# Estimation of geometric anisotropy for clustered data collected from environmental sensor networks

D. T. Hristopulos[a,*] G. Spiliopoulos[a] M. P. Petrakis[a]

A. Chorti[b]

[a] *Geostatistics Research Unit*

*Department of Mineral Resources Engineering*

*Technical University of Crete*

*Chania 73100, Greece*

[b] *Computer Communications Department, Middlesex University, The Burroughs,*

*London, NW4 4BT, UK*

## Abstract

This paper addresses the estimation of geometric anisotropy parameters from scattered spatial data that are obtained from environmental surveillance networks. Estimates of geometric anisotropy improve the accuracy of spatial interpolation procedures that aim to generate smooth maps for visualization of the data and for decision making purposes. The anisotropy parameters involve the orientation angle of the principal anisotropy axes and the anisotropy ratio (i.e., the ratio of the principal correlation lengths). The approach that we employ is based on the covariance Hessian identity (CHI) method (Hristopulos, 2002; Chorti and Hristopulos, 2008), which links the expectation of the first-order sample derivatives tensor with the Hessian matrix of the covariance function (Swerling, 1962). In addition, we use image processing methods for the segmentation of the data set into clusters. The clusters are defined based on sampling density variations and extreme values. Both real and synthetic data are used to investigate the impact of CHI anisotropy estimation on the results of spatial interpolation with ordinary kriging.

*Key words:* segmentation, clustering, spatial interpolation, clustered CHI, non-parametric

## 1 Introduction

Geoinformatics is a science that aims at the development of information science infrastructures for handling problems in geosciences, as well as geotechnical, environmental and mining engineering. It involves the development and management of data structures and databases, the utilization of networking and communication technologies for the transfer of the data, as well as the development and application of statistical methodologies for the processing of the data. Terrestrial environmental monitoring networks involve irregular distribution of measurement stations in space. The local density and characteristics of the networks are influenced by various factors, including national environmental policies, terrain topography and proximity to urban centers. However, in order to visualize the information provided by the network and to generate risk maps that can be used for decision making, smooth maps of the monitored process are required. To generate smooth maps, a spatial model is needed for

* Corresponding author.
*Email addresses:* dionisi@mred.tuc.gr (D. T. Hristopulos), spiliopoulos@gmail.com (G. Spiliopoulos), petrakis@mred.tuc.gr (M. P. Petrakis), ersi.chorti@gmail.com (A. Chorti).
*URL:* http://www.mred.tuc.gr/home/hristopoulos/dionisi.htm (D. T. Hristopulos).

interpolation of the measurements on regular map grids. Remote sensing measurements are also affected by the problem of missing data (Rossi et al., 1994; Foster and Evans, 2008). Geostatistical methods are important for the statistical processing of the information provided by environmental networks, since they permit characterization and quantification of spatial dependencies from scattered data, and subsequently their use for spatial interpolation and map generation.

An accurate spatial model should, among other things, incorporate estimates of anisotropy. In addition, if the model is to be used with minimum used involvement (e.g., in an automatic setting) such estimates should involve a small number of free parameters. Such estimation methods would allow users that are not expert geostatisticians to obtain working estimates of anisotropy. Anisotropy appears in spatial data under at least two forms: trend anisotropy or statistical anisotropy. Trend anisotropy is easily determined from the coefficients of the spatial regression model. Here we focus on the estimation of statistical anisotropy, and in particular geometric anisotropy. Geometric anisotropy implies that the covariance function of the monitored process has different correlation lengths in different spatial directions. In two spatial dimensions, geometric anisotropy is determined from two parameters: the orientation angle of the principal axes and the ratio of the correlation lengths along the principal directions.

Estimation of the anisotropy parameters is typically based on empirical methods such as fitting of experimental directional variograms or on the rigorous but computationally demanding maximum likelihood method. The recently proposed *Covariance Hessian Identity (CHI)* or Covariance Tensor Identity (CTI) method (Hristopulos, 2002; Chorti and Hristopulos, 2008) provides a computationally fast, non-parametric approach for the estimation of the anisotropy parameters. The term non-parametric here refers to the fact that it is not necessary to assume a parametric covariance model in order to determine the anisotropy ratio and orientation angle. The caveat is that CHI assumes a differentiable covariance model. A deterministic interpolation scheme (e.g., bilinear or bicubic, minimum curvature, Savitsky-Golay polynomial filtering) is then used to approximate the sample derivatives of scattered data as described in (Chorti and Hristopulos, 2008). Note that the anisotropy estimation grid is not necessarily the same as the target map grid. The differentiability assumption is not a significant restriction, even though only the Gaussian covariance model is differentiable from the covariance functions traditionally used in geostatistics. In fact, both the Matérn and the Spartan (Hristopulos, 2003; Hristopulos and Elogne, 2007) models provide families of covariance functions with controllable differentiability properties.

The remainder of this manuscript is structured as follows: Section 2 briefly describes how the data used in the case study are obtained. While the clustered CHI method, presented afterwards, aims to provide a general tool for anisotropy analysis, the presentation of the method significantly benefits by focusing on a specific example. Section 3 proposes an approach for the application of CHI to scattered data that require segmentation into different clusters due to sampling density variations or to the presence of extreme values. In section 4 we use the clustered CHI method to determine the anisotropy of the radioactivity data set. We also compare the performance of spatial interpolation with and without anisotropic correction using cross validation measures. Finally, in Section 5 we present our conclusions. This paper extends the scope of the CHI method for anisotropy parameter estimation by combining it with image segmentation techniques for clustering and by developing methods for anisotropy averaging over different clusters.

## 2 Description of the Data

The data set involves scattered measurements of gamma dose rates (GDR) over part of Europe. The GDR data were generated by the German Federal Office for Radiation Protection (Bundesamt für Strahlenschutz) in the framework of the INTAMAP project (INTAMAP, 2009) and are described in Deliverable 5.4. The sampling network is represented by the sites of the European Radiological Exchange Platform (EURDEP). $N = 3626$ sampling sites are used with their positions expressed in the INSPIRE coordinate system (INSPIRE, 2009). GDR is measured in nanoSievert per hour (nSv/h). The network involves both densely sampled areas (e.g., Germany and Austria) and sparsely sampled ones (e.g. in South Europe).

Real background radioactivity measurements are combined with simulations that include systematic errors, local peaks due to to washout effects caused by heavy rainfall, single peaks due to lighting strikes, and areas of extreme values resulting from the dispersion of a radioactive plume caused by a simulated reactor accident in central Europe. The simulations are generated with the RODOS system (Ehrhardt, 1997) using meteorological information from the German weather service. The time of the accident was 23:40 on January 6, 2008. Forecasts of the plume dispersion were produced at +18h, +30h, +42h, and +54h from the starting time, for an area of $2500 \times 2500$ km$^2$ centered at the city of Offenbach. We used the +42h time slice for the spatial analysis. The statistics of the data set are given in Table 1.

