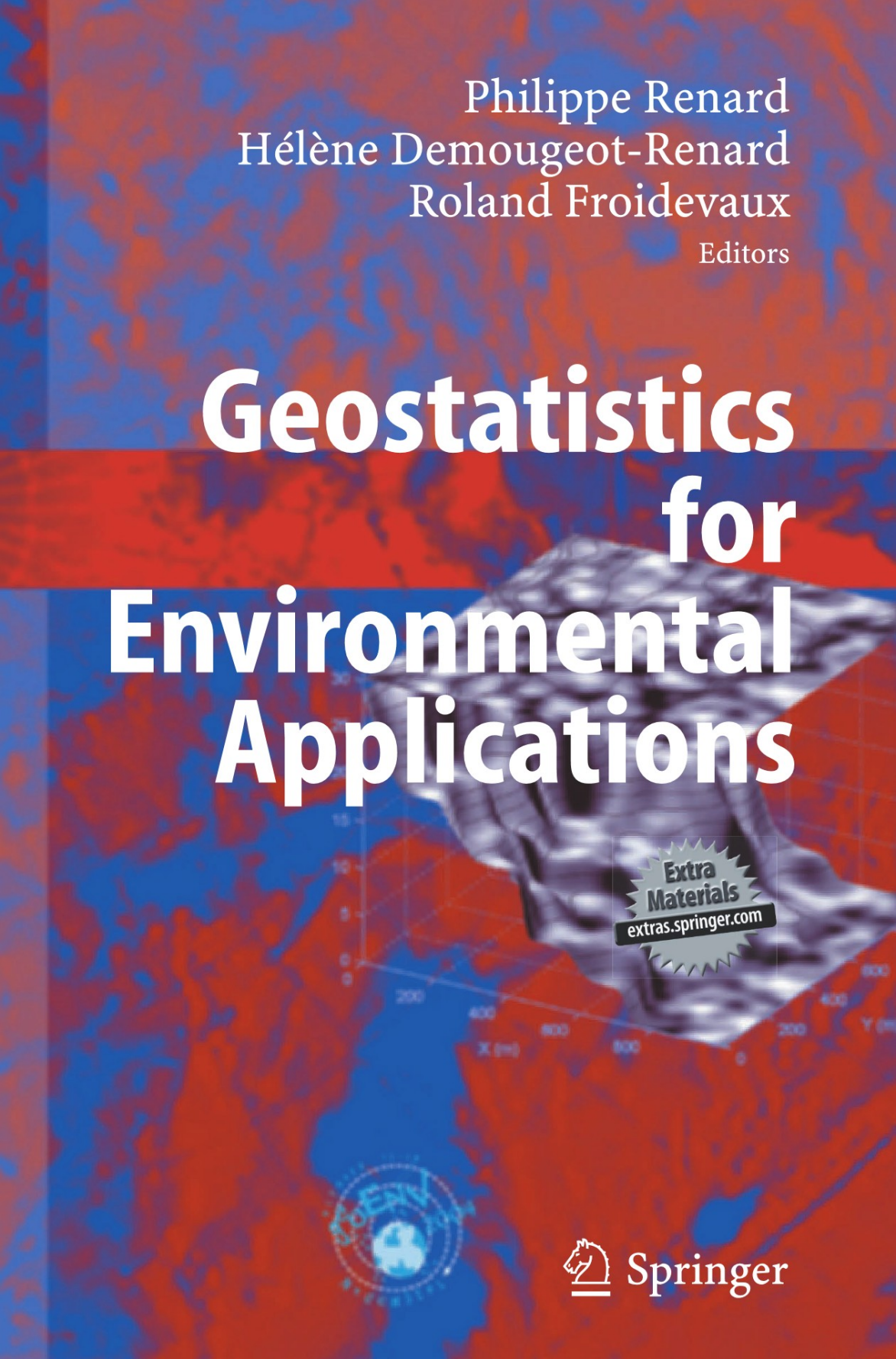
Philippe Renard
Hélène Demougeot-Renard
Roland Froidevaux
Editors

Geostatistics for Environmental Applications



Springer

Philippe Renard
Hélène Demougeot-Renard
Roland Froidevaux

Geostatistics for Environmental Applications

Proceedings of the Fifth European Conference on
Geostatistics for Environmental Applications

Philippe Renard
Hélène Demougeot-Renard
Roland Froidevaux
(Editors)

Geostatistics for Environmental Applications

Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications

With 218 Figures and a CD-ROM

Springer

Dr. Philippe Renard
University of Neuchâtel, Centre for Hydrogeology
11 Rue Emile Argand, 2007 Neuchâtel, Switzerland
E-mail: *philippe.renard@unine.ch*

Dr. Hélène Demougeot-Renard
FSS International
7 Chemin de Mont-Riant, 2000 Neuchâtel, Switzerland
E-mail: *demougeot.renard@fssintl.com*

Dr. Roland Froidevaux
FSS International
9 Rue Boissonnas, 1227 Geneva, Switzerland
E-mail: *roland.froidevaux@fssintl.com*

Library of Congress Control Number: 2005927607

ISBN-10 3-540-26533-3 **Springer Berlin Heidelberg New York**
ISBN-13 978-3-540-26533-7 **Springer Berlin Heidelberg New York**

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in The Netherlands

Cover design: Erich Kirchner
Production: Luisa Tonarelli
Typesetting: Camera-ready by the editors

Printed on acid-free paper 30/2132/LT – 5 4 3 2 1 0

Foreword

While the society becomes everyday more aware of environmental problems, the experts have to deal with a major issue: uncertainty due to incomplete data sets and spatio-temporal variability. Among the techniques used to quantify this uncertainty and to improve environmental management, geostatistics is becoming a recognized standard, applied in fields as different as hydrogeology, remote sensing, ecology or soil contamination. In recent years, the domain of application of these techniques has regularly grown together with the wide availability of Geographical Information Systems and geostatistical software packages.

This book is the outcome of the fifth edition of the European Conference on Geostatistics for Environmental Applications (geoENV V) held in Neuchâtel (Switzerland) from October 13th to October 15th, 2004. The conference attracted more than 100 participants, mostly from Europe, but also from North America, South America, North Africa, Russia and Australia. Among the 140 abstracts submitted to the conference, the organization committee selected 53 for oral presentation and 40 for poster presentation.

The book reflects the current status of the research in the field of geostatistics for environmental applications. It opens with one keynote paper by Carol Gotway-Crawford, senior researcher at the Center for Disease Control and Prevention, Atlanta, which emphasizes the problem of the size of the data support when making spatial statistics. It is then followed by 38 research papers, classified according to their main topics, that describe recent methodological advances and applications. All these papers have been presented orally during the conference and accepted by the reviewers. The final version of the papers were again checked by the editors. Also included in the book is a CDROM containing the original posters and the companion poster papers. This material has not been reviewed.

Finally, the editors wish to thank all the reviewers and the authors without whom this book could not exist, as well as the sponsors of the conference: the Swiss Federal Statistical Office (http://www.bfs.admin.ch), the Swiss Federal Office for Water and Geology (http://www.bwg.admin.ch/e/), the Swiss National Science Foundation (http://www.snf.ch/), the University of Neuchâtel (http://www.unine.ch), the Centre of Hydrogeology of the University of Neuchâtel (http://www.unine.ch/chyn), the Banque Cantonale Neuchâteloise (http://www.bcn.ch), and the NCCR Plant Survival (http://www.unine.ch/nccr/).

Neuchâtel, February 2005

The editors.

Philippe Renard
Hélène Demougeot-Renard
Roland Froidevaux

Organizing and scientific committees

Philippe Renard, *University of Neuchâtel, Switzerland (Chairman)*
Hélène Demougeot-Renard, *FSS International, Neuchâtel, Switzerland*
Roland Froidevaux, *FSS International, Geneva, Switzerland*
Denis Allard, *INRA, Avignon, France*
Jaime Gómez-Hernández, *UPV, Valencia, Spain*
Pascal Monestiez, *INRA, Avignon, France*
Xavier Sánchez-Vila, *UPC, Barcelona, Spain*
Amilcar Soares, *IST, Lisbon, Portugal*

The editors are grateful to the following persons for their work as referees:

Rachid Ababou, *IMFT, Toulouse, France*
Denis Allard, *INRA, Avignon, France*
Peter M. Atkinson, *University of Southampton, United Kingdom*
Ana Bio, *Instituto Superior Tecnico, Portugal*
Patrick Bogaert, *Université Catholique de Louvain, Belgium*
Josè Capilla, *Universidad Politécnica de Valencia, Spain*
Jesus Carrera, *Universidad Politécnica de Catalunya, Spain*
Eduardo Cassiraga, *Universidad Politécnica de Valencia, Spain*
Mario Chica-Olmo, *University of Granada, Spain*
Jean-Paul Chilès, *Ecole des Mines de Paris, France*
George Christakos, *University of North Carolina, USA*
Olaf Cirpka, *University of Stuttgart, Germany*
Noel Cressie, *Ohio State University, USA*
Hélène Demougeot-Renard, *FSS International, Neuchâtel, Switzerland*
Dimitri d'Or, *Université Catholique de Louvain, Belgique*
Souheil Ezzedine, *Lawrence Livermore National Laboratory, USA*
Chantal de Fouquet, *Ecole des Mines de Paris, France*
Luc Feyen, *Katholieke Universiteit Leuven, Belgium*
Aldo Fiori, *University of Roma Tre, Italy*
Roland Froidevaux, *FSS International, Geneva, Switzerland*
Michel Garcia, *FSS International, Paris, France*
Marc Genton , *North Carolina State University, USA*
Tilmand Gneiting, *University of Washington, USA*
Jaime Gómez-Hernández, *Universidad Politécnica de Valencia, Spain*
Pierre Goovaerts, *Biomedware, Ann Arbor, USA*
Carol Gotway-Crawford, *NCEH, Atlanta, USA*
Alberto Guadagnini, *Politecnico de Milano, Italy*
Laura Guadagnini, *Politecnico de Milano, Italy*
Gilles Guillot, *INRA, France*
Harri-Jan Hendricks-Franssen, *ETH Zürich, Switzerland*
Olivier Jaquet, *Colenco Power Consulting AG, Switzerland*
André Journel, *Stanford University, USA*

Phaedon Kiriakidis, *University of California, USA*
Denis Marcotte, *Ecole Polytechnique Montréal, Canada*
Peter Meier, *NOK Baden, Switzerland*
Pascal Monestiez, *INRA, France*
Carla Nunes, *Instituto Superior Tecnico, Portugal*
Maria João Pereira, *Instituto Superior Tecnico, Portugal*
Antonio Pulido-Bosch, *University of Almeria, Spain*
Monica Riva, *Politecnico de Milano, Italy*
Jacques Rivoirard, *Ecole des Mines de Paris, France*
Klaus-Jürgen Röhlig, *GRS, Germany*
Christian Roth, *Airparif, France*
Xavier Sánchez-Vila, *Universidad Politécnica de Catalunya, Spain*
Marc Serre, *University of North Carolina at Chapel Hill, USA*
Amilcar Soares, *Instituto Superior Tecnico, Portugal*
Mohan Srivastava, *FSS Canada, Canada*
Fritz Stauffer, *ETH Zürich, Switzerland*
Alfred Stein, *University of Wageningen, The Netherland*
Hans Wackernagel, *Ecole des Mines de Paris, France*
Richard Webster, *Rothamsted Experimental Station, United Kingdom*

Table of Contents

Groundwater

Soil contamination

Remote sensing

Environment

Contents of the CD-ROM

Change of support: an inter-disciplinary challenge

C. A. Gotway Crawford[1] and L. J. Young[2]

[1] Centers for Disease Control and Prevention, Atlanta, GA USA
[2] Department of Statistics, University of Florida, Gainesville, FL USA

1 An introduction to change of support in geostatistics

One of the fundamental ideas underlying the field of geostatistics is the concept of a *regularized* variable, the average value of Z(s) over a volume B

$$Z(B) = \frac{1}{|B|}\int_B Z(s)ds, \tag{1}$$

where $|B| = \int_B ds$ is the called the *support* of Z(B). The term support reflects the geometrical size, shape, and spatial orientation of the units or regions associated with the measurements (see e.g., Olea 1991). Changing the support of a variable (typically by averaging or aggregation) creates a new variable. This new variable is related to the original one, but has different statistical and spatial properties. Determining how these properties vary with support is called the *change of support problem*. From the beginning, the field of geostatistics has incorporated solutions to change of support problems (Matheron 1963).

The practical problems driving the initial development of geostatistics were those encountered in the mining industry, with a primary problem being the prediction of the average grade of a mining block from drill core samples. Thus, most change of support problems were concerned with "upscaling," the prediction of a variable whose support is larger than that of the observed data. A common example of this is block kriging where Z(B) is predicted from data $Z(s_1),\ldots, Z(s_n)$ that have mean $E[Z(s)]=\mu$ and semivariogram $\gamma(s-u)=1/2\mathrm{Var}[Z(s)-Z(u)]$. The block kriging predictor is given by $\hat{Z(B)} = \sum_{i=1}^{n}\lambda_i Z(s_i)$, where the weights $\{\lambda_i\}$ are obtained by solving the equations (Journel and Huijbregts 1978, Chilès and Delfiner 1999)

$$\sum_{k=1}^{n}\lambda_k\gamma(s_i - s_k) + m = \gamma(B, s_i), \quad i = 1,\ldots,n$$

$$\sum_{k=1}^{n}\lambda_k = 1.$$

Here $\gamma(B,s_i)=\frac{1}{|B|}\int_B \gamma(s_i-u)du$ and m is a Lagrange multiplier from the constrained minimization. The kriging variance is

$$\sigma_K^2(B)=\sum_{i=1}^{n}\lambda_i\gamma(B,s_i)-\gamma(B,B)+m,$$

where

$$\gamma(B,B)=\frac{1}{|B|^2}\int_B\int_B\gamma(s-u)dsdu.$$

There are many more sophisticated geostatistical solutions to this change of support problem, including nonlinear methods and those developed to infer the entire probability distribution of the regularized variable (see, e.g., Journel and Huijbregts 1978, Matheron 1984a and b, Cressie 1993b, Rivoirard 1994, Goovaerts 1997, and the compilations in Chilès and Delfiner 1999 and Gotway and Young 2002). However, most practical applications that use them have data of point support (or data measured on small cores or boreholes), and the inferential goal is *upscaling* by *regularization,* so that the inferential goal is prediction of Z(B) (or some function of it) in Eq. 1. Moreover, the volumes B of interest are rectangular blocks and so the integrations required can be done fairly easily and quickly. However, spatial data come in many forms. Instead of measurements associated with point locations, we could have measurements associated with lines, areal regions, surfaces, or volumes. In many disciplines such as geology and soil science, observations often pertain to rock bodies, stratigraphic units, soil maps, and large-scale land use. In geographic and public health studies, the data are often counts or rates obtained as aggregate measures over geopolitical regions such as census enumeration districts and postal code zones. Moreover, the inferential goal may also not be limited to upscaling. For example, modeling hydrological and soil processes often involves making predictions from models that have relatively coarse spatial resolution and these then need to be downscaled to the watershed level or combined with digital elevation data of point support. In many studies in public health, sociology, and political science, the data are counts or rates aggregated over areal regions (e.g., per postal code or per census unit), but individual-level inference is desired. Finally, the idea of regularization as defined through Eq. 1 is not always an appropriate mathematical description of either the data that are available or the inferential quantity of interest. For example, in geographical studies, the data are totals (e.g., the number of people per enumeration district) or rates that are based on population totals and not on the area of the regions. Developing valid inferential methods for upscaling, downscaling and "side-scaling" (in the case of overlapping spatial units) variables is of critical importance to numerous scientific disciplines. It seems natural to try to extend the relatively rich ideology on change of support developed in geostatistics to more general change of support problems.

In this context, we examine the change of support problem from an interdisciplinary point of view. This viewpoint allows us to extract some key ideas, common themes, and general statistical issues common to change of support prob-

lems. We provide a brief summary of the various types of solutions that have been proposed to various change of support problems over more than five decades of research conducted in numerous fields of study. The goal of this extroverted contemplation is the search for a general framework for statistical solutions to change of support problems.

2 Why is support important?

Changing the support of a variable through regularization creates a new variable with different statistical and spatial properties. In particular, the variability in $Z(B)$ decreases as the support B increases, the histogram of $Z(B_1)$, …,$Z(B_m)$, $m<n$ will tend to be more symmetric and approximately bell-shaped, and the spatial autocorrelation in the regularized values is altered as well (Journal and Huijbregts 1978, Armstrong 1999). Thus, any inferential procedure must take these factors into consideration. There are numerous examples of this *support effect* in the geostatistical literature, and many methods have been suggested for adjusting for support effects in spatial prediction and resource estimation.

While this view of support has served the mining industry quite well, the situation is more complex in other disciplines. Global Positioning Systems, remote sensing technology, and Geographic Information Systems (GIS) allow greater access to a variety of spatial data and easily permit analysis on almost a limitless choice of spatial units: points, postal code polygons, Census tracts and enumeration districts, hydrogeologic regions, raster images with different pixel sizes, and even regions defined by the whim of the user. More often than not, the data of interest in any one analysis are of different supports that are irregularly shaped. Another factor, related to support, comes into play here: the concept of *scale*. From our review work in this area, we have found that the term is used differently in different disciplines. In fact, few good definitions exist. For example, Bierkens *et al.* (2000) use the terms scale and support interchangeably, defining scale to be support. We argue that they while these two concepts are very much related, they are in fact quite different. From our perspective, spatial scale is defined by both the number and the relative size of the spatial units used to partition a spatial domain of interest. Corresponding to every spatial scale is a level of spatial aggregation that represents the particular mixture of sub-units that comprise the larger units of interest. For a fixed domain, increasing the scale results in a fewer number of larger units. Since size is one aspect of support, clearly support and scale are related. However, we prefer the more general definition of support that includes the shape and orientation and the units. It is possible to partition two spatial domains into subunits with the subunits being of essentially the same size in both partitions, but of different shapes and/or different orientations (Fig. 1).

Geographers have long encountered the interplay between support and scale, noting that the choice of spatial units for analysis is “modifiable,” and that statistical results depend heavily on the way the spatial units are created. In geography,

the change of support problem is known as the *Modifiable Areal Unit Problem* (MAUP) (Openshaw and Taylor 1979).

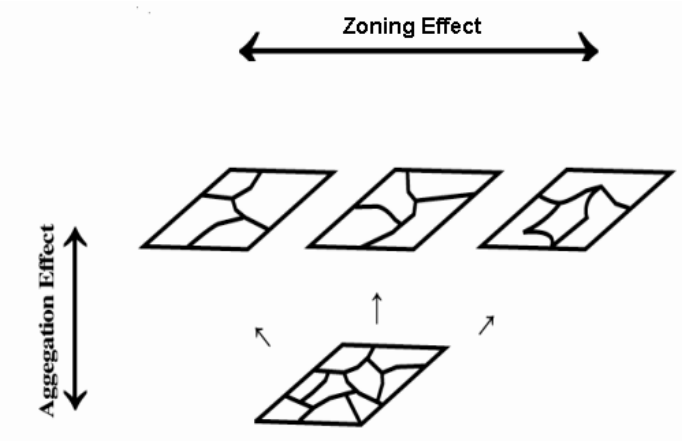


Fig. 1. Components of the support effect and sources of the MAUP. Adapted from Wong(1996).

Thus, the change of support problem and the MAUP are really comprised of two interrelated problems. The first occurs when different inferences are obtained when the same set of data is grouped into increasingly larger areal units. This is often referred to as the *scale effect* or *aggregation effect.* Aggregation reduces heterogeneity among units. The uniqueness of each unit and the dissimilarity among units are both reduced. However, spatial autocorrelation is a mitigating factor: When areal units are similar to begin with, aggregation results in much less information loss than when aggregating highly dissimilar units. Spatial aggregation also affects the spatial variability in the resulting units, often inducing positive spatial autocorrelation, particularly if the aggregation process allows overlapping units. The second, often termed the *grouping effect* or the *zoning effect*, arises from the variability in results due to alternative formations of the areal units that produce units of different shape or orientation at the same or similar scales (Openshaw and Taylor 1979, Wong 1996). The zoning effect is much less pronounced when aggregation of areal units is performed in a non-contiguous or spatially random fashion. It is most apparent only when contiguous units are combined, altering the spatial autocorrelation among the units. Combining smaller units through regularization is analogous to smoothing with different combinations of spatial neighbors. Depending on the similarity of the neighbors, different zoning rules may lead to different analytical results.

In geostatistics, the aggregation effect and the zoning effect are usually treated in a combined fashion through the ideas of the dispersion variance, the regularized semivariogram and its theoretical relationship to the point semivariogram and change of support models that account for both issues simultaneously. However,

to appreciate the solutions to the MAUP and the change of support problem developed in other disciplines, we found it helpful to separate the two components. Most solutions to upscaling problems address the effects of aggregation, and most solutions to downscaling problems recognize the need to reconstruct variation at the smaller scale, but the zoning effect issues associated with both of these problems are often ignored.

3 Solutions to change of support problems

Most solutions to change of support problems require spatial prediction of data associated with one set of units based on data associated with another set of units. In developing solutions to change of support problems, the criteria that such predictions should satisfy varies widely across the different disciplines. A collective list of some of the important considerations includes the following:

1. The ability to explicitly account for the differing supports of the spatial units involved;
2. A general framework that can be used for upscaling (aggregation), downscaling (disaggregation), or side-scaling (overlapping units); The framework should allow for upscaling from points to volumes or from volumes to other volumes with larger support. It should allow for downscaling from volumes to volumes with smaller support, or from volumes to points.
3. The predicted surface generated should be smooth across unit boundaries;
4. Standard errors of the predictions can be computed and these should accurately account for the uncertainty involved;
5. Covariates can be used to improve predictions;
6. The method can be used when the data and predictions are averages (as in Eq. 1) or counts/totals;
7. Predictions should lie in the parameter space (e.g., when predicting an inherently positive quantity, the predictions should not be negative);
8. There should be consistency in predictions across scales: For example, consider predicting $Z(A_{ij})$ from data $Z(B_1)$, …,$Z(B_m)$, where the A_{ij} , j=1, … n_i are nested within volume B_i where $A_j \cap A_k=\varnothing$ for j ≠k, and $\cup_{j=1}^{n_i} A_j = B_i$. Then the predictions within each volume B should add to the observed datum

$$Z(B_i) = \frac{1}{|B_i|}\sum_{j=1}^{n_i} \hat{Z}(A_{ij})$$

Huang *et al.* (2002) call this the *mass balance* property. When downscaling observed data that are totals and not averages to point support, then the predictions $\hat{Z}(s)$ should satisfy the *pycnophylactic* (volume preserving) property (Tobler 1979):

$$Z(A) = \int_A \hat{Z}(s)ds \cdot$$

9. Ideally, the prediction method should be based on a paucity of model and distributional assumptions;

10. The prediction method should be computationally feasible for routine use within a GIS where it is relatively easily to perform computations involving point-in-polygon operations and digital boundaries.

Of course, asking for a solution that satisfies all of these properties is probably unrealistic. However, this list provides a backdrop against which we can evaluate current solutions and understand their advantages and disadvantages. In the following sections, we provide an overview of some of the general types of solutions to change of support problems and briefly outline some of their main advantages and disadvantages. More comprehensive descriptions of the methods are found in the references provided and many of these are reviewed in more detail in Gotway and Young (2002). We deliberately exclude the rich literature on upscaling and downscaling in many of the physical sciences such as hydrology, soil science, and petroleum engineering in which models that adhere to engineering laws often form a basis for solutions to change of support problems.

3.1 GIS operations and raster calculations

Description: Basic geoprocessing operations with a GIS include union, intersection, and dissolve operations applied to the boundaries of the spatial units in order to create new spatial units. Raster calculations include averaging of interpolated values over irregularly shaped regions ("zonal" statistics) and pixel-by-pixel computations.
Main Advantages: Working with digital boundary files is the consummate utility of GIS. The computations are fast, invisible to the user and can explicitly factor in the support of the different units involved. Layers representing different variables can be combined using raster calculations so that covariates can be incorporated, although the effect of the covariate layers on the predictions must be specified, rather than inferred statistically. Smooth surface generation is straightforward and visualization is automatic.
Main Disadvantages: The main disadvantage is the lack of uncertainty measures for the resulting predictions. Moreover, when several layers with different supports are rasterized to the same cell size and then used in subsequent computations, error propagation is a big concern. Volume-volume disaggregation is done using proportional allocation. Depending on how many operations are used and their nature, the resulting predictions may not be aggregation consistent.

3.2 Spatial smoothing

Description: The goal with spatial smoothing methods is to make a smooth map from aggregated data. Methods in this group vary greatly and include point kriging based on centroids, kernel smoothing (Bracken and Martin 1989), support-adjusted locally weighted regression (Brillinger 1990, Muller *et al.* 1997), and pycnophylactic interpolation (Tobler 1979).
Main Advantages: Point kriging and kernel smoothing based on centroids are easily implemented and provide a measure of uncertainty associated with predictions. The kernel smoothing approach developed by Bracken and Martin (1989) and the pycnophylactic interpolation method of Tobler (1979) computationally constrain the predictions to be aggregation consistent. The methods developed by Brillinger (1990) and Müller *et al.* (1997) are more statistically sophisticated and allow adjustment for covariates and provide a measure of uncertainty. The methods developed by Tobler (1979), Brillinger (1990) and Müller *et al.* (1997) explicitly consider the supports of the units involved.
Main Disadvantages: The major disadvantage to these methods is that are concerned only with the volume-point change of support problem. Constraining predictions to ensure aggregation consistency (as in the methods of Bracken and Martin 1989 and Tobler 1979) makes it difficult to adjust for covariates and to obtain a valid measure of uncertainty. On the other hand, the methods developed by Brillinger (1990) and Müller *et al.* (1997) may not give predictions that are aggregation consistent.

3.3 Regression methods

Description: Proposed by Flowerdew and Green (1992), a regression model is assumed for data associated with "target" units, with the response data on target units treated as missing values. Starting values from proportional allocation are used to obtain initial estimates of the regression parameters. Updated estimates of target-unit data are then obtained from the regression model and constrained to satisfy the pycnophylactic property. This process is repeated until the estimates of the target unit data remain essentially unchanged.
Main Advantages: The main advantage is the ability to use covariates to estimate data on the target units. The regression framework can be used for a variety of change of support problems involving different types of data (binary, discrete and continuous) The computations are fairly simple and could be easily programmed into a GIS script.
Main Disadvantages: Because of the iterative process that includes the pycnophylactic constraint, accurate measures of the uncertainty in target-unit predictions cannot be obtained. Also, the regression model must be built on units formed by the intersection of the target units and the "source" units (those for which data are observed), and so covariates on these "atomic" units must be derived. The support of the units is not considered and spatial autocorrelation is ignored.

3.4 Bayesian hierarchical models

Description: A statistical model is specified for the data, given unknown variables, and then prior distributions are specified for the unknown variables. The unknown variables may include unknown data to be predicted. A posterior distribution is derived from the likelihood of the data that is updated by prior information in accordance with Bayes' theorem. Simulation methods are used to generate realizations from posterior distribution (see, Mugglin and Carlin 1998, Wikle *et al.* 2001, Gelfand *et al.* 2001, Kelsall and Wakefield 2002).
Main Advantages: The methodology is based on very elegant statistical theory combing Bayes' theorem, likelihood estimation and Markov chain theory. The posterior predictive distribution provides a comprehensive description of uncertainty. Complex models that include covariates on many different scales can be more easily constructed hierarchically than simultaneously.
Main Disadvantages: The models are computationally intensive. With the exception of the model in Gelfand *et al.* (2001) each model can be used to solve only one type of change of support problem, and solutions to other problems require complex statistical derivations. Most rely too heavily Gaussian distributions and many account for support only through areal weighting and hence ignore the zonal effect completely. The hierarchical specification can induce unknown constraints within the overall model. There has been little evaluation of the resulting uncertainty distribution (e.g., to assess ergodic properties, or the ability to contain a value of a transfer function of interest as described in Deutsch and Journel 1992 and Gotway and Rutherford 1994).

3.5 Multi-scale tree models

Description: Chou *et al.* (1994) developed a scale-recursive algorithm based on a multilevel tree structure for image processing in engineering. Each level of the tree corresponds to a different spatial scale (see Fig. 2). Data are observed at some of the nodes of the tree and the goal is prediction at other nodes of the tree. Algorithms are based on the Kalman filter. To eliminate some of the artifacts imposed by the tree structure and to ensure mass balance, Huang and Cressie (2000) and Huang *et al.* (2002) extend these models to more general graphical Markov models.
Main Advantages: The recursive nature of the Kalman filter (for which kriging is a special case) is extremely computationally efficient for processing huge data sets. It also provides a measure of uncertainty associated with the predictions.
Main Disadvantages: The tree structure ignores spatial support and it is not clear how it can be adapted to more general cases with overlapping spatial units. Statistical parameter estimation can be difficult.

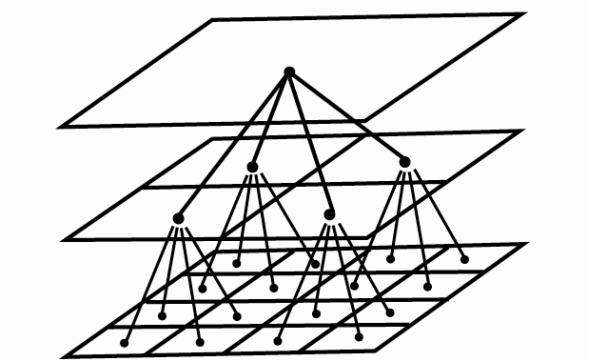


Fig. 2. A tree structure for multiscale processes.

3.6 Geostatistical methods

Description: Includes ``block" kriging, nonlinear geostatistical methods and isofactorial models (Journel and Huijbregts 1978, Matheron 1984a and b, Cressie 1993b, Rivoirard 1994, Goovaerts 1997, and Chilès and Delfiner 1999).
Main Advantages: The field of geostatistics includes many innovative solutions to change of support problems. These solutions have proven themselves in practical applications such as mining where profitability is of primary concern. A measure of prediction uncertainty can be easily obtained. The basic calculations needed for change of support predictions based on kriging and cokriging can be done in GIS.
Main Disadvantages: Most practical solutions were developed only for the upscaling problem. Estimating the semivariogram from data that are not of point support may be problematic. Prediction uncertainty may not adequately reflect estimation error in the semivariogram.

4 Towards a general framework

Clearly, the solutions to change of support problems range from those that are simple and require few assumptions, but are statistically unsophisticated (GIS and proportional allocation), to those that are complex and statistically elegant, but require many assumptions and are difficult to implement (Bayesian hierarchical models). Moreover, many solutions are particular to the change of support problem they were developed to address. We seek a compromise, one that provides a unified framework for the different types of change of support problems encountered in a variety of disciplines, is based on fewer assumptions, and can be implemented in a geographic information system (GIS) using current GIS technology, but also one that can incorporate covariates and provide standard errors for the resulting predictions.

While block kriging was developed for the upscaling problem, a slight modification allows the same ideas to be adapted to more general change of support problems (Journel and Huijbregts 1978, Gotway and Young 2002, Gotway and Young 2004). Consider the linear predictor

$$\hat{Z}(A_j) = \sum_{i=1}^{n} w_i(A_j) Z(B_i)$$

based on data $Z(B_1)$, …,$Z(B_n)$, where each weight $w_i(A)$ measures the influence of datum $Z(B_i)$ on the prediction of another variable with differing support, $Z(A)$. The theory of best linear unbiased prediction can be applied to determine optimal weights, $w_i(A)$ in a manner analogous to that used in the development of the block kriging predictor. The key to this development is the relationship between the semivariogram of $Z(B)$ and that of the underlying process $Z(s)$ (Journel and Huijbregts 1978, p. 77)

$$2\gamma(B_i, B_j) = \frac{2}{|B_i||B_j|}\int_{B_i}\int_{B_j}\gamma(s-u)dsdu - \frac{1}{|B_i||B_i|}\int_{B_i}\int_{B_i}\gamma(s-u)dsdu - \frac{1}{|B_j||B_j|}\int_{B_j}\int_{B_j}\gamma(s-u)dsdu.$$

Given data of point support, $\gamma(s\text{-}u)$ can be estimated and then used to determine the semivariogram of data at any other support, $\gamma(B_i\ B_j)$ and $\gamma(A_i, A_j)$. Although in many applications, data of point support are available, in man others, such data are not available. However, it is possible to still use this relationship. If a parametric model, $\gamma(s\text{-}u;\theta)$, is assumed for point support semivariogram, an estimate of θ can be obtained, and hence $\gamma(s\text{-}u;\theta)$ can be determined, from data of volume support $Z(B_1)$, …,$Z(B_n)$ (Mockus 1998, Gotway and Young, 2004). Computationally, it is easier to use the covariance functions

$$C(B_i, B_j) = Cov(Z(B_i), Z(B_j)) = \frac{1}{|B_i||B_j|}\int_{\mathbf{B}_j}\int_{\mathbf{B}_i} C(u, v; \theta)dudv$$

since only one multidimensional integration is required. Then, if $Y(B_i)= Z(B_i)-\mu$, θ can be estimated by the value that minimizes (Mockus 1998)

$$\sum_i \sum_j \{Y(B_i)Y(B_j) - \frac{1}{|B_i||B_j|}\int_{B_i}\int_{B_j} C(u-v;\theta)dudv)\}^2$$

The optimal weights needed to construct $\hat{Z}(A)$ can then be obtained from

$$\sum_{k=1}^{n} w_k(A)C(B_i, B_k) - m = C(A, B_i), \quad i = 1, \ldots, n$$

$$\sum_{k=1}^{n} w_k(A) = 1,$$

and the minimized prediction mean squared error (kriging variance) is given by

$$\sigma_K^2(A) = C(A, A) - \sum_{i=1}^{n} w_i(A)C(A, B_i) - m.$$

Gotway and Young (2004) extend these ideas to the "external drift" case where $E[Z(S)] = x(s)'\beta$ and develop an iterative generalize least squares approach to estimating the drift parameters and the autocorrelation parameters simultaneously.

If the data are totals instead of averages, so that $Z^*(B) = \int_B Z(s)ds$, this approach can be used with the normalized variables $N(B) = Z^*(B)/|B|$.

Since an optimal predictor is derived in terms of data with general supports A and B, it can be used for upscaling (spatial aggregation), downscaling (spatial disaggregation), or side scaling (overlapping) units, and the spatial units may be of point, areal, or volumetric support. Because the predictor is linear and honors the data, mass balance properties are inherently satisfied.

However, this approach suffers from the same problems encountered in using geostatistical methods with data of point support, namely the variability in the cross-products if not suitably binned and averaged, and the sensitivity of the estimates of θ to a few large cross-product values and choices for the lag spacing. Another disadvantage of the geostatistical framework when applied to count data is that negative predictions can occur; the predictions are not formally constrained be positive.

5 Summary and challenges

In spite of the rather substantial disadvantages associated with using GIS operations to combine spatial data, the ability to easily implement solutions to change of support problems within a GIS is overwhelmingly appealing. Thus, overall, this approach is the most commonly used method for combining incompatible spatial data and solving complex change of support problems. While Bayesian hierarchical models and isofactorial models offer elegant statistical solutions to a variety of change of support problems, their complexity (both statistical and computational) and their dependence on a large number of unverifiable, pedantic assumptions make them unattractive for routine use in most applied sciences at the present time. Thus, as a compromise, we considered a geostatistical approach to general change of support problems that allows downscaling and side scaling. This approach explicitly accounts for the supports of the data, can incorporate covariate information to improve the predictions, and provides a measure of uncertainty for each prediction.

While the geostatistical framework presented here is not without its disadvantages, it offers a way to put the concept of spatial support back into spatial analysis. Subsequent research and development could easily adapt this framework for use as a routine part of many software packages.

References

Armstrong M (1999) Basic Linear Geostatistics. Springer-Verlag New York

Bierkens MFP, Finke PA, De Willigen P (2000) Upscaling and Downscaling Methods for Environmental Research. Kluwer Dordrecht

Bracken I., Martin D (1989) The generation of spatial population distributions from census centroid data. Environment and Planning A 21: 537–543

Brillinger DR (1990) Spatial-temporal modeling of spatially aggregate birth data. Survey Methodology 16: 255-269

Chilès JP, Delfiner P (1999) Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, New York

Chou KC, Willsky AS, Nikoukhah R (1994) Multiscale systems, Kalman filters, and Riccati equations. IEEE Transactions on Automatic Control, 39: 479-492

Cressie N (1993) Statistics for Spatial Data. John Wiley & Sons, New York

Cressie N (1993b) Aggregation in geostatistical problems. In Soares, A (ed.) Geostatistics Troia '92, Kluwer Academic Publishers, Dordrecht, 25-35

Deutsch CV, Journel AG (1992) GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York.

Flowerdew R and Green M (1992) Developments in areal interpolating methods and GIS. Annals of Regional Science, 26: 67-78

Gelfand AE, Zhu L, Carlin BP (2001) On the change of support problem for spatio-temporal data. Biostatistics, 2: 31-45

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, New York

Gotway CA and Rutherford BM (1994) Stochastic simulation for imaging spatial uncertainty:comparison and evaluation of available algorithms. In Armstrong, M and Dowd, PA (eds). Geostatistical Simulations, Kluwer: Dordrecht, 1-21

Gotway CA, Young LJ (2002) Combining incompatible spatial data. Journal of the American Statistical Association, 97: 632-648

Gotway CA, Young LJ (2004) A geostatistical approach to linking geographically-aggregated data from different sources. Technical report # 2004-012, Department of Statistics, University of Florida

Huang H-C, Cressie N (2001) Multiscale graphical modeling in space: Applications to command and control. In Moore, M. (ed.) Spatial statistics. Methodological Aspects and Some Applications, vol. 159 of Lecture notes in Statistics, Springer Verlag, New York

Huang H-C, Cressie N, Gabrosek J (2002) Fast resolution-consistent spatial prediction of global processses from satellite data. Journal of Computational and Graphical Statistics, 11:1-26

Journel AG, Huijbregts CJ (1978) Mining Geostatistics. Academic Press, London

Kelsall J, Wakefield J (2002) Modeling spatial variation in disease risk: A geostatistical approach. Journal of the American Statistical Association, 97: 692-701

Matheron G (1963) Principles of geostatistics. Economic Geology, 58: 1246-1266

Matheron G (1984a) Isofactorial models and change of support. In Verly, G. *et al.* (eds.) Geostatistics for Natural Resources Characterization, Reidel, Dordrecht, 449-467

Matheron G (1984b) The selectivity of the distributions and the "second principle of geostatistics." In Verly, G. *et al.* (eds.) Geostatistics for Natural Resources Characterization, Reidel, Dordrecht, 449-467

Mockus A (1998) Estimating dependencies from spatial averages. Journal of Computational and Graphical Statistics, 7: 501-513

Mugglin AS, Carlin BP (1998) Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. Journal of Agricultural, Biological, and Environmental Statistics, 3: 117-130

Müller HG, Stadtmüller U, Tabnak F (1997) Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps. Journal of the American Statistical Association, 92: 61-71

Olea RA (ed.) (1991) Geostatistical Glossary and Multilingual Dictionary. Oxford University Press, New York

Openshaw S, Taylor P (1979) A million or so correlation coefficients. In Wrigley, N. (ed.) Statistical Methods in the Spatial Sciences, Pion, London, 127-144

Rivoirard J (1994) Introduction to disjunctive kriging and non-linear geostatistics. Clarendon Press, Oxford

Tobler W (1979) Smooth pycnophylactic interpolation for geographical regions (with discussion). Journal of the American Statistical Association, 74: 519-536

Wikle CK, Milliff RF, Nychka D, Berliner LM (2001) Spatio-temporal hierarchical Bayesian modeling: Tropical ocean surface winds. Journal of American Statistical Association, 96: 382-397

Wong DWS (1996) Aggregation effects in geo-referenced data. In Griffiths, D (ed.) Advanced Spatial Statistics, CRC Press, Boca Raton, Florida, 83-106

Combining categorical and continuous information using Bayesian Maximum Entropy

P. Bogaert and M.-A.Wibrin

Université catholique de Louvain, dept. of Environmental Sciences and Land Use Planning – Environmetry and Geomatics, Croix du Sud, 2 bte16, 1348 Louvain-la-Neuve, Belgium, e-mails: bogaert@enge.ucl.ac.be, wibrin@enge.ucl.ac.be

1 Introduction

Usually, in environmental studies, categorical and continuous data are used jointly. Each type of data convey substantial information. For example, a thematic map may often be used as an auxiliary information source for the prediction of a continuous variable. Sometimes, categorical information is collected during the sampling campaign and may also be used as an auxiliary information source for the prediction of a continuous variable.

In environmental sciences literature, several methods have been proposed so far for predicting a continuous variable with auxiliary categorical information. The most common are indicator kriging (Goovaerts 1997), stratified kriging (Stein *et al.* 1988, Goovaerts 1997), kriging with varying local means (Goovaerts 1997) and neural networks (Venables and Ripley 1994).

On the contrary, methods used to predict a categorical variable with the help of a continuous one are less common. Among those methods, one can find generalized linear models (Diggle *et al.* 1998) and fuzzy classification (Wang 1990).

All those methods suffer from limitations that prevent any kind of generalization to combine categorical and continuous variables in an unique framework.

Recently, the Bayesian Maximum Entropy approach has proved to be a powerful tool for processing spatial data sets (Christakos 2000, Christakos *et al.* 2002, D'Or *et al.* 2001). Based on sound information processing rules and classical probability laws, this geostatistical method has been developed first for continuous variables (Christakos 2000) and later for categorical data (Bogaert 2002). Its strongest feature is its ability to manage data of various nature and quality ('hard' and 'soft' data). The paradigm of BME methods is based on a constrained maximization of an entropy. From an epistemic point of view, BME analysis involves three steps, namely the prior, the meta-prior and the posterior steps. The first one yields the prior pdf by maximization of the entropy. This reverts to maximize the expected information content while ensuring that all the available information, called the general knowledge, is taken in account. Constraints are the continuous, categorical and mixed first and second order moments. The posterior step is a bayesian conditionalization that consists in incorporating the specific information

provided by the data (the observed values in a neighborhood around the prediction location).

We tackle the problem of combining categorical and continuous variables for spatial prediction within the BME framework, this method will be named BME/MIX hereafter. We propose a unique framework for estimating the joint distribution of both types of variables based on the entropy maximization. In this article, we will develop a two-points case with two spatial random fields (a continuous and a categorical one) for which the specific information consists in the categorical values observed at the prediction point and at the nearest data point as well as the continuous value at this last point.

The first part of the paper is dedicated to some theoretical developments. The mixed random field is presented as well as the different BME steps applied to the mixed case. In the second part, we present the experimental design with the simulation of the reference map and the sampling strategy. In the third part we compare the BME/MIX results with those obtained from different common geostatistical methods.

2 Theroretical developments

A definition of the mixed random field (RF) and some notational conventions will be first presented. The three different kinds of covariance that can be defined for the mixed case are then explained. The constrained maximization problem is defined and it is shown that the prior probability density function (pdf) is a combination of Gaussian pdf's. The estimation of this prior pdf is finally presented.

2.1 The mixed RF

Given two random variables (RV's) $Z(\mathbf{x})$ and $C(\mathbf{x})$, where $\mathbf{x}=(x_1,\ldots,x_d)' \in \mathcal{D}$ refers to a spatial location over a d-dimensional domain $\mathcal{D} \subseteq \mathbb{R}^d$, such that $F_Z=\{ Z(\mathbf{x}), \mathbf{x} \in \mathcal{D} \}$ and $F_C=\{C(\mathbf{x}), \mathbf{x} \in \mathcal{D} \}$ are two spatial RF's. The first one is assumed to be continuous with $Z(\mathbf{x}) \in \mathbb{R}^1$, whereas the second one is categorical (ordinal or nominal) with $C(\mathbf{x}) \in \{c_1,\ldots,c_m\}$, where the c_k's are forming a complete system of events.

Using a disjunctive coding of the categorical variables, we have $F_{Yk} = \{Y_k(\mathbf{x}), \mathbf{x} \in \mathcal{D}\}$, k = 1,…,m where F_{Yk} 's are binary RF's that are summing to one. Each $Y_k(\mathbf{x})$ is a Bernoulli random variable taking as value 1 if $C(\mathbf{x})=c_k$ and 0 otherwise, subject to the constraint that $\sum_{k=1}^{m} Y_k(\mathbf{x}) = 1$.

2.2 Covariances in a mixed random field

Whatever the RF's we are considering, it is useful to characterize them at least partially by computing the covariance between two RV's taken respectively at locations x_i and x_j. For the continuous RF, we will simply make use of the classical definition

$$Cov[Z(\mathrm{x}_i), Z(\mathrm{x}_j)] = E[Z(\mathrm{x}_i)Z(\mathrm{x}_j)] - E[Z(\mathrm{x}_i)]E[Z(\mathrm{x}_j)] \quad (1)$$

If we compute the covariance between two RV's $Y_k(\mathbf{x}_i)$ and $Y_{k'}(\mathbf{x}_j)$, we get

$$\begin{aligned} Cov[Y_k(\mathrm{x}_i), Y_{k'}(\mathrm{x}_j)] &= E[Y_k(\mathrm{x}_i)Y_{k'}(\mathrm{x}_j)] - E[Y_k(\mathrm{x}_i)]E[Y_{k'}(\mathrm{x}_j)] \\ &= P(C(\mathrm{x}_i) = c_k \bigcap C(\mathrm{x}_j) = c_{k'}) - P(C(\mathrm{x}_i) = c_k)P(C(\mathrm{x}_j) = c_{k'}) \end{aligned} \quad (2)$$

Furthermore, we can also compute the covariance between any Bernoulli RV $Y_k(x_i)$ and any continuous RV $Z(x_j)$, with

$$\begin{aligned} Cov[Y_k(\mathrm{x}_i), Z(\mathrm{x}_j)] &= E[Y_k(\mathrm{x}_i)Z(\mathrm{x}_j)] - E[Y_k(\mathrm{x}_i)]E[Z(\mathrm{x}_j)] \\ &= E[Z(\mathrm{x}_j) \mid C(\mathrm{x}_i) = c_k]P(C(\mathrm{x}_i) = c_k) - E[Z(\mathrm{x}_j)]P(C(\mathrm{x}_i) = c_k) \\ &= \left(E[Z(\mathrm{x}_j) \mid C(\mathrm{x}_i) = c_k] - E[Z(\mathrm{x}_j)]\right)P(C(\mathrm{x}_i) = c_k) \end{aligned} \quad (3)$$

The first two covariances $Cov[Z(x_i),Z(x_j))]$ and $Cov[Y_k(x_i),Y_{k'}(x_j)]$ are well-known and widely used in geostatistics, the second one corresponding to the popular indicator covariance (Journel 1983). The third covariance $Cov[Y_k(x_i),Z(x_j)]$ is more peculiar; it shows that, up to the multiplicative constant $P(C(x_i)=c_k)$, the covariance values between the binary random variable Y_k and the continuous random variable Z are given by the difference between the conditional expectation $E[Z(x_j)|C(x_i)]$ and the unconditional expectation $E[Z(x_j)]$. The mixed covariance function is expected to be monotonic and it will drop down to zero or close to zero for high distances. Indeed, it is expected that the conditional expectation will become more and more similar to the unconditional one when the two points are more distant. The mixed covariance function may be increasingly or decreasingly, depending on which mean is the highest (the conditional one or the unconditional one).

2.3 BME/MIX constrained maximization

In the next part of this paper, we will assume that, under second-order stationarity hypothesis, all first-order and second-order theoretical moments can be inferred with reasonable accuracy from a dataset, so they constitute the general knowledge that we have at hand about the mixed RF. We will consider hereafter that we have all those moments for obtaining the joint multivariate pdf for a set of locations $\mathbf{x}_{map}=\{\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_n\}$, that we will denote as

$$f(\mathbf{y}_{map}, \mathbf{z}_{map}) = f_{\{\bigcup_k Y_k(\mathrm{x}_0),\ldots,Y_k(\mathrm{x}_n)\},Z(\mathrm{x}_0),\ldots,Z(\mathrm{x}_n)}(\bigcup_k \{y_{k0},\ldots,y_{kn}\}, z_0,\ldots,z_n) \quad (4)$$

where y_{ki} refers to the Bernoulli variable $Y_k(x_i)$, and where z_i refers to the continuous variable $Z(\mathrm{x_i})$. What we seek for is the distribution that has maximum Shannon's entropy $H_{\mathbf{Y}map,\mathbf{Z}map}$, with

$$H_{\mathbf{Y}_{map},\mathbf{z}_{map}} = -\sum_{\mathbf{y}_{map}} \int_{-\infty}^{\infty} \ln\left(f(\mathbf{y}_{map},\mathbf{z}_{map})\right) f(\mathbf{y}_{map},\mathbf{z}_{map}) d\mathbf{z}_{map} \tag{5}$$

where the integral is a n-fold integral and $d\mathbf{z}_{map} = \prod_{i=1}^{n} d\mathbf{z}_i$ and that respects a set of (direct and cross-) second-order moments given by those functions

$$E[Z(\mathrm{x}_i)Z(\mathrm{x}_j)] = \sum_{\mathbf{y}_{map}} \int_{-\infty}^{\infty} z_i z_j f(\mathbf{y}_{map},\mathbf{z}_{map}) d\mathbf{z}_{map} \tag{6}$$

$$E[Y_k(\mathrm{x}_i)Y_{k'}(\mathrm{x}_j)] = \sum_{\mathbf{y}_{map}} \int_{-\infty}^{\infty} y_{ki} y_{k'j} f(\mathbf{y}_{map},\mathbf{z}_{map}) d\mathbf{z}_{map} \tag{7}$$

$$E[Y_k(\mathrm{x}_i)Z(\mathrm{x}_j)] = \sum_{\mathbf{y}_{map}} \int_{-\infty}^{\infty} y_{ki} z_j f(\mathbf{y}_{map},\mathbf{z}_{map}) d\mathbf{z}_{map} \tag{8}$$

where $E[Z(\mathrm{x}_i)Z(\mathrm{x}_j)]$, $E[Y_k(\mathrm{x}_i)Y_{k'}(\mathrm{x}_j)]$ and $E[Y_k(\mathrm{x}_i)Z(\mathrm{x}_j)]$ are values that were inferred from the data set $\forall$ i, j, k, and where summation covers all the possible combinations of the binary variables.

2.4 The prior pdf, a mixture of Gaussian distributions

A well-known result states that the general expression for the maximum entropy solution with moments incorporated as constraints has an exponential form (Jaynes 1982). The lagrangian solution is,

$$f(\mathbf{y}_{map},\mathbf{z}_{map}) = \frac{1}{A}\exp\left(\sum_{\alpha} \mu_\alpha g_\alpha(\mathbf{y}_{map},\mathbf{z}_{map})\right) \tag{9}$$

where the μ_α are the lagrangians that must be identified from the constraints, $g_\alpha(y_{map}, z_{map})$ is the set of functions that correspond to the constraints we want to incorporate, and where A is a normalization constant. For our specific problem, it is easier to develop and rearrange the terms by making use of the matrices $\mathbf{B}$, $\mathbf{N}_i$ and $\mathbf{\Gamma}_{ij}$, that are lagrangian coefficients matrices respectively associated with the terms $\mathbf{zz}$, $\mathbf{y}_i\mathbf{z}$ and $\mathbf{y}_i\mathbf{y}_j$, so that

$$f(\mathbf{y}_{map},\mathbf{z}_{map}) = \frac{1}{A}\exp\left(-\mathbf{z}'\mathbf{B}\mathbf{z} - \left(\sum_i \mathbf{y}'_i \mathbf{N}_i\right) z - \sum_{i\neq j} \mathbf{y}'_i \mathbf{\Gamma}_{ij} \mathbf{y}_j\right) \tag{10}$$

Let us define l=1,...,m^{n+1} as an index over the set of all possible combinations of categories at locations $\mathrm{x}_0,...\mathrm{x}_n$. For a given choice $\mathbf{y}_{l,map}$ of $\mathbf{y}_{map}$ at these locations, the joint conditional distribution of $\mathbf{z}_{map}$ can be written as

$$f(\mathbf{z}_{map} \mid \mathbf{y}_{l,map}) = \frac{f(\mathbf{y}_{l,map},\mathbf{z}_{map})}{\int f(\mathbf{y}_{l,map},\mathbf{z}_{map}) d\mathbf{z}_{map}} \tag{11}$$

$$\alpha \quad \exp\left(-\mathbf{z}'\mathbf{B}\mathbf{z} - \boldsymbol{\eta}'_l \mathbf{z}\right)$$

where the values for $\boldsymbol{\eta}'_l = \sum_i \mathbf{y}'_i \mathbf{N}_i$ are depending on the specific choice for the conditioning categories, and the denominator is a normalization constant so that

$f(\mathbf{z}_{map}|\mathbf{y}_{l,map})\ d\mathbf{z}_{map}= 1$. It is not difficult to prove that this is the general expression for a multivariate Gaussian distribution $N(\boldsymbol{\mu}_l,\boldsymbol{\Sigma})$ with a mean vector $\boldsymbol{\mu}_l = (E[Z_0|\mathbf{y}_{l,map}], \ldots, E[Z_n|\mathbf{y}_{l,map}])'$ that depends on the conditioning categories and a covariance matrix $\boldsymbol{\Sigma}=½\mathbf{B}^{-1}$. As the general expression of $f(\mathbf{z}_{map}|\mathbf{y}_{l,map})$ always includes the same $\mathbf{z}'\mathbf{B}\mathbf{z}$ that does not depend on the choice of the conditioning categories, this also shows that the covariance matrix of all conditional distributions $f(\mathbf{z}_{map}|\mathbf{y}_{l,map})$ will be the same, whatever the choice for $\mathbf{y}_{l,\text{map}}$; these distributions will only differ with respect to their mean vector $\boldsymbol{\mu}_l$.

2.5 Prior pdf estimation

Our objective is to find the prior pdf that maximize the joint entropy $H_{\mathbf{Y}map,\mathbf{Z}map}$, which is a mixture of Gaussian conditional distributions. However, both $\boldsymbol{\Sigma}$ and the $\boldsymbol{\mu}_l$'s are unknown and have to be estimated. In order to find the solution, we propose to decompose the entropy in two terms using the well-known formula (Gray 1990)

$$H_{\mathbf{Y}_{map},\mathbf{Z}_{map}} = H_{\mathbf{Z}_{map}|\mathbf{Y}_{map}} + H_{\mathbf{Y}_{map}} \tag{12}$$

where $H_{\mathbf{Y}map}$ is the entropy for the categorical variable and $H_{\mathbf{Z}map|\mathbf{Y}map}$ is the total conditional entropy, with

$$H_{\mathbf{Y}_{map}} = -\sum_{\mathbf{y}_{map}} \ln\left(f(\mathbf{y}_{map})\right) f(\mathbf{y}_{map}) \tag{13}$$

$$H_{\mathbf{Z}_{map}|\mathbf{Y}_{map}} = \sum_{\mathbf{y}_{map}} f(\mathbf{y}_{map}) H_{\mathbf{Z}_{map}|\mathbf{y}_{map}} \tag{14}$$

where $H_{Zmap|Ymap}$ is the conditional entropy defined as

$$H_{\mathbf{Z}_{map}|\mathbf{y}_{map}} = -\int_{-\infty}^{\infty} \ln\left(f(\mathbf{z}_{map}\,|\,\mathbf{y}_{map})\right) f(\mathbf{z}_{map}\,|\,\mathbf{y}_{map})\, d\mathbf{z}_{map} \tag{15}$$

Maximizing $H_{\mathbf{Y}map,\mathbf{Z}map}$ is not equivalent to maximizing separately $H_{\mathbf{Y}map}$ and $H_{\mathbf{Z}map|\mathbf{Y}map}$, but provide that the joint probabilities estimates $\hat{\pi}_{\ell,map}$ that will be obtained from the maximization of $H_{\mathbf{Y}map}$ are not too different from the true $\pi_{\ell,map}$'s, it is a reasonable simplification. It also makes things much easier. Indeed, (i) $H_{\mathbf{Y}map}$ does not depend on $\mathbf{Z}_{map}$ so its entropy can be maximized subject to the constraints about bivariate probabilities (Bogaert 2002), and (ii) as the conditional distributions $f(\mathbf{z}_{map}|\mathbf{y}_{,lmap})$ are all multivariate Gaussian with same covariance matrix $\boldsymbol{\Sigma}$, as seen from eq. 14, the total conditional entropy becomes

$$H_{\mathbf{Z}_{map}|\mathbf{Y}_{map}} = \ln(2\pi e)^{\frac{n+1}{2}} \sqrt{\det(\boldsymbol{\Sigma})} \tag{16}$$

Therefore, the first step of BME/MIX algorithms is equivalent to BME/CAT, that is determining the categorical distribution on the basis of the constrained maximization of $H_{\mathbf{Y}map}$. The second step is based on the maximization of the entropy $H_{\mathbf{Z}map|\mathbf{Y}map}$, which is equivalent to maximizing the determinant of the conditional covariance's, $\det(\boldsymbol{\Sigma})$. Conditional covariances are calculated on the basis of

corresponding general covariance ($\mathbf{\Sigma}_{\mathbf{Z}map}$), BME probability table ($\pi_{l,map}$) and conditional means ($\boldsymbol{\mu}_l$), with

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathbf{Z}_{map}} + \mathbf{1}\mathbf{1}'m_Z^2 - \sum_{\ell=1}^{m^{n+1}} \pi_{\ell,map}\boldsymbol{\mu}_\ell\boldsymbol{\mu}_\ell' \tag{17}$$

and where **1** is a n*1 unit vector. The BME/MIX prior pdf is easily derived from those results.

2.6 Posterior step

When the joint maximum entropy prior pdf has been obtained, the specific knowledge can be used to obtain the conditional posterior pdf at the unsampled location x_0. The specific information is the hard and various types of soft data collected on the study site. For the sake of brevity, we will focus on the prediction of the continuous variable. For example, in the case there is no soft information, the posterior pdf is

$$f(z_0 \mid \mathbf{z}_{map}, c_0, \mathbf{c}_{map}) = \frac{f(z_0, \mathbf{z}_{map}, c_0, \mathbf{c}_{map})}{f(\mathbf{z}_{map}, c_0, \mathbf{c}_{map})} \tag{18}$$

Incorporation of soft information is also possible as for BME/CONT and BME/CAT (Christakos et al. 2002, D'Or 2003).

From this entire pdf, the relevant estimate can be chosen, according to the goals of the study. It could be the mean, the mode, the variance or a quantile for example.

3 Experimental design

In order to compare BME/MIX prediction performance with those of the classical approaches, reference maps are simulated; this will allow us to perform a fair and objective performance comparison. Hard data are sampled from these maps and will be considered as the only available information for each prediction method.

3.1 Simulation of the reference map

To obtain a mixed RF, the first step will be to generate two correlated continuous RF's, *Z* and *Q*. They are jointly simulated at 900 locations over a regularly spaced 30 by 30 grid with a traditional non conditional simulation method based on a Choleski decomposition. The theoretical covariance model is exponential with a range equal to 9 and a sill equal to 1. The covariance between *Z* and *Q* is equal to 0.8. Both RF's are Gaussian and have a mean equal to zero.

Second, for building a categorical RF, namely C, simulated values of Q are replaced by the interval to which they belong. The bounds of these intervals are the 0, 1/3, 2/3 and 1 quantiles of a zero mean unit variance Gaussian distribution.

Both the simulated Z and C (Fig. 1) RF's are the reference to which simple kriging, stratified kriging, simple kriging with varying local means and BME/MIX estimates will be compared.

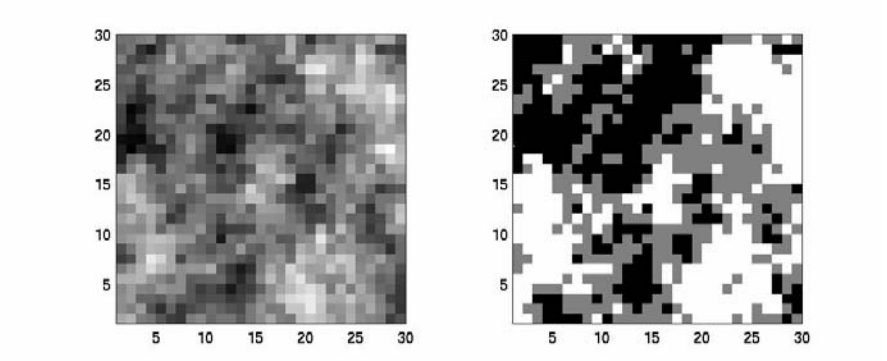


Fig. 1. Maps of the simulated (a) continuous RF and (b) corresponding categorical RF. Black indicates the lowest value and white the highest.

3.2 Sampling strategy

From the 900 simulated data, a hard data set is extracted. The continuous hard data set consists in the continuous simulated values at 82 locations that are randomly sampled over the grid. The categorical hard data set is the whole simulated categorical RF so that classical approaches like stratified or residual kriging can be used. Prediction is conducted over the whole grid and predicted values for the RF will be compared to the reference (simulated) values given in Fig.1.

4 Comparisons between methods

For comparing prediction performances between methods, two cases will be considered. The first one will be to conduct prediction for each method using all the available information at hand in the neighborhood of a prediction location (the neighborhood will consist in the five closest sampled locations around each prediction location); this will allow us to see how well each method can perform in the most favorable situation. The second case will be to conduct prediction using only the information about the sampled location which is the closest one to the prediction location; as the current implementation of BME/MIX only makes uses of this closest information, this allows us to perform a fair comparison between the methods as performances are thus compared on the basis of the same used information.

4.1 Geostatistical methods

The four methods that are compared using specific criteria for measuring prediction accuracy are respectively simple kriging (SK), stratified kriging (StK), simple kriging with varying local mean (SKlm) and Bayesian Maximum Entropy for a mixed case (BME/MIX).

The first three methods are well-known (Goovaerts 1997). The simplest one among them is SK, a method that does not take categorical information into account. As the categorical information is spatially exhaustive, StK and SKlm may also be used in a straightforward way. For StK, prediction is performed within each specific stratum (the stratum corresponding to a category in our case) using the corresponding covariance model and only the data belonging to the considered stratum. In the second variant SKlm, the specific stratum mean is subtracted from hard data before prediction and is added afterward; prediction is therefore performed with all data and a unique covariance model.

4.2 Comparison criteria

Criteria that can be used for comparing the four methods are either local or global. The local criterium used here is the pattern of the map produced by each method. Those maps will allow us to emphasize the ability of the methods to reproduce local features by comparison with the reference spatial pattern.

On the global scale, the methods are compared on basis of their mean errors (ME) and root mean errors (RMSE), respectively computed as:

$$\mathrm{ME} = \frac{1}{n}\sum_{k=1}^{n}(z_k - \hat{z}_k), \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(z_k - \hat{z}_k)^2}.$$

with z_k the reference value at location x_k, $\hat{z}_k$ the predicted value at the same location and n the number of prediction locations. Ideally, ME should be close to zero, so that there is no bias. The closest the RMSE value is to zero, the more accurate will be the method on the average. Those global criteria evaluate the method's ability for yielding predictions as close as possible to the observations "on the average".

4.3 Prediction performances

Table 1 summarizes the ME and RMSE results for SK, StK, SKlm and BME/MIX predictions. As expected from theory, ME's are not different from zero, indicating that the estimators seem to be unbiased. Depending on the way categorical information is incorporated, ME's and RMSE's are slightly different for all the methods. One can make the same observations for maps (Fig. 2).

Table 1. ME and RMSE with five and one data points in the neighborhood, except for BME/MIX that uses only one data point for prediction.

	five data points				one data point		
	SK	StK	SKlm	BME/MIX	SK	StK	SKlm
ME	-0.018	-0.067	-0.077	-0.042	0.033	-0.09	-0.085
RMSE	0.726	0.543	0.518	0.523	0.754	0.659	0.615

Obviously, the prediction that does not use categorical information has the worst performances. The SK map shows only a partial idea of the structure and reproduces only the global pattern of the reference map, and other maps that are using the categorical information have a pattern closer to the reference one. The higher RMSE values for SK compared to other methods also confirms a lower accuracy.

Performances are slightly better for SKlm than for Stk; this can be explained by the differences in the way categorical information is incorporated, as well as by the amount of data they have at hand for prediction.

Single point BME/MIX map is quite similar to the reference map. Their patterns are alike but small structures are not yet identified on the BME/MIX map.

Note also that BME/MIX map is quite similar to the SKlm and StK maps based on a five sampled locations neighborhood. Similar performances with less information comes from the fact BME/MIX is able to explicitly use the spatial link between categorical and continuous variables, whereas this is not the case for the other methods.

RMSE values obtained for prediction based on a single neighboring value decrease from SK to StK, SKlm and finally BME/MIX. Among the corresponding maps, it is also clear that the BME/MIX one is the closest one to the simulated map. So, when all methods use the same information, BME/MIX shows the best performances. The exploitation of the spatial relation between categorical variables and the links between the categorical and the continuous RF can explain this.

As a final remark, when using other ranges, covariance matrix and bounds of the intervals for the categorical variable, the same type of results are obtained: (1) BME/MIX predicts often a good as SKlm does with five data points, and (2) with the same used information, BME/MIX is a better prediction method than classical ones.

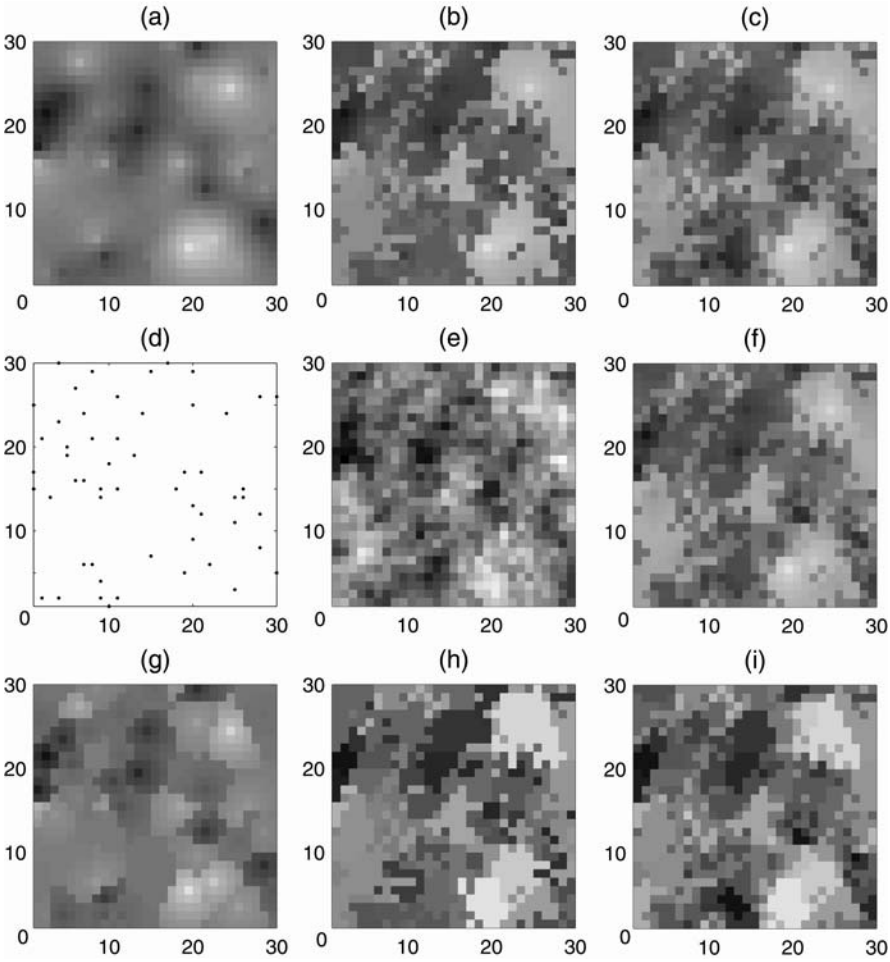


Fig. 2. Maps of (a) simple kriging, (b) stratified kriging and (c) kriging with varying local means where predictions are performed with five data points; (d) hard data points; (e) simulation; (f) BME/MIX one point prediction; (g) simple kriging, (h) stratified kriging and (i) kriging with varying local means where predictions are performed with one data point. Black indicates the lowest values and white the highest (color scale is the same for all figures).

5 Conclusions

The observed differences between the results obtained using different geostatistical methods are coming from the way the secondary information (i.e., the categorical information) is processed. While SK simply ignores it, SKlm and StK consider a different mean value of the continuous variable for each categorical stratum, whereas BME uses spatial links between the categorical and the continuous RF's.

With only one hard data point, BME/MIX is able to perform predictions that are as good as those obtained using StK or SKlm with more data points. With the same information is used (only a single data point), BME/MIX is clearly more accurate than SK and StK and at least as accurate as SKlm.

It is worth also noting that BME/MIX is the only method that is always able to produce a complete posterior pdf, thus allowing an easy computation of a wide range of indicators or estimates, like the mean, the median, the variance, the mode or even confidence intervals. Therefore, the estimation step (obtaining a posterior pdf) can be separated from the decision step (choosing a single representative value).

Although the methodology presented here focused on a single categorical RF in combination with a single continuous one, it can be generalized for several RF's considered at the same time (multivariate case), as well as for combining hard and soft data. E.g., for the categorical RF, soft data may come from the imprecise knowledge that one may have about categories at some locations, whereas for the continuous RF soft data may consist of intervals or pdf's.

For all those reasons, BME/MIX can be considered as a serious challenger compared to traditional geostatistical methods, as the methods does really improve the spatial mapping results by making used of the available information in a more relevant and efficient way.

Acknowledgements

This work has been partially supported by a belgian grant of the Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture.

References

Bogaert P (2002) Spatial prediction of categorical variables: the Bayesian Maximum Entropy approach. Stoch. Env. Res. Risk A., 16 (6) : 425-448

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS. Advanced Functions for Field-Based Applications. Springer-Verlag, New-York

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. Journal of the Royal Statistical Society series C-Applied Statistics, 47 : 299-326

D'Or D, Bogaert P, Christakos G (2001) Application of the BME approach to soil texture mapping. Stoch. Env. Res. Risk A., 15 (1) : 87-100

D'Or (2003) Spatial prediction of soil properties: the BME approach. Thesis dissertation. Université catholique de Louvain, Belgique

Gray RM (1990). Entropy and Information Theory. Springer-Verlag, New York

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, New York

Jaynes ET (1982) On the Rational of Maximum-Entropy Methods. Proceedings of the IEEE, 70(9) : 939-952

Journel AG (1983) Nonparametric estimation of spatial distributions. Mathematical Geology, 45 : 445-468

Stein A, Hoogerwerf M, Bouma J (1988) Use of Soil-Map Delineations to improve (Co)-Kriging of Point Data on Moisture Deficit. Geoderma, 43 : 163-17

Venables WN, Ripley BD (1994) Modern applied statistics with S-Plus. Verlag, New York

Wang F (1990) Fuzzy supervised classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 28 : 194-201.

Geostatistical prediction of spatial extremes and their extent

N. Cressie, J. Zhang and P.F. Craigmile

Department of Statistics, The Ohio State University, Columbus, OH 43210 USA

1 Introduction

The motivating example in this paper involves a region whose soils have been contaminated by tetrachlorodibenzo-p-dioxin (TCDD). Remediation is required when TCDD concentrations are above predetermined levels. That is, for a geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^d\}$ and a given threshold t, we are interested in spatial prediction of such nonlinear functionals as $I(Z(A) \geq t)$ based on data $\mathbf{Z} \equiv (Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n))'$, where $Z(A) \equiv ave\{Z(\mathbf{s}):\mathbf{s} \in A\}$ and $A \subset D$.

The prediction of *linear* functionals of Z, such as $Z(\boldsymbol{s}_0)$ at a known point $\boldsymbol{s}_0$ or $Z(A)$ for a given block A, can be carried out via kriging (e.g., Matheron 1963, Journel and Huijbregts 1978 - Chapter V, Cressie 1993b - Chapter 3). When predicting *nonlinear* functionals of Z, the major thrust in geostatistics has been to use nonlinear predictors such as indicator kriging (Journel 1983), indicator cokriging (Lajaunie 1990), and disjunctive kriging (Matheron 1976). While these methods are appropriate for nonlinear functionals like $I(Z(\boldsymbol{s}_0) \geq t)$, they do not generalize to the problem of predicting $I(Z(A) \geq t)$. Clearly, conditional simulation (e.g., Deutsch and Journel 1992) can be used for inference, but we would like to show in this paper that a more analytic method based on loss functions targeted at spatial extremes goes beyond the usual inference from conditional simulation. This method is compared with two types of kriging.

The first kriging is ordinary or universal kriging (e.g., Journel and Huijbregts 1978). The second kriging is covariance matching constrained kriging (CMCK) due to Aldworth and Cressie (2003), where constraints are added to the kriging equations that force elements of the variance matrix of a vector of linear predictors to match those from the corresponding predictands. The CMCK predictor is unbiased, has approximate optimal mean squared prediction error, handles additive measurement error straightforwardly, and can predict nonlinear functionals of $Z(A)$ just as easily as those of $Z(\boldsymbol{s}_0)$. Its strength is its generality for handling many types of nonlinearity, but it has not been tested properly on highly nonlinear functionals like extrema and their spatial extent. Recently, Craigmile *et al.* (2004) have developed a method of tackling such functionals by directly building loss functions (namely, IWQSELs) that put more weight on values of Z that are spatially extreme according to a target value of α near (but less than) 1.

The IWQSEL predictors are described in Section 2, and Section 3 contains a brief summary of CMCK prediction. Section 4 gives details on the example concerning environmental characterization and remediation of TCDD contamination in soil. This is followed with a comparison of kriging and CMCK to IWQSEL prediction using the TCDD example. Conclusions and discussion are given in Section 5.

2 Loss functions for extremes

Suppose that the geostatistical process Z satisfies,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}); \ \mathbf{s} \in D, \tag{1}$$

where $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$ denotes a linear-trend component for fixed explanatory variables, $\mathbf{x}(\mathbf{s}) \in \mathbb{R}^p$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ are regression parameters. We assume that the process $\{\delta(\mathbf{s}): \mathbf{s} \in D\}$ is a mean-zero, stationary spatial process with covariance function $C_{\boldsymbol{\theta}}(\cdot)$ and spatial parameters, $\boldsymbol{\theta} \in \mathbb{R}^q$. Let $\boldsymbol{\phi} \equiv (\boldsymbol{\beta}', \boldsymbol{\theta}')'$. Recall that the data are $\mathbf{Z} = (Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n))'$ observed at locations $\{\mathbf{s}_1,\ldots, \mathbf{s}_n\}$.

The cumulative distribution function of $Z(\mathbf{s})$ at some spatial location $\mathbf{s}$ is defined by $F_{Z(\mathbf{s})}(z) \equiv \Pr(Z(\mathbf{s}) \le z)$. As in Craigmile et al. (2004), we define the averaged cumulative distribution function (ACDF) over the region $B \subset D$ to be

$$F_B(z) \equiv \frac{1}{|B|}\int_B F_{Z(\mathbf{s})}(z)d\mathbf{s}, \tag{2}$$

where $|B|$ denotes the d-dimensional volume of the region B. The inverse ACDF is then defined by $F_B^{-1}(p) \equiv \inf\{z \in \mathrm{IR}: F_B(z) \ge p\}$. Now we introduce the loss function for predicting the process $Z(\cdot)$ in B using $\widetilde{Z}(\cdot)$:

$$L_B(Z(\cdot),\widetilde{Z}(\cdot)) \equiv \int_B w_B(Z(\mathbf{s}))(Z(\mathbf{s}) - \widetilde{Z}(\mathbf{s}))^2 d\mathbf{s}, \tag{3}$$

where

$$w_B(Z(\mathbf{s})) \equiv \int_0^1 w(p) I(Z(\mathbf{s}) \in (F_B^{-1}(p), F_B^{-1}(p+dp))]), \tag{4}$$

I$(\cdot)$ is the indicator function, and the "importance function" $w(\cdot) : [0,1] \to [0,\infty)$ is prespecified. Craigmile et al. (2004, Sec. 2.3) give examples of possible choices for $w(\cdot)$, such as the sigmoid-type function defined by

$$w(p) = \frac{1}{1+e^{-C(p-\alpha)}}; \ \ p \in [0,1], \tag{5}$$

where $C > 0$ is a scale parameter and $1/2 < \alpha < 1$ is the target value for which we wish to predict high extremes and their extent. For a given α, larger values of C put more weight on larger values of $Z(\cdot)$ in the loss function (3). Thus, C controls

the amount of shrinkage in the predictor. A good choice of C can be obtained by minimizing the bias in estimating the inverse ACDF, $F_B^{-1}(\alpha)$, for known ϕ; in practice, an estimate $\hat{\phi}$ is plugged in.

Upon minimizing the loss function (3) componentwise, Craigmile et al. (2004) show that the IWQSEL predictor of $Z(\cdot)$ at a location $\mathbf{s}^* \in B$ is given by

$$\widetilde{Z}_{wq}(\mathbf{s}^*) = \frac{\int_{IR} w_B(z) f_{Z(\mathbf{s}^*)}(z \mid \mathbf{Z}) z \; dz}{\int_{IR} w_B(z) f_{Z(\mathbf{s}^*)}(z \mid \mathbf{Z}) \; dz} = \frac{E(w_B(Z(\mathbf{s}^*))Z(\mathbf{s}^*) \mid \mathbf{Z})}{E(w_B(Z(\mathbf{s}^*)) \mid \mathbf{Z})}, \tag{6}$$

where $f_{Z(\mathbf{s}^*)}(z \mid \mathbf{Z})$ denotes the conditional density of $Z(\mathbf{s}^*)$ given the data $\mathbf{Z}$, sometimes called the predictive density. Note that (6) is a special functional of this density and hence it could equally well be computed by conditional simulation.

In practice, we need to approximate the integrals (2), (3), and (4). For a finite collection of points $\{\mathbf{s}^*_j : j = 1,\ldots,m\}$ that cover B well, we can approximate the ACDF of the $Z(\cdot)$ process, $F_B(\cdot)$, by

$$\widehat{F}_B(z) \equiv \frac{1}{m}\sum_{j=1}^{m} F_{Z(s_j^*)}(z) = \frac{1}{m}\sum_{j=1}^{m} \Pr(\delta(s_j^*) \le [z - \mathbf{x}(s_j^*)'\boldsymbol{\beta}]). \tag{7}$$

Under Gaussianity, these probabilities are obtained from the Gaussian cumulative distribution function.

The expression (7) is then substituted into (4) to obtain $\widehat{w}_B(\cdot)$. Finally, to approximate (6), let $\{Z^{(1)}(\mathbf{s}^*), \ldots, Z^{(\ell)}(\mathbf{s}^*)\}$ be a random sample of size ℓ from $f_{Z(\mathbf{s}^*)}(\mathbf{z} \mid \mathbf{Z})$. Then we approximate $\widetilde{Z}_{wq}(s^*)$ by

$$\widehat{Z}_{wq}(\mathbf{s}^*) = \frac{\sum_{j=1}^{\ell} \widehat{w}_B(Z^{(j)}(\mathbf{s}^*)) \; Z^{(j)}(\mathbf{s}^*)}{\sum_{j=1}^{\ell} \widehat{w}_B(Z^{(j)}(\mathbf{s}^*))}. \tag{8}$$

All the equations above simplify somewhat if the error process, $\delta(\cdot)$, is Gaussian; in (7), the right-hand side is the average of Gaussian CDFs with parameter estimates plugged in, and the conditional distribution of $Z(\mathbf{s}^*)$ given $\boldsymbol{Z}$ (from which we simulate) is Gaussian.

3 Covariance matching constrained kriging (CMCK)

The universal kriging (UK) block predictor of $Z(A)$ is well known to be,

$$\widehat{Z}_{uk}(A) = \mathbf{x}(A)'\widehat{\boldsymbol{\beta}} + \mathbf{c}(A)'\Sigma^{-1}(\mathbf{Z} - X\widehat{\boldsymbol{\beta}}), \tag{9}$$

where $\Sigma = \mathrm{var}(\mathbf{Z})$; $\mathbf{c}(A) = (C(\mathbf{s}_1,A),\ldots,C(\mathbf{s}_n,A))'$ and $C(\mathbf{s},A) \equiv (\int_A C_{\boldsymbol{\theta}}(\mathbf{s}-\mathbf{u})\, d\mathbf{u}) / |A|$; $X \equiv (\mathbf{x}_j(\mathbf{s}_i))$ is an $n \times p$ matrix of explanatory variables; $\mathbf{x}(A) \equiv \int_A \mathbf{x}(\mathbf{u})\, d\mathbf{u} / |A|$; and $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, namely $\widehat{\boldsymbol{\beta}} \equiv (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{Z}$.

Now consider prediction of $g(Z(A))$, where g is a smooth nonlinear function and $|A| > 0$ or A is countable. Cressie (1993a) proposed the *constrained kriging (CK) predictor* $\hat{Z}_{ck}(A)$, which is an optimal linear predictor that is unbiased for $Z(A)$ (as is $\hat{Z}_{uk}(A)$) and satisfies the extra constraint, $\mathrm{var}(\hat{Z}_{ck}(A)) = \mathrm{var}(Z(A))$. Cressie (1993a) showed that $E(g(\hat{Z}_{ck}(A))) \simeq E(g(Z(A)))$.

In the spirit of wanting the predictor's statistical properties to match those of the target quantity, Aldworth and Cressie (2003) extended this methodology to include further constraints, where (some) covariances are matched in addition to the variance(s). This *covariance-matching constrained kriging (CMCK) predictor* $\hat{Z}_{cm}(\cdot)$ includes $\hat{Z}_{ck}(\cdot)$ as a special case.

Consider the problem of predicting the M-dimensional vector, $\mathbf{Z}(\mathbf{A}) \equiv (Z(A_1),\ldots,Z(A_M))'$, where $A_i \subset D$; $i = 1, \ldots, M$ and $\mathbf{A} \equiv \{A_1,\ldots,A_M\}$. The CMCK predictor, given by Aldworth and Cressie (2003), is:

$$\hat{\mathbf{Z}}_{cm} = X_M\hat{\boldsymbol{\beta}} + K'C'_M\Sigma^{-1}(\mathbf{Z} - X\hat{\boldsymbol{\beta}}), \tag{10}$$

where $X_M \equiv (\mathbf{x}_j(A_i))$ is an $M \times p$ matrix, $C_M \equiv (\mathbf{c}(A_1),\ldots,\mathbf{c}(A_M))$ is an $n \times M$ matrix, and K is an $M \times M$ matrix defined as follows. Write $P \equiv \mathrm{var}(\mathbf{Z}(\mathbf{A})) - \mathrm{var}(X_M\hat{\boldsymbol{\beta}})$ and $Q \equiv \mathrm{var}(\hat{\mathbf{Z}}_{uk}) - \mathrm{var}(X_M\hat{\boldsymbol{\beta}})$. Under the assumption that both of these $M \times M$ matrices are positive-definite, there exist nonsingular matrices P_1 and Q_1 such that $P=P'_1P_1$ and $Q=Q'_1Q_1$. Then $K \equiv Q^{-1}{}_1P_1$. Notice that the two constraints, $E(\hat{\mathbf{Z}}_{cm}) = E(\mathbf{Z}(\mathbf{A}))$ and $\mathrm{var}(\hat{\mathbf{Z}}_{cm}) = \mathrm{var}(\mathbf{Z}(\mathbf{A}))$, *are* satisfied; the latter constraint involves covariances as well as variances.

When the CMCK predictor is not defined (i.e., P or Q is not positive-definite), the covariance constraints can be relaxed until it is defined. Cressie and Johannesson (2001) took the approach of including the $(M - 1)$ nearest data locations to prediction region A, and then predicting $Z(A)$ using the implied $M(M+1)/2$ variance-covariance constraints. This is what we shall do in the next section on TCDD contamination. Also, in the next section, we shall detrend the data first and apply CMCK to the residual process $R(\cdot)$ with constant trend, namely $E(R(\cdot)) = \beta_0$. This amounts to putting $p = 1$ and $\mathbf{x}(\mathbf{s}) \equiv 1$ in the formulas above.

4 TCDD contamination in soil

4.1 Background

In environmental-remediation problems, soil contamination at one location often leads to contamination at other locations because of the conductivity properties of soil. This has been demonstrated by many case studies such as for the TCDD (tetrachlorodibenzo-p-dioxin) data, which were analyzed by Zirschky and Harris (1986) and Waller and Gotway (2004).

In 1971, a truck transporting dioxin-contaminated residues dumped an unknown quantity of waste in a rural area of Missouri in order to prevent citations for being overweight. Although the highest concentration occurred where the waste was dumped, contamination had spread to the shoulders of an adjacent state highway. In November 1983, The US Environmental Protection Agency (EPA) collected soil samples along transects and measured the TCDD concentration (in $\mu g/kg$) in each sample. Samples were composited along the transects. In our analysis, we consider only the data close to the dumping location, where the TCDD concentrations tend to be larger. (Recall that our goal is to estimate high extrema and their extent.) These data are based on 50-foot transects and there are no non-detects.

Thus, we consider measurements of 31 samples of TCDD within a region of D, a 458 × 78 square-foot rectangle. The direction parallel to the highway and the transects is defined to be the x-direction, and the y-direction is then perpendicular to the highway. The x-coordinates of the data are the x-values of the *start* of each transect. A plot of the data and their spatial locations is shown in Fig. 1a, where the decimal point represents the (x,y) coordinate of the data

4.2 Spatial Analysis

The TCDD data appear to be lognormally distributed; see Fig. 1b. Let $\{Z(\mathbf{s}_i) : i = 1,\ldots,31\}$ denote the observed log concentrations. Based on some exploratory plots and regression analysis, we considered the model, $Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$; $\mathbf{s} \in D$, for the log-concentration process, where $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, and $\mathbf{x}(\mathbf{s}) \equiv (1, y^2)'$ denotes a quadratic trend in the y-direction for $\mathbf{s} = (x, y)$. The ordinary least squares estimates of the regression coefficients are $\widehat{\beta}_0 = 2.6703$ and $\widehat{\beta}_1 = -0.0030$. Recall that the x-axis runs along the center of the highway; then we conclude that the quadratic surface of log concentrations in the y-direction is probably due to the drainage system along the highway, which is designed to let water run off the road quickly. See Fig. 1c.

We then removed the trend component and analyzed the spatial process generated by the residuals $\{R(\mathbf{s}_i) \equiv Z(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)'\widehat{\boldsymbol{\beta}} : i = 1,\ldots,31\}$. Histograms of the residuals show that normality is a fairly good approximation (given the small sample size); see Fig. 2a.

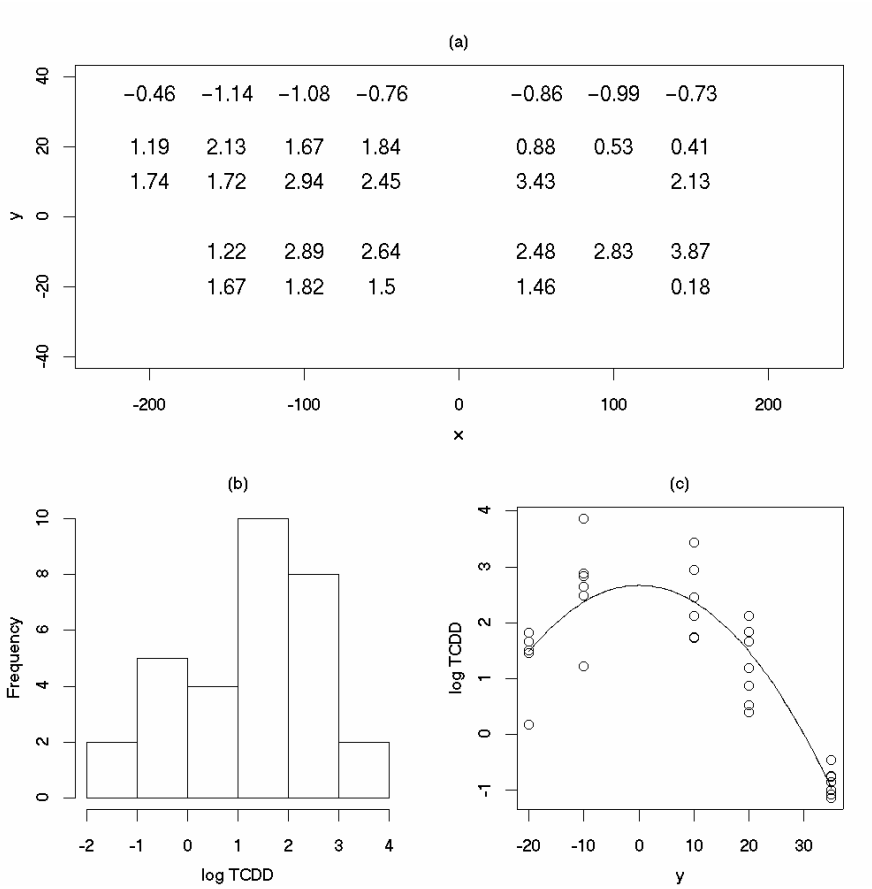


Fig. 1. The top panel displays the spatial location and value of the log TCDD concentrations; the decimal point represents the (x,y) coordinate of the data. The bottom left panel shows a histogram of the log TCDD concentrations. The bottom right panel displays the estimated trend for the log TCDD concentrations.

Directional empirical semivariograms of the residuals demonstrated that the residual process is severely anisotropic. To make the empirical semivariograms look isotropic, we multiplied the y-values by a factor of 7 (and the x-values were left unchanged). With this transformation, the data almost lie on a square coordinate system (the transformed y-values range between [-266, 280] and the x-values range between [-230, 228]). We then considered various semivariogram models for the residual process. Using the weighted-least-squares method (Cressie 1993b,

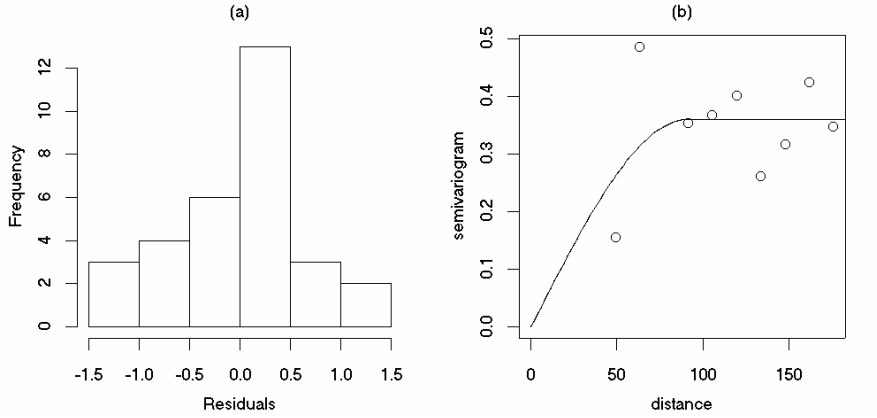


Fig. 2. The left panel displays a histogram of the residuals. The right panel shows the empirical semivariogram for the residuals; the solid line is the estimated spherical semivariogram $\gamma_{\hat{\boldsymbol{\theta}}}(h)$.

p. 99), we deemed that the spherical model fitted best. The general form for the covariance function derived from the isotropic spherical model is,

$$C_{\boldsymbol{\theta}}(h) = \begin{cases} c_0 I(h=0) + \sigma^2(1 - 1.5(h/r) + 0.5(h/r)^3); & |h| < r \\ 0; & \textit{otherwise}, \end{cases} \tag{11}$$

where h denotes the Euclidean distance on the transformed coordinate system and the vector of spatial parameters is $\boldsymbol{\theta} = (c_0, \sigma^2, r)'$. The estimate of the nugget effect, c_0, was 0; the estimate of the range, r, was 91.61; and the estimate of the partial sill, σ^2, was 0.36. Fig. 2b. shows the estimated semivariogram $\gamma_{\hat{\boldsymbol{\theta}}}(h)$. Notice that the estimated value $\hat{c}_0 = 0$ implies a stronger spatial dependence than is apparent from the empirical semivariogram values, also shown in Fig. 2b. The estimated covariance function is $C_{\hat{\boldsymbol{\theta}}}(h) = \hat{c}_0 + \hat{\sigma}^2 - \gamma_{\hat{\boldsymbol{\theta}}}(h)$.

4.3 Comparison of prediction methods

Based on the spatial analysis we carried out in Section 4.2, we now consider prediction of the log-concentration process. Our aim is to estimate the value and extent of high extremes of the log TCDD concentration process in a region B. For this analysis, we let $B \equiv D$, the rectangular region that encloses all the observed sites. We shall predict on a 2 × 2 ft. grid of points throughout D. We denote this discrete approximation to D by $\{\mathbf{s}^*_j : j = 1,\ldots,m = 9200\}$.

As our standard we use the IWQSEL predictor calibrated to predict the 90th percentile of the ACDF, which we shall compare to a CMCK predictor and an ordinary kriging predictor.

We estimated the ACDF of Z in B using Eq. 7. In the sigmoid importance function, the target value was set at $\alpha = 0.9$; then, for various values of C, we calculated the IWQSEL predictor (Eq. 6) based on a Monte-Carlo random sample of size $\ell = 2000$ from $f_{Z(\mathbf{s}^*_j)}(z \mid \mathbf{Z})$, for each prediction location $\mathbf{s}^*_j$, $j = 1, \ldots, m$=9200. For our statistical model,

$$Z(\mathbf{s}^*_j) \mid \mathbf{Z} \sim N(\omega(\mathbf{s}^*_j), \tau^2(\mathbf{s}^*_j)), \tag{12}$$

where (using the notation of Section 3), $\omega(\mathbf{s}^*_j) \equiv \mathbf{x}(\mathbf{s}^*_j)'\boldsymbol{\beta} + \mathbf{c}(\mathbf{s}^*_j)' \Sigma^{-1}(\mathbf{Z} - X\boldsymbol{\beta})$, and $\tau^2(\mathbf{s}^*_j) \equiv c_0 + \sigma^2 - \mathbf{c}(\mathbf{s}^*_j)' \Sigma^{-1}\mathbf{c}(\mathbf{s}^*_j)$. For this dataset, the value of C in $w(\cdot)$ that minimized the bias in estimating $F_B^{-1}(0.9)$ was C=4.75. To calculate all these quantities, we plugged in the parameter estimates, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$.

For the other two predictors, ordinary kriging plus trend, and CMCK plus trend, we first predicted the residual process $R(\cdot)$ using either ordinary kriging or CMCK and then we added the estimated trend component, $\hat{\mu}(\cdot) = \mathbf{x}(\cdot)'\hat{\boldsymbol{\beta}}$, to obtain predictions of the log-concentration process, $Z(\cdot)$.

We start by comparing the extent of shrinkage in the prediction of high extreme values in the three predictors. For a range of proportions p, from 0.85 to 0.99, we calculated the inverse ACDF, $F_B^{-1}(p)$. Then for each predictor we calculated the proportion of prediction locations whose predicted values were smaller than this inverse ACDF. For example, for the CMCK predictor we calculated $m^{-1}\sum_{j=1}^{m} I(\hat{Z}_{cm}(\mathbf{s}^*_j) \le F_B^{-1}(p))$, for each value of p. These numbers are summarized in Fig. 3, where a value above/below the 45° line indicates that the predictor underestimates/overestimates the inverse ACDF (i.e., denotes overshrinkage/undershrinkage). Thus, the ordinary-kriging-based predictor overshrinks prediction of extremes, whereas the CMCK-based predictor undershrinks prediction of extremes. By design, using a sigmoid importance function with $C = 4.75$ and $\alpha = 0.9$, the shrinkage at the 90th percentile of the inverse ACDF is just right for the IWQSEL predictor; and it experiences undershrinkage/overshrinkage for p below/above $p = \alpha = 0.9$.

We now use the three predictors to estimate the spatial extent of exceedances of the log-concentration process as follows. The exceedance set for the process $Z(\cdot)$ in the region B associated with an absolute threshold $t \in \mathbb{R}$ is defined by $e_B(t) = \{\mathbf{s} \in B : Z(\mathbf{s}) \ge t\}$, which we estimate by $\{\mathbf{s}^*_j : \overline{Z}(\mathbf{s}^*_j) \ge t\}$ for some generic predictor $\overline{Z}(\cdot)$; see Craigmile et al. (2004). The four panels of Fig. 4 display the exceedance sets obtained when we use ordinary kriging plus trend, the IWQSEL predictor, and CMCK plus trend with M=1 (i.e., zero neighbors and constraining only the variance of the predictor) and M=2 (i.e., one neighbor and constraining the variance and the covariance between nearest neighbors). In each case, we let the threshold be the 90th percentile of the inverse ACDF; that is, $t = F_B^{-1}(0.9)$.

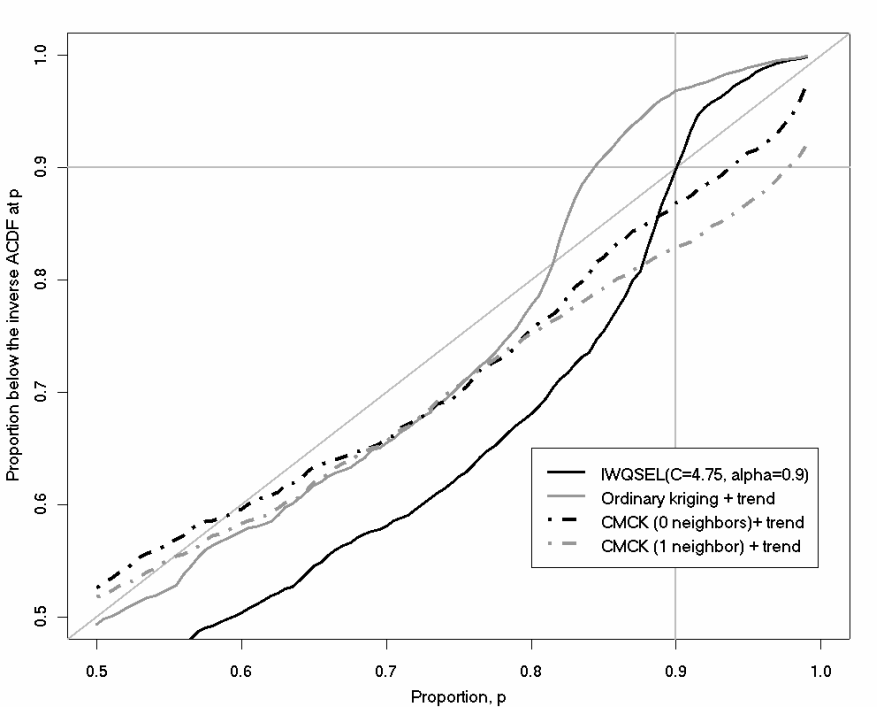


Fig. 3. Plots show the proportion of locations with predicted values smaller than the inverse ACDF evaluated at proportion p for ordinary kriging plus trend, the IWQSEL predictor, and CMCK plus trend. The 45^o line is shown for comparison. A predictor that predicts extremes well will be close to the 45^o line, for large values of p. A predictor that over(under)shrinks when predicting extremes, will be above (below) the 45^o line.

The spatial pattern of exceedances are qualitatively similar; there are two regions of exceedance centered at around $x \approx 100$ feet and there is at most one region of exceedance centered at $x \approx -100$ feet. Since the IWQSEL predictor is calibrated at the 90th percentile of the inverse ACDF, and hence we expect neither undershrinkage nor overshrinkage at that level, we use it as a standard to compare the exceedance sets based on ordinary kriging plus trend and CMCK plus trend.

Ordinary kriging underestimates the extent of the exceedance set because the predictor overshrinks prediction of the extreme values of the log-concentration process. For this predictor, 3.16% of the prediction locations are in the exceedance set, compared to 10.16% for the IWQSEL-based exceedance set. On the other hand, CMCK overestimates the extent of the exceedances; 13.18% (zero neighbors) or 17.09% (one neighbor) of its prediction locations are in the exceedance set. Comparing the points that are in the IWQSEL-based and CMCK-based exceedance sets, it is obvious that the exceedance set centered around the x-coordinate of $x \approx 100$ feet is wider for the CMCK predictor than for the IWQSEL predictor.

