Bagging Voronoi-classifiers for clustering spatial functional data

Piercesare Secchi, Simone Vantini and Valeria Vitelli

1 Introduction

We consider the problem of clustering functional data indexed by the sites of a spatial finite lattice, motivated by the analysis of the environmental data contained in the *Surface Solar Energy* database (NASA 2010). To this purpose, we exploit the bagging Voronoi-classifiers algorithm introduced in Secchi *et al.* (2012), based on repeatedly partitioning the investigated area in random neighborhoods, and on replacing the original data set with a reduced one, composed by local representatives of neighboring data. In this way we obtain many different weak formulations of the analysis, whose results are then bagged together to give a conclusive strong analysis.

The analysis of high–dimensional spatial data is a recent topic in the statistical literature (see the interesting review by Delicado *et al.* 2010), and very often the focus is on prediction rather than being on classification problems. Here we present an application of the bagging-Voronoi strategy for unsupervised classification of functional data (Secchi *et al.* 2011, 2012) to the analysis of the *Surface Solar Energy* database (NASA 2010). This methodology is completely non-parametric, since it does not rely on any explicit parametric assumption for the spatial dependence of functional data. Moreover, the computational cost of this approach is low, since the analysis of the original data set is simply replaced by the analyses of smaller data sets: this fact makes the bagging Voronoi-classifiers approach particularly appealing

Piercesare Secchi
MOX – Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, Milano, Italy e-mail: piercesare.secchi@polimi.it

Simone Vantini
MOX – Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, Milano, Italy e-mail: simone.vantini@polimi.it

Valeria Vitelli
MOX – Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, Milano, Italy e-mail: valeria.vitelli@mail.polimi.it

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

for the analysis of large data sets, and it opens to parallel computing implementations.

The analysis here presented is carried out to investigate the possible exploitation of solar energy for power production in different areas of the planet, which strongly depends on solar irradiance and atmospheric conditions (Richter 2009). In particular we aim at identifying different homogeneous macro-areas, interpretable in terms of the observed phenomenon, and not captured by traditional unsupervised functional classification procedures (Tarpey and Kinadeter 2003). The *Surface Solar Energy* database consists of 47520 solar irradiance patterns along the year observed in 47520 worldwide non-polar districts of a non-uniform global lattice formed by meridians and parallels, and covering the surface of the earth.

This communication is structured as follows. In Section 2, the bagging Voronoi-classifiers algorithm for clustering spatially dependent functional data is described. Section 3 focuses instead on the application of the bagging Voronoi-classifiers algorithm to irradiance data. The analysis of real data sets is performed in R (R Development Core Team 2006).

The properties of the bagging Voronoi-classifiers algorithm are deeply discussed in Secchi *et al.* (2012), where its efficiency - with respect to non-spatial clustering techniques - is tested and its behavior investigated through a large battery of MC simulations. Please refer to the latter for further details.

2 Bagging Voronoi-classifiers for clustering spatially dependent functional data

Consider a possibly non uniform lattice of sites S_0, and consider the situation in which a functional datum is observed in each site $\mathbf{x} \in S_0$. We have in mind the following generating model for our functional data set: a latent field of labels $\Lambda_0 : S_0 \to \{1,\dots,L\}$ is defined in each site $\mathbf{x} \in S_0$, such that $\Lambda_0(\mathbf{x})$ is the true unknown label associated to the site $\mathbf{x}$. From an application perspective, this field is thought to sum up characteristics of the considered area which are interesting for the scopes of the analysis. Then, given the field Λ_0, the functional observable data are thought independently generated in each site $\mathbf{x} \in S_0$ from a distribution indexed by $\Lambda_0(\mathbf{x})$. The main purpose of the analysis is the reconstruction of the unknown field Λ_0 of labels, based on the clustering of the functional observed data indexed by the sites of S_0. Hence, we need a procedure which is able to perform the classification of the observed functional data, thus returning as a final result a label assignment for each site of the lattice.

The following box reports the pseudo-code scheme of the algorithm. The procedure is a bagging-inspired clustering algorithm, composed by a *bootstrap* sampling phase, articulated in three basic steps, and by an *aggregation* phase (see Breiman, 1996 for details on bagging procedures):

Algorithm. Bagging Voronoi-classifiers.