Table 1
Statistics of the radioactivity data set used in the case study. The abbreviations used are as follows: $X_{\min}$ (minimum value), $q_1$ (first quartile), $q_2$ (median), $m_{\mathrm{X}}$ (mean value) $q_3$ (third quartile), $X_{\max}$ (maximum), $\sigma_{\mathrm{X}}$ (standard deviation), $\mu_{\mathrm{X}}$ (skewness), $k_{\mathrm{X}}$ (kurtosis).

| $X_{\min}$ | $q_1$ | $q_2$ | $m_{\mathrm{X}}$ | $q_3$ | $X_{\max}$ | $\sigma_{\mathrm{X}}$ | $\mu_{\mathrm{X}}$ | $k_{\mathrm{X}}$ |
|---|---|---|---|---|---|---|---|---|
| 29.0 | 85.8 | 131.0 | 2442.0 | 3082.0 | 26990.0 | 4371.36 | 2.29 | 5.25 |

## 3 The Clustered CHI method for Anisotropy Estimation

Consider an environmental sensor network (e.g., radioactivity probes) containing $N$ sampling points $\mathbf{s}_i = (x_i, y_i)$, $i = 1, \ldots, N$, where $(x_i, y_i)$ are the spatial coordinates expressed in an equidistant projection system. The sampled process to be mapped is denoted by $X(\mathbf{s})$. We will assume that $X(\mathbf{s})$ is modeled by a spatial random field the realizations of which admit at least first-order partial derivatives in two orthogonal directions. For jointly Gaussian spatial random fields, this class includes fields with Gaussian covariance, or Matérn covariance with smoothness index $\nu > 1$, or Spartan random fields with finite spectral cutoff $k_c < \infty$. The CHI method assumes that the data are generated from a stationary and normal (jointly Gaussian) or log-normal random field. Mild deviations from normality can be handled using the Box-Cox transform. Then, the anisotropy parameters can be estimated for the normalized field.

Often, the stationarity assumption is not supported by the data. Here, we focus on deviations from stationarity due to extreme values. For example, an accident releasing radioactivity in the environment generates a radioactive plume whose statistical properties differ markedly from the background radioactivity. In order to justify using the stationarity assumption, it is necessary to consider separately subsets of the sampling network that contain a large number (e.g., $N_g > 50$) of extreme values compared to the background.

Thus, separate *domains* are defined that contain the “normal” and “extreme” values respectively. Further, if the sampling density varies significantly over a domain, the domain is partitioned into clusters of *similar sampling density (SSD)* using standard segmentation methods from image processing (Gonzalez and Woods, 2006). We define a different anisotropy estimation grid for each cluster by tuning the grid step to the cluster sampling density, as discussed in Section 3.1 below.

### *3.1 Segmentation of the Data Set*

The segmentation procedure aims to divide the network of sensor points into different groups based on three criteria: the first criterion removes all the isolated and distant points from the sample. The second criterion requires that areas containing a number of clustered “extreme values” be treated as a separate group. Possibly, there may be more than one such groups. The third criterion requires that the points in each group be separated into clusters based on the local sampling density values.

#### *3.1.1 Filtering of isolated and distant points*

This step ensures that values which are not correlated with other points are excluded from the geostatistical analysis. To implement this criterion, a rectangular box centered at the network's centroid is defined. The extent of the box in the directions $x$ and $y$ is set to $\pm 4\sigma_x$ and $\pm 4\sigma_y$ where $\sigma_x, \sigma_y$ are the standard deviations of the sample's coordinate locations. Points that lie outside the boundary box and do not have a neighbour within a radius equal to $\min(\sigma_x, \sigma_y)$ are removed. For the case study data, this criterion has the effect of removing mainly values from sensors on remote island locations, e.g., at the Azores or stations in former European colonies. The application of this filtering process is illustrated in Fig. 1.

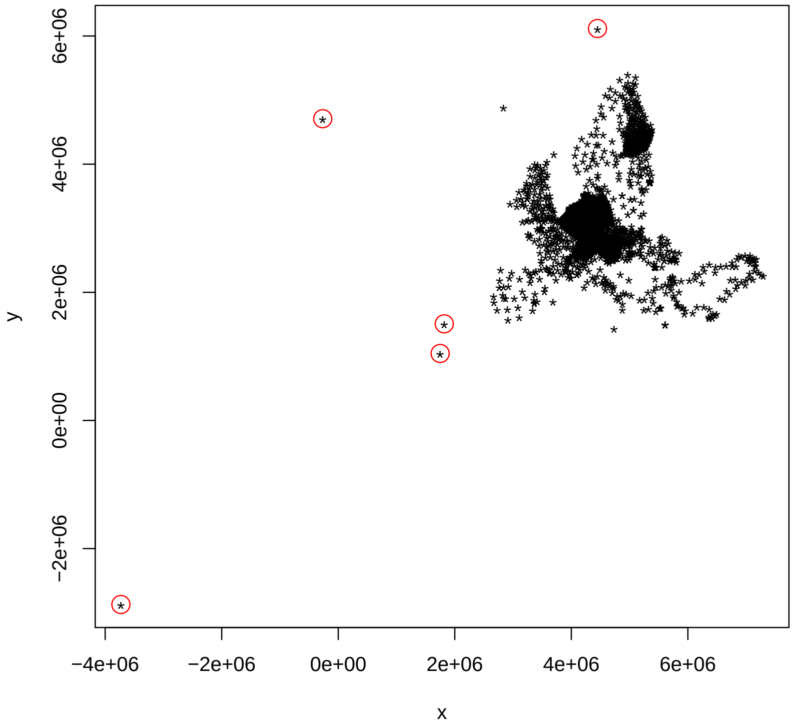


Fig. 1. Map of the sensor grid network. The distant and isolated points, as identified by the filtering algorithm, are marked by circles.

### *3.1.2 Segregation into "normal" and "extreme" value domains*

The second criterion defines a group of extreme values, henceforth called G2, that exceed a process specific threshold. For the present case study, the threshold is set at 250 nSv/h. The "extreme value" group includes the locations affected by the spreading plume of a simulated radioactive release. The remaining sensor points form a group that contains "normal" values, henceforth referred to as G1. In the case study, G1 involves points registering background radioactivity levels, as well as points with instrument malfunctions, spikes generated by lightning, etc. (some such events are also included in G2). G1 contains around 2500 points. The third criterion is necessary in order to construct meaningful anisotropy estimation grids, with a step that adapts to the sampling density in each area. We show the result of the segregation procedure on the pattern obtained at +42h after the simulated release in Fig. 2. Depending on the nature of the problem at hand, this segregation procedure may be modified. For example, if it is known that the fluctuations should follow the Gaussian distribution, one can filter out spikes due to instrument malfunctions using the iterative algorithm proposed in (Hristopulos et al., 2007).