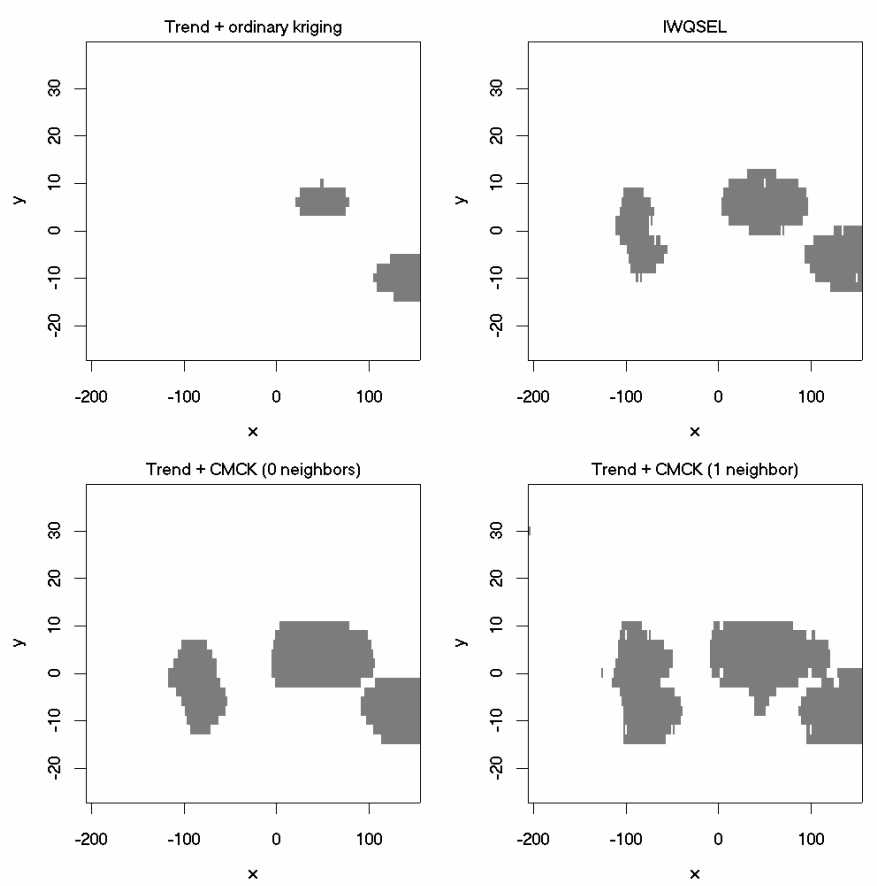


Fig. 4. The gray regions denote those locations with predicted values that exceed the 90th percentile of the inverse ACDF for ordinary kriging plus trend, the IWQSEL predictor, and CMCK plus trend.

5 Discussion

In this paper, we have calibrated the IWQSEL predictor to predict extreme values by choosing the importance function $w(\cdot)$ in the predictor to estimate the α^{th} quantile of the inverse ACDF well. In terms of this standard, we have demonstrated that CMCK undershrinks the predictions of large values of the process, which leads to an overestimation of the spatial extent of the exceedances of the log-concentration process. In practice, overestimating the exceedance set (region of remediation) increases the cost of remediation. However, this tends not to be as

consequential as the environmental impact when one underestimates the exceedance set, such as when ordinary (or universal) kriging is used.

Acknowledgments

This research was partially supported by the Office of Naval Research under grant no. N00014-02-1-0052. We would like to thank Jesse Frey for sharing his CMCK programs with us and Tom Santner for early discussions on geostatistics for extremes.

References

Aldworth J, Cressie N (2003) Prediction of nonlinear spatial functionals. Journal of Statistical Planning and Inference 112, 3-41.

Craigmile PF, Cressie N, Santner TJ, Rao Y (2004) Bayesian inferences on environmental exceedances and their spatial locations. Submitted for review

Cressie N (1993a) Aggregation in geostatistical problems. In A. Soares (ed.), Geostatistics Troia 1992, vol. 1, pp. 25-36. Kluwer Academic Publishers: Dordrecht

Cressie N (1993b) Statistics for Spatial Data (Revised edition). Wiley: New York

Cressie N, Johannesson G (2001) Kriging for cut-offs and other difficult problems. In Monestiez P, Allard D, Froidevaux R (eds.) geoENV III – Geostatistics for Environmental Applications . Kluwer: Dordrecht, 299-310

Deutsch CV, Journel AG (1992) GSLIB: Geostatistical Software Library and User's Guide, Oxford University Press: New York

Journel AG (1983) Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology 15, 445-468

Journel AG, CJ Huijbregts (1978) Mining Geostatistics. Academic Press: London

Lajaunie C (1990) Comparing some approximate methods for building local confidence intervals for predicting regionalized variables. Mathematical Geology 22, 123-144

Matheron G (1963) Principles of geostatistics. Economic Geology 58, 1246-1266

Matheron G (1976) A simple substitute for conditional expectation: the disjunctive kriging. In Guarascio M, David M, Huijbregts C (eds.), Advanced Geostatistics in the Mining Industry, pp. 221-236. Reidel: Dordrecht

Waller LA, Gotway CA(2004) Applied Spatial Analysis for Public Health Data. Wiley: New York

Zirschky J, Harris D (1986) Geostatistical analysis of hazardous waste site data. Journal of Environmental Engineering 112, 770-783.

Monitoring network optimisation using support vector machines

A. Pozdnoukhov[1] and M. Kanevski[2]

[1] IDIAP Research Institute, Martigny, Switzerland, e-mail: pozd@idiap.ch
[2] Faculty of Geosciences and Environment, University of Lausanne, Switzerland

1 Introduction

Monitoring network optimisation is a challenging problem for a number of real-life applications. This problem is closely related to the cost optimisation and is of particular importance for decision making process. Traditional optimisation of spatial data monitoring networks deals with geostatistics, which is a model-dependent approach based on analysis and modelling of spatial correlation structures. Network optimisation is performed by means of the analysis of the kriging/simulation variances.

Recently there has been an explosive growth in development of adaptive methods for learning from data. A number of very important problems of how to handle, understand and model the data if there are too many or too few of them were in the focus of developments. A family of data-driven and model-free contemporary approaches is based on Statistical Learning Theory, Vapnik-Chervonenkis theory (Vapnik 1998). Concerning spatially distributed data these learning methods predict unknown mapping between inputs (spatial co-ordinates and secondary variables) and outputs (random function) from available data and a priori knowledge. One of the most successful paradigms is called Support Vector Machines (SVMs). SVMs provide non-linear and robust solutions by mapping the input space into a higher-dimensional feature space using kernel functions. The strength of the method is that it attempts to minimise simultaneously the empirical risk of the training error and the structural risk (complexity of the model). SVM solutions depend on Support Vectors and not on statistics such as means and variances. Support Vectors (SVs) are the only data samples that influence the prediction and they are uniquely determined from data by solving a quadratic programming optimisation problem.

In the present paper, a new approach to the monitoring network optimisation based on SVM is proposed. The approach deals with categorical data, and the task is to improve the current monitoring network, similar to the task considered in Carrera (1984). The idea is to check new sampling sites as a potential Support Vectors. New Support Vectors have a priority to become sampling sites. The presented method can be improved by the recently developed solutions that incorporate confidence measures into SV-based models. The presented method

proposes an alternative for kriging-based approaches. It is clearly task-oriented and aims to improve the decision boundary directly, thus minimizing the testing error. Kriging-based methods are indirect in the sense that they basically improve the topology of the current network thus rely on improving the classification model.

A real case study deals with the monitoring network optimisation for modelling environmental categorical variables such as soil types.

2 SVM for spatial data analysis

Traditionally, geostatistics (statistics for spatial data) is one of the well-established approaches for working with spatially distributed data. Geostatistics, in general, is a model-dependent approach based on the exploratory analysis and modelling of spatial correlation structures (Kanevski and Maignan 2004).

On the other hand, data-driven and model-free contemporary approaches, based on Statistical Learning Theory, were successfully applied to a number of environmental problems. It should be noted that the challenges in learning from data (biocomputing, hyperspectral remote sensing images, data mining, etc.) have led to a revolution in the statistical sciences during the last decade.

In the early nineties a new learning paradigm emerged called Support Vector Machines (SVM). At first, it was proposed essentially for two-class classification problems (dichotomies), but later it has been generalised to multiclass classification problems, regression tasks, as well as to estimation of probability densities. Concerning spatially distributed data, the learning methods based on Support Vector Machines were applied to tasks, such as soil type classification, contamination level estimation, medium porosity prediction, contaminant concentration predictive mapping, etc (Kanevski *et al.* 2002). SV-based algorithms can be applied to modelling environmental phenomena at different spatial scales. Expert knowledge in the form of maps or confidence measures can be included in SVM models. Next, SV-based regression models have shown promising results in hybrid ML/geostat models for non-stationary multi-scale data (Pozdnoukhov 2002).

Generally, SVM provides non-linear and robust solutions by mapping the input space into a higher-dimensional feature space using kernel functions. By using different kernels one obtains learning machines analogous to well-known architectures such as Radial Basis Function neural networks and multilayer perceptrons. Thus, this method has the advantage of placing into the same framework some of the most widely used models such as linear and polynomial discriminating surfaces; feedforward neural networks; and networks composed of radial basis functions. In contrast to the Bayesian methods based on a modelling of the probability densities of each class, SVMs are focusing on the marginal data samples. SVMs provide the classification model directly, without solving a more general task of modelling the class densities at an intermediate step. SVMs provide sparse models, i.e. only a (small) subset of data possesses nonzero

weights. These data samples, called Support Vectors, usually lie close to the decision surface. They can be considered as a robust characteristic of the problem (given fixed model parameters). The basic facts of SVM theory and some properties of the SVs are considered in more details below.

2.1 Statistical learning theory

In Machine Learning one's aim is to find ("learn") an algorithm (modelling/mapping function) that describes training data and has good generalization abilities (that is, allows for accurate predictions at the new points, where the desired quantity is unknown). Statistical Learning Theory (SLT) is devoted to such problems of extracting knowledge from finite empirical data.

The following bounds of the generalization error were derived in SLT:

$$R(\lambda) \le R_{emp}(\lambda) + R_{conf}(\lambda), \tag{1}$$

where R is a bound of testing error, R_{emp} is an empirical risk on the training data (training error), and R_{conf} is a confidence term which depends on the "complexity" of the modelling function. The complexity can be controlled by the hyper-parameters λ of the modelling functions.

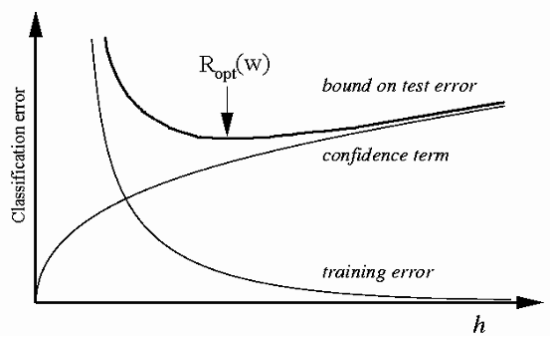


Fig. 1. Bound on test error derived in SLT. The minimum corresponds to some optimal complexity of the model for a given dataset.

The parameter that characterizes the «complexity» is called the VC-dimension of the modelling functions. It is denoted by h on Fig. 1. Hence the relevant strategy for constructing a learning machine is to minimize the training error while maintaining h small (see Fig. 1). This idea is realized for the specific learning tasks and results in a family of Support Vector algorithms.

2.2 Support Vector Classification

The SVM classification algorithm was initially derived for linear discriminating surfaces - hyper-planes. The criteria which controls the model complexity is the width of the margin between samples of different classes. It was proven that in order to minimize the model complexity one has to maximize the margin.

Given the discriminating hyper-plane $f(\boldsymbol{x},\boldsymbol{w},b) = \boldsymbol{w}\cdot\boldsymbol{x}+b$, the following optimisation problem can be formulated to minimize its complexity for given dataset $\{(\boldsymbol{x}_i, y_i), i=1,...L\}$:

$$\max \ \tfrac{1}{2}\|w\|^2 . \tag{2}$$

This should be done under the following constraints, which correspond to correct classification of the training data samples:

$$y_i[w \cdot x_i + b] \geq 1, \quad i = 1,..., L. \tag{3}$$

The algorithm can be extended to allow for training errors (i.e. misclassifying the training data, which is reasonable if the presence of noise in data is probable). Non-linear solutions emerge from applying the kernel trick – substituting the scalar product with kernel functions. These are symmetric positive-definite functions, which correspond to scalar product in some high-dimensional feature space.

The final formulation of the optimisation problem of SVM classification algorithm is presented below.

2.2.1 Optimisation Problem

Given a training set of pairs $\{(\boldsymbol{x}_i, y_i), i=1,...L\}$, non-linear SVM seek the decision function in the form:

$$f(x) = sign\left[\sum_{i=1}^{L} y_i \alpha_i K(x, x_i) + b\right], \tag{4}$$

where $K(.,.)$ is a symmetrical positive definite function – kernel (see also section 2.2.3). The weights α_i a are obtained from the solution of the convex QP optimisation problem:

$$\max_{\alpha} \quad \sum_{i=1}^{L} \alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{L} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{5}$$

with the following constraints:

$$\sum_{i=1}^{L} \alpha_i \cdot y_i = 0, \quad 0 \leq \alpha_i \leq C, \ \forall i . \tag{6}$$

It was observed by a number of researchers, that optimal solutions provided by SVM are often sparse. It means that a larger part of the weights are zero, while only the rest nonzero ones contribute to the decision function.

2.2.2 Support Vectors

Concerning the weights, the following cases are possible according to the Kuhn-Tucker conditions:

- If $\alpha_i = 0$, then $y_i f(x_i) \geq 1$
- If $C > \alpha_i > 0$, then $y_i f(x_i) = 1$
- If $\alpha_i = C$, then $y_i f(x_i) \leq 1$

The two major possibilities are: $\alpha_i = 0$ and $\alpha_i > 0$. Those training data samples that correspond to $\alpha_i > 0$ are called the *Support Vectors.* The Support Vectors with $C > \alpha_i > 0$ are the closest to the decision boundary. Notice, if we remove all other points except the SV from the training data set and train SVM on the SV only, we'll obtain the same decision boundary, i.e. SV have the determinant meaning for the given classification task. In particular, it gives us an opportunity to use the number of SV, their locations and corresponding weights as the criteria for the search for the locations where additional measurements would change (improve) the current model.

The meaning of parameter *C* has to be emphasized. This parameter is an upper bound for weights. It defines the trade-off between model complexity and allowance of training errors. If *C* is set to a sufficiently large value (infinity), the model is forced to discriminate the training data without errors. It can be a doubtful choice if the data are known to be noisy. Noisy data are often better modelled with values of C, which allow for training errors.

2.2.3 Kernel Function

The parameter(s) of the kernel are the hyper-parameter(s) of the support vector machine, and should be tuned using data and available knowledge. Kernel parameter(s) and the constant *C* are the only values that have to be provided by a user.

Gaussian Radial Basis Functions,

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \tag{7}$$

were found to be well suited for environmental applications such as predictive spatial mapping. Its bandwidth σ is proportional to some characteristic distance implied by the data. Anisotropic RBF can also be used.

Kernel parameters, as well as *C*, are usually tuned by minimizing cross-validation or testing error calculated on an independent set.

2.3 Multi-Class Classification

A generally used approach for multi-class classification task is to combine several binary classifiers. Given the basic SVM algorithm, we will apply the class-sensitive "one-to-rest" classification scheme.

2.3.1 One-to-Rest Scheme

A classifier (SVM) is trained for every subset of data of each class, considering all the rest samples of the samples as an opposite class regardless their true class membership. The final result is then obtained by:

$$y = \arg \max_{m} \sum_{i} \alpha_i^{(m)} y_i K^{(m)}(x, x_i) + b^{(m)} \quad (8)$$

where upper index m correspond to the m-th class. In other words, one simply compares the binary decision functions and takes one that provides the maximum output value for a tested sample. An advantage of this approach is the simplicity of realization and easy interpretable results. It was found that this scheme provides reasonable results for a number of environmental classification tasks. However, to obtain a flexible class-sensitive model, one has to tune m sets of parameters, one for each binary model.

A real case study involving the described approach is presented below.

2.3.2 Case Study: Soil Types

The real case study deals with the classification of soil types classification in the Briansk region of Russia. This is the most contaminated part of Russia due to Chernobyl radionuclides. Soil type data often accompanied the data on radionuclide activity.

Migration of radionuclides in soil depends on the properties of radionuclides, precipitation, etc. At the same time, the influence of soil type on ^{90}Sr radionuclide vertical migration was found to be a very important factor. High variability of environmental parameters and the multi-scale nature of initial fallout highly complicates the solution of the general modelling problem. Concerning soil types predictive mapping, geostatistical classification method – indicator kriging – fails since there are too few data in some classes to model the correlation structures adequately.

The original geographical coordinates were transformed to Lambert map projection and then linearly projected to (-1; 1) segment. All data and results are presented in this coordinate system. The training set consists of 310 measurements; the postplot of the data is presented in Fig. 2, left. Different colours represent different soil types. The postplot of SVM prediction mapping (Fig. 2, right) is accompanied with the validation data (500 samples), shown by crosses. Optimal mapping parameters were chosen according to the minima of cross-validation error.

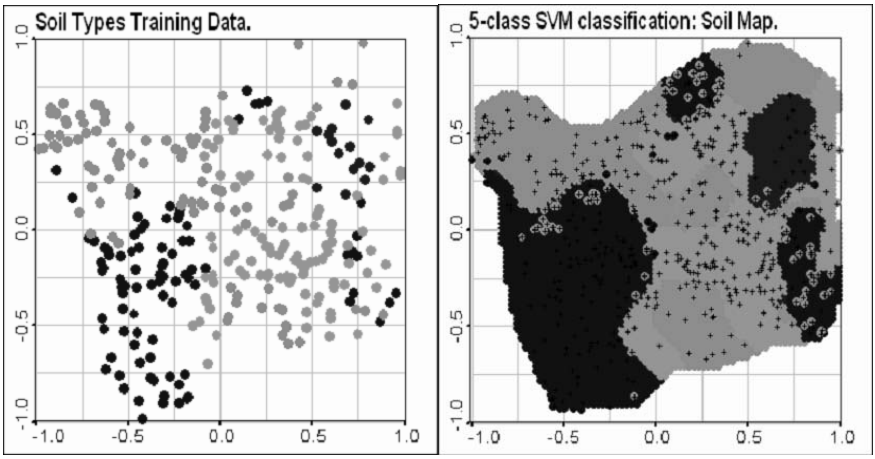


Fig. 2. Left: Training data for 5 types of soils. Right: SVM's solution for soil types classification problem. Validation data are shown by crosses.

Validation error over all classes (the whole number of misclassified validation samples divided by validation set size) obtained with SVM classifier is 12.8%. Other methods such as probabilistic neural network (PNN) and nearest neighbour classifier (NN) provided 18.2% and 17.8% of misclassified samples correspondingly.

3 SV Monitoring network optimisation

A network optimisation task can be approached in a number of ways, depending on the problem statement. We don't consider here the general and the hardest problem of designing a new monitoring network. Our task is to refine the current network to improve the classification model.

Simple geometrical approaches suggest analysis of the monitoring network by means, for example, of Voronoi polygons or other geometrical characteristics.

In geostatistics, network optimisation is performed by means of the analysis of the kriging/simulation variances (Carrera 1984). This method is highly dependent on a proper analysis and modelling of spatial correlation structures. It can be, however, awkward in the case of insufficient data or unobvious spatial structure. Things are even more complicated when one deals with categorical data.

Another geostatistical approach deals with stochastic simulation technique, as described in detail in Kyriakidis (1996).

Next, physics-based models can establish a foundation for monitoring network optimisation. For example, mass transfer models can provide tools to estimate the dependence of the results variability on the spatial locations of the measurements.

The SVM-based approach proposed in this paper is task-oriented in the sense that it directly explores whether the proposed spatial location would influence the

classification model, and how significant this influence can be. Next, since the baseline SVM algorithm is model-free and data-driven, the consequent advantage of the proposed method is its universality.

3.1 The proposed SV approach

The proposed method of monitoring network optimisation is based on the Support Vectors' properties. Suppose one is given a training set and a set of possible locations for taking the additional measurements. These spatial locations can be specified by an expert who takes into account environmental conditions, the significance of prediction accuracy in different sub-regions, etc. Otherwise (if the whole region is investigated) one can consider a dense grid covering the entire region as the set for exploration.

Given a new location for prospective measurement, one includes it into the current model. Two possible labels are consecutively assigned to the sample and the model's weights are updated. The update procedure can be organised in such a way so as to avoid the complete re-solving the optimisation problem, Eq. 3 - 4. If the new measurement obtains zero weight and is not a SV, it doesn't contribute to the prediction model and is somehow "useless". On the other hand, a sample that becomes a SV is of particular importance to the task since it defines the decision function.

The main steps of the method are:

- Take one sample from the examined set; assign it a "positive" label.
- Update the model on the extended training set (with the added "positive" sample).
- Store the weight that the sample obtained as a result of updating the current SV model, then remove the sample.
- Assign the "negative" label to the specified sample.
- Update the model on the extended training set (with the added "negative" sample).
- Store the weight the sample obtained as a result of updating the current SV model, then remove the sample.

- Repeat all the previous steps for all the examined samples

At the output of this scheme we are given two weights for every examined sample: α^+_k and α^-_k, according to the possible labelling of the point. The following cases are possible:

1) $\alpha^+_k = 0$, $\alpha^-_k > 0$. The sample is not a SV when labelled as positive and is a SV when negative. Note that α^-_k might be equal to *C,* which is an upper limit for the weights.
2) $\alpha^+_k > 0$, $\alpha^-_k = 0$. The sample is a SV when labelled as positive and is not a SV when negative. Note that α^+_k might be equal to *C.*

3) $\alpha^+_k>0$, $\alpha^-_k>0$. The sample is a SV while assigned both to positive and negative labels.

Let's take into account that two types of SV are possible: boundary SV ($\alpha=C$) and ordinary SV ($0<\alpha<C$). If the sample becomes a boundary SV for either labelling, its location is out of our interest. The reason comes from the meaning of C parameter mentioned in Section 2.2.2 and can be expressed as follows: the samples with limit weights are either mislabelled or too atypical and can be considered as noise. In the presented scheme those samples that belong to the third case are the points of our interest – these are the desired locations of potential additional measurements.

In fact, the scheme is simplified since it considers samples one by one and all the mutual interactions are being neglected. One can try considering the samples in an ensemble or apply some prior knowledge to overcome this difficulty. The problem vanishes if the measurements are taken consecutively and true labels become known step by step. However, the presented scheme is preferable in the sense of low computational time it takes.

Another important question is how to rank by significance those samples, that were found to become SVs. The magnitude of corresponding weight is not really a true significance measure. However, as the value of α determines some influence of the corresponding sample on the model, we can consider some heuristic values based on α's, such as a sum of α^+ and α^-. Recent results on incorporating confidence measures into SV models can be applied for deriving the desired ranking criteria.

In conclusion, let's mention that the presented scheme also provides a way to remove the unnecessary (inefficient) sites. These are basically the sites, that obtain zero weights according to the SVM model.

4 Case study: soil types

An example of SVM application for multi-class classification of the soil types was presented before in Section 2.2. Suppose that one particular class (class 3, see Fig. 2) correspond to the soil type, which is of crucial importance for radionuclide migration modelling. The task is to improve the current model for selected class given the possibility of obtaining a small number of additional measurements. Current model parameters for class 3 model are: C=100, σ=0.19. The parameters were tuned according to the minima of cross-validation error. The training error is 0%, and validation error for class 3 is 5.8%.

4.1 Importance level mapping

The following results were obtained after applying the algorithms over a dense regular square grid that covers the entire region. The grid contains 4321 points.

SVM prediction map obtained with the current model of soil type 3 (shown in dark) is presented in Fig. 3 (left). Fig. 3 (right), presents a postplot of the proposed indirect "importance" measure discussed in Sect. 3.1.

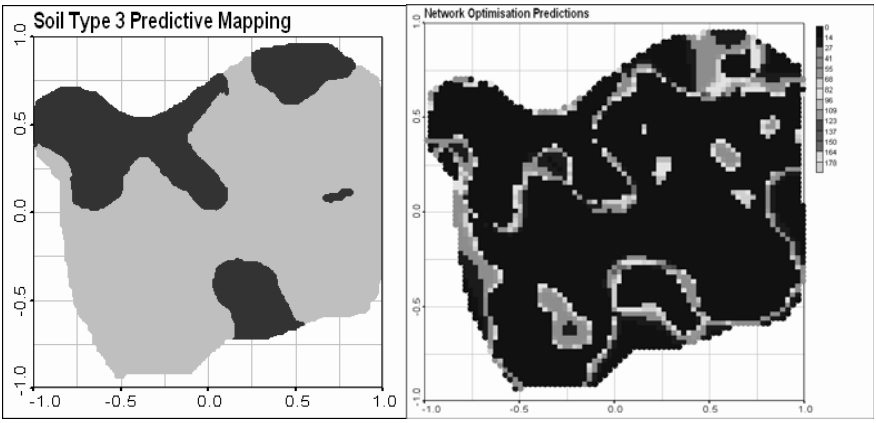


Fig. 3. Left: SVM predictive mapping for class 3. Right: Network Optimisation Results, sum of weights ($\alpha^- + \alpha^+$) is plotted as indirect "importance" measure.

As expected, the regions that are close to decision surface are of most importance for modifying the monitoring grid. At the same time, some regions with small amounts of data are also taken into consideration, as well as regions close to the region border.

4.2 Network optimisation efficiency

The following scheme is proposed to control and illustrate the efficiency of the method. Suppose we are given a set of measurements (310 samples), and a number of spatial locations where some additional (potentially expensive and uneasy) measurements can be carried out (500 samples). Our task is to choose those that would improve the current classification model most efficiently. We will control the model's performance using independent validation set of 500 samples, reserved beforehand. The graph on Fig. 4 presents validation error obtained by using two extended training sets. The first set was supplied by additional samples advised by network optimisation algorithm, where the samples were added according to their importance level. The other training set was extended with the same number of randomly selected samples. The "randomly" taken samples are not absolutely random, however. They are measurements taken from real monitoring network, hence they can be used for providing a correct comparison.

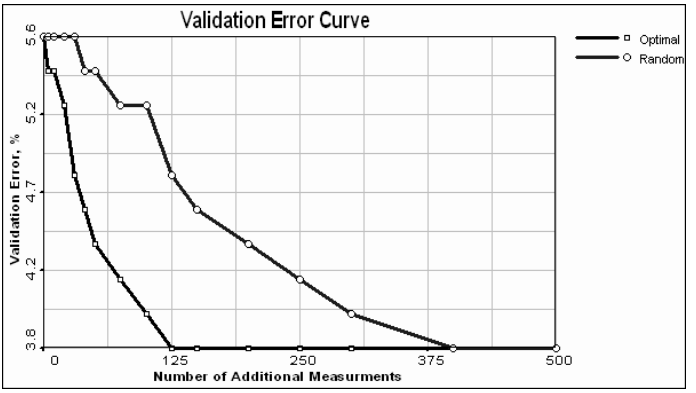


Fig. 4. Evolution of the validation error while including additional measurements.

The presented example illustrates that more then 60% of soil probes taken were "useless" in the sense that they do not give any improvement to SVM classification model for the considered soil type. The result also confirms the assumption of the exceptional significance of Support Vectors, which initiated the presented monitoring network optimisation algorithm.

5 Conclusions

Machine Learning opens promising perspectives for approaching the tasks of monitoring network optimisation. This paper presents a novel method for monitoring network optimisation, which can be used to increase the accuracy of classification models by taking a small number of additional measurements. This problem statement is common for the fields of modelling the hydro-geologic units, reservoir modelling, environmental monitoring, etc.

The method is based on the recent ML technique known as Support Vector Machine. The method is problem-oriented in the sense that it directly answers the question of whether the advised spatial location is important for the classification model. However, the question of ranking the samples according to their "importance" still has to be investigated. Similar ideas involving Support Vector Regression models can be applied for continuous data analysis. Next, hybrid ML\geostat models (Kanevski 2002) offer approaches for further extensions of the proposed SV-based algorithm. Application of geostatistically adjusted kernels can improve both the efficiency and interpretability of the approach proposed.

Further research deals with comprehensive comparisons of the proposed algorithms with different geostatistical approaches for the optimization of monitoring networks, elaboration of the SVM-based approach to multi-class classification and regression problems.

References

Carrera J, Usunoff E, Szidarovsky F (1984) A method for optimal observation network design for groundwater management. Journal of Hydrology, V.73, 147-163

Kanevski M, Maiganan M (2004) Analysis and Modelling of Spatial Environmental Data. EPFL Press, Lausanne

Kanevski M, Pozdnoukhov A, Canu S, Maignan M (2002) Advanced Spatial Data Analysis and Modelling with Support Vector Machines. International Journal of Fuzzy Systems, Vol. 4, No. 1, March 2002, 606-616

Kanevski M, Parkin R, Pozdnukhov A, Timonin V, Maignan M, Yatsalo B, Canu S (2002) Environmental Data Mining and Modelling Based on Machine Learning Algorithms and Geostatistics. iEMSs2002, Lugano, Switzerland, 414-419

Kyriakidis PC (1996) Selecting Soils for Remediation in Contaminated Soils Via Stochastic Imaging, in Proceedings of the Fifth International Geostatistics Congress, September 22-27, Wolongong, Australia

Pozdnukhov A, Kanevski M, Maignan M, Canu S (2002) Robust mapping of spatial data with Support Vector Regression. Preprint IBRAE-2002. Nuclear Safety Institute RAS, p. 15

Savelieva E, Kanevski M, Timonin V, Pozdnukhov A, Murray C, Scheibe T, Xie Y, Thorne P, Cole C (2002) "Uncertainty in the hydrogeologic structure modeling" In the proceedings of IAMG'02, September 15-20, Berlin, Germany

Vapnik V (1998) Statistical Learning Theory. New York: John Wiley & Sons.

Bayesian Kriging with lognormal data and uncertain variogram parameters

J. Pilz, P. Pluch and G. Spöck

University of Klagenfurt, Austria

1 Introduction

The usual proceeding in geostatistical analyses is to assume the model contents as known, i.e. to fix the trend function, the variogram model and the distribution function of the data. For example, the trend is modeled by a low order polynomial (usually only by some constant), the variogram is chosen to be spherical, exponential or Gaussian, and the data are assumed to be normally distributed. The last assumption is often not explicitly stated, but implicit in the model.

First attempts to geostatistical modeling with non-normal data are made in Diggle *et al.* (1998) and De Oliveira *et al.* (1997), in the latter paper Box-Cox-transformations are considered. In our paper we start from lognormal data; the logarithmic transformation is a special case of the Box-Cox-transformation. Assuming the data, after an appropriate transformation, to be normally distributed then the quality of the results of geostatistical (Kriging) interpolation crucially depends on the underlying variogram. Instead of a "true" variogram we only have an empirical estimate of it and the usual practice is then "Plug-in-Kriging", i.e. Kriging on the basis of the estimated variogram. Hence, the plug-in-predictor is only an estimate of the unknown Kriging predictor; the claimed BLUP-optimality (BLUP = best linear unbiased predictor) is therefore no longer valid. The plug-in-predictor, also called EBLUP = empirical BLUP by Stein (1999), is neither unbiased nor linear. The characterization as "best" predictor (in the sense of minimum variance) is thus more than questionable. With the exception of some asymptotic results in Stein (1999), very little is known about the statistical properties of the EBLUP. It is already known from Christensen (1991) that the actual mean squared error of prediction (MSEP) may be much larger than the theoretical one resulting from the assumption that the variogram has been specified exactly.

Pilz *et al.* (1997) proposed a minimax approach to prediction which takes account of the uncertainty with respect to variogram choice and estimation. Instead of focusing on a simple estimated variogram, this approach starts by admitting a whole class of plausible (nonparametric) variogram functions and then looks for a predictor which leads to a minimum MSEP in the "worst case". The numerical computation of the minimax predictor is, however, a rather complex and computationally intensive task. In this paper we present an alternative Bayesian approach to prediction which models the uncertainty by means of suitable (posterior) prob-

ability distributions for the parameters of a flexible class of (nested) Matèrn and Gaudard variogram functions, respectively.

2 The Model

Consider a lognormal random function $Y(x)$ such that

$$Z(x) = \log Y(x) = m(x) + \varepsilon(x); \quad x \in D \subset R^d, d > 1 \tag{1}$$

where $m(x)$ is the trend function and $\varepsilon(x)$ a (nonobservable) random error term with expectation zero for all points *x* from a given region *D*. Usually, a linear regression setup is chosen to model the trend function,

$$m(x) = \beta_1 f_1(x) + \ldots + \beta_r f_r(x) = f(x)^T \beta \tag{2}$$

with given functions $f_1,\ldots,f_r$ and unknown regression parameter vector $\beta = (\beta_1,\ldots,\beta_r)^T$, e.g. a low-order polynomial. We assume the transformed observations to be covariance-stationary, i.e.

$$Cov(Z(x+h), Z(x)) = C(h) \tag{3}$$

for all $x, x+h \in D$, where $C(\cdot)$ is the covariance function. Clearly, $C(\cdot)$ must be positive semidefinite. Having available observations at *n* points $x_1,\ldots,x_n \in D$, it is well-known that the best linear unbiased predictor (BLUP) for $Z(x_0)$ at an unobserved location $x_0 \in D$ takes the form

$$\hat{Z}_{UK}(x_0) = \hat{m}(x_o) + k_0^T K^{-1}(Z - F\hat{\beta}) \tag{4}$$

where

$$k_o = (C(x_o - x_1),\ldots,C(x_o - x_n))^T, Z = (Z(x_1),\ldots,Z(x_n))^T,$$
$$F = (f_j(x_i))_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}}, \quad K = (C(x_i - x_j))_{i,j=1,\ldots,n}$$

and $\hat{m}(x_0) = f(x_0)^T \hat{\beta}$ stands for the estimated trend at x_0 with $\hat{\beta} = (F^T K^{-1} F)^{-1} F^T K^{-1} Z$, the generalized least squares estimator of β. The BLUP (4) is known as the universal Kriging predictor, see e.g. Cressie (1993). A Bayesian analogue to this predictor has been presented by Omre and Halvorsen (1987). Their predictor assumes prior knowledge about the trend parameter vector β such that

$$E\beta = \mu \text{ and } Cov(\beta) = \Phi \tag{5}$$

i.e. prior knowledge allows the specification of the mean μ and covariance matrix Φ of β. The Bayes Kriging predictor then reads

$$\hat{Z}_{BK}(x_0) = f(x_0)^T \mu + (k_0 + F\Phi f(x_0))^T (K + F\Phi F^T)^{-1}(Z - F\mu) \quad (6)$$

this predictor builds the bridge between simple Kriging (corresponding to $\Phi = 0$, i.e. "perfect" knowledge of the trend) and universal Kriging (corresponding to $\Phi^{-1} = 0$, i.e. "nothing" is known a priori about the trend).

Omre and Halvrsen (1987) show that $\hat{Z}_{BK}(x_0)$ minimizes the total MSEP, i.e. the MSE averaged with respect to the trend parameter β.

The weak point of both universal Kriging and Bayes universal Kriging as presented by Omre and Halvorsen (1987), as well as of many other spatial interpolation methods, lies in the fact that the covariance function and thus k_0 and K must be known exactly to guarantee the BLUP-optimality. In practice, however, a plug-in-Kriging is performed, i.e. the unknown covariance (or variogram) function is estimated empirically and then fitted to some covariance model. This means, k_0 and K in Eq. 4 are replaced by empirical estimates $\tilde{k}_0$ and $\tilde{K}$, usually based on the empirical moment estimator for the related variogram function. The resulting plug-in-predictors $\tilde{Z}_{UK}(x_0)$ and $\tilde{Z}_{BK}(x_0)$, where e.g.

$$\tilde{Z}_{UK}(x_0) = f(x_0)^T \tilde{\beta} + \tilde{k}_0^T \tilde{K}^{-1}(Z - F\tilde{\beta})$$

with $\tilde{\beta} = (F^T \tilde{K}^{-1} F)^{-1} F^T \tilde{K}^{-1} Z$, are then no longer linear (in Z), since $\tilde{k}_0$ and $\tilde{K}$ depend on Z in a complicated non-linear manner. Also, unbiasedness no longer holds, which implies that the MSEP no longer coincides with the variance of prediction and an additional (squared) bias term occurs:

$$\text{MSEP}(\tilde{Z}_{\text{UK}}) = \text{Var}\left\{\tilde{Z}_{UK}(x_0) - Z(x_0)\right\} + \left\{E(\tilde{Z}_{UK}(x_0) - Z(x_0))\right\}^2 \quad (7)$$

Hence, the BLUP-optimality is completely lost.

Thus, it is not only necessary to study the empirical MSEP (7) and the consequences of a misspecification of the covariance function $C(\cdot)$ but also to develop "robust" alternatives to the universal and Bayes universal predictors, respectively.

Stein (1999) points out that this requires further elaboration of the model and the underlying distribution law, in particular it is important to model the local behaviour of $Z(\cdot)$ sufficiently well and in a flexible way. The degree of "smoothness" of the random function is of primary importance for the MSEP of the plug-in-predictor, which, in turn, is determined by the analytical properties of the covariance function, especially by its behaviour near the origin. To this purpose, we will consider here two such variogram models: the Matérn variogram and the less known (convex) combined exponential-Gaussian variogram introduced by Gaudard *et al.* (1999).

Before we develop a Bayesian alternative to the plug-in (Bayes) universal predictors, which takes account of variogram model uncertainty, we will briefly summarize some of the sources of this uncertainty.

3 Sources of variogram uncertainty

3.1 Uncertainty due to model assumptions

Implicit in all geostatistical considerations is the assumption of ergodicity of the random function to be considered. This assumption cannot be verified, however. The covariance function cannot be determined completely from the knowledge of paths in a finite region, see e.g. Cressie (1993) for a general discussion on this issue. The situation becomes still worse when we have only few data from a single realization of the random function, which is but the usual case in practical applications.

Moreover, there is the dilemma with the choice of scale, i.e. we have to decide which part of the observed variability is to be attributed to the global scale (trend) and to the local scale (random fluctuations), respectively. In the geostatistical literature, there exists a multitude of proposals for detrending and ensuing modeling of residuals, ranging from e.g. classical trend surface analysis followed by ordinary Kriging of the residuals, universal, median-polish and IRF-Kriging procedures to quite sophisticated nonparametric and local smoothing techniques, again followed by Kriging of the residuals.

A further source of uncertainty results from the various approaches to modeling of anisotropy. Very often directional variograms exhibit different behaviour of the random function in specific directions, but due to insufficient data there are doubts on the reliability of the variograms.

3.2 Different estimation and fitting methods

There are numerous methods for empirical estimation of covariance functions and variograms, respectively, which are all in common use. Besides the well-known moment estimator there exist robust estimation versions (see Cressie (1993), p. 74) and estimators based on variogram clouds. But with all these different methods we are faced with the problems of choosing appropriate group sizes, lag classes, the maximum lag, etc.

The next source of uncertainty results from the process of fitting the empirical variogram to some theoretical variogram model. Usually, this is done on a subjective basis. Apart from stationarity and isotropy assumptions, and even if we confine ourselves to the few well-known parametric variogram models, we still have to decide on the “correct” type of variogram, about possible nested structures and then on the specification of the variogram parameters such as nugget, sill and range. After having chosen a theoretical model, we are again faced with a great variety of parameter estimation methods, e.g. (weighted) least squares, generalized least squares, maximum-likelihood and REML-methods, Bayesian methods, and, finally, classical estimation methods for variance components such as MINQUE or MVUE, see e.g. Mardia and Mashall (1984) and Zimmermann (1989). The misspecification of the model type and model parameters may have dramatic effects

on the prediction. In particular, a critical evaluation of the choice of the nugget effect is necessary, since this requires extrapolation of the variogram into the origin.

3.3 How to specify uncertainty about the covariance structure?

From the above discussion of the various sources of uncertainty about the covariance structure it should be natural to start a geostatistical prediction task with a whole class of plausible covariance functions instead of focusing on a single estimated covariance function, then proceeding with the usual Kriging "apparatus" and, finally, ending up with "nice" but unrealistic smooth Kriging error maps. This way we will also be able to cope with artefacts which arise from fitting with insufficient data.

Some proposals for specifying a sufficiently flexible class of plausible covariance functions may be found in Pilz *et al.* (1997). The approach chosen there is nonparametric and based on spectral decompositions of covariance functions as given e.g. in Yaglom (1986). In the sequel we develop an alternative approach, which is more intuitive and starts from parametric covariance function models; where the uncertainty about the covariance parameters is modeled on the basis of conditional simulation.

4 Flexible parametric classes of covariance functions

Since it was first mentioned in the monograph of Matérn (1986), the four-parameter variogram

$$\gamma_M(h;\theta) = c_0 + c[1-(a\,|\,h\,|)^{\nu} K_{\nu}(a\,|\,h\,|)] \tag{8}$$

named after him has gained increasing attention in geostatistical research and applications. Here $\theta = (c_0, c, a, \nu)$ denotes the vector of covariance parameters, where $c_0 \geq 0$ (nugget effect) and $c, a, \nu > 0$, K_{ν} is the modified Bessel function of order ν. Important special cases include the exponential variogram $(\nu = 0.5)$, Whittle's variogram $(\nu = 1)$, which is widely used in hydrology, and the Gaussian variogram, which results as a limiting case when $\nu \to \infty$.

The parameter ν in (7) is referred to as smoothness parameter, the integer part $k = [\nu]$ indicates the order of differentiability of the random function (in the mean square sense).

As a simple alternative we further consider the variogram

$$\gamma_{Gd}(h;\theta) = c_0 + c\left[1-(1-\nu)e^{-\alpha|h|} - \nu e^{-\alpha|h|^2}\right] \tag{9}$$

$$\theta = (c_0, c, a, \nu); \quad 0 \leq \nu \leq 1$$

which has been proposed by Gaudard *et al.* (1999).

This is a convex combination of the exponential variogram (set $\nu = 0$) and the Guassian variogram (set $\nu = 1$); thus the Gaudard variogram builds the "bridge" between linear and parabolic behaviour of the variogram near the origin.

5 The full Bayesian approach

The Bayesian apparatus provides a general methodology for taking account of uncertainty w.r.t. the model and its components. This is particularly important for the specification of the variogram model and its parameters. The four-parameter classes introduced in the previous section seem very promising, since, on the one hand, they offer more flexibility than the tree-parameter models currently used, and, on the other hand, the number of parameters is still small and manageable and these allow a simple and natural interpretation.

Let denote $\gamma(h;\theta)$ with $\gamma \in \{\gamma_M, \gamma_{Gd}\}$ and $\theta = (c_0, c, a, \nu)$ such a four-parametric variogram where $\nu \in (0,\infty)$ in case of $\gamma = \gamma_M$ and $\nu \in [0,1]$ in case of the Gaudard variogram.

The full Bayesian approach requires a completely specified distributional model, i.e. besides the distribution of the data, displayed by the likelihood function, we need to specify probability distributions for the trend parameter β and the covariance (variogram) parameter vector θ. We get a reward, however, for the bigger efforts necessary for a Bayesian modeling: first, we may handle more flexible distributional models for the data (not only normally or lognormally distributed data), second, the uncertainties with respect to the model parameters can be modeled by appropriate (prior or posterior) probability distributions and, finally, the so-called *predictive density* offered by the Bayesian paradigm gives us a complete probability distribution for the predicted values, not only expected (kriged) values and (Kriging) variances.

The Bayes-optimal prediction of $Z_0 := Z(x_0)$ at an unobserved location $x_0 \in D$ is based on the predictive density

$$p(Z_0 \mid Z) = \int_{\Theta} \int_{B} p(Z_0 \mid \beta, \theta, Z)\, p(\beta, \theta \mid Z)\, d\beta\, d\theta \tag{10}$$

This is the conditional probability density of Z_0 for given data $Z = \left(Z(x_1), \ldots, Z(x_n)\right)^T$, averaged over all trend parameters $\beta = (\beta_1, \ldots, \beta_r) \in B \subseteq R^r$ and variogram parameters $\theta \in \Theta \subset R^4$, where the averaging is done with respect to the posterior probability density $p(\beta, \theta \mid Z)$ of β, θ for given data *Z*. Here $B \subseteq R^r$ and $\Theta \subset R^4$ denote the regions of possible trend and variogram parameters, respectively.

In case of the Matérn variogram model we have $\Theta = [0,\infty) x (0,\infty) x (0,\infty) x (0,\infty)$ and in case of the Gaudard model

$\Theta = [0,\infty)x(0,\infty)x(0,\infty)x[0,1]$. The first factor under the integral sign in Eq. 10 represents the conditional distribution of Z_0 for given parameters β, θ and data *Z*. The second factor is, as mentioned above, the posterior probability density of the parameters, which can be obtained according to Bayes's theorem as

$$p(\beta,\theta \mid Z) = \frac{p(Z \mid \beta,\theta)\, p(\beta,\theta)}{\int_{\theta}\int_{B} p(Z \mid \beta,\theta)\, p(\beta,\theta)\, d\beta d\theta} \tag{11}$$

where $p(Z \mid \beta,\theta)$ is just the likelihood function of the data and $p(\beta,\theta)$ stands for the prior probability density of the parameters.

Since $Z(\cdot)$ is a Gaussian random function, the observation vector *Z* follows an n-dimensional normal distribution with expectation $F\beta$ and covariance matrix $K(\theta)$ having elements $K_{ij}(\theta) = c_0 + c - \gamma(x_i - x_j;\theta);\ i, j = 1,\ldots,n$. Hence, the log-likelihood function reads

$$\log p(Z \mid \beta,\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det[K(\theta)] - \frac{1}{2}(Z - F\beta)^T K(\theta)^{-1}(Z - F\beta)$$

The density $p(Z_0 \mid \beta,\theta,Z)$ occurring in Eq. 10 is then the density of the univariate normal distribution $N(\mu,\sigma^2)$ where $\mu = f(x_0)^T\beta$ and $\sigma^2 = c_0 + c - k_0^T K(\theta)^{-1} k_0$ with $k_0 = (c_0 + c - \gamma(x_0 - x_1;\theta),\ldots,c_0 + c - \gamma(x_0 - x_n;\theta))^T$, where, again, we have used the well-known relationship $\gamma(h) = C(0) - C(h)$ between the variogram and the covariance function.

The only missing "ingredient" to compute the predictive density (Eq. 10) is then the prior probability density $p(\beta,\theta)$. In the literature, the Bayesian approaches to geostatistical prediction assume that the trend parameter β and the variogram parameter θ are independent a-priori, i.e. $p(\beta,\theta) = p(\beta)\cdot p(\theta)$.

Usually, $p(\beta)$ is assumed to be *locally uniform*, i.e. $p(\beta) \equiv 1$ for all $\beta \in B = R^r$. Of course, this is not a proper probability density, since the integral $\int p(\beta)d\beta$ over *B* does not converge. Such densities are often called *noninformative* prior densities; they are simply used to model the situation of total ignorance, where nothing is known a-priori about β. The fact that the noninformative prior density $p(\beta) \equiv 1$ is not proper does not cause problems as long as the resulting posterior density (after applying Bayes's theorem) becomes a proper density. A thorough discussion on the use of noniformative prior densities and the situations under which they lead to proper posterior densities may be found in Berger *et al.* (2001).

We argue that usually we know "more" about the trend than just assuming that any parameter β is equally likely to occur. E.g. we know that the expected level of

radioactivity in a given region is nonnegative and does not exceed a certain threshold. Such type of relatively weak prior knowledge implies, in turn, some restriction on β. Proposals on how to "transform" different types of prior knowledge about the trend into "adequate" and proper probability densities may be found in Pilz (1994), Pilz *et al.* (1997) and Dubois *et al.* (2000). A good part of the proposals in the literature just cited has already been implemented in the statistics program system "R", work in this direction is ongoing (see e.g. Gebhardt (2004)).

Modelling of adequate prior distributions for the variogram paramter vector θ is a much more comlex and difficult task. For the Matérn variogram, Handcock and Wallis (1994) and Qian (1997) proposed to use

$$p(\theta) = (c_0 \cdot c)^{-1}(1+a)^{-2}(1+\nu)^{-2}, \qquad \text{for } c_0, c, a, \nu > 0$$

This is an improper density, too; moreover, it ignores dependencies among the nugget, sill, range and smoothness components. Non-informative priors for the Matérn variogram parameter θ based on Jeffreys's (invariance) rule have been derived in Berger *et al.* (2001). Other "automatic" solutions, as presented e.g. in Cui *et al.* (1995), are non-satisfactory as well.

Our objection against all such "non-informative" and automatic" priors is that existing a-priori knowledge should be used to model proper priors for the trend and to take into account the inherent uncertainty about θ by specifying an informative prior or posterior distribution for the variogram parameter as well. Some results on the distribution of the estimated covariance function can be found in Cressie (1993).

6 Our proposal

We propose to avoid the cumbersome task of specifying a prior density $p(\theta)$ for the variogram parameters, and thus also to avoid the dangers of a possible misspecification, and let the data "speak" themselves about the inherent uncertainty: instead of $p(\theta)$ we generate the posterior density $p(\theta \mid Z)$ via conditional simulation. This is then used to factor the joint posterior density of trend and variogram parameters according to

$$\begin{aligned} p(\beta, \theta \mid Z) &= p(\beta \mid \theta, Z) * p(\theta \mid Z) \\ &= k \cdot p(Z \mid \beta, \theta) \cdot p(\beta \mid \theta) \cdot p(\theta \mid Z) \end{aligned} \tag{12}$$

where *k* denotes the normalization constant.

The posterior density $p(\theta \mid Z)$ of the variogram parameters is computed according to the following algorithm:

(A1) Generate N (= 5000, say) simulated data sets from the random function $Z(\cdot)$, conditional on the actual observations $Z = (Z(x_1), \ldots, Z(x_n))^T$ and based on the (Matérn or Gaudard) variogram γ fitted to the usual empirical variogram of the data. As a result we obtain *N* new empirical variograms.

(A2) The new empirical variograms are fitted to the chosen (Matérn or Gaudard) variogram using nonlinear weighted least squares. We thus obtain N realizations of the variogram parameter

$$\theta^{(1)} = (c_0^{(1)}, c^{(1)}, a^{(1)}, \nu^{(1)}), \ldots, \theta^{(N)} = (c_0^{(N)}, c^{(N)}, a^{(N)}, \nu^{(N)})$$

Having obtained a sufficient number (N) of realizations of the posterior density $p(\theta \mid Z)$ we can then go on with the computation of the predictive density $p(Z_0 \mid Z)$ as given in Eq. 10.

Assuming a normal prior $N(\mu, \Phi)$ for the trend parameter β, the posterior density $p(Z_0 \mid Z, \theta)$ is also normal with mean $\hat{Z}_{BK}(x_0)$ = Bayes Kriging predictor and variance

$$\sigma^2_{BK}(x_0) = c_0 + \mathrm{c} + \mathrm{f}(\mathrm{x}_0)^{\mathrm{T}} \Phi \mathrm{f}(\mathrm{x}_0) - (k_0 + F\Phi f(\mathrm{x}_0))^{\mathrm{T}} (K + F\Phi F^T)^{-1} (k_0 + F\Phi f(\mathrm{x}_0))$$

= Bayes Kriging variance. The final step to be made is then the backtransformation of the density of Z_0 given Z to the original (lognormal) scale of radioactivity $Y_0 = \exp(Z_0)$, i.e. we need to compute the predictive density $p(Y_0 \mid Y)$ of Y_0 given the original observations $Y = (\exp(Z(x_1)), \ldots, \exp(Z(x_n)))$.

With the Jacobian of this transformation, $\partial Z_0 / \partial Y_0 = 1/Y_0$, we thus have $p(Z_0 \mid Z) = \int_\Theta p(Z_0 \mid Z, \theta) p(\theta \mid Z) d\theta$ and $p(Y_0 \mid Y) = p(\log Z_0 \mid Z) / Y_0$.

This means that the final predictive density can be obtained by simply averaging normal densities. Also, the above indicated numerical integrations over the regions B and Θ are accomplished by simple averagings in the Monte Carlo sense. Our numerical experiences so far confirm that these approximations using only a few hundred of simulated parameters work sufficiently well. This way we may avoid time-consuming Gibbs sampling techniques for the computation of the predictive density, as done e.g. in Diggle *et al.* (1998) and Ecker and Gelfandt (1998).

7 Illustration for an example data set

We will illustrate our solutions for a data set consisting of 592 measurements of Cs137 activities in the region of Gomel (approximately 330 km Northeast from Chernobyl) taken in Autumn 1996.

These data are lognormally distributed with logmean = 0.664 and logsigma = 1.475. We fitted a Matérn variogram to these data with parameters

$$\theta = (c_0, c, a, \nu) = (0.066, 2.452, 122.79, 0.5)$$

and for the sake of comparison, a Gaudard variogram with parameters $\theta = (c_0, c, a, \nu), c_0, c$ and a as above and $\nu = 0.02$. Both estimated values of ν clearly favour an exponential variogram.

During the simulations we observed a high variability with respect to the variograms generated, see Fig. 1.

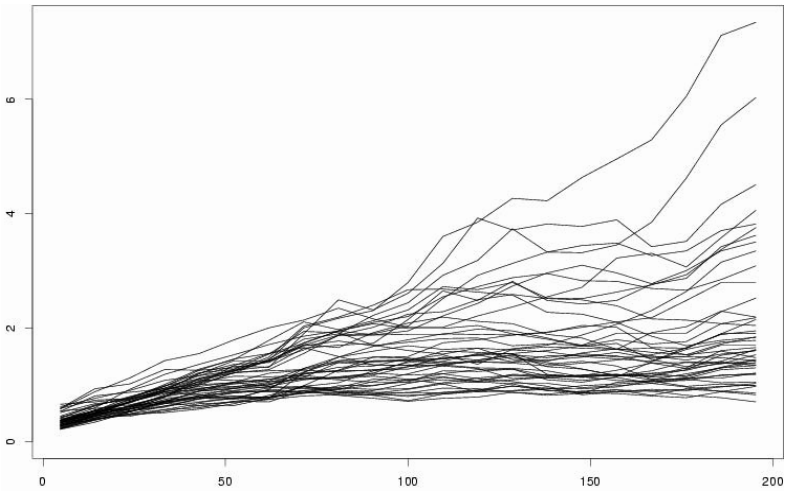


Fig. 1. Simulated variograms

The predictive densities were computed on the basis of 300 simulations each, which gave already satisfactory numerical accuracy. The next figure shows the predictive density at the point x_0 = (Easting, Northing) = (-80,40) which had been identified as a hot spot.

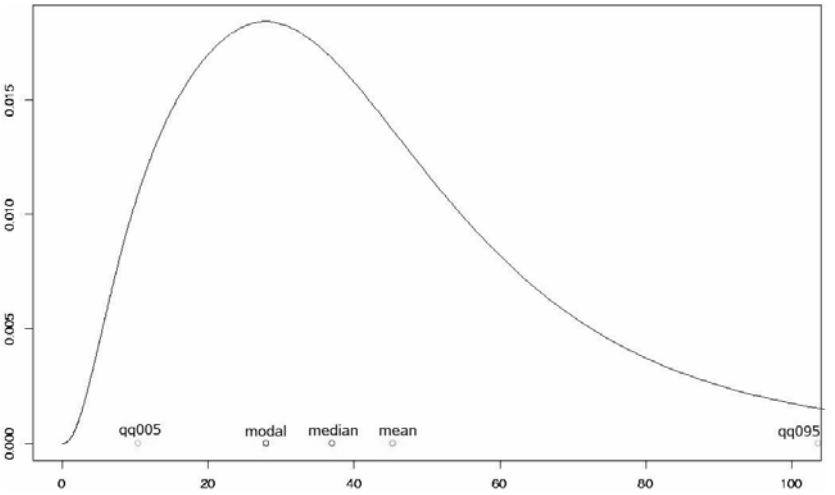


Fig. 2. Predictive density with mode, mean, median and quantiles

This figure demonstrates the great strength of the Bayesian approach, which gives us a complete probability distribution from which we may derive all interesting quantities such as the median and (95%-)quantiles, not only an expected value and variance as with "classical" Kriging methods.
If we are interested in threshold values, e.g. in the 95% threshold value, then we may simply produce a map of the 95% quantiles from the predictive distributions computed at a corresponding grid of points $x_0 \in D$.

Finally, if we are interested in an uncertainty map then we may plot the interquartile ranges (IQR), which are defined as the differences between the upper quartiles (75%-quantiles) and the lower quartiles (25%-quantiles), possibly devided by some constant a. Choosing e.g. a = 1.45 we get a map of approximate standard deviations of the predictive densities.

References

Berger JO, De Oliveira V, Sansó B (2001) Objective Bayesian Analysis of Spatially Correlated Data. Amer J (ed.), Statist. Assoc. 96, 1361–1374

Christensen R (1991) Linear Models for Multivariate, Time Series, and Spatial Data. Springer, New York

Cressie N (1993) Statistics for Spatial Data. 2nd rev. ed., Wiley, New York

Cui H, Stein A, Myers DE (1995) Extension of information, Bayesian kriging and updating of prior variogram parameters. Environmetrics 6, 373–384

De Oliveira V, Kedem B, Short DA (1997) Bayesian prediction of transformed Gaussian random fields. Amer J (ed.), Statist. Assoc. 92, 1422–1433

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussion). Appl. Statist. 47, 299–350

Dubois G, Wood R, Pilz J, Gebhardt A (2000) VSS software: combining GIS and geostatistics for the validation and the mapping of radioactivity measurements. In: CIVERT – Final Report (M. deCort, ed.), Brussels

Gaudard M, Karson M, Sinha ELD (1999) Bayesian spatial prediction. Environmental and Ecological Statistics 6, 147–171

Gebhardt A (2005) Bayesian Methods in Geostatistics: Using prior knowledge about trend. To appear in: Proc. useR Conference 2004

Handcock MS, Wallis JR (1994) An approach to statistical spatial-temporal modelling of meteorological fields (with discussion). Amer J, Statist. Assoc. 89, 368–390

Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71, 135–146

Matérn B (1986) Spatial Variation. 2nd Ed., Springer, Berlin

Pilz J (1994) Robust Bayes linear prediction of regionalized variables. In: Geostatistics for the Next century, Dimitrakopoulos R (ed.), Kluwer, Dordrecht, 464–475

Pilz J, Schimek MG, Spöck G (1997) Taking account of uncertainty in spatial covariance estimation. In: Geostatistics Wollongong, Baafi E and Schofield N (eds.), Vol. I, Kluwer, Dordrecht, 402–413

Qian SS (1997) Estimating the aerea affected by phosphorus runoff in an Everglades wetland: a comparison of universal kriging and Bayesian kriging. Environmental and Ecological Statistics 4, 1–29

Stein, ML (1999) Interpolation of Spatial Data. Some Theory for Kriging. Springer, Berlin

Zimmermann DL (1989) Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. Math. Geol. 21, 655–672

Kriging of scale-invariant data: optimal parameterization of the autocovariance model

R. Sidler and K. Holliger

Institute of Geophysics, Swiss Federal Institute of Technology (ETH),
ETH-Hoenggerberg, CH-8093 Zurich, Switzerland

1 Introduction

A critical aspect of kriging is the choice of a suitable autocovariance model that adequately describes the statistical relation between the observed data. Traditionally, there has been a rather limited repertoire of commonly used autocovariance models, the most popular of which are the exponential, spherical and Gaussian autocovariance functions (Deutsch 2001). The observation that many scientific data and phenomena are inherently scale-invariant or "fractal" (Mandelbrot 1983, Turcotte 1997) led to the introduction of corresponding autocovariance models for geostatistical applications (Hardy and Beier 1994). Despite their clear phenomenological justification and their undoubted potential, such autocovariance models are relatively rarely used and their performance has never been rigorously assessed.

One reason for this could be that scale-invariant autocovariance models are predominantly characterized by their behavior at short lags, which tends to be poorly constrained for typical, rather sparsely sampled geostatistical data. Another reason could be that, compared to more traditional autocovariance models, the parameterization of kriging estimators using pure scale-invariant autocovariance models is rather cumbersome and prone to produce numerical artifacts. Finally, the absence of an outer band-limiting scale implies that the variance of such autocovariance models becomes scale-dependent, which violates one of the key assumptions of linear geostatistics. These problems can be alleviated by using a band-limited scale-invariant autocovariance model (Chilès and Delfiner 1999, Goff and Jennings 1999). The attractiveness of such an approach is further enhanced by the fact that the limited scale of observation inherently introduces a band-limiting effect in the observed data even if the considered phenomenon is truly scale-invariant (Western and Blöschl 1999). In addition, "true" scale-invariance can be readily emulated with corresponding band-limited autocovariance models by simply choosing the outer range of scale-invariance to be larger than the actually considered range.

In this study, we present a kriging approach based on a versatile band-limited scale-invariant autocovariance model. Although we focus on 2-D datasets and linear kriging, the method can be readily extended to 3-D problems and the results obtained also apply to non-linear kriging techniques, as long as the autocovariance

structure of the original data is not fundamentally altered. We first introduce the autocovariance model, explore the optimal parameterization of the corresponding kriging algorithm, and finally interpret the results and discuss their implications.

2 Autocovariance model

This study is based on the anisotropic band-limited scale-invariant von Kármán autocovariance model (von Kármán 1948)

$$C(\mathbf{r}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\mathbf{r}}{a_r}\right)^{\nu} K_{\nu}\left(\frac{\mathbf{r}}{a_r}\right), \tag{1}$$

where $\mathbf{r}$ is the lag vector, a_r is the correlation length in the direction of the lag vector, σ is the standard deviation, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind of order $0 \leq \nu \leq 1$. The correlation length corresponds approximately to the outer range of scale-invariance. This autocovariance model has been successfully used to characterize a wide variety of scientific phenomena, such as turbulent fields (von Kármán 1948), seafloor morphology (Goff and Jordan 1988), and wave propagation in heterogeneous media (Tatarski 1961, Wu and Aki 1985). Despite its intriguingly simple parameterization, the von Kármán autocovariance model is extremely versatile and diverse. It encompasses essentially the entire spectrum of scale-invariant phenomena. For example, the popular exponential autocovariance model represents the special case of $\nu = 0.5$ of the von Kármán family of autocovariance functions. It should also be noted that autocovariance model defined by Eq. 1 remains fundamentally valid for $\nu < 0$ and $\nu > 1$, but in those cases the thus characterized data are not scale-invariant.

Fig. 1 shows von Kármán autocovariance functions for various values of ν. We see that with decreasing ν-value, the autocovariance function decreases more and more rapidly at short lags while still leveling off rather gradually at larger lags. The former implies that the local variability increases with decreasing ν-value, whereas the latter indicates that heterogeneity persists over a wide range of scales. This finds its quantitative expression in the fact that the parameter ν is related to the "Hausdorff" fractal dimension D through $D = E + 1 - \nu$ with E denoting the underlying Euclidean dimension of the stochastic process under consideration (Goff and Jordan 1988). ν and D thus control the roughness and complexity of a stochastic process. For a topographic surface, for example, E is equal to 2 and D thus lies between 2.0, which represents a very smooth, quasi-flat surface, and 3.0, which represents a very rough, quasi-space-filling surface. Stochastic processes with ν-values close to zero are referred to as flicker noise, or, equivalently, as fractional Gaussian noise (fGn) with a Hurst parameter H close to one (Hardy and Beier 1994). Flicker noise behavior is probably the most commonly observed stochastic phenomenon and characterizes a wide variety of data throughout virtually all fields of the natural and social sciences (West and Shlesinger 1990).

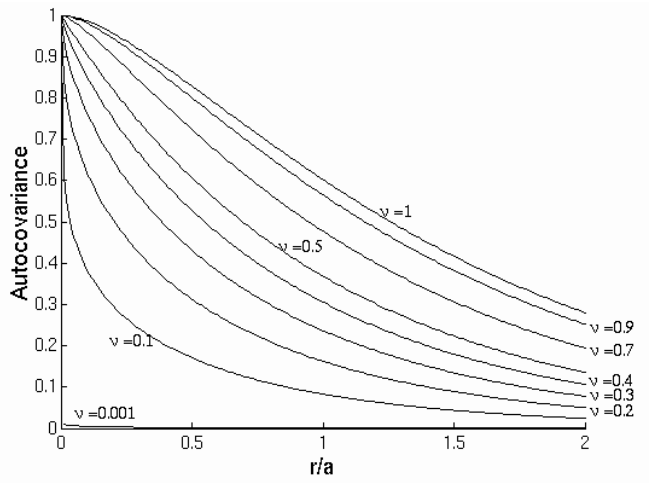


Fig. 1. Plot of a set of normalized von Kármán autocovariance functions. *r* and *a* denote the lag and the correlation length, respectively.

The versatility of the von Kármán autocovariance model is illustrated in Fig. 2a-2c, which show 2-D synthetic spatially anisotropic stochastic data fields for ν-values of 0, 0.5, and 1. The average autocovariance functions of these synthetic data fields are consistent with the corresponding parametric models. To generate these synthetic data fields, we take the amplitude spectrum of the corresponding stochastic process, as defined by the square-root of the Fourier transform of Eq. 1, uniformly randomize the phase spectrum, and take the inverse Fourier transform (Goff and Jordan 1988). The resulting stochastic dataset has a Gaussian probability density function. Von-Kármán-type stochastic data fields characterized by continuous (e.g., lognormal) or discrete (e.g., bimodal) non-Gaussian probability density functions can be obtained through subsequent transformations of the original Gaussian-distributed datasets (Goff *et al.* 1994, Lampe and Holliger 2003).

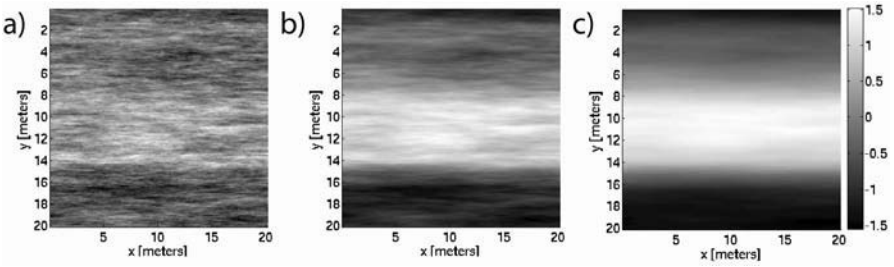


Fig. 2. Dimensionless synthetic data fields characterized by von Kármán autocovariance functions with ν-values of **a)** 0, **b)** 0.5 and **c)** 1. All models have a standard deviation of 0.5 and horizontal and vertical correlations lengths of 100 m and 10 m, respectively. The models were generated using the same seed number to initialize the random number generator and hence exhibit the same overall structure but differ in terms of their "roughness".

3 Kriging of scale-invariant data

3.1 Validation of the algorithm

In the following, we apply the von Kármán autocovariance model outlined above (Eq. 1) to ordinary kriging, which estimates the value $\hat{Y}$ at the non-sampled location $\mathbf{r}_o$ as (e.g., Kelkar and Perez 2002)

$$\hat{Y}(\mathbf{r_0}) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{r}_i)\,, \tag{2}$$

where $Y(\mathbf{r}_i)$ are the observed data, λ_i are the corresponding weighting factors and n the number of observed data used for the estimation. These weighting factors are related to the autocovariance model characterizing the data through the following system of equations (e.g., Kelkar and Perez, 2002)

$$\sum_{j=1}^{n} \lambda_j C(\mathbf{r}_i, \mathbf{r}_j) + \mu = C(\mathbf{r}_0, \mathbf{r}_0) \quad \text{for } i=1,\ldots, n, \tag{3}$$

where $C(\mathbf{r}_i, \mathbf{r}_j)$ denotes the value of the used parametric model for the autocovariance function for the lag vector $\mathbf{r} = \mathbf{r}_i - \mathbf{r}_j$ (Eq. 1) and μ is the Lagrange multiplier. Ordinary kriging does not make any assumptions with regard to the mean value of the data and is probably the most widely used geostatistical estimation technique. Because of the close mathematical analogies between essentially all kriging techniques, the corresponding algorithms can be readily interchanged by applying minor modifications to the governing equations. Please note that $\hat{Y}$ corresponds to the expected value at a given location and hence the autocovariance function of a kriged dataset does not correspond to the used parametric model (e.g., Gelhar 1993).

Ideally, the number of points n used for the interpolation procedure should comprise all available observations. In practice, however, this may be computationally too costly, particularly for large datasets. Moreover, the value of the kriging weights, and thus their relative importance, decreases rapidly with increasing autocovariance lag. This decay is particularly pronounced for rapidly decaying autocovariance models, such as von Kármán autocovariance functions with small ν-values (Fig. 1). Based on extensive tests with subsets of the synthetic data fields shown in Fig. 2a-2c, we decided to use the 32 observations that are closest to the non-sampled location in a search neighborhood, the shape of which is consistent with the anisotropy ellipsoid of the autocovariance model. This approach conforms to the accepted practice for large datasets (Kelkar and Perez 2002).

An effective way to verify the implementation of a kriging algorithm is through cross-validation, that is, by "blindly" estimating data whose values are actually known. To this end, we estimate a number of omitted values from the synthetic stochastic data field shown in Fig. 2b using the basic "leaving-one-out" cross-validation approach (Kelkar and Perez 2002). The results shown in Fig. 3 indicate that the estimated values are uniformly close the actual ones (Fig. 3a), that the magnitudes of the errors are independent of the magnitudes of the values to be es-

timated (Fig. 3b) and that the spatial distribution of the kriging errors is seemingly uncorrelated (Fig. 3c). These results also imply that the corresponding estimation is conditionally unbiased and that the error variance is spatially uncorrelated. All of these observations are consistent with the fundamental assumptions of kriging (Kelkar and Perez 2002).

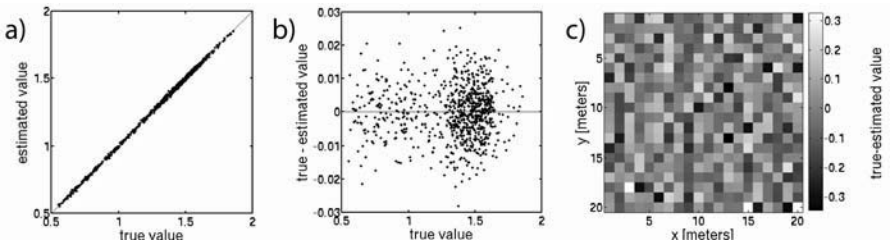


Fig. 3. Results from the cross-validation of part of the dataset shown in Fig. 2b. Shown are **a)** actual versus kriged values, **b)** actual values versus kriging errors, and **c)** the spatial distribution of kriging errors.

3.2 Optimal parameterization

Constraining and parameterizing the autocovariance model based on the observed data is a particularly delicate and time-consuming aspect of geostatistical analyses. Based on the theoretical foundations of kriging, optimal estimation results should be obtained when using the autocovariance model that best fits the observed data (e.g., Journel and Huijbregts 1978). For sparsely sampled datasets, the estimation of the autocovariance model is, however, notoriously uncertain and error prone due to the lack of information at short lags (e.g., Kitanidis 1997). Consequently, it is important to assess the sensitivity of the estimation with regard to the choice of the autocovariance model and to determine whether the autocovariance parameters characterizing the observed data do indeed provide the best results.

For this purpose, we have cross-validated the four-fold decimated stochastic data fields shown in Fig. 2a-2c using a "jackknife" approach (Kelkar and Perez, 2002) for variable ν-values and anisotropic aspect ratios (ratios of horizontal to vertical correlation lengths). In doing so, we initially fixed the horizontal correlation length at the correct value. Fig. 4a-4c show the sums of the absolute estimation errors as functions of the autocovariance parameters. These results indicate that, with increasing ν-value of the input data, the sum of the absolute errors decreases sharply and the minima of these parameter trade-off maps defining the optimal parameterization of the autocovariance model become increasingly broad and ill defined. Regardless of the ν-value of the input data, this trade-off analysis estimates the anisotropic aspect ratio of the input data with remarkable accuracy. The most puzzling and interesting result of this sensitivity analysis, however, is that the optimal ν-value for kriging is always larger than the ν-value actually characterizing the input data. This discrepancy is most pronounced for input data with very small values of ν and decreases with increasing ν-values of the data.

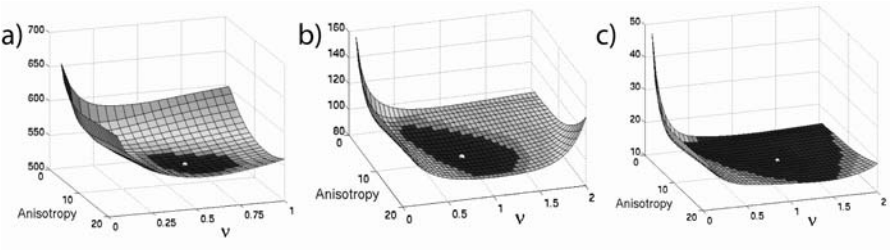


Fig. 4. Sum of absolute cross-validation errors as functions of ν-value and anisotropic aspect ratio used in autocovariance model. The horizontal correlation length in the autocovariance model was fixed at the correct value of 100 m. The input data correspond to those in Fig. 2a-2c after four-fold decimation and are characterized by ν-values of **a)** 0, **b)** 0.5, and **c)** 1 and horizontal and vertical correlation lengths of 100 m and 10 m, respectively. White dots denote the locations where the sum of the estimation errors is at a minimum.

To assess the sensitivity of kriging estimation to the absolute values of the correlation lengths, we have repeated the above analysis for the input dataset shown in Fig. 2b, kept ν fixed at the correct value of 0.5, but varied the horizontal correlation length in the autocovariance model used for the sensitivity analysis between half (50 m) and twice (200 m) the actual value (100 m). The resulting trade-off maps (Fig. 5a-5c) are quite similar to the corresponding trade-off map determined with the horizontal correlation length fixed at the actual value of the input data (Fig. 4b). These results indicate that kriging estimation of spatially anisotropic scale-invariant data is sensitive to the choice of the ν-value and aspect ratio, but remarkably robust with regard to the choice of the absolute values of the correlation lengths. This finding is of significant practical importance because accurate estimates of the correlation lengths are notoriously difficult to obtain and in many cases may be influenced by "filtering artifacts" due to the inherently finite experimental scales (Gelhar 1993, Western and Blöschl 1999).

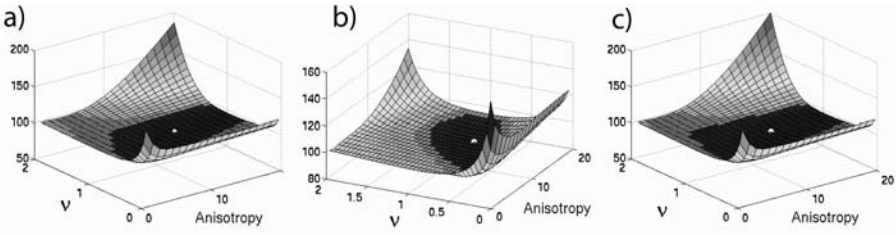


Fig. 5. Sum of absolute cross-validation errors as functions of ν-value and anisotropic aspect ratio used in autocovariance model. ν was fixed at the correct value of 0.5 and the horizontal correlation length at **a)** 50 m, **b)** 100 m and **c)** 200 m. The input data correspond to those in Fig. 2b after four-fold decimation and are characterized by a ν-value of 0.5 and horizontal and vertical correlation lengths of 100 m and 10 m, respectively. White dots denote the locations where the sum of estimation errors is at a minimum.

Finally, Fig. 6-8 compare the data fields shown in Fig. 2a-2c after ten-fold decimation and subsequent kriging estimation of the decimated values with both the actual and the optimal ν-values estimated from Fig. 4a-4c. Differences in the corresponding estimates are quite pronounced for the input dataset characterized by $\nu = 0$ (Fig. 6c and 6d). An interesting observation is that the dataset kriged with the ν-value of the input data is characterized by a subtle but highly systematic "salt-and-pepper" pattern consisting of high-amplitude values coincident with the locations of the data used for kriging. This is illustrated in the blown up part of Fig. 6c. Moreover, the estimate obtained with the ν-value of the input data is much too smooth and exhibits significant distortions with regard to the character of the input data. Kriging the data with the optimal ν-value notably attenuates these artifacts and the result can be clearly identified as an adequately smoothed version of the input data. In contrast, the differences between the kriging estimates obtained with the actual and optimal autocovariance models are quite modest for the input data characterized by $\nu = 0.5$ (Fig. 7c and 7d) and insignificant for the input data characterized by $\nu = 1$ (Fig. 8c and 8d). Moreover, the overall similarity between the original input stochastic data fields and their kriged equivalents increases significantly with increasing ν-value of the input data. This finding is consistent with the systematic decrease in the sum of the estimation errors with increasing ν-values of the input models observed in the parameter trade-off analyses (Fig. 4a-4c).

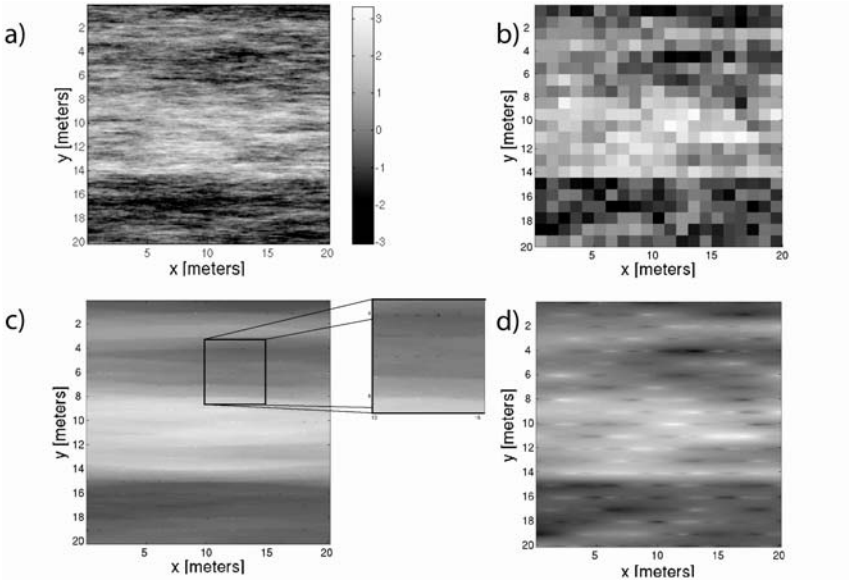


Fig. 6. a) Dimensionless synthetic data characterized by $\nu = 0$ shown in Fig. 2a. **b)** Same data after ten-fold decimation. Kriging of decimated data **c)** with the ν-value of the input data (note the "salt-and-pepper" pattern in the blown up part) and **d)** with the optimal ν-value minimizing the sum of the absolute errors (Fig. 4a).

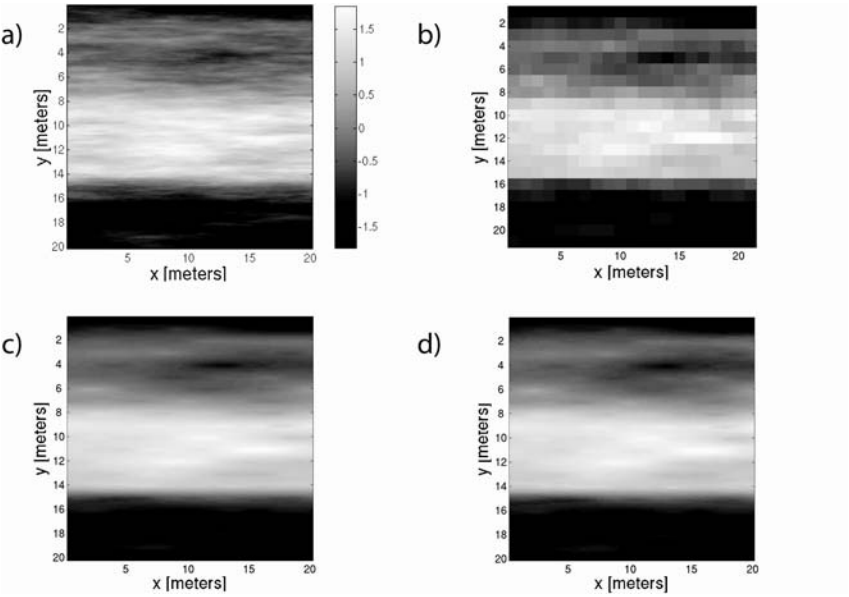


Fig. 7. a) Dimensionless synthetic data characterized by $\nu = 0.5$ shown in Fig. 2b. **b)** Same data after ten-fold decimation. Kriging of decimated data **c)** with the ν-value of the input data and **d)** with the optimal ν-value minimizing the sum of the absolute errors (Fig. 4b).

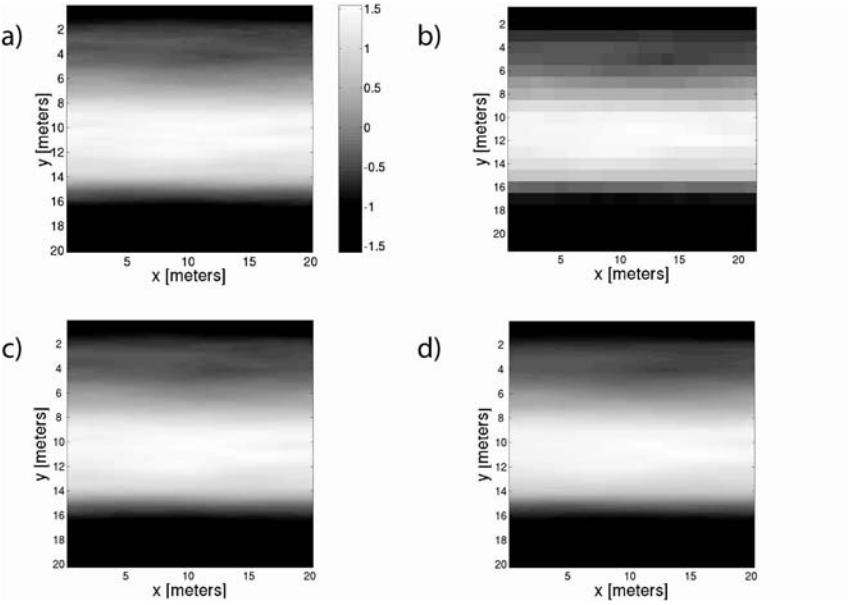


Fig. 8. a) Dimensionless synthetic data characterized by $\nu = 1$ shown in Fig. 2c. **b)** Same data after ten-fold decimation. Kriging of decimated data **c)** with the ν-value of the input data and **d)** with the optimal ν-value minimizing the sum of the absolute errors (Fig. 4c).

4 Discussion

An important outcome of the above analysis is that the best results, in terms of the estimation errors, are obtained when scale-invariant data are kriged with ν-values that are systematically higher than those characterizing the actual input data. This finding is at odds with the theoretical foundations of kriging as well as the common practice in geostatistical data analysis and hence is of considerable practical importance. In the following, we quantify the relationship between the actual and optimal ν-values and explore the origins of this rather puzzling phenomenon.

Fig. 4a-4c show that the difference between the actual and optimal ν-values decreases with increasing ν-value of the input data. To constrain the relationship between the ν-value of the input data and the optimal ν-value for kriging in more detail, we have conducted the parameter trade-off analysis outlined above with for the ν-values only. The synthetic input data fields used for this parameter analysis are the type of those shown in Fig. 2 and only differ in terms of their ν-values, which are incremented in steps of 0.1. These input data fields are then decimated by a factor of four prior to "jackknife"-type parameter trade-off analysis. The average autocovariance functions of the synthetic input data are found to be consistent with the corresponding parametric models. The same is true for the decimated data, although the behavior at short lags, which primarily constrains the ν-value, becomes more ambiguous, particularly for small ν-values.

The results of the analysis are summarized in Fig. 9 and corroborate the findings inferred from the trade-off maps shown in Fig. 4a-4c. As expected from the results shown in Fig. 4a, there are large differences between actual and optimal ν-values for data characterized by very small ν-values. For example, an actual ν-value of 0 (flicker noise) corresponds to an optimal ν-value for kriging of ~0.5. Based on this parameter trade-off analysis, the relationship between the actual ν-values characterizing the input data ν_{true} and the ν-values that provide the optimal kriging results ν_{opt} can be approximated through the following empirical polynomial relation

$$\nu_{opt} = -0.64 \cdot \nu_{true}^2 + 1.20 \cdot \nu_{true} + 0.54 \,. \tag{4}$$

Together with Fig. 9, this empirical relation illustrates that the difference between actual and optimal ν-values diminishes with increasing actual ν-values, such that for ν-values close to 1 the actual and optimal ν-values converge. An intriguing outcome of this analysis is that the range of actual ν-values from 0 to 1 characterizing the input data is mapped into a "compressed" range of optimal ν-values between 0.5 and 1. The reason for this could be that the nature of scale-invariant data, in particular their predictability, changes fundamentally around $\nu = 0.5$ (Hergarten 2002). Data characterized by ν-values smaller than 0.5 are referred to as anti-persistent, such that a positive gradient between two data points is likely to be associated with a negative gradient between adjacent data points. Conversely, scale-invariant data characterized by ν-values larger than 0.5 are referred to as

persistent, such that a positive gradient between two data points is likely to be associated with positive gradients between adjacent data points.

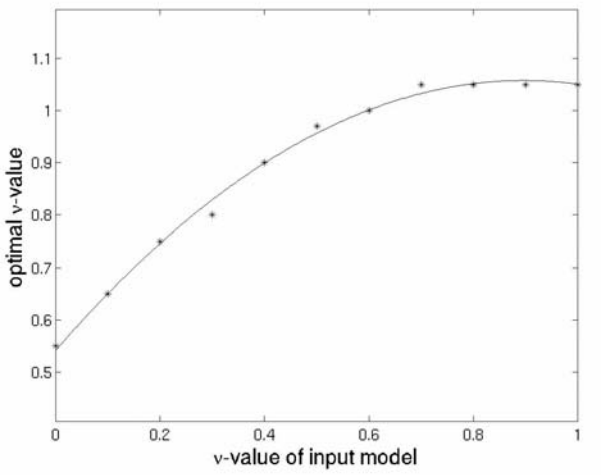


Fig. 9. Plot quantifying the relationship between the actual ν-values of the input data and the ν-values that provide the best kriging estimates. Stars: estimates obtained from parameter trade-off analyses of the kind shown in Fig. 4; solid line: best-fitting polynomial approximation (Eq. 4).

The predictability of anti-persistent data is inherently limited and decreases with decreasing ν-value. For scale-invariant data characterized by ν-values close to 0, the predictability is essentially as poor as for totally random white noise, where the best prediction simply corresponds to the expected or mean value of the entire dataset. Kriging of such a dataset thus simply estimates the data at an unsampled location as the expected value of the corresponding search neighborhood. This results in a smooth data field centered around the global mean value, from which the observed values with their high intrinsic variability stand out as anomalies, forming the "salt-and-pepper" pattern observed in the blown-up part of Fig. 6a.

The predictability of persistent data is inherently better than that of their anti-persistent equivalents and for ν-values of 1 converges to a basic linear estimation. This is consistent with the smooth appearance of the kriged data fields shown in Fig. 7 and 8, which are quite similar to the corresponding input data. It also explains why the sums of the absolute errors displayed in the parameter trade-off maps in Fig. 4a-4c decrease markedly with increasing ν-values of the input data. The finding that the range of optimal ν-values for the kriging of scale-invariant data is limited to values between 0.5 and 1 can thus be interpreted as a reflection of the fact that kriging, like essentially all other estimation techniques, is inherently based on the assumption of persistency of the underlying database.

Another intriguing outcome of this analysis is that actual ν-values close to 0 correspond to optimal ν-values close to 0.5 (Fig. 9). ν-values of 0 and 0.5 define two of the best known parametric autocovariance models in stochastic data analy-

sis. As mentioned above, scale-invariant data characterized by ν-values close to 0 are referred as flicker noise, a ubiquitous and seemingly universal stochastic characteristic of data observed in a wide variety of scientific disciplines (West and Shlesinger 1990). In the earth sciences, for example, there is increasing evidence that the spatial distributions of virtually all petrophysical properties seem to exhibit flicker noise behavior (Hardy and Beier 1994, Kelkar and Perez 2002). Although the origins of this immensely common scaling phenomenon are still enigmatic (Holliger and Goff 2003), its practical potential for providing critical *a priori* information for the conditional stochastic simulation is increasingly being realized (Hardy and Beier 1994). A ν-value of 0.5, on the other hand, corresponds to an exponential autocovariance function, which is widely regarded as one of the most common and robust parametric models, particularly in geostatistics. It is therefore interesting and important to note that, although exponential autocovariance models are commonly used for kriging estimation, there seem to be rather few studies that convincingly applied this model to characterize the observed autocovariance behavior of densely sampled, high quality datasets (Turcotte 1997). Fig. 4a and 9 demonstrate that for a dataset characterized by ν-value close to zero, cross-validation and/or parameter trade-off analyses would unambiguously point towards the use of an exponential autocovariance model for kriging. Conversley, data characterized by an exponential autocovariance function should actually be kriged using a considerably smoother autocovariance function (Fig. 4b and 9). Moreover, it can be shown that a sparsely sampled autocovariance function characterized by a ν-value close to 0 can be adequately matched through an exponential autocovariance model in combination with a nugget effect. The results of this study may therefore indicate that the popularity and robustness of the exponential autocovariance model in geostatistics could, at least in part, be a reflection of the ubiquity of flicker noise ($\nu \approx 0$) behavior in the observed data.

5 Conclusions

This study illustrates the suitability, versatility and robustness of the von Kármán autocovariance model for kriging scale-invariant data and provides clear guidelines for with regard to the optimal choice of the autocovariance parameters. Despite the inherently band-limited nature of the von Kármán autocovariance model, the corresponding kriging algorithm proves to be surprisingly robust with regard to the absolute values of the horizontal and vertical correlation lengths as long as the anisotropic aspect ratio of the observed data is honored. An important finding is that for input data characterized by ν-values between 0 and 1, the range of ν-values that provide optimal kriging estimates is compressed to a range between 0.5 and 1. This phenomenon is interpreted as a reflection of the inherently persistent nature of the estimation process and may suggest that the popularity of the exponential autocovariance model ($\nu = 0.5$) in geostatistics could indeed be an "artifact" related to the inherent sparseness of typical geostatistical data and the ubiquity and universality of flicker noise statistics ($\nu \approx 0$) in natural phenomena.

Acknowledgements

Comments and suggestions by John A. Goff, Alan G. Green and Jens Tronicke helped to improve the quality of this manuscript. ETH-Geophysics Contribution No. 1378.

References

Chilès J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York

Deutsch CV (2001) Geostatistical reservoir modeling. Oxford University Press, Oxford

Gelhar LW (1993) Stochastic subsurface hydrology. Prentice-Hall, Englewood Cliffs

Goff JA, Jennings JW (1999) Improvement of Fourier-based unconditional and conditional simulations for band-limited fractal (von Kármán) statistical models. Math Geol 31: 627–649

Goff JA, Jordan TH (1988) Stochastic modeling of seafloor morphology: inversion of sea beam data for second-order statistics. J Geophys Res 93: 13589–13608

Goff JA, Holliger K, Levander A (1994) Modal fields: a new method for characterization of random seismic velocity heterogeneity. Geophys Res Lett 21: 493-496

Hardy HH, Beier RA (1994) Fractals in reservoir engineering. World Scientific, Singapore.

Hergarten S (2002) Self-organized criticality in earth systems. Springer, Berlin

Holliger K, Goff JA (2003) A generalized model for the 1/f-scaling nature of seismic velocity fluctuations. In: Goff JA, Holliger K (eds.) Heterogeneity in the crust and upper mantle – nature, scaling, and seismic properties. Kluwer Academic/Plenum Publishers, New York, 131-154

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic Press, San Diego

Kelkar M, Perez G (2002) Applied geostatistics for reservoir characterization. Society of Petroleum Engineers, Richardson, Texas

Kitanidis PK (1997) Introduction to geostatistics. Cambridge University Press, Cambridge.

Lampe B, Holliger K (2003) Effects of fractal fluctuations in topographic relief, permittivity, and conductivity on ground-penetrating radar antenna radiation. Geophysics 68: 1934-1944

Mandelbrot BB (1983) The fractal geometry of nature. Freeman, New York

Tatarski VL (1961) Wave propagation in a turbulent medium. McGraw-Hill, New York

Turcotte DL (1997) Fractals and chaos in geology and geophysics. 2nd edition, Cambridge University Press, Cambridge

von Kármán T (1948) Progress in the statistical theory of turbulence. Marit Res J, 7: 252–264

West BJ, Shlesinger M (1990) The noise in natural phenomena. American Scientist 78: 40-45

Western AW, Blöschl G (1999) On the spatial scaling of soil moisture. Hydrol J, 217: 203–224

Wu R-S, Aki K (1985) The fractal nature of the inhomogeneities in the lithosphere evidence from seismic wave scattering. Pure Appl Geophys 123: 805-818.

Scaling Effects on Finite-Domain Fractional Brownian Motion

S. Cintoli[1,2], S. P. Neuman[1] and V. Di Federico[2]

[1]Department of Hydrology and Water Resources, University of Arizona, Tucson, USA
[2]DISTART, Università di Bologna, Bologna, Italy

1 Introduction

There is growing evidence that the hydrogeologic properties of porous and fractured media may be statistically nonhomogeneous and behave as random fractals (Molz *et al.* 2003). Aspects of this behavior are captured by a variety of fractal models such as fractional Gaussian noise (fGn) (Hewett 1986, Robin *et al.* 1991, Molz and Boman 1993, 1995, Tubman and Crane 1995, Liu and Molz 1996, Eggleston and Rojstaczer 1998), corresponding power law approximations (Glimm *et al.* 1993, Dagan 1994), Weierstrass-Mandelbrot fractal function (Molz *et al.* 1998), fractional Levy motion (fLm) (Painter 1996a-b, 1998), or multifractals (Liu and Molz 1997, Molz *et al.* 1997, Boufadel *et al.* 2000). We focus in this paper on power-law variograms that lack a finite sill (asymptotic value representing variance) or correlation scale. Such variograms have been inferred from porosity and/or permeability data at several sites (Hewett 1986, Grindrod and Impey 1992, Desbarats and Bachu 1994, Molz and Boman 1993, 1995, Tubman and Crane 1995, Guzman *et al.* 1996, Liu and Molz 1996) on distance scales ranging from meters to 100 km. They are indicative of a nonhomogeneous (nonstationary) random field with homogeneous spatial increments (differences between values at points separated by some distance or lag).

A statistically isotropic random field $Y(x)$ with homogeneous spatial increments, defined on an infinite domain, is characterized by a power (semi)variogram (PV) or second-order structure function

$$\gamma(\mathbf{s}) = \frac{1}{2}\left\langle \left[Y(\mathbf{x}+\mathbf{s}) - Y(\mathbf{x})\right]^2 \right\rangle = C_0 \mathbf{s}^{2H}, \qquad (1)$$

where $\boldsymbol{x}$ is a vector of spatial coordinates, $\boldsymbol{s}$ is a displacement (lag) vector, s is the magnitude of $\boldsymbol{s}$, $< >$ indicates ensemble mean (expectation), C_0 is a constant and H is the Hurst coefficient. Since the variogram scales as $\gamma(rs) = r^{2H}\gamma(s)$ the field is self-affine and, within the range $0 < H < 1$, constitutes a random fractal with dimension $D = E + 1 - H$ where E is Euclidean (topologic) dimension (Voss 1985). If the field is additionally Gaussian, it constitutes fractional Brownian motion (fBm).

Let $\gamma(s,\lambda)$ be an exponential or Gaussian variogram having variance $\sigma^2(n) = C/n^{2H}$ where $n = 1/\lambda$ is a mode number, λ being the integral scale, C is a constant having dimensions $[L^{-2H}]$, $0 < H < 0.5$ in the exponential case and $0 < H < 1$ in the Gaussian case. Then one can express PV as a weighted integral of such variograms over all modes (Di Federico and Neuman 1997),

$$\gamma(s) = \int_0^\infty \gamma(s,n)\frac{dn}{n}. \tag{2}$$

Introducing a lower cutoff $n_l = 1/\lambda_l$ yields a truncated power variogram (TPV)

$$\gamma(s,n_l) = \int_{n_l}^\infty \gamma(s,n)\frac{dn}{n}, \tag{3}$$

given in the case of exponential modes by

$$\gamma(s,n_l) = \frac{C_0}{\Gamma(1-2H)n_l^{2H}} S_E, \tag{4}$$

where, $S_E = [1-\exp(-n_l s)+(n_l s)^{2H}\Gamma(1-2H,n_l s)]$, $C_0 = C\Gamma(1-2H)/2H$, Γ being the complete gamma function and $\Gamma(a,x)$ the incomplete gamma function (Abramowitz and Stegun 1972, equation (6.5.3), p. 260). This TPV defines a homogeneous field associated with a constant variance $\sigma^2(n_l) = C_0/\Gamma(1-2H)n_l^{2H}$ and finite integral scale $I(n_l) = \lambda_l 2H/(1+2H)$. In the case of Gaussian modes we have

$$\gamma(s,n_l) = \frac{C_0'}{\Gamma(1-H)(\pi/4)^H n_l^{2H}} S_G, \tag{5}$$

where $S_G = [1-\exp(-(\pi/4)n_l^2 s^2)+((\pi/4)n_l^2 s^2)^H\Gamma(1-H,(\pi/4)n_l^2 s^2)]$ and $C'_0 = C(\pi/4)^H \Gamma(1-H)/2H$. Gaussian TPVs have variance $\sigma^2(n_l) = C'_0/[\Gamma(1-H)(\pi/4)^H n_l^{2H}]$ and again integral scale $I(n_l) = \lambda_l 2H/(1+2H)$. In the limit as $n_l \to 0$, Eq. 3 to 5 reduce to Eq. 2.

The integral scale $I(n_l) = \lambda_l 2H/(1+2H)$ of the truncated field is proportional to, and smaller than, the cutoff scale λ_l (integral scale of the lowest mode retained). For $n_l s << 1$, or equivalently $s << \lambda_l$, $\gamma(s,n_l) \approx C_0 s^{2H}$ indicating that TPV coincides with PV over lags much smaller than the lower cutoff scale. In this paper we ask two related questions:

(a) At what value of s/λ_l does the approximation $\gamma(s,n_l) \approx C_0 s^{2H}$ become acceptable?

(b) How is λ_l related to the length scale L of a finite domain (window) superimposed on an fBm defined on an infinite domain? In other words, is there a relationship between the truncation criterion (since $\lambda_l = 1/n_l$) and the domain of a truncated field?

A tentative answer to the second question was proposed by Di Federico and Neuman (1997) and Di Federico *et al.* (1999) (for the implications of their work vis-à-vis flow and transport see Di Federico and Neuman 1998a,b). It had been

noted earlier by Neuman (1994) that upon plotting published integral scales of log hydraulic conductivity and transmissivity in a variety of soils and aquifers versus a characteristic length of the sampling domain on logarithmic paper (Fig. 1), the data appear to delineate a straight line with a 1/10 slope. Di Federico and Neuman (1997) set $\lambda_l = \mu L$ and wrote $I(n_l) = \lambda_l 2H/(1+2H) = \alpha\lambda_l = \alpha\mu L$ to obtain $\alpha\mu \approx 0.1$. From juxtaposed apparent dispersivity data derived on the basis of tracer studies world-wide, Neuman (1990, 1995) deduced a generalized value of the Hurst coefficient, $H \approx 0.25$. This gave $\alpha \approx 1/3$ and, correspondingly, $\mu \approx 1/3$. However, there has been no independent theoretical verification of this quasi-empirical result.

In this paper we address the above two questions through numerical Monte Carlo simulation and variogram analysis of two-dimensional fBm fields.

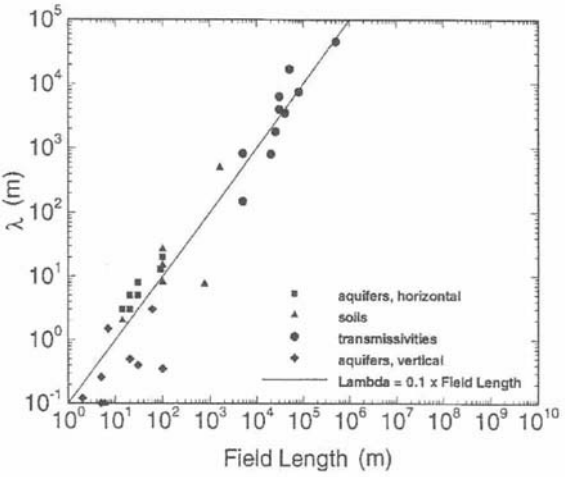


Fig. 1. log-K and log-T integral scales [after Neuman (1994), data from Gelhar (1993, Table 6.1)]

2 Generation of fields characterized by power variograms using HYDRO_GEN

To generate random realizations of fBm fields we used the HYDRO_GEN software of Bellin and Rubin (1996) and Rubin and Bellin (1998). HYDRO_GEN is a sequential Gaussian simulator of stationary or non-stationary, conditional or unconditional random field replicates over a grid of arbitrary geometry. It achieves computational efficiency by relying on a fixed sequence of conditioning points, and associated interpolation coefficients, for all replicates.

Two sets of 200 fBm realizations each were generated on a square grid of 101×101 points coinciding with the centroids of unit cells. One set corresponded to $H = 0.25$, the other to $H = 0.75$, and both to $C_0 = 0.027$ (deduced by Neuman 1990, 1995, from world-wide apparent dispersivity data). Fig. 2a shows one of the

200 fBm realizations with $H = 0.25$, and Fig. 2b one with $H = 0.75$. The realization in Fig. 2a illustrates antipersistence (negatively correlated increments) and that in Fig. 2b persistence (positively correlated increments).

To obtain sample variograms for a given realization over the entire length L of our grid, 202 pairs of points were considered at each horizontal and vertical lag $s = 1, 3, 5, \ldots, L$. We found the resulting sample variograms of individual realizations to vary widely in shape, magnitude and degree of anisotropy as shown in Fig. 3a and 3b, implying a distinct lack of ergodicity. Averaging these sample variograms over 100 and 200 realizations of each field resulted in rapid convergence toward the corresponding isotropic power variogram, as illustrated in Fig. 4a and. 4b. Normalized versions of the 200-sample (dashed) and theoretical (solid) variograms in Fig. 5 suggest that convergence to the power model is slower for $H = 0.75$ than for $H = 0.25$ as a result of persistence.

