Non-parametric estimation of geometric anisotropy from environmental sensor network measurements

Dionissios T. Hristopulos[1], M. P. Petrakis[1], G. Spiliopoulos[1], Arsenia Chorti[2]

[1]Technical University of Crete, Geostatistics Research Unit, Chania 73100, Greece.
dionisi@mred.tuc.gr

[2]Computer Communications Department, Middlesex University, The Burroughs, London, NW4 4BT, UK.
ersi.chorti@gmail.com

Abstract. This paper addresses the estimation of geometric anisotropy parameters from scattered data in two dimensional spaces. The parameters involve the orientation angle of the principal anisotropy axes and the anisotropy ratio (i.e., the ratio of the principal correlation lengths). The mathematical background is based on the covariance Hessian identity (CHI) method developed in [3, 1]. CHI links the expectation of the first-order sample derivatives tensor with the Hessian matrix of the covariance function [6]. The paper focuses on the application of CHI to samples that require segmentation into clusters, either due to sampling density variations or due to systematic changes in the process values. A non-parametric isotropy test is also presented. Finally, a composite (real and synthetic) data set is used to investigate the impact of CHI anisotropy estimation on spatial interpolation with ordinary kriging.

1 Introduction

Scattered samples of a physical process from an environmental sensor network are considered. To generate smooth maps of the process, a spatial model is needed that will be used for interpolation of the measurements on the regular map grid. The spatial model should incorporate estimates of the anisotropy. It is assumed that the process can be modeled in terms of a second-order stationary, Gaussian or lognormal spatial random field, at least within clusters (subsets of the sampling set). Hence, we focus on the estimation of statistical, geometric anisotropy. A method is formulated for classifying scattered data according to their position, sampling density, and the process values in order to estimate the anisotropy parameters.

2 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 expressed in an equidistant projection system. The sampled process to be mapped is denoted by $X(\mathbf{s})$. The covariance Hessian identity (CHI) method requires stationarity and normality or log-normality. To justify the use of stationarity, it is necessary to consider as a separate group subsets of the sampling network that contain a large (e.g., $N_g > 50$) number of extreme values (compared to the background). Thus, separate *stationarity domains* are defined that contain the “normal” and “extreme” values respectively. On-grid deterministic interpolation is used to approximate the sample derivatives of scattered data in the CHI method. If the sampling density varies significantly over a stationarity domain, the data are segregated into clusters of *similar sampling density (SSD)* and define a different interpolation grid for each cluster by tuning the grid step to the cluster sampling density.

2.1 Clustering

The first step of the clustering process removes sensor locations that are isolated and distant. A rectangular box centered at the network's centroid is defined. The extent of the box in directions x and y is $\pm 4\sigma_x$ and $\pm 4\sigma_y$ where σ_x, σ_y are the standard deviations of the coordinate locations. The 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. The stationarity domain of extreme values (G2) that exceed the threshold is then identified. The remaining points form the stationarity domain G1.

Different clusters are identified by constructing a sampling density matrix (SDM) on a regular grid (different than the map grid) that covers the sensor network. The sampling density grid (SDG) consists of N equal area cells. The sampling density of each 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. Image edge detection techniques are used to determine the cluster perimeters. The SDM is smoothed by an averaging 3×3 filter. Then, an 5×5 edge detection logarithmic filter is passed over the grid to detect likely cluster perimeters. Once the candidate cells have been determined, closed perimeters are identified by checking that there is a sequence of edge cells linked sequentially (i.e., each cell should have a neighbour inside a 3×3 neighbourhood centered at the cell's position). Each closed perimeter is labeled and considered as a cluster perimeter. The partitioning of sensor points inside cluster perimeters forms the initial cluster assignment. Some points are not assigned to clusters at this stage.

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 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 distances and the sampling density differences are normalized in the $(0, 1]$ interval. In the cost function, the point-cluster distances are weighed with the coefficient 0.9 and the sampling density differences by the coefficient 0.1. This scheme ensures that points near a specific cluster's perimeter 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.