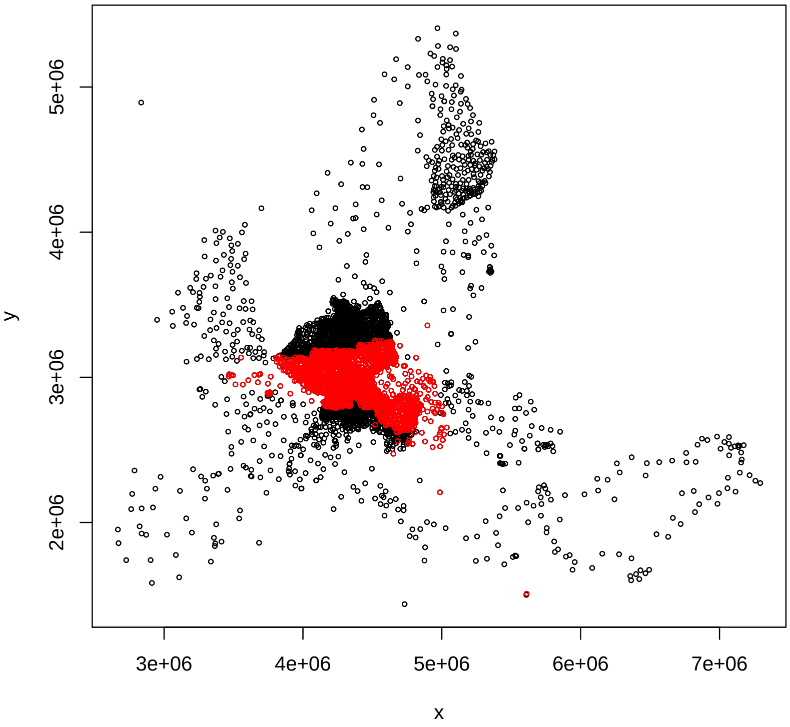


Fig. 2. Segregation of the truncated sensor network into “normal value” (black) and “extreme value” (red) domains.

### *3.1.3 Domain partitioning into SSD clusters*

Clustering of the network sensors based solely on the coordinates $x_i$, $y_i$ can be performed using various standard methods. Such choices include the Mixture of Gaussians (McLachlan and Peel, 2000), which is based on the probability density functions (PDF) of the $x_i$ and $y_i$, the method of k-means (Gan et al., 2007) which clusters the points according to the distances of $x_i$ and $y_i$ from (iteratively defined) cluster centers, Support Vector Machines (Cristianini and Shawe-Taylor, 2000), the method of k-median (Gan et al., 2007), Maximum Entropy classifiers, and Orthogonal Forward Regression with leave-one-out test score. In the case of the radioactivity monitoring network that we consider, different sensors may report at different times, so the network grid varies with time. We need to be able to cluster the sensors dynamically, without a-priory information on the number of clusters. It also makes sense to define clusters based not only on their geographical location, but also on the local sampling density, which may vary across national borders. Hence, we follow a four-step procedure. (1) We construct a sampling density function for the sensor network. (2) We use edge detectors from image processing to identify closed perimeters. (3) We identify different clusters and the sensor points that they include. (4) We reject regions with very few sensors, and then assign the remaining sensor locations to clusters according to an empirical criterion.

**3.1.3.1 Sampling density function** To implement the first step, we define a sampling density grid (SDG), which is in general different than the map grid. The SDG has the same number, $L = \lfloor \sqrt{N} \rfloor$, nodes per side and covers the sensor network area. For example, for the data set studied, $L = 50$ for the SDG in the domain G1. The SDG contains $L^2$ rectangular cells. We then form a sampling density matrix (SDM), the value of which at each grid cell is proportional to the number of sensor points enclosed by the cell. Each sensor point is assigned the sampling density value of the corresponding SDG cell. The spatial variation of the SDM in the domain G1 is shown in Fig. 3. The gap between the two main peaks in the center is due to the fact that the points in G2 are excluded.

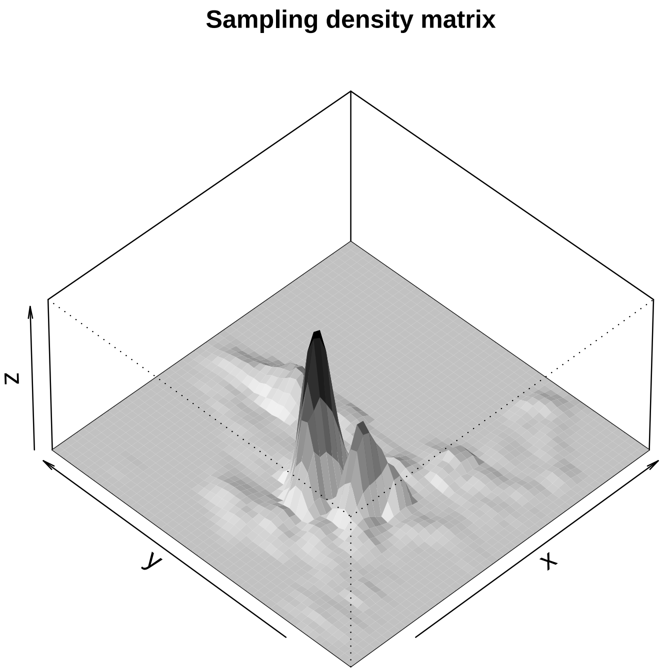


Fig. 3. Map of the sampling density matrix over the "normal value" domain G1.

**3.1.3.2 Edge detection** Next, we determine potential cluster perimeters based on the spatial variability of the SDM. Edge detection techniques are used to determine the cluster perimeters (Gonzalez and Woods, 2006, Chap. 7). The SDM is smoothed by an averaging 3 × 3 filter, and an 5 × 5 edge detection logarithmic filter is passed over the grid to detect likely cluster perimeters.

After identifying the candidate "edge" cells, closed perimeters are determined by searching for sequentially linked edge cells. A cell is considered "linked" if it possesses a neighbour inside a 3 × 3 neighbourhood centered at the cell's location.