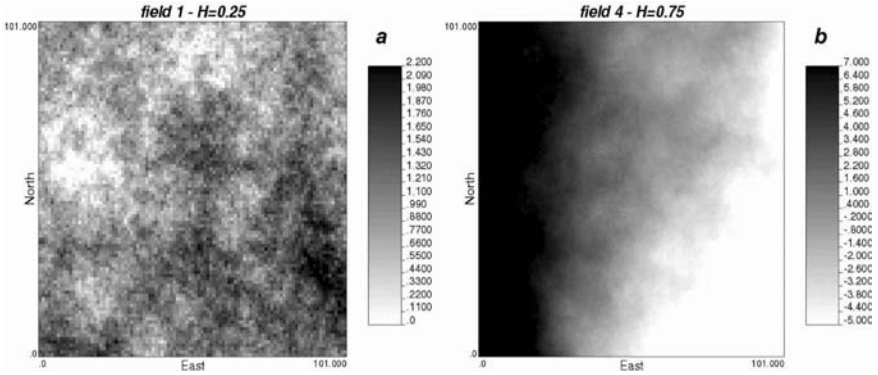


Fig. 2. Example of generated fields with **a**) H=0.25 and **b**) H=0.75

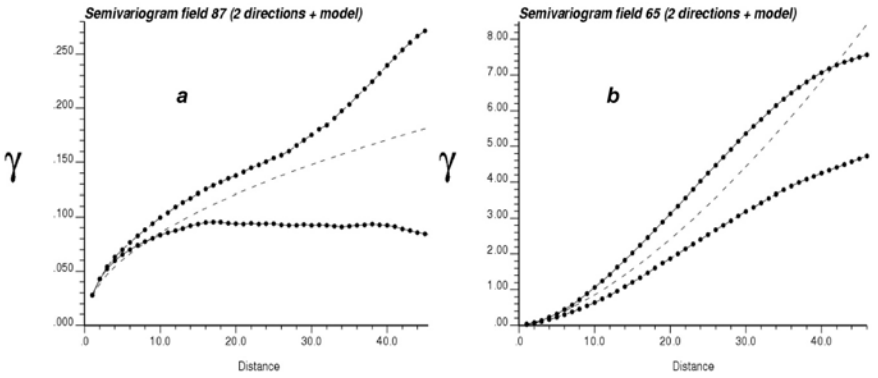


Fig. 3. Example of directional sample variograms extracted **a**) from a single realization with H=0.25 and **b**) from a realization with H=0.75

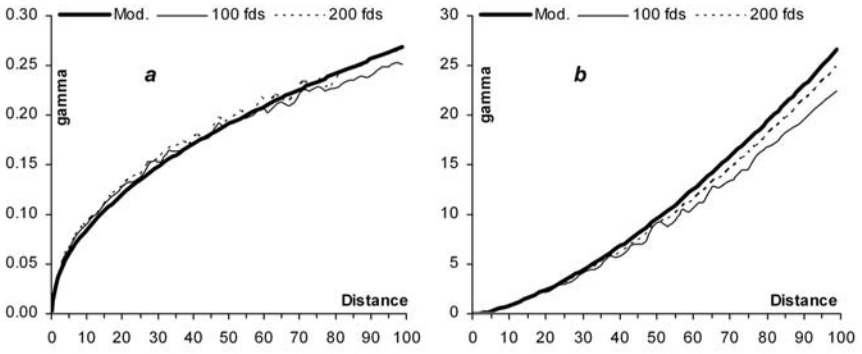


Fig. 4. Model vs average sample variograms over an ensemble of 100 and 200 fields with **a**) H=0.25 and **b**) H=0.75

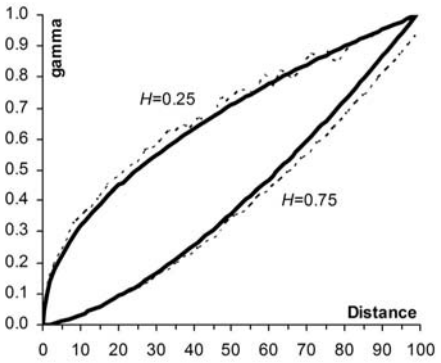


Fig. 5. Normalized model (solid) vs 200 fields average-sample variograms (dashed) for H=0.25 and H=0.75

3 Effect of sampling window size

In practice, sampling is always limited to finite size domains (windows). To investigate the effect of sampling an fBm within a finite size window, it is necessary to generate realizations of the fBm over a domain that is much larger (ideally infinite) than the window. Computational constraints have limited us in this study to square domains of 101×101 points. We have superimposed on each set of 200 realizations (corresponding to H = 0.25 and 0.75) two sampling windows, one containing 60×60 points and the other 40×40 points, both centered about the midpoint of the larger domain. Fig. 6a and 6b show sample variograms across each window corresponding to H = 0.25 and 0.75, respectively, averaged over all 200 realizations. Regardless of window size or H, each 200-sample average variogram is seen to lie very close to the corresponding theoretical power model.

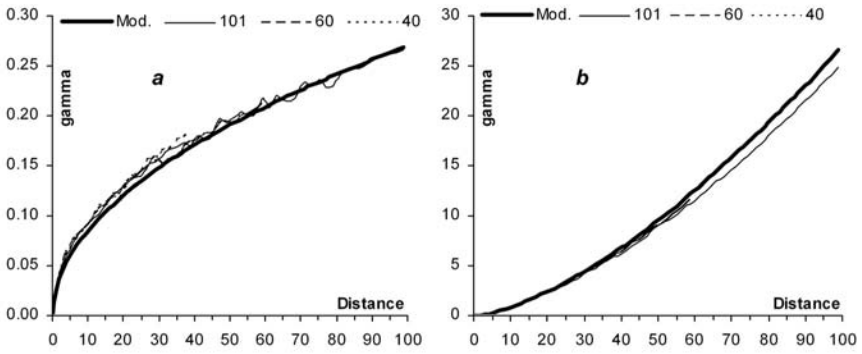


Fig. 6. Sample variograms inferred from square windows of decreasing side extracted from fields **a**) with H=0.25 and **b**) H=0.75

4 Fitting TPVs to power variogram on finite windows

We now investigate the relationship between truncated power variograms (TPVs) and average (over a set of realizations) sample (over a given realization) variograms of fBm fields on finite windows of length scale L. Given our finding in the previous section that such average sample variograms tend to the power model, we compare in Fig. 7a and 7b TPVs characterized by various lower cutoffs with PVs for H = 0.25 and H = 0.75, respectively. These figures show that as the lower cutoff n_l decreases, the TPVs in each case tend asymptotically to the corresponding PV over the entire length L of our grid. When n_l is relatively large, the TPVs lie below the power model; as n_l decreases, they initially rise above this model and then diminish toward it asymptotically. This provides a partial answer to our earlier question at what value of s/λ_l (or sn_l) does the approximation $\gamma(s,n_l) \approx C_0 s^{2H}$ become acceptable? Fig. 11 and 12 suggest that, for the approximation to be valid more or less uniformly over the entire range $0 \le s \le L$, one must have $n_l << 0.0001$. A perfect reproduction of the model over the whole domain (not shown in the figures) is only obtained for $n_l = 0.00001$ or $\lambda_l = 100{,}000$ (corresponding to $\mu = \lambda_l/L \approx 1{,}000$).

Next we examine for L = 101 what n_l (and λ_l) values are required to obtain $\gamma(s,n_l) = C_0 s^{2H}$ to four significant figures when s = 29, 51, 59, 99. Fig. 8a shows such a fit between a Gaussian truncated power model with H = 0.75 and the power model for s = 59. The fit appears acceptable over the entire range $0 \le s \le 59$. Fig. 8b shows that for H = 0.25 and s = 99 the exponential TPV approximates the power model more closely than does the Gaussian TPV. Fig. 9 plots the λ_l values obtained by fitting three TPVs to the power model at the above four lags. For any given TPV these cutoff scales delineate straight lines representing fixed ratios $\mu = \lambda_l/s$. This provides a strong numerical confirmation that, for a given TPV, μ is independent of lag (or, equivalently, window size). Di Federico and Neuman (1997) considered μ to be additionally independent of the choice of TPV model and H;

Fig. 9 and Table 1 suggest that μ may in fact vary somewhat with this choice, though the variations may perhaps be an artifact of our particular way of fitting TPVs to the power model. The theoretical μ values in Table 1 are much smaller than the asymptotic values deduced earlier from Fig. 7a and 7b, and much closer to the semi-empirical value $\mu = \lambda_l/L = 1/3$ deduced by Di Federico and Neuman (1997) for $H = 0.25$ on the basis of hydraulic conductivity, transmissivity and apparent dispersivity data.

Fig. 10a and 10b show how exponential and Gaussian TPV models compare with the power model when μ is set equal to 1/3 and to its value in Table 1 corresponding to $H = 0.25$. It is instructive to note that earth and/or environmental data typically represent a single realization of what is assumed to be some underlying random field (in our case, an fBm). As the number of data pairs often diminishes rapidly with their separation distance (lag), it is common in variogram analyses to disregard or assign very low weights to data pairs with lags in excess of $L/2$. Quite often one infers from the remaining data pairs a variogram which, subject to possible filtering out of an underlying drift, represents a statistically homogeneous field. Our results suggest that, if the fitted variogram is a truncated power model, it may be associated with a relatively low μ value on the order of 1/3 – 2. It is worth noting here that the latter values of μ provide a "good" fit over lags up to half the considered 101×101 domain, while a very high value of the same parameter (i.e. $\mu \approx 1{,}000$) is required in order to achieve a perfect reproduction of the model over the entire window.

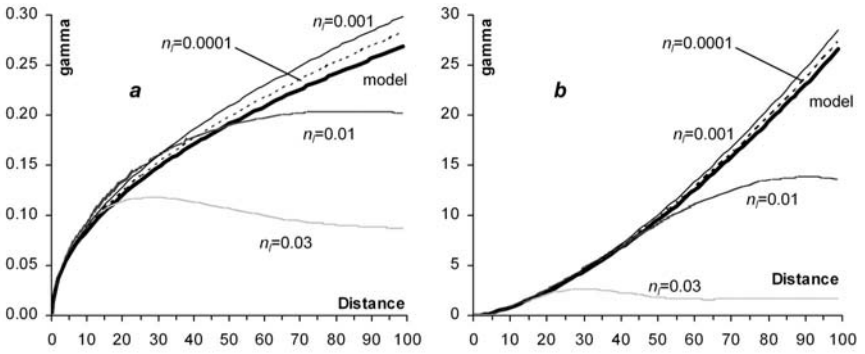


Fig. 7. a) Exponential TPVs with H=0.25 and **b**) Gaussian TPVs with H=0.75 for varying n_l

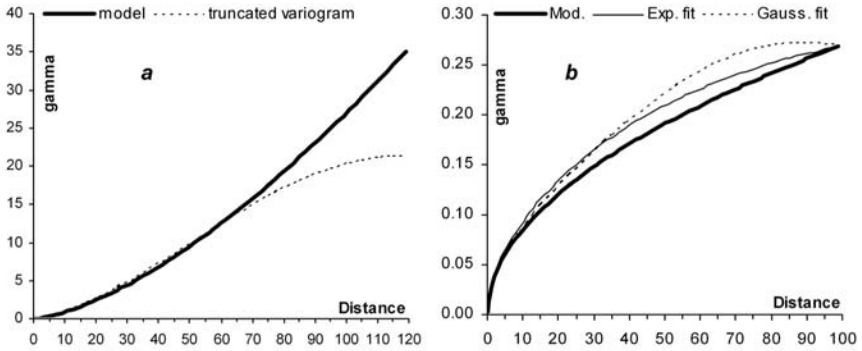


Fig. 8. a) Gaussian TPV with H=0.75 fitted to PV model at s=59; **b**) Exponential TPV with H=0.25 vs Gaussian TPV with H=0.25 at s=99

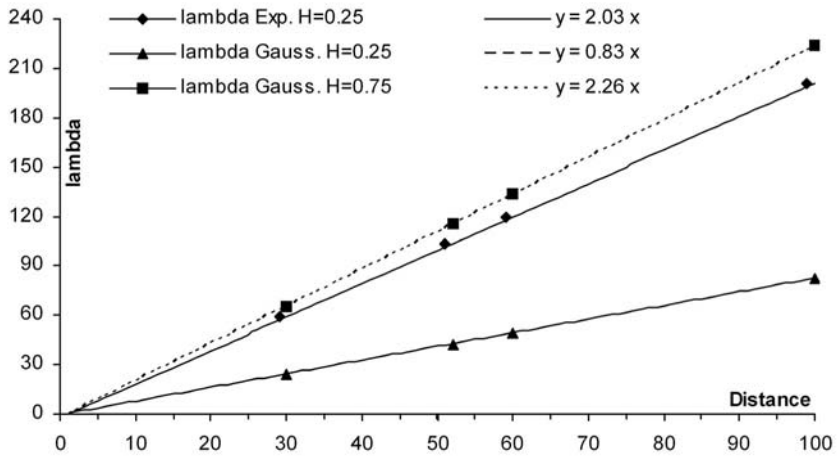


Fig. 9. Discrete integral scales interpolated by lines with constant μ

Table 1. μ values corresponding to Fig. 9

TPV fitted to PV	μ
Exponential modes with H=0.25	2.03
Gaussian modes with H=0.25	0.83
Gaussian modes with H=0.75	2.26

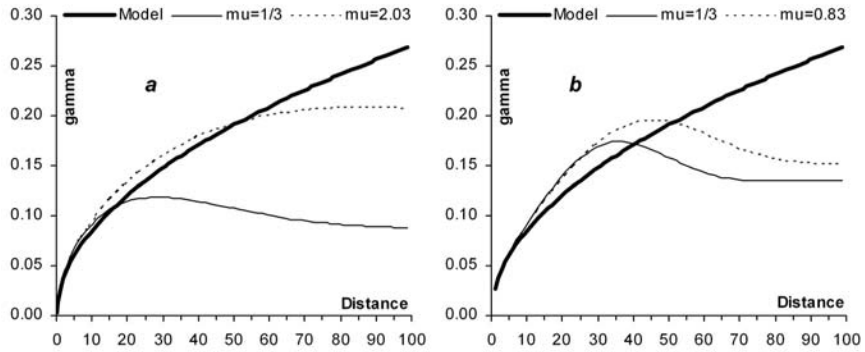


Fig. 10. a) Exponential TPV with H=0.25 and varying μ (light solid: μ=1/3, dashed: μ=2.03); **b)** Gaussian TPV with H=0.25 and varying μ (light solid: μ=1/3, dashed: μ=0.83). In both cases the model PV is provided for comparison.

6 Conclusions and future developments

The following major conclusions can be drawn from this paper:

1. The software HYDRO_GEN can be used with success to generate unconditional random fields characterized by isotropic power variograms on finite domains of length scale L. Due to lack of ergodicity, sample variograms of individual realizations are direction dependent and differ from each other sharply in form and magnitude. However, the average of such sample variograms over 200 realizations lies very close to the theoretical power model, more so in the case of anti-persistent fields with Hurst coefficient H = 0.25 than in the case of persistent fields with H = 0.75.
2. The previous conclusion applies to arbitrary size sampling domains (windows) of length scale smaller than L.
3. Truncated power variograms with a low frequency cutoff scale λ_l approximate the power model more or less uniformly well over a finite domain of length scale L when $\lambda_l \geq 100{,}000$ (corresponding to $\mu = \lambda_l/L \geq 1{,}000$).
4. Fitting such truncated power models to the power model at any discrete lag s yields ratios $\mu = \lambda_l/s$ that are independent of s (and hence window scale). This provides numerical support for a corresponding postulate by Di Federico and Neuman (1997). The latter authors considered μ to be additionally independent of the choice of TPV model and H; our results suggest that μ may in fact vary somewhat with these choices (within a range of about 0.8 – 2), though these variations may perhaps be an artifact of our particular way of fitting TPVs to the power model. Ratios μ within this range are much smaller than the asymptotic ratio μ = 1,000 listed in the previous conclusion and much closer to the semi-empirical value μ = 1/3 deduced by Di Federico and Neuman (1997) for H = 0.25 on the basis of hydraulic conductivity, transmissivity

and apparent dispersivity data previously analyzed by Neuman (1990, 1994, 1995).

5. Earth and environmental data typically represent a single realization of what is assumed to be some underlying random field (in our case, an fBm). As the number of data pairs often diminishes rapidly with their separation distance (lag), it is common in variogram analysis to disregard or assign very low weights to data pairs with lags in excess of $L/2$. Quite often one infers from the remaining data pairs a variogram which, subject to possible filtering out of an underlying drift, represents a statistically homogeneous field. Our results suggest that if the fitted variogram is a truncated power model, it may be associated with a relatively low μ value on the order of 1/3 – 2. It is worth noting here that the latter values of μ provide a "good" fit over lags up to half the considered 101×101 domain, while a very high value of the same parameter (i.e. $\mu \approx 1{,}000$) is required in order to achieve a perfect reproduction of the model over the entire window.
6. Multiscale fields characterized by TPVs are statistically homogeneous and should therefore be easier to generate than fields characterized by power variograms. Individual realizations of the former, when generated over sufficiently large domains, should yield sample variograms that reproduce closely the underlying TPV model. We are therefore modifying HYDRO_GEN so it can generate random fields characterized by truncated power variograms.

Acknowledgements

The authors are grateful to Prof. Alberto Bellin of Dipartimento di Ingegneria Civile ed Ambientale e CUDAM (Trento - Italy) for his generous help in our implementation of HYDRO_GEN.

References

Abramowitz M, Stegun IA (1972) Handbook of Mathematical Functions. Dover, Mineola, New York

Bellin A, Rubin Y (1996) HYDRO_GEN: A spatially distributed random field generator for correlated properties. Stochastic Hydrology and Hydraulics (10), 253-278

Boufadel MC, Lu S, Molz FJ, Lavalle D (2000) Multifractal scaling of the intrinsic permeability. Water Resources Research 36, 3211-3222

Dagan G (1994) Significance of heterogeneity of evolving scales to transport in porous formations. Water Resources Research 30(12), 3327–3336

Desbarats AJ, Bachu S (1994) Geostatistical analysis of aquifer heterogeneity from the core scale to the basin scale: A case study. Water Resources Research 30(3), 673-684

Di Federico V, Neuman SP (1997) Scaling of random fields by means of truncated power variograms and associated spectra. Water Resources Research 33(5), 1075-1085

Di Federico V, Neuman SP (1998a) Flow in multiscale log conductivity fields with truncated power variograms. Water Resources Research 34(5), 975-987

Di Federico V, Neuman SP (1998b) Transport in multiscale log conductivity fields with truncated power variograms. Water Resources Research 34(5), 963-973

Di Federico V, Neuman SP, Tartakovsky DM (1999) Anisotropy, lacunarity, upscaled conductivity and its covariance in multiscale fields with truncated power variograms. Water Resources Research 35(10), 2891-2908

Eggleston J, Rojstaczer S (1998) Inferring spatial correlation of hydraulic conductivity from sediment cores and outcrops. Geophys. Res. Lett. 25(13), 2321–2324

Gelhar LW (1993) Stochastic Subsurface Hydrology. Prentice-Hall, Englewood Cliffs, New Jersey

Glimm J, Lindquist WB, Pereira F, Zhang Q (1993) A theory of macrodispersion for the scale-up problem. Transp. Porous Media 13(1), 97–122

Grindrod P, Impey MD (1992) Fractal field simulations of tracer migration within the WIPP Culebra Dolomite. Rep. IM2856-1, vers. 2, p. 62, Intera Inf. Technol., Denver, Colorado, March 1992

Guzman AG, Geddis AM, Henrich MJ, Lohrstorfer CF, Neuman SP (1996) Summary of Air Permeability Data From Single-Hole Injection Tests in Unsaturated Fractured Tuffs at the Apache Leap Research Site: Results of Steady-State Test Interpretation. Rep. NUREG/CR-6360, prepared for U.S. Nuclear Regulatory Commission, Washington, D.C.

Hewett TA (1986) Fractal distributions of reservoir heterogeneity and their influence on fluid transport. SPE Pap. 15386 presented at 61st Annual Technical Conference, Soc. Petrol. Engin., New Orleans, Los Angeles

Liu HH, Molz FJ (1996) Discrimination of fractional Brownian movement and fractional Gaussian noise structures in permeability and related property distribution with range analysis. Water Resources Research 32(8), 2601–2605

Liu HH, Molz FJ (1997) Multifractal analysis of hydraulic conductivity distributions. Water Resources Research 33(11), 2483–2488

Molz FJ, Boman GK (1993) A stochastic interpolation scheme in subsurface hydrology. Water Resources Research 29(11), 3769–3774

Molz FJ, Boman GK (1995) Further evidence of fractal structure in hydraulic conductivity distribution. Geophys. Res. Lett. 22(18), 2545–2548

Molz FJ, Liu HH, Szulga J (1997) Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. Water Resources Research 33(10), 2273–2286

Molz FJ, Hewett TA, Boman GK (1998) A pseudo-fractal model for hydraulic properties in porous medium. In: Baveye P, Parlange J-Y, Stewart BA (eds) Fractals in Soil Sciences. CRC Press, Boca Raton, Fla, 341–372

Molz FJ, Rajaram H, Lu S (2003) Stochastic fractal-based models of heterogeneity in subsurface hydrology: Origins, applications, limitations, and future research questions. Reviews of Geophysics, in press

Neuman SP (1990) Universal scaling of hydraulic conductivities and dispersivities in geologic media. Water Resources Research 26(8), 1749-1758

Neuman SP (1994) Generalized scaling of permeabilities: Validation and effect of support scale. Geophys. Res. Lett. 21(5), 349-352

Neuman SP (1995) On advective transport in fractal velocity and permeability fields. Water Resources Research 31(6), 1455-1460

Neuman SP, Di Federico V (2003) The multifaceted nature of hydrogeologic scaling and its interpretation. Rev. Geophys., in press

Painter S (1996a) Evidence for non-Gaussian scaling behavior in heterogeneous sedimentary formations. Water Resources Research 32(5), 1183–1195

Painter S (1996b) Stochastic interpolation of aquifer properties using fractional Levy motion. Water Resources Research 32(5), 1323–1332

Painter S (1998) Numerical method for conditional simulation of Levy random fields. Math. Geol. 30(2), 163-179

Robin MJL, Sudicky EA, Gillham RW, Kachanoski RG (1991) Spatial variability on strontium distribution coefficients and their correlation with hydraulic conductivity in the Canadian Forces Base Borden aquifer. Water Resources Research 27(10), 2619-2632

Rubin Y, Bellin A (1998) Conditional Simulation of Geologic Media with Evolving Scales of Heterogeneity. In: Sposito G (ed), Scale Dependence and Scale Invariance in Hydrology, Cambridge University Press, 398-420

Tubman KM, Crane SD (1995) Vertical versus horizontal well log variability and application to fractal reservoir modeling. In: Barton CC, La Pointe PL (eds) Fractals in Petroleum Geology and Earth Processes. Plenum, New York, 279–293

Voss RF (1985) Random fractals: Characterization and measurement. In: Pynn R, Skjeltorp A (eds), Scaling Phenomena in Disordered Systems, NATO ASI Series, p. 133

The delineation of fishing times and locations for the Shark Bay scallop fishery

U. Mueller[1], L. Bloom[1], M. Kangas[2], N. Caputi[2] and T. Tran[1]

[1]Edith Cowan University, Perth, Western Australia
[2]Department of Fisheries, Marine Research Laboratories, Waterman, Western Australia

1 Introduction

In this paper we use scallop survey data and lognormal ordinary kriging (Chiles and Delfiner 1999) to obtain a spatial mapping of estimated scallop density in the Red Cliff and NW Peron regions of the Shark Bay managed scallop fishery in Western Australia. The results can then be used, together with the annual pre-season scallop survey, to inform the management decision as to the opening time of the subsequent scallop fishing season.

The Shark Bay Scallop Fishery is Western Australia's largest scallop fishery. Its outer boundaries encompass the waters of the Indian Ocean and Shark Bay between 23°34' south latitude and 26°30' south latitude and adjacent to Western Australia on the landward side of the 200 m isobath, together with those waters of Shark Bay south of 26°30' south latitude (Department of Fisheries 2002).

The scallop catch depends primarily on the strength of recruitment from the breeding season of the previous year. Spawning commences in mid-April and meat condition declines as spawning continues, so the process of setting the opening date of the season needs to balance breeding stock and the seasonal decline in meat condition. In order to determine the opening date for the fishing season a pre-season survey is conducted in November and December of the previous year. The survey covers the three fishing regions Red Cliff, NW Peron, and Denham Sound but, as there has been little fishing activity in Denham Sound during the years considered, only the Red Cliff and NW Peron regions are used.

2 The survey data

The survey data we considered are for the years 2000 to 2003. The fishing grounds Red Cliff and NW Peron are adjacent and are treated by the Department of Fisheries, Western Australia as one fishing ground for stock prediction and we treat them in the same manner here. Each survey was carried out by FRV Naturaliste, equipped with two six-fathom headrope nets. The combined fishing ground is

north of 25°30' south latitude and south of 24° 40', with the Red Cliff survey locations lying north of 25°10' south latitude. For our analysis the locations were converted to nautical miles and a local coordinate system with origin at 24° south latitude and 113° east longitude was chosen. A map of the three fishing grounds together with the survey locations is shown in Fig. 1. The area outlined in grey shows the region for which estimates were computed.

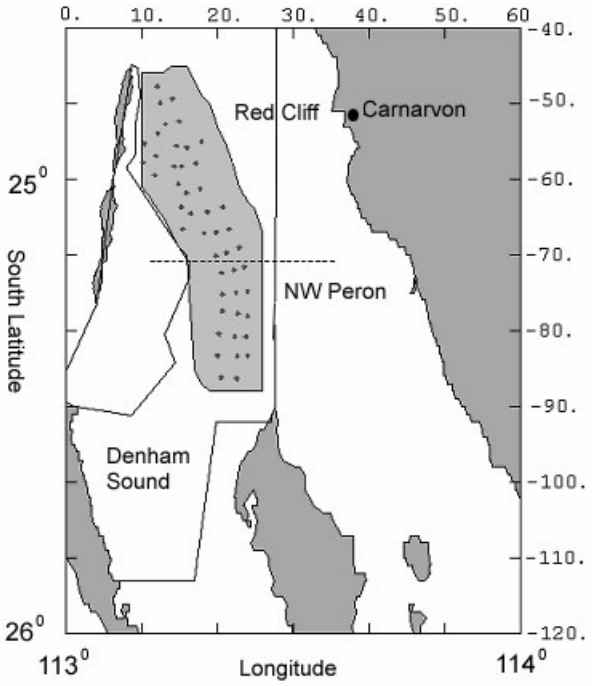


Fig. 1. Shark Bay scallop fishery, the dots indicate survey locations, the legends on the right and on the top give distances in nautical miles relative to the chosen origin

The data comprise the fishing ground, the longitude and latitude in degrees of the start and end locations of each trawl, the counts of recruit and residual scallops caught per net, the trawl duration, distance and speed. The number of survey locations varies from year to year. The numbers, giving the regional split, are shown in Table 1.

Table 1. Number of sample locations by year and data set

Fishing Ground	2000	2001	2002	2003
Red Cliff	23	18	33	30
NW Peron	19	12	12	17

The number of residuals and recruits caught per trawl and net were aggregated into total number of residuals and recruits per trawl respectively. As the trawling speed influences the efficiency of the trawl gear, the catch (by category and total) was standardised to the equivalent catch at a speed of 3.4 knots

$$c_{st} = \frac{c}{3.2331 - 0.6485v}. \quad (1)$$

Here *v, c* and c_{st} denote the trawl speed in knots, the catch and the standardised catch respectively. This formula was derived via a combination of practical experience to decide on a suitable adjustment factor and a subsequent linear regression of this adjustment factor on trawl speed (J. Penn, unpublished) and is deemed reliable by the Department of Fisheries, WA. For this study the standardised number of residuals, recruits and total number of scallops were converted to densities according to

$$d = \frac{c_{st}}{2Tw}, \quad (2)$$

where *T* and *w* denote the trawl distance and the width per net in nautical miles.

The scallop density distributions are highly positively skewed with the 2003 residuals density and 2002 recruits density the most strongly skewed (see Tables 2 to 4).

Table 2. Descriptive statistics of the density of residuals

Residuals Density	2000	2001	2002	2003
Mean	3641	3057	1717	5621
Standard Deviation	7054	3288	1941	13349
Minimum	0	143	0	0
Lower Quartile	0	814	314	713
Median	336	2016	991	2180
Upper Quartile	3960	4470	2539	5013
Maximum	33272	14564	7065	89610
Skewness	2.9	2.1	1.5	5.7

Table 3. Descriptive statistics of the density of recruits

Recruits Density	2000	2001	2002	2003
Mean	9117	10404	14817	17233
Standard Deviation	10561	9619	31808	21117
Minimum	0	369	0	0
Lower Quartile	1555	3848	1428	2382
Median	5865	6737	3541	7784
Upper Quartile	14425	15573	14856	24143
Maximum	55347	34101	196560	86739
Skewness	2.5	1.3	5.0	1.6

A comparison of the mean densities for recruits and residuals during the four years shows that residuals comprise respectively 29, 23, 10 and 25 percent of all scallops caught in the combined Red Cliff and NW Peron fishing region.

Table 4. Descriptive statistics of the density of total scallop catch

Total Scallop Density	2000	2001	2002	2003
Mean	12758	13461	16534	22853
Standard Deviation	14524	11444	32973	27344
Minimum	0	590	142	351
Lower Quartile	3382	5314	2750	6345
Median	7067	9460	4456	10568
Upper Quartile	16576	20906	17192	33097
Maximum	68060	42674	203626	125440
Skewness	2.2	1.1	4.0	2.0

Spatial maps for the densities of residuals, recruits and total catch for Red Cliff and NW Peron for the year 2001 are shown in Fig. 2. There are more locations with high residuals density in the Red Cliff fishing ground than in NW Peron fishing ground. For recruits the locations of high density are more evenly distributed through the two fishing grounds and the locations of low density lie in the centre of the fishing ground. Locations of high density of residuals are not co-located with those of high recruit density. Overall residuals density values are much lower than recruits densities.

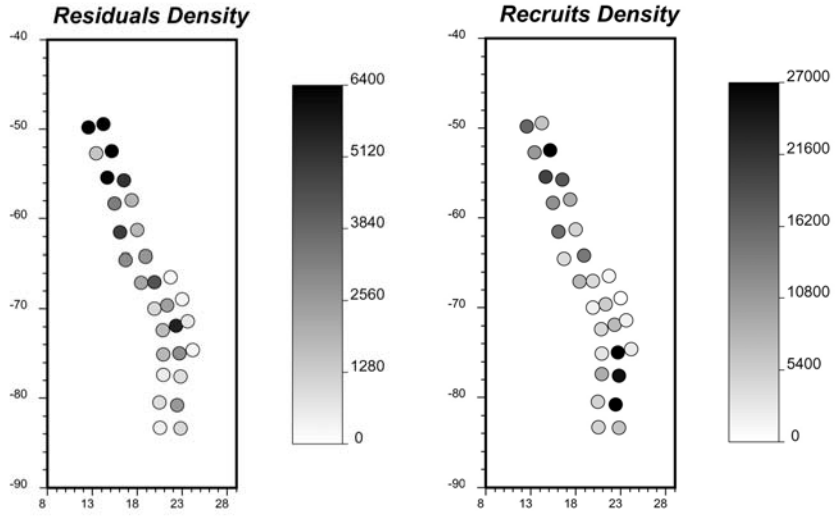


Fig. 2. Spatial maps of residuals density and recruits density 2001

The density patterns for recruits and residuals change from year to year indicating variable settlement patterns in these areas. In 2000 residuals scallop density was highest in the central part of Red Cliff and was low in NW Peron and the northern part of Red Cliff. In 2002 residuals scallop density was low to moderate in NW Peron, high at the western rim of Red Cliff, and low at the eastern rim of

Red Cliff. In the year 2003 the density distribution was similar to that in 2002 except for the occurrence of a high density patch in the south east of NW Peron.

In the year 2000 high recruits densitiy occurred throughout most of the western part of the combined fishing ground, with low density along the eastern rim. In 2002 recruits density was highest in the north west close to the permanent closure area and low throughout NW Peron. In 2003 recruits density in the north west were similar to (and greater in absolute terms than) those in 2002. Values in the NW Peron ground were low overall compared with the rest of the study region with the exception of two locations with high density in the centre.

The spatial distributions of the annual total scallop survey density for each of the four years are strongly influenced by the recruits distributions and follow similar patterns.

3 Estimation

Three-parameter lognormal ordinary kriging was used to obtain estimates for the densities of the three variables. We denote by $y(\mathbf{u})$ the lognormal variable obtained from the attribute $z(\mathbf{u})$ by putting $y(\mathbf{u}) = \ln[z(\mathbf{u}) + c]$, with c being an added constant. In each case the constant was chosen so that the transformed variable follows a normal distribution at the 5% level of significance. The constants for the specific distributions are given in Table 5.

Table 5. Added constants for lognormal distributions

Variable	2000	2001	2002	2003
Residuals	25	1	100	10
Recruits	1500	1	150	50
Totals	2000	1	0	0

The corresponding random variable will be denoted by $Y(\mathbf{u})$. The estimate for the natural logarithm of the value of the attribute at the unsampled location **u** may be expressed as

$$y^*(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) y(\mathbf{u}_i), \tag{1}$$

where $n(\mathbf{u})$ denotes the number of data near **u**, and $\lambda_j(\mathbf{u})$ denotes the ordinary kriging weight of the *j*-th nearby sample. The estimate $z^*(\mathbf{u})$ is then obtained from the logarithmic estimate $y^*(\mathbf{u})$, the ordinary kriging variance $\sigma_Y^2(\mathbf{u})$ and the Lagrange multiplier $\mu(\mathbf{u})$ by

$$z^*(\mathbf{u}) = \exp(y^*(\mathbf{u}) + \sigma_Y^2(\mathbf{u})/2 + \mu(\mathbf{u})) - c \tag{2}$$

with variance

$$\hat{\sigma}^2(\mathbf{u}) = \exp(\sigma_Y^2(\mathbf{u}))(1+\exp(-(\sigma_Y^2(\mathbf{u})+\mu(\mathbf{u})))(\exp(-\mu(\mathbf{u}))-2)) . \quad (3)$$

Spherical models were fitted to the experimental semivariograms. The parameters used in the estimation are summarised in Table 6.

Table 6. Variogram model parameters for the variables (residual, recruit, total)

	2000	2001	2002	2003
Nugget	(0, 0.4, 0.4)	(0.5, 0.2, 0.28)	(0.3, 0.38 0.36)	(0.39, 0.99, 0.7)
$Sill_1$	(6.1, 0.42, 0.42)	(0.2, 0.7, 0.8)	(1.2, 0.53, 0.92)	(2.36, 1.59, 0.95)
$Range_1$	(13, 4.6, 4.6)	(5, 4, 5.3)	(4.8, 2.4, 5.3)	(6.3, 5.3, 4.6)
$Sill_2$		(0.62, 0.28, 0)	(0, 1.53, 0.7)	
$Range_2$		(17, 8.4, 0)	(0, 10.9, 8.6)	

The experimental semivariograms for the residuals and recruits density of 2001 and the corresponding models are shown in Fig. 3. In each case the sample variance has been chosen as the total sill.

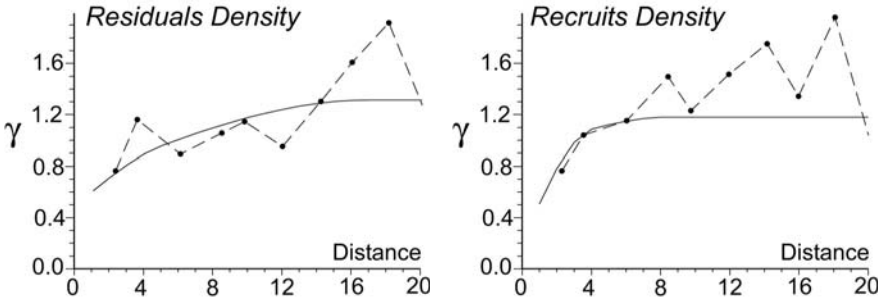


Fig. 3. Experimental semivariograms and models for residual and recruit density 2001

Cross validation results for lognormal kriging using these models are given below. From Table 7 it can be seen that the mean errors for all three variables are close to 0, there is greatest variability in the errors for the residuals density in 2000 and for all densities in 2003.

Table 7. Cross validation results for lognormal kriging

Density	Statistic	2000	2001	2002	2003
Residuals	Mean Error	0.078	0.035	-0.006	0.070
	Variance	2.130	1.04	0.970	2.073
Recruits	Mean Error	0.021	0.071	0.035	0.113
	Variance	0.973	1.124	1.188	2.638
Total Catch	Mean Error	0.023	0.071	0.052	0.087
	Variance	0.876	1.029	0.918	1.984

From Table 8 it can be seen that, with the exception of the results for 2002, the mean square error exceeds the mean kriging variance. The results are typically worse for the residuals density with the mean square error exceeding the mean kriging variance by 48%.

Table 8. Crossvalidation results for lognormal kriging, MSE/ $\overline{\sigma}_Y^2(\mathbf{u})$

Density	2000	2001	2002	2003
Residual	1.48	1.35	0.96	1.11
Recruit	1.26	1.24	0.91	1.17
Total Catch	1.18	1.25	0.86	1.30

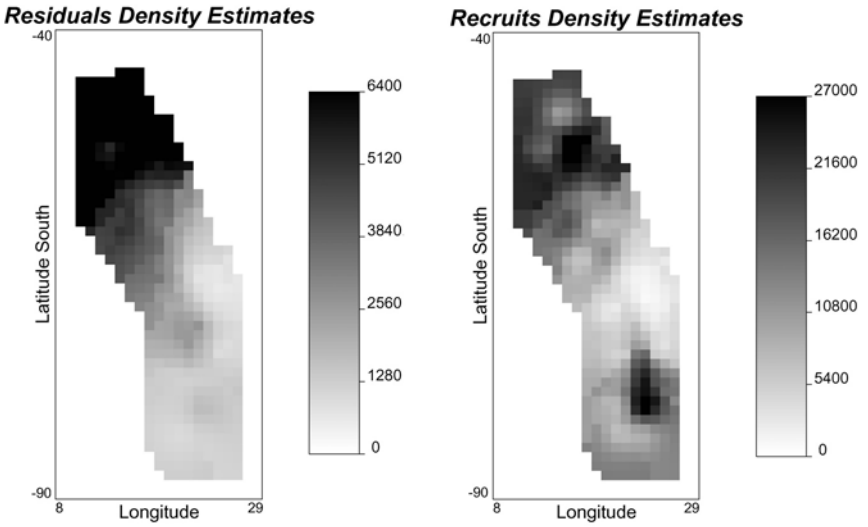


Fig. 4. Estimates of residual density and recruit density 2001

Spatial maps of the density estimates for residuals, and recruits for 2001 are shown in Fig. 4. They exhibit trends similar to those of the sample location maps discussed earlier. In all four years, there was a region of high residuals density in the Red Cliff ground. For the years 2000 and 2001 the residuals density in NW Peron was low. In 2002 and 2003 there was a different pattern in this part, with some high densities emerging in the south east. For recruits there were locations of high density in the south-east of the NW Peron ground in 2000 and 2001. In the remaining years the density was greatest in Red Cliff. Similar trends prevailed for the total catch density. The mean, standard deviation and skewness for the estimates are given in Table 9.

Table 9. Abridged descriptive statistics of the density estimates

Density	2000	2001	2002	2003
Residual				
Mean	3658	3718	1774	6670
Standard Deviation	5902	2934	1368	9295
Skewness	2.20	0.84	1.32	2.89
Recruit				
Mean	10118	12769	11075	22540
Standard Deviation	5775	6944	18180	16218
Skewness	1.15	0.23	3.17	0.95
Total Scallop				
Mean	14389	16464	12713	26062
Standard Deviation	7727	9275	18865	16791
Skewness	0.73	0.36	3.41	1.49

Except in the year 2002, when the proportion of residuals was 14%, the contribution of residuals to the expected total catch was approximately 25%. In all four years the expected total number of scallops in the Red Cliff ground was greater than that for NW Peron (see Table 10). This feature was particularly pronounced in the year 2002, where the total number of scallops in NW Peron was 12% of the estimated number of scallops in the combined Red Cliff and NW Peron ground. In the remaining years the percentage fluctuated about the 30% mark.

Table 10. Estimated percentage of scallops by fishing ground

	2000	2001	2002	2003
Red Cliff	66%	72%	88%	69%
NW Peron	34%	28%	12%	31%

4 Prediction of catch

Currently prediction of the expected annual scallop catch is based on a regression of the actual catch of previous years against the scallop index of the corresponding survey years (Joll and Caputi 1995). The scallop index is computed as the average standardised (as in Eq. (1) of Sect. 2) survey catch in the combined NW Peron and Red Cliff ground. The index treats Red Cliff and NW Peron as a whole and differences in the index between the two fishing grounds are disregarded. The predicted catches for the following years are used to set the opening date for the fishery. If predicted catch is high, an early opening date is set, while for low expected catch a late opening date is chosen. From Table 10 in Section 3 it is apparent that scallop density in the Red Cliff ground is higher than in NW Peron, even though there may be local more dense pockets in NW Peron, as was the case in 2000 and 2001. This may indicate a need to treat NW Peron and Red Cliff separately when setting the opening date.

The contributions to the expected total catch by size class for the two grounds are given in Table 11. In Red Cliff the expected contribution of recruits to the total catch exceeded 70% except in 2000 and in NW Peron this was the case in 2000 and 2001. Setting of the opening date could be further refined by taking into account the percentage contribution of recruits and residuals to the total catch by fishing ground.

Table 11. Expected percentage of scallops by size class and fishing ground

Year	RCRec	RCRes	NWPRec	NWPRes
2000	63	37	95	5
2001	73	27	88	12
2002	90	10	59	41
2003	80	20	68	32

RCRec=Recruits, Red Cliff, *RCRes*=Residuals, Red Cliff, *NWPRec*=Recruits, NW Peron, *NWPRes*=Residuals, NW Peron

To derive a method of setting the opening date based on the spatial estimates for the two size classes, we define abundance as large, if the expected percentage lies above 70%, moderate if it lies between 30 and 70% and small otherwise. The spatial maps of recruits and residuals densities for 2001 in Fig. 5. show this classification for each location.

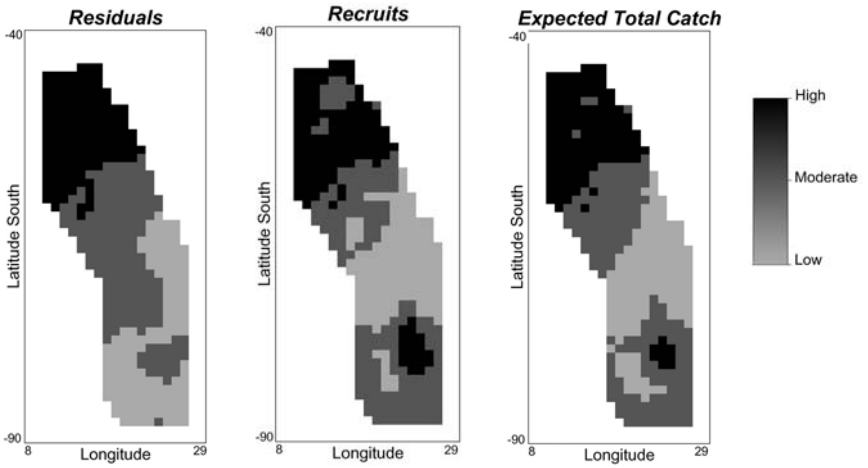


Fig. 5. Spatial maps of residuals, recruits and expected total catch classified as high (above 70th percentile, low (below 30th percentile) or moderate (between 30th and 70th percentile)

First, the current practice scallop index could be used to determine if the expected catch is to be classed as high, moderate or low to decide on an early or late opening date. Then the opening date can then be adjusted to take into account the

relativities between the two size classes. A template indicating possible decisions is shown in Table 12.

Table 12. Possible refinement startegy for setting opening dates

Recruits \ Residuals	High	Moderate	Low
High	open earlier	no change	no change
Moderate	open earlier	no change	open later
Low	open earlier	no change	open later

For the year 2002 the use of this method would have led to an early opening in the north, the opening at the time indicated by the index in the centre and a later opening date than derived from the index in NW Peron.

5 Comparison with actual catch

Fourteen boats with class A licenses (scallop only) and 27 with class B licenses (scallop and prawn) are eligible to fish for scallops in Shark Bay. The annual catch is highly variable, and ranged from 121 to 4414 tonnes meat weight in the last 20 years (Department of Fisheries 2002). The total tonnage of scallops caught in Red Cliff and NW Peron is given in Table 13 together with the contribution from the scallop fleet.

Table 13. Total scallop catch in tonnes meat weight (percentage contribution of scallop fleet to total catch in brackets)

	2001	2002	2003
All boats	205	264	54
Scallop boats	83.3 (41%)	163.3 (62%)	24.8 (45%)

The catch data (in tonnes meat weight) discussed here are those for the scallop-only fishing fleet. For each datum the position at the start of the trawl, the number of shots (a shot is the activation of the trawl gear), the total duration, the total meat weight and the date of the trawl were recorded. For the purposes of this study the temporal aspect was ignored. The duration of the scallop fishing season ranged from 2 weeks in 2003 to 6 weeks in 2001. The actual area fished by the scallop fleet varied from year to year and comprised 30% of the total available area in 2001, 14 % in 2002 (Kangas and Sporer 2002, 2003) and 4 % in 2003 (Kangas, pers. comm.). Consideration of the catch locations of the scallop fleet for 2001 to 2003 indicates that there was a tendency for the scallop fishing fleet to concentrate in the Red Cliff ground. In fact, in 2001 and in 2003 it was the case that 93% of all trawls fell within Red Cliff. In 2002 this percentage was 88%.

In 2001 part of the area fished was not contained within the region for which density estimates were derived using the survey locations, but lay closer to the coast. For a qualitative comparison of the commercial catch data with the density

estimates only data with locations in the estimation grid were considered. The catch was converted to a catch per unit effort measure with the unit time set equal to the duration to the survey trawl (i.e. 20 minutes). The summary statistics for the catch per unit effort for the years 2001 to 2003 are given in Table 14. The number of shots in 2003 is much smaller reflecting the shorter duration of the scallop fishing season in Shark Bay.

Table 14. Descriptive statistics, catch per unit effort (kg/20 min)

	2001	2002	2003
Mean	6.6	8.0	34.1
Standard Deviation	13.6	8.8	58.8
Minimum	0.6	0	0.7
Median	3.8	4.87	15.0
Maximum	151. 7	78.4	344
Count	287	531	42

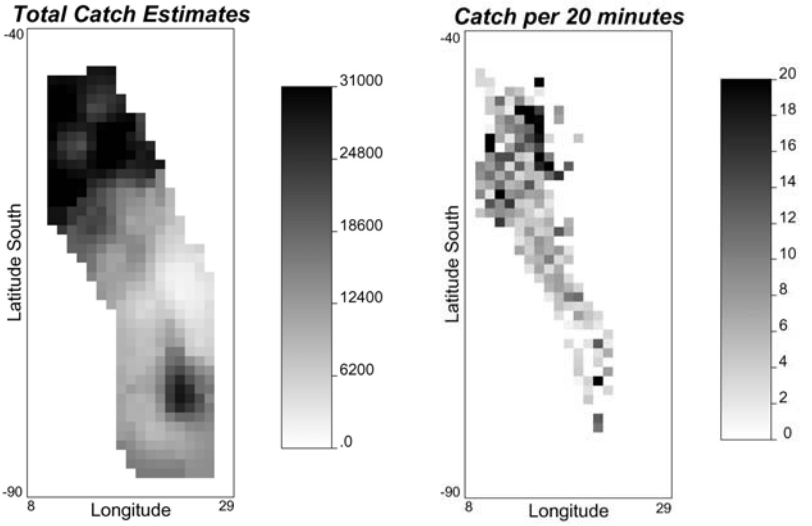


Fig. 6. Spatial maps of total catch density estimates 2001 and catch per unit effort per square nautical mile 2002

The catch per unit effort data were further standardised by moving windows to represent mean catch per unit effort per square nautical mile. The spatial maps of the expected total catch for 2001 and the subsequent (2002) fishing season are shown in Fig. 6 and indicate that the estimates derived from the survey adequately predict locations of large abundance.

6 Concluding discussion

In this paper our objective has been to analyse scallop survey data to help inform fishery management decisions on fishing opening times for the Red Cliff and NW Peron fishing regions. We have seen that the scallop survey data are amenable to analysis by intrinsic geostatistics and we have been able to identify substantial differences in both scallop settlement and scallops-only boats fishing behaviour between the Red Cliff and NW Peron fishing regions and to question the assumption that the results from these two regions be taken together when deciding on the starting date and length of the scallop fishing season. In addition, spatial maps of residuals and recruits density estimates were seen to open up possibilities for the refinement of current practice for setting the opening date for the scallop fishery.

Acknowledgements

The authors acknowledge the assistance *of Errol Sporer* in the co-ordination of scallop surveys and the logbook program and *Joshua Brown* and *Gareth Parry* from the WA Marine Research Laboratories who extracted the survey and logbook catch and effort data.

References

Chilès J-P and Delfiner P (1999) Geostatistics Modeling Spatial Uncertainty. John Wiley and Sons Inc, New York

Department of Fisheries (2002) Application to Environment Australia for the Shark Bay scallop fishery, Perth, //http:www.fish.wa.gov.au

Joll LM and Caputi N (1995) Environmental influences on recruitment in the Shark Bay saucer scallop *(Amusium balloti)* fishery of Shark Bay, Western Australia, *ICES Marine Science Symposia,* 199: 47-53

Kangas M, Sporer E (2002) Shark Bay scallop managed fishery status report, in Penn J (ed) State of the fisheries report 2001/2002, //http:www.fish.wa.gov.au , 52-54

Kangas M, Sporer E (2003) Shark Bay scallop managed fishery status report, in Penn J (ed) State of the fisheries report 2002/2003, //http:www.fish.wa.gov.au , 58-60

A spatial extension of CART: application to classification of ecological data

L. Bel[1], J.M. Laurent[2], A. Bar-Hen[3], D. Allard[4] and R. Cheddadi[2]

[1]Probabilités, Statistique et Modélisation, Université Paris-Sud, Orsay, France
[2]Institut des Sciences de l'Evolution, CNRS and Université Montpellier II, France
[3]Institut National d'Agronomie, Paris, France
[4]Unité de Biométrie, Institut National de la Recherche Agronomique, Avignon, France

1 Introduction

Paleoecology is the science of reconstructing past environments using fossil materials of plants, animals, or other indicators of past environments. These studies are useful for understanding the dynamics of ecosystem changes and thus for predicting their future evolution. They also provide tools to reconstruct conditions that existed before the impacts of industrialized societies on natural ecosystems.

As far as vegetation is concerned, the basic idea is to consider that plant dispersal is in equilibrium with climate and that it is sensitive to climate change. The geographic distribution of pollen frequencies is supposed to reproduce more or less properly plant ranges. Several biases disturb this representation. Some of them are: different pollen production rate between taxa and unequal transport of pollen grains depending on their shapes and densities. Rather than considering specific pollen taxa that may be rare, one approach is to combine taxa with similar environmental envelopes, thus defining Bioclimatic Affinity Groups (BAGs) of plants. The use of these functional groups provides more complete information but increases distortions between observed ranges and those reconstructed from pollen data. Our goal is to provide palaeoecologists with a tool to discriminate between absence or presence of plants using pollen frequencies of each BAG.

We have chosen discrete vegetation information as this database is the most complete, recent and digitally available dataset. Pollen data are continuous variables. From a statistical point of view, map comparison can be considered as predicting a class of vegetation with a continuous variable. This situation is characteristic of discriminant analysis (also known as supervised classification). A direct application of supervised classification would provide biased estimates because the sampling scheme of pollen data is very irregular. By giving the same weight to every record, region with high sampling density would be over-weighted when constructing the discriminant rule. An analogous question in geostatistical literature is the estimation of the regional average of spatial dependent data. Kriging is a classical tool within this framework. We will use the kriging weights to equili-

brate the observed points such that every part of the domain under study has a comparable importance in the construction of the discriminant rule. We propose an adaptation of a non parametric discriminant technique (CART) to the case of spatial data. Our approach will be illustrated with simulated and real data.

2 Methodology

Prediction of a discrete variable by a continuous variable is a classical task of discriminant analysis. A powerful method for discriminant analysis is the Classification and Regression Trees (CART; Breiman *et al.*, 1984). We first describe the method in the usual framework of independent data. After briefly recalling kriging of a regional average, we present how CART can be adapted to the case of spatially dependent data.

2.1 CART

CART is a rule based method that generates a binary tree through binary recursive partitioning, a process that splits a node based on yes/no answers about the values of the predictors. Each split is based on a single variable. Some variables may be used many times while others may not be used at all. The rule generated at each step maximizes the class purity (or minimizes the class heterogeneity) within each of the two resulting subsets. Each subset is split further based on independent rules.

The splitting criterion is based on purity criterion. Let us denote l and m the indices of the two leaves generated by the split of the node k and let n_{il} be the number of observations of leaf l that belong to class i. n_{+l} is the number of observations of leaf l and $p_{il}=n_{il}/n_{+l}$ is the proportion of observations from class i within leaf l. We only consider split with $n_{+l}>0$ for all leaves. The two most popular heterogeneity criteria are the entropy and the Gini index. Since the entropy imposes that $n_{il}>0$ we only consider the Gini index:

$$D_l=\sum_{i\neq j} p_{il}\,p_{jl}=1-\sum_i p_{il}^2 \tag{1}$$

The Gini index is 0 when there is only one class present in leaf l, it is maximum when all classes are present with the same probability.

Among all partitions of the explanatory variables at the node k (here pollen record) the aim of CART is to maximize the heterogeneity difference

$$D_k-\left(\frac{n_{+l}}{n_{+k}}D_l+\frac{n_{+m}}{n_{+k}}D_m\right). \tag{2}$$

The procedure is finished when there is no more admissible splitting. Each leaf is affected to the most present class (conditional mode). This rule can be adapted in case of a cost function. In general the final tree T_n overfits the available data and the error of prediction $R(T_n)$ is typically large. In designing a classification tree,

the ultimate goal is to produce from the available data a tree T whose probability of error prediction $R(T)$ is as small as possible. Thus, in a second stage the tree T_n is "pruned" to produce a subtree whose expected performance is superior to T_n. If Y is the discrete variable and X the continuous variable, then $R(T)=\mathrm{E}[T(X)\neq Y]$. Since the distribution of Y and X is generally unknown, the pruning is based on the empirical risk

$$\hat{R}(T)=\frac{1}{n}\sum_{\alpha=1}^{n}\mathrm{I}(T(X\alpha)\neq Y\alpha), \tag{3}$$

where $\mathrm{I}(A)$ is the indicator function: $\mathrm{I}(A)$ is equal to 1 when A is true, and to 0 otherwise. If the same data are used to construct and to prune the initial tree, $\hat{R}(T)$ underestimates the risk of large subtrees. On the other hand, using separate data sets for growing and pruning is not feasible and additional data are difficult to obtain. The CART pruning algorithm seeks to balance optimistic estimates of empirical risk by adding a complexity term that penalizes larger subtrees. Thus the final tree is

$$S=\underset{T\in Tn}{\operatorname{argmin}}\left\{\hat{R}(T)+\omega.\mathrm{size}(T)\right\} \tag{4}$$

where size(T) is the number of nodes of the tree T. We choose ω=0.01, the default value of *rpart* library in the *R* software.

The main drawback of CART models is that when there are more than just a handful of predictor variables or cases to classify the generated models can be extremely complex and difficult to interpret. This is exemplified by the work on Australian forests by Moore *et al.* (1991), generating a tree with 510 nodes for just ten predictors. Such complexity makes the tree impossible to interpret, whereas in many studies interpretability is a key issue.

2.2 Kriging

For estimating the regional average $\hat{\mu}$ of the variable X over a domain D, using spatially dependent data $(X_\alpha)_{\alpha=1,n}$, the best (unbiased with minimal variance) linear estimator is the kriging (Wackernagel, 2003),

$$\hat{\mu}=\sum_{\alpha=1}^{n}\lambda\alpha X\alpha \quad \text{with} \quad \mathbf{\Lambda}=\begin{pmatrix}(\lambda\alpha)_{\alpha=1,n}\\ m\end{pmatrix} \quad \text{such that} \quad \widetilde{\mathbf{C}}\mathbf{\Lambda}=\widetilde{\mathbf{c}}\,, \tag{5}$$

where

$$\widetilde{\mathbf{C}}=\begin{pmatrix}\mathbf{C} & \mathbf{1}'\\ \mathbf{1} & 0\end{pmatrix} \qquad \widetilde{\mathbf{c}}=\begin{pmatrix}c(x\alpha,D)\\ 1\end{pmatrix}, \tag{6}$$

and $\mathbf{1}$ is a vector of ones of length n, m is a Lagrange multiplier, $\mathbf{C}$ is the covariance matrix with elements $C_{\alpha,\beta}=\mathrm{Cov}(X_\alpha,X_\beta)$ and $c(x\alpha,D)=\frac{1}{|D|}\int Cov(x\alpha,y)dy$ is the average of the covariance between the data point x_α and a point y of the domain D.

The weights λ_α are called the kriging weights, and when assumptions of stationarity and isotropy are made, they only depend on the relative position of the data.

2.3 Spatial CART

When the sampling design is very irregular, using the same weights for all samples leads to uneven weight for regions of equal area: intensively sampled regions will be over-represented and regions with sparse sampling will be under-represented. To correct for this bias, it seems natural to give a smaller weight to clustered samples as they bring similar information; conversely, isolated samples carry much more information and need to have larger weight in the decision rule. Kriging of the mean (or kriging of the regional average) provides natural and optimal weights. It can be interpreted as a "natural declustering": the weight of clustered samples tend to be small or even negative; the weight of isolated samples sufficiently remote to other samples is nearly equal to the inverse of the equivalent number of independent observations.

The CART algorithm is thus adapted so that each sample is weighted using the kriging weights above. Specifically,

$$p_{il}=\sum_{\alpha\in l}\lambda_\alpha \mathrm{I}(\alpha \text{ in class } i)\Big/\sum_{\alpha\in l}\lambda_\alpha \text{ and } n_{+l}/n_{+k}=\sum_{\alpha\in l}\lambda_\alpha\Big/\sum_{\beta\in k}\lambda_\beta \ . \tag{7}$$

The Gini index and the heterogeneity difference are then computed using these new values. For the pruning procedure, the empirical risk is $\hat{R}(T)=\sum_{\alpha=1}^{n}\lambda_\alpha \mathrm{I}(T(X_\alpha)\neq Y_\alpha)$.

The drawback of this method is that the kriging weights can be negative. The partitioning algorithm needs positive weights because they are used to calculate positive indices. Hence we will have to impose a positiveness condition on the λ_α while solving the system of kriging equations.

3 Simulations

The method described in the previous section is firstly tested on a simulated example in order to evaluate the improvement from the standard method. On a 101 x 101 grid of unit 1, a Gaussian field ε with mean 0, variance 1, and exponential covariance of range 10 is simulated.

Each point in the grid is said to belong to class 1 if $(x\text{-}50.5)(y\text{-}50.5)>0$, to class 0 otherwise. This rule generates four subsquares: the upper right and the lower left subsquares are in class 1 while the two remaining ones are in class 0.

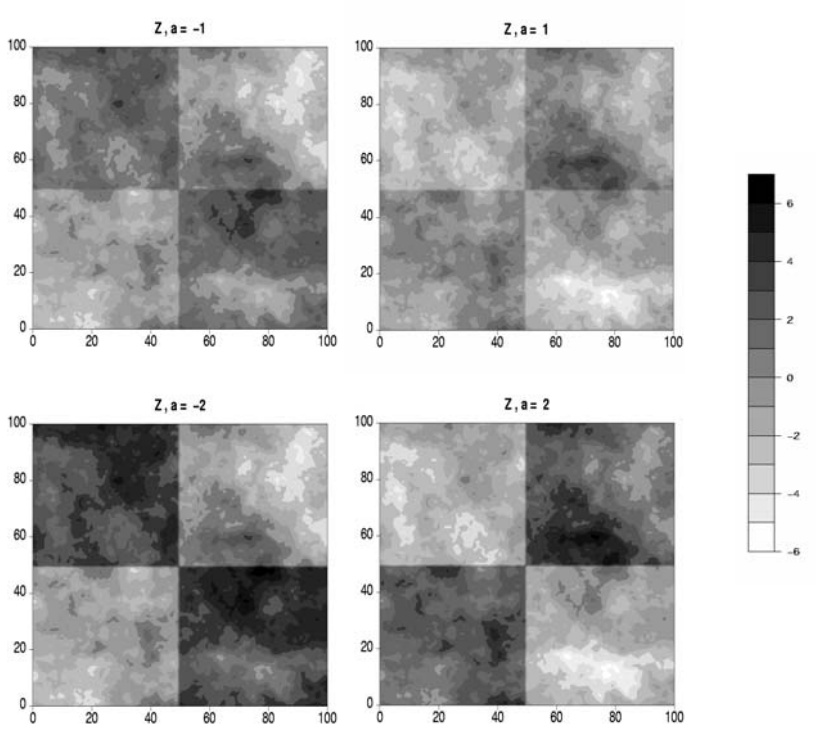


Fig. 1. The variable Z = a*sign(x-50.5)(y-50.5)+ε for various values of a, where ε is a (0,1) Gaussian random field with an Exp(10) covariance function

The variable Z used to classify the data is $Z = a$*sign(x-50.5)(y-50.5)+ε. Since E[ε]=0, we expect the classification rule to allocate to class 1 the points with Z>0 when a>0 (resp. points with Z<0 when a<0) and to class 0 otherwise. We consider four different values for a: -2, -1, 1 and 2, depicted Fig. 1. The sample set is made of 100 points, 75 of them randomly sampled in the entire square, the 25 others in a subsquare of length 10 in the upper right corner (see Fig. 2a). We choose this subsquare because in this area ε<<0. Therefore if a>0, in this subsquare Z can still be negative for many points and this will perturb the classification rule if they all have the same weight. This discrepancy will be minimized if the sum of their weights is not too large. Conversely if a<0, it will be extremely rare to get a positive Z and the clustered points will not perturb the classification rule even with important weights. The test set is made of 100 other points randomly sampled in the entire square (see Fig. 2a).

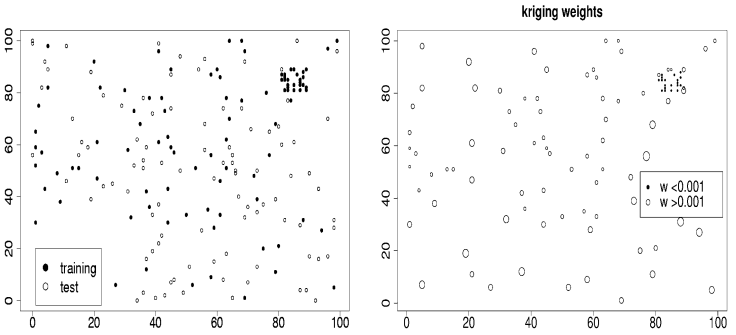


Fig. 2. a) Location of learning and test points; b) Kriging weights for the sample set. Size of points is proportional to the weights

Fig. 2b shows the kriging weights for the sample set when estimating the average $\hat{\mu}$ of Z over the simulation domain. As expected isolated sample points have larger weights than clustered ones. Out of the 25 clustered sample points, 16 of them have null weight, while the other ones have small weights; the latter are usually located on the boundary of the cluster.

Fig. 3 shows the partition generated by the two algorithms for the four different values of a. As expected, when a is negative both algorithms lead to the same partition. In this case the clustered points help for a good classification rule. When a is positive the two methods give quite different results. For a=1 Spatial CART gives a single split at the value Z=-0.09 which is close to the expected value Z=0, while standard CART gives 6 more splits, presumably trying to adapt to the noise. For a=2 both methods have an unique split, Standard CART at the value Z=-0.73, Spatial CART at the value Z=-0.47 which is closer to 0.

To evaluate the quality of the classification procedure we compare these partitioning rules with the theoretical ones on the test set and we look at the number of misclassified points. One may notice that misclassification can have two sources: (i) classification rule can provide wrong prediction if the threshold value is not 0, (ii) perturbation can reverse the sign of Z and change the prediction obtained from the classification rule. Both sources of error will be studied.

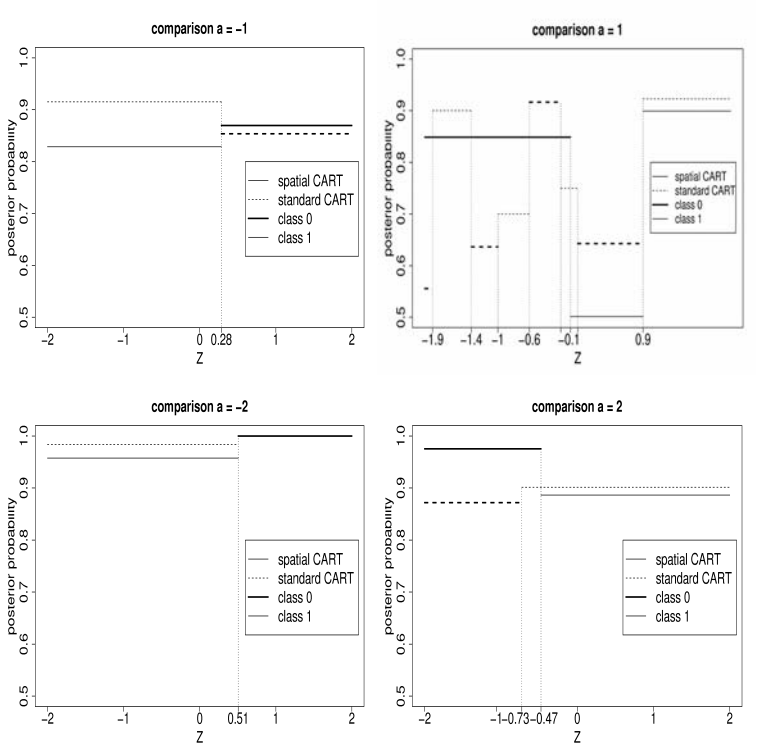


Fig. 3. Comparison of Spatial CART and standard CART on simulations for different values of *a*

For each value of *a*, Table 1 indicates the value of *Z* for all missclassified points, together with the true class (0 or 1) and the origin of the misclassification: C when it is due to the classification rule (i.e., *Z* is of the right sign), and N when it is due to the noise (i.e., *Z* has the wrong sign). When $a<0$, the discrimination rule is strictly identical for standard CART and Spatial CART and leads to only 8 misclassifications (3C and 5N) for $a=-2$. For $a=-1$ the variable *Z* does not discriminates as well and there are many misclassifications due to the presence of noise. In both cases the clustered points do not perturb the classification rule.

When *a* is positive, the clustered sample points have negative *Z* values, instead of positive values expected for points in class 1. Since they have small weights in the Spatial CART method their (bad) influence is lower in the analysis, and most misclassifications for Spatial CART are due to the noise, especially when a=1.This simulated example shows that clustered samples perturb the classification rule of standard CART when the noise has a sign opposite to the class difference. In this case, the spatial extension presented in Section 2 improves the results. When the noise has the same sign as the class difference, both methods are equivalent.

Table 1. Misclassified points for various *a*. The value of the continuous variable *Z* is given with the true class of the point. The origin of the misclassification is indicated by a letter: C for "cut" and N for "noise".

a=-2	*a*=-1	*a*=1		*a*=2	
BOTH	BOTH	Spatial CART	Stand. CART	Spatial CART	Stand. CART
0.1 (0) C	-0.39 (0) N	-0.05 (0) C	-0.05 (0) C	0.09 (0) N	-0.67 (0) C
-0.02 (0) N	-0.9 (0) N	0.01 (0) N	0.81 (1) C	0.28 (0) N	-0.55 (0) C
0.17 (0) C	-1.02 (0) N	0.33 (1) N	-0.05 (0) C	-0.1 (0) C	0.09 (0) N
0.72 (1) N	0.18 (0) C	-0.05 (0) C	-1.82 (0) C	-0.42 (0) C	0.28 (0) N
-1.29 (0) N	0.18 (0) C	-1.02 (1) N	-0.87 (0) C	0.4 (0) N	-0.1 (0) C
1.04 (1) N	-0.2 (0) N	-1.47 (1) N	-1.82 (0) C	-0.09 (0) N	-0.42 (0) C
	0.05 (0) C	0.14 (1) N	-1.57 (0) C	0.21 (0) N	0.4 (0) N
	1.21 (1) N	-0.77 (1) N	-0.05 (0) C	-0.53 (1) N	0.09 (0) N
	0.33 (1) N	-0.05 (0) C	0.75 (1) C	-0.1 (0) C	0.21 (0) N
	0.58 (1) N	0.45 (0) N	-1.51 (0) C	0.36 (0) C	-0.67 (0) C
	0.71 (1) N	0.14 (0) N	-1.51 (0) C	-0.42 (0) C	-0.1 (0) C
	0.11 (0) C	1.09 (0) N	-0.81 (0) C		-0.36 (0) C
	0.52 (1) N	1.28 (0) N	-0.87 (0) C		-0.67 (0) C
	-0.83 (0) N	0.2 (0) N	-1.89 (0) C		-0.42 (0) C
	-0.26 (0) N	0.9 (0) N	0.49 (1) C		
	1.72 (1) N	-0.52 (1) N	-0.94 (0) C		
	0.58 (1) N	0.58 (0) N	-0.75 (0) C		
	-2.29 (0) N	1.4 (0) N	1.09 (0) N		
	2.04 (1) N	1.09 (0) N	1.28 (0) N		
	0.96 (1) N	1.21 (0) N	0.9 (0) N		
	1.34 (1) N	-0.33 (1) N	-0.52 (1) N		
	-0.2 (0) N	0.33 (0) N	1.4 (0) N		
	0.39 (1) N	-1.53 (1) N	1.09 (0) N		
	-1.02 (0) N	0.9 (0) N	1.21 (0) N		
	0.9 (1) N	-0.71 (1) N	-0.33 (1) N		
	-0.08 (0) N	0.2 (0) N	0.9 (0) N		
	-0.64 (0) N	0.64 (0) N	-1.51 (0) C		
	-0.33 (0) N	-0.2 (1) N	-0.81 (0) C		
		0.33 (0) N	-0.45 (1) N		
		0.58 (0) N			
		-0.45 (1) N			
8 3C/5N	28 4C/24N	31 3C/28N	29 19C/10N	11 5C/6N	14 9C/5N

4 Application to ecological data

Future climate change will strongly affect vegetation distribution (Beerling *et al.* 1997). Reconstructing modern and past plant cover is essential to understand vegetation dynamic and to predict their future ranges under changing climate (IPCC 2001). Pollen data are one of the most appropriate proxies to reconstruct modern and past vegetation. They are abundant in fossil records but they give a biased image of surrounding vegetation. Pollen records depend on: population distance from sampling site, population density, pollen production rates (rates are different between species, individuals and even between years), transport (depending on pollen morphology and density) and preservation (more or less resistant according to the thickness of their envelope). Palynological species were gathered into functional groups of plants, the Bioclimatic Affinity Groups of plants (BAGs) (Laurent *et al.*, in press). Laurent *et al.* (in press) georeferenced geographic ranges of 320 European species of plants and gathered these data following palynological taxonomy. Combination of taxa ranges with climate variables (New *et al.* 1999) provided potential distribution of taxa. These inferred ranges correctly reproduced observed ones. Twenty five BAGs were created using hierarchical cluster analyses on potential ranges. They are characterized by different geographical ranges and climatic tolerances and requirements.

The distribution of taxa (families, genera or species) is georeferenced from the database SOPHY (http://sophy.u-3mrs.fr/sommaire.html). For each point of the grid, a binary variable indicates the presence or absence for each species. These binary values for plants belonging to the same BAG were averaged to create one single map for each group.

356 pollen samples were evenly collected in France. We added pollen counts of samples collected at the same place. The resultant 154 samples provided pollen percentages of BAGs. In this work we focus on BAG 8, which is principally located around Mediterranean sea and near the Atlantic coast of south-western France (Fig. 4a).

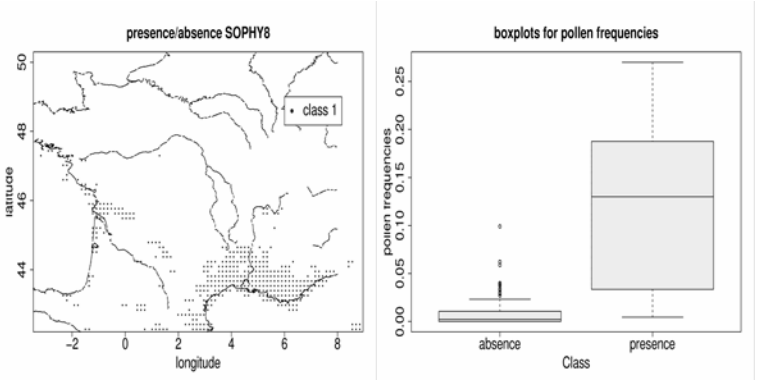


Fig. 4. a) Presence and absence of specie from BAG8; b) boxplots of pollen frequencies according to presence or absence

We affect to each sample site the nearest point class value and we study the link between pollen percentage and the presence/absence of the BAG. Only 11 sites are in class 1, in which the pollen mean rate is 0.126 with a standard deviation of 0.09. The mean for class 0 is 0.008 with a standard deviation of 0.01. Hence class 0 implies low pollen percentage but the converse is not true. Kriging weights of the average of the pollen frequencies over France are computed according to Eq. 5 and 6 for a fitted exponential variogram of range 50km.

Fig. 5a shows the kriging weights for this sampling design. Two points with important weights occur where BAG 8 is present: one in Oriental Pyrenees and the others in the Southern Alps.

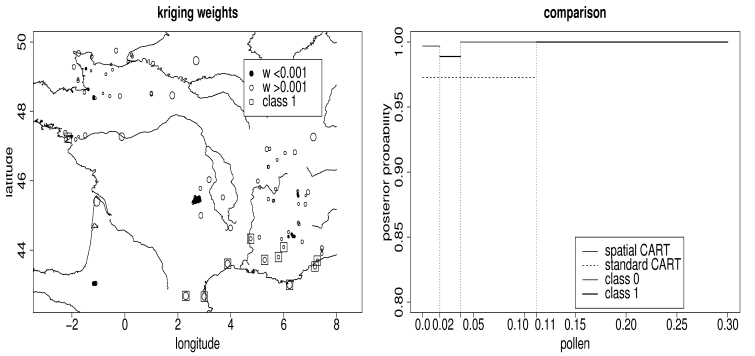


Fig. 5. a) Kriging weights for Spatial CART; b) classification for Standard and Spatial CART

Fig. 5b shows the comparison between the two methods. Standard CART gives only 1 split when the pollen percentage is 0.11, that is it missclassifies four presence points with low pollen percentage and none of the absence points. Despite these four points, this threshold seems relevant to palynologists. Pollen grains from these four sampling sites come from anthropised environments, where population of that BAG is restricted to small inhabited areas. This probably explains such low pollen rates. The sampling design has the particularity that many points are closely clustered especially in Occidental Pyrennees where 39 points are within a distance of 3 kilometers. They are located at points where BAG 8 is not present (0) and pollen percentages are included within 0.002 and 0.06. With standard method, they are classified as absence points. Spatial CART predicts a supplementary interval [0.0169 ; 0.037] in class 1. In this range lie 18 Occidental Pyrennees points with null weight and 3 presence points. These 18 samples in the Pyrennees are in fact almost 18 times the same sample. Hence Spatial CART generates approximatively only two misclassifications, one in the Pyrennees and one false negative on the Atlantic coast, with a very low pollen percentage (0.0004). The weighted misclassification $\sum_{\alpha\, miscl.} \lambda_\alpha$ is equal to 0.01 (0.0001 for Pyrennean sites).

In this case, the use of Spatial CART method allows palynologists to highlight disturbed sampling sites. These sites are located in altitude, where pollen transport may be very different from one valley to another. The effect of ascension winds is known to disturb pollen registration. This implies that they have to be excluded from future analysis.

5 Discussion

We proposed a spatial extension of the CART algorithm in which the samples are weighted according to the kriging weights of the regional average. On simulations it has proven to improve the classification rate in presence of clustered samples. On the pollen data, Spatial CART was useful to decrease the importance of clustered samples in the Pyrennees. The gross number of misclassified points seems to increase, but if the kriging weights are taken into account for counting the number of misclassification, this number actually decreases. Our simulations show the importance of the classification rule in presence of clustered samples.

References

Beerling DJ, Woodward FI, Lomas M, Jenkins AJ (1997) Testing the responses of a dynamic global vegetation model to environmental change: a comparison of observations and predictions. Global Ecology and Biogeography Letters, vol 6, 439-450

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees, Wadsworth, Belmont

Moore DM, Lee BG and Davey SM (1991) A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environ. Manage. Vol 15, 59-71

Intergovernmental Panel on Climate Change (2001) Climate change 2001: The scientific basis, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds). Cambridge University Press, Cambridge

Laurent JM, Bar-Hen A, François L, Ghislain M, Cheddadi R (2004) Bioclimatic Affinity Groups of European plants defined by climate seasonality: statistical analysis for vegetation modeling. Journal of Vegetation Science, in press

New M, Hulme M, Jones P (1999) Representing twentieth-century space-time climate variability. Part I: Development of a 1961-90 mean monthly terrestrial climatology. J. Clim., vol 12, 829-856

Wackernagel H (2003) Multivariate Geostatistics, 3rd ed., Springer, Berlin.

Using a Markov-type model to combine trawl and acoustic data in fish surveys

M. Bouleau and N. Bez

Ecole des Mines de Paris, Centre de Géostatistique, 35 Rue Saint Honoré, F-77305 Fontainebleau, France, tel: +33 1 64694778, fax: +33 1 64694705, e-mail : mireille.bouleau@ensmp.fr, nicolas.bez@ensmp.fr

1 Introduction

Fisheries management is based on estimations of fish abundances derived from commercial catches. Models used to produce these estimates are, most of the time, tuned with indices of abundances estimated from scientific surveys. In the Barents Sea used for application in this paper, the surveys consist in deploying a net every twenty nautical miles (n.mi.). With the objective to compensate for this large distance between catches, acoustic measurements are also collected all along the vessel track when the vessel is shipping from one station to the next. This additional measure of fish concentration does not actually capture fish but estimate their number through their echoes (echoes of all the fish present in the insonified cone beneath the boat). Acoustic echoes are generally integrated over regular distance bins (say one nautical mile) and provide a spatially very dense sampling of fish distribution but different in nature from the spare tows. The purpose of the study is to take as much as possible advantage of this additional information for estimation and mapping purposes.

Here, we consider a partially heterotopic sampling where the target variable is observed on a subset of the auxiliary variable samples. Theoretically cokriging allows performing estimates in such heterotopic configurations. However it can become difficult when the number of samples is high or/and when spatial structures are difficult to model. In such cases, simplifications either assumed or data controlled, are welcome. For instance, for two variables, a Markov-type model, also called model with orthogonal residual, is a well-known simplification (Rivoirard 2001) as one of the two variables is self-krigeable. Two kinds of Markov-type models are mentioned in literature (Schmaryan and Journel 1999): when the cross structure is proportional to the structure of the auxiliary variable or when it is proportional to the structure of the target variable. Here, we consider the first case, The trawl variable is decomposed into an acoustic and a residual components, these two components being spatially uncorrelated, but not independent. In this model, the trawl variable is subordinated to the acoustic, which is the master variable.

After a quick presentation of the data and of the notations, this paper presents the problems of the practical implementation of such a model in the particular case of strong heterotopy (hypothesis testing, structural tools, skew distributions).

2 Data and Notations

Six scientific Norwegian winter surveys (1997-2002) in Barents Sea are used. The sampling scheme (i.e. the tow locations) is targeting a regular grid with a haul every 20 n.mi (Fig. 1a). Sampling size is quite large as surveys get between 200 and 300 hauls. The mean towed distance is 1 n.mi. The acoustic data turned into Nautical Area Scattering Coefficient (NASC) and expressed in $m^2 \cdot n.mi^{-2}$ (MacLennan *et al.* 2002) are collected continuously along the vessel track during and between trawl hauls (Fig. 1b). In this study, acoustic echoes are integrated vertically over the first 40 meters above the bottom (this was found to provide the larger correlation between the two variables) and horizontally over fixed distance bins of 1 n.mi. Given this latter parameter, between 5000 and 7000 acoustic records are available in each survey.

To get variables with comparable units, the fish catches are turned into an equivalent acoustic energy, i.e. the acoustic energy that the fish caught in the trawl hould have generated. Because fish characteristics influence this transformation, two groups of fish have been used: demersal (bottom) fish and pelagic (mid water) fish. For each group of fish, the equivalent NASC of the corresponding fish in the net is provided. The trawl variable will refer alternatively to the demersal or the pelagic equivalent NASC depending on which of these two variables happen to get larger correlation with the acoustic variable.

We get then two measurements of fish abundance (trawl and acoustic) available at equivalent supports (1 n.m.), expressed in the same similar units but sampled differently. They are modelled by two random functions: $T(x)$ the trawl and/or the target variable available at the sampling locations, $x_\alpha \in \{stations\}$ and $A(x)$ the acoustic and/or the auxiliary variable available at the sampling locations $x_\alpha \in \{stations + underways\}$.

When sampling skew distributions, the experimental variance varies considerably with the number of samples, especially when this number is low (for a given number of samples, the sampling fluctuations of the variance are all the more important that the variance is large). We observe (Table 1) that the ratio k^2 between the variance of the underway acoustic observations (few thousands data) and that of on station observations (few hundreds data) diverges from 1. This problem is referred to hereafter as “the variance discrepancy problem”. Proportional effects are examples of this problem.

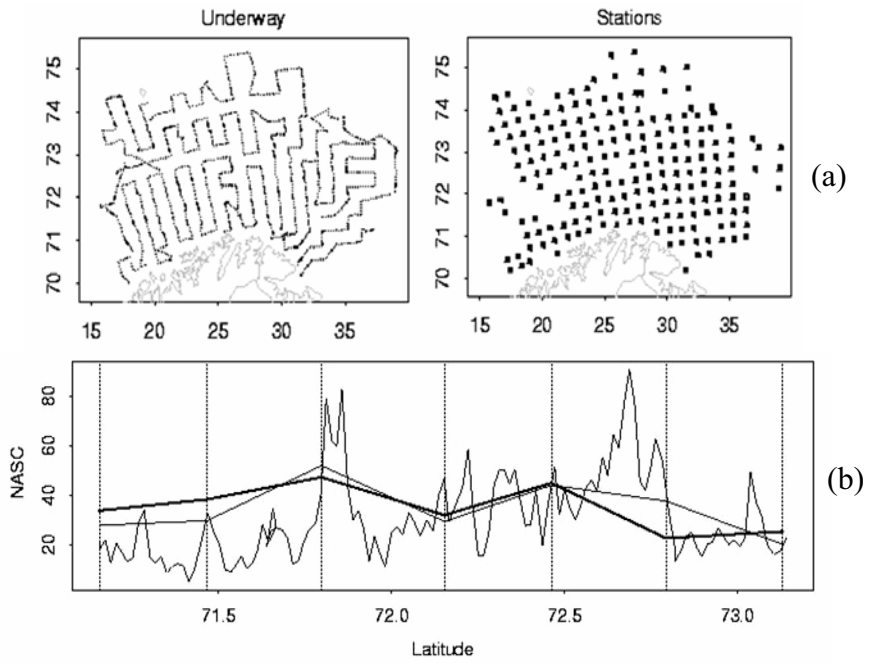


Fig. 1a. Locations of underway recordings (left) and of stations (right). Survey 1998. X-axes unit is in degrees of longitude and Y-axes unit is in degrees of latitude. **b** Representation of a N-S section of the vessel track. The vertical dotted lines represent the stations locations. The fluctuant slight curve is the acoustic underway, the slight line joins the acoustic on-stations values and the bold line joins the demersal NASC-equivalent values collected on-stations. Distances are in degrees of latitude.

Table 1 Ratio between the variance of the underway acoustic observations and the variance of the on station acoustic observations

Year	$k^2 = \frac{\mathrm{var}\left(A(x_\alpha), \alpha \in underway\right)}{\mathrm{var}\left(A(x_\alpha), \alpha \in station\right)}$
1997	1.33
1998	1.83
1999	2.23
2000	1.35
2001	3.55
2002	2.65

Let us consider the entire line followed by the vessel during a survey. This line is made of N underway acoustic values located at the centre of their segment of 1 n.mi. each. Let us also consider a subset of the n segments, to be considered as the

stations following a regular sampling with random origin (given the sampling design N = 20.n). In that case, the additive relation of the dispersion variances applies:

$$D^2(segment|line) = D^2(segment|stations) + D^2(stations|line) \quad (1)$$

The term of the left-hand side is the average variance of underway data while the first term on the right-hand side corresponds to the average variance of station data. In case of pure nugget effect, the third term equals $nugget \cdot \left(\frac{1}{n} - \frac{1}{N}\right)$ and is negligible with regards to the other terms. In this study, we have assumed that the spatial structure is short enough to neglect the dispersion variance of the stations in the line. This amounts to assume that the variances of the underway data and of the on station data are similar on average. Actual differences are then explained by the sole statistical fluctuations and are corrected for by a multiplicative term k^2 (see part 4.1 variance rescaling).

3 Methods

3.1 Model and estimation

One can show (e.g. Rivoirard, 2001) that if the acoustic is autokrigeable, its cross covariance with the trawl variable is proportional to its covariance:

$$C_{A,T}(h) = \alpha\, C_A(h) \quad (2)$$

and the trawl variable is linearly related to the acoustic up to an additive spatially orthogonal residual *R(x)*:

$$T(x) = \alpha . A(x) + \beta + R(x) \quad (3)$$

$$C_{A,R}(h) = 0 \quad \forall h \quad (4)$$

The target variable is then subordinated to the auxiliary but master. This model has a "Markov-type" property as, in Gaussian case with known means, *A(x+h)* and *T(x)* are independent when *A(x)* is given (conditional independence, Chilès and Delfiner, 1999). More generally the screen effect makes the cokriging weight of *A(x+h)* equal to zero when *A(x)* is known, whatever the histogram of the data.
The model is factorized with the two factors *A(x)* and *R(x),* and the cokriging of the target variable reduces to the sum of two krigings as the acoustic variable is known at any location where the trawl variable is known:

$$T^{CK}(x_0) = \alpha\ A^K(x_0) + \beta + R^K(x_0)$$

$$\text{where} \begin{cases} A^K(x_0) = \sum_{\substack{\text{stations} \\ +\ \text{underways} \\ \in \\ \text{neighbourhood}}} \lambda_\alpha^A A(x_\alpha) \\ R^K(x_0) = \sum_{\substack{\text{stations} \\ \in \\ \text{neighbourhood}}} \lambda_\alpha^R R(x_\alpha) \end{cases} \tag{5}$$

and the cokriging variance is:

$$\sigma_T^{CK}(x_0) = \alpha^2\ \sigma_A^K(x_0) + \sigma_R^K(x_0) \tag{6}$$

The constant β is in practice filtered by the ordinary kriging of the residual, and does not need to be assessed.

The estimation of the target variable at a point where the acoustic is known (underway) only uses the acoustic at the target point and on station (by the residual). Then, in the Markov-type model, cokriging is multi-collocated: for estimating an underway point, the auxiliary variable is only used at the target point and on stations. It is the only case where the cokriging is collocated (Rivoirard 2001). In a different model, the previous estimation is only an approximation of cokriging.

3.2 Practical implementation in partially heterotopic samplings

Compared to cokriging in general, an advantage of the previous estimation, based on residual, is that cross structures do not need to be modelled. Cross structures only serve to experimentally test for the validity of the model.

Two tools are used to test for the proportionality between the cross and simple structures; the cross variogram and the cross covariance. The cross variogram, not restricted to stationary cases, uses, only on station data ("isotopic tool") and misses short scale structures. The cross covariance, or preferentially in strong heterotopic cases, the cross correlogram, assumes stationarity but uses all the available information ("heterotopic tool").

The advantage of the estimation based on residual (no cross structure model) is compensated by the need to estimate the parameter α. Equation 4 is general and not specific to any sampling scheme. However, in the particular case of partially heterotopic sampling, Equation 4 is viewed as an "on station" relationship to be parameterised with on station data only and applied to underway data afterwards. In this case, rescaling is required. As a matter of fact, theoretically, the cokriging estimation variance is necessarily less or equal than the kriging estimation variance as long as the same data and the same model for the target variable are used. Here, the model is parameterised on a subset of a data which happens to be less variable (variance discrepancy problem). When applied to the more variable un-

derway acoustic data, it does not protect from inconsistent estimation variances. To solve this problem, the cokriging variance has to be rescaled, so that finally we have:

$$\sigma_T^{CK}(x_0) = \frac{\alpha^2}{k^2}\sigma_{A,un}^{K}(x_0) + \sigma_R^{K}(x_0) \tag{7}$$

where $\sigma_{A,un}^{K}(x_0)$ is the acoustic kriging variance, when all the data available are used, i.e., the stations and the underways.

The estimation of the parameter α can be made by many ways theoretically equivalent. It can be estimated by the slope of the linear regression of T(x) on A(x). This approach has the advantage to allow quantifying the quality of the estimation (e.g. visual inspection of the scatter plots, R-square, etc). A weakness of the regression is that only the pairs of samples at the same location contribute to the estimation. An alternative is to use the mean ratio between the cross and simple experimental variograms computed only with data on station. The gain of this approach is to take into account all distance lags. However, no quality is directly associated to the estimation of α. To enhance the robustness of the estimate, one could have used the simple variogram for all the underway observation or cross covariances. However, the advantage of using all the data is thwarted by loss of statistical coherence. We thus chose not to retain this last estimation.

4 Results

4.1 Variance rescaling

We have simulated 500 sets of 7000 lognormal data (independently) from which 500 subsets of 300 points have been taken randomly (7000 corresponds to the number of underway samples and 300 to the number of stations in 2001).

We are in a special case of pure nugget effect in the equation (1), the variance underway and on-station have to be equal in mean.

The variance and the mean of the simulated lognormal distribution are equal to the mean and the variance of the acoustic underway in 2001 (m = 63 and $\sigma^2 = 23061$). In 80% cases, the ratio k^2 between the empirical variance of the main 7000 samples and the empirical variance of the 300 subsamples is greater than 1 (Fig. 2). The value 3.55 observed in 2001 (represented by a vertical dotted line) is quite singular but not impossible. When a large value is taken, the variance of the subsample becomes extreme because of the small number of samples.

So the observed discrepancy between the experimental variances can be interpreted as a sampling problem (heterotopic sampling of skew distributions) and are in no way particular to the data used in this study. In fact, it can be considered that the variance observed underway is more realistic as it is based on 20 times more data justifying a multiplication of on-station variance. Nevertheless to be compa-

rable to a (monovariate) kriging variance the equation (7) is based on a downscaling of the underway variance.

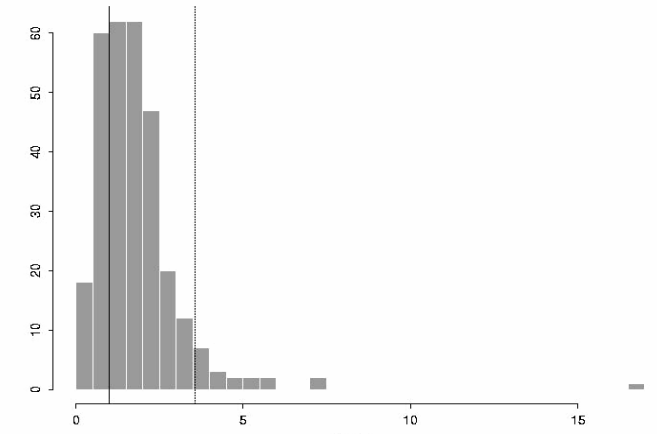


Fig. 2. Histogram of ratio between the empirical variances of the main sample (7000 points) and the subsample (300 points) for 500 draws of a lognormal distribution with the mean and the variance of the acoustic for the 2001 survey. The plain vertical line is equal to 1 and the dotted vertical line is equal to the observed ratio (3.55).

4.2 Hypothesis testing and selection of favourable cases

To test the autokrigeability assumption, experimental simple and cross correlograms have been plotted for each of the variables. Cross correlograms are potentially non symmetrical. They happened to be symmetrical and have been symmetrized before representation (Fig. 3). The single and cross correlograms have been calculated along the vessel track, i.e. in one dimension: In four surveys out of six (1997-1998-2001 with demersal catches and 2000 with pelagic catches), the Markov-type model hypothesis are grounded (Fig. 3, graphs with grey background). They have then been selected for application of a Markov type model.

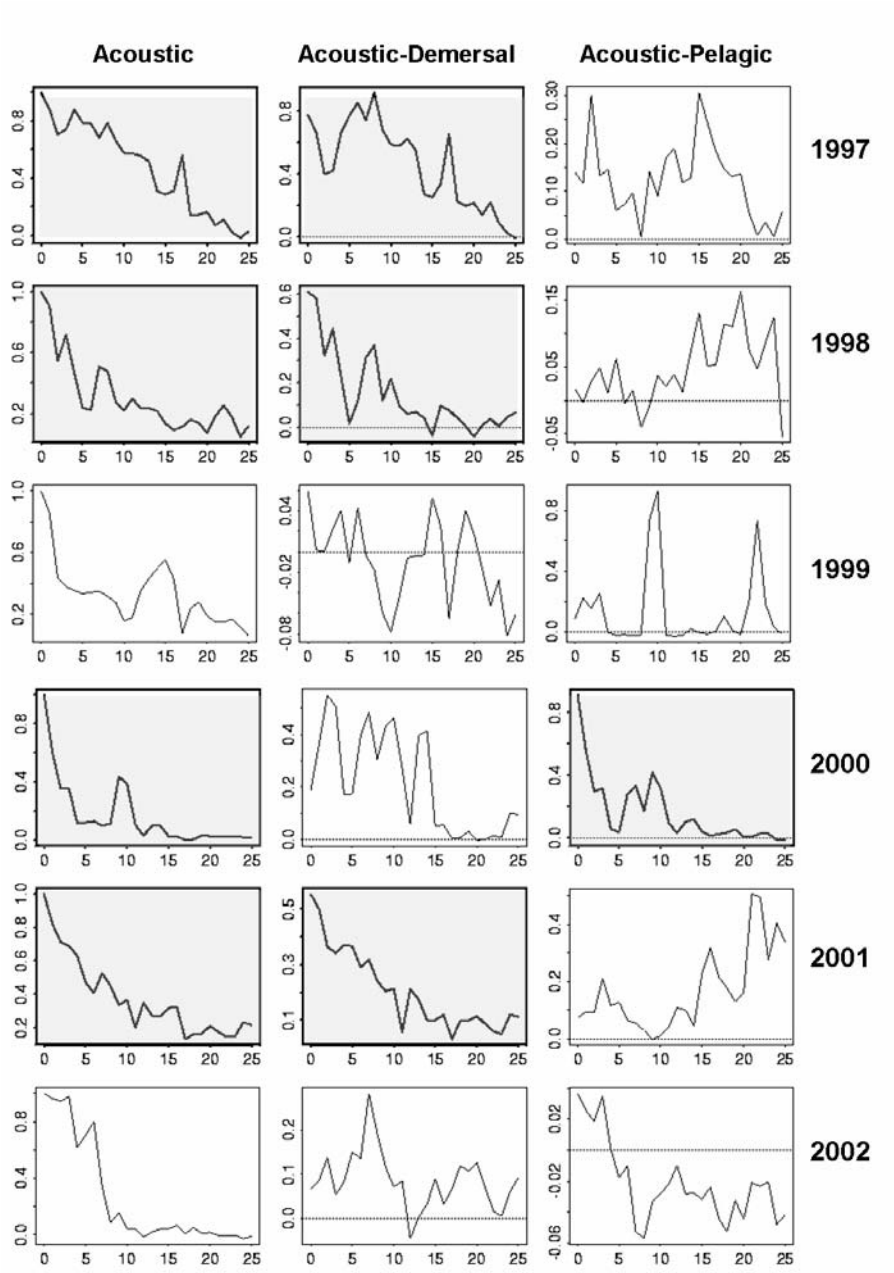


Fig. 3. Symmetrical cross correlograms calculated along the vessel track (1D). The x-axis is the distance from station (in n.mi) and the y-axis is the correlation between the acoustic underway and, according to the column: the acoustic on station (on the left), the demersal NASC-equivalent collected on station (on the middle) and the pelagic NASC-equivalent (on the right).

4.3 Estimation of parameter α

The parameter α is first estimated by the slope of the linear regression of T(x) on A(x). The cross plots between T(x) and A(x) allow evaluating visually the estimation (Fig. 4). We can see that the estimations (and the R-square) are very sensitive to the large values and the fitting of the cloud is not perfect.

The parameter β (Eq. 3) of the linear regressions happens to be very small (about zero in most cases). If the additional assumption β=0 were made, the target variable would be strictly proportional to the on-station acoustic. The α parameter would then be stable for different level of data. The estimation obtained for the whole distribution of data (Fig. 4) could be processed without some outliers, or just for the low values. The estimation should probably be more robust. In fact, the estimation of α changes according to the threshold chosen and is still different for each survey.

The parameter α is also assessed by the mean ratio between the cross and simple experimental variograms computed only with data on station. The gain of this approach is to take into account all distance lags. Results obtained (Fig. 5) are similar to those obtained with the regression.

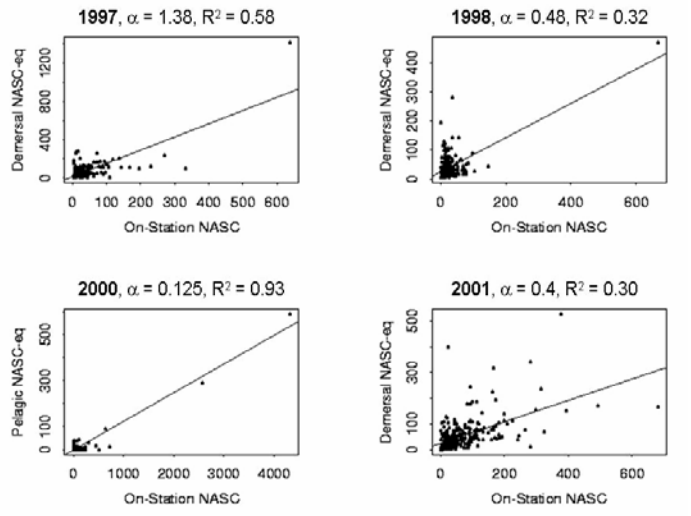


Fig. 4. Cross plot acoustic – catch on- station for the estimation of the multiplicative parameter. The lines represent the linear regressions between the two variables for each year. The values of the multiplicative coefficient α and the value of the R-square of the regression are written above each graph.

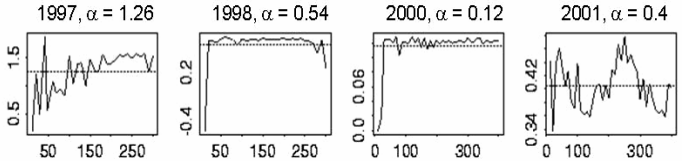


Fig. 5. Estimation of parameter α by the ratio between the $\gamma_{A,T}(h)$ and $\gamma_A(h)$ for on-station data only. The horizontal lines represent the mean value, i.e. the estimation. The x-axis represents distance in n.mi.

4.4 Estimations maps

To evaluate the improvement provided by the acoustic information, the bivariate approach (using a model with acoustic as master variable) is compared with a mono-variate approach based on the sole trawl values. This comparison is meaningful only if the variogram model for the trawl variable is the same for the kriging and the cokriging. It is then totally determined by the models chosen for the acoustic and the residual with the relation:

$$\gamma_T^K(h) = \gamma_T^{CK}(h) = \alpha^2 \ \gamma_A(h) + \gamma_R(h) \tag{1}$$

Given that the sampling grid covers regularly the study area, the cokriged and the kriged maps have the same general long distance patterns and the use of the acoustic variable only impacts the short scale features of the distribution (Fig. 6). In 1997, the kriging interpolation in the south western area where no sample is available amounts to the local mean concentration. A bivariate approach makes it possible to use the underway observations and to suggest some spatial pattern for the fish concentration in this area. In 2000, even if the weight of the acoustic is low (α = 0.12), the cokriged map computed with a Markov-type model honours some rich areas (in the North-East) which are not observed in the kriged map of the trawl data.

4.5 Variance of the estimation error map

For the four surveys, the estimation variance is smaller for the Markov-type approach than for the single variable approach. It is not surprising since the variance of cokriging is always less or equal than the variance of the correspondent kriging.