Bootstrap:
Initialize B, n, p, K. Choose a metric $d(\cdot,\cdot)$.
`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 $i = 1,\ldots,n$, $\mathbf{Z}_i^b \overset{i.i.d.}{\sim} \mathscr{U}(S_0)$, where $\mathscr{U}$ 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.` perform dimensional reduction of the local representatives $\{g_1^b,\ldots,g_n^b\}$ by projecting them on the space spanned by a proper p–dimensional functional orthonormal basis, thus generating the p-dimensional scores vectors $\{\mathbf{g}_1^b,\ldots,\mathbf{g}_n^b\}$, which are then clustered in K groups according to a suitable unsupervised method.
`end for`
Aggregation:
perform `cluster matching`: 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, and match the cluster labels across bootstrap replicates, to ensure identifiability.
`for` $\mathbf{x} \in S_0$ `do`
• calculate the frequencies of assignment of the site to each of the K clusters along iterations, i.e., $\pi_{\mathbf{x}}^k = \#\{b \in \{1,...,B\} : \mathbf{x} \in C_k^b\}/B$, $\forall\, k = 1,\ldots,K$;
• compute spatial entropy $\eta_{\mathbf{x}}^K$ for each site $\mathbf{x} \in S_0$.
`end for`

The general idea of the procedure is finding, at each replicate of the three steps, a single weak classifier, which exploits a different specific structure of spatial dependence; in this way a coarse estimate of the unknown latent field of true labels Λ_0 is obtained. After B replicates, by *bagging* together the estimates given by all weak single classifiers we finally end up with a more accurate global classifier, which includes results of all single replicates. Hence, higher values of B, imply a higher accuracy of the final estimate (the reconstruction of the latent field of labels Λ_0). More precisely, in `step 1` of the bootstrap sampling part of the algorithm, neighboring groups of data are isolated by partitioning the lattice via a random Voronoi tessellation, to capture potential spatial dependence. The property of Voronoi tessellations which justifies their use in the nonparametric treatment of spatial dependence is a consistency property, proven in Penrose (2007) in the context of stochastic geome-

try. In `step 2` local information is summed up via the identification of a local representative for each element of the tessellation, guided by the rationale that neighboring data are most likely drawn from the same functional distribution. In `step 3` relevant functional features in the data are detected via functional dimensional reduction of local representatives; the subsequent clustering in K groups (performed according to a suitable unsupervised method) is based on the projections of local representatives on the space spanned by the obtained basis. In the aggregating phase a final classification map of the lattice S_0 is obtained: results of each replicate are bagged together to ensure a stronger final result. Note that `cluster matching` is needed for the coherence of cluster assignments across replicates. The aggregating strategy in the case of clustering is the computation of the frequency distribution of assignment of each site to each of the K clusters along the bootstrap replicates: in this application, a final assignment to *one* of the K clusters for each site in S_0 is obtained by selecting the label corresponding to a mode of the frequency distribution.

Each particular implementation of the procedure depends on some parameters that require to be carefully tuned. Among these ones, the most relevant are the dimension n of the Voronoi tessellations, and the correct number K of clusters. The parameter n, which sets the dimension of the Voronoi tessellations and thus the number of local representatives to be computed, has great influence on the algorithm behavior, since it induces a strong bias-variance trade-off:

- as n decreases, noise is reduced in the local representatives sample, since local representatives are weighted sample means calculated on sub-samples that are larger on average (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 computation 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 of the representatives increases since the average number of data per element decreases (maximal variance). The limiting case is $n \equiv |S_0|$, when all sites in the finite lattice are nuclei, and thus the local representatives sample coincides with the original dataset.

The optimal value of n determined by this trade-off depends both on the strength of the spatial dependence, and on the mixture components of the distribution generating the functional data. In particular, the quality of the final classification is here evaluated by means of a spatial entropy criterion that consequently drives also the choice for n. Consider the frequency distribution of assignment $\pi_{\mathbf{x}} = (\pi_{\mathbf{x}}^1, \ldots, \pi_{\mathbf{x}}^K)$ of each site $\mathbf{x} \in S_0$ to each of the K clusters generated by the B replicates. The entropy associated to the final classification in the site $\mathbf{x} \in S_0$ is obtained as

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

Hence, the more the frequency distribution $\pi_{\mathbf{x}}$ is concentrated on one particular label, the lower the index (1) is, and the more the classification is precise and stable along replicates. Conversely, when frequencies are more uniformly spread over all labels, the value of the index (1) is higher, and uncertainty associated to the final classification in $\mathbf{x}$ is greater. A global evaluation index can also be computed as the *average normalized entropy*

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

including the contribution to the final classification quality of all sites in S_0. For comparisons over different choices of K, the quantity $\eta^K_{\mathbf{x}}$ in equation (2) has been normalized by its maximum value. Since the index expressed in (1) is a measure of the uncertainty associated to the final classification, we expect the value of η^K to be low if n is properly chosen in accordance to the (unknown) spatial dependence in the latent field of labels, and thus the optimal value of n to be the one minimizing η^K. Indeed, values of n smaller than optimal induce unstable (along replicates) weak classifiers, since many elements are expected to cross boundaries between regions associated to different labels. Values of n larger than optimal induce unstable weak classifiers as well, since they are affected by the high variability of the representatives.