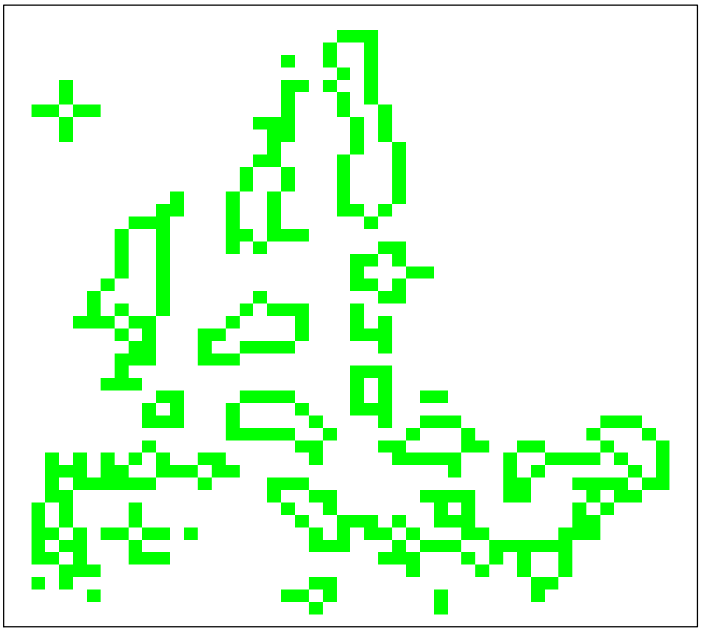


Fig. 4. Identification of potential cluster perimeters for the "normal value" domain (G1). The figure displays the cells of the SDG that lie on the identified perimeters.

**3.1.3.3 Initial SSD cluster identification** After all cells have been searched, each closed perimeter is labeled and considered as a potential cluster perimeter. The sensor points are then assigned to the cluster perimeter that contains them, thus leading to an initial cluster assignment. Some points are not assigned to clusters at this stage. The initial assignment of sampling points to the enclosing cluster perimeters is shown in Fig. 5. Note that the cluster perimeters are defined on nodes of the SDM grid, while the sampling sites do not in general coincide with the nodes of the grid.

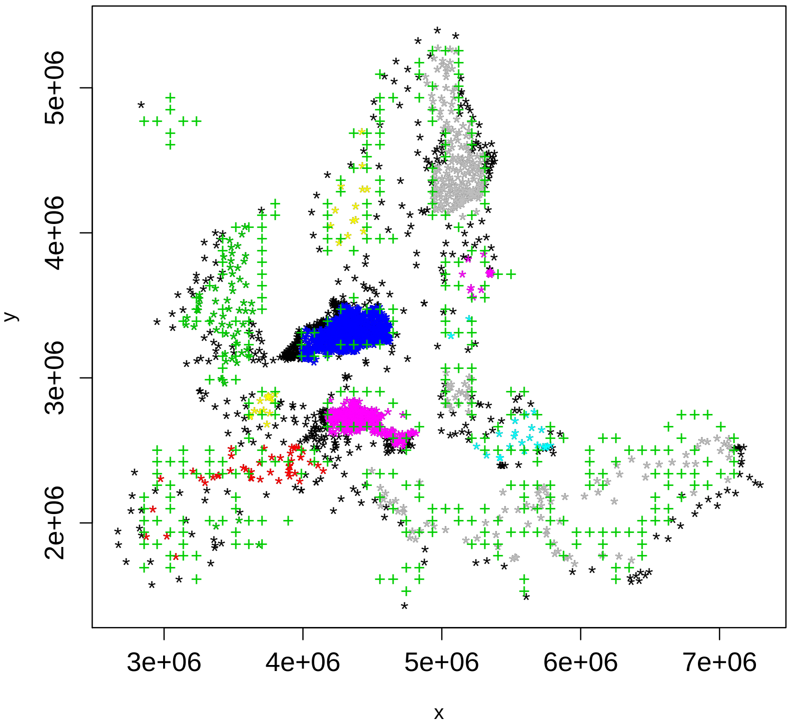


Fig. 5. Initial assignment of sampling sites (stars) in the “normal value” domain G1 to cluster perimeters. The centers of the SDG perimeter cells are denoted by cross marks. Colored sampling points have been assigned at this stage to clusters, while points marked by black circles are unassigned.

**3.1.3.4 Final SSD cluster assignment** Based on our experience, meaningful SSD clusters for CHI anisotropy detection should contain at least 50 sensor points. Smaller clusters are rejected, and the sampling points inside them, as well as unassigned sensor points, are assigned to a neighbouring, sufficiently populated cluster. The assignment of rejected and unassigned points is performed by optimizing a cost function that weighs SDM differences between the sensor point and the three closest neighbour clusters as well as physical distances between the sensor point and the centroids of the clusters.

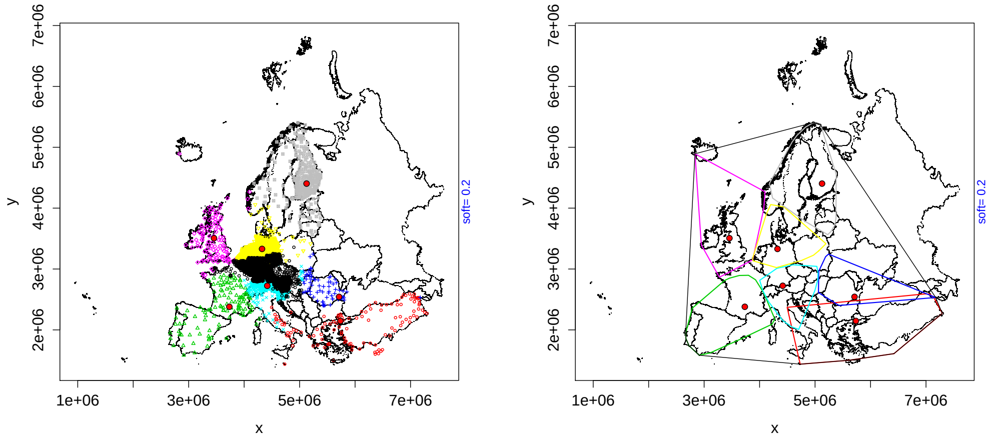
The cost function for assigning point $\mathbf{s}_i$ to the cluster $c$ is given by

$$\phi_{i;c} = \frac{d_{i,c}}{\max\{d_{i,1}, \ldots, d_{i,K}\}} + \beta \frac{\sqrt{\left|\rho_i - \langle\rho\rangle_c\right|}}{\sqrt{\max\{\langle\rho\rangle_1, \ldots, \langle\rho\rangle_K\}}}, \; c = 1, \ldots, K. \qquad (1)$$

In the above, $d_{i,c}$ is Euclidean distance of the point $\mathbf{s}_i$ from the centroid of the cluster $c$, $K$ is the total number of clusters, $\rho_i$ is the sampling density assigned to $\mathbf{s}_i$ based on the (SDM) sample density matrix, and $\langle\rho\rangle_c$ is the average value of the SDM in cluster $c$. The coefficient $\beta$ is empirically determined. For the case studies that we have considered $\beta = 0.2$ works well.