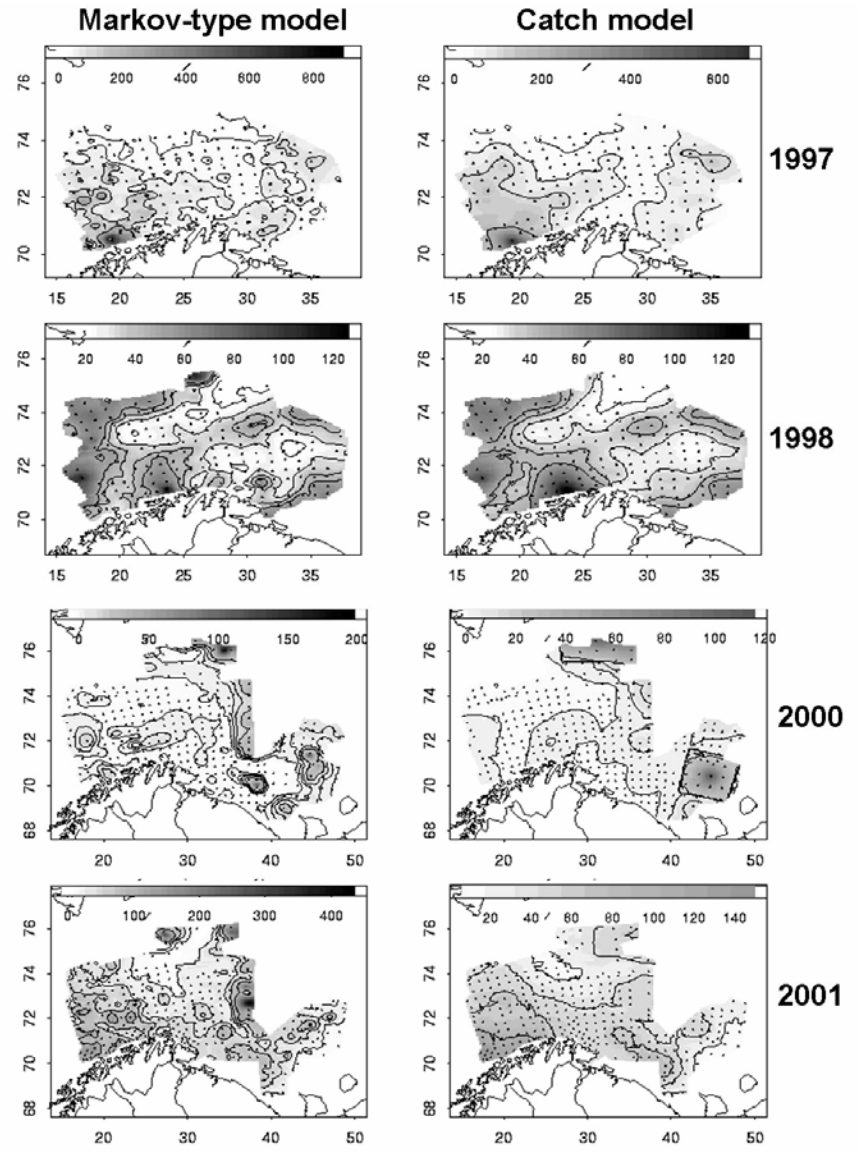


Fig. 6. Estimation maps obtained by the Markov-type model (left column) and a simple model using only the catch information available on station in a compatible model (right column). The maps on the left hand side are very more detailed. To compare the models, the grey scales are identical for each year but different from survey to survey.

4.6 Cross-validations

The cross validation consists in re-estimating a known point. Here we re-estimate each on-station point where the two variables, acoustic and catch, have been removed. It allows appraising the robustness of the model. For each survey the results provided are better in the bi-variate model than with the single variable model. The correlation coefficients between the estimated and observed catch values are shown in the table 2.

Table 2 Correlation coefficient between estimated and observed catch values

Year	Bivariate model	Monovariate model
1997	0.51	0.17
1998	0.30	0.09
2000	0.39	0.06
2001	0.41	0.34

5 Discussion

The estimation of the α parameter is a key step of the process as it this parameter quantifies the weight of the acoustic. In the model, the acoustic drives the catches and the residual allows rescaling the estimation on stations. Such behaviour is physically well understandable: acoustic provides a good representation of the fish abundance and the fish abundance is just obtained by adding a corrective term calculated by the divergence observed on stations between acoustic and catches (the residual). The main structure is then provided by the acoustic and the residual, in the general case, would not be strongly structured. However, in practice, the residual can have a long range structure because of one or few large values at the edge of the sampling area.

The use of an auxiliary variable largely more densely sampled than the target variable improves its estimation. The bi-variate model improves the estimation of the catch by combining acoustic with a simple relation exhibiting the role of each variable. However it is important to mitigate the results at least by the quality of the estimation of the parameter α. This key parameter has to be estimated and the quality of its estimation drives the quality of the whole process.

When variables get skew distributions like in the present study, once again, linear approaches happen to be fragile and we have indeed a weak confidence in the actual value of this parameter. When an estimation routine need to be processed every year, like the estimation of fish abundance, it is important to find a model robust enough to work for all the configurations, not only for a particular year with particular relation between the variables. Here the assumptions of the regression model are funded in four surveys out of six. For the two other cases (1999 and 2002) the erratic cross-structures do not allow to conclude to any model. The fact that the catch is driven by the acoustic, can be considered like a physical property, and we can think that the model will be also pertinent for the next years.

Because of the large skewness of the data, the use of linear approaches is questionable. Linear tools are indeed very sensitive to the large values which often hide the behaviour of the lower values (Rivoirard *et al.* 2000). Some non-linear tools like disjunctive kriging allow minimizing this impact. However the computation of a bivariate disjunctive model is laborious and requires heavy assumptions (Goovaerts 1997). The leading idea of this study has been to find a model simple enough and robust enough to be relevant in most available surveys.

Acknowledgement

The authors thank the Institute of Marine Research of Bergen who carried out all the surveys used in this study, in particular Olav Rune Godø and Vidar Hjellvik for formatting the databases and for helping our understanding of the Barents Sea features.

References

Chiles J-P, Delphiner P (1999) Geostatistics, Modelling Spatial Uncertainty, Wiley, New York, p. 695

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation , Oxford Univ. Press

MacLennan DN, Fernandes PG, Dalen J (2002) A consistent approach to definitions and symbols in fisheries acoustic. ICES Journal of Marine Science, 59: 365-369

Rivoirard J (2001) Which models for collocated cokriging? Math. Geol, v.33, no 2, 117-131

Rivoirard J, Simmonds J, Foote KG, Fernandez P, Bez N (2000) Geostatistics for Estimating Fish Abundance. Ed. Blackwell Science, p. 206

Schmaryan L, Journel AG (1999) Two Markov-type models and their application, Math. Geol. Vol. 31, no 8, 965-98

Mapping unobserved factors on vine plant mortality

N. Desassis[1], P. Monestiez[1], J. N. Bacro[2], P. Lagacherie[3] and J. M. Robez-Masson[3]

[1] Institut National de la Recherche Agronomique, Unité Biometrie, Avignon, France
[2] Université Montpellier II, 13M UMR 5149, Montpellier, France
[3] Institut National de la Recherche Agronomique, UMR LISAH, Montpellier, France

1 Introduction

In the vineyard plain of Languedoc, many parcels with a high rate of vine stock mortality have been observed recently (Legros *et al.* 1998, Lagacherie *et al.* 2001). As this could affect the global vineyard production of the region, it is important to determine the causes of this mortality. Beside a number of well-known permanent causes of vine stock mortality (e.g. vineyard diseases, inadequate rootstocks), a special attention have been paid recently to a new one, namely the degradation of soil physical conditions that have been often observed in association with mortality. Two types of degradations are concerned, both of them potentially affecting the vineyard water balance a) surface crusts that limits infiltration of rainfall water in soil b) compacted zones which reduce the volume of soil available for root activity. These degradations are under the influence of soil physical properties that can be identified as soil texture and soil hydromorphy. The aim of this paper is to investigate the relations between these two soil factors and the vine stock mortality over a small region located in the Languedoc vineyard plain.

2 General Methodology

2.1 Data and statistical framework

96 plots of 15m×20m were sampled on 4 sites of the studied zone (Caux, Neffiès, Pezenas and Roujan). On each one, the state of all vine stock were observed and depicted though two modalities, healthy vs. declining or dead. Note that 3 plots are contained in a parcel. Additionally, two soil samples were made on each plot. The rate of hydromorphy and the textural type, two orderly factors, each with three modalities were deduced. The total number of soil samples, say n is 192. We

have chosen to associate soil factors to the vine stock states in a small neighbourhood of soil sample.

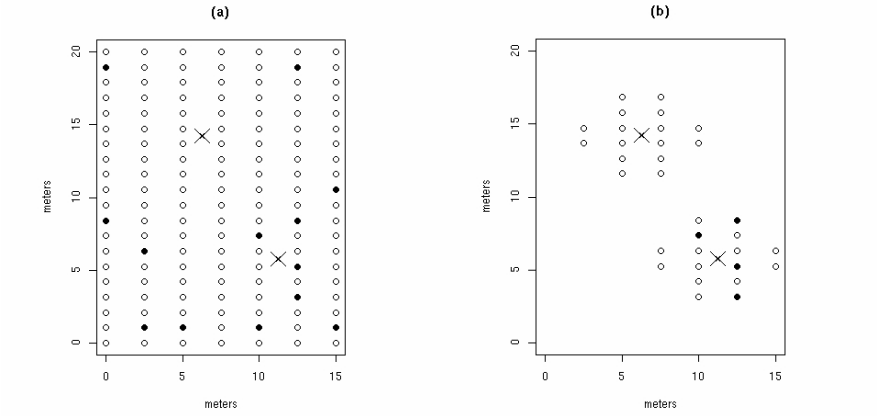


Fig. 1. (a) A schematized plot. Cross = soil sample location; circle = vine stock location : empty = healthy, full = declining. (b) Neighborhoods of vine stocks considered for the study.

Fig. 1 shows a typical plot. In the sequel, we assume that the states of vine stock were mutually independent conditionally to soil factors. As a consequence, the number of declining vine stock around each soil sample is a binomial count and the generalized linear model (GLM) framework suggests itself quite naturally. We refer to Mac Cullagh and Nelder (1989) for a detailed account about this framework.

Let N_i^{tot} , $i = 1,\ldots, n$ the number of vine stock considered around the i^{th} soil sample, and $(N_{i,jk})_{1\le j,k\le 3}$ the corresponding number of declining vine stock where j and k designate respectively the rate of hydromorphy and the textural type for the i^{th} soil sample. We have

$$N_{i,jk} \sim B(N_i^{tot}, p_{i,jk}) \tag{2.1}$$

where

$$h(p_{i,jk}) = \mu + \alpha_j + \gamma_k + \tau_{jk} \tag{2.2}$$

with h is a link function, α_j , γ_k and τ_{jk} are the parameters associated to the rate of hydromorphy, textural type and their interaction. In our study, we take

$$h(x) = \mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right) \tag{2.3}$$

For sake of simplicity, $N_{i,jk}$ and $p_{i,jk}$ will be denoted as N_i and p_i in the sequel.

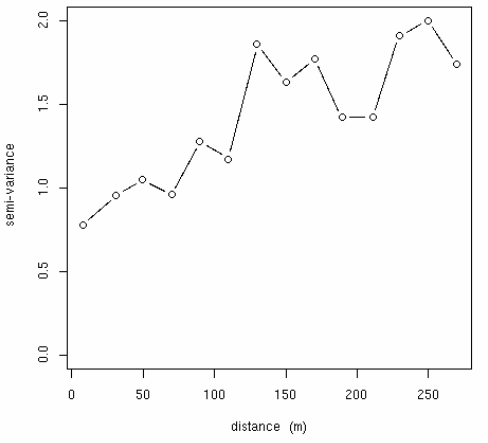


Fig. 2. Experimental semi-variogram of the residuals

Parameters are estimated by likelihood maximization. We set a goodness of fit test based on the difference of deviance between full model and saturated model. The model is strongly rejected. In similar cases, the usual conclusion is that other causes not taken into account in the model have to be investigated. Such a result justifies the introduction in the model of an offset term which intends to capture the effects of those unobserved causes.

Let $r_i=\hat{p}_i-p_i^*$ the residuals, where $\hat{p}_i=N_i/N_i^{tot}$ and p_i^* is given by (1) where the parameters are replaced by their estimation. From Fig. 2.1, it is clear that the residuals are spatially correlated. In order to take into account these spatial dependences, the offset term is considered as a random field. Our objectives are to map this unobserved field and to perform estimations of the soil parameters when working with such a model. Such a map could be a great help to identify new potentials causes of mortality.

2.2 Model formulation

We consider a model similar to the one proposed by Diggle *et al*. (1998). Let S be a stationary gaussian random field with

$$E(S(x))=0 \text{ and } Cov(S(x),S(x+h))=C(h) \tag{2.4}$$

We chose the classical exponential model with nugget effect for C :

$$C(h)=\begin{cases}\sigma^2\exp(-\phi h) & \text{if } h\neq 0\\ \nu+\sigma^2 & \text{if } h=0\end{cases} \tag{2.5}$$

ν and ϕ are respectively nugget effect and range parameter. Conditionally on S, the random variables N_i, $i=1, \ldots, n$, are supposed as mutually independent with distribution function

$$f(n_i | S) = \binom{n_i}{N_i^{tot}} p_i^{n_i} (1-p_i)^{N_i^{tot}-n_i} \tag{2.6}$$

with

$$h(p_{i,jk}) = \mu + \alpha_j + \gamma_k + \tau_{jk} + S(x_i) \tag{2.7}$$

where $S(x_i)$ is the current value of S at the i^{th} soil sample location.

2.3 Method

Since the N_i, $i=1, \ldots, n$ are not independent but only independent conditionally to S, an approach by likelihood maximization can't be used to estimate the parameters. We refer to Diggle *et al.* (1998) for more details.
We used MCMC algorithm in a Bayesian infererential framework.

2.3.1 Bayesian framework and MCMC algorithms

Assume that x is a realization of $X \sim g(x|\theta)$ where θ is a parameter of the distribution g. In the Bayesian framework θ is seen as random and a prior distribution with density $\pi(\theta)$ has to be chosen for this parameter. This prior allows to incorporate the knowledge we have about the studied phenomenon or it can be vague if no information is available or/and if we want to let the data drive the inference.
Information brought by the data x is combined with the prior and summarized in a probability distribution $\pi(\theta|\mathrm{x})$ according to the Bayes formula:

$$\pi(\theta|\mathrm{x}) = \frac{g(x|\theta)}{m(x)} \tag{2.8}$$

where $m(x) = \int g(x|\theta)\pi(\theta)d\theta$ is the marginal density of X. Inference will then be made on the posterior distribution $\pi(\theta|\mathrm{x})$. When $\pi(\theta|\mathrm{x})$ is analytically intractable as in our case, a usual way is to set a MCMC algorithm. The idea is to set up a Markov chain whose transition probabilities are analytically tractable and which has the required multivariate distribution as its equilibrium. By producing a sufficiently long run of this chain, we can therefore simulate a sample from the required distribution. Robert and Casella (2002) describe a large set of MCMC algorithms. Here we combined two types of MCMC: Gibbs sampler and Metropolis-Hasting-algorithm.

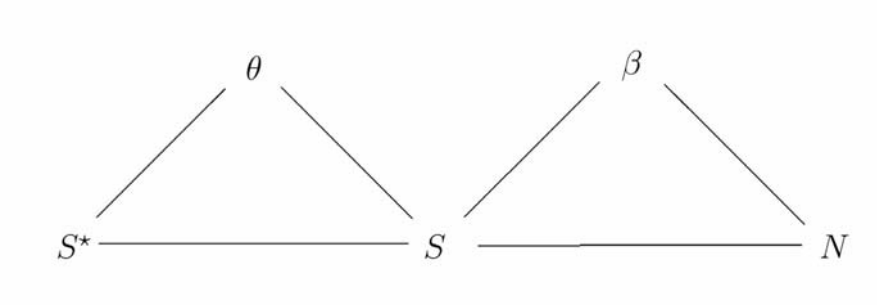


Fig. 3. Schematized dependence structure of our model.

2.3.2 Algorithm Metropolis-within-Gibbs

We use bayesian framework for our study. We note $S=(S(x_1),\ldots,S(x_n))$ the vector of S values at sites x_i , $i=1,\ldots,n$, $N=(N_1,\ldots, N_n)$ the numbers of vine stock declining and $\beta=(\mu,\alpha,\gamma,\tau)$ the vector of parameters of the linear part of the model, and $\theta=(\nu,\sigma^2,\phi)$ the vector of parameters of $C(h)$. Furthermore, let $S^*=(S(x_1^*),\ldots,S(x_m^*))$ denotes the values of S at locations $x_1^*,\ldots,x_m^*$ where we want to estimate the unobserved field S in order to map it.

We choose uniform and independent priors $\pi(\beta)$ and $\pi(\theta)$ for β and θ.
Our concern is now a posterior joint distribution of $(\theta,S,\beta\,|\mathrm{N})$ for inference and of $(S^*\,|\mathrm{S},\theta,\beta)$ for interpolation.

According to Gibbs sampler method, we sample successively θ, S, and β from the conditional distributions $\pi(\theta\,|\mathrm{N},\mathrm{S},\beta)$, $\pi(S\,|\mathrm{N},\theta,\beta)$, and $\pi(\beta|\mathrm{N},\mathrm{S},\theta)$ respectively. The dependence structure we use to compute these distributions is schematised Fig. 3. Since these distributions are not classical, we use a Metropolis-Hasting step. For this algorithm, we only need to know the distributions up to a constant. The likelihood is :

$$\pi(N\,|\mathrm{S},\beta,\theta)=\prod_{i=1}^{n} f(n_i\,|\mathrm{S},\beta) \tag{2.9}$$

and the conditional distributions of the parameters are:

$$\pi(\theta\,|\mathrm{N},\mathrm{S},\beta)=\pi(\theta\,|\mathrm{S})\propto\pi(S|\,\theta)\,\pi(\theta) \tag{2.10}$$

$$\pi(\beta|\mathrm{N},\mathrm{S},\theta)=\pi(\beta|\mathrm{N},\mathrm{S})\propto\pi(N|\mathrm{S},\beta)\,\pi(\beta) \tag{2.11}$$

For S we decompose the problem according to Gibbs sampler algorithm and we sample each component
of the vector from the distribution of this one conditionally to the others.

We note $S_{-i}=(S(x_1),\ldots,S(x_{i-1}),S(x_{i+1}),\ldots,S(x_n))$. Then,

$$\pi(S(x_i)\,|\,\mathrm{S}_{-i},N,\theta,\beta)\propto\pi(N\,|\,\mathrm{S},\beta)\,\pi(S(x_i)\,|\,\mathrm{S}_{-i},\theta) \tag{2.12}$$

where $\pi(S(x_i)|S_{-i},\theta)$ is the univariate gaussian distribution. Finally,

$$\pi(S^*|\theta,S,N,\beta) \propto \pi(S^*|\theta,S) \tag{2.13}$$

2.2.3 Parameters of the algorithm

Here we present the choices that had to be done in order to speed up the algorithm convergence :

-Choice of priors:
The choice of a prior is rather a model specification than a parameter of the algorithm. Nevertheless, this choice is often made when the algorithm is implemented since it plays a role in convergence speed. As we have no information available for soil effects on vine stock mortality and for the spatial characteristics of the unobserved random field, we chose non informative priors. In fact, we took uniform priors on [-5,5] for all soil effects since the inverse function of *logit* is close to 0 and 1 at the range of this set. In the same way we took uniform priors on [0,5] for the values of the nugget effect ν and for the variance parameter σ^2. For an exponential covariance function, the practical range (distance until the process reach 90 % of its variance) is $3/\phi$. For the parameter ϕ we chose an uniform prior on [0,1]. So the practical range can explore $[3,+\infty]$.

-Choice of the initial values:
Initial values has to be chosen for all the parameters.
We use the way proposed by Diggle *et al.* (1998) to do it in order to speed up the convergence.

-Choice of the transitions kernel of the Metropolis-Hasting steps:
We use Metropolis-Hasting step for θ, β and the components of S.
For the mean and the variance of the gaussian transition kernel in updating β, we set a first run with mean given by the generalized linear model procedure without the unobserved random field (see section 2.1) and an arbitrary variance fixed to 1. Then the mean and the variances of this sample are taken for a new run. In order to simplify the expression of acceptance probability for updating the components of θ and S, we chose respectively $\pi(\theta)$ and $\pi(S(x_i)|S_{-i},\theta)$ as transition kernel.

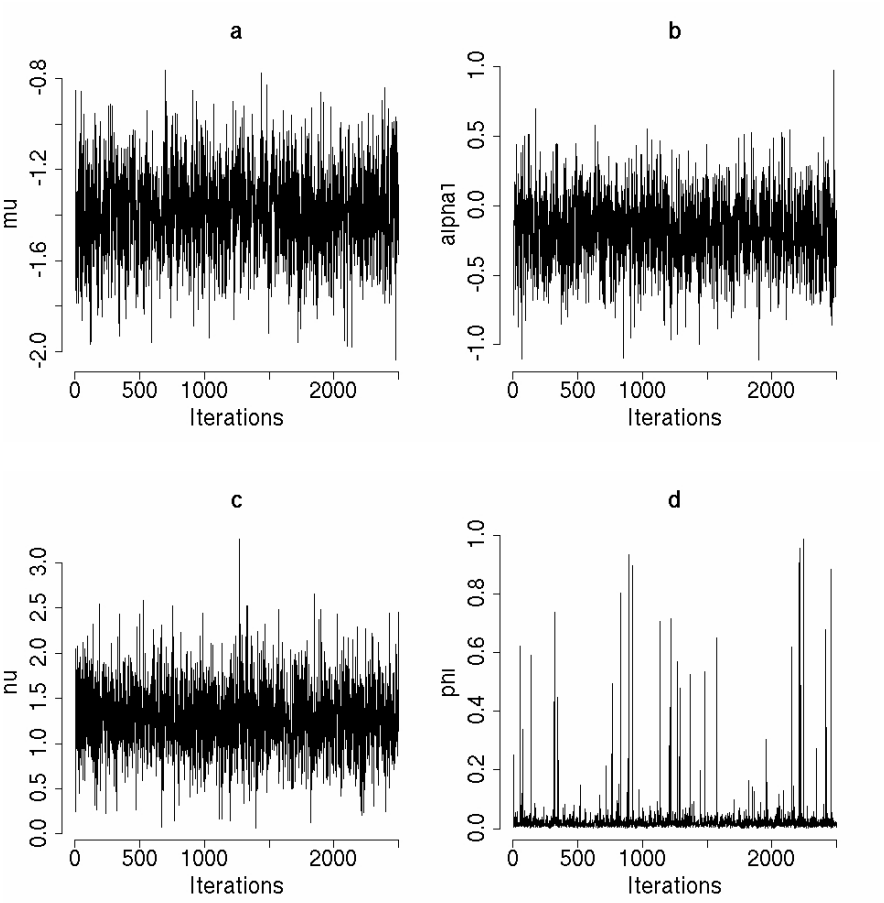


Fig. 4. Time series plots monitoring the MCMC output every 100 iterations for the intercept (a), one soil effect parameter (texture 1) (b), nugget effect (c) and range parameter (d).

For S^* the implementation consists in a simulation of a gaussian random field conditionally to observations (which are in our case the current value of S at the soil sample locations). We used the method proposed by Chilès and Delfiner (1999).

-Other specifications:
We used 500 000 iterations and we sampled every 100 iterations to reduce auto correlations between the iterations. We ignored the first 500 samples, by which time convergence is judged to have occurred. Note that we only need to update S^* every 100 iterations since its components don't play any role for the other parameters update.

We refer to Diggle *et al.* (1998) for more details about the algorithm.

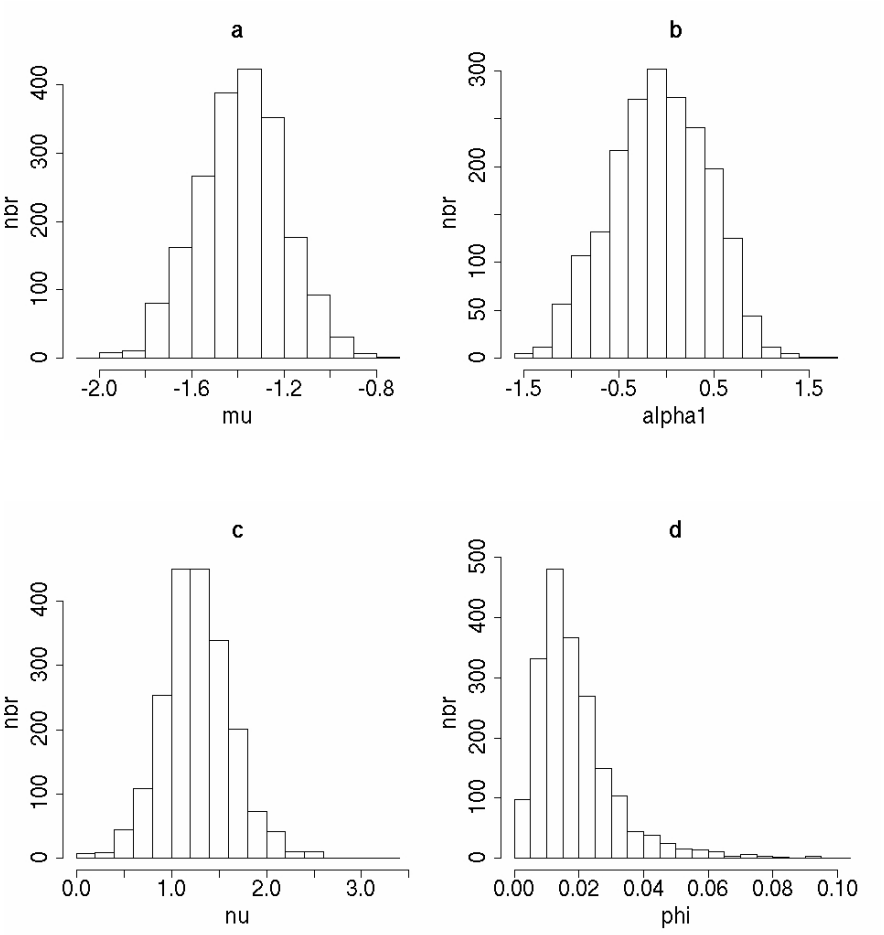


Fig. 5. Histograms of the samples of the intercept (a), of one soil effect parameter (texture 1) (b), of nugget effect (c) and of range parameter (d)

3 Results

In this section, we present the main results by considering a neighbourhood of $N^{tot}=16$ vine stocks around each soil sample. Results given by such algorithm are samples from posterior distributions for each parameters. Fig. 4 shows MCMC output for $(\mu, \alpha_1, \nu, \sigma^2)$ and Fig. 5 the corresponding histograms. Summaries of those distributions are given in table 1 for soil effects and in table 2 for the parameters of covariance function. For the map of the unobserved field we used pos-

terior medians of the distributions. Fig. 6 provides theses maps on Roujan, Caux and Neffies,

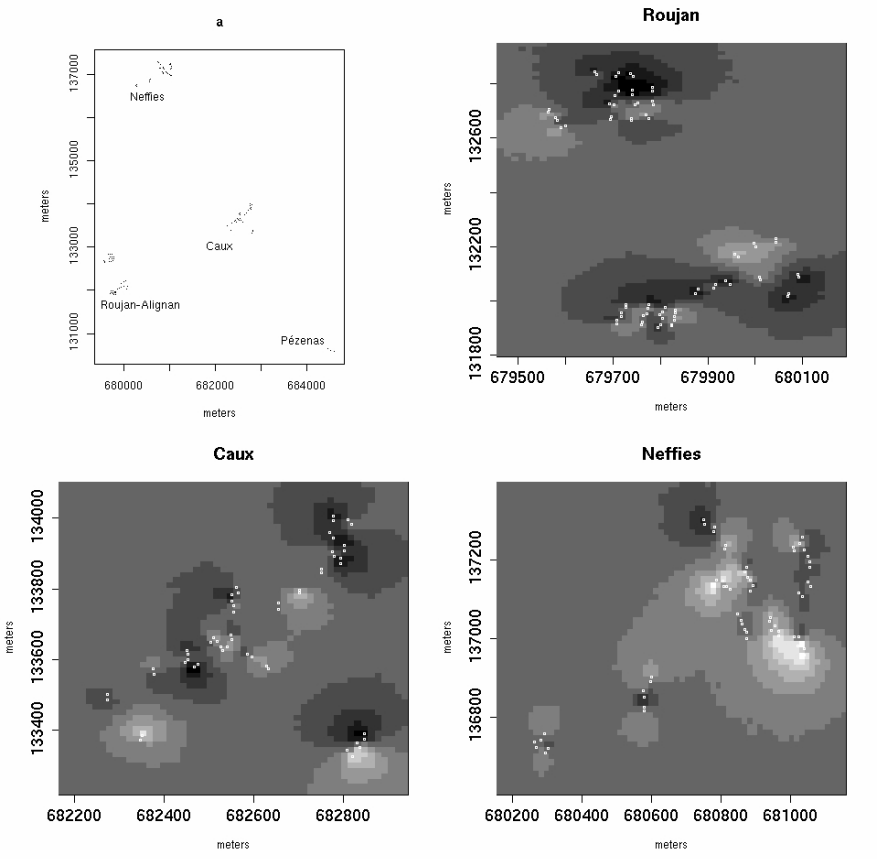


Fig. 6. Locations of the soil samples on the studied zone (a) and maps of the unobserved random field intensity on Roujan, Caux and Neffies. Axis graduating are in Lambert 3. Soil sample locations are indicated by the points. The grey scale indicate the field intensity (dark = low intensity; light = high intensity)

and Fig. 7 a zoom for Roujan. If we apply the inverse *logit* function to the posterior median of the intercept we obtain about 0.20 which is close to the frequency of declining vine stocks (23%). Since zero is included in all the 95 % confidence interval of soil effects parameters, we could conclude that they are not significantly different from zero. Nevertheless, zero is close to the range of the confidence interval for two interaction parameters. So, the interpretation must be done carefully. This two soil classes could have an effect on vine stock declining. Posterior median of ϕ is 0.016 so the practical range of the unobserved random field is about 187.5 meters which is close to the range of the variogram of the vine stocks states (about 150 meters) (see Desassis (2003) for a detailed study of the spatial

dependences of vine stock state process). We exhibited a nugget effect. Nugget effect is characteristic of individually vine stock sensitivity to declining or of micro scale variation of the random field. The fact that this nugget effect is greater than the variance parameter of the unobserved random field tends to show that a large part of the mortality is due to local variation. Nevertheless, the spatial structuration of the structured part of the unobserved random field has to be considered.

Table 1. Posteriors summaries of soil effects

Effects	2.5%	Median	97.5%
Intercept	-1.762	-1.379	-1.009
Texture 1	-0.675	-0.185	0.313
Texture 2	-0.773	-0.317	0.146
Texture 3	-0.240	0.505	1.244
Hydromorphy 1	-0.899	-0.344	0.206
Hydromorphy 2	-0.284	0.284	0.848
Hydromorphy 3	-0.397	0.056	0.506
Text. 1 Hydro. 1	-0.289	0.343	0.997
Text. 1 Hydro. 2	-0.872	-0.254	0.382
Text. 1 Hydro. 3	-0.609	-0.091	0.398
Text. 2 Hydro. 1	-0.489	0.142	0.795
Text. 2 Hydro. 2	-1.151	-0.570	0.036
Text. 2 Hydro. 3	-0.077	0.425	0.929
Text. 3 Hydro. 1	-1.423	-0.479	0.411
Text. 3 Hydro. 2	-0.176	0.820	1.840
Text. 3 Hydro. 3	-1.070	-0.337	0.439

The patterns of the values of the unobserved random field are highly dependent on the soil sample location pattern. Nevertheless, there are zones of marked effects of S, easy to identify, having very different behaviours from one map to another. Remarkably, there is no evident gradient between zones of highest and lowest intensity, but rather sharp variations of S between close soil sample locations (see e.g. Fig. 6, Caux) which probably depict strong heterogeneity in the environmental factors involved in vine stock mortality. In other cases, we observed homogeneity of S on several parcels (see e.g. Fig. 7, Roujan). This could reflect the effect of an agricultural practice, yet to identify.

Table 2. Posterior summaries of the parameters of covariance function

Parameters	2.5%	Median	97.5%
Nugget effect	0.563	1.254	2.046
Variance parameter	0.414	1.057	2.248
Range parameter	0.004	0.0016	0.076

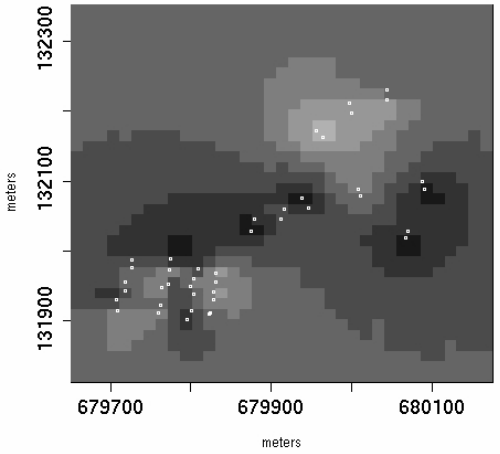


Fig.7. Map of the unobserved field on Roujan (zoom of the Fig. 6, Roujan)

4 Conclusions

The method presented in this paper provides a powerful tool to interpolate an unobserved random field introduced in a hierarchical way in the model. Bayesian framework allows to consider parameters uncertainty in order to make the inference. Nevertheless, several comments have to be formulated. First, the model chosen results of an approximation in our specific configuration. Indeed we supposed that the number of declining vine stock around a soil sample is a binomial count but this assumption requires that vine stock states are independent and identically distributed conditionally to soil factors and *S*. Spatial dependencies between vine stock states exhibited in Desassis (2003) could proceed from soil effects or from the field of unobserved effects. But other causes can be responsible of this observed structure. Our method cannot be used to detect if the exhibited structure arises from a spatially structured factor or from another phenomenon as contagion. The micro scale variations of the random field or the individual sensitivity to mortality of each vine stock makes the hypothesis of smoothness of the unobserved random field unacceptable. So the binomial approximation is rough and over dispersion should be introduced in the model. Finally, we did not detect significative effects of soil factors but we neglected nearly 2/3 of the data when keeping only vine stock in the neighbourhoods of soil samples. On the other hand, in the estimation, we give the same weight for all the vine stock in the neighbourhood of a soil sample without taking account of their distance to soil sample. Vine stock which are far from soil sample location have more risk than the one close,

to be associated to another soil type. So we should propose a model which allows to take into account the entire data set and the distances to soil sample location. In such model, we would consider each vine stock individually. Spatial structures of soil and spatial dependencies between soil and vine stock states should be modelled.

References

Chilès JP, Delfiner P, (1999) Geostatistics Modeling Spatial Uncertainty, Wiley : Chichester

Desassis (2003) Modélisation et analyse statistique des structures spatiales des phénomènes de dégradation d es sols et du dépérissement de la vigne, Mémoire de DEA, Université Montpellier II

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-Based geostatistic, Journal of the Royal Statistical Society, Series C, 47: 299-350

Lagacherie P, Collin-Bellier C, Goma-Fortin N (2001) Evaluation et analyse de la variabilité spatiale de la mortalité des ceps dans un vignoble Languedocien à partir de photographies aériennes à haute résolution, J. Int. Sci. VigneVin, 35 : 141-148

Legros JP, Argillier JP, Callot G, Carbonneau A, Champagnol F (1998) Les sols viticoles du Languedoc. Un état préoccupant, Progrès Agric. Vitic., 13-14, 296-298

Mc Cullagh P, Nelder JA (1989) Generalized Linear Models (second edition). Chapman & Hall: London

Robert CP, Casella G (2002) Monte Carlo Statistical Methods, Springer-Verlag: New York.

Analysis and modelling of spatially and temporally varying phenological phases

D. Doktor[1,2], F.W. Badeck[2], F. Hattermann[2], J. Schaber[2] and M. McAllister[1]

[1] Department of Environmental Science and Technology; Imperial College, London; RSM Building, Prince Consort Road; London, SW7 2BP, UK
[2] Potsdam Institute for Climate Impact Research, Telegrafenberg A51, P.O. Box 60 12 03, D-14412 Potsdam, Germany

1 Introduction

Temperature is one of the main determinants of phenological plant development (Schnelle 1955 and 1973, Worral 1993, Dieckmann 1996 and Sparks 2002). Especially in the temperate middle latitudes plants have to rely on synchronising their growth with the seasons to achieve a maximal degree of reproduction (Kramer 1996). Here we demonstrate the supposed influence of variable temperature patterns on the distribution of observed budburst dates and their dependency on altitude. For this purpose we calculated the dependency on altitude of budburst for every analysed year and developed an interpolation method which uses the calculated gradient. We interpolated phenological point data using another method which also explicitly incorporates elevation (External Drift Kriging) and applied Gaussian Probability Functions to describe the different budburst distributions. Based on meteorological records from 1951-1980 Mendl (1995) calculated the relative frequency of general weather situations in Germany. It was demonstrated that the spring is, unlike to other seasons, characterised by a super-proportional occurrence of the so called weather situation north and east, i.e. often a constant alternation of cold shower weather coming in from Scandinavia and dry continental air mass from Eastern Europe which can already heat up the continent due to the increasing sunshine duration. In consequence, this yields in general to discontinuous and variable temperature patterns. We hypothesise that years with discontinuous and variable temperature patterns produce a bi- or multimodal distribution of the observed budburst dates while an undisturbed leaf unfolding generally produces a unimodal distribution.

The current phenologica1 database of the German Weather Service (DWD) provided continuous time series from 1951-2000 over whole Germany. We analysed 12 randomly chosen years of budburst of Beech (*Fagus sylvatica*), Birch (*Betula pendula*) and Oak (*Quercus robur*). On average about 2000 equally distributed phenological observations over whole Germany were available per year. Goovaerts (2000) as well as Hudson and Wackernagel (1993) used elevation data as an external variable to estimate spatially interpolated rainfall and temperature.

Here we incorporated a Digital Elevation Model (DEM) with a resolution of 1*1km over whole Germany.

2 Application of the outlier detection algorithm

An outlier detection routine (Schaber and Badeck 2002) was applied on three different species of deciduous trees: Beech (*Fagus sylvatica*), Oak (*Quercus robur*) and Birch (*Betula pendula*): 0.42%, 0.44% and 0.46% of the data respectively were identified as outliers. An outlier is detected by the distribution-free 30-day residual rule in combination with a robust estimation procedure based on the minimisation of the sum of absolute residuals. This is due to one of the few detectable mistakes in phenological databases, a so called "month-mistake" resulting from the conversion of the observed budburst date to the absolute day of year (DOY). A deviation above 30 days cannot be explained by the natural variability and/or the observer mistake (i.e. precision of measurement). Schaber and Badeck (2002) postulated a danger of false outlier identification in case of a distinct bimodal distribution of observed budburst dates. This distribution "is produced by an intermittent occurrence of environmental conditions unfavourable to phenological development, e.g. a cold spell when the buds in a part of the population have already broken". We suggest that this phenomenon is caused by the response of phenological development to discontinuous temperature patterns in spring.

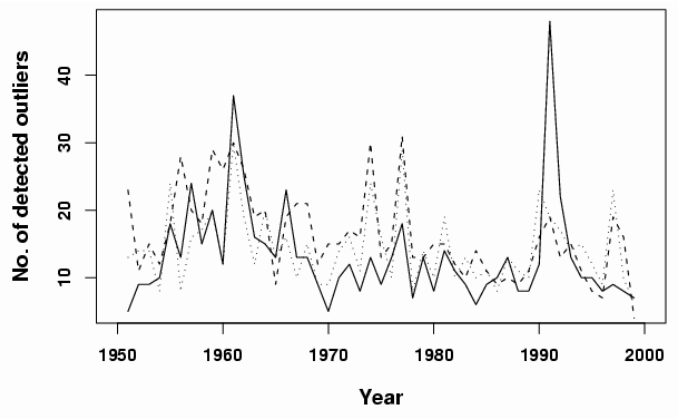


Fig. 1. Number of detected outliers per year for Birch (dashed line), Oak (thick line) and Beech (dotted line).

The high variation of detected outliers between years can be taken as an indicator of the potential of false outlier identification as one would expect an even distribution of outliers in time if they were caused by human errors only (s. Fig. 1).

The years with a great number of detected outliers correspond to relatively high standard deviations of observed budburst dates (correlation coefficient=0.39).

Furthermore, we determined that in regions east of the river Elbe (each region represents a geographical unit in terms of climate, vegetation, geomorphology and water balance, s. Schmithuesen *et al.* (1962)) no or only a few outliers are detected while all other regions of Germany show a considerably larger amount of detected outliers. The extremely low numbers of outliers in the easterly regions correspond to a low variance of observed budburst dates in the very same areas which might be due to a different behaviour of deciduous trees in easterly parts to avoid the danger of late frost regarding the more continental climate. Finally, it should be taken into account that different varieties in easterly regions (with modified behaviour) are the reason for a lower variability, although the types of variety of the trees are not included in the database.

3 Inverse altitudinal gradient of budburst dates close to the sea

The long lasting records of mean daily temperatures in Germany, considering the area between the coasts of the Northern and the Baltic Sea and the northern edge of the low mountain range ("Norddeutsches Tiefland"), show that the general pattern of decreasing temperature within increasing altitude is reversed here. This is because of the thermal influence of the sea, especially in autumn and spring. The sea dampens extreme temperatures in summers and winters on the one hand and delays the cooling down in autumn as well as the warming up in spring of coastal regions on the other. Thus, the delayed warming up of regions close to the coast leads to an increase of mean daily temperatures with increasing altitude of meteorological observation stations in the Northern Germany lowlands (s. Fig. 2). The phenological development of deciduous trees in this very area mirrors these temperature patterns, i.e. an earlier budburst is observed at greater altitude. The mean budburst date of Oak for example in vicinity of the coast (below 20 m) is delayed about five days in comparison with inbound regions (50-100 m, s. Table 1).

Table 1. Mean observed budburst date (DOY) of Beech, Oak and Birch from 1951-2000 in different altitudinal belts in Northern Germany.

	Birch	Beech	Oak
0-20 m	114.7	121.9	129.5
20-50 m	112.6	120.7	126.8
50-100 m	110.8	119.3	124.2

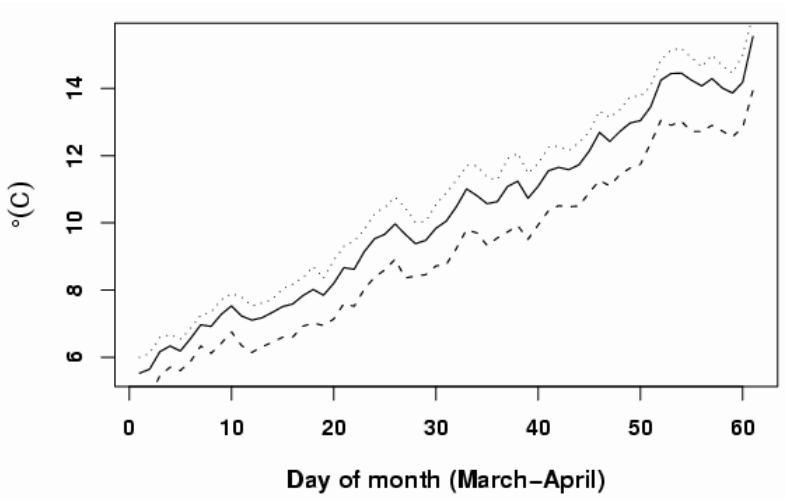


Fig. 2. Mean daily temperatures during March and April from 1950-2000 in Northern Germany. The dashed line represents regions under 20 m asl, the thick line 20-50 m and dotted 50-100m.

4 Calculation of elevation gradient – Detrended Kriging

The dependency of phenological development on temperature is well documented in case of growth rate in different altitudinal belts. The vertical decrease of air temperature of ca. 0.7 ºC per 100 meters of altitude (lapse rate) is reflected in delayed budburst. Baumgartner (1956) showed a delay of budburst of Beech of 1-2 days for every 30 m increase of altitude. These results are corresponding with investigations of Worral (1983), Benecke (1972), Oberarzbacher (1977) and Schnelle (1973). In this paper we will use a simple linear model to determine the dependency of phenological phases on altitude (s. Fig. 3 as well).

$$g_{ha} = \frac{\sum_{i=1}^{n} \left(\frac{\Delta d_{obs}}{\Delta h} \right)_i * w_i}{\sum_{i=1}^{n} w_i} \qquad (1)$$

With n = number of compared observation pairs: $\frac{(A_s)^2 - A_s}{2}$, A_s = number of observation stations, g_{ha} = global dependency on altitude (delayed budburst day per m increase), Δd_{obs} = Difference of budburst date of the two compared ob-

servation stations and Δh = Difference of altitude of two compared observation stations. All possible observation data pairs will be compared in terms of budburst date and altitude. The weights w_i are defined as the inverse distance between the stations. The index of the global dependency on altitude is used to avoid an unneccesarily strong influence of local values. This is essential because of the high variability of phenological data even on a local scale which could lead to a false calculation of the gradient. On the basis of the phenological database of the DWD we determined a delayed budburst dependent on species and year respectively between 0.42 – 4.6 days per 100 m increase of altitude (s. Table 2). Years with a significant high value (delayed budburst) are 1979, 1981, 1990 and 1997.

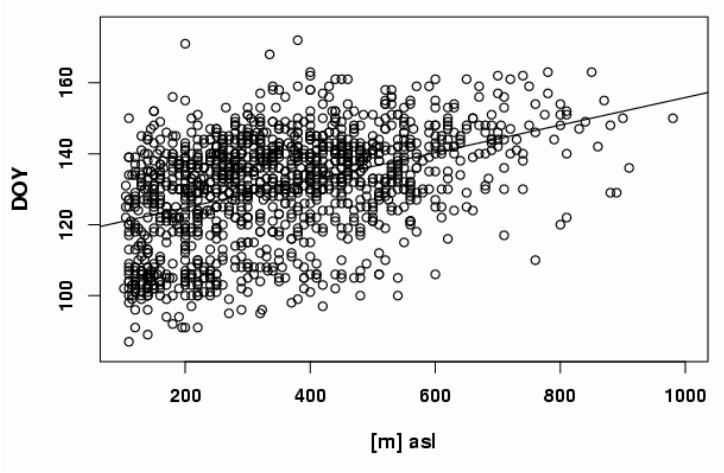


Fig. 3. Linear regression (black line) for the budburst dates of Oak and altitude of the year 1991. Here, only observations stations above 100 m asl are analysed. Residual standard error: 13.64 on 1332 degrees of freedom; Multiple R-Squared: 0.1834

Now the global gradient is used as a component for the so called Detrended Kriging: subtraction of the product of the altitude of the observation station and the calculated global gradient from the original budburst date. The resulting values (residua) are interpolated using Ordinary Kriging as it is more robust than Simple Kriging. Finally, the product of the global gradient and the respective value of an underlying DEM will be added. According to the detected reverse of the global gradient on temperature with elevation in Northern Germany observation stations below 100 m where not included to calculate the dependency of budburst on altitude. Just as well we did not compare observation stations with less than 50 m difference of altitude because of insufficient details of the altitude of the observation stations in the original database.

Table 2. Delayed budburst in days per 100 m increase of altitude of Birch, Beech and Oak.

Year	Birch	Beech	Oak
1956	2.38	1.1	2.2
1961	1	1.22	2.84
1970	1.8	1.79	2.47
1974	0.42	1.3	2.82
1977	1.34	1.61	3.14
1979	2.6	2.27	2.7
1981	1.7	2.47	4.58
1986	1.04	1.87	2.24
1987	1.25	1.67	2.51
1990	2.63	3.26	2.91
1993	1.19	1.24	1.84
1997	2.23	1.69	2.78

5 EDK vs Detrended Kriging

Kriging with External Drift is an approach for incorporating secondary information in statistical interpolation. The variance of the external variable using EDK stands in close relationship to the variance of the estimated variable, i.e. a physical correlation is a basic requirement. Furthermore the external variable should vary smoothly in space and must be known at all locations of the primary data values and at all locations to be estimated (Deutsch and Journel 1998). For a theoretical background please refer to Isaaks and Srivastava (1989) and Wackernagel (1998). Due to high natural variability the interpolation (Kriging with External Drift) of phenological ground observations requires special procedures: In contrast to the common procedure of not using more than 10 surrounding observation stations to predict values at unknown locations (to highlight local effects), we incorporated 30 or more surrounding observation points. In addition a method of the same quality based on the described global gradient of altitude could be established.

Despite the high phenological variance this method is robust but not available to be applied on a local scale. Estimated values for both interpolation methods are in a good agreement with observations and even with phenological models for budburst prediction. An improvement in quality can only be achieved while using additional information, e.g. exposition (aspect) of the analysed objects. However, a good understanding of all mechanisms influencing budburst is still missing. Fig. 4 shows the different behaviour of budburst events of Oak through time. 1981 represents a year with an early mean budburst date in contrast to the year 1986, which is relatively late. In both years the mountainous regions vary only slightly in terms of budburst date while the lowlands show differences of about 20-30 days. Thus lowlands and mountainous regions are acting differently by corresponding to the weather characteristics of the particular year, even in cases of spa-

tial contiguity. The technique of cross-validation allows us to compare estimated and observed values (Isaaks and Srivastava 1989). Thus, the results of different interpolation methods could be checked in terms of interpolation quality. The particular interpolation method is tested at locations where a sample value is available. This sample will be removed from the dataset within the interpolation process and it will be interpolated on exactly this sample point. This procedure is repeated for each sample value.

Table 3. Averages of statistical variables of all cross-validations for Detrended Kriging with mean=mean error, MAE=Mean Absolute Error, SD=standard deviation of MAE, Var=variance of MAE, G=skewness of MAE, min/max=Minimum and Maximum of local estimation errors, Err-sum=sums of estimation errors.

	Mean	MAE	SD	Var	G	min	max	Err-sum
Oak	-0.02	5.85	7.79	63	0.02	-32.86	41.3	-45.2
Beech	-0.01	5.35	7.2	55.06	0.07	-34.75	41.03	-16.83
Birch	-0.03	5.35	7.3	56.36	0.07	-34.59	36.75	-59.85

Table 4. Averages of statistical variables of all cross-validations for EDK with mean=mean error, MAE=Mean Absolute Error, SD=standard deviation of MAE, Var=variance of MAE, G=skewness of MAE, min/max=Minimum and Maximum of local estimation errors, Err-sum=sums of estimation errors.

	mean	MAE	SD	Var	g	min	max	Err-sum
Oak	0	5.78	7.72	61.9	0.03	-32.25	42.96	-4.26
Beech	0.02	5.26	7.1	53.53	-0.06	-34.41	40.66	49.93
Birch	0.01	5.2	7.1	53.6	-0.03	-34.19	38.25	35.02

Both interpolation techniques are of nearly the same quality (s. Tables 3 and 4). On average the MAE of the date of budburst is about 5 days for each tree species and interpolation method. Looking at the sums of errors it appears that Detrended Kriging (DK) tends to underestimate while EDK tends to overestimate the predicted value. In Fig. 5 the MAEs of DK and EDK are compared for every year. The values differ considerably between years (range 3 to 9). This is due to the changing variability of observed budburst dates which is depending on the weather conditions in the spring of each respective year. There is hardly any apparent difference in terms of estimation performance. The correlation coefficient of the standard deviation of observed budburst dates and the MAE is 0.86, 0.90 and 0.97 for Oak, Beech and Birch respectively.

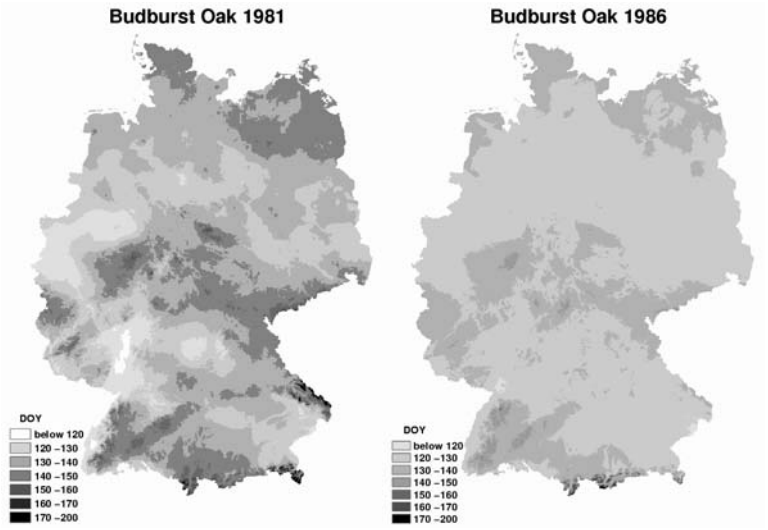


Fig. 4. Maps of the interpolated budburst day of Oak of the years 1981 (left) and 1986 (right) over whole Germany using EDK. The legend represents the julian day (Day Of Year) of budburst.

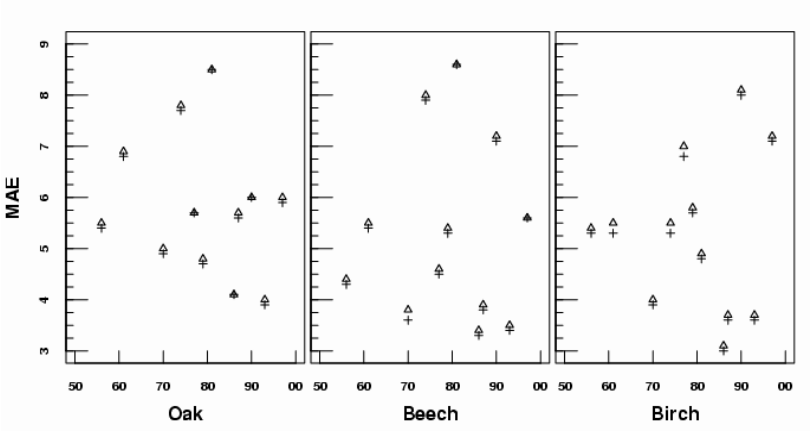


Fig. 5. MAE of EDK (cross) and DK (triangle) for Oak on the left, Beech in the middle and Birch on the right hand side for all analysed years (randomly chosen from 1950-2000).

6 Gaussian probability mixture models

To describe different distributions of observed budburst dates we used Gaussian Probability Mixture Models. The quality of the model fit was tested using a chi-

square test at the 0.05 significance level. A mixture model was applied to the observed budburst date distribution for every year. In most cases the observed shapes suggested that data might arise from an underlying pattern of two or more overlapping bell-shaped distributions maybe dependent on changing weather conditions during spring. Occasionally the observed distribution was similar to a normal unimodal distribution. We fitted a mixture model of one, two or more Gaussian components to the data by the maximum likelihood method, and its parameters, which include the mixing proportions and the parameters of the component distributions, were estimated. Gaussian Mixture Models have the property of being able to represent any distribution as long as the number of Gaussians in the mixture is large enough (Gilardi *et al.* 2002). A mixture distribution with n continuous components has a density of the form (Poland and Shachter 1994):

$$f_m(x) = p_1 f_1(x) + ... + p_n f_n(x) \qquad (2)$$

where $p_1,...,p_m$ are positive numbers summing to one and $f_1(x),...,f_m(x)$ are the component densities. Mixtures of analytically tractable component distributions, such as Gaussians, are useful to model not only true mixtures but any continuous probability distributions with which fast calculations are desirable. Mixture models provide a useful way to identify homogeneous groups within a given population, whenever there is no a priori knowledge of any group structure on the underlying population but heterogeneity is suspected (Mclachan and Peel 1999 after Tentoni *et al.* 2004). Maximum likelihood estimates of the mixture model parameters were derived by the Generalised Reduced Gradient (GR52) nonlinear optimisation algorithm.

The approach can be used to distinguish years with different temporal evolution of the budburst date in a quantitative manner. A correlation of the type of the distribution with the elevation gradient was found so that in cases of relatively high elevation gradients at least two or more Gaussian Mixture Functions with high distance of their corresponding means had to be used to achieve a good model fit. All distributions of budburst dates which could be produced by the behaviour of the weather (described in Sec. 1 and Sec. 2) of the respective year could be characterised: Years with an undisturbed continuous phenology (Fig. 6: 1970, 1977, 1986, 1993), years with bi- or multimodal distributions (Fig. 6: 1974, 1981, 1990) as well as years with steady but slow development (and early budburst: 1961) or late budburst due to a cold spell when budburst is nearly finished in most of the regions (1956). Also, we compared the calculated elevation gradient of each year with the respective standard deviation of the observed budburst dates. The correlation coefficient for Oak, Beech and Birch was 0.79, 0.43 and 0.48 respectively. The results show a correlation of these two parameters. It is obvious that the delayed budburst in higher regions can only explain a part of the variance because there is as well a delayed budburst in cool regions in lower areas, i.e. the general gradient of temperature from the southwest to the northeast of Germany.

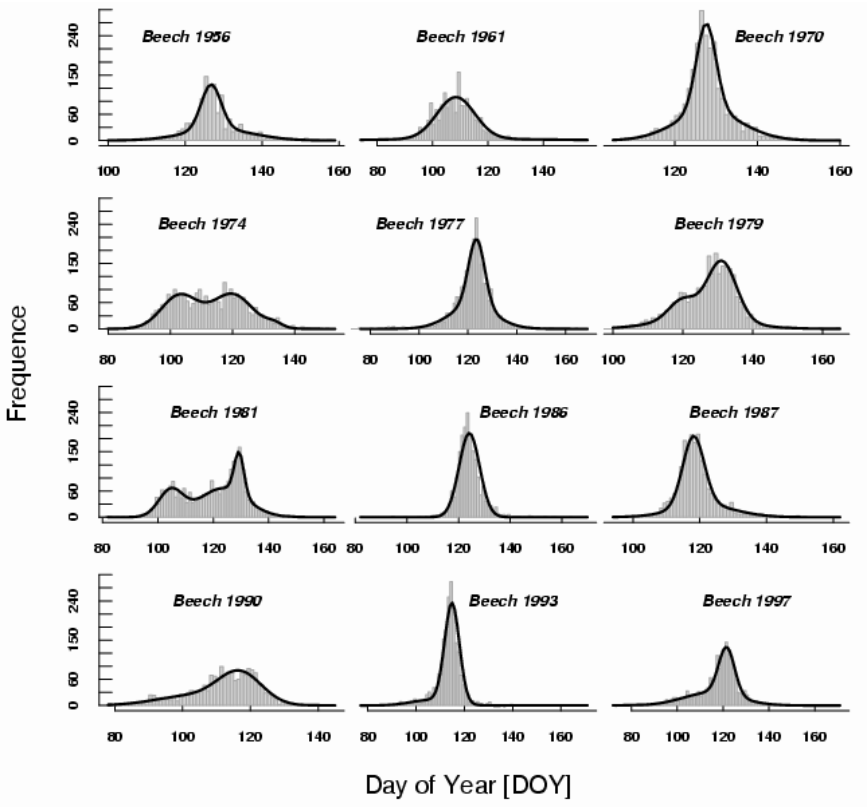


Fig. 6. Histogram of the density of the observed budburst dates and predicted budburst dates (black line) using Gaussian Probability Mixture Functions.

7 Conclusions

In this paper we tested the quality of two methods for interpolation of budburst dates which both incorporate explicitly elevation as additional variable. It could be shown that using Ordinary Kriging together with a simple linear model which calculates a global gradient of dependency of budburst on altitude leads to results with the same quality as using External Drift Kriging. The high variability of phenological data even on a local scale (trees on a south facing slope might have an earlier budburst as trees in a valley despite higher altitude) makes it very difficult to predict values based on a few surrounding observation stations only. Thus, both interpolation methods cannot be applied on a local scale without more detailed data, e.g. aspect. The quality of estimation varies between the analysed years and seems to be based on the general weather conditions during the respec-

tive spring (s. Section 8). Furthermore the computed global gradient on elevation showed a variable dependency of budburst on altitude between successive years. We could also show a reverse altitudinal gradient of budburst dates close to the sea. The calculated delayed budburst in vicinity to the sea is in good agreement with long time temperature measurements within the very same area showing delayed heating of the air masses close to the sea in spring.

8 Interpretations and Recommendations

We tried to show the interrelationship of the applied methods, as they are all affected by interannual variance of budburst. Years with an intermittent cold spell could be identified as well as years with an undisturbed leaf unfolding. Bi- or multimodal distributions are correlated with the computed elevation gradient in terms of the difference of the means of used Gaussian distributions to fit the Gaussian Mixture Model: the correlation coefficient is 0.61 and 0.36 for Oak and Beech respectively. Years with a great number of detected outliers correspond also to bi- or multimodal distributions with a great difference of their corresponding means (1974, 1981 and 1990 for Beech, compare Fig. 1 and 6). The quality of estimation of both interpolation methods tends to be considerably better when applied in years with a continuous and undisturbed budburst (compare Fig. 5 and 6). The description of the interannual variance of budburst via Gaussian Mixtures Distributions can be used for modification of the outlier-detection algorithm. Extreme values of a bi- or multimodal distribution tend to differ more from their correspondent means than the ones in unimodal distributions. Therefore, the algorithm detects more outliers in years with varying temperature patterns, which are eventually part of the natural variability. In years with a cold intermittent spell the outlier detection algorithm might be improved by taking into account the membership in the mixture component considering the relative probability within the overlapping distributions. A better understanding of the devolution of the spatial pattern of the so called “green wave” could be achieved by knowing that variability of budburst is mainly dependent on delayed budburst in lowlands while mountainous regions do not differ considerably in terms of budburst date (s. Fig. 3). This is especially useful for interpreting remote sensing data.

References

Baumgartner AG, Kleinlein G, Waldmann G (1956) Forstlich-phänologische Beobachtungen und Experimente am Großen Falkenstein. Veröffentlichungen aus dem Meteorologischen Institut der Forstlichen Forschungsanstalt München: 290-303.

Benecke U (1972) Wachstum, CO2-Gaswechsel und Pigmentgehalt einiger Baumarten nach Ausbringung in verschiedenen Höhenlagen. Angew. Bot. 46: 117-135.

Dieckmann M (1996) Relationship between phenology of perennial herbs and meteorological data in deciduous forests of Sweden. Can. J. Bot. 74: 528-537

Deutsch CV, Journel AG (1998) GSLIB Geostatistical Software Library and User's Guide. Oxford University Press, New York

Goovaersts P (2000) Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology 228: 113-129

Gilardi NS, Bengio S, Kanevski N (2002). Conditional Gaussian Mixture Models for Environmental Risk Mapping

Hudson G, Wackernagel H (1993) Mapping temperature using kriging with external drift: theory and an example from Scotland. International Journal of Climatology 14: 77-91

Isaaks EH, Srivastava RM (1989) Applied Geostatistics. Oxford University Press, New York

Kramer K (1996) Phenology and growth of european trees in relation to climate change. Institute for forestry and nature research. Wageningen, Landbouwuniversiteitte

McLachland GJ, Peel D (2000) Finite Mixture Models. Wiley, New York

Mendl M (1995) Klima. In: Liedtke H, Marcinek J (ed) Physische Geographie Deutschlands, Perthes, Gotha

Oberarzbacher P (1977) Beiträge zur physiologischen Analyse des Höhenzuwachses von verschiedenen Fichtenklonen entlang eines Höhenprofiles im Wipptal (Tirol) und in Klimakammern. Innsbruck, Universität Innsbruck

Poland WB, Shachter RD (1994) Three Approaches to Probability Model Selection. in Uncertainty in Artificial Intelligence

Schaber J, Badeck FW (2002) Evaluation of methods for the combination of phenological time series and outlier detection. Tree Physiology 22(14): 973-982

Schmithüsen J, Meynen E (1962) Handbuch der naturräumlichen Gliederung Deutschlands. Bundesanstalt für Landeskunde und Raumforschung, Bad Godesberg

Schnelle F (1955) Pflanzen-Phänologie. Geest & Portig, Leipzig

Schnelle F (1973). Die Vegetationszeit von Waldbäumen in deutschen Mittelgebirgen. Erlanger Geographische Arbeiten 32: 2-34

Sparks TH (2002) Observed changes in seasons: an overview. International Journal of Climatology 22 (14): 1715-1725

Tentoni S, Astolfi P, Pasquale De A, Zonta L (2004) A. Birthweight by gestational age in preterm babies according to a Gaussian Mixture model. International Journal of Obstetrics and Gynaecology 111: 31-37

Wackernagel H (1998). Multivariate Geostatistics. Springer, New York, Berlin, Heidelberg

Worral J (1983) Temperature-bud-burst relationships in amabilis and subalpine fir provenance tests replicated at different elevations. Silvae Genet. 32: 203-209

Worral J (1993). Temperature effects on bud burst and leaf fall in subalpine larch. Journal of sustainable forestry 1: 1-18

Detection of spatial clusters and outliers in cancer rates using geostatistical filters and spatial neutral models

P. Goovaerts

BioMedware, Inc. 516 North State Street, Ann Arbor, MI, USA.

1 Introduction

Cancer mortality maps are important tools in health research, allowing the identification of spatial patterns, clusters and disease 'hot spots' that often stimulate research to elucidate causative relationships. Their analysis is typically performed using a statistical pattern recognition approach whereby a statistic (*e.g.* spatial cluster or autocorrelation statistic) quantifying a relevant aspect of spatial pattern is first calculated. The value of this statistic is then compared to the distribution of that statistic's value under a null spatial model. This provides a probabilistic assessment of how unlikely an observed spatial pattern is under the null hypothesis.

Most statistical tests for spatial pattern are based on the "normality" and "randomization" null hypotheses (Waller and Gotway 2004). Under the normality hypothesis all observations follow independent, identically distributed normal distributions. Under the randomization hypothesis, each permutation of the observed values is equally likely. These translate into a null hypothesis of spatial independence (SI) of observed rates and, provided the population sizes of areal units (e.g. ZIP codes) are fairly homogeneous, the assumption of constant or spatially uniform risk. However since some spatial pattern is almost *always* present, rejecting this hypothesis has little scientific value. Also, as emphasized by Ord and Getis (2001), Type I errors may increase when tests of hypothesis using the randomization assumption are applied to spatially correlated data, leading us to reject the null hypothesis of no clustering more often that we should. What are needed are realistic null models that incorporate background pattern. The term "*Neutral Model*" captures the notion of a plausible system state that can be used as a reasonable null hypothesis. The problem then is to identify spatial patterns *above and beyond* that incorporated into the neutral model, enabling, for example, the identification of "hot spots" *beyond* background variation in a pollutant or the detection of clusters beyond regional variation in the risk of developing cancer.

Geostatistical simulation (Goovaerts 1997) provides fast and flexible ways to generate a large number of realizations of the spatial distribution of attribute values that reproduce the sample histogram and spatial patterns displayed by the data, and also account for any auxiliary data or information on the local trend. Its application to the generation of neutral models in health sciences must however ac-

count for specific features of cancer rates, that is 1) the irregular and non-punctual support of the data, and 2) the presence of noise which is often caused by unreliable extreme rates recorded over small areas, such as United States ZIP code areas or census tracts.

There have been relatively few applications of geostatistics to cancer data, with alternative solutions to the problem of non-stationarity of the variance caused by spatially varying population sizes. In his book (p.385-402), Cressie (1993) analyzed the spatial distribution of the counts of sudden-infant-death-syndromes for 100 counties of North Carolina. He proposed a two-step transform of the data to remove first the mean-variance dependence of the data and then the heteroscedasticity. Traditional variography was then applied to the transformed residuals. In another study on the risk of childhood cancer in the West Midlands of England, Oliver *et al.* (1998) developed an approach that accounted for spatial heterogeneity in the population of children to estimate the semivariogram of the "risk of developing cancer" from the semivariogram of observed mortality rates. Binomial cokriging was then used to produce a map of cancer risk. In their review paper Gotway and Young (2002) showed how block kriging can account for differing supports in spatial prediction (aggregation and disaggregation approach), allowing the analysis of relationships between disease and pollution data recorded over different geographies. More recently, Goovaerts *et al.* (2005) presented an adaptation of semivariogram and factorial kriging analysis that accounts for spatially varying population size in the processing of cancer mortality data.

Capitalizing on earlier works on binomial cokriging and weighted semivariograms of cancer mortality data, this paper presents first a geostatistical filtering approach for estimating cancer risk from observed rates. Sequential Gaussian simulation is then used to generate realizations of the spatial distribution of mortality rates under increasingly stringent conditions: 1) reproduction of the sample histogram, 2) reproduction of the pattern of spatial autocorrelation modeled from the data, 3) incorporation of regional background obtained by kriging of the local mean, and 4) integration of local trends in cancer rates inferred from the calibration of an exposure model. These alternate sets of neutral models are incorporated into traditional local cluster analysis algorithms. This approach is similar to the one described in more details in Goovaerts and Jacquez (2004), except that a new filtering procedure is implemented here and exposure data are used to derive the non-uniform risk model. The methodology is illustrated using Long Island, New York, breast cancer and exposure data which have been investigated under the spatial independence hypothesis in Jacquez and Greiling (2003a,b).

2 Setting the Problem

Consider the problem of detecting significant clustering and spatial outliers in the map of breast cancer incidence rates displayed in Fig. 1 (top graph). These data represent newly diagnosed cancer cases in the period 1993-7, and they are calculated as the number of cancers for each 100,000 people in the population. To pro-

tect patient privacy, the New York State Department of Health provided data referenced to ZIP codes rather than individual residences.

The local Moran test (Anselin 1995) evaluates local clustering or spatial autocorrelation. Its null hypothesis is that there is no association between rates in neighboring ZIP codes. The working (alternative) hypothesis is that spatial clustering exists. For each ZIP code, referenced geographically by its centroid with the vector of spatial coordinates **u**=(x,y), the so-called LISA (Local Indicator of Spatial Autocorrelation) statistic is computed as:

$$\text{LISA}(\mathbf{u}) = \left[\frac{z(\mathbf{u}) - m}{s}\right] \times \left(\sum_{j=1}^{J(\mathbf{u})} \frac{1}{J(\mathbf{u})} \times \left[\frac{z(\mathbf{u}_j) - m}{s}\right]\right) \quad (1)$$

where $z(\mathbf{u})$ is the incidence rate for the ZIP code being tested, which is referred to as the "kernel" hereafter. $z(\mathbf{u}_j)$ are the values for the $J(\mathbf{u})$ neighboring ZIP codes that are here defined as units sharing a common border or vertex with the kernel **u** (1-st order queen adjacencies). All values are standardized using the mean *m* and standard deviation *s* of the 214 ZIP codes. Since the standardized values have zero mean, a negative value for the LISA statistic indicates a negative local autocorrelation and the presence of spatial outlier where the kernel value is much lower or much higher than the surrounding values. Cluster of low or high values will lead to positive values of the LISA statistic.

In addition to the sign of the LISA statistic, its magnitude informs on the extent to which kernel and neighborhood values differ. To test whether this difference is significant or not, a Monte Carlo simulation is conducted, which traditionally consists of sampling randomly and without replacement the global distribution of observed rates (i.e. sample histogram) and computing the corresponding simulated neighborhood averages. This operation is repeated many times (e.g. *L*=999 draws) and these simulated values $z^{(l)}(\mathbf{u}_j)$ are multiplied by the kernel value to produce a set of *L* simulated values of the LISA statistic at location **u**:

$$\text{LISA}^{(l)}(\mathbf{u} \mid \text{SI}) = \left[\frac{z(\mathbf{u}) - m}{s}\right] \times \left(\sum_{j=1}^{J(\mathbf{u})} \frac{1}{J(\mathbf{u})} \times \left[\frac{z^{(l)}(\mathbf{u}_j) - m}{s}\right]\right) \quad l=1,\ldots,L \quad (2)$$

This set represents a numerical approximation of the probability distribution of the LISA statistic at **u,** under the assumption of spatial independence (Model I). The observed statistic, LISA(**u**), is compared to the probability distribution, allowing the computation of the probability of not rejecting the null hypothesis. The so-called *p*-value is compared to the significance level α chosen by the user and representing the probability of rejecting the null hypothesis when it is true (Type I error). The smaller α the fewer ZIP codes will be declared significant clusters or outliers. Following Jacquez and Greiling (2003a), an adjusted significance level α=0.01101 was used to account for the fact that the multiple tests (i.e. 214 in this study) are not independent since near ZIP codes share similar neighbors. This level was obtained using the Bonferroni adjustment which divides the chosen level α (here 0.05) by the average number of neighbors in each test. Every ZIP code where the p-value is lower than 0.01101 is classified as a significant spatial outlier

(HL: high value surrounded by low values, and LH: low value surrounded by high values) or cluster (HH: high value surrounded by high values, and LL: low value surrounded by low values). Table 1 and Fig. 1 indicate that most of the significant ZIP codes are declared low-low clusters and located in the western part of Long Island. A couple of high-high clusters occur in the eastern part of the island, nearby the single spatial outlier detected under this Model I.

The use of the SI null hypothesis means that the distribution of cancer rates is assumed to be spatially random with uniform risk over the study area. In most cases, however, rates are spatially correlated while the risk of developing cancer varies regionally as a result of changes in environmental exposure or other demographic, social, and economic factors. Another weakness of the above test is that it ignores whether incidence data are based on many or a few cases, thereby ignoring the instability of rates computed from small population sizes. Several modifications of the local Moran's I test of hypothesis have been proposed to take into account heterogeneous population sizes (e.g. see Assunçao and Reis 1992). An alternative is to randomly shuffle the counts rather than the rates (e.g. see Waller and Gotway 2004). A third option is to transform or standardize the rates prior to the application of the test, thereby removing much of the noise due to the small population size (Anselin *et al.* 2004, Goovaerts and Jacquez 2004). This is the option adopted in this paper and described in Section 3.

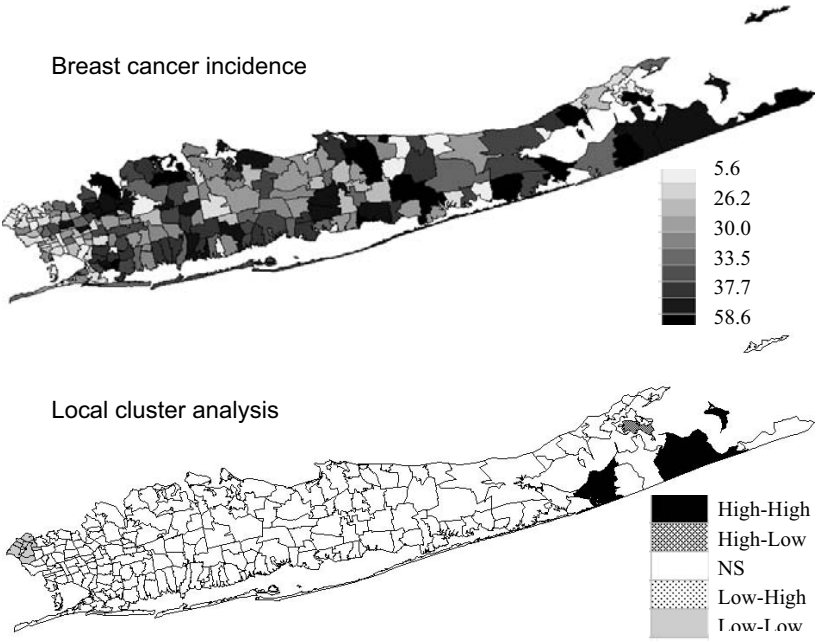


Fig. 1. Map of breast cancer incidence data in Long Island, New York, and the results of the cluster analysis under the hypothesis of spatial independence (Model I).

Table 1. Number of significant zip codes for the different types of cluster/outliers and neutral models. Numbers between parentheses indicate ZIP codes that have similar classification under the reference Model I (spatial independence).

	Neutral Model Type				
	Model I	Model II	Model III	Model IV	Model V
High-High	2	12(2)	2(2)	2(0)	12(0)
High-Low	1	0(0)	0(0)	0(0)	4(0)
Low-High	0	0(0)	0(0)	0(0)	1(0)
Low-Low	9	28(9)	9(9)	3(0)	13(4)
P-value					
Mean	0.222	0.191	0.261	0.257	0.168
Std. dev.	0.155	0.162	0.155	0.149	0.156

3 Geostatistical Analysis of Cancer Rates

The rates recorded at N=295 counties can be modeled as the sum of the risk of developing cancer and a random component (error term ε) due to spatially varying population size, $n(\mathbf{u}_\alpha)$:

$$Z(\mathbf{u}_\alpha)=R(\mathbf{u}_\alpha)+\varepsilon(\mathbf{u}_\alpha) \qquad \alpha=1,\ldots,N \tag{3}$$

Conditionally to a fixed risk function, the counts $d(\mathbf{u}_\alpha)=z(\mathbf{u}_\alpha)\times n(\mathbf{u}_\alpha)$ follow then a binomial distribution with parameters $R(\mathbf{u}_\alpha)$ and $n(\mathbf{u}_\alpha)$. The following relations are satisfied:

$$E[\varepsilon(\mathbf{u}_\alpha)]=0 \quad \text{and} \quad Var[\varepsilon(\mathbf{u}_\alpha)]=R(\mathbf{u}_\alpha)\times\{1-R(\mathbf{u}_\alpha)\}/n(\mathbf{u}_\alpha) \tag{4}$$

$$E[Z(\mathbf{u}_\alpha)]= E[R(\mathbf{u}_\alpha)]=\mu \quad \text{and} \quad Var[Z(\mathbf{u}_\alpha)]= Var[R(\mathbf{u}_\alpha)]+Var[\varepsilon(\mathbf{u}_\alpha)] \tag{5}$$

For estimation purpose and in agreement with Oliver *et al.* (1998), the variance of the error term can be approximated as $Var[\varepsilon(\mathbf{u}_\alpha)]=\sigma_\varepsilon^2= \mu \times(1- \mu)/n(\mathbf{u}_\alpha)$, where μ is estimated by the population-weighted average of rates, $\bar{z}$. The risk is then estimated from $s(\mathbf{u}_\alpha)$ neighboring observed rates using a form of cokriging:

$$\hat{R}(\mathbf{u}_\alpha) = \sum_{i=1}^{s(\mathbf{u}_\alpha)} \lambda_i(\mathbf{u}_\alpha) z(\mathbf{u}_i) \tag{6}$$

The kriging weights are solution of the following system:

$$\sum_{j=1}^{s(\mathbf{u}_\alpha)} \lambda_j(\mathbf{u}_\alpha) C(\mathbf{u}_i - \mathbf{u}_j) + \mu(\mathbf{u}_\alpha) = C_R(\mathbf{u}_i - \mathbf{u}_\alpha) \qquad i = 1,\ldots,s(\mathbf{u}_\alpha)$$
$$\sum_{j=1}^{s(\mathbf{u}_\alpha)} \lambda_j(\mathbf{u}_\alpha) = 1 \tag{7}$$

where $C(\mathbf{u}_i-\mathbf{u}_j)=\{1-1/n(\mathbf{u}_i)\}C_R(0)+\bar{z}\times(1-\bar{z})/n(\mathbf{u}_i)$ if $\mathbf{u}_i=\mathbf{u}_j$ and $C_R(\mathbf{u}_i-\mathbf{u}_j)$ otherwise. The addition of the measurement error variance for a zero distance accounts for variability arising from population size, leading to smaller weights for less reliable (i.e. measured over smaller population) data. System (7) requires knowledge of the covariance of the unknown risk, $C_R(\mathbf{h})$. Oliver *et al.* (1998) proposed an estimator for the semivariogram of the risk but its application to Long Island data leads to negative values, a feature that has been observed on various datasets with different geographies and population sizes. According to simulation studies (Goovaerts 2005) this problem is caused by the overestimation of the variance of the error term by the expression $\bar{z}\times(1-\bar{z})/n(\mathbf{u}_\alpha)$. In other words, all developments (3) through (7) are based on the modeling of the error term (expressed in terms of counts) as a Binomial random variable, an assumption which might not always be consistent with the observed variability. The following empirical modification of the binomial cokriging approach has been tested on simulations and proved to be more robust with respect to misspecification of underlying hypothesis.

A more robust estimator of the semivariogram of the risk is the population-weighted semivariogram, which is similar to the weighted semivariogram described in Rivoirard (2000):

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\sum_{\alpha=1}^{N(\mathbf{h})} n(\mathbf{u}_\alpha) n(\mathbf{u}_\alpha+\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} n(\mathbf{u}_\alpha) n(\mathbf{u}_\alpha+\mathbf{h}) \left[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha+\mathbf{h})\right]^2 \tag{8}$$

The weighting scheme attenuates the impact of data pairs that involve at least one rate computed from small population sizes, revealing structures that might be blurred by the random variability of extreme values.

The second modification relates to the kriging system (7) itself. In particular the term $\bar{z}\times(1-\bar{z})/n(\mathbf{u}_\alpha)$ can become disproportionately large relatively to the variance of the risk $C_R(0)$, leading to very large diagonal elements in the kriging matrix (and indirectly very large nugget effect). Such a severe understatement of the spatial correlation between rates typically results in over-smoothing since the risk becomes a simple population-weighted average of observed rates. An easy way to check for any discrepancy is to compare the sill of the semivariogram of observed rates $C_Z(0)$ with the value of the error variance averaged over all locations:

$$G = \frac{1}{N} \sum_{\alpha=1}^{N} \frac{\bar{z}(1-\bar{z})}{n(\mathbf{u}_\alpha)} \tag{9}$$

For Long Island data $C_Z(0)=4.48910^{-9}$, while G is one order of magnitude larger $G=7.93710^{-8}$. The proposed modification of the binomial cokriging system consists of rescaling the correction of the diagonal term to account for any discrep-

ancy between estimates of the rate and error variances, that is $C(\mathbf{u}_i-\mathbf{u}_j)=\{1-1/n(\mathbf{u}_i)\}$ $C_R(0)+\{\bar{z}\times(1-\bar{z})/n(\mathbf{u}_i)\}\{C_Z(0)/G\}$. Simulation studies (Goovaerts 2005) have shown that: (1) the use of population-weighted semivariogram and empirical rescaling of diagonal terms of the cokriging system yields estimates of the risk that are nearly as accurate as the ones obtained using the true (but unknown in practice) semivariogram of the risk, and (2) the rescaling of the diagonal term systematically outperforms the formulation in Eq. (7).

Fig. 2 shows the omnidirectional semivariogram of breast cancer incidence data obtained using a traditional or population weighted estimator. Incorporating the population size into the estimation reduces the overall variability (lower sill) as well as the nugget variance, and leads to smaller fluctuations around the nested model. The central scattergram highlights the greater variability of rates recorded for small population sizes. These extreme rates disappear after filtering by the modified binomial cokriging algorithm: the standard deviation is halved while the averaged rate is unchanged. The map of filtered rates, displayed at the top of Fig. 3, appears more homogeneous or smoother than the map of raw rates in Fig. 1. Therefore, the cluster analysis leads to the detection of larger clusters of high and low values, located in the eastern and western parts of Long Island, respectively. The only spatial outlier detected on Fig. 1 is non-significant anymore, as it corresponds to the sparsely populated Shelter Island (574 habitants).

4 Cluster Analysis using Spatial Neutral Models

Results in Fig. 1 and 3 are based on spatial independence as the null hypothesis, which means that the spatial distribution of cancer incidence rates is assumed to be random (no autocorrelation) with uniform risk over the study area. This assumption clearly disagrees with the structured semivariogram of Fig. 2, and more realistic neutral models would be ones that reproduce not only the sample histogram, but also the pattern of spatial correlation observed in the data.

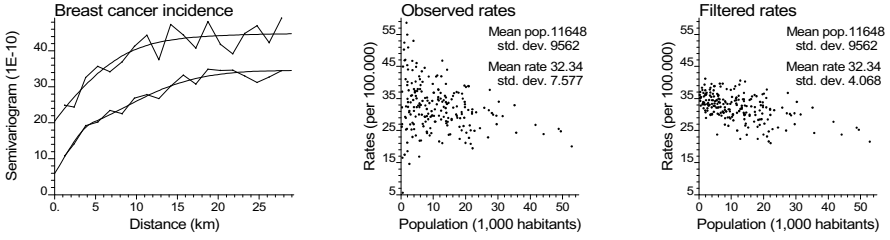


Fig. 2. Experimental semivariogram of observed rates before (top curve) and after weighting by the population size, with the model fitted. Scatterplots illustrate the impact of filtering on variability among rates, as a function of the female population size.

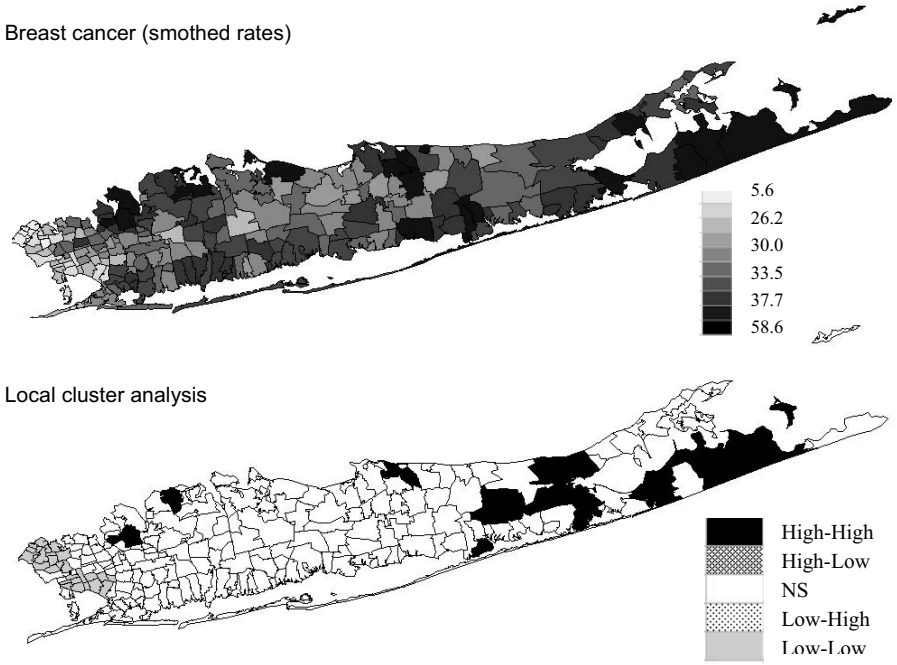


Fig. 3. Map of filtered breast cancer incidence data in Long Island, New York, and the results of the cluster analysis under SI hypothesis (Model II).

Following Goovaerts and Jacquez (2004) spatial neutral models are generated using sequential Gaussian simulation, using either a global conditioning (only the histogram and semivariogram model of filtered rates are incorporated) or a global and local conditioning to reproduce the location of high and low-valued zones. This local conditioning was achieved using simple kriging with spatially varying local means instead of a global constant mean to derive the mean and variance of local probability distribution functions (see Goovaerts 1997 for more details). These local means were identified to the regional background of incidence data (i.e. obtained by setting $C_R(\mathbf{u}_i-\mathbf{u}_\alpha)=0$ in the kriging system (7)) or derived by calibration of an airborne carcinogen exposure model described in Jacquez and Greiling (2003b). In this later case, the relationship between exposure and incidence rates was modeled using linear functions fitted separately to low exposures observed in the eastern part of Long Island and high exposures in the western part.

Since the number of simulated values equals the number of observed values, the normal score transform and back-transform which precedes and follows the simulation are fairly straightforward. The simulation can then be viewed as a two-step procedure: 1) normal scores are simulated in space according to a given spatial covariance and local means, 2) the simulated normal scores are ranked from the smallest to the largest and replaced by the filtered rates that have the same rank in the sample histogram (i.e. equal *p*-quantiles correspondence). In other words, the

rates are not directly simulated but rather their arrangement in space is simulated using sequential Gaussian simulation.

Depending on the type of secondary information used for the local conditioning, the neutral models are referred to as Models IV (background defined by regional incidence) and V (background modeled using environmental exposure). Model III corresponds to non-conditional simulation which reproduces only the pattern of spatial autocorrelation. The different neutral models and the corresponding null hypothesis are summarized in Table 2. Once the L sets of N simulated rate values, $\{z^{(l)}(\mathbf{u}_\alpha); \alpha=1,\ldots,N\}$ have been generated, they are imported into Eq. (2) to compute the simulated values of the LISA statistic at each location u_α and the resulting *p*-value for the test of hypothesis.

For each of the three types of neutral models, 999 realizations were generated and used to compute the LISA statistic defined in Eq. (2). For example, Fig. 4 shows the distribution of simulated LISA values for the ZIP code # 11435 under Models II through IV. Clearly, the variance of the distribution in Model III (central graph) is much larger than the results obtained under spatial independence (Model II, left graph), while the means are very similar and close to zero. The spatial autocorrelation of simulated rates increases the likelihood that the J neighboring values are jointly small or high, causing the neighborhood average, hence the LISA value, to exhibit much larger fluctuations among realizations. Consequently, the probability that the observed LISA statistic lies in the tails of the simulated distribution decreases, leading to a larger p-value (0.116 versus 0.008 for this ZIP code). The same pattern is observed for all ZIP codes, with an average p-value increasing from 0.191 to 0.261 (see Table 1). These larger p-values cause a substantial reduction in the size of significant LL or HH clusters, which confirms previous findings regarding the increased risk of type I error when ignoring the presence of spatial autocorrelation in the data.

Table 2. Typology of neutral models based on the spatial characteristics of the risk being simulated.

Model	Risk
I	Uniform, spatially random (raw rates)
II	Uniform, spatially random (smoothed rates)
III	Uniform, spatially correlated
IV	Heterogeneous (regional background), spatially correlated
V	Heterogeneous (environmental exposure), spatially correlated

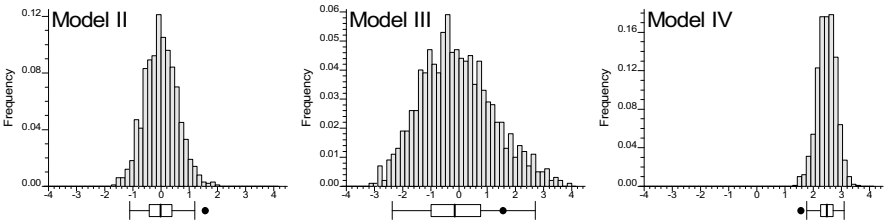


Fig. 4. Histograms of the values of the LISA statistic simulated for ZIP code # 11435, identified in Fig. 5, under different neutral models. The black dot denotes the observed LISA statistic which lies inside the 0.95 probability interval for Model III.

The local conditioning of realizations generated under Models IV and V entails that locations of high and small values are now reproduced by the neutral models. This causes less variation among realizations, leading to the J neighboring values being consistently either small or large across the realizations. Thus the distribution of 999 simulated LISA values is expected to be narrower than for the two previous models with a shift in the mean. This is illustrated for the ZIP code # 11435 in Fig. 4 (right graph). Because this unit has a small rate and is located in a low-valued area, the use of neutral models reproducing the regional background yields large positive simulated LISA values (average=2.48 instead of -0.05). If this unit had a high rate, the shift would have been in the opposite direction.

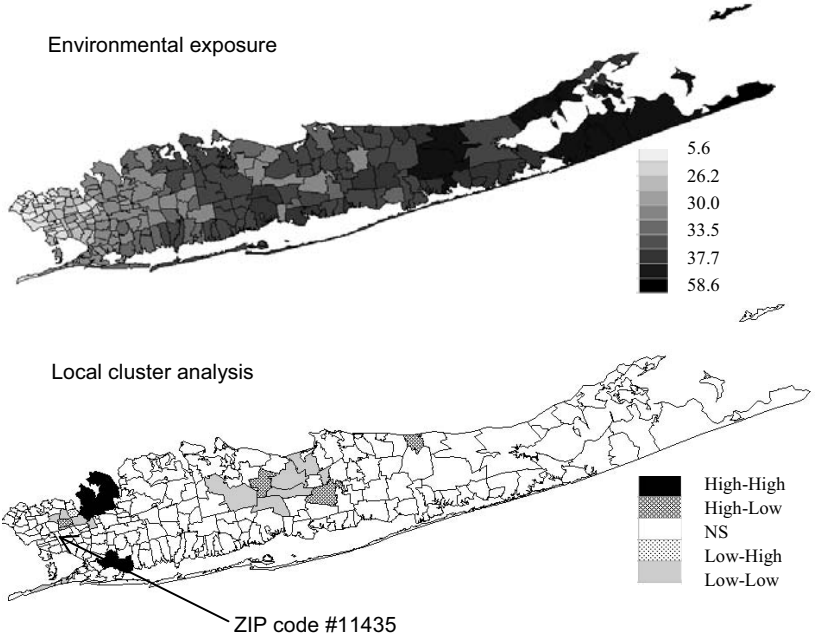


Fig. 5. Background risk inferred from the environmental exposure model, and the results of the cluster analysis to detect spatial pattern above and beyond this risk (Model V).

Incorporation of local information in the generation of neutral models allows the testing of more complex null hypotheses. For example, Model IV is useful to detect any departures from the regional background of incidence values. Model V reflects the situation where environmental exposure makes the risk of developing cancer non-uniform. In this instance the researcher wishes to detect spatial pattern above and beyond this non-uniform risk. These questions are more specific than the ones tackled under the previous neutral models; hence the cluster analysis leads to substantially different results, see Table 1. For example, the exposure model leads to a very different map of spatial clusters and outliers. Fig. 5 (bottom graph) reveals a series of ZIP codes that are significant high clusters in the North western part of the Island. Cancer incidences in these ZIP codes are higher than expected under the environmental exposure model and should warrant further investigation to identify additional cofactors. Note that for situations where health professionals are mostly interested in identifying areas with generally high (or low) disease rates, the focus would be on the detection of cancer clusters above and beyond a null hypothesis of constant risk.

5 Conclusions

The approach presented in this paper enables researchers to assess geographic relationships using more realistic null hypotheses that account for spatial correlation and background variation modeled from observed rates and any ancillary information, such as exposure. An immediate consequence of using spatially correlated neutral models are larger p-values, leading to a substantial reduction in the number of ZIP codes declared significant outliers or clusters across Long Island. This result confirms earlier finding that the SI hypothesis often leads to an over-identification of the number of significant spatial clusters or outliers. When the constraint of local conditioning of neutral models is superimposed to the reproduction of spatial autocorrelation (i.e. models IV and V), the approach allows one to detect local departures from the incidence background specified by the user.

Another issue, which often impacts the results of cluster analysis, is the lack of reliability of rates inferred from small populations. If ignored, large differences in population size decrease the ability of Moran's I to detect true clustering or departures from spatial randomness. Binomial cokriging has been adapted to the situation where the variance of observed rates is smaller than expected under the binomial model, thereby avoiding negative estimates for the semivario-gram of the risk. The smoothing of local fluctuations, in particular the ones recorded in sparsely populated ZIP codes, resulted in the detection of larger and more compact clusters of low or high SMR values as well as the disappearance of some unreliable spatial outliers. Other methods could be used (i.e. Empirical Bayes smoother) and a performance comparison with binomial cokriging is under way.

Acknowledgments

This research was funded by grant 1R43CA105819-01 from the National Cancer Institute. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI.

References

Anselin L (1995) Local indicators of spatial association – LISA. Geographical Analysis 27: 93–115

Anselin L, Syabri I, Kho Y (2004) GeoDa: An Introduction to Spatial Data Analysis. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Champaigne-Urbana, IL

Assunçao RM, Reis EA (1999) A new proposal to adjust Moran's I for population density. Statistics in Medicine 18:2147-2162

Cressie N (1993) Statistics for Spatial Data. Wiley, New York

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press New York

Goovaerts P, Jacquez GM (2004) Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. International Journal of Health Geographics 3:14

Goovaerts P (2005) Simulation-based assessment of a geostatistical approach for estimation and mapping of the risk of cancer. In: Leuangthong O and Deutsch CV (eds) Geostatistics Banff 2004. Kluwer Academic Publishers, Dordrecht, The Netherlands, in review

Goovaerts P, Jacquez GM, Greiling DA (2005) Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. Geographical Analysis, 37, in press

Gotway CA, Young LJ (2002) Combining incompatible spatial data. Journal of the American Statistical Association 97:632-648

Jacquez GM, Greiling DA (2003a) Local clustering in breast, lung and colorectal cancer in Long Island, New York. International Journal of Health Geographics 2:3

Jacquez GM, Greiling DA (2003b) Geographic boundaries in breast, lung and colorectal cancer in relation to exposure to air toxics in Long Island, New York. International Journal of Health Geographics 2:4

Oliver MA, Webster R, Lajaunie C, Muir KR, Parkes SE, Cameron AH, Stevens MCG, Mann JR (1998) Binomial cokriging for estimating and mapping the risk of childhood cancer. IMA Journal of Mathematics Applied in Medicine and Biology 15:279-297

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. Journal of Regional Science 41: 411-432

Rivoirard J. *et al.* (2000) Geostatistics for Estimating Fish Abundance. Blackwell Science, Oxford

Waller LA, Gotway CA (2004) Applied Spatial Statistics for Public Health Data. John Wiley and Sons, New Jersey

Geostatistical assessment of long term human exposure to air pollution

N. Jeannée[1], V. Nedellec[2], S. Bouallala[3], J. Deraisme[1] and H. Desqueyroux[3]

[1] GEOVARIANCES, 49bis av. Franklin Roosevelt,77212 Avon, France, e-mail : jeannee@geovariances.com
[2] Vincent Nedellec Consultants, 15 rue Firmin Gillot, 75015 Paris, France
[3] ADEME, Direction de l'Air et des Transports, Département Air, 27 rue Louis Vicat, 75737 Paris Cedex 15, France

1 Introduction

Health Impact Assessment (HIA) on air pollution is a scientific based approach that allows to forecast impact of air pollution on public health. Epidemiological studies investigate the relationship between temporal variation of pollutant air concentrations (data from air monitoring network) and health outcomes in the population (data from hospitals or other public health institutions or measured in a representative sample of the population). Exposure response functions (ERF) are derived from these studies; these functions estimate the number of cases (morbidity or mortality) for a given atmospheric concentration of a given air pollution indicator. This approach can be used to compare different scenarios or the efficiency of measures introduced to reduce air pollutant concentration. Complementary to epidemiological or other research efforts, HIA studies are frequently used for decision making and evaluating the economic consequences of the impact of air pollution on health (APHEIS 2000).

Specific HIA on transport-related air pollution requires the accurate assessment of the population exposure to chemical compounds that are indicators of transport-related pollution. This assessment implies the ability to separate traffic-related air pollution from other air pollution sources (building, industry, energy, etc.). The paper focus on a preliminary step of this HIA: the assessment, with data from the French air monitoring network, of ambient air concentrations of particulate matter PM_{10} (particulate matter with an aerodynamic diameter less than 10 micron). These concentration results are issued from the French national air quality databse managed by ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie).

Numerous concordant epidemiological study results established exposure response functions between PM_{10} air concentration and an increased frequency in many health outcomes (see Mosqueron *et al.* 2003 for a comprehensive list). Though black smokes and $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than 2.5 micron) seem to be better indicators of exposure to traffic emissions

than PM_{10}, the lack of available measures and exposure functions did not permit us to base the HIA on these pollutants.

As the interest is put on long-term exposure effects, the study is based on average annual PM_{10} concentrations from existing measuring stations in France. The exposure estimation ignores the day to day variability of the ambient PM_{10} air pollution. Special attention is paid to the heterogeneous nature of the available data (rural, roadside, industrial and urban stations), excluding in particular proximity stations because of their lack of representativity.

Geostatistics is applied in air pollution since a few years, mainly to map air pollutants at urban (Cressie 1998), regional (Roth 2001) or national scale (Deraisme *et al.* 2002). Computing the exposure of population to given levels of air ambient pollutants with geostatistics is more recent (Deraisme *et al.* 2002). This computation requires quantifying the local uncertainty associated with air pollutant levels. Linear estimation techniques such as kriging or cokriging are not adapted to solve non linear problems. Furthermore, though non linear techniques allow the computation of specific characteristics such as the probability to exceed a threshold, only simulations provide a general framework if the interest is to quantify populations exposed to specific pollutant levels (Deraisme *et al.* 2002).

A significant increase in the reliability of the results is obtained by taking into account the existence of: (i) a correlation between PM_{10} concentrations and more densely acquired NO_2 data, and (ii) more recent PM_{10} data that supplement the PM_{10} monitoring network in otherwise entirely non sampled areas.

Air concentration results are then coupled with geo-data from the last national census (1999). The population is stratified in 5 years age classes equal to those used in epidemiological studies from which the exposure response function are derived. The population exposed to different levels of average annual concentrations is then calculated. Statistical parameters from the resulting distributions are derived in the perspective of carrying out the HIA study on transport related air pollution.

The exposure and health assessment case studies are part of the French research effort contributing to the UNECE-WHO Pan European Program for Transport, Health and Environment (THE PEP Project): "Transport-related health impacts and their costs and benefits with a particular focus on children". The aim of the international project is to:

- provide sound scientific information on social costs and the impact of road traffic to the European Environment Minister,
- recommend action or regulation that can decrease external costs or protect population health.

Firstly the paper recalls the main geostatistical methods involved in the suggested approach in order to (i) take into account additional data though kriging with variance of measurement error, (ii) introduce auxiliary variable with cokriging and (iii) compute the exposure of population through a stochastic simulation approach. Then the two major aspects of the case study are presented: spatial modeling of PM_{10} and evaluation of population exposure to PM_{10} pollution. The relevance of the approach in the framework of an HIA is finally discussed.

