field is multivariate.

Here, S_0 is a two-dimensional square lattice of 50×50 sites and the latent field of labels is generated by a Ising Markov Random field $\Lambda_0 : S_0 \to \{-1, 1\}$: the Ising model has been extensively studied in statistical physics to describe the behavior of magnetic materials, and according to this model the probability of a configuration of sites $\mathbf{x} \in S_0$ depends on its energy. Precisely, we have

$$\mathbb{P}\left(\Lambda_0(S_0) = \boldsymbol{\lambda}\right) = \frac{1}{Z} \exp\left\{\beta \sum_{\mathbf{x} \in S_0} \sum_{\mathbf{x}' \in \mathcal{N}_{\mathbf{x}}} \lambda(\mathbf{x})\lambda(\mathbf{x}')\right\},$$

where Z is a normalizing constant, $\boldsymbol{\lambda} = \{\lambda(\mathbf{x}), \mathbf{x} \in S_0\}$ collects the realizations of the field on each site of the lattice S_0 and $\mathcal{N}_{\mathbf{x}}$ is a proper neighborhood of $\mathbf{x} \in S_0$. The strength of spatial dependence is controlled by the parameter β, a physical constant characterizing the influence of neighboring sites on the realization observed in a particular site: higher values of β imply a stronger spatial dependence (see Kunsch, 1995 [8] for more details on this model). Hence, for each site $\mathbf{x} \in S_0$, denote by $\Lambda_0(\mathbf{x})$ its true label drawn from a Ising field. For our simulation studies, we let β range in the interval $(0.5, 1)$. Conditionally on the realization of the latent field, in each site of S_0 we generate independently

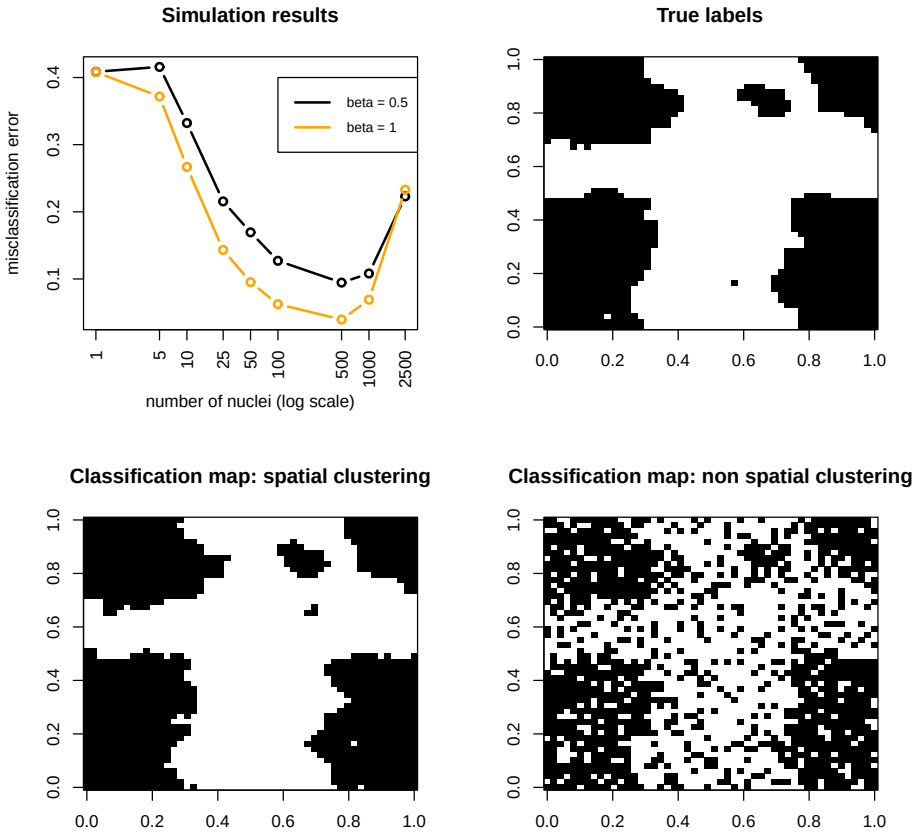

Figure 2: Results of multivariate simulation study: misclassification error obtained via bagging Voronoi classifiers algorithm over different choices for n and β – mean over 30 repetitions of the procedure (top, left); true labels, one of the realizations of the Ising field with $\beta = 1$ used for the analysis (top, right); final classification map obtained via spatial clustering with $n = 500$ (bottom, left); final classification map obtained via non–spatial clustering (bottom, right).