The scheme presented above ensures that points near a specific cluster's centroid are preferably assigned to that cluster, while points that are equally far from all three clusters are assigned to the cluster that has a similar sampling density. All sensor sites are finally assigned to an SSD cluster that includes more than 50 sensor points, as shown in Fig. 6(a). The convex hulls of the final clusters are shown in Fig. 6(b). The stars inside the convex hulls represent the centroids of the initial clusters (before the reassignment). Note that the cluster around Iceland has been rejected based on the minimum number of points criterion, and its stations have been incorporated with the UK cluster.

We also note that the cost function (1) can lead to non-intuitive assignment of sensor points in the case of elongated cluster shapes. For example, consider two neighbouring clusters: one in containing the points marked as blue crosses (mostly Poland) and the other containing the points marked as red circles (Southeast Europe and Turkey) in Fig. 6(a). The first cluster contains one point in Eastern Turkey that is clearly closer to the perimeter of the second cluster. By taking a look at Fig. 5, one can distinguish a number of unassigned sampling points along the eastern border of Turkey. These points are assigned to clusters in the final SSD assignment stage. Due to the elongated shape of the second cluster, the distance of the point in question from the centroid of the first cluster is smaller than its distance from the centroid of the second cluster. Such inconsistencies can be removed if the cost function is based on the distance between each unassigned point and its nearest neighbour that has been assigned to a non-rejected cluster in the initial SSD cluster identification. However, this option leads to an increase of the computational time, due to the calculations of near neighbour structures. Thus, we opted for the simpler approach which is based on the distance from the cluster centroid. This approach seems to provide overall reasonable cluster structures, and its speed makes it suitable for automatic mapping purposes.



(a) Final cluster structure.

(b) Convex hulls of clusters.

Fig. 6. Final assignment of sampling points in G1 to SSD clusters (a). Cluster ownership is color coded. The "black" area in central Europe represents the domain G2. Convex hulls of the final SSD clusters (b). The centroids of the initial clusters are marked by stars.

## *3.2 Anisotropy estimation*

### *3.2.1 SSD Cluster estimates*

The estimates of the anisotropy parameters $(R, \theta)$ in each cluster are based on the CHI method (Chorti and Hristopulos, 2008). The angle $\theta$ represents the angle between one of the principal axes, arbitrarily called $M_1$, and the horizontal axis of the coordinate system. In geography, it is preferred to define the anisotropy orientation in terms of the complementary angle. $R = \xi_1/\xi_2$ is the ratio of the correlation lengths along $M_1$ and its orthogonal direction $M_2$. If $\hat{Q}_{i,j}$ are sample-based estimates of the slope tensor of $X(\mathbf{s})$, and $q_{\text{diag}} = \frac{\hat{Q}_{22}}{\hat{Q}_{11}}$, $q_{\text{off}} = \frac{\hat{Q}_{12}}{\hat{Q}_{11}}$ represent the diagonal and off-diagonal ratios, respectively, $\hat{R}$ and $\hat{\theta}$ are given by

$$\hat{\theta} = \frac{1}{2} \tan^{-1} \left( \frac{2q_{\text{off}}}{1 - q_{\text{diag}}} \right), \quad \hat{R}^2 = 1 + \frac{1 - q_{\text{diag}}}{q_{\text{diag}} - (1 + q_{\text{diag}}) \cos^2 \theta}. \tag{2}$$

Equations (2) are valid if the random field is Gaussian or lognormal, second-order stationary, and differentiable.

In each SSD cluster, the $q_{\text{diag}}$ and $q_{\text{off}}$ are estimated by means of finite differences on the *rectangular anisotropy estimation grid* that covers the cluster domain. The grid extends from $x_{c;\min}$ to $x_{c;\max}$ in the $x-$direction and from $y_{c;\min}$ to $y_{c;\max}$ in the $y-$direction, where $x_{c;\min} = \min(x_1, \ldots, x_{n_c})$, $x_{\max} = \max(x_1, \ldots, x_{n_c})$, are respectively the smallest and largest values of the $x$ coordinates for all points in the cluster $c$ (similarly for $y$). To avoid introducing bias related to the shell shape, the grid cells are squares with side length equal to

$$\alpha_c = \min(|x_{c;\min} - x_{c;\max}|, |y_{c;\min} - y_{c;\max}|)/\sqrt{N}.$$

The interpolation is conducted using a non-parametric, deterministic approach, such as triangle-based linear or minimum curvature interpolation. This introduces some bias in the anisotropy estimation, since the interpolation model does not account for the anisotropy. However, the field generated on the anisotropy estimation grid incorporates the anisotropy properties imparted by the data. A detailed analysis of the impact of the interpolation method on the anisotropy estimation is given by Chorti and Hristopulos (2008).

### *3.2.2 Coarse-grained domain estimates*

Using individual cluster estimates of anisotropy to interpolate $X(\mathbf{s})$ on the map grid would require a smoothing filter, e.g., moving windows. Alternatively, one can seek an average estimate of the anisotropy in the domains of "normal" and "extreme" values. Given the nonlinearity of the expressions in (2), a simple average of the cluster anisotropy parameters is not appropriate. Let us assume that each domain involves $K_g$ clusters ($g = 1, 2$), and that $\hat{Q}_{ij}^{g;c}$, $i, j = 1, 2$, $c = 1, \ldots, K_g$ represent the estimates of the slope tensor for the $c$-th cluster in the $g$-th domain. Anisotropy estimates are based on the *weighted average*, $\overline{Q_{ij}^{g}}$, of the slope tensor:

$$\overline{Q_{ij}^{g}} = \frac{\sum_{c=1}^{K_g} w_{g;c}\, \hat{Q}_{ij}^{g;c}}{\sum_{c=1}^{K_g} w_{g;c}} \tag{3}$$

The weights $w_{g;c}$ are set equal to the area $A_k$ enclosed by the convex hull of each cluster.

### *3.3 A Statistical Test for Isotropy*

The anisotropy parameter estimates given by (2) are sample statistics and thus exhibit sample-to-sample fluctuations. If the estimates indicate significant anisotropy, this should be taken into account in the interpolation procedure used to generate the map of the process. There are various ways in which $(\hat{R}, \hat{\theta})$ can be used. If the estimate of the spatial structure is based on the experimental variogram, $(\hat{R}, \hat{\theta})$ can be used to rotate and rescale the coordinate system to render the spatial dependence isotropic (Hristopulos, 2002; Chorti and Hristopulos, 2008). Then, the omnidirectional variogram can be estimated and modeled. Spatial interpolation should be performed in the transformed coordinate system using the optimal isotropic variogram model. In this case the transformed values of the map grid coordinates should be used. Alternatively, $(\hat{R}, \hat{\theta})$ can be used as initial guesses of the anisotropy parameters in a maximum likelihood optimization procedure.