2 Methodological aspects

2.1 Kriging with Variance of Measurement Error

Numerical values with varying levels of precision might be available for the variable of interest. For example, the data may come from several surveys: old ones and new ones, the latter being more accurate due to advances in measurement techniques. In such cases error variances albeit different for each sub-population may be known. Certain data might be assumed to have an error variance of 0, whilst some indirect or old measures are uncertain with a known error variance.

Suppose that, instead of the "true" concentration value z_i we only know $z_i + e_i$ where e_i is a random error satisfying the following conditions for each sampling point i: $E[e_i]=0$, $Cov[e_i, e_j]=0$ for $j \neq i$, $Cov[z_i,e_i]=0$ and $Var[e_i]=v_i$, where the constant value v_i may differ for each i. Kriging with variance of measurement error (VME) consists of integrating these error variances. From a kriging system point of view, the variance of measurement error simply consists of adding the v_i values to the diagonal covariance terms, or in replacing the 0 diagonal values by $-v_i$ in variogram terms (Geovariances 2004).

2.2 Cokriging

Secondary information about the phenomenon is usually available in addition to the pollutant concentrations available over a set of sample points: concentrations of correlated pollutants potentially measured at other locations, cofactors exhaustively known over the area of interest, etc. Cokriging techniques aim at integrating this secondary information and therefore reducing the uncertainty about the variable of interest at non-sampled locations. Ordinary cokriging is the classical multivariate extension of ordinary kriging and is therefore not described here (see for example Chilès and Delfiner 1999).

2.3 Stochastic simulations

Stochastic simulations require the variable of interest to follow a gaussian distribution, which is generally not the case. It is then recommended to transform the raw distribution into a standard gaussian one. Therefore, we consider the stationary random function $Z(x)$ as a function $Z(x)=\Phi[Y(x)]$ of a gaussian one $Y(x)$, where x is the spatial location in 2D. The "anamorphosis" function Φ is determined by the coefficients of its truncated development in orthogonal Hermite polynomials. Finally, we associate to each raw value z_i a gaussian transform value having the same cumulate frequency as z_i (Rivoirard 1994).

Moreover, stochastic simulations require the multivariate distribution of variables $(Y(x), Y(x_1), ...)$ to be multigaussian, i.e. any linear combination of these variables should be normally distributed. Except in the case of an exhaustive sys-

tematic sampling, the validation of the multigaussian assumption is quite inextricable and is most often reduced to the validation of the bigaussian assumption. Several tests exist to evaluate the bigaussian assumption: examination of h-scatterplots, computation of the ratio $\sqrt{\gamma(h)}/\gamma_1(h)$ between variogram and madogram (first order variogram), which has to be constant and equal to $\sqrt{\pi}$, validation of the relationship between raw and gaussian covariances (Lajaunie 1993, Chilès and Delfiner 1999).

Co-simulations are performed using the Turning Bands (TB) technique. The basic idea of the TB algorithm consists of simplifying the 2D simulation in several 1D simulations along randomly generated lines, and then reconstructing the 2D simulation by averaging the projected values from the 1D simulations (Matheron 1973). The only parameter required to ensure consistency of the resulting simulations (i.e. histogram and variogram reproduction) with the TB technique is the number of turning bands. Although theoretical results may give some hints about this number (Lantuéjoul 2002), a pragmatic approach usually consists in analyzing visually the quality of the simulations, that should not reveal the existence of the generated bands, and then to check the quality of the simulations.

All the geostatistical results are obtained using version 5.0 of the Isatis software (Geovariances 2004).

2.4 Estimation of population exposure

The number of inhabitants is known for each 4km x 4km cell of the estimation grid. This spatial resolution is thought to be a good compromise between precision and representativity. From each stochastic simulation of air pollutant, the knowledge of the population within each cell easily allows the computation of the total population exposed to a given interval of pollution over France, e.g. total population exposed to PM_{10} concentrations between 5 and 10 $\mu g/m^3$. Repeating this operation for all the simulations leads to the distribution of the French population exposed to an average annual PM_{10} concentration between 5 and 10 $\mu g/m^3$. Classical characteristics about this statistical distribution (mean, standard deviation, median, quantiles) are finally derived for conducting the HIA.

3 Spatial modeling of PM_{10} concentrations

3.1 Data analysis

Air quality monitoring in France is based on a network of 740 measuring stations designed to respect the European and National regulations on air quality (Directive 96/62/EC and LAURE, Dec. 30 1996). Each station belongs to one of the following classes: urban, near-city background, regional rural, national rural, roadside, industrial and specific observation. Urban, peri-urban and rural

(regional/national) stations constitute the "background stations", as opposed to "proximity stations" (ADEME 2002).

185 monitoring stations provided PM_{10} average annual concentrations in 2000 (see Fig. 1a). Among them, 54 proximity stations were not considered for modeling the annual background pollution in PM_{10}, due to their lack of spatial representativity. The most noticeable fact is the absence of measures in most of the rural stations.

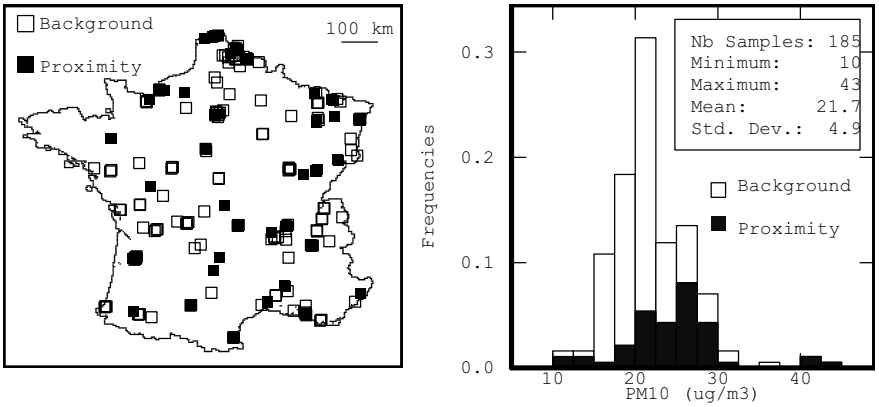
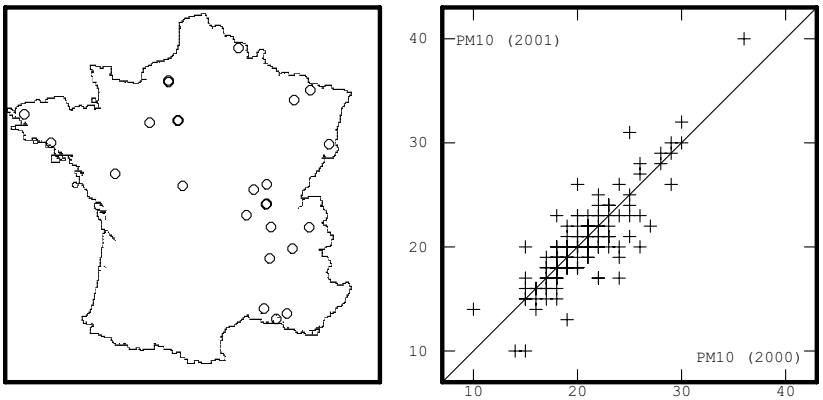


Fig. 1. Monitoring network for PM_{10} in 2000, background (empty squares) and proximity stations (filled black squares): (a) base map, (b) histogram of PM_{10} with global statistics.

The 131 PM_{10} concentrations from background stations are lying between 10 and 36 $\mu g/m^3$ with an average concentration of 20.7 $\mu g/m^3$, slightly inferior to the overall (from the 185 stations) PM_{10} average concentration of 21.7 $\mu g/m^3$ (Fig. 1b). This difference is due to the fact that we have removed proximity stations, associated to large particle emissions. Peri-urban stations have an average concentration of 18.8 $\mu g/m^3$ slightly less than the urban stations (21.1 $\mu g/m^3$).

The sampling density of the monitoring network is highly heterogeneous, particularly away from the urbanized areas. The development of the monitoring network led to more abundant PM_{10} measures in 2001; we therefore planned to integrate this indirect information for mapping PM_{10} in 2000.

Fig 2a illustrates the location of the 23 new monitoring stations of PM_{10} in 2001. In order to quantify the information brought by these additional stations, we used linear regression between existing 2000 and 2001 stations, calculated on the 129 background stations where both data are present (see the Fig 2b, correlation coefficient equal to 0.84). On the 23 stations only available in 2001, the missing PM_{10} values have been replaced by the result of the linear regression.



(a) (b)

Fig. 2. a) Stations where PM_{10} measurements are only available for 2001. **b)** Scatter diagram of the 2000 vs. 2001 PM_{10} background concentrations, first bisector indicated.

As these additional data cannot be put at the same level as real measurements, they have been "penalized" by a VME equal for all the stations to the value of the variance of residuals around the linear regression, i.e. 4. A standard cokriging approach could similarly have been used to introduce these 23 additional data. The interest here is to evaluate whether or not kriging with VME, which avoids the bivariate modeling required by the cokriging, leads to satisfactory results. The efficiency of kriging with VME and cokriging will be compared.

In order to introduce additional information for improving PM_{10} modeling we finally studied the potential correlation with more densely sampled NO_2, this pollutant being measured at 296 stations in year 2000 (see Fig. 3).

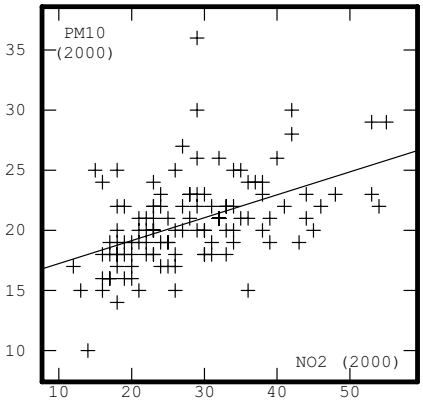


Fig. 3. Scatter diagram of PM_{10} and NO_2 (both in $\mu g/m^3$), regression line indicated.

Both pollutants having a partly common origin linked to traffic emissions, their positive correlation (correlation coefficient equal to 0.49), although not strong, justifies the integration of NO_2 data in the our PM_{10} modeling.

3.2 Modeling spatial variability

Because of the significant correlation between PM_{10} and NO_2, a bivariate variogram calculation is performed on this heterotopic dataset (see Fig. 4).

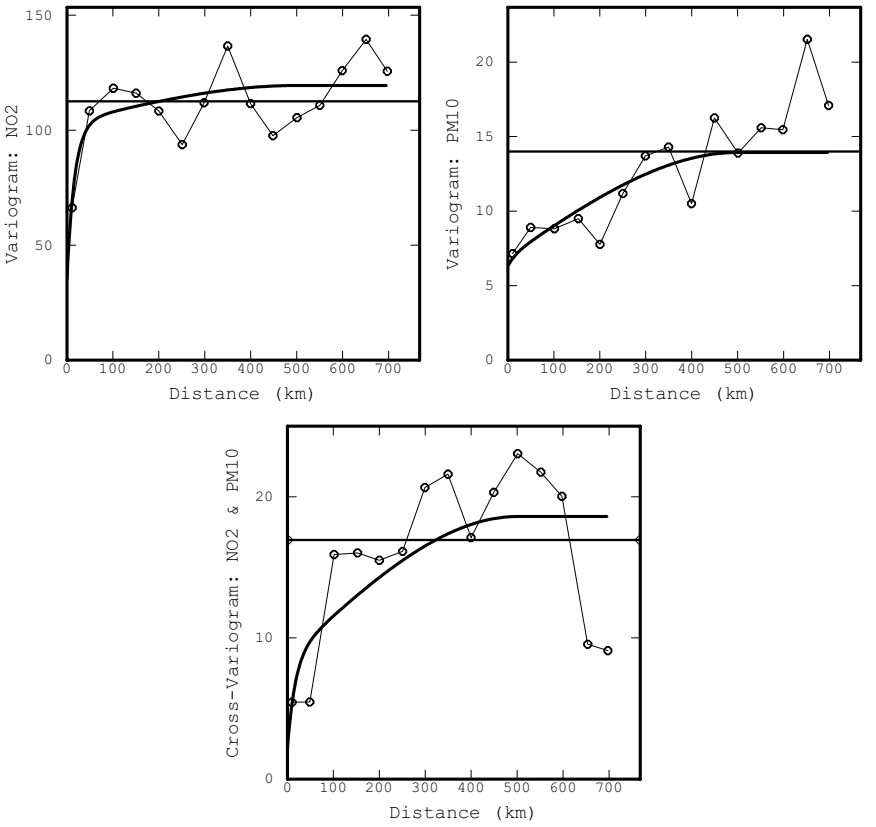


Fig. 4. Simple and cross experimental variograms for PM_{10} and NO_2. Statistical (a priori) variance / covariance indicated by a dash line; fitted variogram models in bold.

The bivariate variogram model is obtained by an automatic sill fitting procedure (Lajaunie and Behaxétéguy 1989) of the following basic structures to the experimental variograms:

- Nugget effect: Variability at very small scale, due to potentially important local variations;
- Exponential structure of range 50 km: Steep increase of the variability at the scale of 50 km, particularly close to agglomerations;

- Spherical structure of range 500 km: Slight correlation at large scale due to the large representativity of the rural stations.

The procedure ensures that the obtained linear model of coregionalization is authorized.

3.3 Model validation

Different models have been compared to predict PM_{10} concentrations in 2000: a quick interpolation by Inverse Squared Distances, an ordinary kriging of PM_{10} (2000), an ordinary kriging of PM_{10} (2000) completed by PM_{10} (2001) with a VME approach, an ordinary cokriging of PM_{10} (2000) and PM_{10} (2001), an ordinary cokriging of PM_{10} (2000 completed by 2001 with a VME approach) and NO2.

The comparison is performed by randomly dividing the dataset into five unconnected validation sets (constituted of 22 to 30 monitoring stations); each technique is sequentially used to re-estimate the PM_{10} concentrations on these validation sets. A moving neighborhood involving the 50 closest data within a circle of radius 300 km centered at the target point is systematically used. Median squared errors between estimated and real PM_{10} concentrations are computed and ranked for each validation set from 1 (smallest MSE) to 5 (largest MSE); MSE and mean rank results are summarized in Table 1.

Table 1. For each estimation technique, Mean Squared Error (MSE) from each validation set and Mean Rank. Abbreviations: inverse squared distance (ID2), ordinary kriging (IK) and ordinary cokriging (OCK).

Estimation Technique	MSE for validation set					Mean Rank
	1	2	3	4	5	
ID2 PM_{10} 2000	25,95	13,88	16,30	12,44	21,63	4,6
OK PM_{10} 2000	10,09	7,26	15,34	11,91	6,80	2,8
OK PM_{10} 2000 comp2001	11,77	6,52	12,18	14,89	4,35	2,8
OCK PM_{10} 2000 / PM_{10} 2001	10,13	6,78	12,35	13,13	6,41	3,0
OCK PM_{10} 2000comp / NO2	10,53	6,67	11,79	12,21	4,10	1,8

Several conclusions can be drawn from these results. Firstly, kriging with VME and cokriging lead to similar validation results to integrate the 23 additional PM_{10} data from 2001; kriging with VME is therefore preferred because of its simplicity. Then, the ordinary cokriging of PM_{10} in 2000 (completed by data from 2001) and NO_2 leads to the best results and therefore constitutes the recommended approach for this case study. The resulting PM_{10} map is illustrated in Fig. 5.

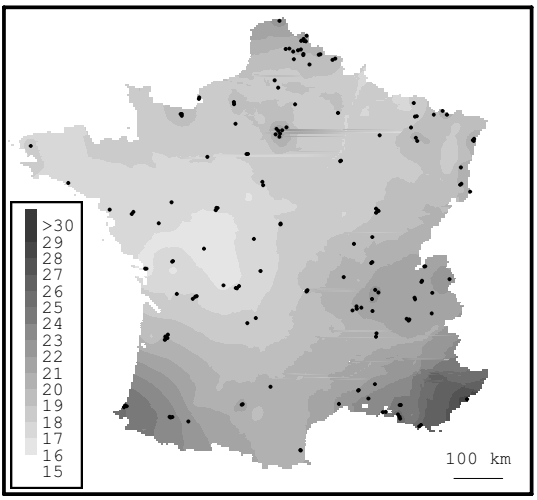


Fig. 5. Estimation of PM_{10} (integrating 2001 data) by cokriging with NO_2 (in $\mu g/m^3$).

4 Population exposure to PM_{10}

4.1 Analysis of gaussian transforms (anamorphosis)

PM_{10} data have been transformed into gaussian data. Simultaneously the anamorphosis function has been modeled. Because the spatial distribution of PM_{10} concentrations is highly heterogeneous, this clustering has been taken into account using a standard cell declustering algorithm (Isaaks and Srivastava 1989), to avoid any bias in the gaussian transformation and on the resulting simulations. The validity of the underlying bigaussian assumption has been checked with the computation of h-scatter plots and of the ratio $\sqrt{\gamma(h)}/\gamma_1(h)$, which is reasonably constant and equal to $\sqrt{\pi}$. Even though we are only interested in PM_{10}, the gaussian transform is recommended for both PM_{10} and NO_2 variables, as:

- analyzing the correlation and bivariable spatial structure between two gaussian transforms usually yields to better results, and ensures the homogeneity of the process,
- the Turning Bands co-simulation algorithm requires first the non conditional simulation of both variables in the gaussian space.

Variograms of PM_{10} and NO_2 gaussian transforms have been calculated and modeled using the same basic structures as those used for the raw concentrations. Attention is paid to the fact that the simple variograms sills should not be larger than 1 (variance of a standard gaussian variable), to avoid unrealistic simulated results after back-transformation.

Note that the VME approach presented for the spatial modeling is in general not applicable to gaussian transforms. Indeed, we usually only know the error variance of raw measurements, and not its counterpart for gaussian transforms. However, in our particular case, the approach used to compute this error variance for the 2001 data is based on regression using gaussian transform values, not actual knowledge of measurement error, and is therefore integrated in the analysis.

4.2 Stochastic simulations of PM_{10} concentrations

200 conditional co-simulations have been performed using the Turning Bands algorithm, with 500 turning bands. The number of turning bands is the only key parameter required to ensure the quality of the simulations. The latter has been verified on a few simulations, in terms of histogram and spatial structure reproduction, before the back-transformation in raw scale.

4.3 Exposure frequency of the population

The number of inhabitants per grid cell is derived from the last national census (1999). For the HIA purpose, the focus is on the population exposed to consecutive PM_{10} classes of 5 $\mu g/m^3$ and the results are stratified according to several criteria (age class, restricted to urban areas only, etc).

The statistical results for the total population over the entire French territory are presented in Table 2. Mean exposed populations show a major exposure to PM_{10} levels comprised between 15 and 30 $\mu g/m^3$. These values might be compared with the current limit value for PM_{10} annual average concentration (based on scientific knowledge, maximum concentration value accepted to avoid, prevent or reduce harmful effects on human health), equal to 40 $\mu g/m^3$ while the French quality objective is equal to 30 $\mu g/m^3$.

Table 2. Population (in billions) exposed to classes of annual PM_{10} concentrations (in $\mu g/m^3$) over the entire French territory.

PM_{10} class ($\mu g/m^3$)	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Mean	0.22	1.88	26.87	23.12	5.11	0.75	0.29
Standard deviation	0.12	0.44	1.48	1.27	0.78	0.26	0.18
Percentile 2.5%	0.07	1.15	24.01	20.66	3.84	0.30	0.06
Percentile 97.5%	0.49	2.73	29.43	25.29	6.78	1.31	0.74

5 Conclusions

This paper illustrated the efficiency of geostatistics in providing the basic figures of a specific Health Impact Assessment (HIA) on air pollution based on PM_{10}. A significant correlation with NO_2 has been exploited. The capability of geostatistical methods to go beyond a mere mapping, by providing a quantification of the uncertainty has been emphasized. The geostatistical framework offers the possibility to generate PM_{10} stochastic simulations that take into account the correlation between PM_{10} and NO_2. The population exposure to different levels of PM_{10} concentrations is derived from these simulations. The statistical analysis of the results will be used for carrying out a health and economic impact assessment. The computation of the part of the PM_{10} pollution specifically attributable to traffic will be considered in future work.

Even though the use of auxiliary variables like the NO_2 data may lead to more realistic results, the geostatistical approach highly depends on the availability and spatial distribution of PM_{10} measurements. When this information is scarce, the approach would really benefit from the knowledge of the physico-chemical process of the pollution. Such information could be obtained from detailed analysis of the emissions and transformation process, through a classical numerical simulation of air pollutant transport. This simulation output could be incorporated in the geostatistical framework as an accurate cofactor through a collocated cokriging or kriging with external drift approach (Blond 2002). The final model would then present the advantage of integrating the actual data from the air monitoring network and the best knowledge on the pollution phenomenon.

Acknowledgements

The financial support of the French agency ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie) through contracts n° 03 62 C0023 and n° 03 62 C0053 is gratefully acknowledged.

References

ADEME (2002) Classification and Criteria for Setting Up Air-Quality Monitoring Stations. ADEME Editions, Paris, p. 63

APHEIS (2000) Air pollution and Health: a European Information System. Monitoring the Effects of Air Pollution on Public Health in Europe. Scientific report 1999-2000. DG SANCO G/2"Polution related diseases" program; p. 135

Blond N (2002) Assimilation de données photochimiques et prévision de la pollution troposphérique. Thèse de doctorat de l'Ecole Polytechnique, Palaiseau, p. 204

Chilès JP, Delfiner P (1999) Geostatistics: modelling spatial uncertainty, Wiley Series in Probability and Mathematical Statistics, p. 695

Cressie N, Kaiser MS, Daniels MJ, Aldworth J, Lee J, Lahiri SN, Cox LH (1998) Spatial Analysis of particulate matter in an urban environment. In: Second European Conference on Geostatistics for Environmental Application, (eds Gomez-Hernandez JJ, Soares A, Froidevaux R), Kluwer Academic Publishers, 41-51

Deraisme J, Jaquet O, Jeannée N (2002) Uncertainty management for environmental risk assessment using geostatistical simulations. In: Fourth European Conference on Geostatistics for Environmental Application, (eds Sanchez-Villa X, Carrera J, Gomez-Hernandez JJ), Kluwer Academic Publishers, 139-150

Geovariances (2004) Isatis Software Manual, 5th Edition, Geovariances & Ecole des Mines de Paris, p. 710

Isaaks EH, Srivastava RM (1989) An introduction to Applied Geostatistics, Oxford University Press, p. 561

Lajaunie C (1993) L'estimation géostatistique non linéaire. Cours C-152, Centre de Géostatistique, Ecole des Mines de Paris

Lajaunie C, Béhaxétéguy JP (1989) Elaboration d'un programme d'ajustement semi-automatique d'un modèle de corégionalisation – Théorie. Technical report N21/89/G. ENSMP Paris, p. 6

Lantuéjoul C (2002) Geostatistical Simulation - Models and Algorithms. Springer-Verlag, p. 256

Matheron G (1973) The intrinsic random functions and their applications. Advances in Applied Probability, 5, 439-468

Mosqueron L, Nedellec V, Desqueyroux H, Annesi-Maesano I, Le Moullec Y, Medina S (2003) PEP project "Transport-related health impacts and their costs and benefits with a particular focus on children"; The Hague Workshop Input Reports; State of the Art - Review of exposures and health effects from Epidemiological studies Focused on Children, p. 53

Rivoirard J (1994) Introduction to disjunctive kriging and non-linear geostatistics. Oxford University Press, Oxford, p. 181

Roth C, Bournel-Bosson C (2001) Mapping diffusive sampling results: including uncertainty and indirect information. International conference Measuring Air Pollutants by Diffusive Sampling, Montpellier, France, Sept. 26-28 2001

Air quality models resulting from multi-source emissions

A. Russo [1], C. Nunes [1, 2] and A. Bio [1]

[1] Environmental Group of the Centre for Modelling Petroleum Reservoirs. CMRP-IST, Avenida Rovisco Pais, 1049-001 Lisbon, Portugal. arusso@ist.utl.pt
[2] Universidade de Évora, Portugal

1 Introduction

Air quality is normally characterized using different indicators, generally expressed by the concentration of a certain pollutant for a determined time period. The most frequently used indicators are: sulphur dioxide (SO_2), nitric oxides (NO_x), carbon monoxide (CO) and total suspended particles (TSP).

Air pollution systems integrate three main components: emission source, transport medium (atmosphere) and receptor. Pollution reaching a receptor depends not only on the emitted quantity, but also on atmospheric dynamics (Seinfeld 1986, de Nevers 2000). The impact on the receptor can be estimated by developing source-receptor linkages through the atmosphere. In some cases its transport may occur over great distances until it reaches ground level, reason for which these substances are also object of agreements and international conventions.

It is well known that air pollutants at ground level can be harmful to human health, if their concentrations exceed certain limits (de Nevers 2000). As pollutants accumulate in, or near, large metropolitan areas, populations are typically more exposed to unhealthy pollutant concentrations (Seinfeld 1986, Cobourn *et al*. 2000, Kolehmainen *et al*. 2000).

Considering the effect that air pollutants' concentrations have on human health, a study that allows the identification of regional emission-receptions patterns for some pollutants and the quantification of the contribution of local industrial units, is of great interest for the health system, the environment, the economy and also to local management (Cobourn *et al*. 2000, Kolehmainen *et al*. 2000)

Even so, in order to develop robust predictive air quality (AQ) models, wide-range monitoring systems are necessary. Modelling therefore often needs to be used in conjunction with other objective assessment techniques, including monitoring, emission measurement and inventories, interpolation and mapping (WHO 1999). However, obtaining suitable and representative AQ samples can be quite difficult.

A predictive model of the different emissions' contribution to the pollutant concentrations captured at each monitoring station, will allow an analysis of the im-

pact caused in the monitoring station's area and its translation into an air quality index.

The purpose of this study is to analyse possible relations between sulphur dioxide (SO_2) emissions from industrial complexes located in the Sines area (Portugal), and air quality data colleted at monitoring stations, by means of linear and non-linear modelling.

2 Objectives

The objective of this study consists in developing and implementing a methodology that allows classifying the contribution of different emission sources to air quality (AQ) in the region of Sines (Fig. 1). This methodology is based on the use of artificial neural networks (ANN's), in order to identify non-linear relations between meteorological parameters, emissions and air quality data measured at monitoring stations. Within the scope of this work, the following tasks were performed:

- Identification of regional emission-reception patterns for some pollutants;
- Quantification of the individual contribution of local industrial units;
- Development of an instrument that allows, simultaneously, to evaluate temporal forecasts of AQ parameters and, also, to simulate situations of extreme atmospheric pollution.

We present a case study where possible relations between sulphur dioxide (SO_2) emissions, generated by three industrial complexes (Petrogal, Borelis and CPPE) located in the Sines area (Portugal), and air quality data colleted at four air quality monitoring stations (Santiago do Cacém, Sonega, Monte Chãos, Monte Velho), are analysed by means of linear and non-linear modelling, as described in the section 4.

Typically, the distributions of daily data of the emissions of the three industrial complexes and air quality data colleted at the four AQ monitoring stations have large values of skew. Thus, it is natural to consider that the physical processes related to these distributions are of a non-linear nature.

Presently, ANN's constitute the best technique (as flexible mathematical structure) that is able to identify complex non-linear relations between inputs and outputs, without previous integral understanding of the phenomenon's nature.

In recent years there has been a tendency to use more statistical methods instead of traditional deterministic modeling (Kolehmainen *et al*. 2000). A number of linear methods have been applied to time-series for air pollutants (Simpson and Layton 1983, Ziomas *et al*. 1995, Shi and Harrison 1997), including comparisons with neural network methods (Yi and Prybutok 1996, Comrie 1997, Gardner and Dorling 1999, Cobourn *et al*. 2000, Kolehmainen *et al*. 2000). Gardner and Dorling (1998) concluded, in their overview of applications of ANN's to the atmospheric sciences that ANN's generally provide as good or better results than linear methods.

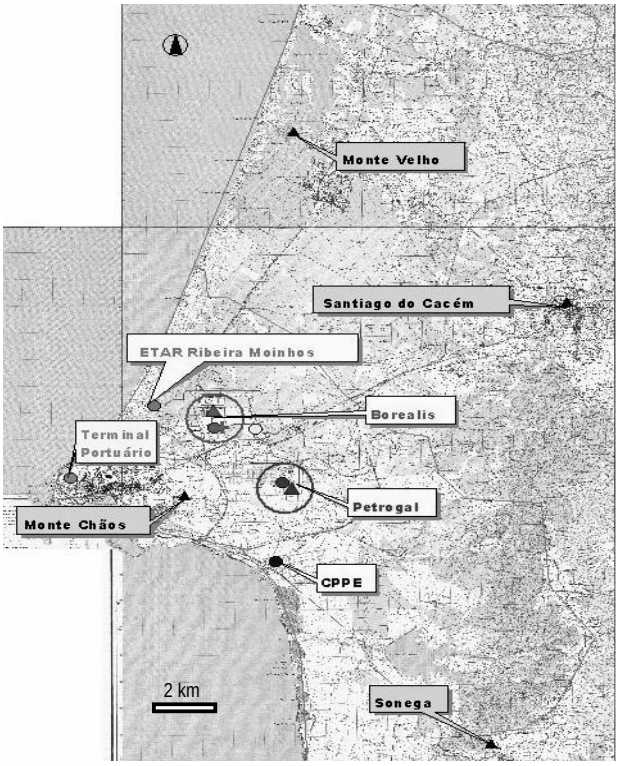


Fig. 1. An overview of the Sines Peninsula (Petrogal, Borealis and CPPE industrial complexes in light gray; AQ monitoring stations in Santiago do Cacém, Sonega, Monte Chãos, Monte Velho in dark gray).

3 Data

Sulphur dioxide emissions (mg/m3) from three industrial complexes – Petrogal, Borealis and CPPE – are periodically measured in a set of monitoring stations – Santiago do Cacém, Sonega, Monte Chãos, Monte Velho – and converted to daily averages for a period of 12 months (from 1/1/2002 to 31/12/2002) (Fig. 2 and 3). Meteorological data – wind speed and direction on an hourly basis, for the same period – was also collected and analysed (Fig. 4).

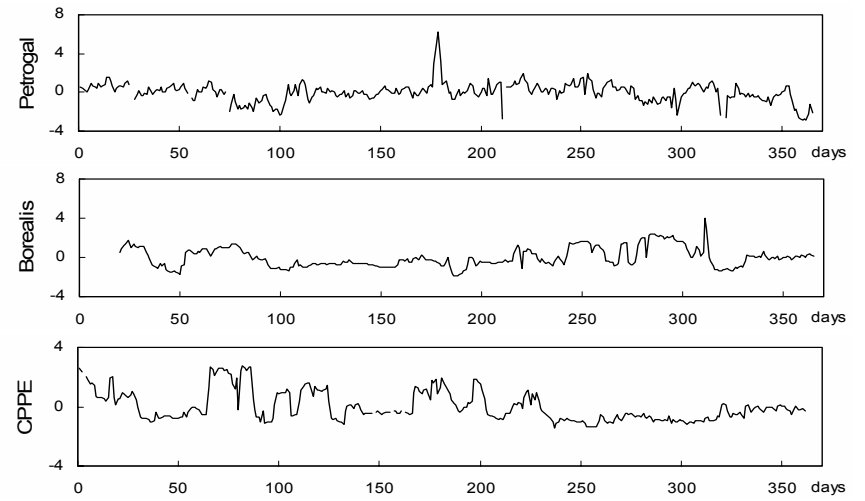


Fig. 2. SO_2 emitted by the three industrial complexes, standardized to zero mean and unit standard deviance.

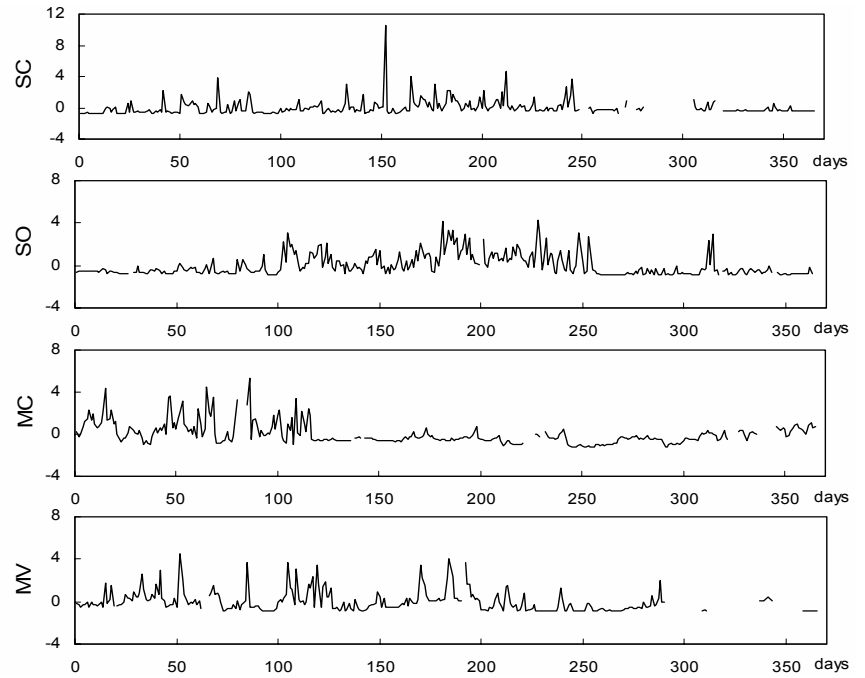


Fig. 3. SO2 measured by the monitoring stations. MC - Monte Chãos, MV - Monte Velho, SC - Santiago do Cacém, SO - Sonega, standardized to zero mean and unit standard deviation.

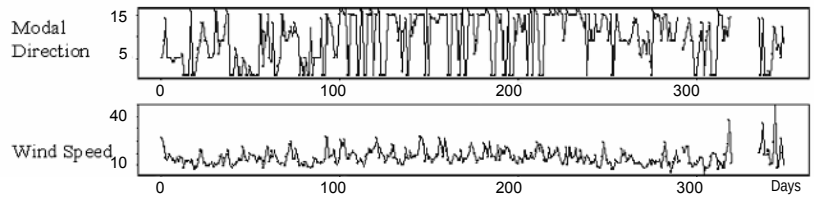


Fig. 4. Wind speed and of the modal wind direction registered, standardized to zero mean and unit standard deviance.

The available data was standardized in order to minimize the effect of different local means and variances in the evaluation of the emissions/AQ measurements relationships. Afterwards, the days, which did not have any register of data in at least one of the emission-reception stations, were annulled. Thus, only the period common to all the available data sets or to each pair emission-monitoring station was used (365 days - *N* error values).

4 Methodology

A two steps methodology was approached for this study. First, the time series of each data pair – industrial emission and monitoring station records – was filtered out, in order to obtain contiguous time periods with high correlation of that specific industrial emission with the equivalent monitoring-station measurements. For this purpose, an iterative optimisation process was developed, using the correlogram between industrial emissions standardized data and monitoring station time series as the objective function (*e.g.* Fig. 5). After this classification of "historical" data, the process becomes automatic for any future pair of data (same industrial emission source and monitoring station) through the use of a probabilistic neural network (PNN). The PNN automatically classifies the time series into two classes:

– Class 1: Pairs of highly correlated points;
– Class 2: Pairs of points without correlation.

In a second step, artificial neural networks (ANN's) were applied in order to identify non-linear relations between the SO_2 emitted by the industrial complexes, AQ parameters measured at the monitoring stations and meteorological information.

4.1 Classification of time periods with high correlation between emission and monitoring station records

After the first attempts of including meteorological variables into prediction models, we concluded that the available data of wind speed and direction wasn't responsible for the dynamics of the different pollutant plumes. The main reason is

that the meteorological data was often collected at an altitude and locations inadequate to capture emissions from the different industries' chimneys.

Hence, in a first step, the time series of each data pair – industrial emissions and monitoring stations records – was filtered with the purpose of obtaining contiguous time periods with high correlation between that specific industrial emission and the equivalent monitoring station measurements.

Basically, a simple iterative procedure was implemented. The variogram of each pair emissions-monitoring station AQ measurements along a period T (365 days - *N* error values):

$$\gamma(z_1, \Psi_1) = \frac{1}{2T} \sum_{i=1}^{T} \left[z_1(i) - \Psi_1(i) \right]^2, \tag{1}$$

where $z_1(i)$ and $\Psi_1(i)$ are the measurements of the emission source z_1 and of the monitoring station Ψ_1 for the instant i after standardization, was assumed as an objective function that tends to decrease (increasing the correlation between z_1 and Ψ_1) as, iteratively, pairs of points with less contribution are removed.

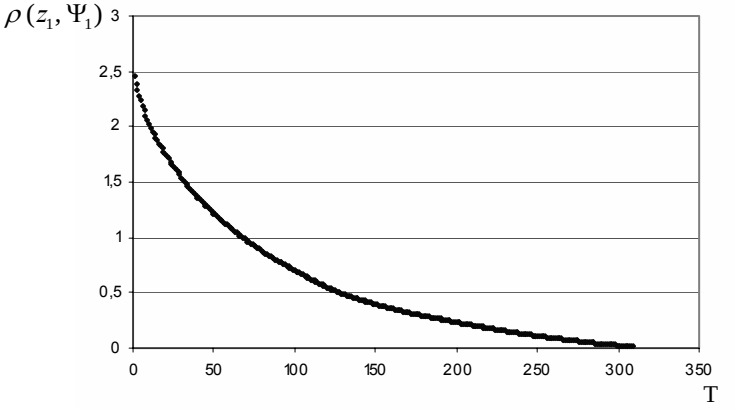


Fig. 5. Borealis and Sonega class 1 correlogram.

A probabilistic neural network (PNN) was used for an automatic data classification into the two classes described above. PNN's can be useful for classification problems and have a straightforward design. A PNN is guaranteed to converge to a Bayesian classifier providing it is given enough training data, and these networks generalize well. PNN's have many advantages, but they suffer from one major disadvantage. They are slower to operate because they use more computation than other kinds of networks to do their function approximation or classification (Haykin 1994, Beale and Demuth 1998).

4.2 Prediction Model Formulation

Neural networks are composed by a number of interconnected entities, similar in many ways to biological neurons – the artificial neurons. These artificial neurons may be associated in many different ways – the network architecture. The network function is determined largely by the connections between neurons. An ANN may be trained to perform a particular function by adjusting the values of the connections (weights) between neurons (Haykin 1994, Beale *et al*. 1996, Gurney 1997, Beale and Demuth 1998).

Commonly, neural networks are adjusted, or trained, so that a particular input leads to a specific target output. The network is adjusted, based on a comparison of the output and the target, until the network output matches the target. The choice of the network's architecture depends on the task to be performed (Haykin 1994, Sarle 1994, Beale *et al*. 1996, Gurney 1997, Beale and Demuth 1998, Dorling and Gardner 1998). To model a physical system such as an air pollution system, a feed-forward layer is normally employed (Wasserman 1989). It consists of layers of input neurons, and one or more hidden layers. For this study the MATLAB's neural networks toolbox was used (Beale and Demuth 1998).

Considering the construction of a neural network model, an air pollution system is looked upon as a system that is under various sets of inputs (*e.g.* weather parameters, AQ parameters), and will respond by producing different sets of outputs (*e.g.* pollutant concentrations). Such a model assumes no prior knowledge about the structure of the relationship that exists between input and output variables. The neural network model is trained and tested using the AQ data.

After the classification process, the sets of industrial emission-monitoring station pairs of points belonging to Class 1, were processed by a multiple layers neural network with feed-forward propagation (feed-forward multi-layer perceptron) trained by a back-propagation algorithm. Considering that the available data sets were limited, a simple random validation was used. Some of the models' details are addressed below:

- Weights randomly initialised;
- One hidden layer;
- Number of hidden neurons limited to a minimum of two (for the non-linear model);
- Number of hidden neurons tested through trial and error (non-linear model);
- Number of epochs to train limited to 100 (non-linear model);
- Sum-squared error goal set to 0,5 (non-linear model);
- Learning rate set to 0,001 (non-linear model);
- Activation functions:
- Linear model: linear
- Non-linear model: log-sigmoid and linear.

The linear and non-linear models were integrated. On the non-linear model, the number of neurons (s) in the hidden layer (s={*1, 2, 3, 4, 5, etc*.}) was systematically altered.

5 Results and Discussion

With the purpose of obtaining contiguous time periods with high correlation between each pair of industrial emission and monitoring station measurements, the time series of each data pair was filtered (*c.f.* Section 4.1). Scatters plots of the standardized values of the data series, before and after being filtered, are presented in Fig. 6 and 7, for the case of Borealis' emissions and Sonega's monitoring station. Fig. 8 shows the respective time series after filtering.

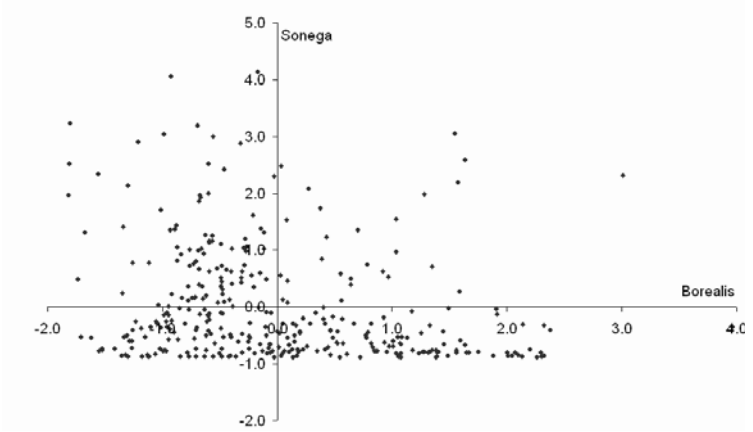


Fig. 6. Borealis (x-axis) and Sonega's (y-axis) SO_2 concentrations before being filtered.

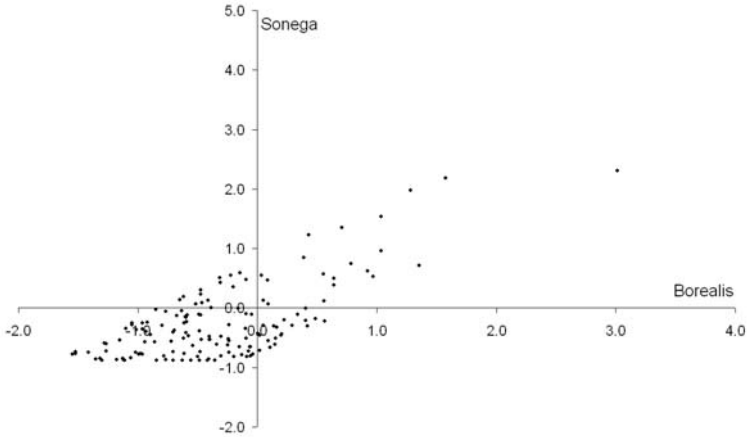


Fig.7. Borealis (x-axis) and Sonega's (y-axis) class 1 data points.

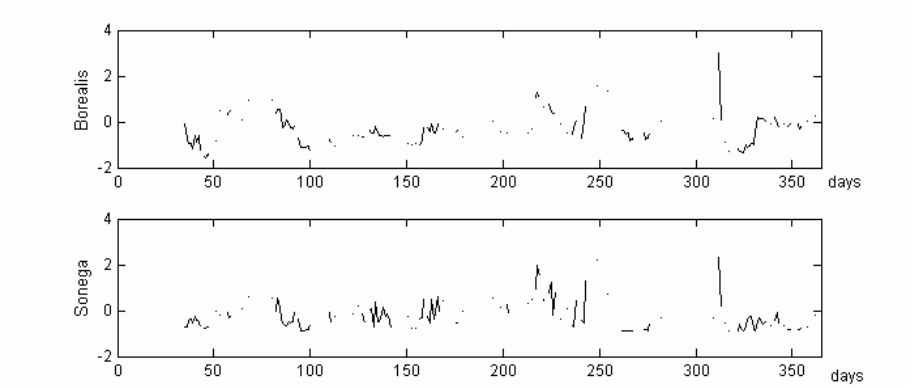


Fig. 8. Borealis and Sonega's Class 1 data points.

After the classification process, the sets of industrial emission-monitoring station pairs of points belonging to Class 1 were processed by a feed-forward multilayer perceptron (*c.f.* Section 4.2).

Correlation coefficients between observed and modelled AQ values attained with the linear and non-linear models, for the same day (t=0) and for the day after (t=1), are represented in table 1.

Table 1: Correlation coefficient between observed and modeled AQ values attained with the linear and non-linear models for the same day (t=0) and for the following day (t=1).

		Correlation coefficient (%) (t = 0)		Correlation coefficient (%) (t = 1)	
		Linear Model	Non-Linear Model	Linear Model	Non-Linear Model
MC	Borealis	56	63 (9)	61	63 (2)
MC	Petrogal	72	73 (3-7)	58	66 (2)
MC	CPPE	78	80 (4)	75	76 (7)
MV	Borealis	59	62 (2)	64	66 (2)
MV	Petrogal	59	63 (2)	50	57 (2)
MV	CPPE	65	57 (1)	76	77 (2)
SC	Borealis	62	70 (2)	64	70 (2-4)
SC	Petrogal	70	75 (2)	61	63 (3)
SC	CPPE	67	75 (2 and 7)	64	62 (2)
SO	Borealis	67	69 (2)	72	82 (2)
SO	Petrogal	72	73 (2-3)	73	74 (2)
SO	CPPE	61	67 (2-3)	72	74 (2)

Note: The number(s) between parentheses, on the non-linear models' columns, refer to the number of hidden neurons.

6 Discussion

From the analysis of table 1 we may conclude that:
- The best AQ-emission correlations for Monte Velho and Monte Chãos, are obtained with the CPPE's industrial complex. For Monte Velho, the best correlations attained for the same day, result from the integration of the linear model.
- The best correlations for Sonega, for the same day are obtained with Petrogal. The best correlations for Sonega and for the following day are reached with Borealis, and result from the integration of the non-linear model.
- For Santiago do Cacém, the best correlations for the same day are attained with Petrogal and for the following day with Borealis, by the non-linear model.

The developed neural network models establish a reasonable relationship between the values emitted by the tree industrial units and the values measured in the AQ monitoring stations.

7 Conclusions

The models developed present satisfactory correlations between pollutant values emitted by the tree industrial units (Petrogal, Borealis and CPPE) and those values measured at the AQ monitoring stations (Monte Chãos, Monte Velho, Santiago do Cacém and Sonega).

Models' performance could improve using longer AQ data series, and other types of meteorological data series.

Note that this study merely aims to assess possible relationships between pairs of emission sources and air quality monitoring stations records. However, the results suggest that, more robust prediction models can be developed.

Finally, we find it appropriate to point out that neural nets are far from being the solution to all statistical modelling problems. The use of such complex models should be considered with some caution. In particular, the possibility of adjusting a high number of coefficients in the non-linear ANN's may lead to apparent yet spurious improvements in the results.

8 Further Work

The methodology developed and applied to SO2 will be used for NOx measurements, from the same period (2002). Subsequently, the same methodology will be applied for the same variables but for the year 2003.

Based on the results attained with the calibration and validation models, we intend, later on, to develop an operational air-quality assessment tool – time prediction models.

These time prediction models may in future be improved coupling spatial information for the same pollutants captured by passive monitors (diffusive tubes).

References

Beale MH, Demuth HB (1998) Neural Network Toolbox for Use with MATLAB, User's Guide, version 3. The MathWorks, Inc.

Beale MH, Demuth HB, Hagan MT (1996) Neural Networks design. PWS Publishing Company, Boston

Cobourn WG, Dolcine L, French M and Hubbard MC (2000) Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting. J. Air & Waste Manage. Assoc., 50, 1999-2009

Comrie AC (1997) Comparing neural networks and regression models for ozone forecasting. Journal of Air and Waste Management Association, 47, 653-663

De Nevers N (2000) Air Pollution Control Engineering, 2nd edn. McGraw-Hill

Dorling SR, Gardner MW (1998) Artificial Neural Networks (the Multi-layer Perceptron) - A review of applications in the atmospheric sciences. Atmospheric Environment, 32, 2627-2636

Dorling SR, Gardner MW (1999) Neural network modelling and prediction of hourly Nox and NO2 concentrations in urban air in London. Atmospheric Environment 33 709-719

Gurney K (1997) An Introduction to Neural Networks. UCL press, London

Haykin S (1994) Neural Networks: A Comprehensive Foundation. Macmillan Pub, New York

Kolehmainen M, Martikainen H and Ruuskanen J (2000) Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35, 815-825

Sarle W (1994) Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference, Cary, NC: SAS institute

Seinfeld JH (1986) Atmospheric Chemistry and Physics of Air Pollution. John Wiley & Sons

Shi JP and Harrison RM (1997) Regression modelling of hourly NO*x* and NO2 concentrations in urban air in London. Atmospheric Environment, 31, 4081-4094

Simpson RW and Layton AP (1983) Forecasting peak ozone levels. Atmospheric Environment, 17, 1649-1654

Wassermann PD (1989) Neural Computing theory and Practice. New York Van Nostrand Reinhold

Ziomas I, Melas D, Zerefos CS, Bais AF and Paliatsos AG (1995) Forecasting peak pollutant levels from meteorological variables. Atmospheric Environment, 29, 3703-3711

WHO (1999) Ambient Air Quality Monitoring and Assessment - Guidelines for Air Quality. World Health Organization, Geneva

Yi J and Prybutok VR (1996) A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution, 92, 349-357

Variogram estimation with noisy data in the space-time domain: application to air quality modelling

C. Nunes[1,2] and A. Soares[2]

[1] Math Department, Évora University, Portugal
[2] Environmental Group of the Centre for Modelling Petroleum Reservoirs, CMRP/IST, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, e-mail: carlanunes@mail.ist.utl.pt

1 Introduction

The main issue addressed in this paper is related to the robust estimation of variograms, in space and time, representative of a given period and region. Several studies have addressed this issue: robust variograms, in the sense of less sensitivity to very high and low values (Armstrong 1984, Chauvet 1982, Cressie *et al.* 1980, Cressie 1984); variograms estimation of non stationary phenomenon of space-time lattices (Switzer 1989, Sampson *et al.* 1992, Perrin *et al.* 1999); and robust measures of data affected by heteroscedasticity and clustering (Srivastava *et al.* 1989).

When dealing with space-time data, e.g. a set of monitoring stations measuring the water quality of a river or the air quality of a region, some problems arise in modelling the space-time pattern of the main pollutants by a robust and representative variogram. Measurement errors (Wiersma 2004) or different factors conditioning the dynamics of different systems (water or wind) can substantially affect the deposition and mask the main space-time patterns of the pollutant in a given period.

The objective of this study is to propose two simple approaches to filter the space-time data, removing the effects of assumed measurement errors and turbulent influences of external factors.

The first approach, assumes that the data have some (few) measurement errors that significantly affect variogram values for short distances. The objective is to identify those measurement errors, in order to avoid them in the variogram computation. An optimisation technique (simulated annealing) is used to remove pairs of points using as criterion the assumption that the variograms have a linear behaviour near the origin. Obviously this approach is valid only if a small proportion of points is removed, to obtain the average behaviour. Here, the objective is to capture the major and representative space-time features, and not to drive data to a given continuity model.

When dealing with the noisy influence of secondary variables that condition the spatiotemporal patterns, a second methodology is proposed. Here, the problem results from the fact that the air pollution dispersion is dependent on wind speed, wind direction, atmospheric stability classes or turbulent conditions. Spatiotemporal dependencies can exist during some periods within some range of wind speed, and can, for example, disappear for high wind speeds or severe turbulence conditions. Hence a classification of short time periods with anomalous spatial continuities is performed. For a new set of data corresponding to a new period (which was not taken into account in the variogram estimation) each time can be classified in terms of its probability to belong to an anomalous period of time.

2 Methodologies

Assuming a value $Z(x_i,t_j)$ of variable Z, measured at monitoring station x_i for time t_j, this value can be correlated with the concentrations measured in previous time periods at the same monitoring station, and with concentrations measured at neighbouring monitoring stations during the same or previous time periods.

The spatial continuity for a given period of time can be characterized using a mean spatial variogram, $\gamma_s(h)$, computed by averaging the spatial variograms of each time t slice and representing the mean spatial pattern for that given period of time:

$$\gamma_S(h) = \frac{1}{2NtNh} \sum_{j=1}^{Nt} \sum_{i=1}^{Nh} \left[Z(x_i, t_j) - Z(x_{i+h}, t_j) \right]^2 \quad (1)$$

where Nt is the number of time periods and Nh the number of pairs of monitoring stations at distance h from each other.

A linear behaviour of the variogram, near the origin, is assumed in the presence of spatial and temporal dependencies. Both methodologies here proposed are developed based on this notion. The Pearson correlation coefficient was used as criterion to evaluate these linear behaviours. The concept of "near the origin" is theoretically vague, and the selected number of points depends on each case study.

2.1 Optimisation technique

This approach was developed under the assumption that only few data have measurement errors concealing the existing spatiotemporal patterns, and aims at identifying these data in order to remove them from the variogram calculation. Two basic principles underlie the proposed methodology: the data to be removed must be validated (*a posteriori*); the number of data removed should be minimal.

For this purpose, an optimisation technique is used to remove pairs of points, in order to minimize the square differences from a linear behaviour of the variogram near the origin; being the approach based on the maximization of the corresponding correlation coefficient.

The idea of this process is to choose the values that affect most negatively the correlation coefficient of a linear behaviour of the variogram near the origin, to understand and clarify the mean spatiotemporal pattern. The values $Z(x_i,t_j)$, at a given spatial location x_i at a time t_j, are iteratively removed until a pre-defined coefficient of correlation is achieved or until the maximum number of points allowed is removed. This maximum number must be an insignificant proportion of the data to preserve their specific spatiotemporal patterns.

After defining the threshold for the correlation coefficient ρ (i.e. the objective function value considered satisfactory) or the maximum number of points allowed to be removed, and defining the "near the origin" range, the process can be summarized as follows:

1. Iteratively remove a value $Z(x_i,t_j)$, for a given i and j.
2. Estimate the spatiotemporal variogram $\gamma_{ij}(h)$, according to Eq. 1, after $Z(x_i,t_j)$ has been removed.
3. Compute the corresponding correlation coefficient ρ_{ij}, between $\gamma_{ij}(h)$ and h. This coefficient represent the linear behaviour fitting quality of the spatiotemporal variograms near the origin, when $Z(x_i,t_j)$ is not being considered.
4. Return to step 1 and repeat for all $Z(x_i,t_j)$, to identify the most influent point in the convergence of the objective function. The value $Z(x_i,t_j)$, corresponding to i and j, that results in the maximum correlation coefficient when omitted, is definitively removed.
5. The process continues until the pre-defined objective function limit is reached.

2.2 Classification of short time periods

The second approach has different goals and different applications. It deals with situations where a considerable proportion of data must be removed to obtain a good spatiotemporal variogram; situations that the first methodology is not appropriate to deal with.

In those cases the problem is probably not the existence of measurement errors but, for instance, the influence of external variables that are conditioning the spatiotemporal patterns. For instance, air pollution dispersion can be dependent on wind speed, wind directions, or atmospheric stability classes.

An alternative methodology was developed to characterize the spatial continuity in these situations. The process is similar to the first one, but instead of removing one anomalous value, an entire set of values of one period, which is assumed to be disturbed by the meteorological conditions of that period, is removed. The final goal is to obtain contiguous periods of evident space-time patterns and distinct periods with poor correlation between monitoring stations.

The classification of time periods with different space-time patterns is identical to the previous.

Note 1: Both methods can be generalized to achieve more robust space-time variograms, especially when time delays are verified between monitoring stations. In these cases the variogram in equation (1) is replaced by the space-time variogram:

$$\gamma(h,\delta) = \frac{1}{2NhNt} \sum_{j=1}^{Nt} \sum_{i=1}^{Nh} \left[Z\left(x_i, t_j\right) - Z\left(x_{i+h}, t_{j+\delta}\right) \right]^2 \quad (2)$$

where *Nt* is the number of time periods and *Nh* the number of pairs of monitoring stations at distance *h* from each other.

The methodological sequences are identical to 2.1 and 2.2.

Note 2: In those cases where measurement errors are assumed to disturb the average model and the first method (2.1) is applied to remove a few points to obtain a robust and representative variogram, the space time dispersion can be characterized for past or simulated for future scenarios with patterns identical to those of the recent past (Nunes *et al.* 2004).

When different periods are classified using the second method – periods with evident space-time patterns and periods without any special or temporal structure – it is necessary to identify which class each period belongs to, before applying the space-time variogram model. Russo *et al.* (2004) suggest the use of a Probabilistic Neural Network to identify those periods prior to any estimation or simulation.

3 Applications

3.1 Air quality of Setúbal Peninsula case study

The proposed methodologies are illustrated with an air quality assessment case study. This case study aims at characterizing air quality in the Setúbal Peninsula (Portugal). Particulate emissions from three main non-diffuse sources – a cement factory, a power plant and a pulp mill – are periodically measured in a set of monitoring stations, on a daily average basis (Fig. 1).

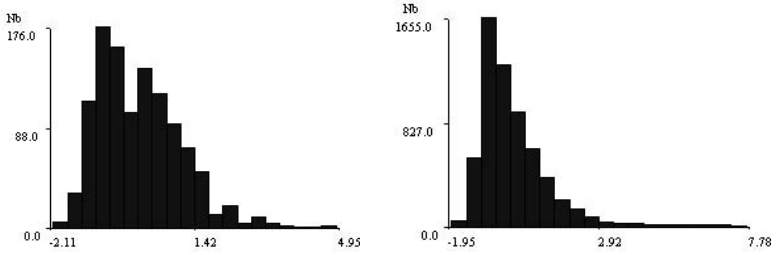
The first approach (2.1) was applied to the data measured in monitoring stations during a relatively homogeneous period of 6 months (from 1/2/1997 to 31/7/1997).

To ensure the presence of different meteorological conditions for the application of the second methodology (2.2), daily data of particulate concentrations measured during 3 years (from 1996 to 1998) was used.

Fig. 1. View of part of the Setúbal peninsula with monitoring stations (○) and pollutant sources (✧)

Given the varying local means and variances, the original variables were, in both cases, standardised by their local means and standard deviations. These statistical measures are very sensitive to extreme values. Note that, noisy data are not necessarily extreme values but values that are not in agreement with the average spatiotemporal patterns. Therefore standardisation before filtering, to treat local heterogeneities, is a valid approach.

Global statistics of the transformed experimental data, for the referred periods of time, are shown in Fig. 2.

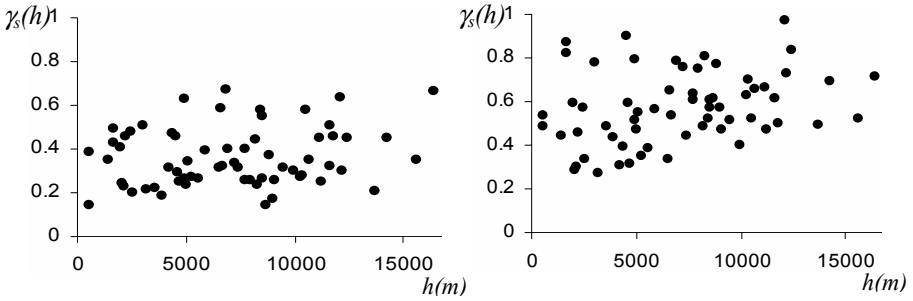


a) 6 Months b) 3 Years

	Mean	Variance	Min	Max	Skewness	P95
6 Months	0	0.99	-2.11	4.95	0.71	1.67
3 Years	0	1	-1.95	7.78	1.62	1.89

Fig. 2. Histograms and descriptive statistics of the transformed datasets

Mean spatial variograms (1) of the transformed experimental datasets are shown in Fig. 3.



a) 6 Months b) 3 Years

Fig. 3. Mean spatial variograms of standardized datasets, for the given periods of times

Both of the mean spatial variograms do not reveal any spatial pattern; it seems that the particulate concentrations do not display any mean spatial structure, during the given periods of times.

To identify potential measurement errors, the first methodology was applied to the 6-months dataset. In the 3-years dataset the existence of some short time periods, where the global space-time structures are disturbed by external variables (e.g. adverse meteorological conditions), is expected. To identify and classify those periods of times, the second methodology was applied.

3.2 Estimation of robust variogram: method 1

The methodology presented in section 2.1, was applied to the 6 months standardized data. The dimension of the aureole, inside which the optimisation algorithm is applied, depends on the monitoring network, the spatial distribution of the pairs of points etc.. Several ranges were essayed.

Fig 4 shows the application of best linear fit "near the origin" inside a 5000 meters range aureole.

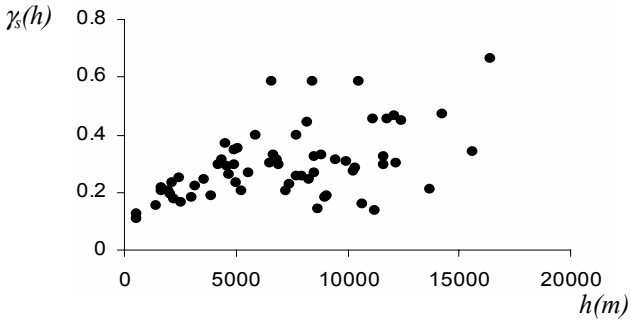


Fig. 4. Robust mean spatial variogram, with a 5000 meters range aureole, for the given period of time.

After removing 2% of the data (22 out of 1087), the mean spatial variogram shows a clear mean spatial pattern, near de origin, allowing the use of geostatistical techniques in future phases of this work (estimation or simulation). It became possible to identify the monitoring station(s) responsible for the erroneous measures, and to understand the main causes for those problems.

In Fig. 5 an identical test, but with a 10000 m range aureole, is shown.

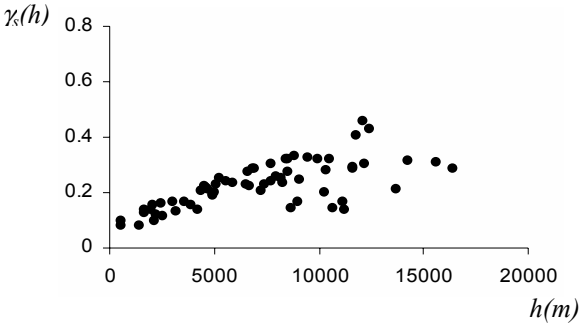


Fig. 5. Robust mean spatial variogram, with a 10 000 meters range aureole, for the given period of time.

The mean spatial variogram reaches an identical structure to that in Fig. 4, after removing 4,5% of the data (50 out of 1087). Both have 10 % of nugget effect and similar ranges. The slope in the Fig. 5 is lower than that of Fig. 4. In general, we observed that increasing the "near the origin" range beyond 5000 meters does not change significantly the variogram behaviour. The small-scale variability and/or some remaining measurement errors can be responsible for the 10% nugget effect.

To test this methodology, a dataset with a uniform distribution and totally random spatial and temporal structures, was simulated. In Fig. 6, the initial mean spatial variogram (a) and the obtained mean spatial variogram (b), after filtering, are shown.

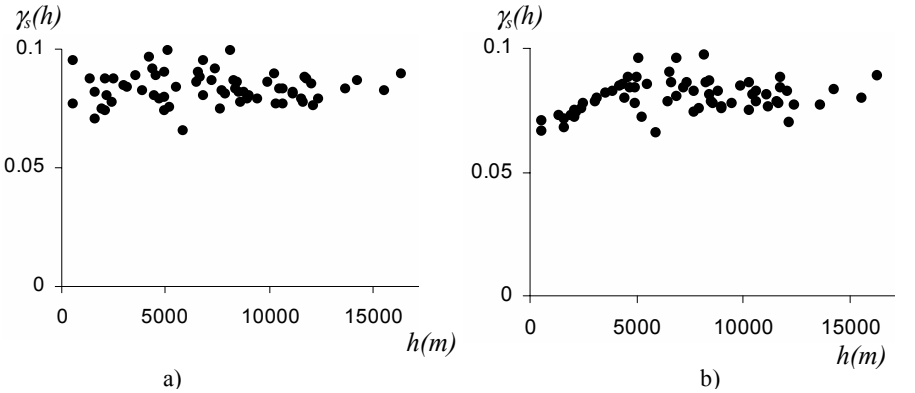


Fig. 6. a) Initial mean spatial variogram and b) final filtered mean spatial variogram (5 000 meters range aureole), using a random test dataset.

The initial variogram (Fig. 6a) and filtered variogram (Fig. 6b) do not reveal any spatial structures. After removing 5 % of data using a 5000 near the origin range, a linear behaviour was reached (0.9 as correlation coefficient) but with 81% of the sill as nugget effect, proving that if there are no structures in the model, this methodology is useless, and can only capture spatial structures if they are really existing although masked by few measurements. This exercise demonstrates that the methodology does not "create" any spurious structure, by picking the best fitting data for a linear near-the-origin behaviour.

3.3 Estimation of robust variogram: method 2

The methodology presented in section 2.2 was applied to the 3 years standardized dataset, using identical parameters to those in section 3.2: 5000 and 10000 meters "near the origin" ranges. The detailed evolutions of the correlation coefficients as functions of the number of time slices (days) removed are shown in Fig. 7.

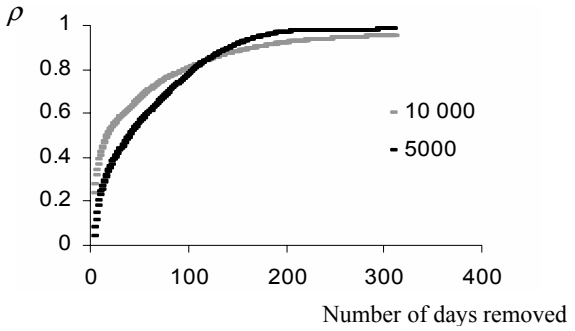


Fig. 7. Evolution of the correlation coefficients, for 5000 and 10000 meters "near the origin" ranges.

In both cases, after removing 308 days, 28% of the total data set used in the computation of the mean spatial variogram (Fig. 7), the correlation coefficients reach approximately the value 0.98. A 0.8 correlation coefficient can be obtained removing 100 data (9% of the total), and after 200 omissions the correlation coefficient function becomes practically constant at about 0.97. For the remaining 72% of daily measurements it is possible to fit a mean spatial variogram. The mean spatial variograms, using a 5000 and a 10 000 meters range aureoles, are shown in Fig. 8.

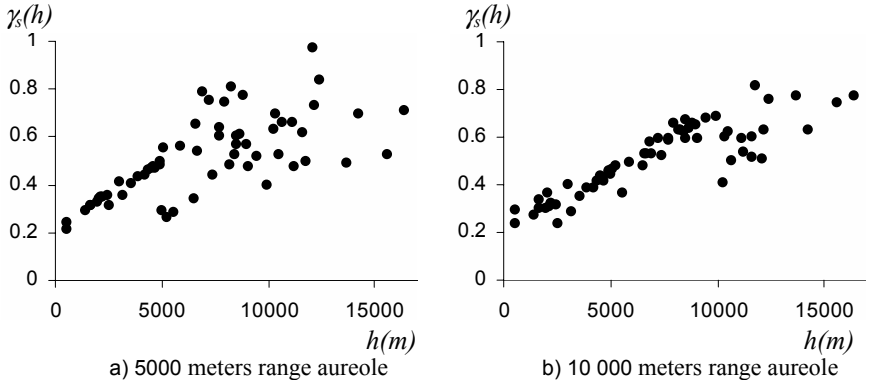


Fig. 8. Mean spatial variogram omitting 308 daily measures

The second variogram reveals a more clear spatial pattern. Using the second method it seems that the dimension of the range aureole has a more evident effect on the variogram structures. For the selected periods of time, the variograms, revealing a spatial pattern, can be used in future estimation or simulation applications. Considering, for example, a new set of data corresponding to a new period,

each time period can be classified, in terms of the probability of belonging to an anomalous period of time (for instance using neural networks, Russo *et al.* 2004).

In a posterior analysis, no direct relation between the identified periods of time and the available meteorological information (wind speed, wind direction, precipitation, temperature, humidity) was found. Note that the available meteorological information is measure only in one location, and has to be considered as representative for the entire area; a very questionable assumption considering this case study and the region's irregular topography.

4 Final remarks

The presented methodologies constitute simple algorithms to identify (and (possibly remove) erroneous data or anomalous time periods, for the identification of critical situations (measurement errors or the influence of external factors) and for the estimation of robust and representative spatial variograms for space-time data.

The approach is based on the assumption that external factors (measurement errors, turbulent meteorological conditions) affect data at specific locations or an entire set of data from a given period. Instead of analysing all factors (most of which unknown or unmeasured) that could possibly interfere with the data measurements, we propose to filter the data according to a given optimisation criterion.

Both methodologies here proposed are developed under the assumption that a variogram has a linear behaviour near the origin in the presence of spatial and temporal dependencies. Note that these methodologies can be easily adapted to others models (for instance, linear, exponential, Gaussian, or other prior variogram model), incorporating their respective parameters in the algorithms and using a correlation coefficient to evaluate the near-the-origin behaviours. The advantages of a linear behaviour assumption are that it describes the most common situations and that it is not necessary to adjust model parameters.

In the here presented case study the original variograms are standardized by local means and standard deviations, as a way to deal with local heterogeneities.

Notice that, although the methodologies presented did produce good results for a set of airborne pollution data, there are some limitations that must be taken into account:

- Regarding the first method, the assumption of measurement errors can stand as long as just a few points are removed; otherwise one risks to obtain a spurious variogram, that is neither robust nor representative of a consistent time period.
- The second method consists in the classification of continuous time periods with a given structured spatial pattern. If the obtained variogram is representative of a dispersed set of time periods, it can be useless for any posterior application.

References

Armstrong M (1984) Improving the estimation and modelling of the variogram. In: Verly G, David M, Journel AG and Marechal A (eds) Geostatistics for Natural Resourses Characterization. NATO ASI Series. Series C: Mathematical and Physical Sciences Vol. 122, 1-19

Chauvet P (1982) The Variogram Cloud. 17th APCOM Symposium, Colorado School of Mines, Golden, Colorado

Cressie N (1984) Towards resistant geostatistics. In: Verly G, David M, Journel AG and Marechal A (eds) Geostatistics for Natural Resourses Characterization. NATO ASI Series. Series C: Mathematical and Physical Sciences Vol. 122, 21-44

Cressie N and Hawkins D. (1980) Robust Estimation of the Variogram, I. In: Journal of the International Association for Mathematical Geology, 12, 115-125

Nunes C and Soares A (2004) Geostatistical Space-Time Simulation Model, paper accept in Environmetrics (in press)

Perrin O and Monestiez P (1999) Modelling of non-stationary spatial structure using parametric radial basis deformations. In: Gómez-Hernández J, Soares A and Froidevaux R (eds) GeoENV II - Geostatistics for Environmental Applications. Kluwer Academic Publishers, 175-186

Russo A, Nunes C and Bio A (2004) Air quality models resulting from multi-source emissions. GeoENV 2004 - Fifth European Conference on Geostatistics for Environmental Applications (accepted), Centre for Hydrogeology, University of Neuchâtel, Switzerland

Sampson PD and Guttorp P.(1992) Nonparametric Estimation of Nonstationary Spatial Covariance Structure. In: Jounal of the American Statistical Association, 87, 108-119

Srivastava RM and Parker HM (1989) Robust Measures of Spatial Continuity. In: Armstrong M (ed) Geostatistics, vol. 1. Kluwer Academic Publishers, 296-308

Switzer P (1989) Non-stationary spatial covariances estimated from monitoring data. In: Armstrong M (ed) Geostatistics, vol. 1. Kluwer Academic Publishers, 127-138

Wiersma GB (ed) (2004) Environmental Monitoring. CRC Press.

Multiple-point geostatistics: a powerful tool to improve groundwater flow and transport predictions in multi-modal formations

L. Feyen[1] and J. Caers[2]

[1] Katholieke Universiteit Leuven, Hydrogeology & Engineering Geology, Leuven, Belgium
[2] Stanford University, Department of Petroleum Engineering, Stanford University, USA

1 Introduction

Groundwater flow and transport models rely on a detailed description of the hydraulic properties of the subsurface. Because of financial and physical limitations to data collection, the subsurface heterogeneity cannot be described in detail deterministically. In recent decades numerous stochastic approaches have been developed to overcome this problem. These methods interpolate between hard data and use geologic, hydrogeologic and geophysical information to create images of the property of interest. An excellent review (up to 1995) of structure-imitating, process-imitating and descriptive methods is presented by Koltermann and Gorelick (1996). To date, most applications of geostatistics in hydrogeology have employed variogram-based techniques. The use of two-point correlation methods can be justified to describe the heterogeneity within a single statistically homogeneous stratigraphic unit. However, they are too limited to adequately characterize the spatial continuity for multimodal distributions, such as sand-shale formations, fractured rock masses or dolomite rocks with dissolution channels. Mimicking such complex geological settings requires relations between variables at three or more locations at a time. Also, variogram-based methods cannot take full advantage of existing prior geological knowledge or depositional information.

Multiple-point (mp) geostatistics aims to overcome the limitations of the variogram-based techniques in representing the geological continuity. The premise of mp-geostatistics is to move beyond two-point correlations between variables and to obtain relations between variables involving jointly several locations at a time. Strongly connected, curvilinear structures often constitute preferential flow paths that largely affect groundwater flow and transport. Conductivity barriers of various sizes and shapes may be present and need to be adequately represented. Mp-geostatistics is an active area of research that recently emerged in the field of petroleum engineering (see e.g., Caers and Zhang 2003, Strebelle 2002, Strebelle *et al.* 2002). In this paper we show that some of the techniques developed could prove to be powerful tools for a wide range of hydrogeological applications.

Therefore, we employ a synthetic non-stationary bimodal reference field representing a typical fluvial deposition that consists of permeable sand channels embedded in less permeable fine-grained floodplain material. We show results of a numerical analysis to evaluate groundwater flow and transport behavior in these types of settings and compare the mp-geostatistical approach with a more traditional 2-point variogram-based method.

2 Multiple-point geostatistics

The premise of mp-geostatistics is to generate models/images of the subsurface by borrowing patterns of geological heterogeneity from training images. Training images are merely conceptual and depict the expected patterns of geological heterogeneity. They need not be conditioned to any local data nor carry other locally accurate information. Several training images may be used to reflect different scales and styles of heterogeneities, or alternative conflicting geological interpretations to account for uncertainty about the subsurface architecture. 3-D training images can be obtained from unconditional object-based or pixel-based techniques, 3-D interpretation of outcrop data or high resolution geophysical data from analog fields of study. An example training image is presented in Fig. 1. It represents a fluvial setting of W-E oriented sand channels with an average channel width of 8-10 m. The training image was generated with the object-based algorithm fluvsim (Deutsch and Tran 2002).

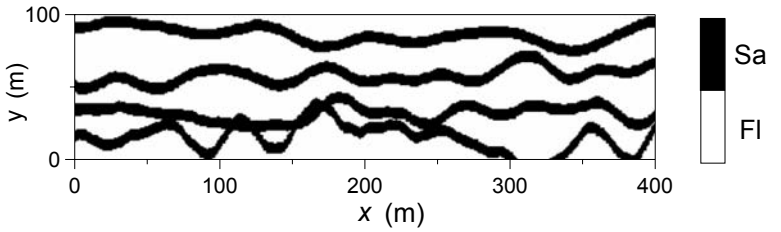


Fig. 1. Training image representing a fluvial deposit (generated with fluvsim, Deutsch and Tran 2002): Sa = sand, Fl = floodplain

Similar to variogram-based geostatistics, the training images are bound by the same principles of stationarity and ergodicity. They are essentially databases of geological architectures, and if patterns are to be extracted from them enough repetitivity and consistency of patterns is required. Ergodic considerations dictate the minimum size of the training image. Reproduction of large scale patterns like sand channels require training images of at least 2 times the size of the area in the direction of the channel continuity. Small training images will result in large ergodic fluctuations and will deteriorate pattern reproduction (Caers and Zhang 2002).

3 Single normal equation simulation algorithm

The single normal equation simulation algorithm (*snesim*) developed by Strebelle (2000, 2002) is an efficient pixel-based sequential simulation algorithm that obtains multiple-point statistics from the training image(s), exports it to the geostatistical model and anchors it to the actual subsurface data, both hard and soft. For each location $\mathbf{u} = (x, y)$ along a random path, the set of local data values and their spatial configuration, termed 'data event', is recorded. The training image is scanned for replicates that match this event. The central node values corresponding to the replicates are used to calculate the conditional probability of the central value, given to the data event. Current implementations of *snesim* acquire significant CPU efficiency by performing this scanning prior to simulation and storing the conditional probabilities in a dynamic data structure, called the search tree. In summary, the *snesim* algorithm works as follows:

- construct a 2-D (or 3-D) grid for the area, assign hard data to closest grid cells
- scan the training image for data events and store them in a search tree
- define a random path
- until each non-datum cell with coordinates $\mathbf{u} = (x, y)$ on the random path is visited
    1. search for the closest nearby well data and previously simulated cells (this set is the 'data event');
    2. obtain the probability distribution for the property to be simulated from the search tree; and
    3. draw an outcome from the probability model in step 2 and assign that value to the current grid cell.

In two-point geostatistical methods, the probability distribution in step 2 is obtained through some form of kriging based on a variogram model. In the *snesim* approach no kriging or variogram is involved and the probability distribution is obtained directly from the training image. For details of this procedure the reader is referred to the works of Strebelle (2000, 2002). Soft data can be included through an extension of Bayes' theorem, as discussed in Strebelle *et al.* (2002). Caers (2003) describes how production data can be incorporated using history matching.

The stationarity requirement for the training image does not imply that only stationary fields can be generated. Similar as to building complex variogram models from basic variograms, the well known principles of nesting models, rotation and affinity transformation can be used to build complex strongly non-stationary fields, such as sand channels with locally varying channel widths or changing channel directions. Nesting of models is obtained by using different training images for different scales of observations (see Strebelle and Journel 2001). For the rotation and affinity transforms, each single datum with original coordinates $\mathbf{u}^{orig} = (x^{orig}, y^{orig})$ in the entire data event is rotated and affinely transformed along the center node to the new coordinates $\mathbf{u}^{new} = (x^{new}, y^{new})$ according to

$$\mathbf{u}^{new} = A(\mathbf{u})\, R_\theta(\mathbf{u})\, \mathbf{u}^{orig} \tag{3.1}$$

where $$A(\mathbf{u}) = \begin{pmatrix} a_x(\mathbf{u}) & 0 \\ 0 & a_y(\mathbf{u}) \end{pmatrix}$$

contains the major and minor range of continuity, $a_x(\mathbf{u})$ and $a_y(\mathbf{u})$, respectively; and

$$R_\theta(\mathbf{u}) = \begin{pmatrix} \cos(\theta(\mathbf{u})) & -\sin(\theta(\mathbf{u})) \\ \sin(\theta(\mathbf{u})) & \cos(\theta(\mathbf{u})) \end{pmatrix}$$

contains the rotation angle azimuth $\theta(\mathbf{u})$. Example maps for the affinity factors and rotation angles are presented in Fig. 2. Location-dependent rotation and affinity information can be obtained from well-data, seismic, geological, or depositional information.

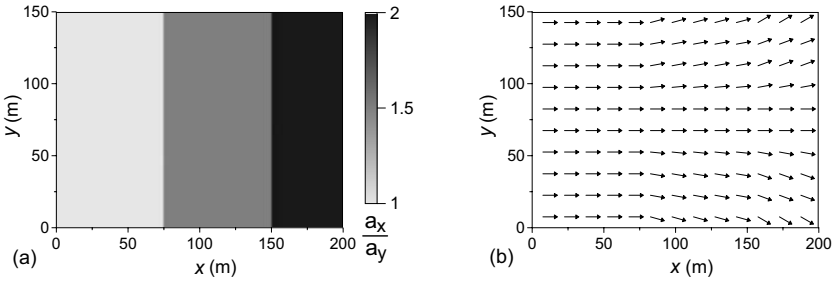


Fig. 2. a) affinity factors (ratio) $a_x(\mathbf{u})/a_y(\mathbf{u})$, $a_x(\mathbf{u}) = 1$; and **b)** channel rotation angles

4 Synthetic fluvial case study

A typical fluvial fan depositional system is presented in plan view in Fig. 3a. The area of interest is 200m by 150m, and is discretized in 1x1 m blocks. The system is characterized by high permeable sand channels embedded in less permeable fine-grained floodplain material. Sand channels compose 30% of the system and form an interconnected network. The channels are oriented W-E and diverge north- and southwards when moving along the *x*-axis. The channel width in the area decreases from 8-10 m to 3-5 m moving from west to east. The synthetic field is generated with *snesim* using the training image from Fig. 1 and the angle rotation and affinity data presented in Fig. 2. Within each facies the natural log of the hydraulic conductivity ($Y = \ln K$) is modeled as a realization of a second-order stationary Gaussian random field using the sequential Gaussian simulation algorithm *sgsim* (Deutsch and Journel 1998). The statistics of both random fields are presented in Table 1. The ln *K* histogram and experimental facies variogram for the

reference field are plotted in Fig. 3b and c, respectively. Despite the strong connectivity of the sand channels, the facies variogram is characterized by short ranges. This is because the variogram is only a measure of rectilinear connectivity that does not capture the curvilinearity of the sand channels. Krishnan and Journel (2003) indroduced multiple-point connectivity measures that better value the continuity of these curvilinear structures. The histogram clearly shows that a unimodal approach would be inappropriate and that the two facies composing the system should be modeled as distinct units.

Table 1. Statistical and hydraulic parameters for the sand and floodplain facies

	Sand	floodplain
variogram type	exponential	exponential
mean ln K	2	-3
geometric mean K	7.39 m/day	0.05 m/day
sill (σ_Y^2)	1	0.25
correlation length (λ_{Y_x})	25 m	10 m
anisotropy $(\lambda_{Y_x}/\lambda_{Y_y})$	2.5	2
effective porosity (θ)	0.3	0.2
dispersivity (α_L, α_T)	0.1 - 0.01 m	0.05 - 0.005 m

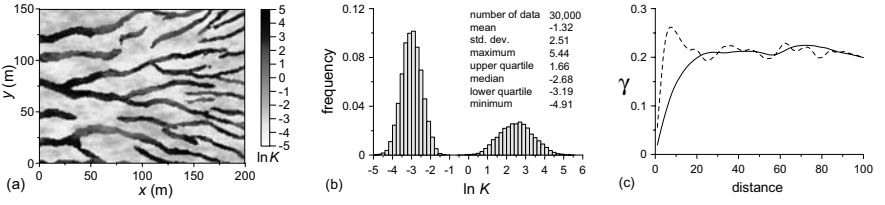


Fig. 3. a) ln K distribution for the reference field; **b)** ln K histogram; and **c)** experimental facies variogram

The reference field was randomly sampled at 100 locations, with 30 samples located in the sand channels. A sample consists of the facies type and the ln K value at that location. To compare the mp-statistics approach with a more traditional 2p-correlation approach, a random realization, conditioned on the extensive sample data set, was generated using the sequential indicator simulation program *sisim* (Deutsch and Journel 1998). The variogram used to generate the *sisim* realization is that of the training image shown in Fig. 1. The resulting ln K image, $\ln K$ histogram and experimental facies variogram are presented in Fig. 4a, b and c, respectively. The corresponding results for a random conditional realization generated with the *snesim* algorithm are given in Fig. 4d, e and f. It is important to note that only the facies geometries differ, and that the conditional ln K realizations of the sand and floodplain formations are the same in the conditional facies

realizations generated with *sisim* and *snesim*. Both methods very closely reproduce the ln *K* histogram and experimental facies variogram of the reference field. However, results clearly indicate that the 2p-approach fails to reproduce the channel network, in contrast to the mp-approach. Hence, using a variogram model accounting only for 2p-correlation fails to mimic the interconnected channel network, even for extensive conditioning data sets. Also presented in Fig. 4, in plates g, h and i, are the results of an unconditional realization generated with *snesim* and *sgsim*. Again, the statistics and the channel structures are very well reproduced. However, the exact locations of the channels are not reproduced without conditioning data.

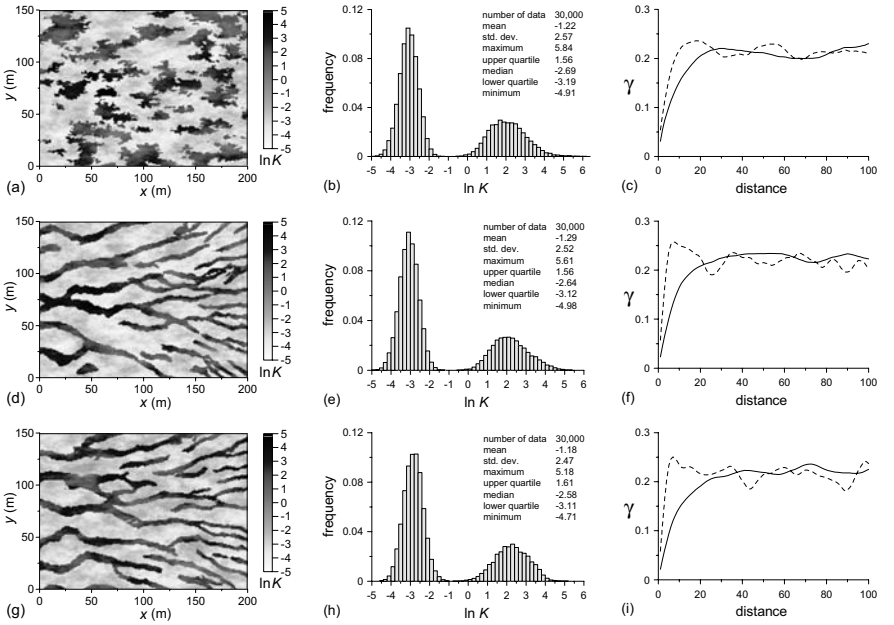


Fig. 4. a), **d)**, **g)** ln *K* distribution; **b)**, **e)**, **h)** ln *K* histogram; and **c)**, **f)**, **i)** experimental facies variogram: **a)**, **b)**, **c)** = *sisim*, conditional realization; **d)**, **e)**, **f)** = *snesim*, conditional realization; and **g)**, **h)**, **i)** = *snesim*, unconditional realization.

5 Some observations on flow and transport

To investigate the impact of the interconnected channel structure on groundwater flow and transport we performed a numerical analysis for which the results are presented in this section. We consider the case of a confined aquifer. The governing equations for steady-state confined groundwater flow and non-reactive single species transport are

$$\nabla.(\mathbf{T}\nabla h) - q = 0 \tag{5.1}$$

$$\mathbf{T} = \mathbf{K}b \tag{5.2}$$

$$\mathbf{v} = -\frac{\mathbf{T}}{b\theta}\nabla h \tag{5.3}$$

$$b\theta\frac{\partial c}{\partial t} = \nabla.(b\theta\mathbf{D}\nabla c) - b\theta\mathbf{v}\nabla c + q(c_s - c) \tag{5.4}$$

where $\mathbf{v}$ is the groundwater flow velocity vector (L/T); $\mathbf{T}$ is the transmissivity tensor (L^2/T); $\mathbf{K}$ is the hydraulic conductivity tensor (L/T); θ is the effective porosity (dimensionless); h is the hydraulic head (L); q are fluid sources/sinks (L/T); c is the solute concentration (L/T); c_s is the solute concentration in the fluid sources/sinks (L/T); and $\mathbf{D}$ is the hydrodynamic dispersion tensor (L^2/T). Neglecting molecular diffusion, the principal terms of the dispersion tensor are

$$D_{xx} = \alpha_L\frac{v_x^2}{|\mathbf{v}|} + \alpha_T\frac{v_y^2}{|\mathbf{v}|} \quad \text{and} \quad D_{yy} = \alpha_L\frac{v_y^2}{|\mathbf{v}|} + \alpha_T\frac{v_x^2}{|\mathbf{v}|} \tag{5.5}$$

where α_L and α_T are the longitudinal and transverse dispersivity (L), respectively.

The simulation model used to predict groundwater flow behavior is MODFLOW-2000 (Harbaugh *et al.* 2000). Transport is simulated with MT3DMS (Zheng and Wang 1999), using the Third-Order TVD solution scheme. The confined aquifer has a uniform thickness b = 25 m. At the north and south boundaries of the area no-flow boundary conditions are specified. Constant head values are set along the west (h = 22 m) and east (h = 20 m) boundaries. At location (x = 25, y = 75) a spill of an inert contaminant occurred during a 10-day period with a low constant flow rate of 100 l/day and a source concentration c_s = 20 mg/l. It is assumed that at the location of the spill the facies type is sand, and that this information is known in all cases evaluated in the numerical analysis. Numerically solving the groundwater flow and transport model requires specification of values for the unknown parameters θ, K, α_L and α_T in each cell throughout the model area. The spatial distribution of $\ln K$ is generated as described above using a combination of *snesim* and *sgsim*. The other hydraulic parameters are assumed homogeneous within the facies. The values used for the hydraulic parameters are given in Table 1. The small dispersivity values imply that transport is dominated by advection.

The 'true' head distribution is obtained by running the groundwater flow model for the reference field and is presented in Fig. 5a. The head contours clearly show the effect of the permeable sand channels, which dominate flow through the system. The head distributions for the conditional *sisim* and *snesim*, and the unconditional *snesim* realizations, are given in Fig. 5b, c and d, respectively. The conditional *snesim* realization yields a fairly good prediction of the reference head distribution. For the unconditional *snesim* realization, the effect of the sand channels on heads can also clearly be seen. However, the different positioning of the sand channels results in a less accurate prediction of the heads. Despite the large number of conditioning data, the conditional *sisim* realization fails to reproduce

the reference head field. This can be attributed largely to the inability of the method to represent the channel structure.

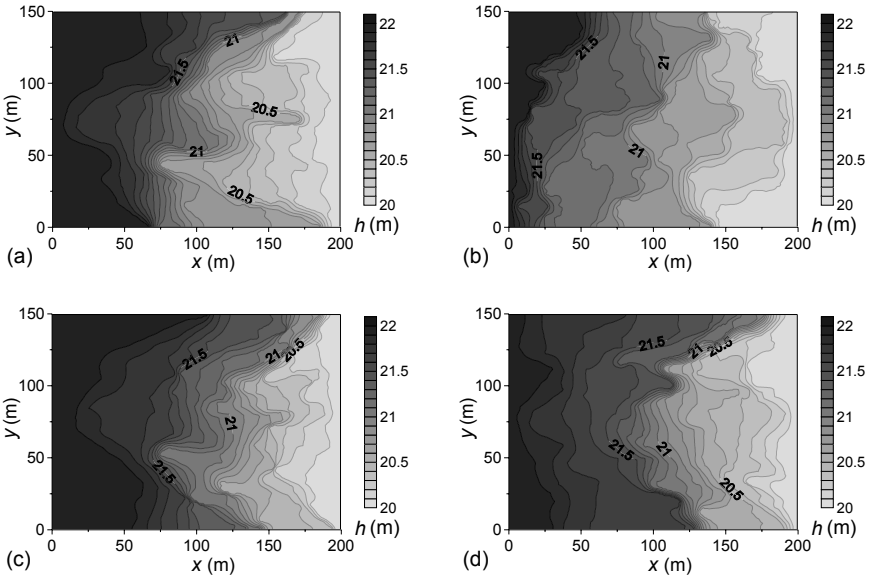


Fig. 5. Simulated head distributions: **a)** reference field; **b)** conditional *sisim* realization; **c)** conditional *snesim* realization; and **d)** unconditional *snesim* realization.

With the transport model we simulated the behavior of the released contaminant under natural steady-state flow conditions for a period of 500 days. Plates a, b, and c in Fig. 6 display the distribution of the contaminant plume for t = 100, 300 and 500 days, respectively. The bulk of the released contaminant moves through the permeable sand channels. Fig. 6 also shows the transport predictions for the conditional *sisim* (plates d, e and f) and *snesim* (plates g, h and i) realizations, and the unconditional *snesim* (plates j, k and l) realization. The variogram-based method underestimates solute movement in the direction of flow, as it is not able to reproduce the interconnected preferential flow paths. Once the solute mass enters into the floodplain material it moves downstream very slowly, until perhaps, a new permeable sand body is encountered. The conditional *snesim* realization yields a fairly good prediction of the location of the contaminant plume through time. Results for the unconditional *snesim* realization indicate that for transport predictions it is very important to accurately determine the location of the sand channels. The training image, angle and affinity information allow characterizing the structural features of the system, but conditioning is needed to precisely locate the sand channels.

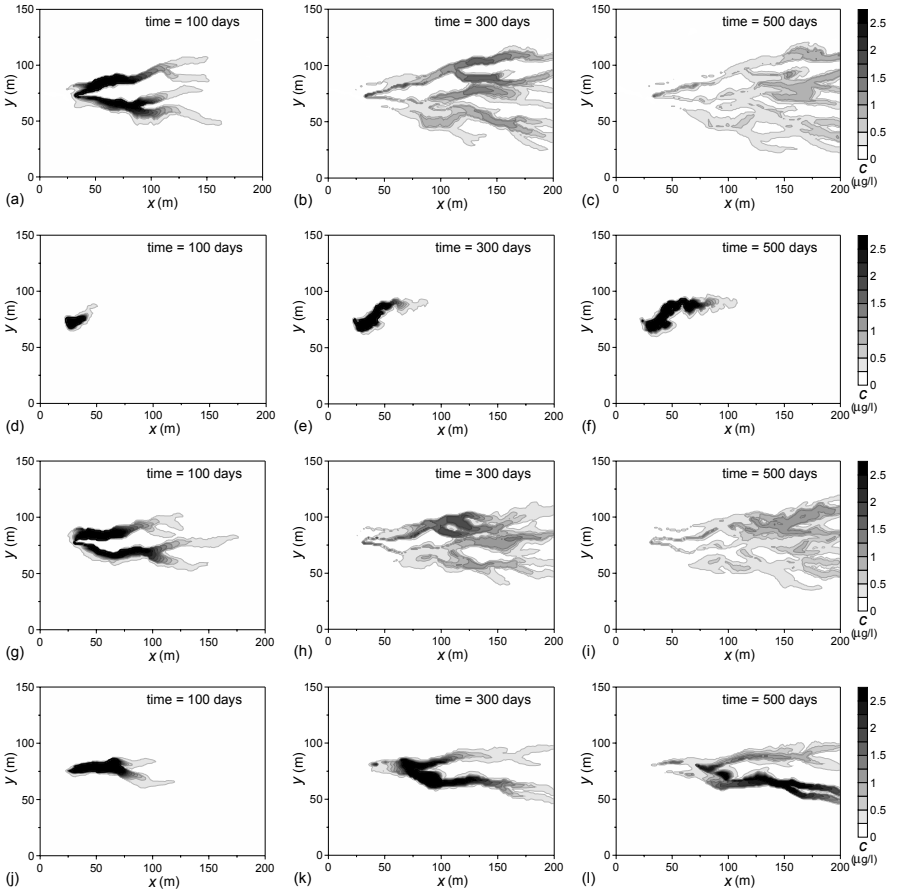


Fig. 6. Simulated contaminant concentrations for t = 100, 300 and 500 days: **a)**, **b)**, **c)** reference field; **d)**, **e)**, **f)** conditional *sisim* realization; **g)**, **h)**, **i)** conditional *snesim* realization; and **j)**, **k)**, **l)** unconditional *snesim* realization.

6 Conclusions

Results shown in this paper indicate that multiple-point geostatistics is potentially a very powerful tool to characterize subsurface heterogeneity for hydrogeological applications in a wide variety of complex geological settings. Geological structures or features such as sand channels or clay lenses often constitute preferential flow paths or obstacles to flow. Accurately representing and locating these structures is of high importance when predicting groundwater flow and transport, as was shown in this work. Because data are scarce, the mp-statistics are borrowed from training images that depict the expected patterns of geological heterogeneity.

The mp-statistics are exported to the geostatistical model and anchored to hard and/or soft data. Strongly non-stationary fields can be generated using several training images, angle rotation and affinity information. Mp-geostatistics should bring geological interpretation closer to hydrogeological modeling.

We note that the analysis presented herein was based upon comparing the simulation results of the example field with those of individual realizations. Further analysis is needed to systematically evaluate and quantify the effects of the different levels of geological uncertainty on groundwater flow and transport predictions in multi-modal settings.

Acknowledgements

The first author wishes to acknowledge the Fund for Scientific Research – Flanders (Belgium) for providing a Postdoctoral Fellowship and a mobility grant.

References

Caers J (2003) Geostatistical history matching under a training image-based geological model constraints. SPE Journal: SPE 77429: 218-226.

Caers J, Zhang T (2003) Multiple-point geostatistics: a quantitative vehicle for integration geologic analogs into multiple reservoir models. In: "Integration of outcrop and modern analog data in reservoir models" AAPG memoir, in press

Deutsch CV, Journel AG (1998) *GSLIB***,** Geostatistical Software Library and User's Guide. Oxford University Press, New York

Deutsch CV, Tran TT (2002) FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems. Computers and Geosciences 28: 525-535

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, U.S. Geological Survey modular ground-water model-user guide to modularization concepts and the ground-water flow process: U.S. Geological Survey Open-File Report 00-92

Koltermann CE, Gorelick SM (1996) Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches. Water Resources Research 32(9): 2617-2658

Krishnan S, Journel AG (2003) Spatial connectivity: from variograms to multiple-point measures. Mathematical Geology 35(8): 915-925

Strebelle S (2000) Sequential simulation drawing structures from training images. Ph.D. thesis, Department of Geological and Environmental Sciences, Stanford University, Stanford

Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point geostatistics. Mathematical Geology 34(1): 1-22

Strebelle S, Journel AG (2001) Reservoir Modeling using multiple-point geostatistics. SPE Journal: SPE 71324

Strebelle S, Payrazyan K, Caers J (2002) Modeling of a deepwater turbidite reservoir conditional to seismic data using multiple-point geostatistics. SPE Journal: SPE 77425

Zheng C, Wang PP (1999) MT3DMS, A modular three-dimensional multi-species transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems. Documentation and user's guide, U.S. Army Engineer Research and Development Center Contract Report SERDP-99-1, Vicksburg, MS

Simulation of radionuclide mass fluxes in a heterogeneous clay formation locally disturbed by excavation

M. Huysmans[1], A. Berckmans[2] and A. Dassargues[1, 3]

[1]Hydrogeology & Engineering Geology Group, Department of Geology-Geograpy, Katholieke Universiteit Leuven, Belgium,
e-mail: marijke.huysmans@geo.kuleuven.ac.be
[2]ONDRAF/NIRAS, Brussel, Belgium
[3]Hydrogeology, Department of Georesources, Geotechnologies and Building Materials, Université de Liège, Belgium

1 Introduction

The safe disposal of nuclear waste is an important environmental challenge. Several countries are investigating deep geological disposal as a long-term solution for high-level waste. The Belgian nuclear repository program, conducted by ONDRAF/NIRAS (Belgian agency for radioactive waste and enriched fissile materials), is in the process of characterizing the host rock capacities of the Boom Clay. This is a marine sediment of Tertiary age (Rupelian) (Wouters and Vandenberghe 1994). The research activities are concentrated at SCK CEN (Belgian Nuclear Research Centre) located on the nuclear zone of Mol/Dessel (province of Antwerp) where an underground experimental facility (HADES-URF) was built in the Boom Clay at 225 m depth. In this area, the Boom Clay has a thickness of about 100 m and is overlain by approximately 180 m of water bearing sand formations.

The isolation of waste from the biosphere is obtained by means of a multi-barrier concept, composed of engineered and natural barriers. In this study, the radionuclide migration through the most important natural barrier, the Boom Clay, is investigated. The average hydraulic conductivity value of this formation is very low (K=2.10^{-12} m/s), but the clay is not completely homogeneous. It contains alternating horizontal sublayers of silt and clay with an average thickness of 0.50 m and a large lateral continuity (Vandenberghe *et al.* 1997). Furthermore, the clay exhibits excavation-induced fractures around the excavated galleries (Dehandschutter *et al.* 2002). The sublayers have hydraulic conductivity values up to 10^{-10} m/s (Wemaere *et al.* 2002) and the fractures may have even higher hydraulic conductivity values. Therefore, the aim of this study is to model the transport of radionuclides through the clay, taking into account the geological heterogeneity and the excavation induced fractures around the galleries in which the waste will be stored.

2 Method

2.1 Data analysis

In order to analyze and simulate the heterogeneous hydraulic conductivity, measurements of the hydraulic conductivity and several secondary variables were collected. All measurements were carried out in the Mol-1 borehole (Wemaere *et al.* 2002). The resulting data set comprises of 52 hydraulic conductivity values, 71 grain size measurements, an electrical resistivity log, a gamma ray log and a description of the lithology variation. Hydraulic conductivity and grain size were measured in the laboratory on cores of 3 to 10 cm and 10 to 20 cm respectively. Borehole resistivity and gamma ray logging was performed with a vertical spacing of 15 cm. The lithology description was derived from a Fullbore Formation MicroImager log with a vertical resolution of 5 cm. The scales of all different measurements are of the same order of magnitude.