a random multivariate vector of dimension $p = 5$ from a multivariate Gaussian distribution; the distribution of the random vector depends exclusively on the site label. For $\mathbf{x}$ ranging in S_0, we thus obtain

$$\mathbf{Y}_{\mathbf{x}}|(\Lambda_0(\mathbf{x}) = l) \sim N_p(\boldsymbol{\mu}_l, \Sigma),$$

where $\Sigma = \sigma \mathbf{I}_p$, being $\mathbf{I}_p$ the identity matrix in p-dimensions, and $\sigma = 2$. This means that the conditional distribution of the observed signal given the label differs only in the mean for different values of the label. In particular we choose $\boldsymbol{\mu}_{-1} = \mathbf{0}$ and $\boldsymbol{\mu}_1 = (-2, -1, 0, 1, 2)$.

The parameters controlling the algorithm are fixed as follows: $B = 50$, $K = 2$ and $n \in \{1, 5, 10, 25, 50, 100, 500, 1000, 2500\}$. The Voronoi tessellations are obtained uniformly drawing the set of nuclei in S_0, and generated by means of the Euclidean distance. The n representatives are identified as weighted means with Gaussian isotropic weights while no dimensional reduction of the representatives is performed. Finally, clusterization of the representatives is obtained through

K-means. The final classification map is obtained through a majority vote on cluster assignment, and the result is evaluated by computing a misclassification error rate with respect to the true realization of the field. The final evaluation of the algorithm is obtained by repeating the simulation 30 times, and by calculating a mean misclassification rate.

Results are illustrated in Figure 2. Consider the top/left panel of the picture, showing the mean misclassification error for different values of β and n. First, consider the behavior of the mean misclassification error with respect to n: we appreciate the existence, for both values of β, of a value of n that minimizes the misclassification error. Moreover, looking at the top/left panel of the picture, we notice that misclassification error is uniformly smaller (with respect to n) for higher values of β: hence the improvement introduced by the bagging Voronoi classifiers algorithm is stronger in the presence of a stronger spatial dependence in the latent field of labels, for any chosen value of n. Note that the limiting case $n = 50 \times 50 = 2500$ corresponds to the application of a non–spatial clustering procedure, namely a standard k-mean clustering. It is clear from the picture that this technique has a poor behavior in terms of misclassification error compared to spatial clustering, for nearly all values of n. Finally, in the top/right panel, one realization of the true latent field of labels obtained for one of the simulated data sets is shown ($\beta = 1$), while in the bottom panels the results obtained via spatial clustering with $n = 500$ (left), and via non–spatial clustering (right) are depicted. It is evident from the picture that, although the final map obtained via non–spatial clustering is suggestive of the true label pattern, the final result obtained via bagging Voronoi classifiers algorithm is far more precise, and the classification map nearly coincides with the true label field (misclassification errors are 3.28% and 23.08% for spatial and non–spatial clustering, respectively).

5 A case study: clustering irradiance data

We now illustrate an application of our classification algorithm to irradiance data, carried out to investigate the possible exploitation of solar energy in different areas of the planet. In particular, power production via collectors that are able to track the sun diurnal course is strongly influenced by solar irradiance and atmospheric conditions. In fact, solar thermal power employs only direct sunlight and it is therefore best positioned in areas, such as deserts, steppe or savannas, where large amounts of humidity, fumes or dust, that may deviate the sunbeams, do not occur (see Richter, 2009 [14]).

Insolation is a measure of solar radiation energy received on a given surface area in a given time. It is commonly expressed as average irradiance in kilowatt–hours per square meter per day (kWh/(m^{2}day)). We consider *direct insolation*, i.e., the solar irradiance measured at a given location on earth with a surface element perpendicular to the sunbeams, excluding diffuse insolation (the solar radiation that is scattered or reflected by atmospheric components in the sky). Direct insolation is equal to the solar constant minus the atmospheric losses due to absorption and scattering: while the solar constant varies with the earth–sun distance and solar cycles, the losses depend on the time of day (length of

light's path through the atmosphere depending on the solar elevation angle), cloud cover, moisture content, and other impurities.