In both cases, it is advantageous to know whether the anisotropy of the data is significant in order to incorporate it in the map grid interpolation. To this end, a non-parametric joint probability density function has been developed and its confidence regions have been calculated (Petrakis and Hristopulos, 2009). These can be used to test (a) if two sets of anisotropy parameters are statistically different and (b) if the isotropy assumption can be rejected at a given confidence level. We use the isotropy test to determine if it is necessary to perform an isotropy restoring transformation (rotation and rescaling) of the coordinates. For nearly isotropic data, this helps to reduce the computing time of map generation without significant impact on the accuracy of the interpolation. More specifically, the isotropy hypothesis can not be rejected if

$$\hat{R}^2 \in \left( \frac{n_c - 2\sqrt{(n_c - r_\alpha) r_\alpha}}{n_c - 2r_\alpha}, \frac{n_c + 2\sqrt{(n_c - r_\alpha) r_\alpha}}{n_c - 2r_\alpha} \right), \tag{4}$$

where $r_\alpha$ is a constant that depends on the desired confidence level: for a 95% confidence level $r_\alpha \simeq 6$. In (4) $n_c \geq 50$ is the number of sampling points implicated in the estimates: in the case of a single domain and a single cluster $n_c = N = 1$, while for a single domain with multiple clusters $\sum_{c=1}^{K} n_c = N$. The test is conservative (as shown by theoretical arguments and numerical simulations), leading to wider confidence intervals than the true ones, due to the underestimation of correlation effects. The accuracy of the test is compromised for small data sets or sparsely sampled areas, due to poor estimation of the anisotropy parameters in such cases.

## 4 Cross validation analysis of anisotropy estimates

### *4.1 Study design*

We have conducted tests on single-cluster synthetic data and densely sampled real data (not shown herein), which show that application of CHI improves interpolation performance. In the context of the radioactivity case study, we test the potential benefits of anisotropy estimation for mapping using a cross validation approach. To generate cross validation measures 60 *training set* realizations from the domains G1 and G2 are used. Each training set contains 2/3 of the total number of points, and the sampled points are replaced at the end of each run. The remaining 1/3 of the points are used for validation purposes.

### *4.2 Spatial model parameter estimation*

The reference prediction values are obtained using an isotropic variogram model, estimated from the empirical variogram of the training set by means of a weighted least squares (WLS) fit. To include anisotropy, $(R, \theta)$ are estimated for the training set points in the domains G1 and G2. Bilinear interpolation, implemented by means of the `akima` package, is used to obtain estimates of the GDR on the anisotropy estimation grids. The GDR values thus obtained for G1 and G2 are shown in Fig. 7.

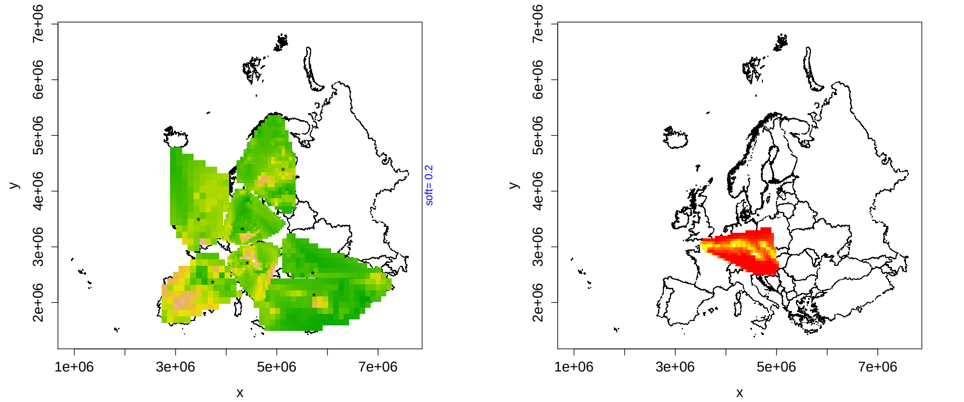
Since G1 contains seven clusters, the anisotropy estimates are based on the cluster average of the slope tensor as given by Eq. (3). For each training set realization we test based on (4) if the isotropic hypothesis is supported, and then an isotropy restoring coordinate transformation is used. Next, the range and sill of the variogram are estimated in the isotropic coordinate system using the R function `automap` (Hiemstra et al., 2009). For each training set the optimal variogram is selected from among the exponential, Gaussian, spherical and Matérn models. The estimates $(\hat{R}, \hat{\theta})$ are then incorporated to obtain the anisotropic variogram model.

### *4.3 Spatial interpolation and cross validation*

The method of ordinary kriging (OK) is used for interpolation using the `gstat` package (Pebesma, 2004). Validation measures compare the estimates with the “true” values at the validation locations. The validation measures are obtained by calculating first the spatial average over the validation set followed by an average over the realizations, e.g.,

$$ME = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_v}\sum_{i=1}^{N_v}\left[\hat{X}(\mathbf{s}_i^j) - X(\mathbf{s}_i^j)\right],$$

where $M = 60$, $N_v$ is the number of points in the validation set, $\hat{X}(\mathbf{s}_i^j)$ denotes the ordinary kriging prediction at $\mathbf{s}_i^j$, and the latter represents the $i-$th sampling point in the $j-$th realization.

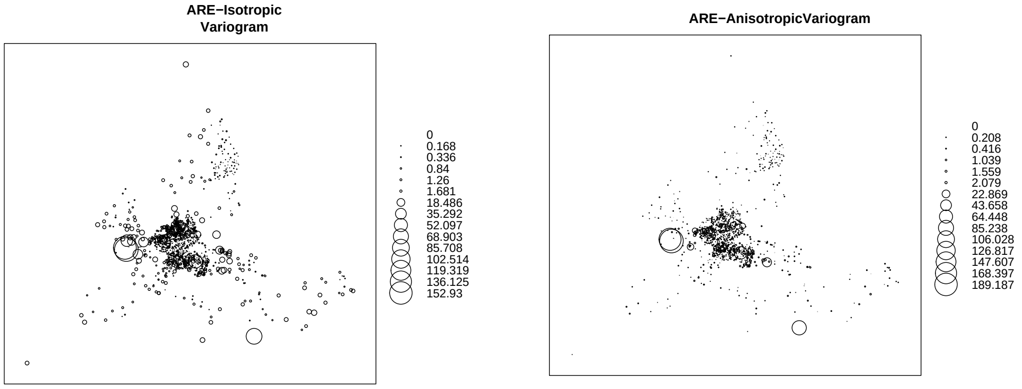


(a) Interpolated GDR on anisotropy estimation grid in G1.

(b) Interpolated GDR on anisotropy estimation grid G2.