All secondary measurements show a clear correlation with hydraulic conductivity (Table 1). Electrical resistivity and hydraulic conductivity have a correlation coefficient of 0.73. Gamma ray, on the other hand, shows a negative and smaller correlation with hydraulic conductivity (R=-0.65). This lower correlation is probably caused by the presence of organic matter and glauconite in the Boom Clay, which both affect the gamma ray measurements. Grain size is observed to be well correlated with hydraulic conductivity. The correlation coefficient between d_{40} (i.e., the grain size for which 40% of the total sample has a smaller grain size) and hydraulic conductivity is 0.95. The lithostratigraphic column, determined on the basis of a Formation Micro Imager (FMI) log (Mertens and Wouters 2003), also shows a relationship with hydraulic conductivity: the mean log hydraulic conductivity of the clay layers (-11.7) is smaller than the mean log hydraulic conductivity of the silt layers (-11.3). All secondary parameters are thus well correlated with hydraulic conductivity and were therefore incorporated in the simulation of hydraulic conductivity.

In previous work (Vandenberghe *et al.* 1997), the Boom Clay formation was divided into three zones. This subdivision was confirmed by the statistical analysis. The deepest zone (Belsele-Waas Member: 278m – 292.4m) shows a large variability of hydraulic conductivity and the secondary variables, the middle zone (Putte and Terhagen Member: 216m – 278m) shows a small variability and the upper zone (Boeretang Member: 190.4m – 216m) shows an intermediate variability. Variograms and cross-variograms of all primary and secondary variables were calculated and modeled for the three separate zones. Table 2 shows the fitted log K variograms of the three zones of the Boom Clay formation. Fig. 1 shows two examples of experimental and fitted variograms and cross-variograms: the variogram of gamma ray of the Belsele-Waas Member and the cross-variogram of gamma ray and resitivity of the Belsele-Waas Member.

Table 1. Correlation coefficients between hydraulic conductivity and secondary parameters

Secondary parameter	Correlation coefficient with hydraulic conductivity
Electrical resistivity	0.73
Gamma ray	-0.65
Grain size (d_{40})	0.95

Table 2. Fitted log K variograms of the three zones of the Boom Clay formation

	Model	Nugget	Range	Sill
Boeretang Member	Spherical	0.035	4.6 m	0.03
Putte and Terhagen Member	Spherical	0.003	4.8 m	0.0056
Belsele-Waas	Spherical	0.23	5.5 m	0.38

2.2 Simulation of hydraulic conductivity

Detailed input fields reflecting the heterogeneity of hydraulic conductivity were simulated. These hydraulic conductivity fields serve as input for the hydrogeological model. Since the Boom Clay shows a large lateral continuity (Wouters and Vandenberghe 1994) and since the hydrogeological model is a local scale model, it could be assumed that the properties of the Boom Clay do not change considerably in the horizontal direction. Therefore, one-dimensional vertical simulations of the hydraulic conductivity were calculated.

These fields were generated using geostatistical sequential simulation which allows to take spatial variability and secondary data into account. The simulation algorithm is iterative and contains the following steps:

1. The location to be simulated is randomly chosen. The spacing between the locations to be simulated was 0.2 m, which is of the same order of magnitude as the measurement scale of the different variables.
2. The simple co-kriging estimate and variance are calculated using the original primary and secondary data and all previously simulated values using COKB3D (Deutsch and Journel 1998).
3. The shape of the local distribution is determined in such a way that the original histogram of hydraulic conductivity is reproduced by the simulation. This is achieved by the following approach. Before the start of the simulation, a look-up table is constructed by generating non-standard Gaussian distributions by choosing regularly spaced mean values (approximately from -3.5 to 3.5) and variance values (approximately from 0 to 2).

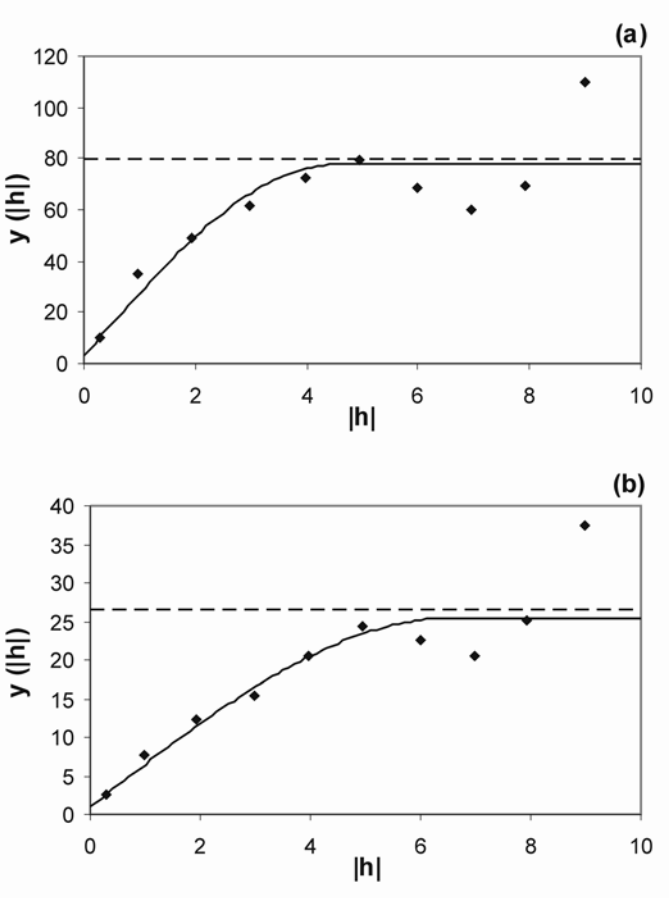


Fig. 1. Experimental and fitted a) variogram of gamma ray of the Belsele-Waas Member in the vertical direction and b) cross-variogram of gamma ray and resistivity of the Belsele-Waas Member in the vertical direction

The distribution of uncertainty in the data space can then be determined from back transformations of these non-standard univariate Gaussian distributions by back transformation of L regularly spaced quantiles, p^l, $l=1,...,L$:

$$K^l = F_K^{-1}\left[G\left(G^{-1}\left(p^l\right)\sigma_y + y^*\right)\right], \quad l = 1,...,L \tag{1}$$

where $F_K(K)$ is the cumulative distribution function from the original K variable, $G(y)$ is the standard normal cumulative distribution function, y^* and σ_y are the mean and standard deviation of the non-standard Gaussian distribution and the p^l, $l=1,...,L$ are uniformly distributed values between 0 and 1. From this look-up table the closest K-conditional distribution is retrieved by searching for the one with the closest mean and variance to the co-kriging values (Oz *et al.* 2003).

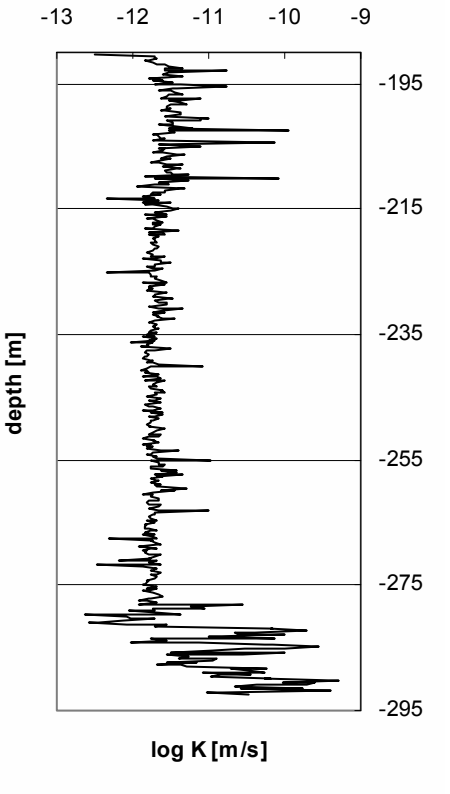


Fig. 2. Simulation of the vertical hydraulic conductivity of the Boom Clay

4. A value is drawn from the K-conditional distribution by Monte-Carlo simulation and assigned to the location to be simulated.
This approach creates realizations that reproduce (1) the local point and block data in the original data units, (2) the mean, variance and variogram of the variable and (3) the histogram of the variable (Oz *et al.* 2003). Fig. 2 shows one realization of the hydraulic conductivity of the Boom Clay.

2.3 Simulation of fractures

Around the galleries in the Boom Clay, excavation-induced fractures are observed. About 90% of the discontinuities are approximately parallel planes that are part of a twofold conjugated fault set (Fig. 3). The excavation-induced fractures around the future disposal galleries were modeled as discrete fractures. Their properties (i.e., extent, aperture, spacing, dip and strike) are simulated using Monte Carlo simulation. Since these fractures will probably have similar properties to the frac-

tures observed in previously excavated galleries in the Boom Clay, the input probability distributions of the fracture properties were derived from measurements carried out during recent tunnel excavation in the Boom Clay (Dehandschutter *et al.* 2002, Dehandschutter 2002, Mertens *et al.* 2004).

Examination of a mounting chamber excavated in the Boom Clay revealed that the rock mass seemed to be damaged up to approximately 2 m, with a lot of small scale disturbances. The fractures were open and pyrite oxidation was present on the surfaces up to a depth of 2 m (Mertens *et al.* 2004). To account for potential variations in clay properties, tunnel design or excavation techniques, some variation of the extent of the fractured zone was allowed and the extent of the fractures was simulated as a random number between 1 m and 3 m.

Fracture apertures were examined using microtomography and scanning electron microscopy (Dehandschutter *et al.* 2004). Values of tens of micrometers were measured. The aperture could be as large as 1 mm at the tunnel walls and decreased rapidly as the distance to the excavation increased (Dehandschutter B. personal communication). Therefore, fracture aperture was simulated as a random number between 0 μm and 50 μm.

Faulting is very intense over most part of the excavation zone. The distance between subsequent fractures is generally less then 1 m. The average spacing is about 70 cm (Mertens *et al.* 2004). The fracture spacing was drawn from a distribution reflecting these observations, i.e., a normal distribution with a mean of 0.70 m and a standard deviation of 0.12 m.

Fracture dip angle varies between 30 and 80 degrees. 82 fracture dip measurements of shear faults were carried out (Dehandschutter 2002). The average fracture dip was 53° and the standard deviation was 11°. The fracture dip was therefore drawn from a normal distribution with a mean of 53° and a standard deviation of 11°.

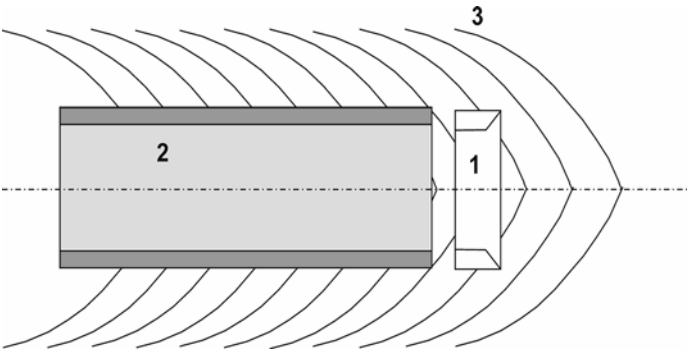


Fig. 3. Schematical representation of a vertical cross section through the Connecting Gallery showing the typical symmetrical form of the encountered shear planes (1. Tunneling maching; 2. Supported tunnel; 3. Induced shear planes)

Examination of the strike of discontinuities surrounding boreholes and larger excavations in the Boom Clay revealed that the strike of most discontinuities was perpendicular to the borehole or gallery axis (Dehandschutter 2002). The orientation of the fractures was fairly constant and all fractures were therefore assumed to have a strike perpendicular to the gallery axis.

The fracture geometry and properties were simulated by independent sampling from the proposed marginal distributions of fracture extent, aperture, spacing, dip and strike.

2.4 Hydrogeological model

A local 3D hydrogeological model of the Boom Clay, including the simulations of matrix hydraulic conductivity values and fractures, was constructed (Fig. 4). The model width in the x-direction is 20 m, i.e., half the distance between the disposal galleries. The model length in the y-direction is 15 m. This length was a compromise between including as many fractures as possible and keeping the computation time manageable. The model dimension in the z-direction is 102 m, i.e., the thickness of the Boom Clay in the nuclear zone of Mol-Dessel. The grid spacing is 1 m in the x-direction, approximately 0.17 m in the y-direction and between 0.2 m and 1 m in the z-direction. This fine grid was necessary to include the high resolution simulations of hydraulic conductivity and the geometry of the fractures. The vertical boundary conditions for groundwater flow are zero flux boundary conditions since the hydraulic gradient is vertical. The horizontal boundary conditions for groundwater flow are Dirichlet conditions. The specified head at the upper boundary is 2 m higher than the specified head at the lower boundary since the downward vertical hydraulic gradient is approximately 0.02 in the 100 m thick Boom Clay (Wemaere and Marivoet 1995). This gradient could vary in magnitude or even change direction over the long time period associated with radioactive waste disposal. In this study the gradient was however assumed to be constant. The boundary conditions for transport of the upper and lower boundaries are zero concentration boundary conditions (Mallants *et al.* 1999) since the hydraulic conductivity contrast between the clay and the aquifer is so large that solutes reaching the boundaries are assumed to be flushed away by advection in the aquifer.

The model was calculated for the radionuclide Se-79 since previous calculations revealed that this was one of the most important in terms of dose rates from a potential high-level waste repository for vitrified waste (Mallants *et al.* 1999). This radionuclide has a half-life of 65000 years, a solubility limit of 5.5e-8 mole/l, a diffusion coefficient of 2e-10 m^2/s and a diffusion accessible porosity of 0.13. The transport processes that were taken into account in the model are advection, dispersion, molecular diffusion and radioactive decay.

The nuclear waste disposal galleries are assumed to be situated in the middle of the Boom Clay. This radionuclide source is modeled as a constant concentration source with a prescribed concentration equal to the solubility limit. The radionuclides slowly dissolve into the groundwater until all available radionuclides are

dissolved. The source term is thus a constant concentration equal to the solubility limit until exhaustion of the source.
The radionuclide migration in the clay and the fluxes through the clay boundaries into the surrounding aquifers were calculated with FRAC3DVS, a simulator for three-dimensional groundwater flow and solute transport in porous, discretely-fractured porous or dual-porosity formations (Therrien *et al.* 1996, Therrien *et al.* 2003). The fractures were modeled as discrete planes with a saturated hydraulic conductivity of (Bear 1972):

$$K_f = \rho g (2b)^2 / (12\mu) \tag{2}$$

where ρ is the fluid density (kg/m³), g is the acceleration due to gravity (m/s²), $2b$ is the fracture aperture (m) and μ is the fluid viscosity (kg/(ms)). The model was run with ten different simulations of hydraulic conductivity and fractures as input. The computation time of one run of the model with a PC with a 1.8 GHz CPU and 512 MB RAM was approximately 6 to 8 hours.

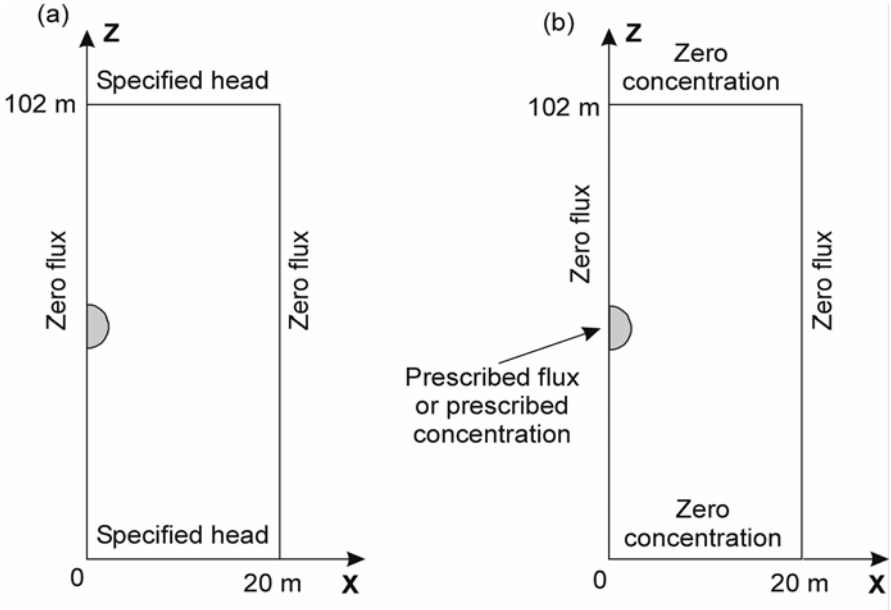


Fig. 4. Boundary conditions for **a)** flow and **b)** transport of 3D local hydrogeological model.

3 Results and discussion

Fig. 5 and 6 show the total Se-79 fluxes through the lower and upper clay-aquifer interfaces for 10 different simulations. The fluxes through the clay-aquifer interfaces gradually increase until they reach a maximum after approximately 200'000

years and decrease slowly afterwards due to exhaustion of the source. The difference between the fluxes of the 10 different simulations is the largest in the time period from 100'000 till 200'000 years. The total amount of Se-79 leaving the clay was calculated as the flux integrated over time for each simulation. The total Se-79 masses leaving the clay vary between 2.200e+12 Bq and 2.438e+12 Bq through the lower clay-aquifer interface and between 2.045e+12 Bq and 2.252e+12 Bq through the upper clay-aquifer interface.

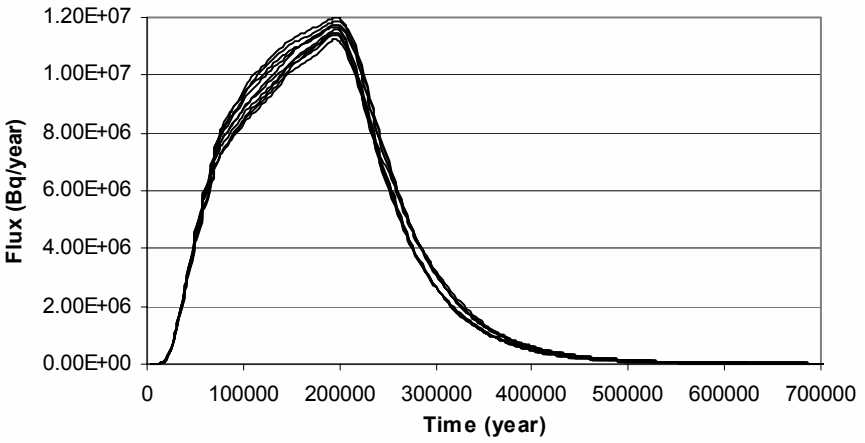


Fig. 5. Total Se-79 flux (Bq/year) through the lower clay-aquifer interface

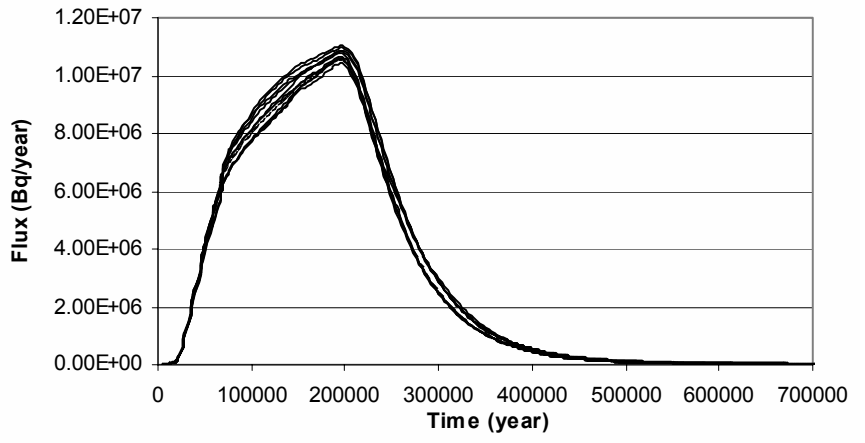


Fig. 6. Total Se-79 flux (Bq/year) through the upper clay-aquifer interface

The range of total Se-79 masses leaving the clay is thus rather small. The difference between the largest and the smallest calculated mass is 10%. This result is important for the evaluation of the suitability of the Boom Clay Formation as a

host rock for vitrified nuclear waste storage. The total mass fluxes leaving the clay, taking excavation induced fractures and high-conductivity sublayers into account, are not very different from the mass fluxes calculated by a previous simple homogeneous model. This is probably caused by the relatively small importance of transport by advection compared to transport by diffusion in such media. Changes in the heterogeneity of hydraulic conductivity do not change the output fluxes significantly and do not affect its main safety function. This again shows that the Boom Clay is a very robust barrier.

4 Conclusions

In this study, the transport of radionuclides through a potential host formation for the disposal of vitrified nuclear waste was calculated, taking the geological heterogeneity and the excavation induced fractures into account. The calculated fluxes through the clay boundaries into the surrounding aquifers were very similar for all the different simulations. The difference between the largest and the smallest calculated mass leaving the clay was 10%. These results show that changes in the heterogeneity of hydraulic conductivity do not change the output fluxes significantly. The robust barrier function of the Boom Clay formation is thus confirmed by these results.

Acknowledgements

The authors wish to acknowledge the Fund for Scientific Research – Flanders for providing a Research Assistant scholarship to the first author. We also wish to thank ONDRAF/ NIRAS (Belgium agency for radioactive waste and enriched fissile materials) and SCK-CEN (Belgian Nuclear Research Centre) for providing the necessary data for this study. We also thank René Therrien and Rob McLaren for providing Frac3dvs and for their assistance.

References

Bear J (1972) Dynamics of fluids in porous media. American Elsevier, New York, p. 764

Dehandschutter B, Sintubin M, Vandenberghe N, Vandycke S, Gaviglio P, Wouters L (2002) Fracture analysis in the Boom Clay (URF, Mol, Belgium). Aardk. Mededel., 12, 245-248

Dehandschutter B (2002) Faulting and Fracturing during Connecting Gallery tunnelling at the URL at Mol (SCK-CEN). ONDRAF/NIRAS unpublished internal report, p. 19

Dehandschutter B, Vandycke S, Sintubin M, Vandenberghe N, Gaviglio P, Sizun J-P, Wouters L (2004) Microfabric of fractured Boom Clay at depth: a case study of brittle-ductile transitional clay behaviour. Applied Clay Science, in press

Deutsch CV, Journel AG (1998) GSLIB geostatistical software library and user's guide. Oxford University Press, New York

Mallants D, Sillen X, Marivoet J (1999) Geological disposal of conditioned high-level and long lived radioactive waste: Consequence analysis of the disposal of vitrified high-level waste in the case of the normal evolution scenario. Niras, Brussel R-3383, p. 82

Mallants D, Marivoet J, Sillen X (2001) Performance assessment of vitrified high-level waste in a clay layer. Journal of Nuclear Materials, 298, 1-2, 125-135

Mertens J, Wouters L (2003) 3D Model of the Boom Clay around the HADES-URF. NIROND report 2003-02, p. 48

Mertens J, Bastiaens W, Dehandschutter B (2004) Characterization of induced discontinuities in the Boom Clay around the underground excavations (URF, Mol, Belgium). Applied Clay Science, submitted

Oz B, Deutsch CV, Tran T T, Xie Y (2003), DSSIM-HR: A FORTRAN 90 program for direct sequential simulation with histogram reproduction: Computers & Geosciences. v. 29, no.1, 39-51

Therrien R, Sudicky EA (1996) Three-dimensional analysis of variably-saturated flow and solute transport in discretely-fractured porous media. Journal of Contaminant Hydrology, 23, 1-2, 1-44

Therrien R, Sudicky EA, McLaren RG (2003) FRAC3DVS: An efficient simulator for three-dimensional, saturated-unsaturated groundwater flow and density dependent, chain-decay solute transport in porous, discretely-fractured porous or dual-porosity formations, User's guide. p. 146

Vandenberghe N, Van Echelpoel E, Laenen B, Lagrou D (1997) Cyclostratigraphy and climatic eustacy, example of the Rupelian stratotype. Earth & Planetary Sciences, Academie des Sciences, Paris, vol. 321, 305-315

Wemaere I Marivoet J (1995) Geological disposal of conditioned high-level and long lived radioactive waste: updated regional hydrogeological model for the Mol site (The north-eastern Belgium model) (R-3060). Niras, Brussel, p. 72

Wemaere I, Marivoet J, Labat S, Beaufays R, Maes T (2002) Mol-1 borehole (April-May 1997): Core manipulations and determination of hydraulic conductivities in the laboratory (R-3590). Niras, Brussel p. 56

Wouters L, Vandenberghe N (1994) Geologie van de Kempen: een synthese: Niras. NIROND-94-11, Brussel, p. 208

Modeling density-dependent flow using hydraulic conductivity distributions obtained by means of non-stationary indicator simulation

K.-J. Röhlig, H. Fischer and B. Pöltl

Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) mbH,
Schwertnergasse 1, 50667 Köln, Germany

1 Introduction

Within the framework of Safety Assessments for deep (geologic) repositories of radioactive waste, the behavior of the repository components (engineered, geotechnical and natural barriers) is described using model calculations with regard to the potential migration of radionuclides and to hazards for man or the environment. An essential part of this modeling framework is the assessment of the retention potential of the host rock and the overburden ("geosphere", "far-field"), which forms an important component of the barrier system.

Assessing the post-closure safety of deep repositories by means of safety analyses implies the necessity to treat a variety of inevitable uncertainties caused by the complexity of the phenomena and systems under consideration and by the long timeframes of concern. It is necessary to demonstrate that assessment results are defendable and safety is not compromised even in the presence of uncertainties. This demonstration is based on multiple lines of reasoning, e.g. robustness arguments, the use of conservative assumptions, and confidence-building in data and models.

Recognized approaches are available to characterize uncertainties, to propagate them through safety analyses, and to present results. Namely, probabilistic uncertainty analyses based on Monte Carlo simulations have been proven to be an adequate means to explore ranges of possibilities, to propagate uncertainties amenable to characterization by probability density functions through numerical analyses and to assess uncertainties of potential (detrimental) consequences and sensitivities (Cadelli *et al.* 1996, Marivoet *et al.* 1997, Baltes and Röhlig 2004). However, in other areas such as (hydro-)geological modeling there is still a need to develop and improve appropriate methods. Although it is recognized that the utilization of geostatistical methods in hydrogeology might contribute to a consistent treatment of this problem (Bonano and Thompson 1993), most existing analyses are still based on manually derived hydrogeological models even though some attempts to utilize geostatistical methods have been undertaken (LaVenue Marsh and RamaRao 1992, Zimmerman *et al.* 1998, Jaquet *et al.* 1998, Jaquet *et al.* 2001).

The paper presented here describes a case study which illustrates how a variety of geological information can be incorporated stepwise into hydrogeological models derived by means of non-stationary indicator simulation. The aim of the study was

- to explore possibilities of addressing uncertainty and spatial variability in hydrogeological models by means of geostatistical methods,
- to check how such methods can be used in the framework of probabilistic Safety Assessments,
- to obtain conclusions about uncertainty ranges and bandwidths of calculation results for a given hydrogeological setting and database.

Due to the availability of a large amount of hydrogeological, geophysical and other data and information, the Gorleben site (Northern Germany) has been used for the case study in order to demonstrate the approach. The study is a practical application of aquifer characterizations by firstly simulating hydrogeological units and then the hydrogeological parameters. Earlier work performed within the study was presented at the geoENV conferences in 1998 and 2000 (Röhlig 1999, Röhlig and Pöltl 2001).

2 The Gorleben site and the database

The Gorleben site is located in the federal state of Lower Saxony (Niedersachsen, Northern Germany). Its suitability for the final disposal of all kinds of solid and solidified, especially heat-producing radioactive waste had been investigated since 1979. Since 2001, the exploration has been interrupted to clarify conceptual and safety questions.

The repository is planned at a depth of about 850 m in the Gorleben salt dome. The tertiary clay cover of the salt dome has been partially removed by subglacial erosion forming a system of channels, one of which is the ”Gorleben Erosion Channel” (Fig. 1). This channel has a length of more than 16 km and its width ranges between 1 and 2 km. At a depth of about 250 m the erosion processes had reached the caprock and at some locations the rock salt. The channel is filled with sandy and gravely sediments forming a system of two aquifers separated by clay layers. The main clay-bearing structure, the so-called Lauenburg Clay Complex, contains apart from clay also the other materials present at the site (sand, silt, gravel).

For the evolution expected for a repository in rock salt, salt creep will close the repository vaults (“convergence”) and the waste will be completely isolated by the surrounding rock salt. Only in a scenario which leads to a release of radionuclides from the repository and the salt dome, these radionuclides would migrate through the aquifer system of the Gorleben channel to the surface. Therefore, groundwater regime and a possible radionuclide transport through the channel have to be studied in a safety assessment.

Hydrogeological investigations were performed in an area of more than 300 km^2 around the salt dome. 340 borehole logs have been compiled and inter-

preted (Fig. 1). The compilation contains a consistent stratigraphic and petrographic classification, remarks concerning the genesis and colour of materials, and a hydrogeological classification. In addition, the following information is available (Schelkes *et al.* 1990, Ludwig 1994, Ludwig and Kösters 2002):

- geological and hydrogeological interpretation,
- results of pumping tests,
- results of salt concentration measurements,
- groundwater ages,
- seismic and geoelectrical data.

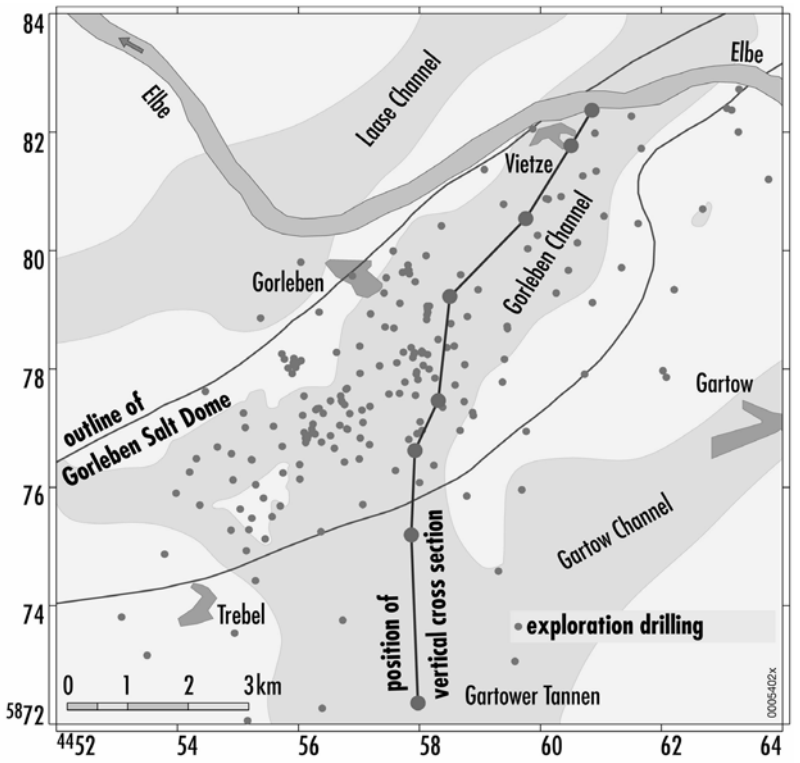


Fig. 1. The Gorleben site: Positions of the *Salt Dome*, the *Gorleben Channel*, the *exploration drillings* and of the *vertical cross section* used for modeling

3 Stepwise approach for the integration of information

Due to the different nature of the several information types to be integrated in hydrogeological models, an approach for utilizing them might take credit from a va-

riety of methods. Spatial statistical methods have the potential to generate images of structures which can be *conditioned* using hard data and to generate either best estimates for such images or *series of realizations* which are equally probable under given assumptions. The latter would allow fitting such methods into a framework of probabilistic uncertainty analyses. Therefore, spatial statistical methods (especially indicator simulation) have been chosen as a basis for the case study. The steps undertaken within the study are sketched in Fig. 2 and described in the following paragraphs.

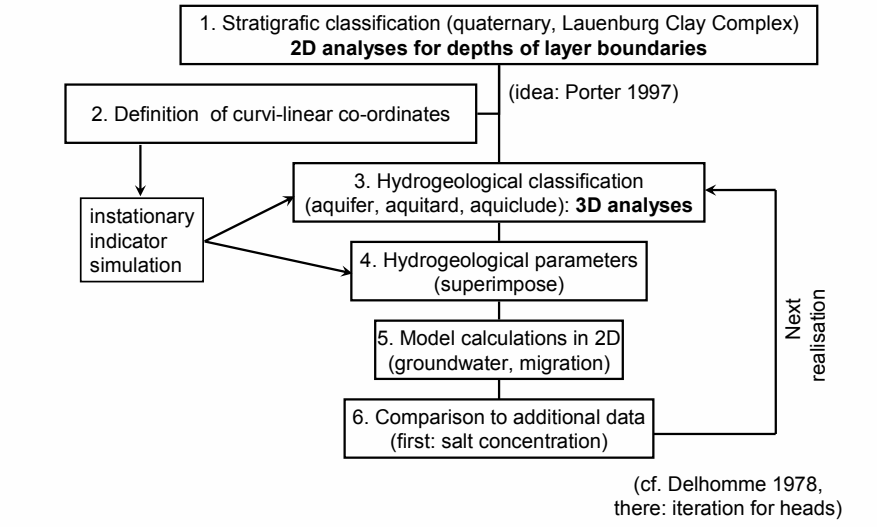


Fig. 2. Stepwise approach for the integration of information

3.1 Two-dimensional analyses

In order to identify trends and anisotropies, a preliminary two-dimensional horizontal analysis of borehole data has been performed. The total thicknesses of petrographic and stratigraphic units, the portions of units in relation to the borehole lengths, the depth of the base of quaternary (stratigraphic), and the depths of the upper and lower boundaries of the Lauenburg Clay Complex have been analysed as functions of two variables (easting and northing). The analysis included uni- and bivariate statistics, variography, and kriging (Röhlig 1999). While most of the obtained results served only to explore basic features of the dataset, the kriging results for the base of quaternary and for the depths of the boundaries of the Lauenburg Clay Complex were directly used as input for the definition of the curvi-linear co-ordinates described in the next section.

3.2 Curvi-linear co-ordinates

Porter and Hartley (1997) stated that they achieved better kriging results for spatial distributions of hydrogeologic units when they took into account the stratigraphic information about the Lauenburg Clay Complex and the base of quaternary. It was considered useful to account for the features of the layered Gorleben channel by performing variogram analyses which allow for the changes in the shape of strata. Porter and Hartley (1997) defined a co-ordinate system mapping the boundaries of the Lauenburg Clay Complex and the base of quaternary to constant values of the transformed vertical co-ordinate. "This can be thought of as using the geological age of the horizon to define a vertical coordinate." The variograms for variables indicating the presence or absence of materials showed, if obtained in the transformed system, a much better "horizontal" correlation structure than for Cartesian co-ordinates.

This approach can be further justified by the fact that the indicator variables characterizing the presence of the „Lauenburg Clay Complex" (stratigraphic) show a strong correlation to the ones characterizing the presence of the material „clay" (petrographic). The clay is mainly forming aquicludes the distribution of which influences the groundwater regime significantly.

Developing the ideas of Porter and Hartley (1997), a curvi-linear co-ordinate system has been defined as follows: The boundaries of the outcrop of the Gorleben Erosion Channel were transformed into surfaces of constant co-ordinate values for the "horizontal" co-ordinate. The kriging results obtained in the two-dimensional analyses described above for the base of the quaternary as well as the lower and the upper boundaries of the Lauenburg Clay Complex have been assumed to be surfaces of constant co-ordinate values for the „vertical" co-ordinate (Fig. 3) (Röhlig 1999).

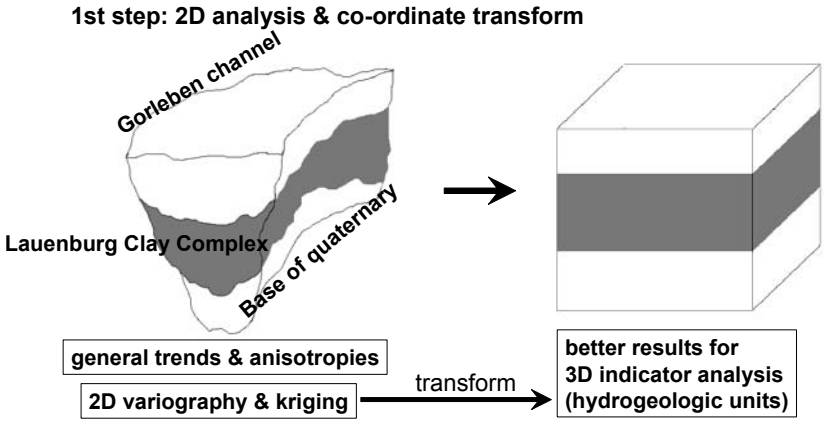


Fig. 3. Definition of curvi-linear co-ordinates

The simulations of hydrogeological units and parameters described in the following two chapters have been performed in curvi-linear co-ordinates. The variogram models used do not depend on the position in transformed co-ordinates. However, after being transformed back to Cartesian co-ordinates, they become dependent on spatial location. Thus, the simulations based on theses variogram models can be seen as non-stationary in the original co-ordinates.

3.3 Three-dimensional geostatistics for hydrogeological units

Using the curvi-linear system described above, variography and non-stationary conditional simulation of categorical variables characterizing 3 hydrogeological units (aquifer, aquitard, aquiclude) were carried out (Fig. 4 center). The hydrogeological classification of borehole data (Ludwig 1994) served for conditioning. As discussed in Röhlig (1999), the curvi-linear co-ordinates allow, compared to Cartesian co-ordinates, a better fitting of variogram models and the simulation results coincide much better with the hydrogeological site interpretation given in (Schelkes *et al.* 1990).

The variograms obtained show a much better horizontal correlation structure than the ones in Cartesian co-ordinates. Accordingly, the simulated units have a higher level of continuity which was seen as more realistic (Röhlig 1999).

3.4 Simulation of hydrogeological parameters

The only available data concerning hydrogeological parameters are bandwidths for each hydrogeological unit. These were obtained based on grain size analyses, pumping tests and literature reviews (Ludwig 1994). Therefore, no direct variography, kriging, or conditioned simulation is possible for these data.

Dependent on the range of the variables (hydraulic conductivity, porosity, dispersion lengths), log-uniform or uniform distributions were assumed for the hydrogeological parameters. After having divided the range of each variable into intervals, indicator variables have been defined for these intervals. The theoretical variance of each indicator function has been chosen as sill for the corresponding indicator variogram. Concerning the spatial continuity of parameters, varying assumptions have been made. A spherical variogram model with a range of 10 % of the model area of 16 km x 16 km x 400 m has been chosen for most of the calculations. Spherical variograms with 50 % range as well as models with a pure nugget effect were also tested. In addition, as recommended by Journel and Alabert (1988), variants with high connectivity (spherical variograms with 50 %) for high conductivities and low connectivity (pure nugget effect) for low conductivity values and vice versa were tested in unconditional indicator simulation runs.

However, the influence of these choices for the spatial continuity on the calculation results of hydrogeological modeling was, as discussed in the next chapters, not very significant.

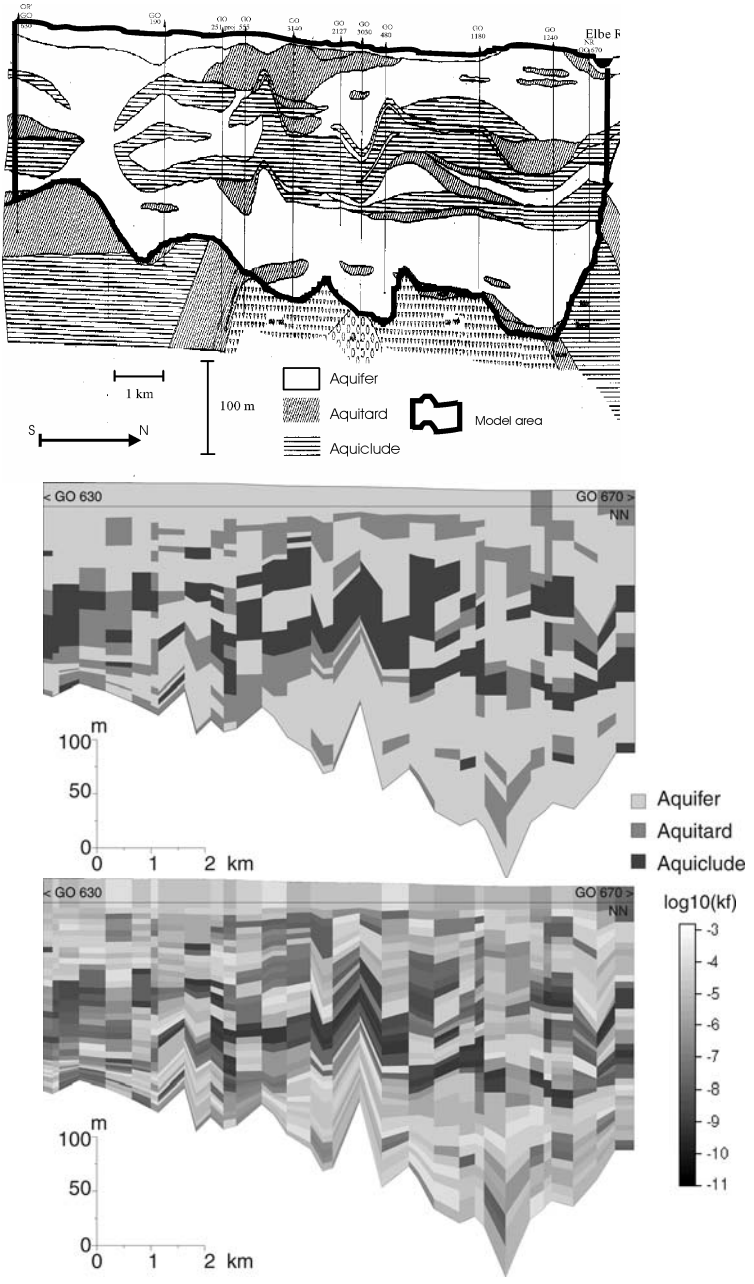


Fig. 4. Representative cross section for groundwater and transport calculations: Hydrogeological interpretation modified after Schelkes *et al.* (1990) (top), conditional simulation of categorical variables (center), and conductivity distribution (bottom)

The simulation of parameter distributions was carried out separately for each of the previously simulated hydrogeological units. (Fig. 4 bottom). Several algorithms were tested with the result that indicator simulation using a GSLIB algorithm (Deutsch and Journel 1992) was most suitable for the purpose of the study.

3.5 Probabilistic uncertainty analyses for freshwater models

The analyses described previously resulted in three-dimensional spatial distributions of hydrogeological parameters. Ideally, these distributions should serve for three-dimensional simulations of density-dependent flow, because the groundwater regime at Gorleben is highly influenced by the salinity gradients. However, those calculations are presently not practicable within acceptable calculation times for such large models. Therefore, most studies carried out for Gorleben are based on a vertical-plane (two-dimensional) cross section through the Gorleben Channel (Fig. 1 and 4). It is considered that such an approach captures major features since the cross-section follows the main groundwater flow direction (Schelkes *et al.* 1990, Ludwig 1994, Ludwig and Kösters 2002).

However, even two-dimensional density-dependent calculations require calculation times of several days per run. Because this would prohibit a full probabilistic uncertainty analysis, such an analysis has been only undertaken for freshwater models. The purpose of the analyses was rather the demonstration of methodology than an adequate assessment of the features of Gorleben.

Probabilistic uncertainty analyses were carried out for performance measures like advective groundwater travel times and contaminant fluxes. For the contaminant fluxes the uncertainty of the flux evolution with time (Fig. 5) as well as the flux maximum and the time of its occurrence were assessed.

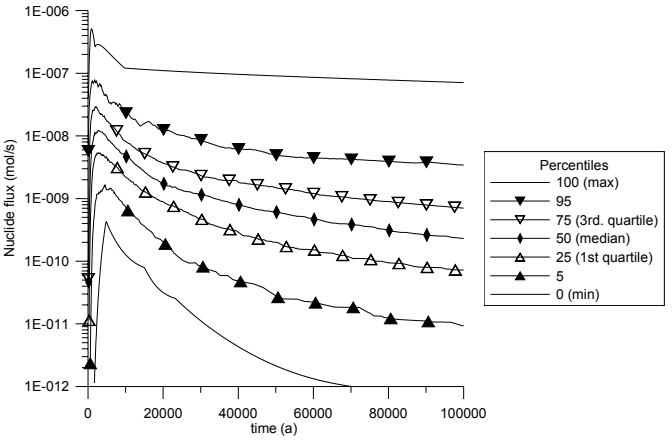


Fig. 5. Freshwater calculations: Percentiles of integrated nuclide fluxes (evolution with time) crossing the model boundary for 300 simulation runs (spherical model for conductivity distributions, range 10 % of model area)

In addition, a method to localize regional sensitivities for variables varying with position has been developed and tested (Röhlig and Pöltl 2001).

As expected, the uncertainty bandwidths increased with increasing spatial continuity of the parameter distributions (cf. the preceding paragraph). However, the differences of the bandwidths were rather marginal, thus indicating that the models were mainly determined by the distributions of the hydrogeological units while parameter distributions are comparably less important. In comparison to earlier analyses based on stationary indicator simulation for only 2 hydrogeological units (Röhlig and Pöltl 2001) the uncertainty bandwidths were remarkably reduced. This reduction is caused by the integration of a greater amount of site-specific information (e.g. the stratigraphic data used for the definition of the co-ordinate transformation and the simulation of 3 instead of 2 hydrogeological units).

3.6 Modeling of density-dependent groundwater flow

In order to allow a comparison of simulation results with site-specific information, the freshwater models mentioned previously were replaced by density-dependent groundwater flow models. This was necessary because the groundwater density which depends on the salt content is an important feature for the groundwater movement on site. However, density-dependent models are based on coupled nonlinear equations. This causes a remarkable effort for each calculation run. Several numerical codes were tested for the density-dependent flow simulations with the result that the SPRING® code (König 2002) was most suitable for the purpose of the study.

For most of the runs manual interventions for adjusting numerical parameters dependent on the numerical performance of the specific realization were necessary. Therefore and due to the high calculation effort it was not possible to run automated probabilistic Monte Carlo uncertainty analyses for the density-dependent models.

However the use of density-dependent models allows comparing the calculated salinity profiles with the ones measured on site. Thus it was possible to evaluate the results of geostatistical simulation and judge about the quality of realizations best fitting the measured density profiles as recommended by Delhomme (1979). It was also possible to compare the results of calculations based on models with different spatial connectivities for the hydrogeological parameters (cf. 3.4). For these comparisons, several penalty functionals indicating either differences in the function values or in the function values and the derivatives (the latter similar to Sobolev norms) were tested. The latter are the ones shown in Fig. 6.

Problems were caused by the variety of degrees of freedom for the modeling assumptions. In addition to the variations caused by the spatial distributions of facies and parameters varying from realization to realization and the choice of the spatial connectivity (paragraph 3.4), the results also showed to be sensitive against choices concerning boundary conditions and parameters like transversal dispersivity.

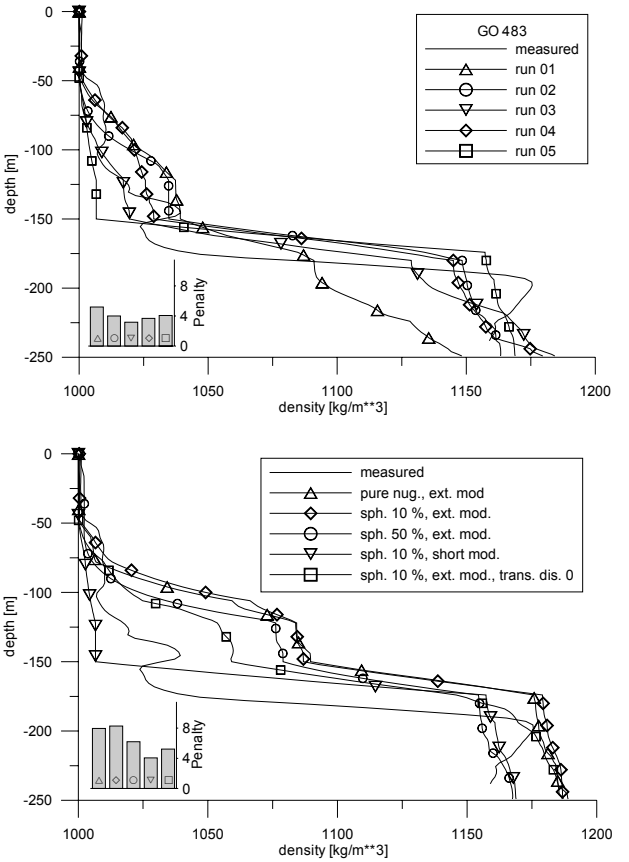


Fig. 6. Density profiles: measured values versus results and Sobolev penalty functionals calculated for several realizations of hydrogeological units with constant underlying modeling assumptions (top) and for the same realization but with varying modeling assumptions (bottom)

This is illustrated in the density profiles shown in Fig. 6: The diagram on top shows a comparison of the density profiles calculated for 5 realizations of hydrogeological units (cf. 3.3) with the measured profile at a certain location. The range of variation of the results is rather small compared to the one in the bottom figure where the realization was left constant and different modeling assumptions were used. It is also evident that a change in the boundary conditions ("short mod." vs. "ext. mod.") or of assumptions concerning the transversal dispersivity ("trans. dis. 0") causes more significant changes than changing assumptions concerning the spatial connectivity of the hydrogeological parameters (cf. 3.4, "pure nug.", "sph. 10 %", "sph. 50 %"). A similar outcome about the choice of parameter distributions being of relatively small influence was already achieved for freshwater calculations.

4 Conclusions

The study has demonstrated that geostatistical analyses are promising as a first step towards an integrated assessment of the hydrogeological features of repository sites covered or surrounded by sedimentary systems. Plausible hydrogeological models consistent with the information used could be derived. It could be demonstrated that information additional to the "hard" borehole data could be accounted for e.g. by performing simulations in transformed co-ordinates based on this information. The tested methods are especially efficient for sites like Gorleben where detailed data are given at a high density.

The groundwater and contaminant migration calculations performed using the derived models can in principle be fitted into the frame of probabilistic safety assessments and support the arguments used in a Safety Case. Using freshwater models, it has been shown how such analyses can contribute to a consistent treatment of uncertainties coming from spatial variability and lack of knowledge in probabilistic safety assessments. However, in the case of a complex density-dependent flow system the high computational effort still prohibits the performance of the high amount of density-dependent flow model runs necessary for probabilistic analyses. Instead, an approach where modeling results are compared with other site-specific information (here: measured density profiles) was used.

The relatively small uncertainty bandwidths obtained for several realizations of hydrogeological units with constant underlying modeling assumptions showed that uncertainties caused by spatial variability could be narrowed down.

However, the results showed to be sensitive against various model assumptions which were hard to test on reality. This caused remarkable problems which could only be resolved by means of more detailed (e.g. three-dimensional) hydrogeological studies as described and envisaged by Ludwig and Kösters (2002).

The methodology used for the case study strongly depends on the specific site under consideration. A "generic" approach or methodology for the integration of various geoscientific information into hydrogeological models will hardly be achievable.

References

Baltes B and Röhlig K-J (2004) Longterm safety of final repositories: German experiences concerning the rôle of uncertainty and risk in assessments and regulations. PSAM 7 – Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management, 2004. Springer, London

Bonano EJ and Thompson BGJ (1993) "Guest Editorial", Reliability Engineering & System Safety. Special issue on Probabilistic Risk Assessment for Radioactive Waste, Vol 42 Nos 2-3, 103-109

Cadelli N, Escalier des Orres P, Marivoet J, Martens, K-H, Prij, J (1996) Evaluation of elements responsible for the effective engaged dose rates associated with the final storage of radioactive waste: Everest project. EUR 17122 EN. EC, Luxembourg

Delhomme JP (1978) Kriging in the hydrosciences. Adv. Water Res., 1, 251-266.

Deutsch CV and Journel AG (1992) GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press. New York

Jaquet O, Schindler M, Voborny O, Vinard P (1998) Modelling of Groundwater Flow at Wellenberg Using Monte Carlo Simulations, Mat. Res. Soc. Symp. Proc., Vol. 506, 865-872.

Jaquet O, Rivera A, Genter M, Fillion E (2001) Site du Gard: Héterogénéités, Simulations Géostatistiques et Modélisation Hydrodynamique, Cahiers de Géostatistique, Fascicule 7, Fontainebleau

Journel AG and Alabert FG (1988) Focusing on Spatial Connectivity of Extreme-Valued Attributes: Stochastic Indicator Models of Reservoir Heterogeneities. 63rd Annual Technical Conference and exhibition, Soc. of Pet. Eng., Richardson, Texas. SPE paper 18324

König C (2002) SPRING®. Benutzerhandbuch Version 3.00. delta h Ingenieurgesellschaft mbH, Bochum

LaVenue Marsh A and RamaRao BS (1992) A Modelling Approach to Address Spatial Variability within the Culebra Dolomite Transmissivity Field, SAND92-7306, Sandia National Laboratories, Albuquerque.

Ludwig R (1994) Projekt Gorleben. Hydrogeologische Grundlagen für Modellrechnungen. Kenntnisstand 1994. BGR-Archiv No. 112 002. BGR Hannover, unpublished report

Ludwig R and Kösters E (2002) Hydrogeologisches Modell Gorleben – Entwicklung bis zum paläohydrogeologischen Ansatz. Schriftenreihe der Deutschen Geologischen Gesellschaft, Heft 24, Hannover

Marivoet J, Wemaere I, Escalier des Orres P, Baudoin P, Certes C, Levassor A, Prij J, Martens K-H, Röhlig, K-J (1997) The EVEREST project: sensitivity analysis of geological disposal systems. Reliability Engineering and System Safety 57, 79-90

Porter JD and Hartley LJ (1997) The Treatment of Uncertainty in Groundwater Flow and Solute Transport Modelling. Application of Indicator Kriging to Stratigraphic and Petrographic Data from the Gorleben Site. EUR 17829 EN, Luxembourg

Röhlig K-J (1999) Geostatistical Analysis of the Gorleben Channel. In: Gómez-Hernández *et al.* (ed) GeoENV II – Geostatistics for Environmental Applications. Proceedings of the Second European Conference on Geostatistics for Environmental Applications held in Valencia, Spain, November 18-20, 1998. Kluwer Academic Publishers, Dordrecht Boston London,319-330

Röhlig K-J and Pöltl B (2001) Uncertainty and Sensitivity Analyses for Contaminant Transport Models Based on Conditional Indicator Simulations. In: Monestiez P *et al.* (ed) GeoENV III – Geostatistics for Environmental Applications. Proceedings of the Third European Conference on Geostatistics for Environmental Applications held in Avignon, France, November 22-24, 2000. Kluwer Academic Publishers, Dordrecht Boston London, 251-262

Schelkes K, Knoop R-M, Geißler N (1990) INTRAVAL PHASE II. Test Case: Saline Groundwater Movement in an Erosional Channel Crossing a Salt Dome (Part 1). BGR, Hannover

Zimmerman DA *et al.* (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modelling advective transport by groundwater flow, Water Resour. Res., 34(6), 1373-1413

Random field approach to seawater intrusion in heterogeneous coastal aquifers: unconditional simulations and statistical analysis

A. Al-Bitar and R. Ababou

Institut de Mécanique des Fluides de Toulouse, Allée Camille Soula,
31400 Toulouse, France.

1 Introduction and summary

Seawater intrusion in coastal aquifers is a growing concern in mediterranean regions, due to over-population and over-exploitation of coastal groundwater resources. Under these circumstances, it is essential to model the extent of seawater intrusion and to locate the saltwater-freshwater interface taking into account heterogeneity and parameter uncertainty. There are different ways to couple salt transport and freshwater flow in groundwater models. We choose here the vertically integrated sharp interface approach, with two immiscible fluid regions (freshwater and seawater). We use this model to analyze the effects of aquifer variability on the saltwater wedge in plane view, based on large numerical simulations of 2D seawater intrusion in randomly heterogeneous unconfined aquifers.

2 Groundwater flow equations with seawater intrusion

We consider an unconfined coastal aquifer with an impervious bedrock at $z = Z_{INF}(x,y)$ and a fresh water table of elevation $z = Z_S(x,y)$. In addition, because we use a plane flow model, the vertically averaged freshwater hydraulic head $H(x,y)$ coincides with the free surface elevation, i.e. : $H(x,y) \approx Z_S(x,y)$. In this 2D framework, all variables and parameters are spatially distributed in (x,y). A schematic representation of the coastal aquifer and its salt wedge is shown in Fig. 1.

We assume that seawater and freshwater are separated by a sharp interface. More precisely, we rely on the Ghyben-Herzberg approximation(s), that is:

- the seawater and freshwater fluids are assumed immiscible (sharp interface);
- the subsurface seawater wedge is assumed quasi-hydrostatic;
- the freshwater is assumed vertically hydrostatic (negligible vertical velocities).

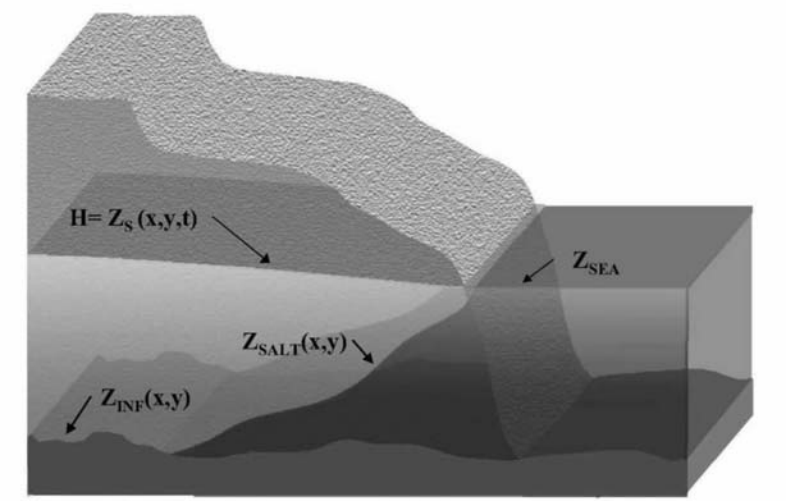


Fig. 1 Schematic view of seawater intrusion (sea level ZSEA shown at right) into a free surface aquifer (shown at left), with saltwater interface ZSALT(x,y) and substratum ZINF(x,y)

Now, let $Z_{SALT}(x,y)$ be the elevation of the salt/fresh water interface. Applying the hydrostatic assumptions and the pressure continuity condition at the interface, and modifying the Badon-Ghyben-Herzberg configuration to account for a finite outflow face of height ΔZ located undersea, we obtain:

$$\rho_F\, g\,(H - Z_{SALT}) = \rho_S\, g\,(Z_{SEA} - Z_{SALT} - \Delta Z) + \rho_F\, g\, \Delta Z \qquad (1)$$

This gives finally the desired closure relation:

$$Z_{SALT} = Z_{SEA} - \frac{(H - Z_{SEA})}{\varepsilon} - \delta Z \qquad (2)$$

In these equations, ρ_F is freshwater density, ρ_S is saltwater density, and ε is the saltwater-to-freshwater density contrast:

$$\varepsilon = \frac{\rho_S - \rho_F}{\rho_F} \approx 1/40 \qquad (3)$$

Parameter ΔZ is the vertical depth of the freshwater outflow face at the shoreline, assumed much smaller than aquifer thickness. It can be obtained from exact solution of seawater intrusion in a vertical slice (x,z) of a homogeneous confined aquifer, without depth-averaging. Here, ΔZ is about 0.77 m, compared to 30 m aquifer thickness.

For freshwater flow, we use the Dupuit-Boussinesq plane flow approximation. The freshwater thickness is defined as:

$$\eta(x,y) = H(x,y) - Z_{\inf}(x,y) \;\; or \;\; \eta(x,y) = H(x,y) - Z_{salt}(x,y) \qquad (4)$$

depending on spatial location (x,y), within the salt wedge *or* not. The freshwater transmissivity $T(x,y)$ is then inferred from freshwater thickness $\eta(x,y)$ as follows :

$$T(x,y) = K(x,y) \times \eta(x,y) \qquad (5)$$

Note that T is spatially variable *via* K and Z_{INF}, and also, nonlinear *via* the unknown variables H and Z_{SALT} on which it depends. Finally, we obtain the following system of vertically integrated flow equations (steady state case):

1. Steady-state mass conservation (freshwater):

$$\frac{\partial \Theta}{\partial t} = -div(\mathbf{Q}) \quad \text{where } \Theta \text{ is the water content} \tag{6}$$

2. Darcy's law (vertically integrated):

$$\mathbf{Q} = -T\left(H, Z_{SALT}, Z_{INF}, x, y\right)\mathbf{grad}\left(H\right) \tag{7}$$

3. Freshwater transmissivity:

$$T = \begin{cases} K(x,y)\times(H - Z_{INF}(x,y)) \; if \; Z_{SALT} < Z_{INF} \\ K(x,y)\times(H - Z_{SALT}(x,y)) \; if \; Z_{SALT} \geq Z_{INF} \end{cases} \tag{8}$$

3 Unconditional random aquifers (single replicates)

In this paper, we choose to study uncertainty *without* regard for specific data. That is, we choose to simulate seawater intrusion in large *unconditional* single-replicates of the heterogeneous aquifer. We use the XIMUL code to generate isotropic log-normal random fields K(x,y) on 1 million node grids (1000×1000).

The XIMUL code deals more generally with Bayesian estimation and conditional simulation of 1,2,3-D random functions of space or time (Ababou *et al.* 1994). The unconditional generator uses the Fourier Turning Band method based on a representation theorem of Matheron (1973): see (Tompson *et al.* 1989) and references therein.

4 Numerical solution with the BIGFLOW code

Numerical simulations of seawater intrusion are carried out using the BigFlow code BF 2000 (Ababou and Trégarot 2002). It solves a generalized model equation for flow in heterogeneous, anisotropic, partially saturated media. It can efficiently follow multiple interacting free surfaces in 3D, and it can represent "open" or "macroporous" media (Trégarot 2000). A vertically integrated 2D flow module is also available, including Boussinesq-Dupuit aquifer flows, free surface hydraulics based on kinematic-diffusive wave, Darcy-Forchheimer flow in rough fractures (Spiller 2004), *and* the seawater intrusion module SWIM2D used here.

The BF 2000 code is based on implicit 3D finite volume formulation of flux divergence equations in conservative form (mixed form). It solves fully coupled transient and steady flow problems, using a single infinite time step for steady state. It uses Preconditioned Conjugate Gradients for matrix solution, and modified Picard iterations for nonlinear solution. The matrix-vector data structure is very sparse. For more details on the numerics, see (Ababou *et al.* 1992; Ababou and Bagtzoglou 1993; Ababou 1996).

5 Simulation results and statistical analysis of salt wedge

5.1 Aquifer flow configuration and statistical inputs

We consider steady flow in a heterogeneous unconfined coastal aquifer in a square domain (1 km × 1 km). The mean freshwater flow is directed along the x axis. We apply constant head boundary conditions (Dirichlet) on boundaries orthogonal to mean flow: Z_{SEA}=30 m (seawater level) and H_1 = 31 m (freshwater inland). The hydraulic gradient, directed along (x), is 0.001 m/m, a typical value for regional flow in coastal regions. Lateral boundaries orthogonal to seashore are assumed impervious. Other statistical-geometric parameters concerning the planar grid and the random log-permeability field lnK (x,y) are shown in Table 1.

Table 1 Summary of statistical parameters for two sets of simulations (small and large)

Parameters	Set 1	Set 2
ni (number of nodes)	300 × 300	1000 × 1000
Δxi (discretization cell size) (m)	10/3	1
Li (domain length) (m)	1000	1000
λ (lnk- correlation scale) (m)	100/3	10
Δxi/λ (grid resolution)	1/10	1/10
Li/λ (sampling number)	30	100
ΔH/Lx (mean gradient)	1/1000	1/1000
σ (standard deviation of lnK)	1, $\sqrt{2}$, 1.6, 2, ln(10)	1, $\sqrt{2}$, 1.6, 2, ln(10)

A statistically isotropic log-normal random field K(x,y) was generated on a one million node grid (1000 × 1000 cells), with either smooth (gaussian) or noisy (exponential) covariance structure. A good fit was obtained when comparing theoretical vs computed spatial autocorrelation function of lnK(x,y) for the gaussian covariance with σlnK = 1 on a 1000 × 1000 grid. Smaller 300 × 300 fields were then extracted from the center of the domain, and single replicate simulations of seawater intrusion were conducted, with variability ranging from σlnK = 1 to ln10.

In order to increase numerical accuracy, we used an iterative continuation method (or homotopy method) with respect to the σ parameter, where σ is the standard deviation of lnK (degree of heterogeneity). Thus, the output of a heterogeneous problem is used as initial condition for simulating a "more heterogeneous" problem. This procedure adds an external loop to the flow solver. Mass balance errors, in terms of net discharge rate normalized by global outflow rate, did not exceed about 1%, for all simulations presented here.

5.2 Effect of heterogeneity level on mean salt wedge

Fig. 2 and Fig. **3** display perspective views of simulated seawater intrusion for a highly variable permeability ($\sigma_{lnK} = \ln 10 \approx 2.30$). Two surfaces are displayed in each figure : $Z_{SALT}(x,y)$, the salt/fresh interface level (mapped with color-coded logK values), and $Z_s(x,y)$, the freshwater piezometric surface (or hydraulic head), also mapped with the same color-coded or grey-scale logK values.

On Fig. 2, one can clearly observe the sharp local gradients of the saltwater interface occuring in low permeability zones, which act as barriers to seawater (it should be kept in mind, however, that K(x,y) is the depth-averaged permeability).

Fig. 4 depicts the effect of heterogeneity level on the mean penetration of the salt wedge, for a 300 × 300 grid. The mean $Z_{SALT}(x)$ profile is plotted versus distance from sea (x), after averaging $Z_{SALT}(x,y)$ along the shorewise direction (y). The three profiles correspond to: $\sigma_{lnK} = 0$ (homogeneous), $\sigma_{lnK} = 1$ (moderate heterogeneity) and $\sigma_{lnK} = \ln 10$ (high heterogeneity). As the level of variability σ_{lnK} increases, the mean elevation $Z_{SALT}(x)$ increases and the mean salt wedge penetrates farther inland. The *extra* penetration of the mean wedge due to heterogeneity is about 200 m, for the most heterogeneous case.

A similar result (*not shown here*) was obtained for the larger 1000 × 1000 grid with heterogeneity levels $\sigma_{lnK} = 0$, 1.0, $\sqrt{2}$, 1.6, and ln10. It confirms the monotonic increase of the mean penetration length of the salt wedge as σ_{lnK} increases, compared to a homogeneous aquifer with geometric mean permeability.

5.3 Statistical analysis of salt wedge fluctuations (1000 x 1000 grid)

As a first step towards uncertainty analysis (next section), let us develop further the statistical analysis of the simulated salt wedge, based on single replicate unconditional simulations obtained on the largest grid (1000 × 1000 cells).

The salt wedge is characterized by the shape of the saltwater interface elevation $Z_{SALT}(x,y)$ and its horizontal extension inland. We consider $Z_{SALT}(x,y)$ as a random field and we analyze it statistically. We focus in particular on the first and second order moments of $Z_{SALT}(x,y)$, including its mean and its standard deviation. This analysis is applied to the 1000x1000 grid, with large variability ($\sigma_{lnK} = \ln 10$).

Given the symmetries of the problem and the statistical stationarity of K(x,y), we expect the surface $Z_{SALT}(x,y)$ to be stationary (statistically homogenous) along the y direction parallel to the seashore. However, it will not be stationary along the x direction parallel to flow (transverse to seashore).

Indeed, Fig. 5 shows 100 transects $Z_{SALT}(x,yn)$ along with the average profile, all plotted as functions of the x-coordinate (perpendicular to sea shore). The profiles $Z_{SALT}(x)$ are clearly non-stationary.

6 Stochastic analysis of salt wedge via Φ-transform

It is clear from both Fig. 4 (mean Z_{SALT}) and Fig. 5 (random Z_{SALT}) that the interface elevation $Z_{SALT}(x,y)$ follows a nonlinear trend along x (for fixed y) and cannot be a *stationary* random function of x. This observation has two consequences:

1. Given a single replicate of the coastal aquifer in (x,y), we can only sample in the shorewise direction (y) to produce a statistical description of the salt wedge.
2. For theoretical purposes, we may seek a convenient transformation $Z_{SALT} \rightarrow \Phi$ to obtain an approximate stationary field Φ from the non-stationary field Z_{SALT}.

6.1 Transformation of Z_{SALT} into a potential Φ

Following this idea, consider first the analytical solution of the homogenous problem ($\sigma = 0$) using the Badon-Ghyben-Herzberg assumption, modified to include a submarine outflow face of height ΔZ at the seashore :

$$Z_{SALT}(x) = Z_{SEA} - \left(\frac{\varepsilon\,\Delta Z}{\varepsilon+1}\right) - \Delta Z - \sqrt{\frac{x}{\varepsilon\,L_X}\frac{H_1^2 - Z_{SEA}^2\,(\varepsilon+1)}{\varepsilon\,(\varepsilon+1)} + \left(\frac{\varepsilon\,\Delta Z}{\varepsilon+1}\right)^2} \qquad (9)$$

Eq. 9 holds for $0 \le x \le L_{SALT}$, where $x=L_{SALT}$ is the intersection of the saltwater interface with the substratum. Thus, if the bedrock is at z = 0, the value of L_{SALT} is defined by $Z_{SALT}(x) = 0$. Other variables in Eq. 9 are defined below:

- L_X is the domain size in the x-direction, between the two fixed head boundaries $H=H_0$ (sea at left) and $H=H_1$ (freshwater at right);
- $Z_{SEA} = H_0$ is the elevation and depth of the sea level above the substratum, at the sea boundary x=0;
- H_1 is the depth of the freshwater level above the substratum at the inland boundary $x=L_X$;
- Z_{SALT} is the elevation of saltwater/freshwater interface above the substratum;
- L_{SALT} is the x-wise penetration length of the salt wedge inland, on the bedrock;
- ΔZ is the vertical length of the submarine freshwater outflow face into the sea.

The term $(\varepsilon\,\Delta Z)/(\varepsilon+1)$ can sometimes be neglected in Eq. 9; we have here: $\varepsilon\,\Delta Z/(1+\varepsilon) = 0.018$ m , which is indeed small compared to ZSEA = 30 m and to ΔH=H1-ZSEA = 1 m. Also, to simplify the above expression Zsalt(x), we define a new parameter h_0:

$$h0 = \sqrt{\frac{H_1^{\,2} - ZSEA^2(\varepsilon+1)}{\varepsilon\cdot(\varepsilon+1)}} \qquad (10)$$

This parameter, h_0, is a length scale on the same order as the thickness of the freshwater lens imposed at the inland boundary (upstream). The solution of the homogenous problem $\sigma = 0$ can now be expressed as:

$$Z_{SEA} - \delta Z - Z_{SALT}(x) \approx \frac{h_0}{\sqrt{\varepsilon}} \cdot \sqrt{\frac{x}{L_X}} \tag{11}$$

This simple analytical expression (11) shows that there exists, for the homogeneous case, a *quadratic transform* which makes the saltwater profile exactly linear in *x*. For the heterogeneous case, with random field permeability, this suggests applying the same quadratic transform to the nonlinear random function $Z_{SALT}(x,y)$. The transformed field is a new random "potential" field $\phi_{SALT}(x,y)$ with:

$$\Phi_{SALT} = (Z_{SEA} - \delta Z - Z_{SALT})^2 \tag{12}$$

We may expect that the random $\Phi_{SALT}(x,y)$ has a roughly linear trend. Furthermore, it is possible to derive *analytically* the mean and variance of Z_{SALT} from the moments of the random field $\Phi_{SALT}(x,y)$. Let us first normalize Z_{SALT} and Φ_{SALT} by Z_{SEA} as follows: $Z=(Z_{SALT}-\delta Z)/Z_{SEA}$; $\phi = \Phi_{SALT}/Z_{SEA}$. The Φ-transform is now:

$$\phi(x,y) = (1 - Z(x,y))^2 \tag{13}$$

with $\phi = 0$ (exactly) on the sea boundary $x = 0$, and $\phi = 1$ at some fixed distance L_1, the characteristic length of penetration of the salt wedge. The latter is given, to order $O(\sigma)$, by the analytical solution for a homogeneous aquifer:

$$L_1 = L_{SALT}(\sigma) \approx L_{SALT}(0) \times (1 + O(\sigma)) \tag{14}$$

Thus, we may write the (approximate) boundary condition of the random case as:

$$x = 0 : \phi = 0; \quad x = L_1 : \phi \approx 1 + O(\sigma) \tag{15}$$

The main idea, here, is that we prefer to solve for the Φ-field because it is more easily amenable to statistical analysis than the Z-field (more on this below). With this goal in mind, let us define the random fluctuations of ϕ and Z:

$$\varphi(x,y) = \phi(x,y) - \overline{\phi}(x) \text{ and } z(x,y) = Z(x,y) - \overline{Z}(x) \tag{16}$$

where the mean potential is given by: $\overline{\phi}(x) = \langle (1 - Z)^2 \rangle$.

The brackets <•> represent either the shorewise spatial average (spatial mean of a single replicate along direction "y"), or the mathematical expectation E(•) over an ensemble of replicates : the two are equivalent if ergodicity is assumed.

Now, substituting the random fluctuations in Eq. 13 and taking averages, we obtain:

$$\overline{\phi} = \left(1 - \overline{Z}\right)^2 + \langle z^2 \rangle \text{ and } {\sigma_Z}^2 = \overline{\phi} - \left(1 - \overline{Z}\right)^2 \tag{17}$$

where the mean <Z> remains to be determined. On the other hand, from Eq. 13:

$$Z = 1 - \phi^{1/2} \text{ (for } 0<x<L_1 \text{ and } 1>Z>0). \tag{18}$$

$$\overline{Z} = 1 - \overline{\phi}^{1/2} \left\langle (1 + \kappa)^{1/2} \right\rangle \text{ with } \kappa = \frac{\varphi}{\overline{\phi}} < 1 \tag{19}$$

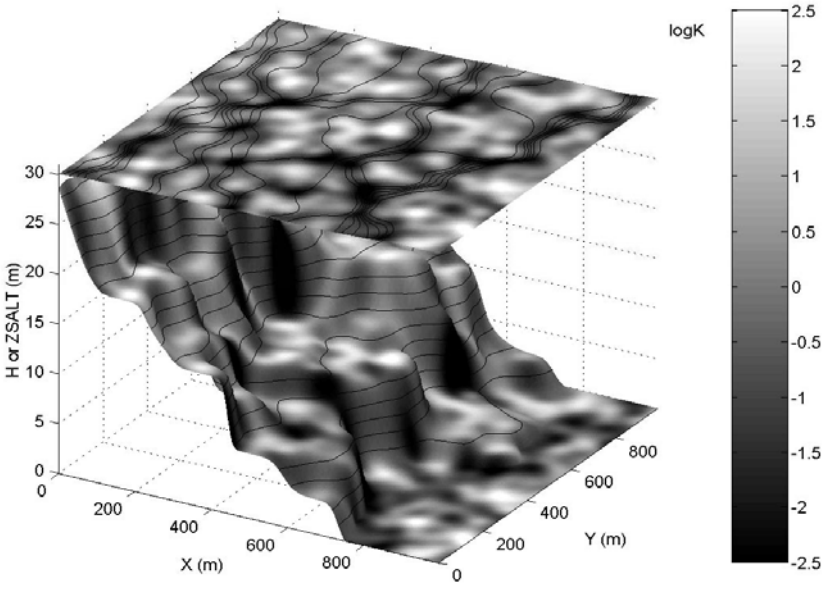


Fig. 2 Perspective view of Z_{SALT} (x,y), H(x,y), and log K(x,y) for a gauss-shaped isotropic covariance with $\sigma = \ln 10$ and $L/\lambda = 30$. Simulation grid: 300×300

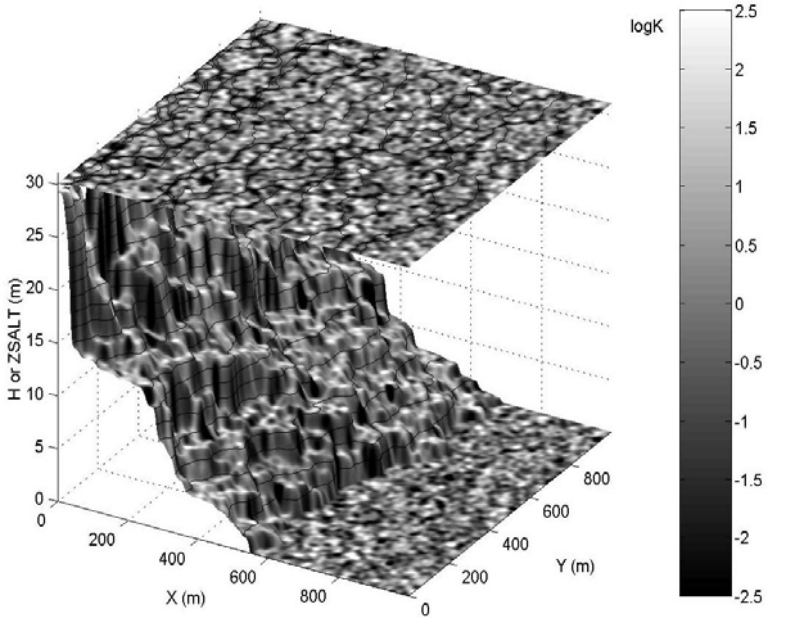


Fig. 3 Perspective view of Z_{SALT} (x,y), H(x,y), and log K(x,y) for a gauss-shaped isotropic covariance with $\sigma = \ln 10$ and $L/\lambda = 100$. Simulation grid: 1000×1000

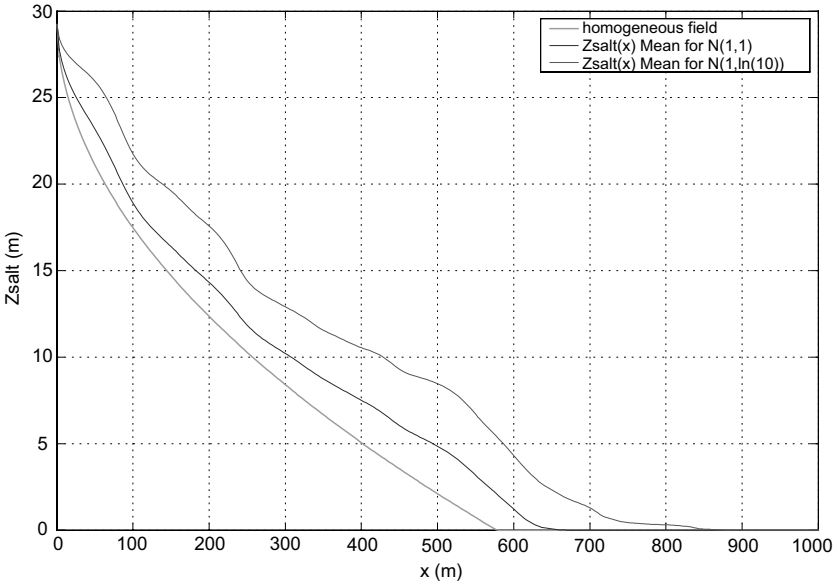


Fig. 4 Mean $Z_{SALT}(x)$ profile transverse to seashore for a 300×300 grid : analytical solution for $\sigma = 0$ and computed mean profiles for $\sigma = 1.0$ to ln10 (Z_{SALT} increases with σ)

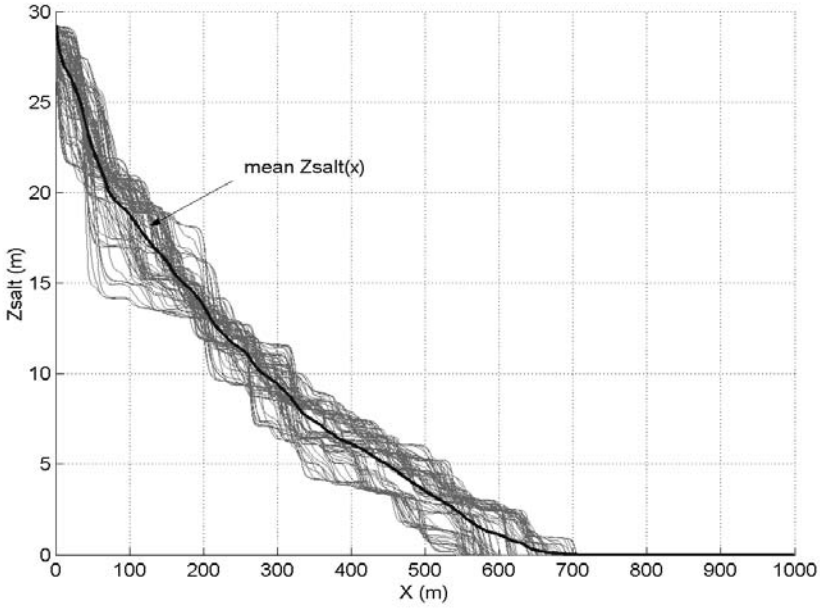


Fig. 5 Transverse profiles of $Z_{SALT}(x)$ (the seashore is at left) : comparison of mean Z_{SALT} (x) (shorewise average) with 100 distinct transects of $Z_{SALT}(x,y_n)$ sampled at equally spaced shorewise positions (y_n). Simulation grid : 1000x1000. Heterogeneity: σ = ln10

Using a Taylor expansion cut off to 2nd order, yields:

$$\overline{Z} \approx 1-\overline{\phi}^{1/2}\left(1-\frac{1}{8}\tau^2+O\left(\left\langle \kappa^3 \right\rangle\right)\right);\ \tau=\frac{\sigma_\phi}{\langle\phi\rangle};\ \kappa=\frac{\varphi}{\langle\phi\rangle} \tag{20}$$

We finally substitute Eq. 20 into Eq. 17 to calculate the standard deviation of Z. Neglecting $\tau^4/64$ and other "*higher order terms*" *("h.o.t.*"), we obtain:

$$\sigma_Z \approx \frac{1}{2}\frac{\sigma_\phi}{\sqrt{\overline{\phi}}} + h.o.t. \tag{21}$$

This analytical expression can be used to predict σ_Z using either numerical estimates or theoretical spectral estimates of ϕ-statistics : the two procedures yield similar results (see comments about Fig. 6 further below).

6.2 Statistics of transformed potential via spectral theory

We know need to determine the statistical moments of Φ, e.g. mean and variance. Two approaches are possible concerning the transformed potential Φ:

a) *Empirical* evaluation of Φ-moments (sampling numerical simulation);
b) *Theoretical* evaluation of Φ-moments (analytical spectral perturbation).

Empirically, the first two lines in Table 2 show some of the numerically computed moments of Φ_{SALT} , assuming a linear trend $<\Phi>$, and stationary fluctuations $\varphi(x,y)$ around the linear trend:

$$\left\langle \Phi(x,y) \right\rangle = \overline{\Phi}(x) \approx ax \quad \varphi(x,y) \approx \Phi(x,y) - ax \tag{22}$$

$$\sigma_\Phi = \left\langle \varphi(x)^2 \right\rangle^{1/2} \approx \textbf{constant} \tag{23}$$

Note: These relations hold only in a subdomain comprised between the sea boundary $x = 0$ (where $\phi = 0$) and the tip of the salt wedge $x \approx L_1+O(\sigma)$ (where $\phi \approx 1+O(\sigma)$).

On the other hand, we demonstrate that the Φ-equation in the salt wedge zone is a stochastic PDE, analogous to the Boussinesq equation for vertically averaged groundwater flow with random K(x,y). Indeed, from Eq. 5, 6 and 7, we have:

$$\frac{\partial}{\partial x_i}\left(-K(x_1,x_2)\,(H-Z_{SALT})\frac{\partial H}{\partial x_i}\right)\ \textit{(i=1,2)} \tag{24}$$

The freshwater head H is given by the Ghyben-Herzberg relation Eq.(2):

$$H = (1+\varepsilon)\,Z_{SEA} - \varepsilon\,\Delta Z - \varepsilon\,Z_{SALT} \tag{25}$$

Substituting H in Eq. 24, and using the Φ-transform, we obtain:

$$\frac{\partial}{\partial x_i}\left(-K(x_1,x_2)\,\frac{\partial \phi}{\partial x_i}\right) = 0\ \textit{(i=1,2)} \tag{26}$$

We observe that this ϕ-equation is equivalent to a stochastic groundwater flow equation with 2D random field transmissivity in a confined aquifer (cf. "infinite

domain" spectral perturbation solutions by (Mizell *et al.* 1982)). Thus, σ_Φ can be evaluated from the spectral solution of Eq. 26, at least far enough from the sea and the saltwedge tip. The "theoretical" standard deviation of ϕ is deduced from the Mizell *et al.* (1982) solution, for a "modified Wittle" correlation structure:

$$(\sigma_\Phi)_{THEORY} \approx c\ \sigma_{\ln K}\ \lambda_{\ln K}\ J_x\ \approx c\ \sigma_{\ln K}\ \lambda_{\ln K}\ a \tag{27}$$

where J_x is the mean Φ-gradient denoted "*a*" in this paper. The coefficient "c" is a dimensionless constant of order 0(1) (Mizell *et al.* 1982). For the problem at hand, the value of "c" can be obtained by matching numerical and theoretical "σ_Φ" at low levels of heterogeneity ($\sigma_{\ln K} \leq 1$). This procedure gives:

$$c \approx 1.10 \tag{28}$$

Similarly, the relevant value of the mean Φ-gradient, $a = \langle d\Phi/dx \rangle$, can be obtained from the exact analytical solution $\Phi(x)$ in a homogeneous aquifer, which corresponds to the asymptotic case $\sigma_{\ln K} \to 0$. Thus, asymptotically:

$$\sigma \to 0: \quad (a)_{THEORY} = a_0 + O(\sigma) = \frac{h_0^2}{\varepsilon\ L_x} + O(\sigma) \tag{29}$$

Table 2. Empirical and theoretical moments of the transformed potential ΦSALT(x,y).

$\sigma_{\ln K}$	0	1	1.60	2.30
$\hat{a}_{NUM} = \left\langle d\Phi / dx \right\rangle$	$a = 1.54$	$\hat{a} \approx 1.54$	$\hat{a} \approx 1.40$	$\hat{a} \approx 1.33$
$\hat{\sigma}_{\Phi\ NUM}$	$\sigma_\Phi = 0$	$\hat{\sigma}_\Phi \approx 17$	$\hat{\sigma}_\Phi \approx 27$	$\hat{\sigma}_\Phi \approx 42$
$\hat{\sigma}_{\Phi\ THEORY}$	$\sigma_\Phi = 0$	$\sigma_\Phi \approx 17$	$\sigma_\Phi \approx 27.2$	$\sigma_\Phi \approx 39.1$

To check whether "*a*" is nearly constant and close to its predicted value "a_0", consider the results summarized in Table 2. We conclude that the theoretical prediction of σ_Φ given by Eq. 27 with $a \approx a_0$ is robust.

Finally - after some manipulations involving statistics from the Z-Φ transform (Eq. 20 and 21) and the spectral solution for σ_Φ - one obtains, to first order:

$$\text{(a)}: \sigma_{Z_{SALT}}(x) \approx \frac{c}{2} \sigma_{\ln K}\ \lambda_{\ln K} \sqrt{\frac{a}{x}} \ \text{ or (b)}: \sigma_{Z_{SALT}}(x) \approx \frac{c}{2} \sigma_{\ln K}\ \lambda_{\ln K} \left| \frac{\partial \bar{Z}}{\partial x} \right| \tag{30}$$

Both versions of this equation require mean gradient information: the first equation (a) requires knowledge of the (stationary) mean ϕ-gradient "*a*", while the second version (b) requires knowledge of the (non-stationary) mean interface elevation gradient.

6.3 Numerical moments of seawater of interface and comparisons

Fig. 6 shows 100 superimposed transects of the "potential" $\Phi_{SALT}(x,y_n)$, sampled at equally spaced shorewise locations "y_n", and plotted versus (*x)*, for $\sigma = \ln 10$. The figure also shows the analytical profile $\Phi_{SALT}(x)$ for a homogeneous aquifer ($\sigma = 0$), as well as the numerical average of $\Phi_{SALT}(x,y)$. The fluctuations of $\Phi_{SALT}(x,y)$ around its mean trend were also plotted as transects (*not shown here*). These numerical plots indicate the level of fluctuation of the salt interface in terms of the transformed field $\Phi_{SALT.}$ They also confirm the quasi-linear trend of Φ_{SALT}.

We computed the fluctuations of Z_{SALT} around its nonlinear mean trend, and we estimated σ_Z by sampling Z_{SALT} parallel to the seashore and plotting the resulting moment σ_Z as a function of distance (*x)* from the sea. One result is shown in Fig. 7 for *large heterogeneity (σ=ln10)*. The standard deviation of Z_{SALT} seems approximately stationary far enough from the seashore ($x = 0$) and far enough from the salt wedge tip ($x \approx 700$ m). In the stationary region of Fig. 7, we find $\sigma_{ZSALT} \approx 1.3$ m. The 95% confidence band of the salt interface is several meters, which represents a rather significant fraction of the total aquifer thickness of 30 m.

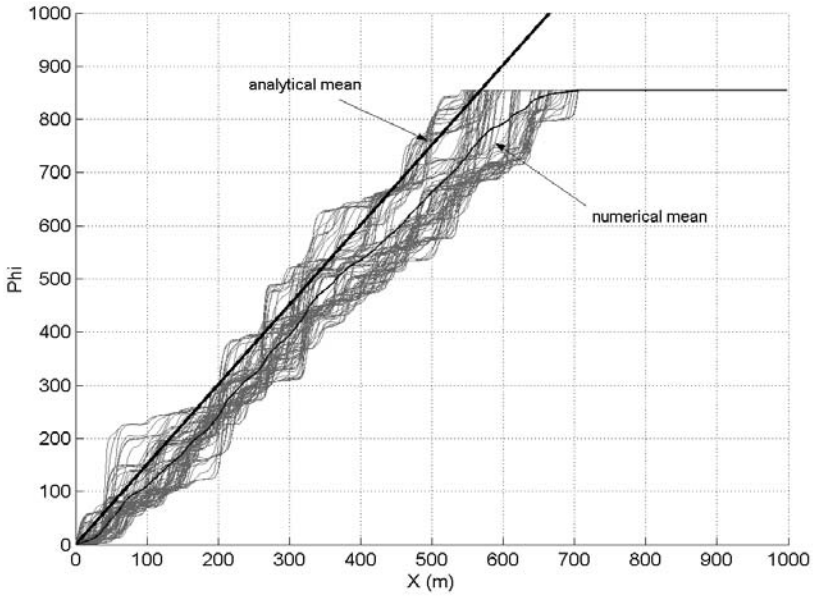


Fig. 6 One hundred transects of ΦSALT (transformed from ZSALT); "analytical mean" curve ΦSALT (homogeneous aquifer); and "numerical mean" curve ΦSALT (mean of ΦSALT sampled shorewise along "y"). The sea shore is at left. Grid: 1000x1000. Heterogeneity: $\sigma = \ln(10)$

The results appear different for *lesser heterogeneity*: see Fig. 8 for $\sigma = 1.60$, and note that similar results were obtained for σ in the range $0 \le \sigma \le 2.0$. In all these cases, $\sigma_{ZSALT}(x)$ is non-stationary with respect to (x) and decreases with (x),

as predicted by the theoretical Φ-transform analysis. This can be seen by comparing the "numerical" and "analytical" (Eq. 30a) $\sigma_{Z_{SALT}}(x)$ curves in Fig. 8.

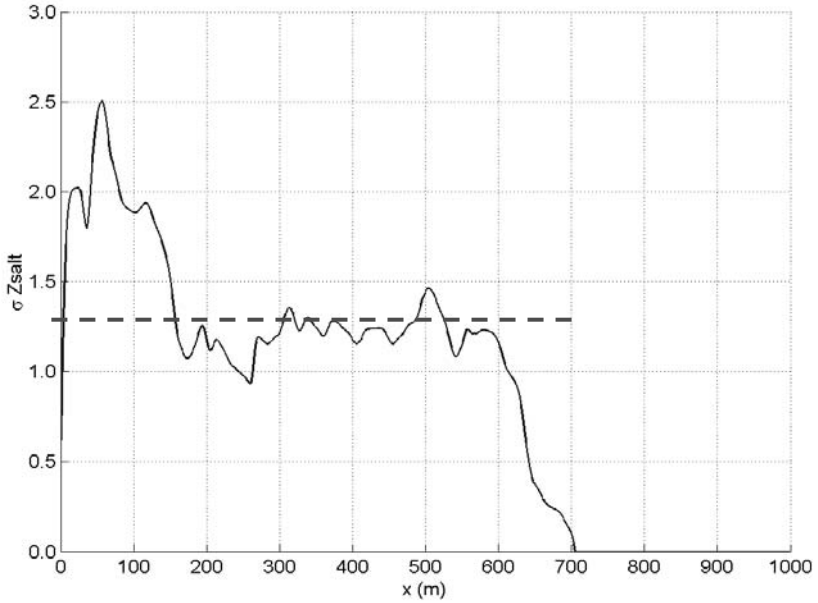


Fig. 7 Standard deviation of Z_{SALT} *vs.* distance (x) from seashore (sea located at left), obtained by sampling Z_{SALT} fluctuations in the shorewise direction (y). The global value of σ_Z appears to be about $\sigma_Z \approx 1.3$ m. Grid: 1000x1000 cells. Heterogeneity: $\sigma_{lnK} = \ln 10 = 2.30$

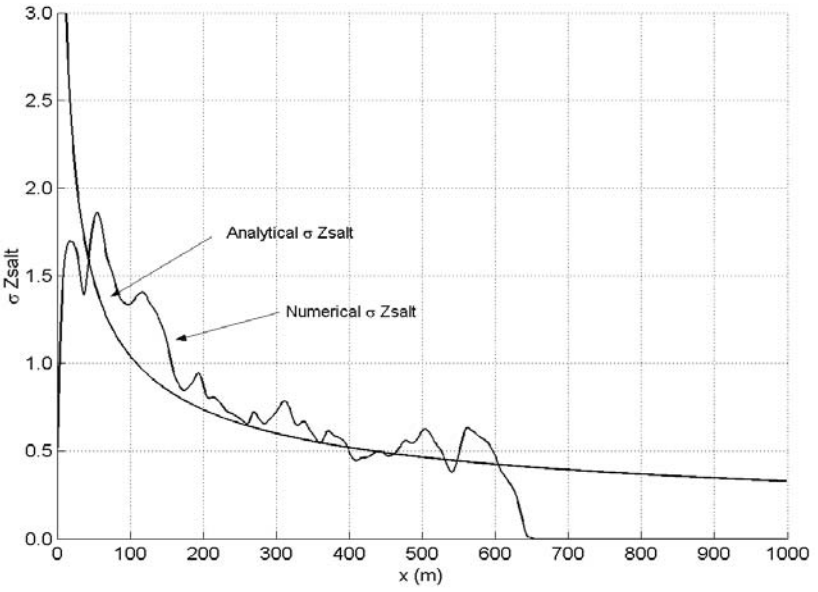


Fig. 8 Numerical and theoretical $\sigma_{Z_{SALT}}$ *vs* distance from sea (x) for $\sigma_{lnK} = 1.60$

7 Summary and conclusions

We have presented numerical experiments of seawater intrusion based on unconditional simulations of random permeability fields K(x,y), where K(x,y) represents a depth-averaged permeability. The effects of planar heterogeneity on the extent and shape of the salt wedge were discussed, and we presented a statistical study of interface elevation $Z_{SALT}(x,y)$ on a 1 million node grid (single replicates). The statistic (σ_{ZSALT}) can be viewed as the *root-mean-square* vertical uncertainty of the seawater interface due to heterogeneity. It is found to be more or less proportional to the mean gradient of Z_{SALT}, at least for low and moderate variability, which validates our perturbation theory. However, for higher variability, the observed standard deviation of Z_{SALT} tends to a constant value more or less independent of x, which is not quite reproduced by the same perturbation theory. Overall, our results indicate that σ_{ZSALT} can be typically on the order of several meters.

Acknowledgment

This study is part of the european project SWIMED on coastal aquifer management, funded by the European Commission (Sustainable Water Management In MEDiterranean coastal aquifers): http://www.crs4.it/EIS/SWIMED

References

Ababou R (1996) Random Porous Media Flow on Large 3D Grids: Numerics, Performance, and Application to Homogenization, Chap.1, 1-25. In: IMA Vol 79 Mathematics and its Applications: Environmental Studies (Math. Comput. Statist. Anal.). Wheeler MF (ed.), Springer, NY, p. 410

Ababou R, Bagtzoglou AC (1993) BIGFLOW: a Numerical Code for Simulating Flow in Variably Saturated, Heterogeneous Geologic MediaTheory and User's Manual Ver.1.1. NUREG/CR-6028, US NRC Report, Washington DC

Ababou R, Bagtzoglou AC, Wood EF (1994) On the Condition Number of Covariance Matrices Arising in Kriging, Estimation & Simulation of Random Fields. Math.Geol.26(1), 99-133, 1994

Ababou R, Trégarot G (2002) Coupled Modeling of Partially Saturated Flows : Macro-Porous Media, Interfaces, and Variability. Proc. CMWR 02, Comput Meth Water Resour, 23-28 June 2002, Delft, The Netherlands, Elsevier, p. 8

Ababou R, Sagar B, Wittmeyer G (1992) Testing Procedures for Spatially Distributed Flow Models. Advances in Water Resources, Vol.15, 181-198

Matheron G. (1973) The Intrinsinc Random Functions & Applications. Adv.Appl.Prob., 5, 439-468

Mizell SA, Gutjahr AL, Gelhar LW (1982) Stochastic Analysis of Spatial Variability in Two-Dimensional Steady Groundwater Flow Assuming Stationary and Nonstationary Heads. Water Resour Res 18(4) 1053-1067

Spiller M (2004) Physical and Numerical Experiments of Flow and Transport in Heterogeneous Fractured Media : Single Fracture Flow at High Reynolds and Reactive Particle Transport. PhD thesis, Aachen Univ. (Germany) & Institut Nat. Polytech. Toulouse (France), October 2004

Tompson AFB, Ababou R, Gelhar LW (1989) Implementation of the Three-Dimensional Turning Bands Random Field Generator. Water Resour. Res., 25(10), 2227-2243

Trégarot G (2000) Modélisation Couplée des Ecoulements à Saturation Variable avec Hétérogénéités, Forçages, et Interfaces Hydrologiques. PhD thesis, Institut Nat. Polytech. Toulouse, May 2000.

Uncertainty estimation of well catchments: semi-analytical post-processing

F. Stauffer and H.-J. Hendricks Franssen

Institute of Hydromechanics and Water Resources Management, ETH Zurich, CH-8093 Zurich, Switzerland

1 Objectives

Regulations for the protection of drinking water wells require the designation of the recharge area or catchment of wells. Very often in practice only limited information is available for their delineation. Therefore, we may ask: How uncertain are well catchments resulting from deterministic groundwater modeling? Stauffer *et al.* (2002) formulated a first-order, unconditional semi-analytical Lagrangian method, which allows to approximately evaluate the uncertainty in the location of two-dimensional, steady state catchments of pumping wells due to the uncertainty of the spatially variable hydraulic conductivity field. They applied their method successfully to a set of simple rectangular flow configurations. Stauffer *et al.* (2004) extended this method by incorporating conditioning by transmissivity and head measurements in observation wells.

In this paper we investigate the effect of conditioning by transmissivity data from observation wells alone. The uncertainty bandwidth of the catchment boundary is approximated in first order by formulating the conditional transversal second moment of the tracer particle displacements along the expected mean catchment boundary. Special relationships have to be developed for the estimation of the uncertainty in the location of the stagnation point. Applications of the approach are presented for a synthetic test case with four different arrangements of a total ten measurement locations. The results are compared with the results from conditional numerical Monte Carlo simulations. The comparison should allow an assessment of the accuracy, and the applicability of the method. Moreover, it should enable an assessment of the effect of conditioning on the reduction of uncertainty, which is achieved by the chosen observation networks.

The following assumptions are adopted for this paper for simplicity and/or feasibility reasons:

- The flow field can be modeled as a horizontal plane system according the flow equation $\nabla\cdot[T(\mathbf{x})\nabla h(\mathbf{x})]+P(\mathbf{x})=0$, where $T(\mathbf{x})$ is the transmissivity, which is variable in space with location $\mathbf{x}$, $h(\mathbf{x})$ is the hydraulic head, and $P(\mathbf{x})$ is a source/sink term, which includes the effect of areal recharge of rate N, and/or of the pumping rate Q_w of wells. For unconfined aquifers the above equation is often referred to as linearized equation, or as semi-confined model.

- The flow domain is a rectangular region, which is characterized by two parallel boundaries with prescribed head, and two impermeable boundaries.
- The pumping rate Q_w of the well is constant.
- The areal recharge rate N is constant and is homogeneously distributed over the domain, which contains the capture zone and the catchment of the well.
- The porosity n of the aquifer is constant.
- The spatial variability of transmissivity $T(\mathbf{x})$ can be described by an exponential covariance function of the form $C_Y(\mathbf{r}) = \sigma_Y^2 \exp(-|\mathbf{r}|/I_Y)$, with $Y(\mathbf{x})=ln(T(\mathbf{x}))$ and the variance σ_Y^2 and correlation length I_Y (integral scale).
- The ensemble mean flow field is approximated by that for equivalent homogeneous transmissivity T_g (geometric mean). This assumption may lead to deviations close to wells.
- Only advective transport mechanisms are considered thus neglecting local dispersion and molecular diffusion.
- The velocity covariance can be locally approximated by a scaling procedure based on the analytical first order approximation for uniform mean flow conditions according to Dagan (1989).