One might guess that spatial entropy is a good criterion also for the selection of the most proper value for K, since we expect the final classification to be less uncertain also for an optimal choice of K. Indeed, the simulation studies described in Secchi *et al.* (2012) clearly point out that the entropy criterion generally leads to a choice for K more parsimonious than necessary. Indeed, the problem of the choice of an optimal K is a well–known issue in cluster analysis, and a general strategy to tackle it has not yet been proposed. A possible approach, which we will adopt in the application described in Section 3, is based on the analysis of the following index associated to the final classification:

$$\theta = \frac{tr(S_B)}{tr(S_B + S_W)}, \tag{3}$$

where S_B and S_W are the final *between* and *within* cluster sum of squares matrix, respectively.

3 A case study: clustering irradiance data

We now illustrate an application of our classification algorithm to irradiance data to investigate the possible exploitation of solar energy in different areas of the planet. In particular, we try to identify areas of the planet which are optimal with respect to the positioning of solar power collectors by considering parameters, which depend on direct insolation, suited for sizing batteries or other energy-storage systems. Stor-

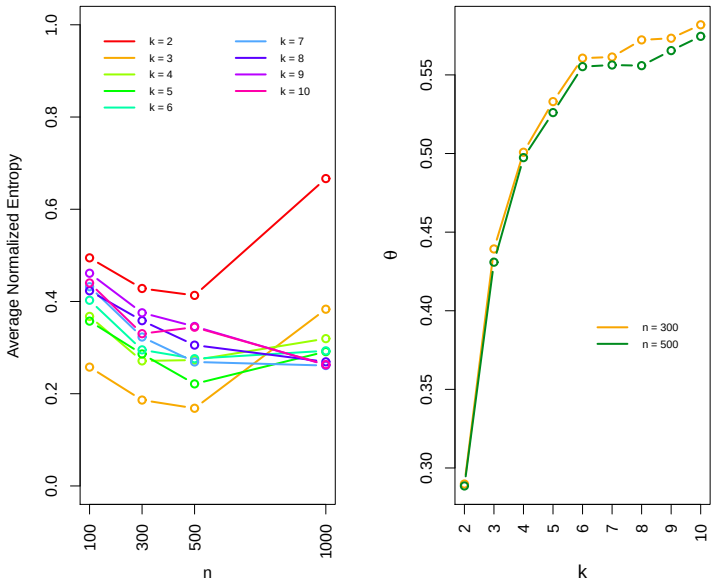

Fig. 1 Results of Bagging Voronoi-classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*: in the left panel, average normalized entropy obtained with different choices of K and for $n = 100, 300, 500, 1000$. In the right panel, values of the index θ introduced in (3) associated to the final classification with $K = 2, \ldots, 10$, and for $n = 300$ and $n = 500$.

age devices must indeed be designed to withstand continuous below–average conditions in various regions of the globe. More precisely, we analyze the maximum deficit below average value of solar radiation 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 NASA 2010 for details). From an engineering point of view, this quantity is considered as a proxy of the buffer extra-capacity that is needed to be installed in order to fulfill the possible gaps in energy supply provided by solar power plants. These gaps, in a particular site at a particular time of the year, can be due to unfavorable environmental conditions. From now on, we will name this quantity *buffer capacity*.

Rough data consist of vectors in $\mathbb{R}^{12}$ indexed by the sites of a spatial lattice. In each site, the 12 measures correspond to the values of the monthly maximum energy deficit with respect to the monthly average. Both the maximum and the average values are computed over the 22 years time period from July 1983 to 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 $Z_1 = \{-180, -179, ..., 178, 179\}$ and $Z_2 = \{-66, -65, ..., 65\}$: 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 47520 worldwide