Fig. 7. Interpolated GDR fields used in the clustered CHI anisotropy estimation. The range of values in G1 is 29.0 – 248.8 nSv/h, while in G2 it is 251.0 – 26992.5 nSv/h.

The cross validation results are reported in Table 2. The first row is obtained using isotropic variogram models. The second row is obtained by estimating the anisotropy parameters, performing an isotropy restoring transformation (rotation and rescaling of coordinate axes), and then determining the variogram model. It is shown that incorporation of the anisotropy improves the validation measures overall, except for the linear correlation coefficient. Of course, to an extent the improvement is due to the fact that our anisotropy method handles separately the two domains G1 and G2. Note that reported values of the mean absolute relative error are large. This is due mainly to overestimates of the dose rate in areas that are close to the plume but not in the plume; the local absolute relative error is given by $\left[\hat{X}(\mathbf{s}_i^j) - X(\mathbf{s}_i^j)\right]/X(\mathbf{s}_i^j)$, and this becomes large if $\hat{X}(\mathbf{s}_i^j) >> X(\mathbf{s}_i^j)$ . In the isotropic model, this is caused by prediction points outside the plume which are influenced by neighboring sampling sites inside the plume. In the anisotropic case, this is due to the false assignment of predictions points to G2 instead of G1. A typical situation is illustrated in Fig. 8.



(a) ARE using the isotropic variogram model.

(b) ARE using the anisotropic variogram model.

Fig. 8. Local average relative error (ARE) for one realization of the training and validation sets.

### *4.4 Computation time*

The computation time is 32 sec for ordinary kriging without anisotropy correction and 17 sec for ordinary kriging with the anisotropy correction. The difference in computational time is due to numerical complexity of ordinary kriging, which scales as the third power of the number of sampling points, and the fact that in the first case kriging uses all the sampling points. In the second case, the training set is split into two domains (G1 and G2). There is an additional cost for assigning the prediction points to either G1 or G2, but interpolation in each domain now involves a smaller number of points. Each prediction point is assigned to a group based on its nearest neighbour in the training set. Nearest neighbours are determined using the computationally efficient kd tree structures, which have a numerical complexity of $N \log_2 N$. These were implemented using the `ann` function from the R package `yaImpute`. The simulations ran on an Intel Core2 Duo CPU with 2Gb RAM, under the Ubuntu 8.10 Operating System.

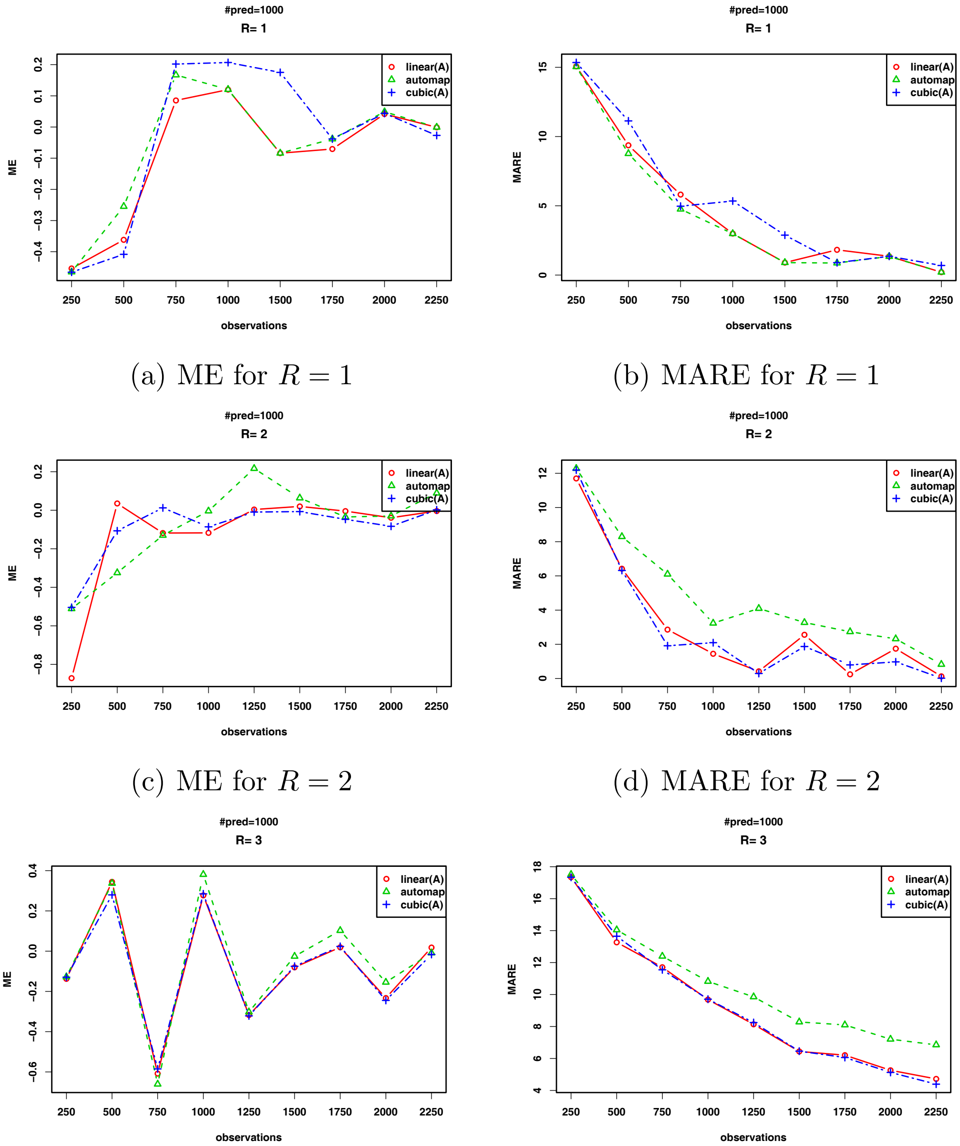
Table 2
Validation measures obtained by taking the mean over 60 realizations of spatially averaged statistics over the validation set. The figures are rounded to the second decimal place. ME: Mean error. MAE: Mean absolute error. MARE: Mean absolute relative error. MRSE: Mean root square error. MRSRE: Mean root square relative error. R: linear correlation coefficient.

| | ME | MAE | MARE (%) | MRSE | MRSRE | R |
|---|---|---|---|---|---|---|
| | (nSv/h) | (nSv/h) | | (nSv/h) | (%) | |
| Isotropic variogram | $-13.12$ | 600.95 | 130 | 1428.94 | 5.84 | 0.95 |
| Anisotropic variogram | $-4.55$ | 538.46 | 77 | 1402.35 | 5.76 | 0.95 |

### *4.5 Synthetic data*

To further investigate the benefits of incorporating anisotropy estimates in the interpolation procedure we apply CTI to synthetic data. We generate realizations from a Gaussian SRF with mean $m_X = 98$, $\sigma_X = 20$ and Gaussian covariance on a square $256 \times 256$ grid. We use $R = 1, 2, 3$ with minimum correlation length $\xi_1 = 12$ and orientation angles randomly generated in $[-45°, +45°]$. For all $R$ we generate 20 realizations of the SRF, we randomly sample a subset of the grid nodes $N$, and then predict the field using ordinary kriging at 1000 validation locations. In the application of CHI we interpolate the field on the anisotropy estimation grid using both bilinear and bicubic interpolation. A single domain with a single cluster is obtained in this synthetic example.