Note that the method can also be formulated for more general conditions.

The prerequisite for the uncertainty analysis is to first establish a deterministic steady state groundwater model for equivalent transmissivity. Then the second task consists of finding the boundary streamlines of the well catchment, which usually start at a stagnation point S. The location of S can be easily found by an appropriate numerical procedure searching for the location with minimum (zero) velocity. The number of relevant stagnation points of a well flow field depends on the prevailing flow conditions.

2 Lagrangian approximation of the uncertainty in the location of well catchment boundaries

2.1 Unconditional expected location of a well catchment

The aim is to formulate the uncertainty in the location of the well catchment boundary by considering the trajectory $\mathbf{X}(t)$ starting at the stagnation point S and proceeding upstream. The ensemble mean trajectory is $<\mathbf{X}(t)>$ and is determined for the flow field calculated with equivalent transmissivity T_g. For constant areal recharge rate N and constant thickness H of the aquifer the lateral second moment of the particle displacements $X'_{pp}(t)$ along $<\mathbf{X}(t)>$ can be approximated by the integral (Stauffer *et al.* 2002):

$$X'_{pp}(t) \approx \frac{1}{\left[U_l(L(t))\right]^2} \int_0^{L(t)} \int_0^{L(t)} u_{pp}(l',l'') \; exp\left[\frac{N}{nH} \cdot \left(2t(L) - t(l') - t(l'')\right)\right] dl' dl'' \qquad (1)$$

The symbols l and p denote the longitudinal and lateral directions along and perpendicular to the mean flow direction, $L(t)$ is the end position along the mean trajectory, n is the porosity, and $U_l(l)$ is the longitudinal velocity along the mean trajectory. The function $u_{pp}(l',l'')$ is the transversal velocity covariance.

2.2 Conditional expected location of a trajectory or well catchment

As a result of the conditioning of the random velocity field by measured transmissivity data both the ensemble mean and the second moment of the particle displacements may be affected. The conditional expected particle displacement is approximated by:

$$\left\langle \mathbf{X}^{cond}\left(t\right)\right\rangle \approx \mathbf{x}_0^{cond} + \int_0^t \mathbf{U}^{cond}\left(\mathbf{X}^{cond}\left(t'\right)\right)dt' \tag{2}$$

based on the conditional ensemble mean velocity $\mathbf{U}^{cond}(\mathbf{x})$. Accordingly, the conditional lateral second moment of the particle displacements may be approximated by Eq. 1 using the conditional transversal velocity covariance function $u_{pp}^{cond}\left(l',l''\right)$ instead of the unconditional one. For a well catchment the trajectory starts at the conditional expected mean stagnation point S and the lateral conditional second moment $X_{pp,s}^{cond}$ of the location of the stagnation point has to be added to the integral in Eq. 2.

For given n_Y transmissivity measurements $T_i(\mathbf{x}_i)$, i=1,..n_Y, the conditional expected mean velocity $\mathbf{U}^{cond}(\mathbf{x})$ and covariance function $u_{pp}^{cond}\left(\mathbf{x},\mathbf{x}'\right)$ can be determined by the method of conditional probabilities (Rubin 1991). The conditional mean velocity component $U_i(\mathbf{x})$ gets:

$$\left\langle U_i^{cond}\left(\mathbf{x}\right)\right\rangle = \left\langle U_i\left(\mathbf{x}\right)\right\rangle + \sum_{m=1}^{n_Y}\lambda_m\left(\mathbf{x}\right)\cdot\left(Y_{meas}\left(\mathbf{x}_m\right) - \left\langle Y\right\rangle\right); \qquad i = l, p \tag{3}$$

The conditional lateral velocity covariance $u_{pp}^{cond}\left(\mathbf{x},\mathbf{x}'\right)$ is:

$$u_{pp}^{cond}\left(\mathbf{x},\mathbf{x}'\right) = u_{pp}\left(\mathbf{x},\mathbf{x}'\right) - \sum_{m=1}^{n_Y}\lambda_m\left(\mathbf{x}\right)C_{U_pY}\left(\mathbf{x}',\mathbf{x}_m\right) \tag{4}$$

The weight coefficients λ_m are determined by the following set of equations:

$$\sum_{m=1}^{n_Y}\lambda_m\left(\mathbf{x}\right)C_{YY}\left(\mathbf{x}_m,\mathbf{x}_k\right) = C_{U_iY}\left(\mathbf{x},\mathbf{x}_k\right); \quad k = 1, n_Y \tag{5}$$

The unconditional lateral velocity covariance $u_{pp}(\mathbf{x}, \mathbf{x}')$ is analytically approximated for quasi-uniform flow and scaled according to Stauffer *et al.* (2002):

$$\tilde{u}_{pp}(\mathbf{r}=\mathbf{x}-\mathbf{x}') \approx u_{pp}(\mathbf{x},\mathbf{x}')/\left(U\left(\mathbf{x}\right)U\left(\mathbf{x}'\right)\sigma_Y^2\right) \tag{6}$$

The scaled covariance is evaluated according to Rubin (1990).

2.3 Unconditional location uncertainty of the stagnation point

The approximate ensemble mean location of the stagnation point S is easily found for equivalent transmissivity. However, we are not aware of a theoretical approach to formulate the uncertainty in the location of the stagnation point. Therefore an empirical approach is suggested to fill the gap. Such a procedure may be based on principles of dimension analysis.

Consider stagnation point S_1 in Fig. 1. Provided that the distances to the boundaries are large enough and not relevant, the lateral variance $X_{pp,S}$ of the location of the stagnation point depends essentially on L, σ_Y^2, I_Y, Q_w, and N, where L is a length scale. Comparison with various Monte Carlo simulations resulted in the following (crude) approximation for $X_{pp,S}$ for stagnation point S_1:

$$X_{pp,S} \approx 0.0043 \frac{Q_w}{N} \cdot \left(\sigma_Y^2\right)^{2/3} \exp\left(-0.09 \frac{\sqrt{Q_w / N}}{I_Y}\right) \tag{7}$$

For stagnation point S_2 at the water divide in Fig. 1, the relevant length scale is expected to be the distance D from S_2 to the nearest prescribed head boundary, in this case the eastern boundary. This holds true if the recharge rate N is of minor importance and the distance to the well is large compared to S_1. Therefore, a (crude) approximation for $X_{pp,S}$ for S_2 may be:

$$X_{pp,S} \approx 0.0115 \cdot \left(\sigma_Y^2\right)^{2/3} D^2 \exp\left(-0.05 \frac{D}{I_Y}\right) \tag{8}$$

Eq. 7 and 8 need that $N>0$. In the case of $N=0$, a similar equation can be found. In an approximate manner these relations may also be applicable to non-rectangular conditions.

2.4 Conditional expected location of the stagnation point

A conditional variance $X_{pp,S}^{cond}(L, \sigma_Y^2, I_Y)$ may be obtained by conditioning the variance $\sigma_Y^2(\mathbf{x})$ in Eq. 7 and 8, neglecting an influence of conditioning on the correlation length I_Y. Conditioning the variance $\sigma_Y^2(\mathbf{x})$ at a location $\mathbf{x}$ can be accomplished by kriging, given n_Y transmissivity measurements. The kriging system is given by (de Marsily 1986):

$$\sum_{m=1}^{n_Y} \lambda_m(\mathbf{x}) C_Y(\mathbf{x}_m, \mathbf{x}_k) = C_Y(\mathbf{x}, \mathbf{x}_k); \quad k = 1, n_Y \tag{9}$$

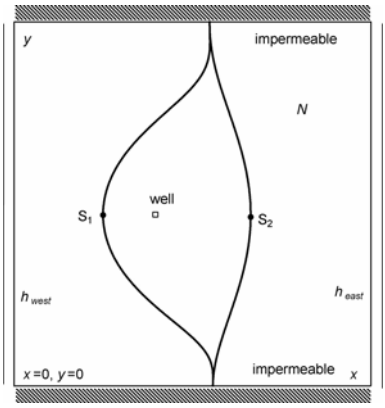


Fig. 1. Boundaries of the catchment of a pumping well for equivalent transmissivity and uniform recharge; they are trajectories starting at stagnation points S_1 and S_2.

The resulting conditional variance $\sigma^2_{Y,cond}(\mathbf{x})$ at location **x** is:

$$\sigma^2_{Y,cond}(\mathbf{x}) = \sigma^2_Y(\mathbf{x}) - \sum_{m=1}^{n_Y} \lambda_m(\mathbf{x}) C_Y(\mathbf{x}, \mathbf{x}_m) \tag{10}$$

Since the possible stagnation points of the realizations are distributed close to the conditional expected mean location, the variance $\sigma^2_{Y,cond}(\mathbf{x}_S)$ is computed as weighted average within this distribution assuming Gaussian distribution of the location of all possible stagnation points.

3 Numerical evaluation of flow related covariance functions

The two-point covariance relations needed are $C_{U_lY}(\mathbf{x}_i,\mathbf{x}_j)$ and $C_{U_pY}(\mathbf{x},\mathbf{x}_i)$, given the covariance $C_{YY}(\mathbf{x},\mathbf{x'})$. Since the desired locations are not known in advance the covariance relations are evaluated for a regular finite difference grid. For all pairs of the n_{cells} cell centers the two-point covariance $C_{YY}(\mathbf{x}',\mathbf{x}'')$ can be expressed in matrix form $[\mathbf{C}_{YY}]$ as a matrix of size n_{cells} *by* n_{cells}. The two-point covariances $C_{U_lY}(\mathbf{x}_i,\mathbf{x}_j)$ and $C_{U_pY}(\mathbf{x},\mathbf{x}_i)$ can be approximately determined by differentiation (Zhang 2002):

$$C_{U_lY}(\mathbf{x},\mathbf{x}') = U_l C_{YY}(\mathbf{x},\mathbf{x}') - \frac{T_g}{n}\frac{\partial}{\partial l} C_{hY}(\mathbf{x},\mathbf{x}') \tag{11}$$

and

$$C_{U_p Y}(\mathbf{x},\mathbf{x}') = -\frac{T_g}{n}\frac{\partial}{\partial p} C_{hY}(\mathbf{x},\mathbf{x}') \tag{12}$$

where T_g is the geometric mean transmissivity, and n the porosity. The matrix corresponding $[\mathbf{C}_{hY}]$ can be approximated in first order by (Zhang 2002):

$$[\mathbf{C}_{hY}] = [\mathbf{C}_{YY}] \cdot [\mathbf{D}_{Yh}^T] \tag{13}$$

where $[\mathbf{D}_{Yh}]$ is the sensitivity matrix with elements $\partial h_i / \partial Y_j$. The sensitivity matrix can be evaluated for the above mentioned finite difference grid with the help of the adjoint state method (Zhang 2002).

4 Numerical Monte Carlo simulation of conditional expected location of a well catchment

The expected location of a well catchment based on given transmissivity data is numerically analyzed using a Monte Carlo based method (Gómez-Hernández and Journel 1993). The analysis consists of the following steps:

1. Multiple equally likely log-transmissivity realizations are generated conditioned to the log-transmissivity measurements.
2. The groundwater flow equation is solved for each of the generated transmissivity fields. In its current implementation, these equations are solved by block-centered finite differences.
3. For each of the realizations a particle is released at the center of a grid cell and it is recorded whether the particle is captured by the pumping well. Averaging over the ensemble of realizations yields the probabilistic well catchment.

5 Test case

The methodology was applied in a synthetic study to investigate the uncertainty of the determination of a well catchment. The two-dimensional domain (Fig. 1) has extensions of 4900m x 5000m and is discretized by 50 x 50 squared grid cells of size $\Delta x = \Delta y = 100$m. The northern and southern boundaries are impervious, along the western boundary a fixed head of h_{west}=0m is imposed, and along the eastern boundary a fixed head of h_{east}=5m prevails. A pumping well with pumping rate Q_w=5000m^3/d is located at a distance of 1900m from the western boundary, and 2450m from the southern boundary. The area receives a spatially uniform recharge of N=1mm/d. Porosity is taken as n=0.1. Steady-state groundwater flow in a semi-confined aquifer is simulated. A reference transmissivity field was generated with a geometric mean transmissivity equal to T_g=86.4m^2/d and an exponential covariance function with variance σ_Y^2=1 and a correlation length of I_Y=500m. For the chosen conditions a water divide along the eastern part of the

area is present and the well pumps water from a considerable area located west of the water divide (see Fig. 1). The groundwater flow and mass transport equations were solved for the reference field. The reference transmissivity fields were sampled according to four different sampling designs. In the first sampling design ten locations of transmissivity measurements were randomly chosen (see Fig. 2). In the second design the measurement locations were systematically located in zones with largest uncertainty in the capture probability (see Fig. 3). In the third design the measurement locations were systematically located in zones with no uncertainty in the capture probability (see Fig. 4). In the fourth design the measurement locations were systematically located in zones with uncertainty in the capture probability (see Fig. 5).

The uncertainty bandwidth b of the semi-analytical approach in Fig. 2 - 5 (upper part) is taken as $b = 4\sqrt{X_{pp}}$ normal to the mean catchment boundary, assuming Gaussian distribution. The results of the semi-analytical approach are confronted with those from unconditional and conditional numerical Monte Carlo solutions, in which the ensemble average and the variance were evaluated for 100 realizations each. The ensemble mean catchment boundary is characterized by the line with the probability of 0.5. The uncertainty bandwidth is presented in Fig. 2 - 5 (lower part) as the two lines with a probability of 0.025 and 0.975 that a location belongs to the catchment.

6 Discussion and conclusions

The semi-analytical results (Fig. 2 - 5) are generally in rather good correspondence with the Monte Carlo solutions. However, the validity of the results is limited to a minimum distance of about two correlation lengths I_Y from the domain boundary. This limitation is mainly due to the chosen analytical velocity covariance approximation. Nevertheless, a certain distance to a boundary is anyway needed in order to justify the assumption of a Gaussian probability density of the lateral particle's location given its variance. Furthermore, it should be kept in mind that the solution is acceptable as long as advection is the dominant transport process and that therefore local dispersion and molecular diffusion can be disregarded.

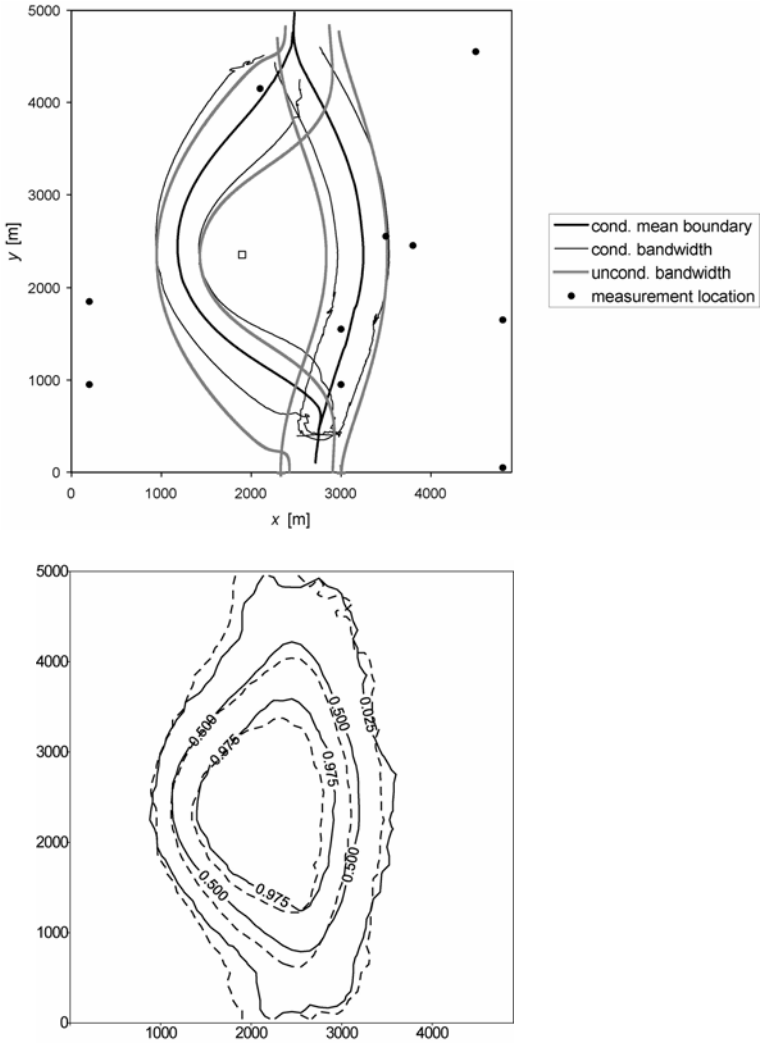


Fig. 2. Semi-analytical solution (top) for the uncertainty well catchment, conditional to 10 randomly located transmissivity measurements; Monte Carlo solution (bottom, solid lines) together with unconditional solution (dashed lines).

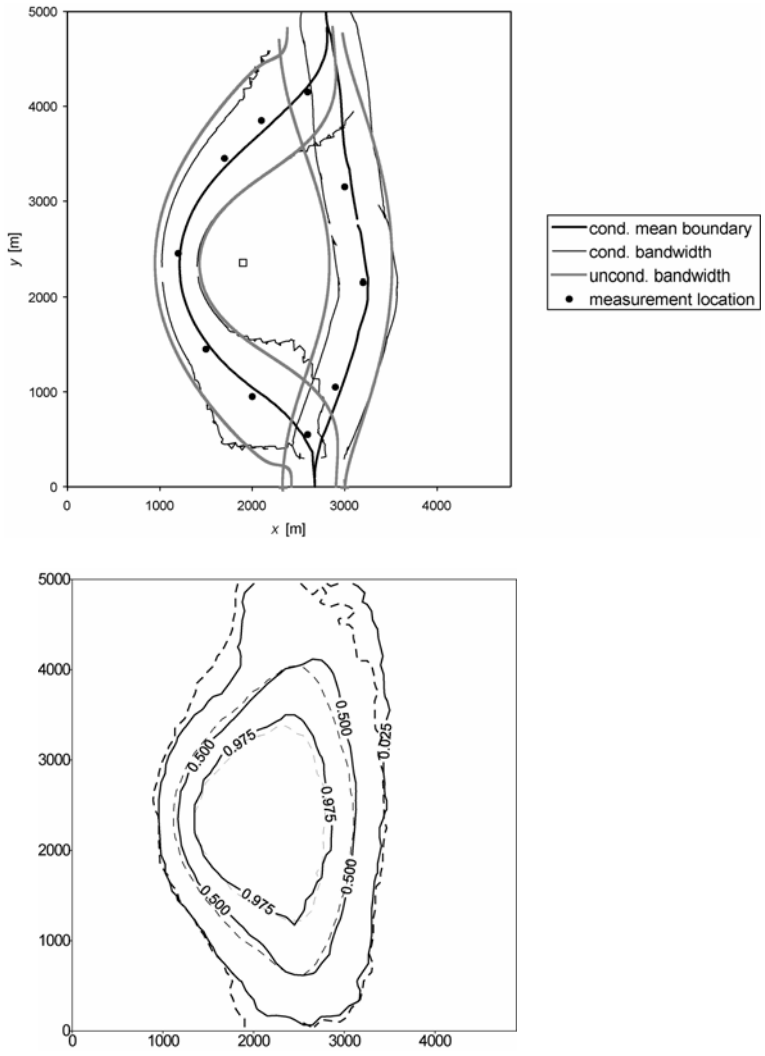


Fig. 3. Semi-analytical solution (top) for the uncertainty well catchment, conditional to 10 transmissivity measurements, located in zones with largest uncertainty in the capture probability; Monte Carlo solution (bottom, solid lines) together with unconditional solution (dashed lines).

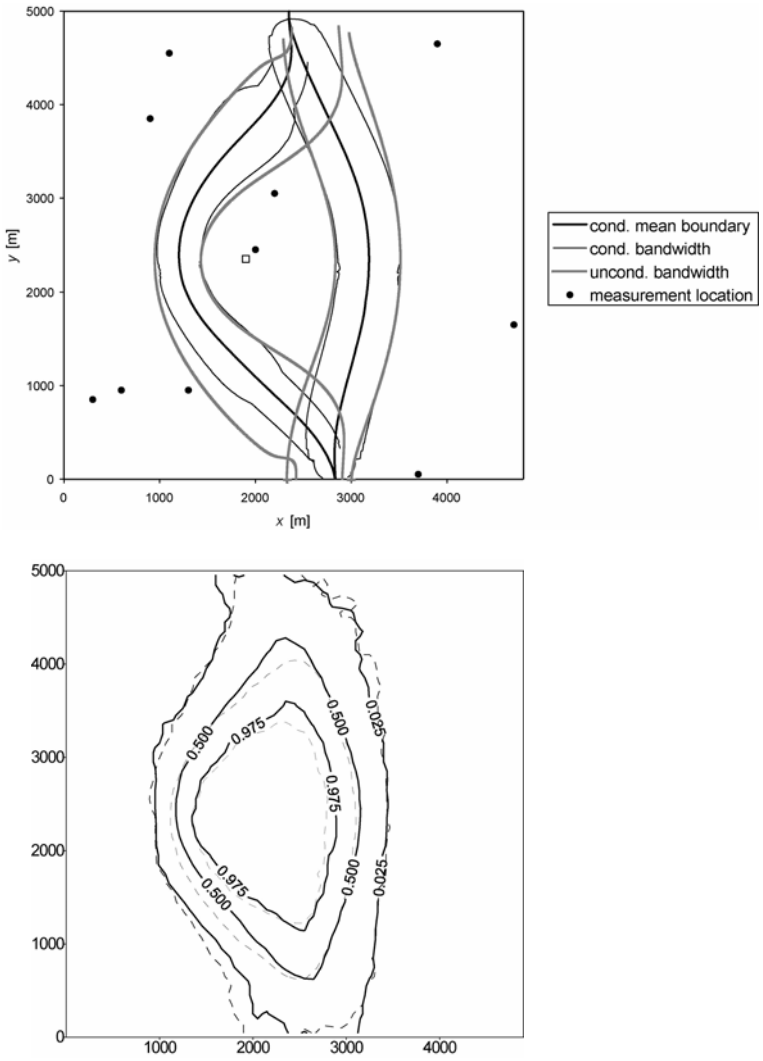


Fig. 4. Semi-analytical solution (top) for the uncertainty well catchment, conditional to 10 transmissivity measurements, located in zones with no uncertainty in the capture probability; Monte Carlo solution (bottom, solid lines) together with unconditional solution (dashed lines).

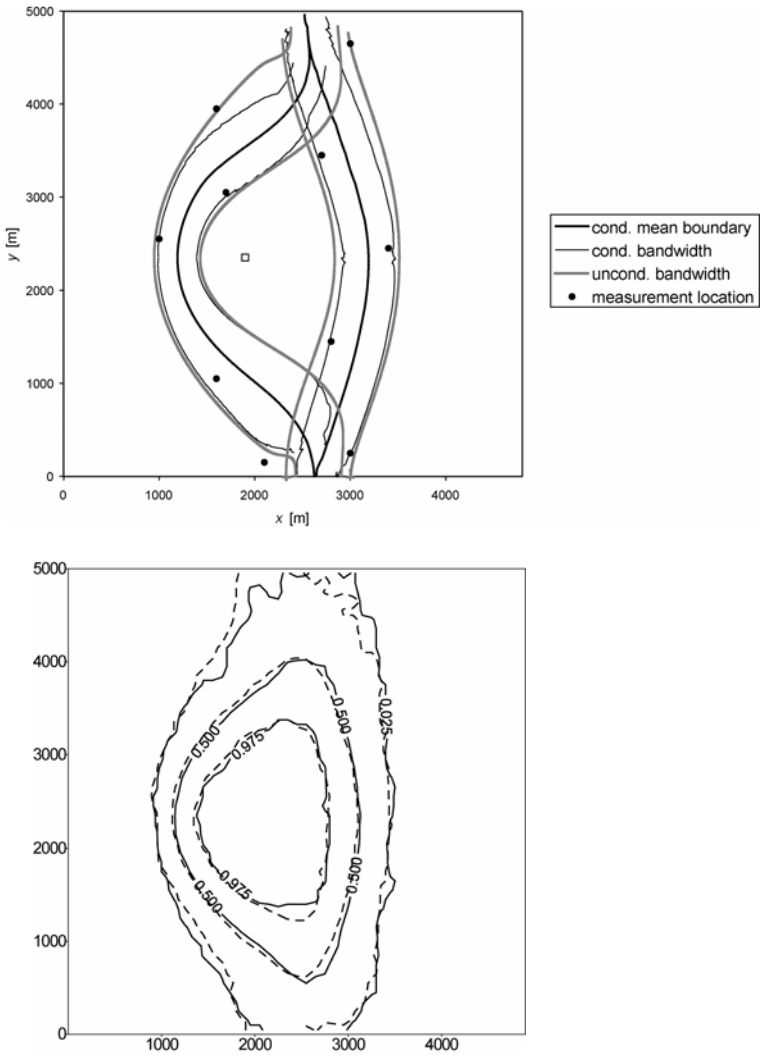


Fig. 5. Semi-analytical solution for the uncertainty well catchment, conditional to 10 transmissivity measurements, located in zones with uncertainty in the capture probability; Monte Carlo solution (bottom, solid lines) together with unconditional (dashed lines).

The synthetic study shows that the consideration of ten transmissivity measurements resulted in an only marginal reduction of the uncertainty bandwidth of the boundary of the well catchment. Obviously, ten transmissivity measurements are not enough to enhance the precision of the results. None of the sampling designs showed a clear improvement compared to the unconditional case. However this does not necessarily mean that the design of the monitoring network is not relevant in general.

The results (Fig. 2 - 5) can directly be compared with those from Stauffer *et al.* (2004), which were based on the same definition of the test problem. They used 25 transmissivity and 25 co-located head data for conditioning in two sampling designs. They found a substantial reduction in uncertainty in both cases. However, in the case where the measurement locations were randomly chosen within the unconditional uncertainty bandwidth close to the catchment boundary, the uncertainty reduction was larger.

The application of the proposed semi-analytical method to non-rectangular domains is in principle possible and straightforward. The method can be used for post-processing of deterministic steady-state models of well catchments.

Acknowledgements

The study was performed within the European Research Project "Stochastic Analysis of Well Head Protection and Risk Assessment" (W-SAHaRA). This project has been supported by the Swiss Federal Office for Education and Science (BBT), project 99.0543.

References

Gómez-Hernández JJ, Journel AG (1993) Joint sequential simulation of multi-Gaussian fields. In: Geostatistics Troia'92 volume 1, Soares A (ed.), 85-94

Dagan G (1989) Flow and transport in porous formations. Springer, Berlin

de Marsily G (1986) Quantitative Hydrogeology. Academic Press, San Diego

Rubin Y (1990) Stochastic modeling of macrodispersion in heterogeneous porous media. Water Resources Research, 26: 133-141

Rubin Y (1991) Prediction of tracer plume migration in disordered porous media by the method of conditional probabilities. Water Resources Research, 27: 1291-1308

Stauffer F, Attinger S, Zimmermann S, Kinzelbach W (2002) Uncertainty estimation of well catchments in heterogeneous aquifers. Water Resources Research, 38(11), 1238, doi:10.1029/2001WR000819

Stauffer F, Hendricks Franssen H-J, Kinzelbach W (2004) Semianalytical uncertainty estimation of well catchments: Conditioning by head and transmissivity data. Water Resources Research, 40(8), W08305, doi:10.1029/2004WR003320

Zhang D (2002) Stochastic methods for flow in porous media: Coping with uncertainties. Academic Press, San Diego

Conditional moments of residence time of sorbent solutes under radial flow

C. Castillo-Cerdà[1], X. Sanchez-Vila[1], L. Nuñez-Calvet[1] and A. Guadagnini[2]

[1]Department of Geotechnical Engineering and Geosciences, Technical University of Catalonia, Gran Capità S/N, 08034 Barcelona
[2]Dipartimento di Ingegneria Idraulica, Ambientale, Infrastrutture Viarie, Rilevamento (DIIAR), Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano

1 Introduction

A most typical aquifer remediation scheme is that of extracting the solutes dissolved in groundwater through pumping. Assuming that the natural background flow is not very important with respect to the flow regime imposed by the abstraction, a pseudo radial flow develops. The objective of this paper is to analyze the influence of a number of hydraulic and hydrogeochemical parameters, which play a role in the processes governing displacement of a pollutant in an aquifer remediation scenario. The remediation method considered consists on extracting groundwater from a pumping well. Water advected to the well would carry the pollutants dissolved to the biosphere, where they can be removed before disposing (or reinjecting) the treated water. An accurate knowledge of the processes controlling the space-time evolution of the pollutant is needed in order to perform an effective aquifer remediation. It is also essential to properly recognize the natural heterogeneity of the medium. A convenient way to accomplish this relies on the assumption that all relevant natural properties (physical and chemical) can be treated as Spatial Random Functions. Here we consider that the solute undergoes reversible linear instantaneous equilibrium sorption (LIE), with spatially variable distribution coefficient (and hence, retardation factor).

Analysis of travel time for conservative and sorbing solutes in heterogeneous media has been broadly addressed in the literature. Shapiro and Cvetkovic (1988) and Dagan *et al.* (1992) study travel time to a plane perpendicular to (uniform) mean flow direction in a 2- or 3-D aquifer, providing the first analytical expressions. Selroos and Cvetkovic (1992) extend the previous work to non-conservative solutes undergoing non-instantaneous sorption, again for uniform flow conditions and using a numerical methodology. Analytical results are provided by Cvetkovic *et al.* (1998) for the non-conditional case and by Sanchez-Vila and Rubin (2003) for the conditional one.

The main objectives of this work are: (1) studying the variation of the travel time (time taken by a solute particle to reach the pumping well) as a function of a

number of parameters such as statistical moments (mean, variance, and integral scale) of transmissivity, T and distribution coefficient, K_d, plus their cross-correlation; (2) determining until which extent the uncertainty in T and K_d is transmitted into the uncertainty in travel time of a given solute particle; and (3) analyzing the impact of measurements (in terms of both location and actual measured values) in the reduction of uncertainty. The methodology is based on a numerical Monte Carlo analysis, and the results for unconditional and conditional travel time are presented in statistical form.

2 Statement of the problem

We consider a thin aquifer of porosity ϕ which is polluted by a sorptive solute. Furthermore, we assume an individual solute particle to be initially located at a given distance r_0 from an existing well. We focus on the process of forcing the extraction of this particle by pumping. The particle trajectory is uncertain due to heterogeneity in hydraulic conductivity (or transmissivity). Further, the solute undergoes retardation due to sorption, which is characterized by the local values of the distribution coefficient. The heterogeneous nature of the medium is modeled by assuming that transmissivity and distribution coefficient are spatially auto- and cross-correlated.

This study can be seen as the kernel for a real pollution problem, where a certain area is considered polluted initially. Since the first step in a real remediation study is to eliminate the source, in our analysis we consider no pollution source term, while the only sink term is the pollutant removed through pumping. We also disregard areal recharge (representing either a confined aquifer or fast remediation problem during a dry season). Furthermore, no dispersion is considered. In summary, the transport equation reduces to:

$$R\phi\frac{\partial c}{\partial t} = -\mathbf{q}\nabla c\,, \tag{1}$$

where ϕ is porosity [-] and R is the retardation factor [-] which, for a solute undergoing linear reversible instantaneous equilibrium in a saturated medium, is given by (e.g. Domenico and Schwartz 1990):

$$R = 1 + \frac{\rho_b}{\phi}K_d\,, \tag{2}$$

where ρ_b is the bulk density [ML^{-3}] and K_d the distribution coefficient [L^3M^{-1}].

Travel time for a sorptive solute, t^R, is defined as the time needed for a particle to travel from point A to point B along a flow path assuming that no dispersion occurs and is given by:

$$t^R = R\int_A^B \frac{d\eta}{V(\eta)}\,, \tag{3}$$

where $V(\eta)$ is the velocity of a particle traveling from point A to point B along trajectory η. Under perfectly radial flow conditions, Eq. 3 can be integrated analytically. If location B corresponds to the pumping well and point A is located at a

distance r_0 from a well pumping a total flow per unit width of aquifer, Q [L^2T^{-1}], the travel time is:

$$t^R = R\,\phi\pi\, r_0^2 \,/\, Q, \tag{4}$$

In physically and geochemically heterogeneous domains, the particle trajectory and the travel time would be uncertain. Uncertainty in travel time would arise from the limited knowledge of the transmissivity distribution in space (and thus, that of the particle trajectory), and from the local variability in the retardation factor.

A most common way to address a problem involving heterogeneity is by means of regionalized variables. Here we would assume that both T and K_d are Spatial Random Functions. Then the output, in this case the travel time from a given release point, becomes a random variable. In general we would be interested in finding the complete probability density function (pdf) of the output variable. However, in most cases we would only be able to find the first few statistical moments. Sanchez-Vila and Rubin (2003) provided the following analytical expression of the expected value of travel time for a solute injected at a distance r_0 from the well:

$$< t^R(r_0) >=< R >< t^C(r_0) > + \Psi(r_0)\,, \tag{5}$$

where $t^C(r_0)$ is the travel time corresponding to a conservative solute, <R> is the mean retardation factor (< > accounts for expectation throughout the text), given as $< R >= 1 + \frac{\rho_b}{\phi} < K_d >$. The last term in Eq. 5 corresponds to the contribution of the physico-chemical correlation, and is given by

$$\Psi(r_0) = \frac{\rho_b}{\phi}\left\langle \int_0^{\eta(r_0)} \frac{K_d(\mathbf{x}) - < K_d >}{V(\eta')}\, d\eta' \right\rangle, \tag{6}$$

The expression in Eq. 6 vanishes if T and K_d are independent variables. However, there is empirical evidence that these two variables should be correlated (Roberts *et al.* 1986; Robin *et al.* 1991; Allen-King *et al.* 1998). Robin *et al.* (1991) postulate the following relationship between T and K_d:

$$Z = \ln(K_d) = \ln(K_{d,G}) + \beta Y' + W\,, \tag{7}$$

where Y=ln(T); Y' = Y – <Y>; $K_{d,G}$ is the geometric mean of the local distribution coefficient; β is a coefficient reflecting the degree of linear correlation between the variables (Y and Z); and W is a Gaussian process, $W = N(0, \sigma_W^2)$, to account for imperfect correlation between Z and Y. Even though Robin *et al.* (1991) used a white noise for W, their expression can easily be generalized to include a variable with a non-zero correlation length. The actual value of β would depend on the solute and the medium mineralogy. It can be either negative (Robin *et al.* 1991) or positive. As an example, Allen-King *et al.* (1998) found positive values of β (in some cases even larger than 1) for PCE and in some given facies.

Some expressions for the first and second moment of K_d can then be written after Eq. 7:

$$< K_d >= K_{d,G} \exp\left(\sigma_Z^2 / 2\right),$$

$$\gamma_Z(h) = \beta^2 \gamma_Y(h) + \gamma_W(h), \tag{8}$$

$$\sigma_Z^2 = \beta^2 \sigma_Y^2 + \sigma_W^2,$$

For the cross-variogram we obtain:

$$\gamma_{YZ}(h) = \beta \gamma_Y(h), \tag{9}$$

Expressions for the variance of travel time that account for the correlation between Y and Z are available in Sanchez-Vila and Rubin (2003).

3 Monte Carlo approach

The numerical approach undertaken in this work can be summarized as follows:

1. Generation of a number of conditional simulations of correlated Y and Z fields.
2. Solution of the groundwater flow equation for convergent flow conditions. Study of transport of sorbing solutes by means of a particle tracking code (see Fig. 1 for the set-up of the problem).
3. Statistical analysis of travel time.

Using this methodology we analyzed different scenarios accounting for variations in the integral scale of the Spatial Random Functions, and for value and location of the conditioning data. For each scenario 2,000 simulations were run. Sensitivity to the number of simulations was checked against the results from 10,000 simulations in selected scenarios, with similar values for the travel time moments.

Simulations were performed within a square domain of size 2L = 20 (arbitrary units), discretized into squares of size 0.2 (10,000 elements in total). The variograms used are spherical and isotropic for both variables. The mean values used are <Y> = 0 and <Z> = –1.27. Additional parameters are $\sigma_Y^2 = 1$, $\rho_b = 1.6$ and $\phi = 0.3$. With the values selected the mean retardation factor is a function of β: when $\beta = 0$, then <R> = 3.46; when $|\beta| = 1$, then <R> = 5.06. The variance of Z depends on β through (8). We consider two main cases for the integral scale, λ: (1) $\lambda = 1$ ($L/\lambda = 10$), so that the integral scale is small with respect to the domain size; and (2) $\lambda = 10$ ($L/\lambda = 1$), rendering an integral scale of the order of the domain size. Throughout the text we denote them as short range and long range heterogeneity, respectively.

Radial flow is analyzed by locating a steady-state pumping well (Q = 100 in consistent units) at the center of the domain, while keeping a fixed head at the external boundary. A particle is then placed at a point located at radial distance r_0=8.9 (see Fig. 1), and it is tracked since it reaches the well. Particle tracking is

performed by evaluating the location of the particle at a given time. Velocity is computed from the nodal heads through finite elements. Velocity is then minored by the local retardation factor at that particular position. Then particle is displaced and travel time is updated.

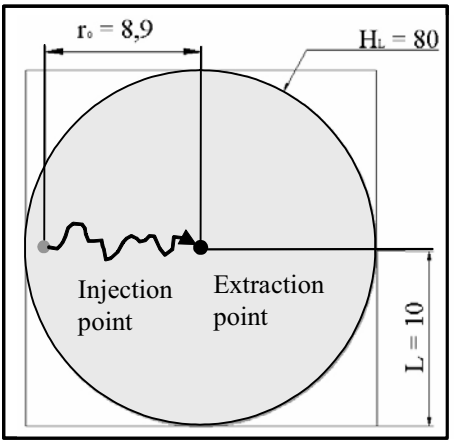


Fig. 1. Set-up of the numerical analysis.

4 Effect of conditioning. Constant K_d

In order to separate the impact of conditioning upon transmissivity from that of heterogeneity in K_d, we first study the displacement of a solute with a constant retardation factor. We analyze the impact on mean and variance of travel time of one transmissivity datum located either at the injection or at the pumping point. This is a common situation in reality, where these are two of the few available points where transmissivity measurements can be taken through hydraulic testing. In our study, we analyze dimensionless travel time, t*, which is obtained by dividing the actual computed time by the equivalent time for homogeneous media provided by (4). Our result is then insensitive to the value of Q, ϕ, and R. Results of normalized travel time mean, <t*>, and variance, Var t*, are displayed in Table 1, together with the values for the variance of the natural logarithm of t*, ln-t*. The latter is much more informative than Var t*, as the pdf of travel time is asymmetrical and positively skewed.

From Table 1 we can see that the mean travel time is larger than the one for homogeneous media (t*>1) in almost all cases. For the unconditional case this result was already presented by Riva *et al.* (2004). Due to the asymmetry in the travel time pdf, the probability of normalized travel time being smaller than 1 is 0.37 for the case L/λ=10 and 0.40 for L/λ=1. On the other hand, there is a small probability that very large times occur, thus leading to a large asymmetry.

For short range heterogeneity (L/λ=10), conditioning on transmissivity at the well has almost no effect on mean travel time or variance of ln-t*, while conditioning on the value at the injection point produces a remarkable effect. The physical explanation of this result is the following: on one side, we note that solute at the injection point travels at a low velocity and spends quite a large time in that vicinity. Contrariwise, the solute spends less time close to the well. Therefore, any modification in the local velocity, due to conditioning, close to the injection point would have a very significant effect in increasing (or decreasing) travel time.

Table 1. Mean and variance of normalized travel time (and natural logarithm of travel time) for different conditioning values and integral scales. The unconditional cases are presented for reference. Y_w and Y_{inj} are the conditioning values of log-transmissivity at the well and at the injection point, respectively.

			<t*>	Var t*	Var ln t*
L/λ=10	Yw	-3	1.313	0.525	0.238
	Yw	-1	1.316	0.562	0.251
	Yw	1	1.329	0.580	0.256
	Yw	3	1.335	0.569	0.247
	Unconditioned		1.356	0.620	0.256
	Yinj	-3	3.229	2.254	0.185
	Yinj	-1	1.604	0.674	0.210
	Yinj	1	1.063	0.307	0.221
	Yinj	3	0.830	0.192	0.218
L/λ=1	Yw	-3	1.170	0.202	0.125
	Yw	-1	1.167	0.248	0.141
	Yw	1	1.232	0.314	0.160
	Yw	3	1.224	0.350	0.175
	Unconditioned		1.228	0.408	0.188
	Yinj	-3	3.017	1.894	0.174
	Yinj	-1	1.551	0.361	0.127
	Yinj	1	0.896	0.096	0.100
	Yinj	3	0.562	0.023	0.066

Finally, from the table we can also identify that conditioning is more significant when the integral scale is comparable to the travel distance. In such a case, the impact of a conditioning datum extends throughout the aquifer, and the transmissivity field is somehow homogenized in each individual realization, thus leading to smaller variances (both in t* and in ln-t*).

5 Effect of conditioning. Heterogeneous T and K_d

A similar numerical methodology is used to estimate the combined impact of heterogeneity in transmissivity and distribution coefficient, which are assumed cross-correlated through Eq. 7. In this case we should differentiate between the effects of conditioning in each variable separately. Notice that due to the existence of

cross-correlation, conditioning on a primary variable would have an indirect effect on the value of the secondary variable.

First, some simulations are performed for the unconditional case. We consider different values for β ranging from −1.0 to 1.0. In Table 2 the mean values are presented, normalized with respect to an equation similar to Eq. 4 but using the mean retardation value given by Eq. 8. The remaining parameters selected for the simulations are: <Y> = 0, <Z> = −1.27, Var(Y) = 1.0, Var(W) = 1.0

Table 2. Normalized mean travel time and variances for different correlation parameter β, and two L/λ values. The scenarios corresponding to constant K_d are reported for reference.

			<t*>	Var ln t*
L/λ=10	β	-1	1.507	0.872
		-0.5	1.387	0.583
		0	1.271	0.364
		0.5	1.137	0.243
		1	1.006	0.216
	Constant Kd		1.356	0.256
L/λ=1	β	-1	1.509	1.099
		-0.5	1.266	0.665
		0	1.166	0.463
		0.5	1.111	0.461
		1	0.957	0.552
	Constant Kd		1.224	0.188

Two conclusions can be drawn from Table 2. First, similar to the constant K_d case, the mean travel times are larger for short range heterogeneity (L > λ). Second, travel times decrease monotonically for increasing β. If β < 0, not accounting for cross-correlation between Y and Z would lead to an underestimation of the rehabilitation time. The opposite holds for positive β.

Regarding the variance, while in the constant K_d case it was larger for the scenario corresponding to L > λ, in the cross-correlated case we obtain larger variances for L = λ. Conditioning in this set-up has been performed both at the injection and the pumping well. We consider now conditioning both in Y or Z.

5.1 Conditioning at the well

When conditioning is performed at the well, the behavior of the system strongly depends on the ratio L/λ. When this ratio is large (results not presented in the paper), the impact of conditioning upon transmissivity on the moments of travel time is almost negligible. When conditioning upon K_d measurements, there is a slight tendency of increasing mean travel time for increasing measured values of K_d.

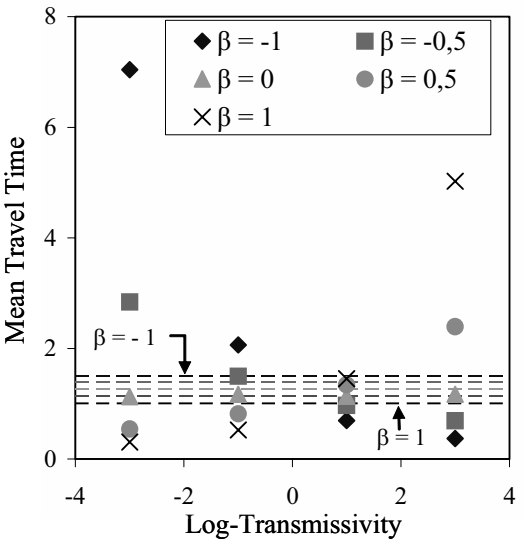


Fig. 2. Impact of conditioning on transmissivity at the well when L/λ=1. Broken lines represent the values corresponding to the non-conditional case for each β value (top line β = –1 and increasing β for downward lines).

On the other hand, for large range correlation (L = λ), even when the only conditioning point is the pumping well the effect of conditioning extends throughout most of the trajectory. For this reason, the effect of conditioning has a major impact on travel time mean (see Fig. 2) and variance. As an example, when we condition only on transmissivity the impact on mean travel time is different depending on the value of β. When β is negative, mean travel time decreases with increasing the transmissivity, T_w, measured at the well and used for conditioning. On the other hand, a positive β leads to an increase of the mean travel time with increasing conditioning value. The reason for this particular behavior is an effect of indirect conditioning on K_d. A negative correlation ($\beta < 0$) causes K_d to decrease when transmissivity increases, thus resulting in a generally shorter mean travel time. On the other hand, large T_w values lead to large retardation factors in the case of positive correlation ($\beta > 0$), thus dispalying a tendency to increase mean travel time. The effect of conditioning on a K_d value is always monotonic, since increasing $K_{d,W}$ (K_d measured at the well) leads to larger travel times, independently of β (see Fig. 3). In this case, the effect of β can be clearly recognized, in that mean travel times are generally smaller when β is larger. The results for the variance are presented in Table 3. In this case, conditioning clearly affects travel time variance, particularly for large range heterogeneity. When conditioning on T_w, variance reduces in the presence of larger T_w values whenever $\beta < 0$, while the opposite happens for $\beta > 0$. Again, when conditioning upon $K_{d,w}$, travel timevariance always increases with increasing value of the conditioning parameter.

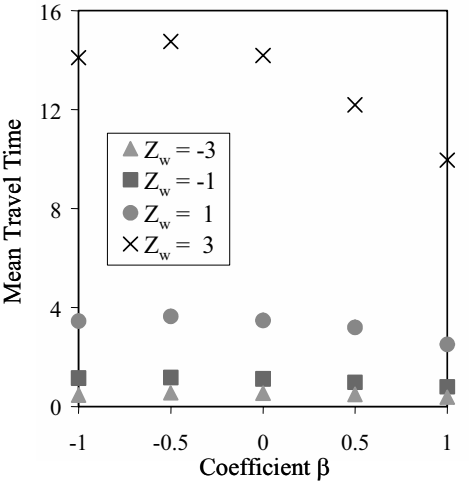


Fig. 3. Dependence of thre mean travel time on the correlation coefficient, β, and the value of the conditional log-distribution coefficient at the well, Z_w, when L/λ=1.

Table 3. Variance of ln-t* as a function of β for different conditional values and L/λ=1.

β	Conditioning on ln T				Conditioning on ln (Kd)			
	-3	-1	1	3	-3	-1	1	3
-1.0	1.284	1.028	0.619	0.359	0.406	0.723	0.873	0.873
-0.5	0.789	0.646	0.535	0.431	0.287	0.460	0.551	0.594
0.0	0.415	0.437	0.433	0.447	0.186	0.278	0.396	0.418
0.5	0.206	0.322	0.470	0.574	0.141	0.211	0.308	0.318
1.0	0.125	0.278	0.555	0.762	0.125	0.242	0.368	0.379

5.2 Conditioning at the injection point

In this case only two values of β were analyzed. Table 4 reports the mean variance of ln-t* for different conditioning values of log-transmissivity and ln(K_d) at the injection point when β = – 0.5, 0.5, and L/λ=10. We start by noting that a major difference with respect to conditioning at the well is that now we see an important impact of conditioning also when L/λ=10. From Table 4 we can observe that a larger T_{inj} (conditional value at the injection point) results in a smaller mean travel time when L/λ=10. This behavior is independent of β, even though it appears to be more evident when β < 0. When β = – 0.5, an increase in transmissivity causes K_d to dimish as well, thus resulting in a strong decrease of the resulting contaminant travel time. Conversely, β = 0.5 indicates that K_d increases with transmissivity. Even though this produces contrasting effects on the travel time it appears that the

distribution coefficient effect dominates that of transmissivity. The results obtained when conditioning on the distribution coefficient (Table 4) reveal that the larger $K_{d,inj}$ (conditional value at the injection point), the larger mean travel time, regardless of the β value.

Table 4. Dependence of mean and variance of ln-t* on β for various conditioning values at the injection point. L/λ=10. Each cell displays mean value (top) and variance (bottom). For comparing purposes, unconditional variances of ln-t* are 0.583 (β = –0.5) and 0.243 (β = 0.5).

β	Conditioning on ln Tinj				Conditioning on ln Kd,inj			
	-3	-1	1	3	-3	-1	1	3
-0.5	6.640	1.938	0.992	0.747	0.932	1.386	3.324	17.50
	0.668	0.534	0.498	0.537	0.487	0.522	0.573	0.682
0.5	1.985	1.342	1.003	0.908	1.046	1.168	1.590	3.141
	0.207	0.243	0.250	0.270	0.236	0.249	0.263	0.314

When L/λ = 10, variance of ln-t* (Table 4) tends to be larger for negative β. Moreover, increasing in T_{inj} in the presence of β>0 leads to a slight increase in the log travel time variance. Contrariwise, β<0 leads to a non-monotonic behavior of the variance. In any case, the most relevant result is that conditioning leads to very small variation in the variance.

The effect of conditioning at the injection point on travel time variance for larger log-transmissivity correlation scales (L/λ=1) is depicted on Fig. 4. As opposed to the non-monotonic behavior observed for L/λ=10, increasing the log transmissivity causes the variance of travel time to decrease when β = 0.5. On the contrary, a negative correlation (β = – 0.5) results in a clear increase of the travel time variance with increasing log-transmissivity conditioning value.

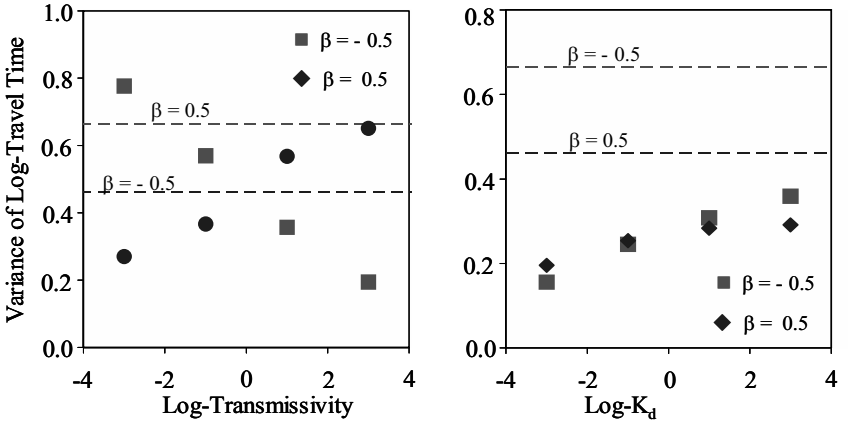


Fig. 4. Variance of log-travel time when conditioning is performed at the injection point and L/λ=1. The conditioning parameter is either transmissivity (left) or distribution coefficient (right). Dashed lines correspond to the unconditional values.

Conditioning on log-transmissivity does not result in a reduction of the travel time variance with respect to the unconditional case in all situations tested. On the other hand, conditioning on the distribution coefficient always results in a reduction of the travel time variance, with respect to the unconditional case.

6 Conclusions and Final Discussion

Our work leads to the following major conclusions:

- Additional information about the expected value of the travel time is not obtained by conditioning at the pumping well location when the transmissivity integral scale is much smaller than the solute travel distance. On the other hand, when both distances are of similar order, conditioning upon measurements taken at the pumping well influences visibly the travel times; however, this influence is in all cases smaller than in the cases where conditioning is performed at the injection point.
- Conditioning upon transmissivity values, when performed at the injection point, has a lesser impact on travel time that doing it upon distribution coefficients.
- The linear correlation coefficient between the transmissivity and the distribution coefficient plays an important role in the behavior of the travel time (both in the conditional and unconditional cases).
- In general, conditioning reduces travel time variance (increases confidence in predictions), even though this is not true in all the cases. We provide an example of this counterintuitive finding. Our results suggest that it is not possible to insure in all cases that conditioning at the injection point renders results which are on the safe side with respect to the time needed for aquifer remediation. A conclusive answer to this problem might be provided by a more extensive series of simulations.

A method to help evaluate whether or not the Pump-and-treat method could be efficient in a given context has been presented. Efficiency is measured here in terms of rehabilitation time, remnant concentrations and cost. The challenge is to be able to define a priori, given a polluted site, whether the method is going to be feasible. The ultimate aim of our work is to develop methods to simulate the real medium and the behavior of a given pollutant, in order to, first, postulate a priori the aquifer reclamation time, and second, to recommend whether pump-and-treat is a good alternative for any given site remediation problem.

References

Allen-King RM, Halket RM, Gaylord DR, Robin, MJL (1998) Characterizing the heterogeneity and correlation of perchloroethene sorption and hydraulic conductivity using a facies-based approach. Water Resources Research, 34 (3), 385-396

Cvetkovic V, Dagan G, Cheng H (1998) Contaminant transport in aquifers with spatially variable hydraulic and sorption parameters. ProceedingsRoyal Soc. London A, 454, 2173-2207

Dagan G, Cvetkovic V, Shapiro A, (1992) A solute flux approach to transport in heterogeneous formations, 2, Uncertainty analysis. Water Resour. Res., 28(5), 1369-1376

Domenico PA, Schwartz FW (1990) Physical and chemical hydrogeology. John Wiley and Sons

Riva M, Sanchez-Vila X, De Simoni M, Guadagnini A, Willmann M (2004) Effects of heterogeneity on aquifer reclamation time. Kluwer Academic Publishers, 259-270

Roberts PV, Goltz MN, Mackay DM (1986) A natural gradient experiment on solute transport in a sand aquifer. 3. Retardation estimates and mass balance for organic solutes. Water Resour. Res., 22(13), 2407-2057

Robin, MJL, Sudicky EA, Gillham R, Kachanoski RG (1991) Spatial variability of Strontium Distribution Coefficients an.W.d their correlation with hydraulic conductivity in the Canadian Forces Base Borden Aquifer. Water Resour. Res., 27(10), 2619-2632

Sánchez-Vila X, Rubin Y (2003) Travel time moments for sorbing solutes in heterogeneous domains under nonuniform flow conditions. Water Resour. Res., 39(4), 1086

Selroos JO, Cvetkovic V (1992): Modeling solute advection coupled with sorption kinetics in heterogeneous formations. Water Resour. Res., 28(5), 1271-1278

Shapiro AM, Cvetkovic VD (1988) Stochastic analysis of solute arrival time in heterogeneous porous media. Water Resour. Res., 24(10), 1711-1718.

Impact of the choice of the variogram model on flow and travel time predictors in radial flows

M. Riva[1], M. De Simoni[1] and M. Willmann[2]

[1] Dipartimento Ingegneria Idraulica, Ambientale, Infrastrutture Viarie, Rilevamento (DIIAR), Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy.
[2] Department of Geotechnical Engineering and Geosciences, Technical University of Catalonia, Gran Capità S/N, 08034 Barcelona, Spain.

1 Introduction

Prediction of hydraulic head, flux and contaminant travel time/trajectories in natural aquifers is uncertain due to the geologic media complexity and lack of information. Hence it is appropriate to cast the equations that govern groundwater flow and contaminant transport within a stochastic framework. The latter is oriented towards rendering ensemble moments of the analyzed quantities. In this view the spatial variable transmissivity is usually modeled as a Stochastic Continuum, characterized by a set of parameters (covariance shape, geometric mean, variance and correlation length). These are generally assumed to be known with certainty even though they are usually derived using a limited amount of experimental data, which are often not enough for a complete characterization.

Full-Bayesian approaches (e.g. Woodbury and Rubin 2000, Woodbury and Ulrych 2000) take into account the uncertainty in the knowledge of the variogram parameters (geometric mean, variance and correlation length). Feyen *et al.* (2002) illustrate an application of these methodologies to determine the uncertainty associated with the delineation of well capture zones. Hendricks Franssen *et al.* (2002) investigate the impact of the uncertainty of variogram parameters on the same topic using sequential Gaussian simulation (Gómez-Hernández and Journel 1993) to generate transmissivity fields and the sequential self-calibrated method for inverse conditioning. In all these works the shape of the correlation structure of the natural logarithm of transmissivity is fixed and assumed known without uncertainty. Salandin and Rinaldo (1989) analyze the influence of the form of the log-conductivity covariance on dispersion coefficients in random permeability fields under mean uniform flow conditions.

Here, we focus on the impact of the choice of the functional form for the log-transmissivity variogram on (ensemble) moments of hydraulic head and contaminant residence time under convergent flow conditions, such as those created by a single pumping well.

Although of high relevance in practical applications, problems associated to contaminant transport in the vicinity of extraction wells in heterogeneous media have been tackled only recently (e.g. Guadagnini and Franzetti 1999, Riva *et al.* 1999, Dagan and Indelman 1999, van Leeuwen *et al.* 2000, Feyen *et al.* 2002).

We perform a numerical Monte Carlo analysis of (a) the predictors of hydraulic head and residence time (rendered by their means) for conservative solute particles injected at various radial distances from the well, and (b) the associated prediction errors (rendered by the variance of the state variables investigated).

The natural logarithm of aquifer transmissivity, Y, is modeled as a statistically homogeneous Gaussian random field. Three functional forms of the variogram (namely Exponential, Gaussian and Spherical), chosen amongst the most common models used in the literature, are considered. The impact of the choice of the variogram model on flow and travel time predictors is analyzed for different domain sizes in terms of correlation scale of Y (i.e. extent of the aquifer within which the effects of pumping are not negligible) and degrees of heterogeneity (in terms of the variance of Y, σ_Y^2).

2 Statement of the problem and numerical Monte Carlo simulations

We consider incompressible groundwater steady state convergent flow created by a well of zero radius, located at the center of a circular randomly heterogeneous domain of radial extent L. The well pumps at a constant deterministic rate, and the head drawdown is assumed negligible at a given distance (L) from the well. This Dirichlet-type of boundary condition is based on the work of Sanchez-Vila *et al.* (1999), who showed that the differential drawdown between the pumping point and any observation point becomes constant with time for large pumping times, so that the drawdown conus keeps its shape. Therefore, a surface can be defined with all points having the same drawdown. Mathematically this is analogous to define this surface as zero drawdown. The shape of this surface would depend on the actual spatial distribution of transmissivity, but in the mean it will be a circumference.

We focus on the evaluation of (a) the hydraulic head distribution and (b) the travel time of solute particles released at time $t_0 = 0$ at a general point of polar coordinates $\boldsymbol{r_0} \equiv (r_0, \theta_0)$ centered at the well. In order to obtain the (ensemble) moments of the variables of interest, Monte Carlo simulations were conducted using the same code of Riva *et al.* (1999), with different boundary conditions. Flow and transport are simulated in a square domain with 100 rows and 100 columns of uniform size ($\Delta x = \Delta y = \delta = 0.2$). A circular boundary of radius $L = 50\ \delta$ was defined around the well by designating all cells outside it as inactive. The hydraulic head along the circular boundary was set constant. A pumping well at a constant rate $Q = 100$ (in consistent units) was placed at the central node of the grid.

We model the log-transmissivity, $Y(\boldsymbol{r}) = \ln T(\boldsymbol{r})$, as a statistically homogeneous and isotropic random function of space. Three different variograms between two points has been adopted in this study:

- Gaussian model (GV)

$$\gamma_Y(d) = \sigma_Y^2 \left[1 - exp\left[-\frac{\pi}{4\lambda^2} d^2 \right] \right] \tag{1}$$

- Exponential model (EV)

$$\gamma_Y(d) = \sigma_Y^2 \left[1 - exp\left[-\frac{d}{\lambda} \right] \right] \tag{2}$$

- Spherical model (SV)

$$\gamma_Y(d) = \begin{cases} \sigma_Y^2 \left[\frac{3}{2}\frac{d}{\lambda} - \frac{1}{2}\frac{d^3}{\lambda^3} \right] & for\ 0 \le d \le \lambda \\ \sigma_Y^2 & d > \lambda \end{cases} \tag{3}$$

where σ_Y^2 is the variance of Y, λ is the correlation length and d is the Euclidean distance. To isolate the influence of the choice of the variogram model we contrast results obtained by keeping fixed the spatial integral scale

$$I_Y = \frac{1}{\sigma_Y^2} \int_0^\infty \left[1 - \gamma_Y(r) \right] dr \tag{4}$$

A Gaussian sequential simulator code (GCOSIM3D, Gómez-Hernández and Journel 1993) was used to generate unconditional random realizations of Y on the above defined two-dimensional grid. Each realization constituted a sample from a multivariate Gaussian, statistically homogeneous field, with mean $\langle Y \rangle = 0$, variances σ_Y^2 ranging from 0.1 to 1.0 and two different values of the spatial integral scale

- Test case 1 – TC1: $I_Y = 5\delta$ (=1 = 0.1L)
- Test case 2 – TC2 : $I_Y = 50\delta$ (= 10 = L)

$I_Y = \lambda$ for the for the Gaussian and Exponential variogram and $I_Y = 3/8\ \lambda$ for the Spherical model.

The effective porosity, n, was taken as a constant and set to 0.3. Flow in each realization was solved by Galerkin finite elements using bilinear shape functions and solute transport was simulated by particle tracking, modeling only the convective component of motion and disregarding diffusive effects at smaller scales. To compute residence times, conservative solute particles were located at grid nodes of radial distances from the well, r_0, ranging from δ to 50δ and various angular positions, θ_0. Tracking was stopped when the particles reached one of the cells sharing the well node. To obtain the total travel time we added the time to get

from the trajectory end-point to the well (separated by a distance r), which is computed by means of the well known equation for the steady state radial flow in a homogeneous and isotropic field, as $t = n\,\pi\,r^2 / Q$.

A crucial point was the determination of the number of Monte Carlo realizations (NMC) needed to obtain the convergence of the ensemble moments analyzed. Due to the radial symmetry of the problem (domain, flow and boundary conditions), the statistical moments of hydraulic head, h , and residence time, t, are independent of the angular coordinate, when the convergence is attained. We applied the methodology proposed by Ballio and Guadagnini (2004) with reference to the mean and variance of hydraulic head and particle residence time. As an example of the results, Fig. 1 depicts the stabilization analysis of mean, $\langle h \rangle$, and variance, σ_h^2, of the hydraulic head (black lines) computed at a monitoring point located at radial distance from the well $r = 5 = L/2$ for TC1, using the Exponential Variogram model and $\sigma_Y^2 = 1.0$.

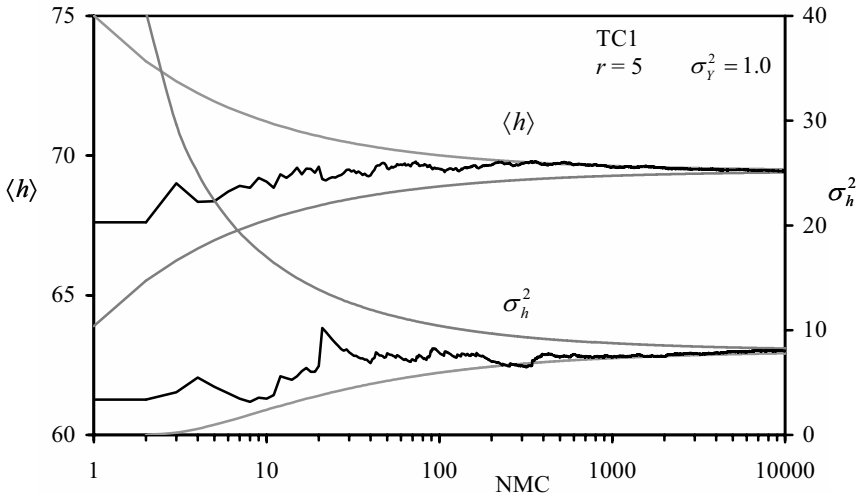


Fig. 1. Dependence of convergence of hydraulic head mean and variance on the number of Monte Carlo simulations (black lines) together with the 95% estimated confidence intervals (grey lines).

Fig. 1 also reports the 95% estimated confidence intervals (grey lines) which have been computed on the basis of the following expressions

$$Pr\left[\overline{\Re}_{NMC} - t_{NMC-1}\left(1-\frac{\alpha}{2}\right)\frac{S_{NMC}}{\sqrt{NMC}} \le \mu \le \overline{\Re}_{NMC} + t_{NMC-1}\left(1-\frac{\alpha}{2}\right)\frac{S_{NMC}}{\sqrt{NMC}}\right] = 1-\alpha \quad (5)$$

$$Pr\left[\frac{NMC-1}{\chi^2_{NMC-1}(1-\alpha/2)}S^2_{NMC} \le \sigma^2 \le \frac{NMC-1}{\chi^2_{NMC-1}(\alpha/2)}S^2_{NMC}\right] = 1-\alpha \quad (6)$$

respectively for the mean (Eq. 5) and for the variance (Eq. 6) of the variable of interest (hydraulic head in our example), where $1 - \alpha$ ($\alpha = 0.05$ in our example) is the probability that the value of the process mean μ (or variance σ^2) lies within the confidence interval around the sample mean $\overline{\Re}_{NMC}$ (or sample variance S^2_{NMC}), $t_{NMC-1}()$ is the Student distribution with (NMC–1) degrees of freedom and $\chi^2_{NMC}()$ is the chi-square distribution with NMC degrees of freedom.

The confidence intervals provide the order of magnitude of the uncertainty associated to the first and second moments evaluated on the basis of a finite sample of Monte Carlo realizations. Obviously, their width decreases as NMC increases.

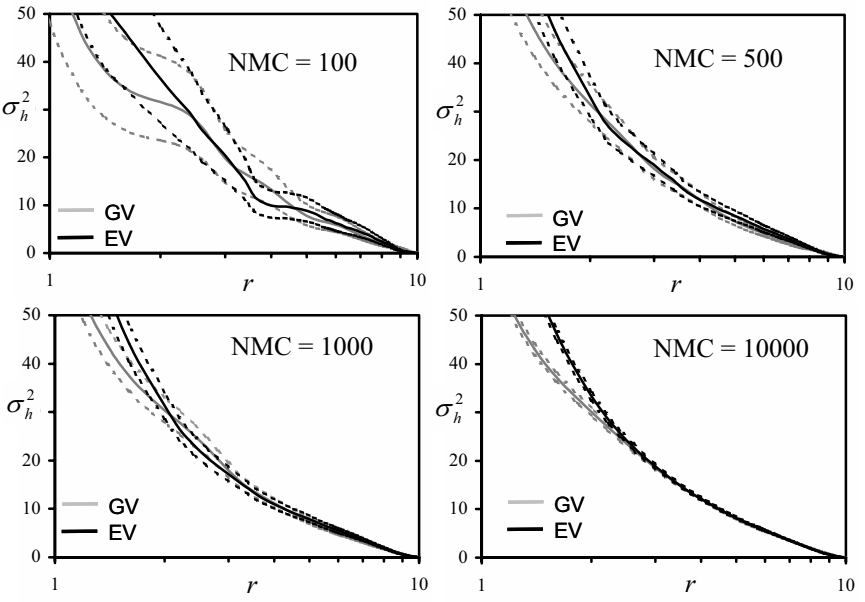


Fig. 2. Hydraulic head variance as a function of the radial distance from the well, r, and NMC for TC1 with $\sigma_Y^2 = 1$ evaluated using an Exponential (black continuous line) and a Gaussian (grey continuous line) variogram. Dashed curves indicate the 95% estimated confidence intervals (black for EV and grey for GV).

As an example of the rate of convergence of the Monte Carlo procedure, Fig. 2 depicts σ_h^2 and its 95% estimated confidence intervals as a function of the radial distance from the well for TC1 and σ_Y^2=1, computed using a Gaussian and an Exponential variogram model and increasing NMC from 100 to 10,000.

The crossings of the confidence intervals clearly evidences that not only a few hundreds, but also 1,000 Monte Carlo simulations are not sufficient to fully identify the effects of the variogram shape on σ_h^2, even though Fig. 1 suggests that 1,000 Monte Carlo iterations allow attaining (quasi-)stable values of σ_h^2. Similar results have been noted for the other cases and statistical moments considered.

An acceptable compromise between CPU time requirements and accuracy of reproduction of the statistical moments of hydraulic head and travel time was obtained with 10,000 Monte Carlo runs. The 95% confidence intervals computed after 10,000 iterations have a maximum width of 3.9% × σ and 5.5% × σ^2, respectively for the first and second order statistical moments, σ^2 being the variance of the variable of interest (hydraulic head or travel time). We also note that stopping at 1,000 iterations causes the width of such confidence intervals to increase up to 12.4% × σ and 17.6% × σ^2, respectively.

3 Results and discussion

In this Section the effect of the choice of the variogram model on hydraulic head and residence time statistical moments is investigated for TC1 and TC2. This is performed by introducing the quantities:

$$\wp^E = \frac{\psi^E - \psi^G}{\psi^G} \times 100; \qquad \wp^S = \frac{\psi^S - \psi^G}{\psi^G} \times 100 \tag{7}$$

where ψ represents a given ensemble statistical moment (i.e. either mean or variance of hydraulic head or travel time); the superscripts *G*, *E*, *S* indicate moments computed on the basis of Gaussian (GV, Eq. 1), Exponential (EV, Eq. 2) or Spherical (SV, Eq. 3) variogram model, respectively.

3.1 Mean hydraulic head

From a practically standpoint, the choice of the variogram model has not meaningful effects on the mean hydraulic head for all the cases analyzed. Larger $\wp^E$ and $\wp^S$ values occur near the well, where the maximum $\wp^E$ absolute value of about 8% has been obtained for TC2 and σ_Y^2 = 1.0. Consistently with our results, Riva *et al.* (2001) show that the heterogeneity effect on mean hydraulic head is relevant only near the well and increases as σ_Y^2 increases and L/I_Y decreases. In particular, for L/I_Y = 10 (Riva *et al.* 2001, their Fig. 8) the effect of σ_Y^2 is detectable only at radial positions $r < 0.5\ I_Y$.

However, a qualitative analysis reveals that the mean drawdown close to the well is larger using the Gaussian variogram for both TC1 and TC2. This behavior is explained by noting that two points are more correlated using the Gaussian model than other variogram models. Thus, at location close to the well, the source effect is stronger in the presence of a Gaussian variogram. At the same time, we also observe that the use of an Exponential variogram renders the smallest mean drawdown near the pumping well.

3.2 Hydraulic head variance

The effect of the variogram model on the assessment of the hydraulic head variance, σ_h^2, is more pronounced. Fig. 3 shows σ_h^2 as a function of the radial distance from the well, *r*, computed with three variogram models and three values of σ_Y^2 (0.1, 0.5 and 1.0) for TC1 (Fig. 3a) and TC2 (Fig. 3b). Fig. 3 also depicts the relative differences evaluated by Eq. (7) for TC1 (Fig. 3c) and TC2 (Fig. 3d) and σ_Y^2 = 1.0, together with the uncertainty bandwidths. These have been calculated by 10,000 Monte Carlo iterations. Here and in the following pictures we also report the uncertainty bandwidths since they allow discriminating between the relative differences due to the effect of the variogram model choice and those due to an incomplete convergence of the Monte Carlo results.
Similar results (not reported) have been obtained for the other σ_Y^2 values.

In general, $\wp^E$ and $\wp^S$ increase with σ_Y^2 for TC2, while for TC1 both the percentage differences seem not to be significantly affected by the heterogeneity of the conductivity field, at least until $\sigma_Y^2 = 1$.

Fig. 3a - c shows that the three variogram models provide very similar values of σ_h^2 near the boundary in TC1 (where the domain is larger, in terms of I_Y). In the middle portion of the aquifer ($3 < r < 7$) the results obtained with GV (Eq. 1) and EV (Eq. 2) coincide for all practical purposes ($\wp^E \cong 0$), while SV (Eq. 3) results in larger values of σ_h^2. Maximum values of the percentage differences (about 40%) have been obtained at radial distances from the well between one and two integral scales for all the values of σ_Y^2 tested.

In TC2 and at locations close to the well, the effect of the source on σ_h^2 is stronger using GV (Fig. 3d), resulting in the largest σ_h^2 values. Adoption of EV results in the smallest σ_h^2 values in most of the domain (except within a region close to the boundary), in analogy to what already observed for the mean hydraulic head. Similarly, the effect of the imposed deterministic hydraulic head at the external boundary on σ_h^2 is stronger when adopting a Gaussian variogram, as evidenced by the smallest values obtained for σ_h^2.

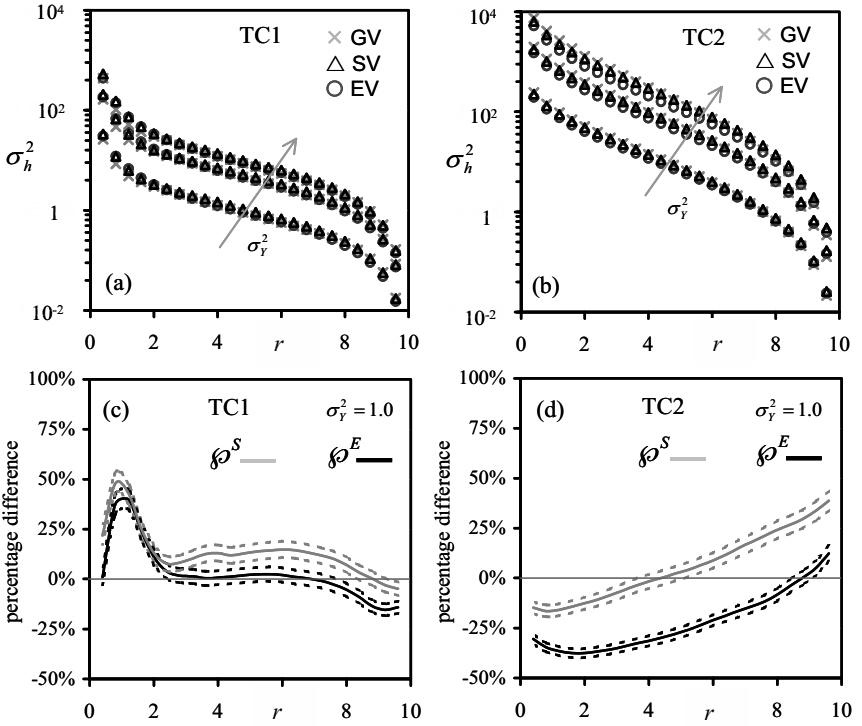


Fig. 3 Variance of the hydraulic head computed using 3 variogram models for **a)** TC1, and **b)** TC2 with σ_Y^2 = 0.1, 0.5, 1.0. Relative percentage differences in the hydraulic head variance for **c)** TC1 and **d)** TC2 with σ_Y^2 = 1.0. Dashed curves indicate the uncertainty bands.

3.3 Mean Travel time

Fig. 4 depicts the mean travel time as a function of the particle injection point, r_0, computed with the three variogram models adopted and three σ_Y^2 values for TC1 (Fig. 4a) and TC2 (Fig. 4b); for comparison, with a continuous line is also reported the solution obtained with an homogeneous deterministic field with constant transmissivity equal to the geometric mean of T. We observe that $\langle t \rangle$ increases with σ_Y^2; thus, as opposed to what observed for the mean hydraulic head, the aquifer heterogeneity affects the mean residence time, $\langle t \rangle$, not only for TC2 but also for TC1.

Fig. 4 also depicts the relative percentage differences in the mean residence time for TC1 (Fig. 4c) and TC2 (Fig. 4d) with σ_Y^2 = 1.0, together with the corre-

sponding uncertainty bandwidths. As in the case of the mean hydraulic head, the mean travel time is not significantly influenced by the variogram model adopted.

The percentage differences (Eq. 7) generally increase with σ_Y^2 regardless the domain size and reach a maximum value of about 5% for $\sigma_Y^2 = 1$. This value is of the same order of magnitude as the uncertainty bandwidths.

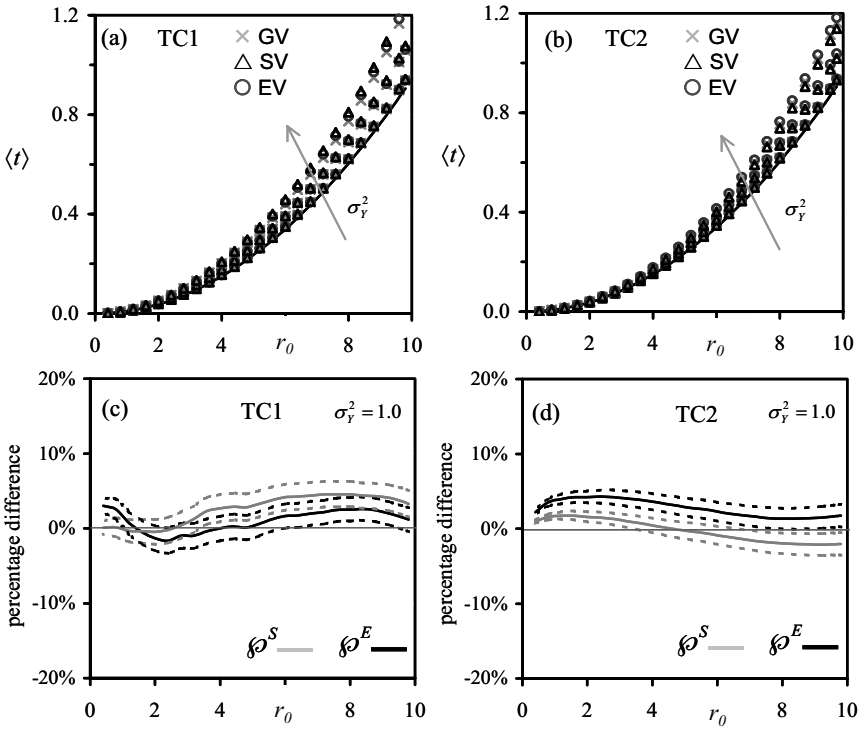


Fig. 4. Mean residence time as a function of particle released position computed using 3 variogram models for **a)** TC1 and **b)** TC2 with $\sigma_Y^2 = 0.1, 0.5, 1.0$. Relative percentage differences in the mean residence time for **c)** TC1 and **d)** TC2 with $\sigma_Y^2 = 1.0$. Dashed curves indicate the uncertainty bands.

3.4 Travel time variance

Similarly to what observed for σ_h^2, the impact of the variogram model on the travel time variance, σ_t^2, is significant. Fig. 5 shows σ_t^2 as a function of the particle injection point computed with the three variogram models and three values of σ_Y^2 for TC1 (Fig. 5a) and TC2 (Fig. 5b). Fig. 5 also shows the relative percentage

differences, $\wp^E$ and $\wp^S$, for TC1 (Fig. 5c) and TC2 (Fig. 5d) and $\sigma_Y^2 = 1.0$ together with the uncertainty bandwidths. In general $\wp^E$ and $\wp^S$ increase with σ_Y^2 for both the test cases analyzed.

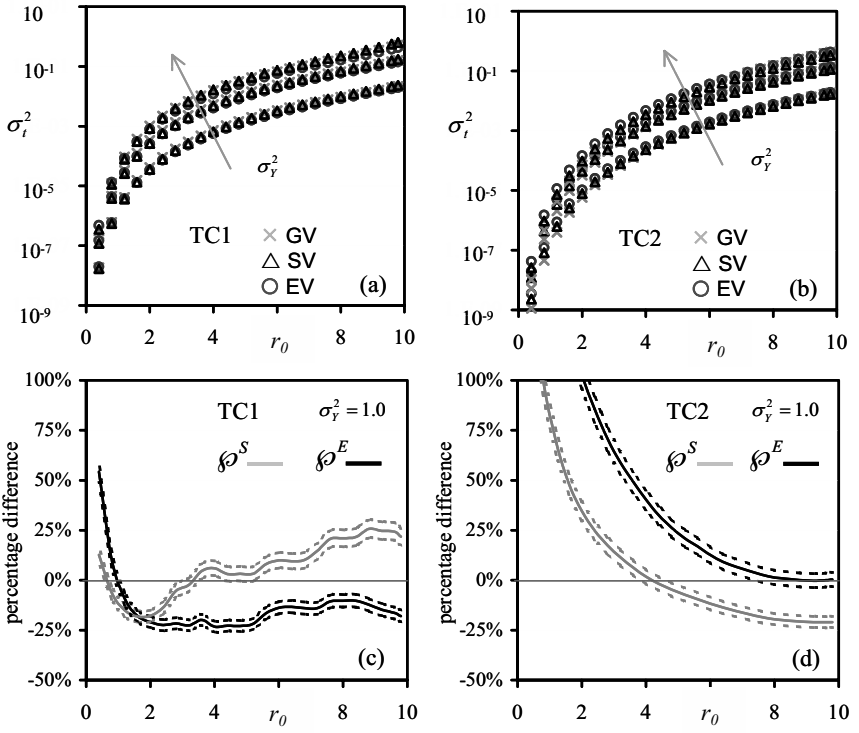


Fig. 5 Residence time variance as a function of particle released position computed using 3 variogram models for **a)** TC1 and **b)** TC2 with $\sigma_Y^2 = 0.1, 0.5, 1.0$. Relative percentage differences in the residence time variance for **c)** TC1 and **d)** TC2 with $\sigma_Y^2 = 1.0$. Dashed curves indicate the uncertainty bands.

Fig. 5a and 5c show that, for TC1 at locations close to the well, the choice of a Gaussian model yields smaller values of σ_t^2, in analogy to what we observed for σ_h^2. On the other hand, at larger distances from the well, the adoption of a Spherical model provides the largest σ_t^2 values, while the Exponential model results in the lowest prediction variances. The maximum differences have been obtained close to well, where $\wp^E$ attains values larger than 50%.

For TC2 (Fig. 5b and d), as observed for TC1, the maximum differences are reached for release position close to the source. Adopting a Gaussian model re-

sults in the smallest values of σ_t^2 in this part of the domain. The largest travel time variances are obtained using the Exponential model ($\wp^E$>100%). This large value for $\wp^E$ is related to the fact that all injection points in TC2 are very close to the well (in terms of the integral correlation scale, $r_0 \leq I_Y$) and the well (where $\sigma_t^2 = 0$) has a larger impact than in TC1.

4 Conclusions

We consider the effect of the choice of the log transmissivity variogram model on the estimation of the mean and variance of hydraulic head, *h*, and contaminant residence time, *t*. This is presented for two-dimensional steady convergent flow created by a well, located at the center of a randomly heterogeneous domain. The natural logarithm, *Y*, of transmissivity is modeled as a statistically homogeneous Gaussian random field. The effect of three different functional forms of γ_Y (Exponential, Gaussian and Spherical) is examined. Flow and particle movement were solved in a Monte Carlo framework. An extensive analysis of the stability of the moments (mean and variance) of *h* and *t* highlighted that as many as 10,000 iterations were necessary to carry out the present analysis. Our work leads to the following major conclusions:

- Ensemble mean of hydraulic head and travel time is not affected by the choice of the variogram model for fixed transmissivity geometric mean (T_G), variance (σ_Y^2) and integral scale (I_Y). This implies that numerical Monte Carlo results and analytical solutions which are usually presented in the literature for a particular covariance structure (e.g. Riva *et al*. 2001, Riva *et al.* 2002) together with their applications in pumping test analysis in order to obtain the statistical properties of conductivity fields (e.g. Neuman *et al.* 2004), have a more general validity.
- The influence of the model of γ_Y is relevant on the second order statistical moments of the variables of interest. Differences in the spatial distribution of hydraulic head (σ_h^2) and travel time (σ_t^2) variance due to different γ_Y models increase at locations near the well. Variance of hydraulic head and travel time are also affected by the domain size (*L*) and by the heterogeneity level. Generally, the choice of the variogram functional form has a more profound impact on σ_t^2 than on σ_h^2.

References

Ballio F, Guadagnini A (2004) Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology. Water Resources Research, (40)4, W04603, doi:10.1029/2003WR002876

Dagan G, Indelman P (1999) Reactive solute transport in flow between a recharging and a pumping well in a heterogeneous aquifer. Water Resources Research, 35(12): 3639-3647

Feyen L, Ribeiro Jr PJ, De Smedt F, Diggle PJ (2002) Bayesian methodology to stochastic capture zone determination: Conditioning on transmissivity measurements. Water Resources Research, 38(9), 1164, doi:10.1029/2001WR000950

Guadagnini A, Franzetti S (1999) Time-related Capture Zones for Contaminants in Randomly Heterogeneous Formations. Ground Water, 37(2): 253-260

Gómez-Hernández JJ, Journel AG (1993) Joint sequential simulation of multi-Gaussian field. In: Geostatitics Troia'92, vol 1. Ed Soares, 85-94

Hendricks Franssen HJ, Stauffer F, Kinzelbach W (2002) Influence of uncertainty of mean transmisivity, transmissivty variogram and boundary condition on estimation of wll capture zones. 4th European Conference on Geostatistics for Environmental Applications, Geoenv2002, 223-234

Neuman SP, Guadagnini A, Riva M (2004) Type-curve estimation of statistical heterogeneity, Water Resour. Res., 40, W04201, doi: 10. 1029 / 2003 WR 002 405

Riva M, Guadagnini A, Ballio F (1999) Time related capture zones for radial flow in two-dimensional randomly heterogeneous media Stochastic Environmental Research and Risk Assessment, 13(3): 217-230

Riva M, Guadagnini A, Neuman SP, Franzetti S (2001) Radial flow in a bounded randomly heterogeneous aquifer. Transport in Porous Media, 45: 139–193

Riva M, Sanchez-Vila X, De Simoni M, Guadagnini A, Willmann M (2002) Effect of heterogeneity on aquifer reclamation time, 4th European Conference on Geostatistics for Environmental Applications, Geoenv2002, 259-270

Salandin P, Rinaldo A (1989) The influence of the form of the log-conductivity covariance on non-fickian dispersion in random permeability fields, International Journal for Numerical Methods in Engineering, 27: 185-193

Sanchez-Vila X, Meier PM, Carrera J (1999) Pumping tests in heterogeneous aquifers: An analytical study of what can be obtained from their interpretation using Jacob's method. Water Resources Research, 35 (4): 943-952

van Leeuwen M, Te Stroet CBM, Butler AP, Tompkins JA (2000) Stochastic determination of well capture zones conditioned on regular grids of transmissivity measurements. Water Resources Research, 36(4): 949-957

Woodbury AD, Rubin Y (2000) A full-Bayesian approach to parameter inference from tracer travel time moments and investigation of scale effects at the Cape Cod experimental site. Water Resources Research 36(1): 159-171

Woodbury AD, Ulyrich TG (2000) A full-Bayesian approach to the groundwater inverse problem for steady state flow. Water Resources Research 36(8): 2081-2093.

Strategies to determine dispersivities in heterogeneous aquifers

D. Fernàndez-Garcia and J. Jaime Gómez-Hernández

Universidad Politécnica de Valencia, Institute of Water and Environmental Engineering, Camino de Vera, s/n. 46022 Valencia, SPAIN.

1 Introduction

Prediction of the fate and transport of dissolved contaminants in groundwater is required in conducting risk analysis and in decision-making in problems involving hazardous waste management and remediation of contaminated sites. In order to make such predictions, it is necessary to estimate dispersivity, which is the aquifer parameter that measures the spread of a contaminant plume. Field tracer tests constitute a practical tool to estimate field-scale dispersivities by which the underlying heterogeneous structure is directly incorporated into an effective parameter. However, tracer tests can be materialized in the field using different site-specific schemes that involve different flow configurations, tracer source sizes, and methods of sampling and interpretation. Aquifer parameters estimated from a given selected tracer test scheme may significantly differ from one scheme to another due to natural heterogeneity. In this paper, the influence of the method of interpretation on dispersivity estimated from tracer tests is investigated in a heterogeneous porous medium, which is generated considering the natural log of hydraulic conductivity as a second-order stationary random function.

Two major types of field tracer tests may be distinguished based on the flow configuration: (a) In natural-gradient tracer tests, the tracer is added into the subsurface by means of injection wells and is let to freely move along with the natural groundwater flow system. This type of tracer tests has the disadvantage that groundwater natural velocities are frequently very small giving large test duration times, i.e. hundreds of days (Mackay *et al.* 1986, LeBlanc *et al.* 1991), and that the groundwater flow direction is not well controlled being difficult to design a good well layout configuration that truly captures the tracer plume; (b) In forced-gradient tracer test, the tracer is forced to migrate through an artificial flow system that is conveniently modified by means of water injection and/or pumping wells. It provides better controlled conditions but the groundwater flow system is significantly different from the natural one. Many different forced-gradient flow configurations can be envisioned: convergent flow tracer tests, divergent flow tracer tests, single-well tracer tests, dipole tracer tests, etcetera. In this respect, recent uniform flow tracer tests and forced-gradient tracer tests conducted under well controlled laboratory conditions in the same heterogeneous test aquifer have

shown that dispersivities for conservative and sorptive tracers and retardation factors can be significantly modified by simply changing the flow configuration (Fernàndez-Garcia *et al.* 2004).

Three different methodologies to interpret tracer field concentration data can be distinguished: (a) curve matching techniques (b) the method of moments, and (c) model calibration. The first approach is the most common among practitioners. Field data is somehow fitted to analytical or numerical model solutions that are commonly given in the form of dimensionless type curves (Sauty 1980, Carrera and Walters 1985, Welty and Gelhar 1994). The second approach uses statistical information on the concentration distribution (either in space or in time) to yield aquifer properties (Freyberg 1986, Goltz and Roberts 1987, Das *et al.* 2002). Moments provide information about the size and shape of the concentration distribution such that if all moments were known the distribution would be completely defined (Das *et al.* 2002). Aquifer properties such as dispersivity may be inferred using the first two moments: the first moment is a measure of the center of the distribution and is related to the advective solute transport. The second moment is a measure of the spread of the concentration distribution and describes the solute spreading mechanism in the subsurface. Finally, in the third approach, calibration of an analytical or numerical model is used to infer aquifer parameters (Poeter and Hill 1997, Carrera *et al.* 1997). Numerical simulations by Jan Vanderborght *et al.* (1998) showed that curve matching techniques can yield dispersivities considerably smaller than those obtained using the method of moments.

It has been widely recognized nowadays that the influence of the source size on aquifer parameters such as dispersivity is of great importance. Tracer source sizes smaller than the characteristic scale of heterogeneity sample a small portion of the spatial variability of groundwater flow velocity, which is expected to result in a reduction of "macrodispersivity" (e.g., Rajaram and Gelhar 1993, Dentz *et al.* 2000). This theoretical result has been experimentally validated by Fernàndez-Garcia *et al.* (2004).

Two different sampling procedures can be distinguished. Several authors (Kreft and Zuber 1978, Parker and van Genuchten 1984) have established that concentrations obtained from observation wells are better described by flux-averaged concentrations (C_F), whereas concentrations from multilevel samplers that extract a certain volume of pore water are believed to yield volume-averaged concentrations (C_V). Flux-averaged concentrations are defined as the ratio between the fluxes of the solute mass and the groundwater mass, that are passing through a representative elemental area (REA) at a given time; at the fluid continuum scale, they are physically seen as an average concentration weighted by the fluid velocities (Parker and van Genuchten 1984).

Volume-averaged concentrations are seen as the arithmetic average of the concentrations in the pore space of the Representative Elemental Volume (REV),

$$C_F = \frac{\int_{REA} \vec{v}_p \cdot C_p \cdot \vec{n} \cdot \phi \cdot dA}{\int_{REA} \vec{v}_p \cdot \vec{n} \cdot \phi \cdot dA} \qquad C_V = \frac{\int_{REV} C_p \cdot \phi \cdot dV}{\phi V_{REV}} \tag{1}$$

where C_F and C_V stand for flux- and volume-averaged concentrations, v_p and C_p are respectively the fluid velocity and concentration at the fluid continuum scale, ϕ is the porosity, and n is a unit vector perpendicular to the Representative Elemental Area (REA). The relationship between these two types of concentration is given by Kreft and Zuber (1978), and is written after neglecting molecular diffusion as

$$\sum_i v_i C_F n_i = \sum_i v_i C_V n_i - \sum_i \sum_j \alpha_{ij} |\vec{v}| \frac{\partial C_V}{\partial x_j} n_i \tag{2}$$

Where v_i is the ith-component of the average interstitial velocity, and α_{ij} is the dispersion tensor. In addition to the difference between flux- and volume-averaged concentrations, the sampling volume used during field tracer test can also largely influence the ultimate parameter estimate. For instance, the spread in a breakthrough curve obtained from a sampling volume much smaller than the characteristic scale of heterogeneity is a measure of mixing and dilution occurring along the travel path of the tracer, whereas the information expressed by the averaged breakthrough curve over a control plane quantifies the overall spreading of the tracer plume (Cirpka and Kitanidis 2000).

Traditionally, it is implicitly assumed that those dispersivity estimates resulting from field tracer tests designed with deep-penetrating observation wells that sample flux-averaged concentrations with time at fix locations, and those based upon multilevel samplers which monitor the distribution of volume-averaged concentrations in space at given times, leads to similar effective input parameters. In an ideal homogeneous porous media, both dispersivities are equal provided that the mean plume travel distance (L) is much larger than the Peclet number, $Pe = L/\alpha_L > 100$ (Parker and van Genuchten 1984), considering that pore-scale dispersivities are in the range of 0.1 to 1 cm, the latter restriction is generally always the case. However, in a heterogeneous porous media in which dispersivity is a scale dependent parameter and solute particles may be favoring different preferential paths, the relationship between these two dispersivities is unclear.

The objective of this paper is to provide some insights into the true meaning of these two conceptually different dispersivity estimates in heterogeneous aquifers. To achieve this, natural-gradient tracer tests were numerically simulated in a heterogeneous porous media using different source sizes. The advection-dispersion equation is solved in a three-dimensional lnK random field; the heterogeneous structure is represented by a spatially variation of the natural log of hydraulic conductivities within the domain, which follows a second-order stationary random field with an isotropic exponential covariance function. Temporal moments from flux-averaged concentrations breakthrough curves are calculated at several known control planes that are perpendicular to the mean flow direction. These temporal

moments are used to evaluate dispersivities from temporal moments of breakthrough curves (recorded flux-averaged concentrations with time at given locations) as a function of distance. In addition, spatial moments of tracer plumes are computed as a function of travel times using several snapshots of the solute plume (volume-averaged concentrations distributed in space). These spatial moments are then used to estimate dispersivities from spatial moments. Comparison between dispersivities from temporal and spatial moments is then examined as a function of the source size for a given realization of the lnK stochastic process.

2 Design of computational investigations

2.1 Numerical features

The computational domain was discretized into a regular mesh formed by 250×250×200 parallelepiped cells. The hydraulic conductivity field was generated with a variance of the natural log of hydraulic conductivity (σ^2_{lnK}) of 1.0 and a correlation scale (λ) of 176 cm. The resolution was of five grid cells within a correlation scale in all directions.

The hydraulic conductivity field was incorporated into a seven-point finite difference ground-water flow model, MODFLOW2000 (Harbaugh *et al.* 2000). The model calculates the flow rates at the grid interfaces, which were employed to compute the velocity field considering porosity as a homogeneous variable with a value of 0.35. This velocity field was then used in a Random Walk Particle Tracking transport code (Tompson and Gelhar 1990, Wen and Gómez-Hernández 1996, Labolle *et al.* 1996) that solves the advection-dispersion equation. This method simulates the solute migration by partitioning the solute mass into a large number of representative particles (the number of particles used for all simulations is 10000); moving particles with the velocity field simulates advection, whereas a Brownian motion is responsible for dispersion. For the type of analysis done in this research, this method provides better computational efficiencies than traditional numerical models. This method is well suited for large number of simulations for the simultaneous computation of temporal and spatial moments. In addition, it is free of numerical dispersion. Longitudinal and transverse pore-scale dispersivities were respectively 0.2 and 0.02 cm, which are based on laboratory homogeneous sand column tracer studies (Fernàndez-Garcia *et al.* 2004). Upstream and downstream boundaries are specified as constant heads, such that the hydraulic gradient in the mean flow direction is of 0.004. Mean flow direction is aligned with the x_1 coordinate. No-flow conditions are prescribed at the transverse, top and bottom boundaries.

Initially, particles were randomly distributed (uniformly) in a plane transverse to the mean flow direction. This plane was located three correlation scales away from the upgradient boundary to avoid boundary effects (Rubin and Dagan 1988 and 1989). The shape of the particle source is a rectangle centered within this plane. The source size was progressively increased from a point source to 40 cor-

relation scales in the transverse direction to the mean flow and 35 correlation scales in the vertical direction. Spatial moments and temporal moments were based on particle distributions at times when none of the particles has exited the domain.

2.2 Evaluation of dispersivities from temporal moments of breakthrough curves

Temporal moments of flux-averaged concentrations breakthrough curves obtained at given control planes can be easily calculated with particle tracking codes. Flux-averaged concentrations are proportional to the arrival time probability distribution function of particles passing through a control plane (Shapiro and Cvetkovic 1988), such that it is only necessary to track the first passage time of particles passing through the control plane to estimate these temporal moments. Statistical moments of breakthrough curves can be estimated without having to evaluate the probability density function of the arrival times as

$$\mu'_n(x_1) = \frac{\int_0^\infty t^n C_F(x_1,t)dt}{\int_0^\infty C_F(x_1,t)dt} \approx \frac{1}{NP_a}\sum_{k=1}^{NP_a}\left(t_p^{(k)}(x_1)\right)^n \tag{3}$$

where $\mu'_n(x_1)$ is the nth temporal moment of the breakthrough curve obtained at the x_1-control plane, $C_F(x_1,t)$ denotes flux-averaged concentrations obtained at the x_1-control plane, $t_p^{(k)}(x_1)$ is the first arrival passage time of the kth particle crossing the x_1-control plane, and NP_a is the total number of particles arrived at the control plane. This approach bypasses the need to compute smooth concentrations and avoids the problems involved in constructing a histogram. The nth central temporal moment is defined as the moment about the mean

$$\mu_n(x_1) = \frac{\int_0^\infty \left(t-\mu'_1(x_1)\right)^n C_F(x_1,t)dt}{\int_0^\infty C_F(x_1,t)dt} \tag{4}$$

First-passage arrival times were computed by seeking the time at which the particle intersects the control plane. This was done based upon the particle location information right before $X_p(t^n)$ and after $X_p(t^{n+1})$ passing through the control plane. Central temporal moments were computed using the relationship by Kendall and Stuart (1977), written as

$$\mu_n(x_1) = \sum_{r=0}^{n}\binom{n}{r}\cdot\mu'_{n-r}(x_1)\cdot\left(-\mu'_1(x_1)\right)^r \tag{5}$$

Dispersivities from temporal moments of breakthrough curves were calculated following Aris (1958), Valocchi (1985), and Goltz and Roberts (1987) as

$$A_{11}(x_1) = \frac{x_1}{2} \frac{\mu_2(x_1) - \mu_2(0)}{\left(\mu_1'(x_1) - \mu_1'(0)\right)^2} \tag{6}$$

where x_1 is the x_1-distance from tracer source to the control plane, and $\mu'_1(x_1)$ and $\mu_2(x_1)$ are the first temporal moment and the second central temporal moment at the x_1-control plane, respectively.

2.3 Evaluation of dispersivities from spatial moments of tracer plumes

Spatial moments of volume-averaged concentrations distributed in space were calculated following the approach developed by Tompson and Gelhar (1990). Spatial moments were calculated from snapshots of the distribution of particles at given times as follows:

$$X_{G,1}(t) = \frac{\int x_1 C_V\left(\vec{x}, t\right) dV}{\int C_V\left(\vec{x}, t\right) dV} \approx \frac{1}{NP_t} \sum_{k=1}^{NP_t} X_{p,1}^{(k)}(t) \tag{7}$$

$$S_{11}(t) = \frac{\int \left(x_1 - X_{G,1}\right)^2 C_V\left(\vec{x}, t\right) dV}{\int C_V\left(\vec{x}, t\right) dV} \approx \frac{1}{NP_t} \sum_{k=1}^{NP_t} X_{p,1}^{(k)}(t) X_{p,1}^{(k)}(t) - X_{G,1}(t) X_{G,1}(t) \tag{8}$$

where integrals with respect to dV denote the integration over all space, $X_{G,1}$ is the x_1 coordinate position of the plume center of mass, S_{11} is the second central spatial moment in the mean flow direction associated with the distribution of particles at a given time, NP_t is the number of particles in the system at time t, and $X_{p,1}^{(k)}$ is the x_1 coordinate position of the kth particle.

Dispersivities associated with spatial moments and volume-averaged concentrations were calculated following Freyberg (1986):

$$A_{11}(t) = \frac{1}{2}\frac{S_{11}(t) - S_{11}(0)}{X_{G,1}(t) - X_{G,1}(0)} \tag{9}$$

where t is the elapsed time from particles injection, and $S_{11}(t)$ is the second spatial moment in the mean flow direction at elapsed time t.