We try to identify areas of the planet which are optimal with respect to the positioning solar power collectors by considering parameters, which depend on direct insolation, suited for sizing batteries or other energy-storage systems: the typical solar insolation parameters are the *minimum available insolation over a consecutive–day period (%)*, the *solar radiation deficits below expected values incident on a horizontal surface over a consecutive–day period (kWh/m^2)* and the *equivalent number of NO-SUN or BLACK days that must be supplied by the storage backup system (days)*. All these choices are fully described in the NASA website at `http://eosweb.larc.nasa.gov/sse/text/definitions.html`. These parameters are desired because of the fact that unusually cloudy conditions occurring over a number of consecutive days continually draw reserve power from batteries or some other storage device for solar systems not connected to an electrical power grid. Storage devices must be designed to withstand continuous below–average conditions in various regions of the globe. More precisely, we analyze the maximum solar radiation deficit below expected value incident on a horizontal surface over a consecutive–day period (kWh/m^2), which is strictly related to the equivalent number of NO–SUN or BLACK days, and which is also increasing in the monthly average irradiance (see the NASA website [10] for details on available datasets). This quantity, from an engineering point of view, is considered as a proxy of the buffer extra-capacity that is needed to be installed – in a particular site at a particular time of the year – in order to fullfill the possible gaps in energy supply provided by solar power plants due to unfavorable environmental conditions (from now on, we will name this quantity *buffer capacity*).

Rough data consist of 12 monthly observations in each site measuring for each month the maximum energy deficit, with respect to the monthly average, observed in between July 1983 and June 2005. Sites are located on a non–uniform lattice $S_0 = \bigcup_{\lambda \in Z_1; \theta \in Z_2} A_{\lambda\theta}$, where $Z_1 = \mathbb{Z} \cap [-180; 179]$ and $Z_2 = \mathbb{Z} \cap [-66; 66]$: each element $A_{\lambda\theta}$ is the portion of the earth surface which is included between the meridians at longitude λ and $\lambda + 1$ in degrees, and between the parallels at latitude θ and $\theta + 1$ in degrees; this lattice is of course non–uniform, and includes 47880 worldwide non–polar districts. In each site of the lattice, we observe the buffer capacity $Y^{\nu}_{\lambda,\theta}$ at a given month ν during the year. A set of functional data $Y_{\lambda,\theta}(t)$ can be obtained from raw data $Y^{\nu}_{\lambda,\theta}$, via a Gaussian kernel smoother with a bandwidth equal to 1.5: in this way we reconstruct the annual functional pattern of buffer capacity in each site, which is the input of the bagging Voronoi classifiers algorithm.

For this application, we set the algorithm parameters as follows: $B = 100$ and $n = 300$; a set of n elements is repeatedly drawn from a uniform distribution on S (the surface of the sphere of diameter equal to the earth), and the set of nuclei for the Voronoi diagram is then chosen by selecting the n sites among $\mathbf{x} \in S_0$ nearest in terms of geodesic distance to each of the n generated elements. We then use a Gaussian isotropic kernel to calculate local representatives, and we choose the first $p = 3$ functional principal components to project data. Finally, we use k-mean clustering with the L^2 semi–metric induced by the principal components,