2.2 Anisotropy estimation

The estimates of the anisotropy parameters (R, θ) in each cluster are based on the CHI method [1]. The angle θ represents the angle between one of the principal axes, arbitrarily called $M1$, and the horizontal axis of the coordinate system. $R = \xi_1/\xi_2$ is the ratio of the correlation lengths along $M1$ and its orthogonal direction $M2$. 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

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

Equations (1) are valid if the process is Gaussian or log-Gaussian, second-order stationary, and differentiable. The q_{diag} and q_{off} are estimated by finite differences on an interpolated square grid covering the domain. The interpolation is conducted using a non-parametric, deterministic approach (e.g., triangle-based linear interpolation or minimum curvature).

Anisotropy can be estimated separately for each SSD cluster. However, using individual cluster estimates 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 within the stationarity domains. Given the nonlinearity of the expressions in (1), simply averaging the anisotropy parameters of the clusters is not appropriate. Let us assume that each stationarity domain involves K_g clusters ($g = 1, 2$), and that $\hat{Q}^g_{ij;c}$, $c = 1, \ldots, K_g$ represents the estimate of slope tensor for the c-th cluster in the g-th domain. Anisotropy estimates are based on the *weighted average*, $\overline{Q^g_{ij}}$, of the slope tensor:

$$\overline{Q^g_{ij}} = \frac{\sum_{c=1}^{K_g} w_{g;c}\hat{Q}^g_{ij;c}}{\sum_{c=1}^{K_g} w_{g;c}} \tag{2}$$

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

2.3 A Statistical Test for Isotropy

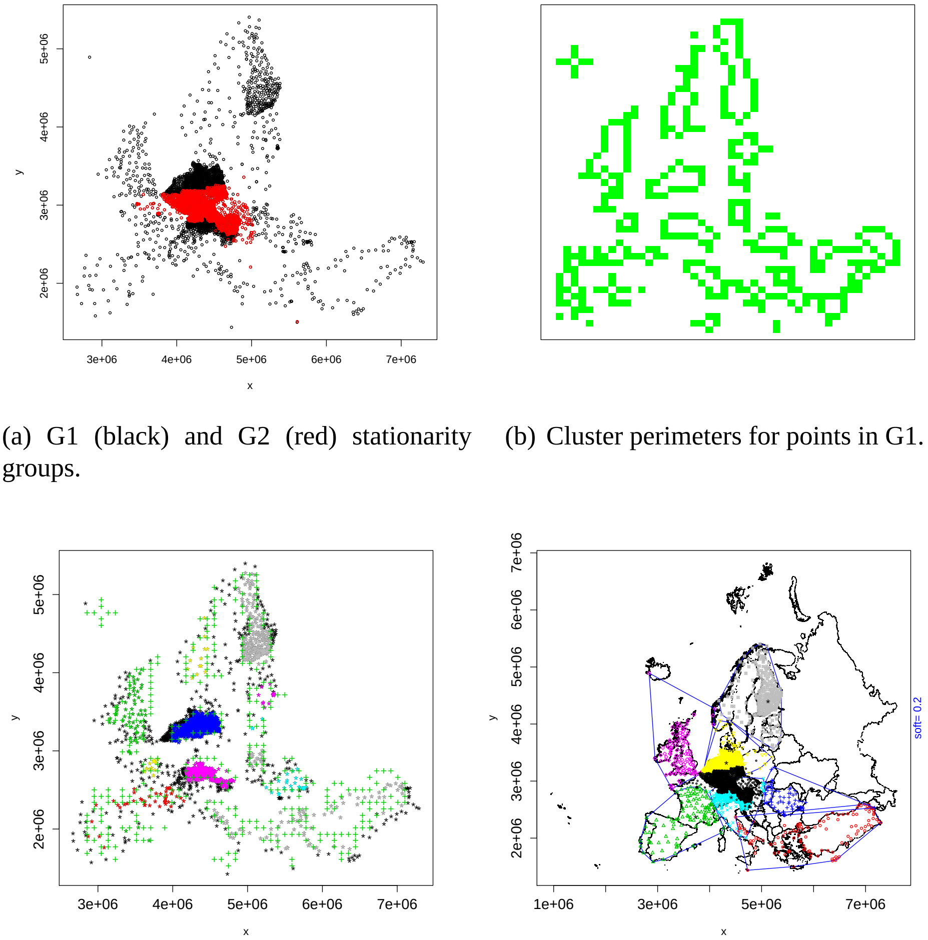
The anisotropy parameter estimates are statistics and exhibit sample-to-sample fluctuations. A non-parametric joint probability density function (jpdf) has been developed and its confidence regions have been calculated [5]. These can be used to test (a) if two sets of anisotropy parameters are statistically different and (b) if the isotropy assumption can be rejected at a given confidence level. The isotropy test is used to determine if it is necessary to perform an isotropy restoring transformation (rotation and rescaling) of the coordinates. This helps to reduce the computing time of map generation for isotropic data. The isotropy hypothesis can not be rejected if