The dependence of the validation measures ME and MARE as a function of $N$ is shown in Fig. 4.5 for three cases: (i) isotropic variogram (ii) anisotropic variogram using bilinear spline interpolation on the anisotropy estimation grid (iii) anisotropic variogram using bicubic spline interpolation on the anisotropy estimation grid. We find no systematic dependence of the bias on the interpolation method used. The MARE tends to decrease with increasing $N$ for all three methods. For $R = 1$, as expected, there is no systematic difference between the outcomes of the three methods. Incorporating the anisotropy estimates for the case $R = 2$ tends at first to decrease the MARE with increasing $N$ in comparison to the isotropic case. The three values seem to converge again for the higher values of $N$. There is practically no difference between the results obtained with the bilinear and bicubic interpolation on the anisotropy grid. For $R = 3$ the difference between the MARE obtained with and without anisotropy estimation is more systematic and it continues to increase for all $N$ considered in this study. In conclusion, this study on synthetic data shows that there is a benefit in interpolation performance by using anisotropy estimation versus an isotropic variogram model. At the same time, the bilinear interpolation on the anisotropy grid seems to perform just as well as the bicubic interpolation.



(a) ME for $R = 1$

(b) MARE for $R = 1$

(c) ME for $R = 2$

(d) MARE for $R = 2$

(e) ME for $R = 3$

(f) MARE for $R = 3$

Fig. 9. Cross validation measures for $R = 1, 2, 3$ (i) using anisotropy parameter estimates obtained with bilinear (circles) and bicubic (crosses) interpolation versus (ii) using an isotropic variogram model (triangles). The horizontal axis measures the number of points $N$ used in the sample.

## 5 Conclusions

This paper introduces the clustered CHI method for the estimation of geometric anisotropy parameters from scattered data in two spatial dimensions. This approach resonates with the development of sound statistical methods for the processing of spatial data, which is one of the pillars of geoinformatics. In particular, from the perspective of environmental surveillance, it is necessary to develop computationally efficient methods that can provide near real-time warnings for developing environmental threats. The proposed method incorporates the computational efficiency of the single-cluster CHI method with a segmentation procedure. The latter partitions the study area into domains according to exceedance of threshold vales, and its domain into clusters based on variations of local sampling density and proximity to established cluster centres. The interpolation performance of the clustered CHI method depends on the sampling density, the presence or lack of stationarity, and the differentiability of the mapped process. Increasing the sampling density, the "degree of stationarity" and the differentiability of the sampled process lead to more accurate estimates of the anisotropy.

In the case study investigated above, we illustrate the clustered CHI method by application to a "difficult" data set which involves deviations from Gaussianity due to several factors and significant variations of the sampling density across the study area. Clustered CHI leads to improved interpolation validation measures compared to estimates that are based on the isotropic variogram hypothesis. It is also shown that the method is computationally fast, requiring only a fraction of a second to determine anisotropy parameters on a data set with as many as $\approx 2000$ data points and two domains, one of which contains several clusters. The R codes that implement clustered CHI are part of the `Intamap` and `IntamapInteractive` R packages, which can be downloaded from `http://sourceforge.net/projects/intamap/develop`.

## Acknowledgment

This work is funded by the European Commission, under the 6th FP, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and not necessarily of the European Commission. We thank the scientists Ulrich Stoehlker, Martin Bleher, and Sven Burbeck of the German Federal Bureau for Radiation Protection (Bundesamt für Strahlenschutz) for providing the data used in the case study.

## References

Chorti, A., Hristopulos, D. T., 2008. Non-parametric identification of anisotropic (elliptic) correlations in spatially distributed data sets. IEEE Transactions on Signal Processing 56 (10), 4738–4751.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge.

Ehrhardt, J., 1997. The RODOS system: Decision support for off-site emergency management in Europe. Radiation Protecition Dosimetry 73 (1-4), 35–40.

Foster, M. P., Evans, A. N., 2008. An evaluation of interpolation techniques for reconstructing ionospheric TEC maps. IEEE Transactions on Geoscience and Remote Sensing 46 (7), 2153–2164.

Gan, G., Ma, C., Wu, J., 2007. Data Clustering Theory, Algorithms and Applications. SIAM.

Gonzalez, R. C., Woods, R. E., 2006. Digital Image Processing (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Hiemstra, P., Pebesma, E., Twenhöfel, C., Heuvelink, G., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers and Geosciences 35 (8), 1711–1721.

Hristopulos, D. T., 2002. New anisotropic covariance models and estimation of anisotropic parameters based on the covariance tensor identity. Stochastic Environmental Research and Risk Assessment 16 (1), 43–62.

Hristopulos, D. T., 2003. Spartan Gibbs random field models for geostatistical applications. SIAM Journal of Scientific Computing 24 (6), 2125–2162.

Hristopulos, D. T., Elogne, S., 2007. Analytic properties and covariance functions of a new class of generalized Gibbs random fields. IEEE Transactions on Information Theory 53 (12), 4667–4679.

Hristopulos, D. T., Mertikas, S., Arhontakis, I., Brownjohn, J., 2007. Using gps for monitoring tall-building response to wind loading: filtering of abrupt changes and low-frequency noise, variography and spectral analysis of displacements. GPS Solutions 11 (2), 85–95.

INSPIRE, 2009. Infrastructure for Spatial Information in the European Community.
URL `http://inspire.jrc.ec.europa.eu/`

INTAMAP, 2009. Interoperability and Automated Mapping.
URL `http://www.intamap.org/`

McLachlan, G., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers and Geosciences 30, 683–691.

Petrakis, M., Hristopulos, D. T., 2009. On the joint probability density function of geometric anisotropy statistics for two dimensional differentiable random fields and a non-parametric test of statistical isotropy. IEEE Transactions on Signal ProcessingTo be submitted.

Rossi, R., Dungan, J., Beck, L., 1994. Kriging in the shadows: Geostatistical interpolation for remote sensing. Remote Sensing of Environment 49, 32–40.

Swerling, P., 1962. Statistical properties of the contours of random surfaces. IRE Transactions on Information Theory, 315–321.