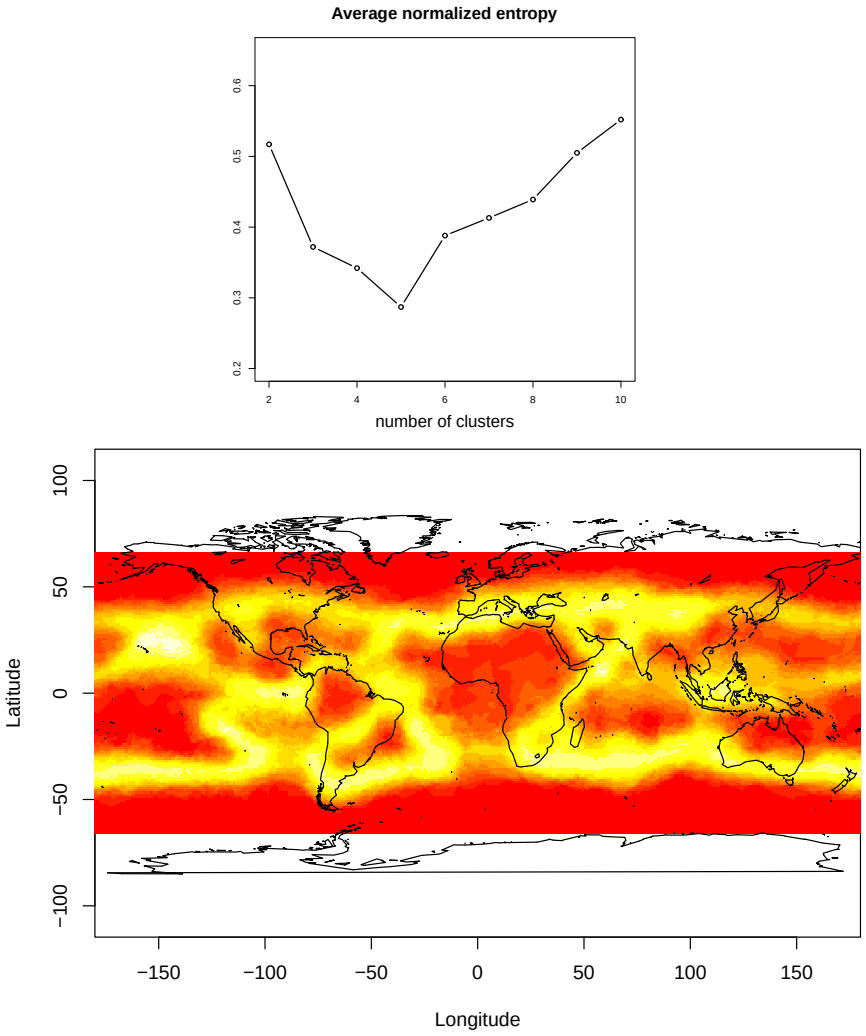

Figure 3: Results of spatial clustering on buffer capacity data from the *Surface meteorology and Solar Energy database*: in the top panel, average normalized entropy obtained via spatial clustering with different choices of K; in the bottom panel, normalized spatial entropy associated to the classification with $K = 5$. In the bottom panel, colors from red to white correspond to values from 0 to 1; higher values identify areas where classification is more uncertain.

and we choose the optimal K through the evaluation of the classification map by means of entropy.

The best classification according to spatial entropy evaluation is obtained for $K = 5$: in the top panel of Figure 3 the average normalized entropy obtained via spatial clustering with different choices of K is shown, while in the bottom panel of the same picture the map of normalized spatial entropy obtained setting $K = 5$ is shown (colors from red to white correspond to a normalized entropy from 0 to 1). We notice that the choice of $K = 5$ is the one that gives the better classification result according to entropy minimization: for $K < 5$ entropy is higher on average due to the random merging of true underlying groups,