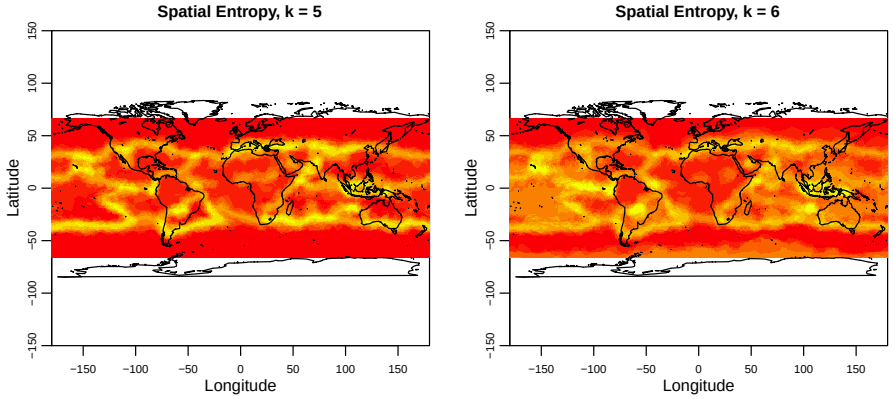

Fig. 2 Results of Bagging Voronoi-classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*: normalized spatial entropy maps associated to the classification with $K = 5$ (left) and $K = 6$ (right). Colors from red to white correspond to values from 0 to 1; higher values identify areas where classification is more uncertain.

non–polar districts. In each site of the lattice, we observe the buffer capacity $Y^{\nu}_{\lambda,\theta}$ for each month $\nu = 1,\ldots,12$. For each site, we obtain a functional datum $Y_{\lambda,\theta}(t)$ by smoothing $\{Y^1_{\lambda,\theta},\ldots,Y^{12}_{\lambda,\theta}\}$ with a Gaussian kernel with bandwidth equal to 1.5: the collection of these functional data, indexed by the sites of S_0, is the input of the Bagging Voronoi-classifiers algorithm.

For this application, we fix the number of bootstrap replicates to $B = 100$ and we test different values for the number n of elements of the Voronoi tessellation and the number K of clusters initializing the clustering algorithm. The n elements are drawn from a uniform distribution on S, the surface of the sphere of diameter equal to the earth. The set of nuclei for the Voronoi tessellation is then chosen by selecting the n sites among those 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, since they explain a proportion of total variance that exceeds 95%. Finally, for clustering the n representatives we use K-means with the L^2 semi–metric induced by the principal components.

In Figure 1, for different values of n and K, the performance of the Bagging Voronoi-classifiers algorithm is evaluated both in terms of average normalized entropy (i.e., sharpness of the image) and in terms of the index θ defined in (3) (i.e., differences among clusters). In the left panel of Figure 1 the values of the average normalized entropy are reported. The first fact to be noticed is that for most values of K, $n = 500$ provides a good choice to obtain a neat image. For small values of K, $n = 500$ is actually a minimum over the tested values, and it is hence chosen for setting the algorithm. Secondly, given $n = 500$, one can see that, in terms of classification sharpness, good values of K seem those between 3 and 7. In particular, two local minima are observed for $K = 3$ and $K = 5$.