$$\hat{R}^2 \in \left(\frac{N - 2\sqrt{(N-r_\alpha)r_\alpha}}{N-2r_\alpha}, \frac{N + 2\sqrt{(N-r_\alpha)r_\alpha}}{N-2r_\alpha}\right), \tag{3}$$

where r_α is a constant defined by the confidence level; for a 95% confidence level $r_\alpha \simeq 6$. 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.

3 Study Design and Results

To test the benefit of anisotropy estimation for mapping, a cross validation approach is employed: validation errors obtained by ordinary kriging (OK) with an isotropic variogram



(a) G1 (black) and G2 (red) stationarity groups.

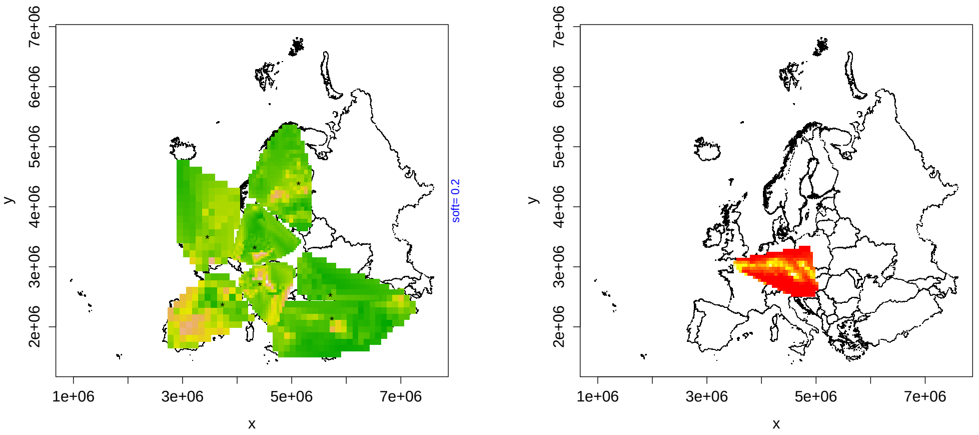
(b) Cluster perimeters for points in G1.

(c) Sampling sites inside cluster perimeters.

(d) Final assignment of points to clusters.

Figure 1: Segmentation of the data set into two groups of normal (G1) and extreme (G2) values after removal of distant points, and segmentation of G1. Crosses (+) denote edge points on the sampling density grid.