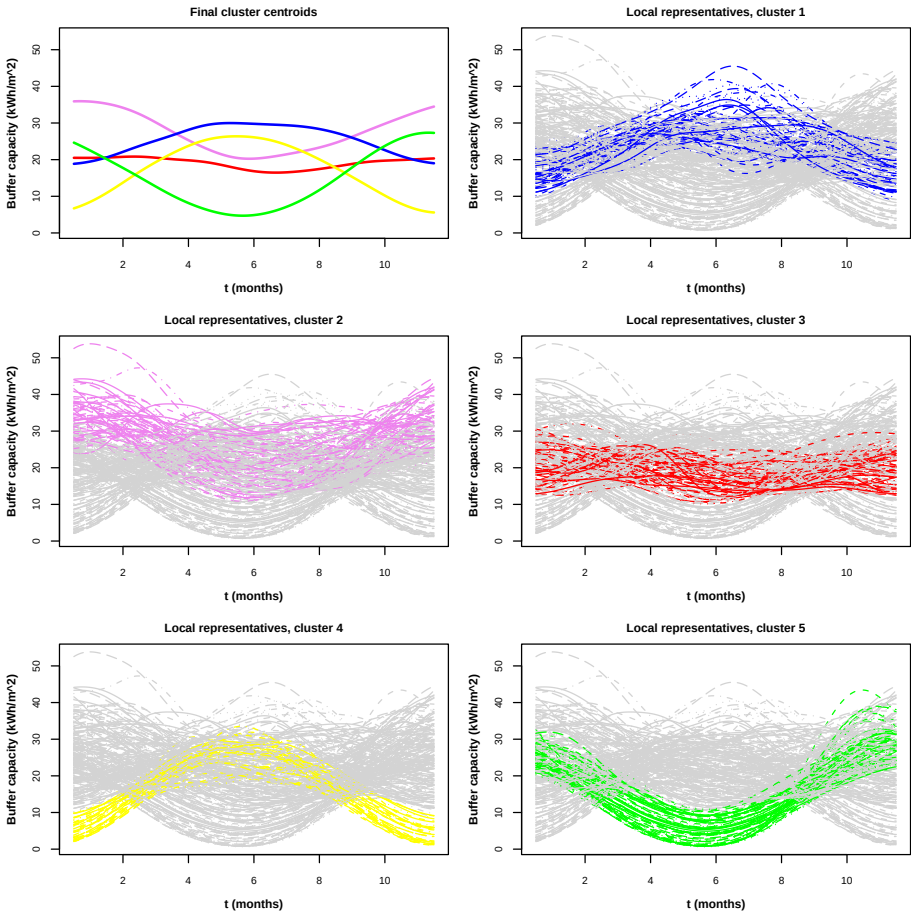

Figure 4: In the top left panel, set of final cluster centroids obtained via spatial clustering procedure with $K = 5$: different colors correspond to the different labels in the final clustering associated to the macro–areas in Figure 5. In the other five panels, result of clustering on a set of functional local representatives obtained with $n = 300$ in one of the iterations of the procedure: a single cluster is coloured in each panel, and the color is chosen coherently with the final classification map in Figure 5, whilst all other data are shown in gray.

while for $K > 5$ artificial clusters arise. With this particular choice for the number of clusters, the spatial clustering algorithm identifies different homogeneous macro-areas which – prima facie – seem interpretable in terms of the observed phenomenon, even though a climatological analysis, which is beyond the scopes of this paper, could deepen their explanation; indeed, the same macro-areas are not captured by customary unsupervised classification procedures, that do not take into proper account the spatial dependence among data. The final results for the choice of $K = 5$ are shown in Figures 4 and 5. In Figure 4 a sample of local representatives is shown, each representative colored with a label corresponding to the macro-area in Figure 5 it belongs to. The red cluster is characterized by a non-seasonal pattern, and by high average buffer capacity along the year; it covers Africa, Middle-East and equatorial America and its presence is not ex-

plained only in terms of latitude. From North to South we can then identify four clusters with seasonal patterns depending on the hemisphere and on the average buffer capacity along the year: north-low (yellow), north-high (blue), south-high (violet), south-low (green). It is interesting to note that, while in the Americas all 5 clusters are present, the north-high and south-high clusters are absent in Europe and Africa, and the red cluster is absent in Asia. Note also that the red cluster is the one that shows an annual buffer capacity pattern which is optimal from an engineering point of view: it needs the minimal-energy installation (the maximal annual need for energy is the lowest among the 5 detected patterns), associated to a constant reliability along the year.

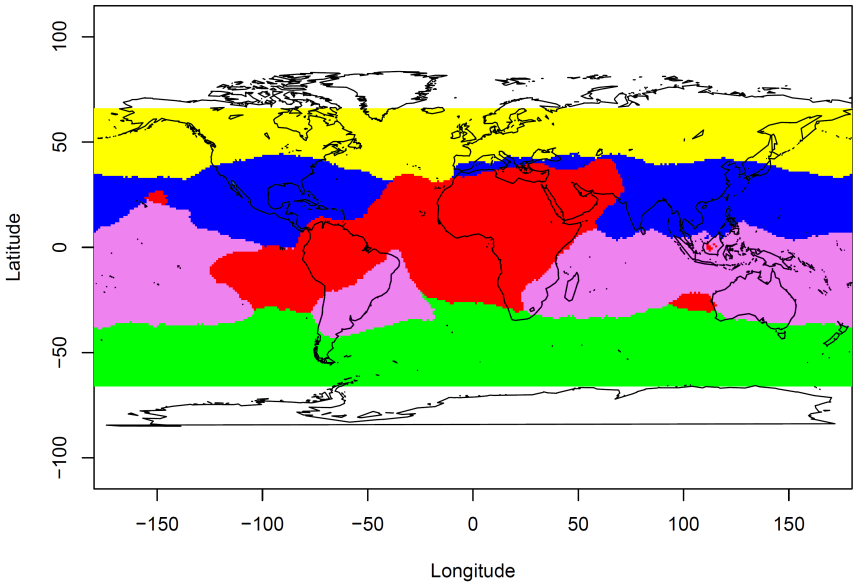

Figure 5: Results of spatial clustering on buffer capacity data from the *Surface meteorology and Solar Energy database*: final classification map obtained by setting $K = 5$ via a majority vote on frequencies of assignment.