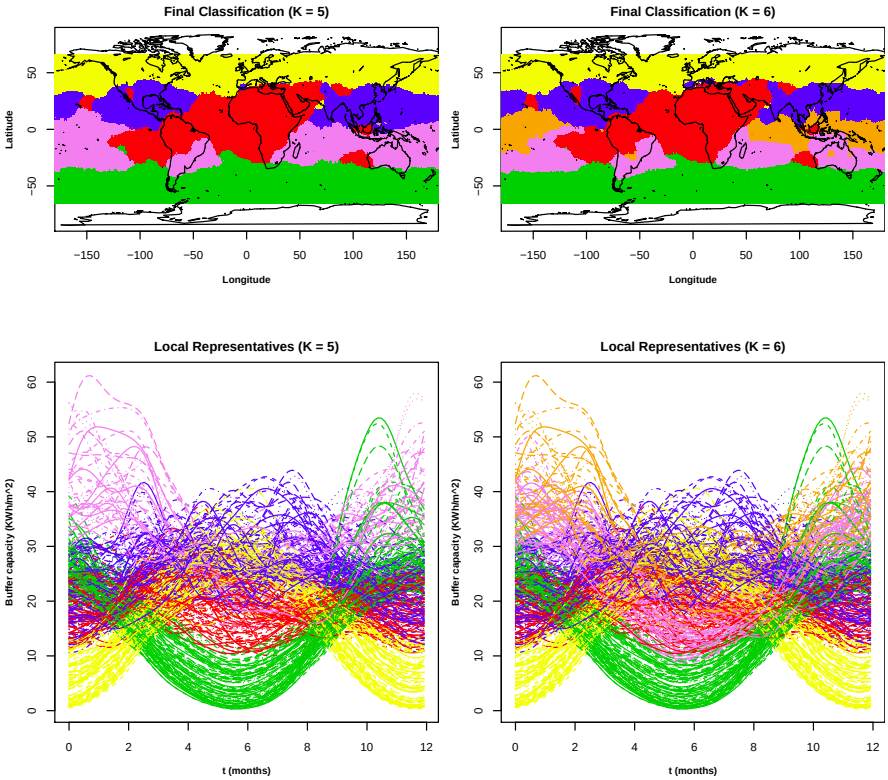

Fig. 3 Results of Bagging Voronoi-classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*. In the top panels, final classification maps obtained via a majority vote on frequencies of assignment, and by setting $K = 5$ (left) or $K = 6$ (right). In the bottom-left panel, a set of functional local representatives obtained with $n = 500$ in one of the iterations of the algorithm and clustered with $K = 5$; the colors for the plot are chosen coherently with the final classification map in the above panel. In the bottom-right panel, the same set of functional local representatives is clustered with $K = 6$, and colored coherently with the final classification map in the above panel.

To further investigate the choice of K, in the right panel of Figure 1 the values of the index θ defined in (3) are reported as a function of K for two values of n. Notice that the shape of the graph is robust with respect to the dimension of the Voronoi tessellation. The plot suggests $K = 6$ as the maximum reasonable number of clusters. Greater values of K are not paid off by a significant improvement in the description of data. Slightly smaller values for K ($K = 4, 5$) seem admissible as well, even though minor features are probably lost. Values $K = 2, 3$ are definitely not suggested. On the whole, $K = 5$ and $K = 6$ seem to be good choices for obtaining a spatially neat classification, with important differences among clusters. This is confirmed by inspection of Figure 2, where the two maps of spatial normalized entropy obtained for $K = 5$ and $K = 6$ are reported (left and right panel, respectively). Both

plots show a neat classification, even though the one corresponding to $K = 5$ seems more reliable. In Figure 3 (top panels), the final clusters obtained with $K = 5$ and $K = 6$ are reported on the Earth surface. The two classifications do not contradict each other; on the contrary, the latter is a refinement of the former supporting the robustness of the obtained classification.

In particular, for $K = 5$, the Bagging Voronoi-classifiers algorithm identifies different homogeneous macro-areas which – prima facie – seem interpretable in terms of the observed phenomenon. A climatological analysis, which is beyond the scopes of this paper, could of course 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 Figure 3 (left panels). In the bottom-left panel of the picture, a sample of local representatives is shown, each representative colored with a label corresponding to the macro-area it belongs to (Figure 3, top-left panel). The red cluster is characterized by a non-seasonal pattern, and by intermediate average buffer capacity along the year. It covers Africa, Middle-East and equatorial America and its presence is not explained 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 five clusters are present, the north-high and south-high clusters are absent in Europe and Africa, and the red cluster is almost absent in Asia.

The main difference obtained by choosing $K = 6$ is the fact that a new cluster, which mostly spurts from the former south-high (violet) cluster, appears along the equator (see Figure 3, top-right panel). This cluster, depicted in orange, is characterized by a very high seasonality of the buffer capacity (see Figure 3, bottom-right panel) that makes it strongly unsuited for electricity production by solar power. All other clusters remain unaffected while moving from $K = 5$ to $K = 6$, in particular the red one. Interestingly, from an engineering point of view, the red cluster is the one that shows an annual buffer capacity pattern which is optimal in terms of electricity production by solar power: it needs the minimal buffer capacity installation (the maximal annual need for energy is the lowest among the five detected patterns), associated to a constant reliability along the year.

References

1. Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24:123–140.
2. Delicado, P., Giraldo, R., Comas, C., Mateu, J., 2010. Statistics for spatial functional data: some recent contributions. *Environmetrics.* 21:224–239.
3. NASA 2010. *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].
4. Penrose, M. D., 2007. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli.* 13(4):1124–1150.

5. 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.
6. 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.
7. Secchi, P., Vantini, S., Vitelli, V., 2011. Spatial Clustering of Functional Data, in *Recent Advances in Functional Data Analysis and Related Topics*, *Contributions to Statistics*, Springer Physica-Verlag, pp. 283–290.
8. Secchi, P., Vantini, S., Vitelli, V., 2012. Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*. doi: 10.1016/j.jag.2012.03.006..
9. Tarpey, T., Kinateder, K. K. J., 2003. Clustering functional data. *J. Classification*. 20:93–114.