model are compared with those obtained after an isotropy restoring coordinate transformation which is based on $\hat{R}, \hat{\theta}$. 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. The impact of anisotropy estimation for a realistic and rather complicated data set is studied below. 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 (http://inspire.jrc.ec.europa.eu/). $X(\mathbf{s})$ represents gamma dose rates (GDR) measured in nSv/h (nanoSievert per hour). The network involves both densely sampled areas (e.g., Germany and Austria) and sparsely sampled ones (e.g. in South Europe). The GDR exercise data were generated by the German Federal Office for Radiation Protection (BfS) for workpackage 5.4 of INTAMAP (www.intamap.org). They combine real background radioactivity measurements with simulated effects that include systematic errors, as well as extreme values due to lighting strikes and the dispersion of a radioactive plume caused by a severe reactor accident in central Europe. The clustering process is illustrated by means of Fig. 1 for the entire data set. First, isolated stations remote to the European continent are removed. Fig. 1(a) shows the partitioning of the remaining points into two stationarity groups. In Fig. 1(b) the identified edges are demonstrated, while Fig. 1(c) displays the sampling points inside the cluster perimeters. Fig. 1(d) shows the final cluster assignments as well as the perimeters of the convex hulls.



(a) Linear interpolation grid in G1. Range of values $29.0 - 248.8$ nSv/h.

(b) Linear interpolation grid in G2. Range of values $251.0 - 26992.5$ nSv/h.

Figure 2: Interpolated fields used in the clustered CHI anisotropy estimation.

To calculate validation measures, 60 *training set* realizations from the remaining sites 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. To include anisotropy, (R, θ) are estimated for each stationarity group, using linear interpolation for finite differences estimation. The interpolated fields for G1 and G2 are shown in Fig. 2. Since G1 contains a number of clusters, the anisotropy estimates are based on Eq. (2). Then, if the isotropy test (3) does not support the isotropic hypothesis, 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` [2]. The estimates $(\hat{R}, \hat{\theta})$ are incorporated to obtain the anisotropic variogram. Finally, OK is applied using the `gstat` package [4]. For each training set the optimal variogram is selected from among the exponential, Gaussian, spherical and Matérn models. Validation measures compare the estimates with the "true" values at the prediction locations ($1/3$ of the points). These measures involve a spatial average over the prediction set followed by an average over the realizations. The results are reported in Table 1. The 1st row is obtained using isotropic variogram models. The 2nd row is obtained by estimating the anisotropy parameters, performing an isotropy restoring transformation (rotation and rescaling of coordinate axes), and then determining the variogram. Incorporation of anisotropy improves the validation measures. [1]

[1] The computation time is 32 sec for OK without and 17 sec for OK with the anisotropy correction. In the first case, OK uses all the sampling points. In the second case, the training set is split into G1 and G2. There is an additional cost for assigning prediction points to G1 or G2, based on the ownership group of their nearest neighbours in the training set. Nearest neighbours are efficiently determined using kd trees (function ann, yaImpute package). The code ran on an Intel Core2 Duo CPU with 2Gb RAM, using Ubuntu 8.10 OS.

Table 1: Average validation measures 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
With isotropic hypothesis	−13.12	600.95	1.30	1428.94	5.84	0.95
With isotropy correction	−4.55	538.46	0.77	1402.35	5.76	0.95

4 CONCLUSIONS

The clustered CHI method for estimating anisotropy parameters from scattered data sampled on irregular supports is presented. The performance of the method depends on the sampling density, the presence or lack of stationarity, and the differentiability of the mapped process. Application of the CHI method to a "difficult" data set leads to improved interpolation validation measures compared to the isotropic hypothesis.

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.

REFERENCES

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

[2] P.H. Hiemstra, E.J. Pebesma, C.J.W. Twenhöfel, and G.B.M. Heuvelink. Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. *Computers and Geosciences*, 2008. Accepted for publication.

[3] D. T. Hristopulos. 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, 2002.

[4] Edzer J. Pebesma. Multivariable geostatistics in S: the gstat package. *Computers and Geosciences*, 30:683–691, 2004.

[5] M. Petrakis and D. T. Hristopulos. A non-parametric test of statistical isotropy for differentiable spatial data random fields in two dimensions. *IEEE Transactions on Signal Processing*, 2009. to be submitted.

[6] P. Swerling. Statistical properties of the contours of random surfaces. *IRE Transactions on Information Theory*, pages 315–321, jul 1962.