References

[1] Banerjee, S., Carlin, B., Gelfand, A., 2004. *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, Monographs on Statistics and Applied Probability.

[2] Besag, J., 1986. On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 48(3):259–302.

[3] Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24:123–140.

[4] Cuesta-Albertos, J. A., Fraiman, R., 2007. Impartial trimmed k-means for functional data. *Comput. Statist. Data Anal.* 51:4864–4877.

[5] Delicado, P., Giraldo, R., Comas, C., Mateu, J., 2010. Statistics for spatial functional data: some recent contributions. *Environmetrics.* 21:224–239.

[6] Ferraty, F., Vieu, P., 2006. *Nonparametric Functional Data Analysis.* Physica-Verlag/Springer, Heidelberg.

[7] Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6:721–741.

[8] Kunsch, H., Geman, S., Kehagias, A., 1995. Hidden markov random fields. *Ann. Appl. Prob.* 5(3):577–602.

[9] Møller, J., 1994. *Lectures on Random Voronoi Tessellations*, Springer, New York.

[10] NASA, Surface meteorology and Solar Energy, A renewable energy resource web site (release 6.0). http://eosweb.larc.nasa.gov/cgi-bin/sse/sse.cgi?#s01, [accessed on the 25^{th} of November, 2010].

[11] Penrose, M. D., 2007. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli.* 13(4):1124–1150.

[12] R Development Core Team (2006), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, `http://www.R-project.org`.

[13] Ramsay, J. O., Silverman, B. W., 2005. *Functional Data Analysis*, second ed., Springer, New York.

[14] Richter, C., Teske, S., Nebrera, J. A., 2009. Concentrating solar power global outlook 09. Technical report, Greenpeace International / European Solar Thermal Electricity Association (ESTELA) / IEA SolarPACES.

[15] Tarpey, T., Kinateder, K. K. J., 2003. Clustering functional data. *J. Classification.* 20:93–114.

MOX Technical Reports, last issues

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

26/2011 Secchi, P.; Vantini, S.; Vitelli, V.
Bagging Voronoi classifiers for clustering spatial functional data

25/2011 de Luca, M.; Ambrosi, D.; Robertson, A.M.; Veneziani, A.; Quarteroni, A.
Finite element analysis for a multi-mechanism damage model of cerebral arterial tissue

24/2011 Manzoni, A.; Quarteroni, A.; Rozza, G.
Model reduction techniques for fast blood flow simulation in parametrized geometries

23/2011 Beck, J.; Nobile, F.; Tamellini, L.; Tempone, R.
On the optimal polynomial approximation of stochastic PDEs by Galerkin and Collocation methods

22/2011 Azzimonti, L.; Ieva, F.; Paganoni, A.M.
Nonlinear nonparametric mixed-effects models for unsupervised classification

21/2011 Ambrosi, D.; Pezzuto, S.
Active stress vs. active strain in mechanobiology: constitutive issues

20/2011 Antonietti, P.F.; Houston, P.
Preconditioning high–order Discontinuous Galerkin discretizations of elliptic problems

19/2011 Passerini, T.; Sangalli, L.; Vantini, S.; Piccinelli, M.; Bacigaluppi, S.; Antiga, L.; Boccardi, E.; Secchi, P.; Veneziani, A.
An Integrated Statistical Investigation of the Internal Carotid Arteries hosting Cerebral Aneurysms

18/2011 Blanco, P.; Gervasio, P.; Quarteroni, A.
Extended variational formulation for heterogeneous partial differential equations

17/2011 Quarteroni, A.; Rozza, G.; Manzoni, A.
Certified Reduced Basis Approximation for Parametrized Partial Differential Equations and Applications