3 Results of computational investigations

Fig. 1 compares the effect of the source size on the scale-dependence of dispersivities estimated from temporal moments of flux-averaged concentrations breakthrough curves obtained at control planes with those dispersivities calculated from spatial moments of volume-averaged concentrations of the tracer plume for a single realization of the lnK random field. Dispersivities are presented as the ratio of the simulated A_{11} values to the product of $\sigma^2_{lnK}\lambda$, and are plotted as a function of x_1/λ (where x_1 is the x_1-distance to the control plane) for temporal moments and as a function of $X_{G,1}/\lambda$ for spatial moments. Simulated A_{11} values are plotted with Dagan's (1984) first-order stochastic analytical solution of dispersivity to better appreciate the effect of the source size on 'macrodispersivity'. In accordance to stochastic theories (e.g., Rajaram and Gelhar 1993, Dentz *et al.* 2000) and experimental tracer studies (Fernàndez-Garcia *et al.* 2004), Fig. 1 shows that, for the same test scale, small source tracer tests exhibit smaller dispersivities than those obtained from larger sources, reflecting that small sources cannot entirely capture the underlying heterogeneous structure of the porous media. As the size of the plume increases (e.g., for large source sizes and mean travel distances), the tracer plume samples a larger portion of heterogeneity and simulated dispersivities approach a more representative value for the entire system. Likewise, as the plume size increases, the erratic behavior of dispersivity expected in a single realization of the aquifer due to non-ergodic effects is reduced and simulated dispersivities approach Dagan's model (1984).

It is seen that significant discrepancies between dispersivities estimated from two different methods of sampling and interpretation of field tracer tests can exist for all source sizes. For instance, dispersivities estimated from temporal moments and flux-averaged concentrations associated with a source of size $I_{x2}=40\lambda$ and $I_{x3}=35\lambda$ (I_{x2} and I_{x3} denote respectively the source size in the x_2 and x_3 directions) were 1.2 to 1.7 times larger than those dispersivities calculated from spatial moments and volume-averaged concentrations within the first 10 correlation scales. However, these differences seem to diminish for large sources and travel distances. At large travel distances, dispersivities estimated from temporal and spatial moments in transport simulations with source sizes greater than $I_{x2}=I_{x3}=20\lambda$ were very similar.

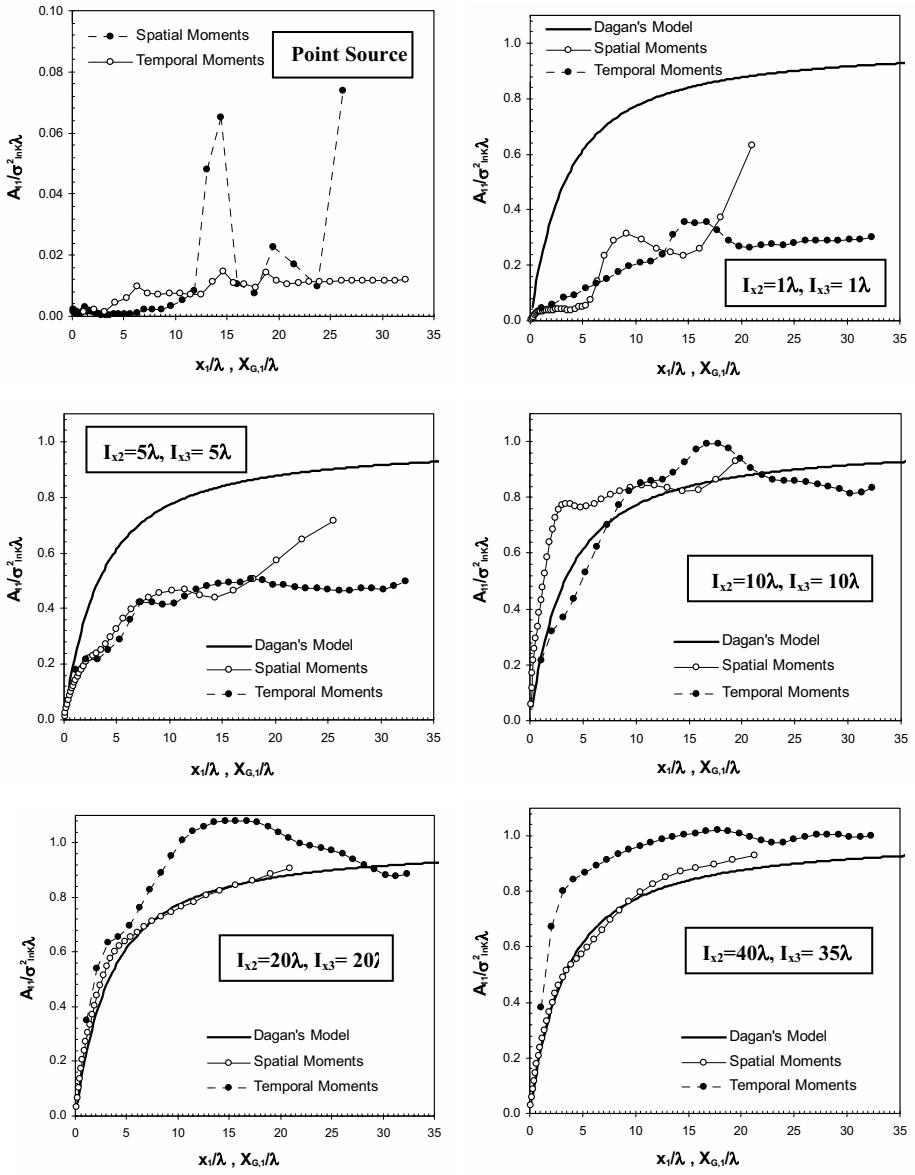


Fig. 1. Effect of the source size on the scale dependence of dispersivity estimated from temporal and spatial moments in a single realization of an isotropic heterogeneous random lnK field with σ^2_{lnK}=1.0 (I_{x2} and I_{x3} denote the source size transverse and vertical to the mean flow direction, respectively), Comparison of simulated dispersivities with Dagan's model (1984).

4 Summary and conclusions

Numerical simulations of uniform-flow conservative tracer tests associated with two different methods of sampling and interpretation to estimate aquifer dispersivities were conducted to examine the importance of the method of sampling and interpretation on the design of field tracer tests in heterogeneous porous media. The scale-dependence of dispersivity associated with temporal moments of flux-averaged concentrations and spatial moments of tracer plumes were evaluated and compared as a function of the source size for a given realization of the lnK random field. It is shown that caution must be taken during the interpretation of field tracer tests. Dispersive processes occurring in the aquifer due to heterogeneity are captured differently in breakthrough curves than in snapshots of the tracer plume. Consequently, dispersivities from temporal moments were found significantly different than dispersivities from spatial moments for all source sizes. For instance, dispersivities estimated from temporal moments of breakthrough curves were seen 1.2 to 1.7 times larger than those dispersivities calculated from spatial moments of volume-averaged concentrations snapshots within the first 10 correlation scales for large sources. However, these differences diminished as the size of the plume increases (e.g., for large source sizes and travel distances). Hence, from the practical standpoint, it is seen that the method of interpretation can play an important role on the final estimated dispersivity values that will be used to assess groundwater remediation problems. It is concluded that the selection of the tracer test scheme should be contemplated during characterization of the aquifer by means of field tracer tests.

It is noted that these simulations were based upon a single realization and thereby further analysis should be conducted to systematically evaluate these effects in an average sense examining numerous realizations of the lnK random field.

References

Aris R (1958) On the dispersion of linear kinematic waves. Proc. R. Soc. London, Series A, 245: 268-277

Carrera J, Walters GR (1985) Theoretical developments regarding simulation and analysis of convergent-flow tracer tests: Technical Report for Sandia National Laboratories. Technical University of Catalonia, Barcelona

Carrera J, Medina A, Axness C, Zimmerman T (1997) Formulations and computational issues of the inversion of random fields, in Subsurface Flow and Transport: A stochastic Approach. International Hydrology Series. Dagan G and Neuman SP (eds.), 62-79

Cirpka OA, Kitanidis PK (2000) Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments, Water Resour. Res., 36(5): 1221-1236

Dagan, G (1984) Solute transport in heterogeneous porous formations. J. Fluid Mech., 145: 151-177

Das BS, Govindaraju RS, Kluitenberg GJ, Valocchi AJ, Wraith JM (2002) Theory and applications of time moment analysis to study the fate of reactive solutes in soil. Stochastic Methods in Subsurface Contaminant Hydrology, ASCE press, Govindaraju RS (ed.), 239-279

Dentz M, Kinzelbach H, Attinger S, Kinzelbach W (2000) Temporal behavior of a solute cloud in a heterogeneous porous medium 1. Point-like injection, Water Resour. Res., 36(12): 3591-3604

Fernàndez-Garcia D, Illangasekare TH, Rajaram H (2004) Conservative and sorptive forced-gradient and uniform flow tracer tests in a three-dimensional laboratory test aquifer. Water Resour. Res., 40, in print

Freyberg DL (1986) A natural gradient experiment on solute transport in a sand aquifer. 2. Spatial moments and the advection and dispersion of nonreactive tracers. Water Resour. Res., 22(13): 2031-2046

Goltz MN, Roberts PV (1987) Using the method of moments to analyze three-dimensional diffusion-limited solute transport from temporal and spatial perspectives. Water Resour. Res., 23(8): 1575-1585

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, The U.S. Geological Survey Modular Ground-Water Model–user guide to modularization concepts and the ground-water flow process. U. S. Geol. Surv. Open File Rep., 00-92, p. 121

Kendall M, Stuart A (1977) The Advanced Theory of Statistics. Macmillan, New York.

Kreft A, Zuber A (1978) On the physical meaning of the dispersion equation and its solution for different initial and boundary conditions. Chem. Eng. Sci., 33: 1471-1480

Labolle EM, Fogg GE, Tompson AFB (1996) Random-walk simulation of transport in heterogeneous porous media: local mass-conservation problem and implementation methods. Water Resour. Res., 32(3): 583-593

Mackay DM, Freyberg DL, Roberts PV, Cherry JA (1986) A Natural Gradient Experiment on Solute Transport in a Sand Aquifer, 1. Approach and Overview of Plume Movement. Water Resour. Res., 22 (13): 2017-2029

LeBlanc DR, Garabedian SP, Hess KM, Gelhar LW, Quadri RD, Stollenwerk KG, Wood WW (1991) Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 1. Experimental Design and Observed Tracer Movement. Water Resour. Res., 27(5): 895-910

Parker JC, van Genuchten MT (1984) Flux-averaged and volume-averaged concentrations in continuum approaches to solute transport. Water Resour. Res., 20(7): 866-872.

Poeter EP, Hill MC (1997) Inverse Models: A necessary next step in ground-water modeling. Ground Water, 35(2): 250-260

Rajaram H, Gelhar LW (1993) Plume scale-dependent dispersion in heterogeneous aquifers 2. Eulerian analysis and three-dimensional aquifers, Water Resour. Res., 29(9): 3261-3276

Rubin Y, Dagan G (1989) Stochastic analysis of boundary effects on head spatial variability in heterogeneous aquifers, 1. Impervious boundary, Water Resour. Res., 25(4): 707-712

Rubin Y, Dagan G (1988) Stochastic analysis of boundary effects on head spatial variability in heterogeneous aquifers, 1. Constant head boundary, Water Resour. Res., 24(10): 1689-1697

Sauty J P (1980) An analysis of hydrodispersive transfer in aquifers. Water Resour. Res, 16(1): 145-158

Shapiro A M, Cvetkovic V D (1988) Stochastic analysis of solute arrival time in heterogeneous porous media. Water Resour. Res., 24(10): 1711-1718

Tompson A F B, Gelhar L W (1990) Numerical simulation of solute transport in three-dimensional, randomly heterogeneous porous media. Water Resour. Res., 26(10): 2541-2562

Valocchi A (1985) Validity of the local equilibrium assumption for modeling sorbing solute transport through homogeneous soils. Water Resour. Res., 21(6): 808-820

Vanderborght J, Mallants D, Feyen J (1998) Solute transport in a heterogeneous soil for boundary and initial conditions: Evaluation of first-order approximations, Water Resour. Res., 34(12):3255-3270, 1998

Welty C, Gelhar L W (1994) Evaluation of longitudinal dispersivity from nonuniform flow tracer tests. Journal of Hydrology, 153: 71-102

Wen X-H, Gómez-Hernández J J (1996) The constant displacement scheme for tracking particles in heterogeneous aquifers. Ground Water, 34(1): 135-142.

Solving the groundwater inverse problem by successive flux estimation

P. Pasquier and D. Marcotte

École Polytechnique de Montréal, Département des génies civil, géologique et des mines, C.P. 6079, Succ. Centre-ville, Montréal, Qc, CANADA, H3C 3A7,
e-mail : philippe.pasquier@polymtl.ca

1 Introduction

Numerical modelling of groundwater systems is a task frequently performed by hydrogeologists to determine groundwater flow paths and travel times and to asses the efficiency of possible remediation measures. The numerical solution of such systems requires the transmissivity field to be perfectly known in the whole domain. Scarce and often unreliable transmissivity data hinders direct reconstruction of transmissivity fields which reproduce numerically the observed heads. In order for the simulation prediction to be as accurate as possible, the spatial distribution of aquifer hydraulic parameters must be known. This is usually done by solving the so-called inverse problem. Several approaches have been suggested for the solution of this ill-posed problem. The zonation method (Carrera and Neuman 1986), the pilot point method (de Marsily *et al.* 1984) and the self-calibrated method (Gòmez-Hernàndez *et al.* 1997) are among the most used inverse procedures. Inverse methods which assume the real head field to be perfectly known are presented in Guo and Zhang (2000), Ponzini and Lozej (1982) and Sagar *et al.* (1975).

The inverse problem is typically an ill-posed problem. Solution non-uniqueness is often addressed in a stochastic framework to take into account the uncertainty of the solution. Since direct evaluation is time consuming, a fast and robust inversion algorithm is essential to assess uncertainty for practical applications.

The main purpose of this paper is to present an adaptation of the comparison model method (Ponzini and Lozej 1982) which efficiently calibrates groundwater numerical models under steady-state conditions. The modification consists of the inclusion of a damping function which ensures rapid calibration while preserving the structure of the transmissivity field initially given to the algorithm. Inversion is performed using kriged head field. To account for the effect of pumping wells and boundary conditions, kriging is modified following the methodology described by Brochu and Marcotte (2003). The effect of head field estimation errors is evaluated using a synthetic study. The approach is tested on a synthetic model.

2 Successive Flux Estimation

2.1 Methodology

Without loss of generality, consider a saturated and incompressible groundwater flow in a confined two-dimensional aquifer with regional and radial flow to a pumping well. Under appropriate boundary conditions, the state equation describing this model is given by:

$$\begin{aligned} &\nabla \cdot (T \cdot \nabla h) = 0, \quad (x,y) \in (\Omega) \\ &h\,|_{(\Gamma_1)} = \hat{h}(x,y), \qquad (x,y) \in (\Gamma_1) \\ &-(T \cdot \nabla h) \cdot n|_{(\Gamma_2)} = 0, \qquad (x,y) \in (\Gamma_2) \\ &-(T \cdot \nabla h) \cdot n|_{(\Gamma_3)} = \frac{Q}{2\pi r_w}, \qquad (x,y) \in (\Gamma_3) \end{aligned} \tag{1}$$

where h is the hydraulic head [L], T is the isotropic transmissivity tensor [L^2T^{-1}], Q is the well pumping rate [L^3T^{-1}], r_w is the well radius [L], n is the unit vector normal to the boundary (Γ) and (Ω) is the flow domain bounded by $(\Gamma) = (\Gamma_1) \cup (\Gamma_2) \cup (\Gamma_3)$.

Assuming h to be known over (Ω), Emsellem and de Marsily (1971) show that the inverse problem reduces to a Cauchy problem if the flux along a line intersecting every streamline in (Ω) is known or if one T value is known on every flow line. The solution to such a problem is given by:

$$T(s,\psi) = \frac{-q(\psi)}{\partial h(s,\psi)/\partial s} \tag{2}$$

where s is the distance along a flow line, ψ is the isopiezometric line and q [L^2T^{-1}] is the flux per unit height along a streamline.

To solve Eq. 2, an estimation of h and q must be available on (Ω). Realistic head field estimations can be easily obtained by kriging (Brochu and Marcotte 2003, Tonkin and Larson 2002) but obtaining fluxes geostatistically is difficult under realistic conditions. The flux is seldom known far from Neumann's boundary and even if flux measurements are available, the estimated vector field must be conservative, a condition which is, presumably, difficult to satisfy in real cases.

To obtain a robust flux field respecting the boundary conditions as well as the condition $\nabla \vec{q} = 0$, the comparison model approach (Ponzini and Lozej 1982) is adopted here. This numerical model has the same geometry, the same boundary conditions, the same numerical discretization and the same state equation as the real numerical model that one attempts to calibrate. Using a seed transmissivity field (T^o) as input in the comparison model method, the flux vector field can be obtained by solving the direct problem (Eq.1) once. To obtain a formulation consistent with the numerical discretization of the direct problem, Eq. 2 is expressed discretely in terms of the mesh element j:

$$T_j = \frac{\left\|\vec{q}_j\right\|_2}{\left\|\nabla h_j\right\|_2} \tag{3}$$

Using Darcy's law to express the flux obtained by the comparison model method results in:

$$T_j = T_j^o \frac{\left\|\nabla h_j^o\right\|_2}{\left\|\nabla \hat{h}_j\right\|_2} \tag{4}$$

where h^o is the solution of Eq. 1 obtained with $T = T^o$, and $\hat{h}$ is the real head field or its estimation. Since an estimation of q is used, the Cauchy problem cannot be solved directly. Thus, the transmissivities given by Eq. 4 will not reproduce the head data if input in Eq. 1. Instead, successive flux field estimations are obtained iteratively. New transmissivity estimations are obtained at every iteration by solving as many "discrete" Cauchy problems. Rearranging Eq. 4 yields:

$$T_j^{i+1} = T_j^i \cdot \frac{\left\|\nabla h_j^i\right\|_2}{\left\|\nabla \hat{h}_j\right\|_2} \quad i \geq 0, 1 \leq j \leq n_{el} \tag{5}$$

where i is the iteration counter, j is the cell index and n_{el} is the total number of elements in the model.

At this point, Eq. 5 is similar to that of Ponzini and Lozej (1982). This approach was found capable of reproducing the head data but the corresponding transmissivity field was often very unrealistic. To reduce this effect, a damping factor (β) is used to suppress strong transmissivity modifications during the first iterations. It was found that reducing the damping factor at each iteration ensures rapid calibration while maintaining the structure of the seed transmissivity field. Moreover, the damping factor ensures convergence and suppresses oscillation of the solution in areas of low gradient. It also accounts for errors in the estimation of $\hat{h}$. The damping factor is reduced exponentially:

$$\beta^i = \beta^o e^{-3i/a} \tag{6}$$

where β^o is the initial damping factor, β^i is the damping factor at iteration i and a is the iteration range. When $i=a$, the damping factor is reduced to 5% of its initial value. Incorporating β^i in Eq. 5 results in:

$$T_j^{i+1} = T_j^i \cdot \frac{\left\|\nabla h_j^i\right\|_2 + \beta^i}{\left\|\nabla \hat{h}_j\right\|_2 + \beta^i} \quad i \geq 0, 1 \leq j \leq n_{el} \tag{7}$$

Using a seed transmissivity field at $i=0$, this relation is iteratively applied simultaneously at every mesh element until a satisfactory match between calculated

and observed head fields is reached. A schematic description of the algorithm applied to the finite element method is presented in Fig. 1.

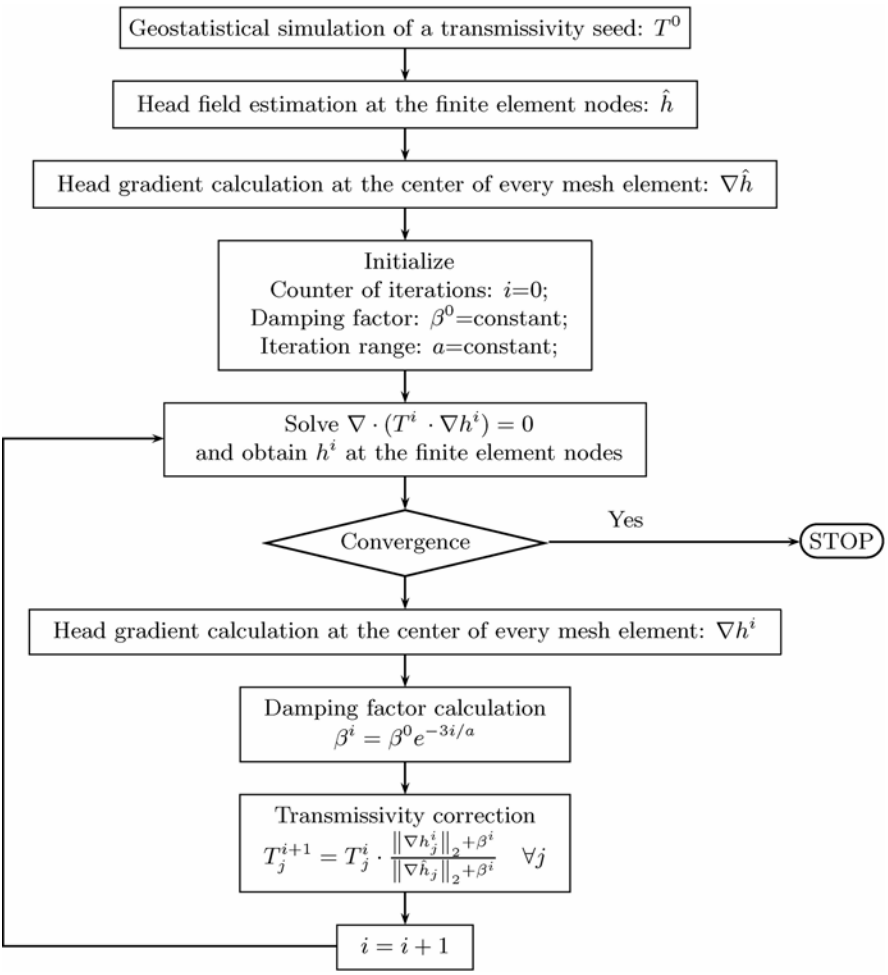


Fig. 1. Schematic description of the algorithm

2.2 Head field estimation

To obtain a realistic estimation of the head field under steady-state conditions in the presence of a pumping well and impervious boundaries, Brochu and Marcotte (2003) and Brochu (2002) used the Bear and Jacob (1965) analytical solution and the concept of double points (Chilès and Delfiner 1999, reporting Delhomme 1979).

Bear and Jacob (1965) showed that under steady-state conditions, the head of a bi-dimensional confined aquifer can be expressed by:

$$h(x,y) = H_w - (x - x_w)\cdot\frac{\partial H}{\partial x} - (y - y_w)\cdot\frac{\partial H}{\partial y} - \frac{Q}{4\pi T}\ln\left(\frac{r^2}{r_w^2}\right) \tag{8}$$

where H is the head field before pumping, H_w is the initial head in a well of radius r_w located at (x_w,y_w), Q is the pumping rate, T is the aquifer constant transmissivity and r is the distance between a point and the well.

To obtain an unbiased kriging estimate of $h(x,y)$, matrices of drift coefficients are added to the covariance matrix (Eq. 9 and 10). The first column corresponds to the unbiasedness conditions for any constant term. The next two columns take care of the regional head gradients $\partial H/\partial x$ and $\partial H/\partial y$. The fourth column describes the influence of the well at an observation point located at a distance r. The matrix corresponding to n_{obs} head observation data is:

$$F_h = \begin{bmatrix} 1 & x_1 & y_1 & \ln(r_1^2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_{obs}} & y_{n_{obs}} & \ln(r_{n_{obs}}^2) \end{bmatrix} \tag{9}$$

and the matrix for double points is given by:

$$F_\Delta = \begin{bmatrix} 0 & \Delta x_1 & \Delta y_1 & \Delta\ln(r_{\Delta 1}^2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \Delta x_{n_\Delta} & \Delta y_{n_\Delta} & \Delta\ln(r_{\Delta n}^2) \end{bmatrix} \tag{10}$$

where n_Δ is the number of double points, Δx and Δy are the coordinate differences of the extremities of a double point and $\Delta\ln(r_\Delta^2)$ is the difference in the well constraints of the extremities of a double point . Using the dual (co)kriging formalism, an estimation of the head field can be expressed as:

$$\hat{h} = \begin{bmatrix} h_{obs} & \Delta & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{hh} & \Sigma_{h\Delta} & F_h \\ \Sigma_{h\Delta}^{'} & \Sigma_{\Delta\Delta} & F_\Delta \\ F_h^{'} & F_\Delta^{'} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{h0} \\ \sigma_{\Delta 0} \\ f_0 \end{bmatrix} \tag{11}$$

where the subscripts h and Δ refer to observed heads and double points respectively; h_{obs} is the $1 \times n_{obs}$ vector of observed head values; Δ is the $1 \times n_\Delta$ vector of double point head differences (0 for an impervious boundary); Σ_{hh} is the $n_{obs} \times n_{obs}$ head covariance matrix; $\Sigma_{h\Delta}$ is the $n_{obs} \times n_\Delta$ head-double point covariance matrix; $\Sigma_{\Delta\Delta}$ is the $n_\Delta \times n_\Delta$ double point covariance matrix; σ_{ho} is the $n_{obs} \times 1$ vector of head covariances between observation points and estimation points; $\sigma_{\Delta o}$ is the $n_\Delta \times 1$ vector of head covariances between double points and estimation points and f_o is the 4×1 drift vector evaluated at the estimation point.

For common hydrogeological application, only a few piezometers are installed in the field. It is thus difficult to infer an experimental variographic model with so few data. Brochu and Marcotte (2003) justify the use of a gravimetric covariance function (Chilès and Guilhen 1984, Marcotte and Chouteau 1993) to represent the behaviour of the hydraulic head. To construct the covariance matrix $\Sigma_{\Delta\Delta}$, the gravimetric covariance function is used with *ad hoc* range, sill and nugget values that result in a realistic head map. The range is chosen large enough to produce a continuous head field. The nugget is specified based on existing knowledge and the sill is approximated based on the variance of the residuals.

3 Synthetic study

The algorithm depicted in Fig. 1 is tested on a synthetic model (Fig. 2). The north and south boundaries are impervious while the east and west boundaries have prescribed head values of 125 m and 110 m, respectively. A pumping well, located at (500 m, 1000 m), has a radius of 0.5 m and a pumping rate of 0.0125 m^3/s.

The field extents over 2000 m by 2000 m and includes 3088 quadratic finite elements. A heterogeneous isotropic $\log_{10} T$ field (Fig. 2b), noted Y^R, was generated on the mesh and constitutes the reference field. The direct problem is solved under steady-state conditions with the software Femlab© 3.0a (Comsol inc. 2004).

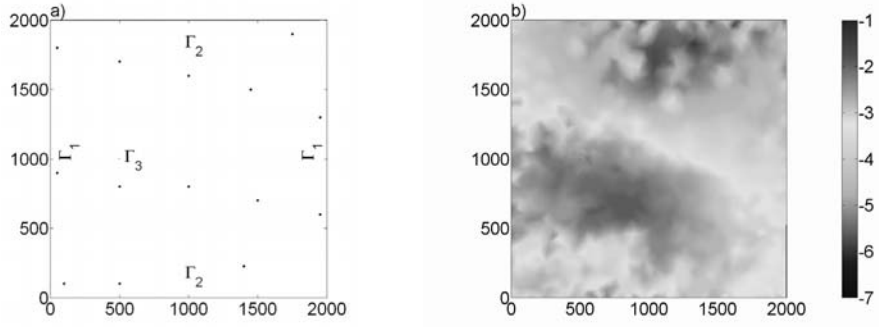


Fig. 2. Synthetic model. **a)** Boundary conditions and T data locations. **b)** Reference $\log_{10}$ transmissivity field.

3.1 Evaluated Scenarios

To evaluate the influence of the head field estimation error in the proposed method, Eq. 7 is applied on one hundred seed transmissivity fields (T^o) using four different $\hat{h}$ fields (Fig. 3). In scenario A, $\hat{h}$ is the reference head field (Fig. 3a) while for scenarios B, C and D (Fig. 3b - d), $\hat{h}$ is obtained by solving Eq. 11 with various numbers (50, 25 and 10 data) of head observations, h_{obs}. To adequately represent the impervious boundaries, 20 double points, separated by 50 m, were

placed on each (Γ_2). A gravimetric covariance function with *ad hoc* nugget of $0.001m^2$, sill equal to the variance of the centered data and range of 400m were used to construct $\Sigma_{\Delta\Delta}$. No measurement error is considered and the constant head boundaries (Γ_1) are assumed known.

The seed transmissivity fields are generated by geostatistical conditional simulation using Cholesky decomposition of the covariance matrix (Davis 1987). Since in practice the lack of data, the scale effect and the large uncertainty on the measurements hinder the inference of the $\log_{10} T$ covariance model, the seed fields were obtained with fourteen T data (Fig. 2a) and a covariance model chosen to be different from the covariogram of Y^R. The iteration parameters used are $a=15$ iterations and β^o =0.025. During the calibration process, the fourteen T data were assumed unknown.

3.2 Evaluation Measures

Each calibrated field was compared to the reference T and h fields to evaluate the efficiency of the method. To evaluate the fit between the reference fields and the calibrated fields (Y and h), the following performance measures are defined:

$$MAE(x^i, x^R) = \sum_{j=1}^{n_{ele}} \left| x_j^i - x_j^R \right| \frac{A_j}{A_\Omega} \tag{13}$$

where MAE is the mean absolute error between parameter x evaluated at iteration i and the reference parameter x^R, A_j is the area of the j^{th} element mesh and A_Ω is the area of the whole domain. At the observation points, the MAE is given by:

$$MAE(h_{obs}^i, h_{obs}^R) = \sum_{l=1}^{n_{obs}} \frac{\left| h_{obs}^i - h_{obs}^R \right|}{n_{obs}} \tag{14}$$

where the subscript *obs* refers to an observation and where n_{obs} is the total number of head observations, h_{obs}. The ensemble average $EMAE(x^i,x^R)$ over the n_{sim} inversions is also defined:

$$EMAE(x^i, x^R) = \sum_{m=1}^{n_{sim}} \frac{MAE(x^i, x^R)}{n_{sim}} \tag{15}$$

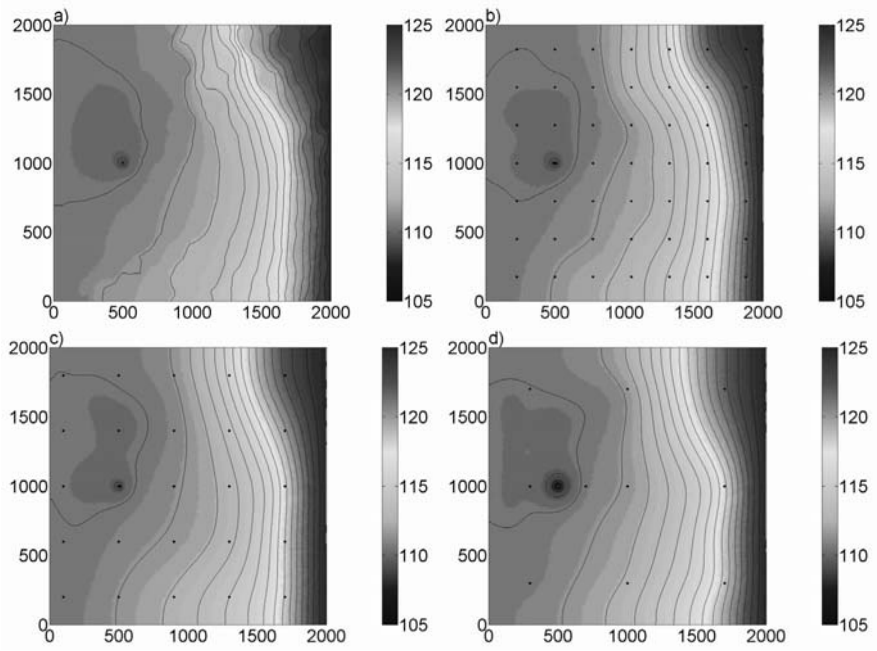


Fig. 3. Head fields used in scenarios A to D and h_{obs} locations. **a)** Reference head field, **b)** head field estimated from 50 observations, **c)** head field estimated from 25 observations, **d)** head field estimated from 10 observations.

4 Results

4.1 Head and Transmissivity calibration

Fig. 4 shows a single realization obtained for scenario D. Even if the seed field (Fig. 4a) is very different from the reference field, the main features of the real field (Fig. 2b) are retrieved in the calibrated transmissivities (Fig. 4c). Also, the head field used to perform the inversion and the reference head field are reproduced quite well (compare Fig. 4d to 3d and 3a). For this realization, the $MAE(h_{obs}^{i}, h_{obs}^{R})$ and the $MAE(Y^{i}, Y^{R})$ after 30 iterations are respectively of 4.98 cm and 0.55.

Mean initial and calibrated errors after 30 iterations are shown in Table 1 for scenarios A to D. The cumulative frequency of the individual realizations is presented in Fig. 5. For all scenarios, an improvement in transmissivity estimation is observed. Moreover, the mean head error at data locations and the mean discrepancy between $\hat{h}$ and the final head field is always less than 5 cm. The fit of the head field is improved by a factor of 10 to 100 when comparing initial and

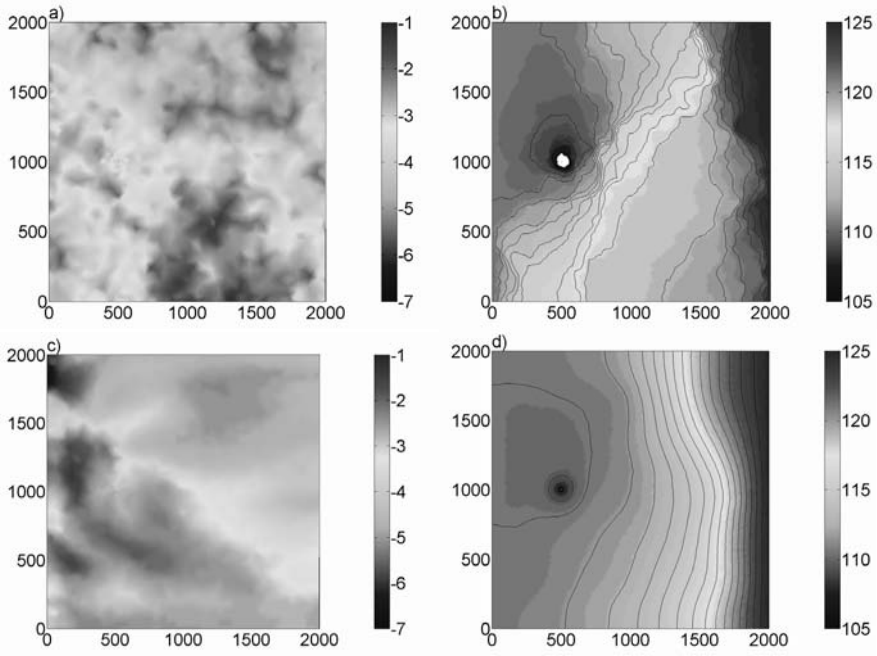


Fig. 4. Single realization for scenario D. **a)** Seed $\log_{10} T$ field and **b)** corresponding head field, **c)** calibrated $\log_{10} T$ field and **d)** corresponding head field.

calibrated heads. The best *EMAE* statistics are for scenario A which uses the real head field to perform the inversion. It would appear that with more head data available, the transmissivities are more constrained and the final fields estimate well the reference fields. However, the results for scenarios B to D show that even when an estimated head field is used to perform the inversion, the method is able to reconstruct the transmissivity field adequately.

The method tends to better reproduce the field used to perform the inversion ($\hat{h}$) than the reference field $EMAE(h^i, h^R)$. However, even in scenario D (10 head data), the reference head field is well reproduced. This indicates that few head data combined with some knowledge of the boundary conditions is sufficient to obtain a realistic transmissivity field.

Table 1. Ensemble average mean absolute error of $\log_{10} T$, head field and observed head for seed and calibrated fields

	$EMAE(Y^i, Y^R)$ ($\log_{10} T$)		$EMAE(h^i_{obs}, h^R_{obs})$ (m)		$EMAE(h^i, \hat{h})$ (m)		$EMAE(h^i, h^R)$ (m)	
	i=0	i=30	i=0	i=30	i=0	i=30	i=0	i=30
Scenario A	0,89	0,41	2,73	0,028	2,73	0,028	2,73	0,028
Scenario B	0,89	0,55	3,12	0,036	2,74	0,034	2,73	0,14
Scenario C	0,89	0,58	3,00	0,043	2,73	0,041	2,73	0,16
Scenario D	0,89	0,56	2,86	0,039	2,64	0,040	2,73	0,29

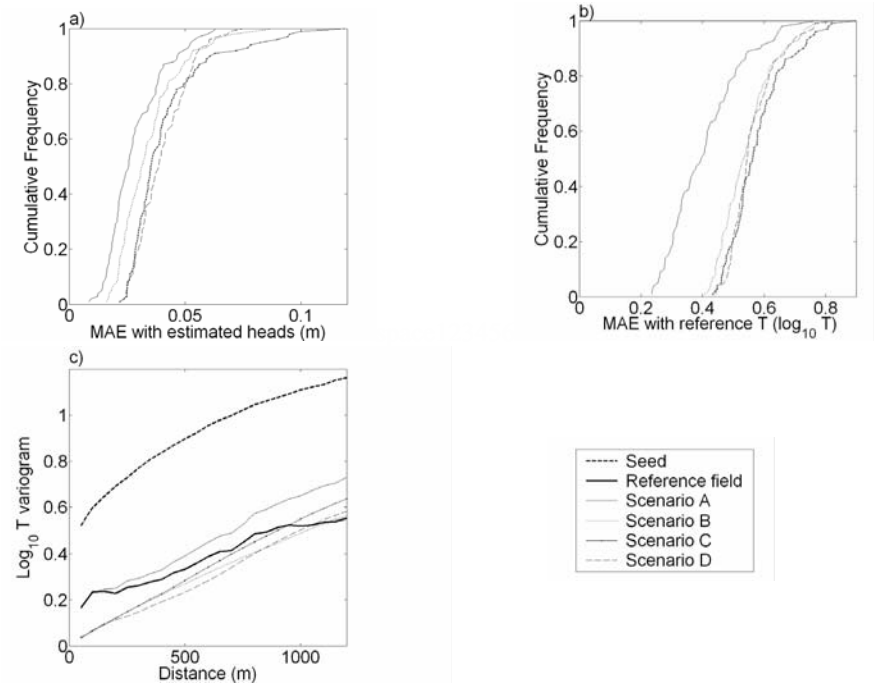


Fig. 5. Cumulative frequency of **a)** $MAE(h^i, \hat{h})$, **b)** $MAE(Y^i, Y^R)$. **c)** Mean variogram of $\log_{10}$ T (evaluated on a regular grid) for the seed fields, the reference field and the calibrated fields of scenarios A to D.

Even if the variographic structure was not imposed during the calibration process, an improvement in the experimental $\log_{10} T$ variograms is observed for each scenario (Fig. 5c). The variogram obtained for scenario A is similar to the reference variogram. Due to smoothing of the estimated head fields in scenarios B to D, the experimental variograms are smoother and do not show a nugget effect.

4.2 Execution Time

A strength of the method is its fast execution time. The inversion shown in Fig. 4 was completed in less than 17 seconds on a 2.66 Ghz Pentium© IV computer with 512 Mb of RAM. Most of the time was used to solve the direct problem involving 3088 quadratic finite elements. Since no time-consuming function evaluations (ex. matrix inversion) are performed when solving Eq. 7, the transmissivity is updated quickly. Table 2 shows the time partition for this single realization.

In the proposed method, the direct problem is solved only once per iteration. Thus, the overall cost of the inversion is very small. By comparison, methods based on non-linear optimization of an objective function require the costly com-

putation of the objective function gradient followed by a line search which typically requires many more solutions of the direct problem.

Table 2. Time partition for the inversion shown in Fig. 4 after 30 iterations

Operation	Time	Ratio
Direct problem solution	13,2 sec	78,6%
Correcting T (Eq. 7)	3,6 sec	21,4%
Total calculation time	16,8 sec	

5 Conclusions

A simple and computationally efficient method to solve the ground water inverse problem is proposed. The method consists of iterative estimation of the flux field and a discrete solution of the Cauchy problem. The method assumes that a realistic head map can be obtained from the available data. It is demonstrated that with as few as 10 data to perform the inversion, the method is able to generate a realistic transmissivity field. Also, even while transmissivities at data locations and variographic structure were not imposed during the calibration process, the final transmissivity variograms reproduced quite well the reference variogram.

The fast execution time of the method allows significant reduction in head data misfit in a few seconds. This enables quick uncertainty assessment in stochastic inverse modelling.

The method relies on the use of an estimated head map which is sometimes difficult to obtain. However, in most practical applications, it is common practice to draw preliminary piezometric maps in order to obtain conceptual models of the groundwater system. The synthetic study presented shows that such a preliminary map is sufficient to retrieve quite well the main characteristics of the transmissivity field.

The algorithm could be modified to fix transmissivities where it is assumed known or confined within a given interval of possible values. Generalisation to the transient case does not pose any particular problem and extension to 3D flow is possible. If the real variogram of Y is known, geostatistical conditional simulation can be introduced in the algorithm, but the problem of passing from measurement scale to the finite element scale still needs be addressed. In the present study, the problem of boundary condition determination and head measurement error was not evaluated.

Acknowledgements

This research was financed by an NSERC research grant and an NSERC student scholarship. Thanks to Maria Annecchione for editing the manuscript.

References

Bear J, Jacob M (1965) On the movement of water bodies injected into aquifers. Journal of Hydrogeology, vol. 3, 37-57

Brochu Y (2002) Estimation directe des charges hydrauliques d'un aquifère par krigeage, Master's thesis (in french), École Polytechnique de Montréal, p. 118

Brochu Y, Marcotte D (2003) A simple approach to account for radial flow and boundary conditions when kriging hydraulic head fields for confined aquifers, Math. Geol., vol. 35, no.2, 111-136

Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, Water Resources Research, vol. 22, no. 2, 199-210

Chilès JP, Delfiner P (1999) Geostatistics: Modeling Spatial Uncertainty. Wiley, New York

Chilès JP, Guilhen A (1984) Variogrammes et krigeages pour la gravimétrie et le magnétisme. Série Informatique Géologique, vol. 1, 455–46

Comsol AB. (2004) Femlab 3.0 User and reference manual, Stockholm, Sweden

Davis MW (1987) Production of conditional simulations via the LU triangular decomposition of the covariance matrix, Math Geol., vol.176, no. 3, 149-265

Delhomme JP (1979) Kriging under boundary conditions, Presented at the American Geophysical Union fall meeting, San Francisco, December 1979

Emsellem Y, de Marsily G, (1971) An automatic solution for the inverse problem, Water Resources Research, vol. 7, no. 5, 1264-1283

Gòmez-Hernàndez JJ, Sahuquillo A, Capilla JE (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data - I. Theory, Journal of Hydrology, vol. 203, no. 1-4, 162-174

Guo X, Zhang C-M (2000) Hydraulic gradient comparison method to estimate aquifer hydraulic parameters under steady-state conditions. Ground Water, vol. 38, no. 6, 815-826

Marcotte D, Chouteau M (1993) Gravity data transformation by kriging. In Geostatistics Tròia 1992, Soares A (ed.), 249–269, vol. 1, Kluwer Academic, Dordrecht, The Netherlands

de Marsily G, Lavedan G, Boucher M, Fasanino G (1984) Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, in Geostatistics for Natural Resources Characterization, Verly G, David M, Journel AG, Marechal A (eds.), 831-849, Reide D, Norwell, Mass.

Ponzini G, Lozej A (1982) Identification of aquifer transmissivities : The comparison model method, Water Ressources Research. vol. 18, no. 3, 597-622

Sagar B, Yakowitz S, Duckstein L (1975) A direct method for the identification of the parameters of dynamic nonhomogeneous aquifers, Water Resources Research, vol. 11, no. 4, 563-570

Tonkin MJ, Larson SP (2002) Kriging water levels with a regional-linear and point-logarithmic drift, Ground Water, vol. 33, no. 1, 185-193

Inverse problem for highly heterogeneous porous media: the factorial geostatistical analysis in differential system method

B. Ortuani

Institute of Agricultural Hydraulics, University of Milan – via Celoria, 2 – 20133 Milan, Italy, e-mail: bianca.ortuani@unimi.it

1 Introduction

The accuracy and the reliability of mathematical models for groundwater systems depend on the comprehensive knowledge about the physical properties and the dynamics of the systems. As the knowledge on transmissivity field *cannot* be complete, a statistical approach is used to describe the uncertainty in transmissivity estimates. The presence in porous media of different materials, whose statistical properties are not uniform, is a critical issue. Much of the existing literature on stochastic hydrogeology, and geostatistical approach in inverse problem (Kitanidis and Vomvoris 1983, Hoeksema and Kitanidis 1984, Dagan 1985) accounts for the assumption of statistically uniform porous media with relatively small variance of log-transmissivity, which limits the applicability of the stochastic approach in aquifer modelling to mildly heterogeneous porous media. The more recent literature considers the challenge in estimating transmissivity fields with large and complex variability (high and low-transmissivity areas, discontinuities...), as well as not stationary multi-gaussian log-transmissivity fields (Zimmerman *et al.* 1998, Hendricks Franssen *et al.* 1999, Guadagnini *et al.* 2002). The study provides a possible way in discrete inverse problem, to deal with the identification of both the geometry and the spatial variability of transmissivity field for different materials in highly heterogeneous porous media, by considering a geostatistical approach within the Differential System Method (Parravicini *et al.* 1995, Giudici *et al.* 1995). DSM calculates the transmissivity values along an integration path beginning at any point with known transmissivity value. If source terms are negligible, the variability in transmissivity field is completely described by a spatially distributed parameter, a^L, depending on hydraulic heads, integration path and cell size L of the numerical grid. The non-uniqueness of discrete solution, that is the uncertainty in transmissivity estimates, depends on the discrete approximations on a^L values, the errors in hydraulic heads, as well as the uncertainty in determining the integration path. The geostatistical approach within DSM takes into account the structural analysis of a^L, which is a variable including all the factors of uncertainty in transmissivity estimates. All the possible a^L values at any location correspond to all the possible transmissivity estimates in that point, depending on directions of

integration. Assumptions of stationarity, and multi-gaussian distribution of a^L variable are requested, being not restrictive on transmissivity distribution. The geostatistical analysis of a^L takes into account i) head data from multiple independent flow conditions, ii) transmissivity variability in different directions at any location, by running the multivariate analysis of a^L distributions for different integration paths; moreover, it provides iii) factorization of error in a^L values, which can be filtered out to improve transmissivity estimates, as well as geometry identification. The analysis of error makes effective iv) the identification of geometry, and v) the estimate of equivalent transmissivity values for each statistically homogeneous area. Finally, the geostatistical analysis of a^L is meaningful in aquifer modeling, as the variability in transmissivity field is considered at the cell size L (Lunati *et al.* 2001). The geostatistical approach in DSM was applied for a synthetic confined aquifer, with nine homogeneous areas, whose transmissivity values are constant and vary of two orders of magnitude within the flow domain, with differences between adjacent areas which are up to two orders of magnitude. In this numerical application, the variability of transmissivity within the statistically homogeneous areas was not considered.

2 The geostatistical approach in DSM

2.1 The DS method

Let us consider a confined aquifer with zero source terms. The following differential system for the unknowns $\partial_x T$ and $\partial_y T$ can be written, starting from the flow equations for two steady head fields (Parravicini *et al.* 1995):

$$\mathbf{A}(\mathbf{x})\nabla T(\mathbf{x}) = -T(\mathbf{x})\Delta\mathbf{h}(\mathbf{x}) \tag{1}$$

T is the transmissivity field, $\mathbf{A}$ is the matrix of the hydraulic gradients, and $\Delta\mathbf{h}$ is the vector of the Laplace operator of the two head fields. Solution for linear system (1) exists if the hydraulic gradients are non-zero and different for the two flow conditions (i.e., independence condition between head data), and it is:

$$\nabla T(\mathbf{x}) = -T(\mathbf{x})\mathbf{a}(\mathbf{x}) \tag{2}$$

where the vector parameter $\mathbf{a}$ is defined as $\mathbf{A}^{-1}(\mathbf{x})\Delta\mathbf{h}(\mathbf{x})$. Starting from Eq. 2, the solution T for *each* point $\mathbf{x}$ in the space is calculated along *any* integration path between points $\mathbf{x_0}$ (where the transmissivity value has to be known) and $\mathbf{x}$. If the integration path consists of a few straight segments, the solution comes from:

$$T(\mathbf{x}) = T(\mathbf{x}_0)\exp\left(-\sum_i \int_0^1 a_{i,i+1}(s)ds\right) \tag{3}$$

where $a_{i,i+1}$ is the projection of the $\mathbf{a}$ vector on direction of the segment between two consecutive points, $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$; $a_{i,i+1}$ value varies within the segment, as s varies from 0 to 1. Finally, the discrete solution at any node (i, j) of a numerical grid

with cell size L is (Giudici *et al.* 1995):

$$T_{i,j} = T_0 \exp\left(- \sum_{n=1}^{N} \left(a^L \right)_n \right) \tag{4}$$

T_0 is the known transmissivity value at the starting node of an integration path, (i_0, j_0), the parameter a^L is derived from $a(s)$ values within the internode segment n, and the sum is extended to the N internode segments which constitute the integration path, between nodes (i_0, j_0) and (i, j). The discrete solution and its stability depend on the integration path (Giudici *et al.* 1995).

2.2 The structural components of a^L

Let us consider the function of space $A(\mathbf{x}) \equiv - a^L(\mathbf{x})$, which is discontinuous and defined just in the grid nodes. It depends on the cell size L, as well as on the integration path. Moreover, the error components in a^L, which derive from the errors on **a** values estimated in a discrete domain, depend on the integration path. A is a random function, whose spatial structure is related to the spatial distributions of transmissivity and error as well. The parameter a^L *completely* describes the spatial variability of transmissivity field (which is considered at the cell size L), as it can be derived from Eq. 4:

$$a^L = -\left(\pm \ln T_{i,j} / T_{i',j'} \right) \tag{5}$$

$T_{i,j}$ and $T_{i',j'}$ are the transmissivity values estimated in two consecutive nodes along the integration path, and the sign depends on the verso of integration along the internode. The geostatistical approach in DSM accounts for the spatial correlation structure of the random function A, which is assumed to be multigaussian and intrinsic; the variogram has a nested structure:

$$\gamma^L(\mathbf{h}) = \gamma_N^L(\mathbf{h}) + \gamma_{Sc}^L(\mathbf{h}) \tag{6}$$

$$A(\mathbf{x}) = N^L(\mathbf{x}) + Sc^L(\mathbf{x}) \tag{7}$$

where **h** is the vector of distance. γ_N^L and γ_{Sc}^L represent the variability, respectively, at the local, and the small and large scales; the structural components, N^L and Sc^L, are the nugget and the spatially correlated components, respectively. If a few correlated functions A^i (two at least) are considered, each of them related to a different integration path, one of the possible decompositions (Eq. 7) in uncorrelated structural components is provided by factorial kriging analysis (Wackernagel 1998).

The nugget component, N^L, considers the variability of A *at each node*, depending on discontinuities in the field of transmissivity (i.e. the Uncorrelated Gradient Component of A), which stem from the presence in porous medium of different materials. Moreover, N^L depends on the uncorrelated errors which affect the a^L values *at each node* (i.e. the Uncorrelated Noisy Component of A). The component Sc^L considers the spatial correlation in variability of A, which depends on the

geometry of the different materials (i.e. the Correlated Geometric Component of *A*), and the spatially correlated errors on a^L values (i.e. the Correlated Noisy Component of *A*). CNC is due to the spatial distributions of the directions of integration (i.e., the integration path). The occurrence of periodic behaviour in variogram exhibits the repetition in space of *some structures*, which are related to the extensions of the different materials. These *structures* are lines along which a^L values are maximum, or larger than the neighbouring ones; they correspond to the discontinuity lines between different materials, which are crossed by the integration path. The characteristic lengths of periodicity in spatial structure of Sc^L describe the geometry of different materials. Moreover, the same sign in Sc^L values along discontinuity lines crossed by integration path - which stem from correlated directions of integration across those lines - gives information on transmissivity gradients between different materials.

3 The equivalent parameter

Equivalent transmissivity values characterise the different materials, which are statistically homogeneous. The transmissivity field T^H, within each statistically homogeneous area *H* of a porous medium with transmissivity *T*, is described by:

$$T^H(\mathbf{x}) = \overline{T}_H + \delta_H(\mathbf{x}) \tag{8}$$

$T^H(\mathbf{x})$ is a random function, $\overline{T}_H$ is its mean value (i.e., the equivalent parameter), and the spatially correlated residuals $\delta_H(\mathbf{x})$, with zero mean value, satisfy the condition $\sigma_{\delta_H}{}^2 < \sigma_T{}^2$. Factorial analysis of the structural components of *A* is used to make reliable the $\overline{T}_H$ estimate by identifying the error on *A*. Let us consider the following decomposition of *A* into independent factors:

$$A(\mathbf{x}) = Grad^L(\mathbf{x}) + Inner^L(\mathbf{x}) + \varepsilon(\mathbf{x}) \tag{9}$$

the random function $Grad^L$ is non zero only at the extremes of internodes crossing discontinuity lines between homogeneous areas, the random function $Inner^L$ is non zero only at the nodes within each area; ε is the error on a^L. $Grad^L$ and $Inner^L$ depend on those components of N^L and Sc^L, which are not error components:

$$Grad^L(\mathbf{x}) = c_N^G UGC(\mathbf{x}) + c_{Sc}^G CGC(\mathbf{x}) \tag{10}$$

$$Inner^L(\mathbf{x}) = c_{Sc}^I CGC(\mathbf{x}) \tag{11}$$

being the coefficients c_N^G, c_{Sc}^G, and c_{Sc}^I the fractions of variance of $Grad^L$ and $Inner^L$, respectively, that are due to the components *UGC* and *CGC*.

The functions $Grad^L$ and $Inner^L$ can be expressed as linear combinations of some random functions, mutually spatially uncorrelated, each of them related, respectively, to the discontinuity line between two specific homogeneous areas, and to the inner part of a specific homogeneous area, being zero-value elsewhere:

$$Grad^{L}(\mathbf{x}) = \sum_{k=1}^{K} \left(Grad^{L}_{H_i H_j}(\mathbf{x}) \right)_k \tag{12}$$

$$Inner^{L}(\mathbf{x}) = \sum_{j=1}^{J} \left(Inner^{L}_{H_j}(\mathbf{x}) \right)_j \tag{13}$$

K is the number of discontinuity lines for adjacent homogeneous areas, and J is the number of homogeneous areas. The mean value of $Grad^{L}_{H_i H_j}$, $\bar{a}_{H_i H_j}$, depends *only* on the discrete transmissivity gradients, across the discontinuity line between H_i and H_j areas, while the mean value of $Inner^{L}_{H_j}$, $\bar{a}_{H_j}$, depends *only* on the spatial variability of transmissivity within the H_j area. If the inner variability of transmissivity is much less than the variability between the different areas, the variances of $Inner^{L}_{H_j}$ and $Grad^{L}_{H_i H_j}$ are negligible, with respect to the spatial variability of their mean values, then the estimate of equivalent transmissivity values can take into account just the mean values $\bar{a}_{H_j}$ and $\bar{a}_{H_i H_j}$, as illustrated hereinafter. The transmissivity value $T_{i,j}$ at any node (i, j) within the H area is:

$$\begin{aligned} &T_0 \left[\exp\left(\sum_{n_l=1}^{N} \left(Inner^{L}_{H_l} \right)_{n_l} \right) \right] \left[\exp\left(\sum_{m_k=1}^{M} \left(Grad^{L}_{H_k H_h} \right)_{m_k} \right) \right] \left[\exp(\varepsilon') \right] = \\ &= T_0 \left[\exp\left(\sum_{n_l=1}^{N} \left(\bar{a}_{H_l} \right)_{n_l} \right) \right] \left[\exp\left(\sum_{m_k=1}^{M} \left(\bar{a}_{H_k H_h} \right)_{m_k} \right) \right] \left[\exp(\delta) \right] \end{aligned} \tag{14}$$

according to Eq. 4, 12 and 13. N is the number of nodes within the homogeneous areas, and M is the number of internodes crossing the discontinuity lines between homogeneous areas, along the integration path between the starting node (i_0, j_0) (in the area H_0) and the node (i, j). The indexes n_l and m_k are referred, respectively, to the inner nodes for which just the function $Inner^{L}_{H_l}$ is non zero, and the nodes across the discontinuity line for which just the function $Grad^{L}_{H_k H_h}$ is non zero. The quantities ε' and δ depend, respectively, on the errors ε, and the residuals of $Inner^{L}_{H_l}$ and $Grad^{L}_{H_k H_h}$ together with the errors ε, which are cumulated along the integration path. Eq. 14 is equivalent to the following one:

$$T_{i,j} = T_0 \left[\exp\left(\sum_{l=1}^{N_l} \left(\bar{a}_{H_l} \right)_l \right) \right] \left[\exp\left(\sum_{k=1}^{M_k} \left(\bar{a}_{H_k H_h} \right)_k \right) \right] + \delta_H \equiv \left(\bar{T}_H + \varepsilon_H \right) + \delta_H \tag{15}$$

In Eq. 15, the error ε_H on the equivalent parameter $\bar{T}_H$ is due to errors on a^L values, and the sums are extended not to the internodes along the integration path between nodes (i_0, j_0) and (i, j), but to the N_l homogeneous areas and the M_k discontinuity lines which are crossed by the integration path, between areas H_0 and H. That is a *new* integration path, each *step* of which crosses a discontinuity line between adjacent areas (i.e. *upscaled* integration steps). The variability of transmissivity within area H is completely described by δ_H. Finally, the equivalent transmissivity value $\bar{T}_H$ depends on i) the known value T_0, ii) the *upscaled* integration path, iii) the mean values $\bar{a}_{H_l}$ and $\bar{a}_{H_k H_h}$. If the ergodicity hypothesys is assumed,

the estimate of $\bar{a}_{H_l}$ and $\bar{a}_{H_k H_h}$ values considers the spatial means:

$$\bar{a}_{H_l} \equiv E\left[Inner^{L}_{H_l}\right] = \frac{1}{n}\sum\nolimits_{n_l=1}^{n}\left(Inner^{L}_{H_l}\right)_{n_l} \tag{16}$$

$$\bar{a}_{H_k H_h} \equiv E\left[Grad^{L}_{H_k H_h}\right] = \frac{1}{m}\sum\nolimits_{m_k=1}^{m}\left(Grad^{L}_{H_k H_h}\right)_{m_k} \tag{17}$$

where n is the number of nodes within the area H_l, and m is the number of nodes across the discontinuity line between areas H_k and H_h. The determination of a *reference* integration path is required to calculate Eq. 16 and 17; that is an integration path including a statistically significant number of internodes for each of the discontinuity lines that are crossed by the upscaled integration path. Each upscaled integration step has the *mean* direction among those of the internodes across the respective discontinuity line, which belong to the reference path. If ε_I and ε_G are zero-mean errors, affecting $Inner^{L}_{H_l}$ and $GradT^{L}_{H_k H_h}$, respectively, the following equations hold:

$$\bar{a}_{H_l} = E\left[Inner^{L}_{H_l} + \varepsilon_I\right] \equiv \frac{1}{n}\sum\nolimits_{n_l=1}^{n}\left(A\right)_{n_l} \tag{18}$$

$$\bar{a}_{H_l H_h} = E\left[Grad^{L}_{H_l H_h} + \varepsilon_G\right] \equiv \frac{1}{m}\sum\nolimits_{m_k=1}^{m}\left(A\right)_{m_k} \tag{19}$$

where the n and m nodes are just located after having identified geometry, through the factorial analysis of the structural components N^L and Sc^L.

4 The numerical application

The geostatistical approach in DSM was applied to a synthetic confined aquifer with unit thickness (Ortuani 2002), and nine homogeneous areas, each of them with constant transmissivity value. In this case, Eq. 15 becomes:

$$T_{i,j} = \bar{T}_H + \varepsilon_H \tag{20}$$

The transmissivity values vary of two orders of magnitude throughout the flow domain, with differences between adjacent areas, which are up to two orders of magnitude (Fig. 1). The cell size L is equal to 0.05 m. Independent head fields have been calculated with different Dirichlet conditions on boundaries, by using a model with a finite difference numerical scheme. The independence condition between two head fields is shown in Fig. 2. It determines the errors on A values: the smaller the determinant of matrix **A** (i.e., det**A**), the worse the independence condition, and the larger the errors on *A could* be.

The spatial distributions of the random function A^1, related to a particular integration path, is shown in Fig. 3. The spatial distribution of the actual values (i.e. free error values) correspondent to A^1 values is represented in Fig. 4; they were calculated from the 'true' values of transmissivity (Eq. 5).

By comparing Fig. 2 with Fig. 3, it can be noticed that the large values of A^1 in some nodes within homogeneous areas, where A^1 *should* be zero and the determinant is almost zero, are due to errors.

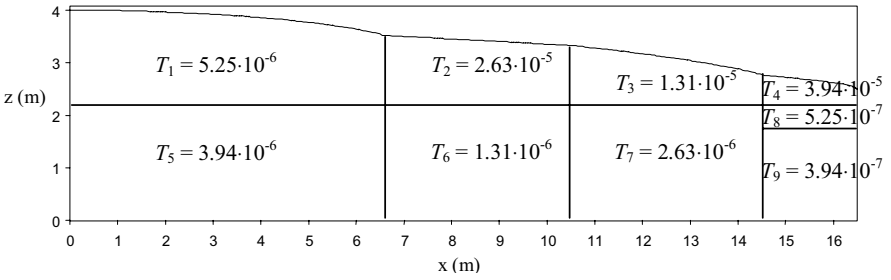


Fig. 1. The synthetic confined aquifer. The nine homogeneous areas, and the relative transmissivity values T_1-T_9 [m^2s^{-1}] are shown.

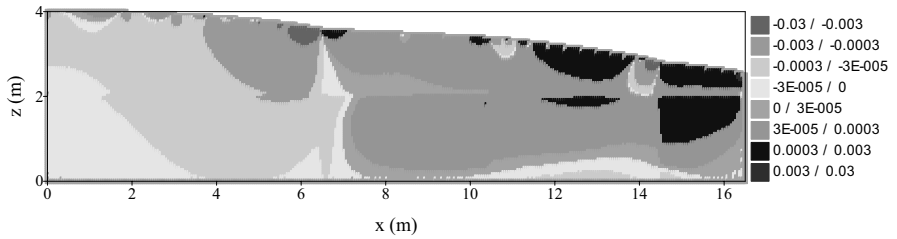


Fig. 2. The spatial distribution of det**A** values.

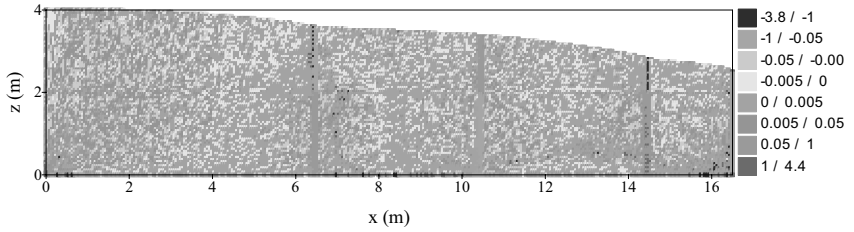


Fig. 3. The spatial distribution of the random functions A^1

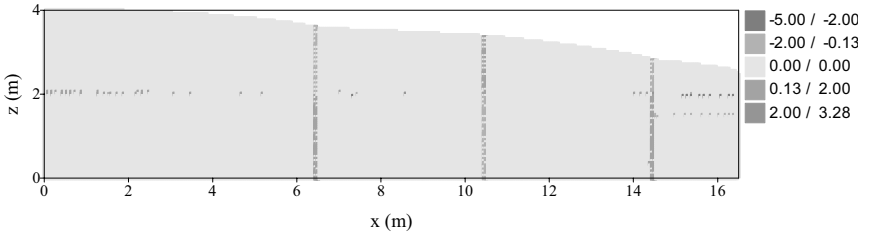


Fig. 4. The spatial distribution of the free error values correspondent to A^1

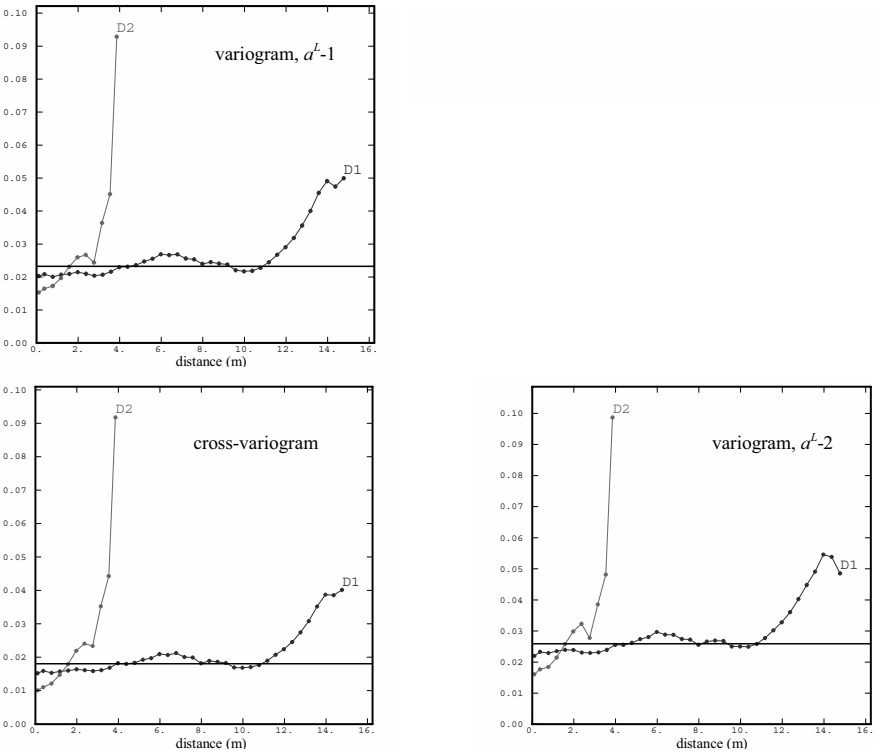


Fig. 5. Multivariate geostatistical analysis. Experimental variograms, variables a^L-1, a^L-2. D1 and D2 are the directions along the axes, respectively, x and z

4.1 The regionalised factorial analysis of a^L

The identification of geometry, as well as the estimation of equivalent transmissivity values depend on the factorisation of A in structural components. The more accurate the determination of *actual* components N^L and Sc^L, the more accurate the identification of geometry, and the estimates of equivalent trasmissivity values. The accuracy on N^L and Sc^L depends on the errors which affect the A values. A multivariate geostatistical analysis was carried out, in order to identify the error components of A and filter them out.

4.1.1 Evaluation of the structural components

Two random functions, A^1 and A^2, were considered. Their experimental variograms and the cross-variogram are represented in Fig. 5. The structural com-

ponents, N^L and Sc^L, were calculated by factorial kriging analysis, as linear combinations of the principal components of the respective co-regionalisation matrixes. Figs. 6 and 7 show the spatial distributions of N^L-1 and Sc^L-1, which are the structural components of A^1. The two principal components for each of the structural components are related to the common factors of variability which induce correlation between N^L-1 and N^L-2, and Sc^L-1 and Sc^L-2. They depend on the same direction and verso of integration, and the same direction and opposite verso of integration.

4.1.2 Identification of the error component

The errors in A^i functions are correlated, even if the directions of integration are different, because the head fields and, consequently, the independence condition are the same. The above factorisation of N^L and Sc^L was not useful in order to identify the error component of A^i; in fact, it considers the correlation due to the same directions of integration. Another factorisation of N^L and Sc^L was determined, by considering four different random functions A^1 (Ortuani 2002): A^1 and A^2, $(A^1)'$ and $(A^2)'$ which differ from the previous ones only for the sign as they don't consider the verso of integration. N^L and Sc^L were expressed as linear combinations of four principal components, each of them related to an independent factor of variability. The factors related to error were recognised from the correlation circles of the standard structural components (Ortuani 2002); the error components in N^L are correlated even if the versos and the directions of integration are different at the same node, while the error components in Sc^L are not positively correlated even if the spatially correlated versos and directions of integration are the same. The principal components associated to error were filtered from N^L and Sc^L. A reduction of the error components was obtained, mainly where the independence condition between head data was the worst and the errors were the largest (Fig. 8). Table 1 reports the statistical properties of the A^1 and A^2 samples, as well as the properties of the $(A^1)^{\text{filtered}}$ and $(A^2)^{\text{filtered}}$ samples, with reduced error components. The sample variances decrease (Tables 1 and 2), because the variability due to the error components was reduced. Besides, the correlation between variables decreases (Table 2). In fact, the correlation between N^L components decreases (it depends mainly on the occurrence of the same directions of integration, and weakly on the common sources of error), more than the correlation between Sc^L components increases (it depends just on the extension of homogeneous areas, and nomore on the spatially correlated errors).

4.2 Identification of geometry and equivalent parameters

The identification of geometry accounts for the lines along which the $(N^L\text{-}1)^{filtered}$ and $(Sc^L\text{-}1)^{filtered}$ values are larger than the neighbouring ones; these lines were identified as the boundaries of homogeneous areas which are crossed by integration path, assuming that the major variability was nomore due to error components, but just to discontinuities in transmissivity field as well as to the correlated

directions of integration along them. The nodes for calculation of the equivalent parameters (Eqs. 18 and 19) were selected through the analysis of the $(N^L\text{-}1)^{filtered}$ and $(Sc^L\text{-}1)^{filtered}$ distributions: they were localised where $(N^L\text{-}1)^{filtered}$ and $(Sc^L\text{-}1)^{filtered}$ values are minimum within homogeneous areas, and (relatively) maximum along the boundaries crossed by integration path; furthermore, $(N^L\text{-}1)^{filtered}$ and $(Sc^L\text{-}1)^{filtered}$ values at the selected nodes along the *crossed* boundaries have the same sign along the *mean* direction of integration.

Finally, the $(A^1)^{filtered}$ values were used to evaluate the equivalent parameters for each homogeneous area. The results are given in Table 3; the T_1 and T_5 values were assumed known, so that the *upscaled* integration path required eight steps. The comparison with the 'true' transmissivity values is rather good, as the order of magnitude was identified for each homogeneous areas, except for value T_8.

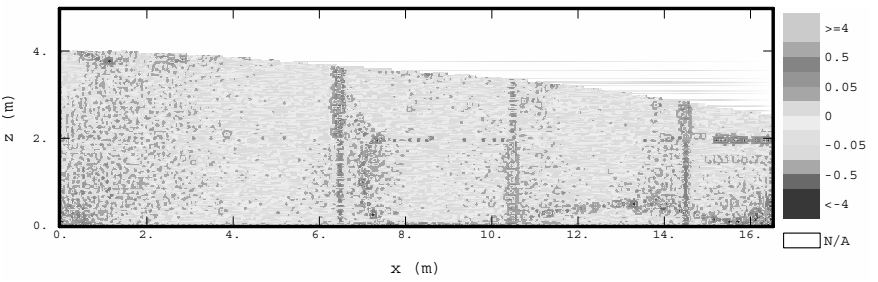


Fig. 6. The spatial distribution of N^L-1.

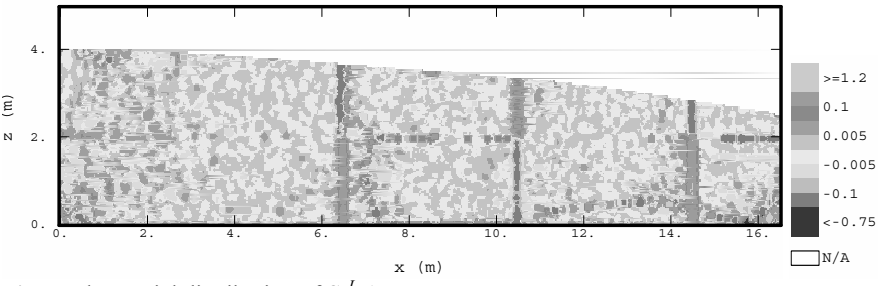


Fig. 7. The spatial distribution of Sc^L-1.

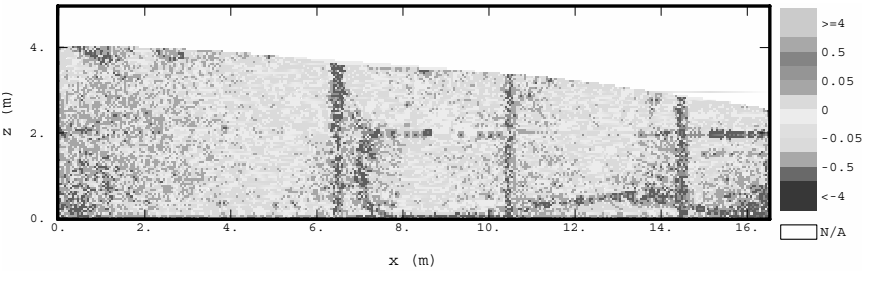


Fig. 8. Spatial distribution of the $|N^L\text{-}1|\text{-}|(N^L\text{-}1)^{filtered}|$ values, where $(N^L\text{-}1)^{filtered}$ is the nugget component with error reduction. In dark gray the nodes where the reduction is the largest.

Table 1. Statistical description of the $(A^1)^{filtered}$, $(A^2)^{filtered}$, A^1 and A^2 samples.

	Minimum	Maximum	Mean	Variance	Corr. Coeff.
$(A^1)^{filtered}$	-2.013	1.591	0.004	0.006	0.44
$(A^2)^{filtered}$	-1.347	2.200	0.007	0.011	
A^1	-4.296	4.423	0.004	0.023	0.61
A^2	-3.756	4.423	0.007	0.026	

Table 2. Statistical description of the structural components samples.

	Minimum	Maximum	Mean	Variance	Corr. Coeff.
$(N^L\text{-}1)^{filtered}$	-1.862	1.256	0.000	0.003	-0.08
$(N^L\text{-}2)^{filtered}$	-1.477	2.212	0.000	0.006	
N^L-1	-3.805	3.449	0.000	0.018	0.55
N^L-2	-3.507	3.449	0.000	0.018	
$(Sc^L\text{-}1)^{filtered}$	-0.687	0.936	0.004	0.003	0.97
$(Sc^L\text{-}2)^{filtered}$	-0.729	1.141	0.007	0.005	
Sc^L-1	-0.750	1.011	0.004	0.004	0.71
Sc^L-2	-0.707	1.101	0.007	0.006	

Table 3. The estimates of equivalent parameters.

	True value [m^2s^{-1}]	Equivalent parameter [m^2s^{-1}]
T_1	$5.25\ 10^{-6}$	$5.25\ 10^{-6}$
T_2	$2.63\ 10^{-5}$	$1.34\ 10^{-5}$
T_3	$1.31\ 10^{-5}$	$8.58\ 10^{-6}$
T_4	$3.94\ 10^{-5}$	$1.89\ 10^{-5}$
T_5	$3.94\ 10^{-6}$	$3.94\ 10^{-6}$
T_6	$1.31\ 10^{-6}$	$2.10\ 10^{-6}$
T_7	$2.63\ 10^{-6}$	$3.42\ 10^{-6}$
T_8	$5.25\ 10^{-7}$	$1.33\ 10^{-6}$
T_9	$3.94\ 10^{-7}$	$9.47\ 10^{-7}$

5 Conclusions and further developments

The results showed the effectiveness of the geostatistical approach within DSM in order to identify the geometry of different materials in highly heterogeneous porous media and to estimate the equivalent transmissivity values. The factorial analysis of each structural component of the random function A provided the identification of *some* error components which have been filtered out to make more reliable the determination of both the geometry and the equivalent transmissivity values. The identification of other factors of variability, through a more accurate analysis of correlation, could produce further reduction in error. Through the mul-

tivariate analysis of a few random functions A^i, each of them related to a different integration path, the discrete transmissivity gradients have been explored in different directions at the same node, increasing the information on transmissivity field, and making weaker the dependence of its characterisation on a particular integration path. Then, the description of heterogeneity in porous media is more consistent with the real condition, while error components can be identified by correlation analysis of random functions A^i. The identification of geometry and the estimation of equivalent parameters have been operated considering the analysis of stability, through: the cell size L, the error on A^i values as well as the independence condition between head data sets.

Further studies should consider the uncertainty in head fields, and the transmissivity variability at small scale, within the statistically homogeneous area.

References

Dagan G (1985) Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem. Water Resour. Res. Vol. 21, 1: 65-72

Giudici M, Morossi G, Parravicini G, Ponzini G (1995) A new method for the identification of distributed transmissivities. Water Resour. Res. 31: 1969-1988

Guadagnini L, Guadagnini A, Tartakovsky D (2002) A geostatistical model for distribution of facies in highly heterogeneous aquifers. In: GeoENV IV – Geostatistics for environmental applications, Kluwer Academic Publisher: 211-222

Kitanidis PK, Vomvoris EG (1983) A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. Water Reasour. Res. Vol. 19, 3: 677-690

Hendricks Franssen HJ, Gomez-Hernandez J.J., Capilla J.E., Sahuquillo A. (1999). Joint simulation of transmissivity and storativity fields conditional to steady-state and transient hydraulic head data. Adv. Water Resour. 23: 1-13

Hoeksema RJ, Kitanidis PK (1984) An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling. Water Resour. Res. Vol. 20, 7: 1003-1020

Lunati I, Bernard D, Giudici M, Parravicini G, Ponzini G (2001) A numerical comparison between two upscaling techniques: non-local inverse based scaling and simplified renormalization. Adv. Water Resour. Vol. 24, 8: 913-929

Ortuani B (2002) Processi di costruzione e validazione di modelli per la simulazione di sistemi acquiferi. PhD Thesis, University of Milan (Italy)

Parravicini G, Giudici M, Morossi G, Ponzini G (1995) Minimal a priori assignment in a direct method for determining phenomenological coefficients uniquely. Inverse Probl. 11: 611-629

Wackernagel H (1998) Multivariate Geostatistics, 2nd completely revised edition. Springer

Zimmerman DA, de Marsily G, Gotway CA, Marietta MG, Axness CL, Beauheim RL, Bras RL, Carrera J, Dagan G, Davies PB, Gallegos DP, Galli A, Gomez-Hernandez J, Grindrod P, Gutjahr AL, Kitanidis PK, Lavenue AM, McLaughlin D, Neuman SP, RamaRao BS, Rivenne C, Rubin Y (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resour. Res. Vol. 34, 6: 1373-1413

Inverse stochastic estimation of well capture zones with application to the Lauswiesen site (Tübingen, Germany)

H.-J. Hendricks Franssen and F. Stauffer

Institute of Hydromechanics and Water Resources Management, ETH Zurich, CH-8093 Zurich, Switzerland

1 Introduction

Capture zones of drinking water wells are estimated in order to know from which areas the contaminants could reach the well. However, these estimates are always uncertain due to the spatial variability of transmissivity, in the first place. Also uncertainty with respect to the mean recharge, the spatio-temporal variability of recharge and the boundary conditions are important. Depending on the situation there may be other important sources of uncertainty.

The uncertainty of the well capture zone estimate is quantified with stochastic methods. Here Monte Carlo methods will be used. Conditioning to transmissivity data reduces the uncertainty, and is also straightforward with Monte Carlo methods. Conditioning to hydraulic head data is more cumbersome, and is done here with the sequential self-calibrated method (Gómez-Hernández *et al.* 1997, Hendricks Franssen 2001). As a result, an ensemble of equally likely well capture zones is generated, each of them conditioned to transmissivity and hydraulic head data.

The methodology has been applied on the Lauswiesen site close to Tübingen, in Germany. This site is characterized by fluvial deposits that show large contrasts in hydraulic conductivity. A river, that is well connected to the aquifer is also present, and is also of concern in the inverse calibration.

2 Methodology

A stochastic well capture zone characterization yields for all grid cells at which the transport problem is solved the probability that that grid cell belongs to the well capture zone. This stochastic well capture zone is obtained in the following steps:

1. Equally probable realizations of the input parameters to the groundwater flow problem are built. In case of a stochastic characterization at least one parameter is modeled as a random variable. Frequently, only the uncertainty of the

logtransmissivity is considered, while other parameters are deterministic. Nevertheless, also recharge, storativity and boundary conditions can be random variables. The equally probable transmissivity realizations are generated by the sequential Gaussian simulation algorithm that is implemented in the software GCOSIM3D (Gómez-Hernández and Journel 1993). The realisations could also be generated by other methods.

2. For each of the realizations the groundwater flow equation is solved by finite differences, by INVERTO (Hendricks Franssen 2001). The computed heads are compared with the measured heads. The following formula is evaluated:

$$J = \sum_{i=1}^{N_h} \xi_i (h_i^{SIM} - h_i^{MEAS})^2 \tag{1}$$

where N_h is the number of head measurement locations, h_i the head at a measurement location, the weight ξ_i is chosen inverse proportional to the estimated measurement error, *SIM* refers to simulated and *MEAS* to measured. If J is smaller than a pre-defined tolerance value the measured heads are reproduced close enough. In case J is larger than the tolerance value the simulations continue with step 3. In case also concentration data would be present, the objective function would be extended with an additional term. In this paper we do not consider concentration data.

3. J was too large and therefore the gradient of the objective function with respect to the perturbation parameters (a parametrization of the random variables) is calculated. This gradient is minimized by a combination of non-linear optimization and geostatistics, and the parameters are updated. Then step 2 is repeated, and steps 2 and 3 are repeated until the objective function is reproduced close enough. As a final result, updated equally likely realizations of the random input parameters are obtained. These equally likely realizations are also conditioned now to hydraulic head data.
4. For each of the updated realisations the well catchment is determined. The particle tracking problem is solved by releasing a particle from the centre of each grid cell and tracking it until a boundary or the well has been reached.
5. The ensemble variances are calculated over the generated realisations. The following definitions hold:

$$AESD(Z) = \frac{1}{N} \sum_{i=1}^{N} \sigma_{Z_i} \tag{2}$$

where *AESD* is the average ensemble standard deviation, N the number of grid cells, i a grid cell index, Z stands for either log transmissivity or hydraulic head and σ is the ensemble standard deviation. The uncertainty with respect to the capture probability is given by:

$$AESD(CZ) = \frac{1}{N} \sum_{x=1}^{N} \min(CZ(x),\ 1 - CZ(x)) \tag{3}$$

where $AESD(CZ)$ is the domain averaged uncertainty with respect to the capture probability. For instance: if $CZ(x)=0$ or $CZ(x)=1$ the grid cell x does not contribute to $AESD(CZ)$; if $CZ(x)=0.5$ the contribution is 0.5 (the largest con-

tribution possible; the maximum uncertainty). With respect to the capture probability: $CZ(x,i)=0$ if a particle released from grid cell x for realisation i does not reach the pumping well and $CZ(x,i)=1$ if the particle reaches the pumping well. The average capture probability $CZ(x)$ for that grid cell is determined by averaging the obtained $CZ(x,i)$.

3 Case study

The Lauswiesen site is located in the Neckar Valley close to Tuebingen in Southwest Germany. The site consists mainly of fluvial deposits with a large hydraulic conductivity. The fluvial deposits result in a layered sediment with many very thin layers that may show a relatively strong variation in hydraulic conductivity in the vertical direction. On top of these fluvial deposits is a clay layer with a very low permeability. It is believed that no recharge passes through this layer. The aquifer is unconfined (the clay layer does not form a confining layer; the upper level of the aquifer is somewhat below the bottom of the clay layer) and has an estimated thickness of approximately 5 meters. The river Neckar passes through the area and is a boundary of the modelling area. The river bottom acts as a resistance for water flow between the river and the aquifer. In the central part of the Lauswiesen site there is a main pumping station of the Tuebingen water works. In the Northern part of the area there is a new pumping station installed, which pumps during the tracer tests. The tracers are injected southwest of the new pumping station, far enough from the Tuebingen water works, so that the tracer is not transported southwards.

The aim was to estimate the drinking water well catchment, to predict the travel times from the injection wells to the pumping well and compare these estimates with the measured values.

3.1 Experimental data

The following data are available for conditioning:

- 11 hydraulic conductivity data.
- 46 steady-state hydraulic head data (when the two main pumping wells were not working).
- the breakthrough curves from six tracer injections.
- porosity data. A few core samples have been analysed from which the porosity is estimated to be 8%.
- the water levels in the Neckar river.
- pumping rates. All the pumping rates at the wells are known.

The hydraulic conductivity data and the hydraulic head data have indeed been used for conditioning. The breakthrough curves served for verification. The po-

rosity and the pumping rates have been taken deterministically, although there is also some considerable uncertainty regarding the porosity value.

Besides these data some initial estimates were available concerning the boundary conditions:

- inflow rates. The amount of water inflow at some points of the border of the Lauswiesen site is available.
- prescribed heads. At some other boundaries of the Lauswiesen site prescribed head values are known.
- leakage factors. The leakage factors between the river bottom and the aquifer are given for the whole river length.

The inflow rates were not subject to calibration, but the prescribed heads on the boundaries and the leakage factors were, in a manner that will be explained in the section 3.3.

3.2 Model setup

The discretization of the domain has been adopted from Martac and Ptak (2002). The model has an irregular mesh with grid cells that vary in size between 20 cm and 25 m. See Figure 1. The grid is refined around pumping wells and the zones close to the river. The modelled area is about 3km x 1km. The total number of grid cells is 68,036. The aquifer has been modelled as confined, although it is unconfined. It is assumed that the decreases due to pumping are relatively limited as compared to the aquifer thickness. A constant saturated thickness of 5.08 m has been assumed.

The sequential self-calibrated method cannot handle the calibration of Cauchy type boundary conditions. Therefore, as an alternative, the prescribed head values on the river grid cells have been subjected to calibration. The modelling of the river as a prescribed head boundary can be justified by the fact that the aquifer shows an immediate response to modifications in the water level of the river.

3.3 Stochastic modelling

The well catchment and its uncertainty, and the results of the tracer tests, have been estimated for the unconditional simulations and the inversely conditioned simulations.

Unconditional simulations were generated by GCOSIM3D (Gómez-Hernández and Journel 1993). The adopted variogram model of logT is exponential, with a correlation length of 50 m and the sill is 0.10. The geometric mean of logtransmissivity is –2.70. The logtransmissivities had to be generated on an irregular grid. First the logtransmissivities were generated on a 5m x 5m grid. As a next step, the logtransmissivities on the irregular grid were obtained by taking the logtransmissivity value of the 5m x 5m grid that was the closest to the centre of a grid cell of the irregular grid. It means that if many small grid cells are contained

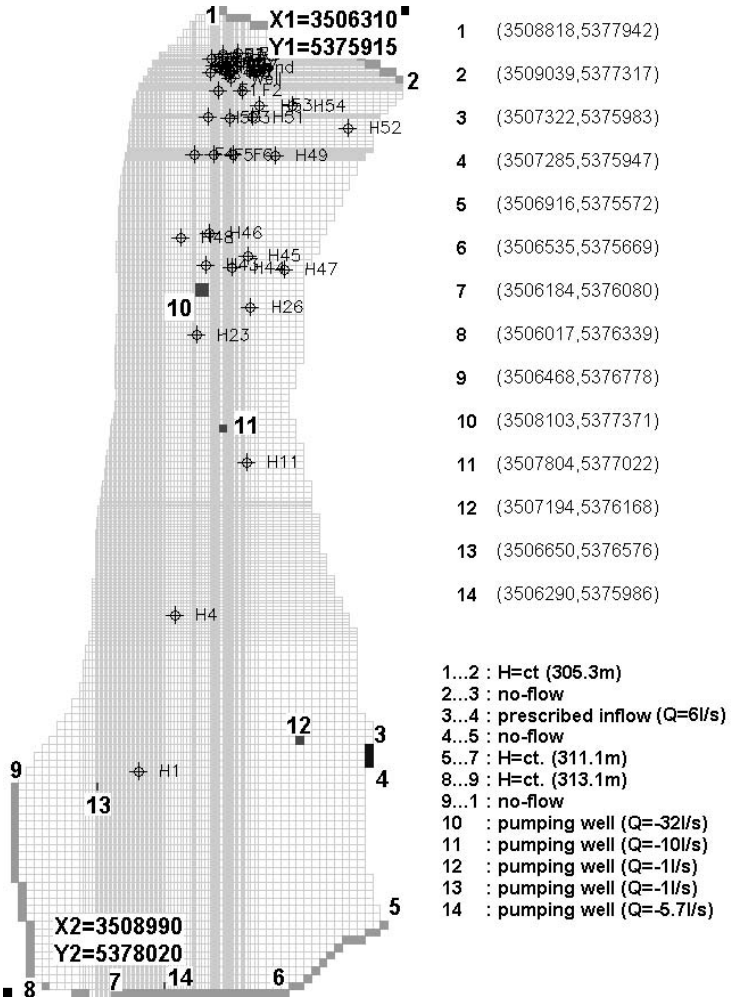


Fig. 1. Set-up of the model including information with regard to pumping wells (numbers 10-14), prescribed inflow (between 3 and 4), prescribed heads (between 5 and 7, between 8 and 9, between 1 and 2). Given are also the coordinates of the locations 1-14 and observation locations.

in a 5m x 5m grid cell, all these smaller grid cells become the same logtransmissivity value. If on the contrary many 5m x 5m grid cells are contained in a larger grid cell of the irregular grid, only one of these values is selected. This procedure is not ideal as the grid cells do not have all the same support, and the logtransmissivity statistics that are used to generate the random fields are support dependent (scale dependent), but it is thought to be an acceptable compromise of getting a good numerical solution, a not too large number of grid cells and a reasonably good representation of the spatial heterogeneity of logtransmissivity.

Although these simulations were unconditional, they were used in order to compare the simulated heads with the measured heads. The objective is to see whether the unconditional simulations reproduce on average the measured heads. It was found that there was a systematic bias in the head values. For nearly all the measurement data the simulated heads were above the measured heads, for (nearly) all the realisations. This suggests a bias, and it was found that this was due to the interaction between the river and the aquifer, and it pointed to the need to calibrate the prescribed heads along the river. Table 1 summarizes some results for the unconditional realisations.

The equally likely conditional realizations are conditioned to 11 transmissivity measurements, obtained from pumping tests, and the 46 head measurements under natural flow conditions. The conditioning to the hydraulic head data has been carried out in two steps. In the first step the prescribed heads along the river were calibrated. During this first calibration step the transmissivities are fixed, so that the calibration of the river heads only corrects the bias of the water balance. This is done in such a way that for all cells of the river the perturbation of the prescribed head value is the same, as such avoiding the development of lo-cal depressions. By calibrating the prescribed heads along the river the water balance of the model area is corrected. After this first step the hydraulic
head measurements are conditioned by modifying the transmissivity field, and fixing the boundary heads. Table 1 also summarizes some results for the conditional simulations. Results are again evaluated according the average ensemble standard deviation.

It is found that the ensemble averaged standard deviations for the transmissivity field and the hydraulic head field are considerably larger for the simulations conditioned to hydraulic head data, than for the unconditional realisations. This surprising observation may be related to the fact that the estimated prior log-transmissivity variance has been too low and the range too long. However, due to the limited number of calibrated realisations (20) the calculated statistics can differ considerably from the “true” statistics. In case of an additional conditioning to 11 transmissivity data, the ensemble variances decrease. The larger transmissivity variance propagates through the groundwater flow equation and the mass transport equation and also results in increased ensemble variances of head. The fact that the ensemble head variance after inverse modelling is larger than before inverse modelling is unique in the sense that with the Self Calibration Method this observation never could be made, not for synthetic studies and neither for practical case studies.

Table 1. Scores on the evaluation criteria for pre-pumping calibration with the Sequential Self Calibration Method. These values are calculated over the active cells only

	AESD(Y)	**AESD(h)**	**AESD(CZ)**
Unconditional	0.61	0.064	
0 T, 46 h data	1.43	0.141	
11 T, 46 h data	1.08	0.104	2.08×10^{-2}

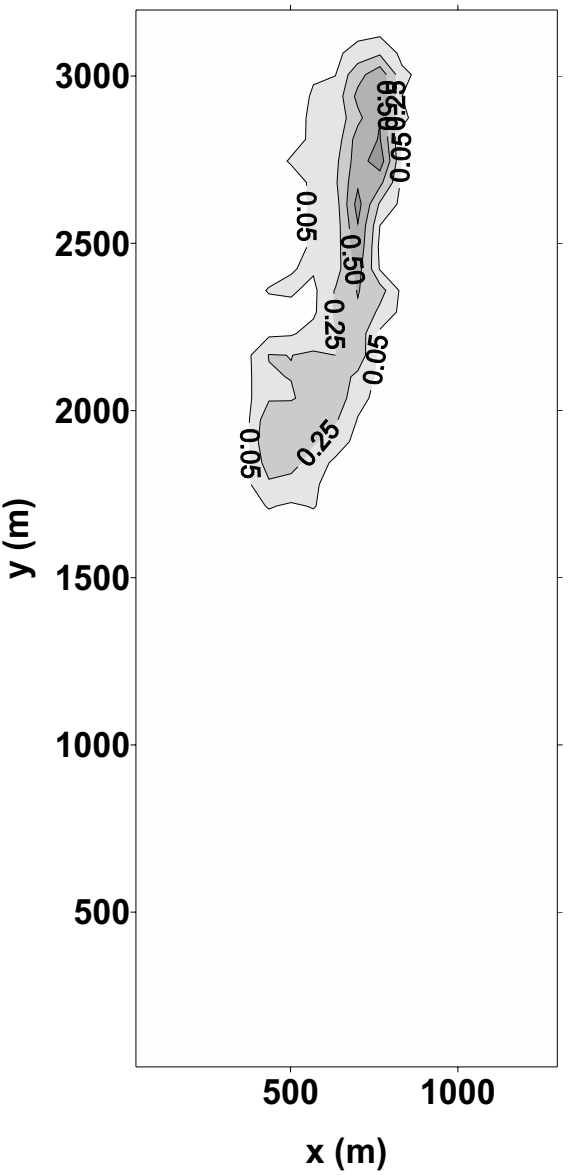


Fig. 2. Map of the catchment of abstraction well F0 for calibration with head measurements obtained from the pre-pumping stage. The mapped area corresponds to the area plotted in Figure 1. Only very locally the capture probability was 100%.

The conditioned transmissivity realisations have been input to forward groundwater flow simulations with the well pumping at F0. The calibrated initial heads along the river (different for each realisation) were left unmodified and a change in the water level of the Neckar has not been taken into account. The calculated hydraulic head fields have been used as input, together with the log-transmissivity fields, for calculating the well catchment of F0. For this purpose the software 3DTRANSP (Hendricks Franssen 2001) has been modified, so that it can handle the particle tracking for non-squared grid cells. Nevertheless, it was found that the implemented procedure is not very efficient for such a grid. Therefore, only the catchment for the conditioned realizations could be calculated, and only for the very limited amount of 20 realizations.

Also the mean and standard deviation of travel times to well F0 of conservative particles injected in wells F1-F6 were calculated and are presented in Table 2. It should be stressed that for non of these injection wells the capture probability was 100%, but always above 50%. In Figure 2 the capture probability of conservative particles is mapped.

The calculated mean travel times are in the same order of magnitude as the measured ones. The calculated catchment, the catchment uncertainty and the travel times are therefore not very surprising results. Main question remains why inverse conditioning results in such a dramatic variance increase in these stochastic inverse calculations.

Table 2. Predicted travel time in days (and standard deviation) from conservative tracers to abstraction well F0 for pre-pumping calibration.

	F1	**F2**	**F3**	**F4**	**F5**	**F6**
Mean	6.6	3.7	4.1	19.1	13.8	10.8
St dev	3.1	2.4	0.7	9.7	3.4	2.3

4 Discussion and conclusions

A main conclusion of the stochastic inverse simulations is that the methodologies are indeed capable of yielding conditional stochastic estimates of the well catchment, for a real world case. However, there exist several complications that point to the need of future research and developments in the nearby future.

A main issue in this particular case is the increase of the variance during conditioning, yielding posterior ensemble variances (after conditioning) that are larger than prior ensemble variances (before conditioning). In synthetic studies this behaviour also occasionally occurs, and it may be related to the fact that the "true" variogram differs considerably from the model variogram, or that measurements were taken by chance at locations that were very little representative (for example very local extreme values). Nevertheless, the increases are normally very limited and for a larger amount of measurement data the variance always will decrease again. In this case ensemble standard deviations were more than double as large after conditioning than prior to conditioning. Also the ensemble

head standard deviation was more than double as large after inverse modelling than prior to conditioning. The following explanations are possible:

The variance increase might be related to an error in the estimated prior sill and/or the range. If the estimated sill is too low, the ensemble variances of output variables of the flow and transport equation will also be too low. Also a range that is much longer than the real range will have a similar effect. At the same time, it is shown that inverse modelling is able to correct – partly – this error. As a result, the posterior ensemble variance is larger than the prior one. It is important to stress that the estimated variance is a model variance, and can even be based in some cases on postulations (e.g., a postulated variogram without any hydraulic conductivity measurements). Therefore the output variances – especially in cases with only few measurement data – should be looked at with some reservation. In relation to the variogram estimation, a research need remains to develop inverse stochastic methodologies that are more robust regarding the estimation of the hydraulic conductivity variogram, or are able to handle the uncertainty of the variogram (e.g. Bayesian methodologies).

Another possible explanation is that the modelling of the river-aquifer interaction was conceptually not completely correct. Further research is needed to handle the inverse stochastic modelling of river-aquifer interactions in an adequate and concise manner. In addition, in order to enable a successful inverse stochastic simulation of the river-aquifer flow conditions more measurement locations along the river are needed.

The little amount of inverse conditioned realisations. It is possible that 100 realisations would yield a considerably smaller posterior ensemble variance. Although there would still be an increase of the ensemble variance during conditioning, the increase could be less dramatic.

In general, a research issue remains the calibration of complex regional models, where not only the logtransmissivity but also the boundary conditions and the recharge are uncertain. In this case, averaging out the strong hydraulic conductivity variations in the vertical plane in a 2-D model may also be problematic. However, a fine discretization along the vertical axis was not feasible, and remains extremely CPU-intensive for inverse models.

Acknowledgements

The study was performed within the European Research Project "Stochastic Analysis of Well Head Protection and Risk Assessment" (W-SAHaRA). This project has been supported by the Swiss Federal Office for Education and Science (BBT), project 99.0543.

References

Gómez-Hernández JJ, Journel AG (1993) Joint sequential simulation of multi-Gaussian fields. In: Geostatistics Troia'92 volume 1, ed. Soares A: 85-94

Gómez-Hernández JJ, Sahuquillo A, Capilla JE (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data. 1. Theory. Journal of Hydrology, 1-4(203): 162-174

Hendricks Franssen HJ (2001) Inverse stochastic modelling of groundwater flow and mass transport. PhD dissertation. Technical University of Valencia

Martac E, Ptak T (eds) (2002). Data sets for transport model calibration/validation, parameter upscaling studies and testing of stochastic transport models/theory; Part I. Deliverable D16, W-SAHaRA

Stauffer F, Hendricks Franssen HJ, Kinzelbach W (2004) Semi-analytical uncertainty estimation of well catchments: Conditioning by head and transmissivity data. Water Resources Research, 40(8), W08305, doi:10.1029/2004WR003320.

“Soft” geostatistical analysis of radioactive soil contamination

R. Parkin[1], E. Savelieva[1] and M. Serre[2]

[1] Nuclear Safety Institute Russian Academy of Sciences, B.Tulskaya 52, 113191, Moscow, Russia
[2] Center for the Integrated Study of the Environment, School for Public Health, University of North Carolina, USA

1 Introduction

Monitoring data used in exposure mapping and risk assessment is usually associated with uncertainty caused by measurement equipment error and measurement methodology problems. These measurement uncertainties are difficult to process in the framework of classical geostatistics. As a result, the value at a monitoring site is often replaced by a “hard” datum that is treated as exact, and is determined by an official using an expert interpretation that takes into account the measurement methodology and raw measurements. In reality this value is not a “hard” datum, it is truly just an “expert” value and it is associated with some uncertainty. Such value we refer to as a “soft” datum.

The geostatistical methods pay special attention to the description of the uncertainty of the obtained result. The tradition starts from the classical measure of uncertainty – kriging variance, which later evolved to probabilistic description provided, for example, by indicator (Goovaerts 1997) or disjunctive kriging (Rivoirard 1994). Usually input information obligatorily includes “hard” data, and the “soft” data are used as additional information that allows to improve the estimation, but still it is treated as an addition to exact “hard” data (Savelieva et al 2003). The current work is an attempt to work exclusively with “soft” data – different repeated raw measurements, minimizing any preliminary expert interpretation.

We used two different approaches to work exclusively with “soft” data – the Bayesian Maximum Entropy (BME) and the “soft” indicator kriging methods. Both these methods allow to incorporate “soft” probabilistic information and they provide probabilistic interpretation of the result. Since this work is mainly devoted to an attempt to deal rigorously with situations that do not involve any “hard” data, we refer to it as “soft” geostatistics.

When a sufficient number of repeated raw measurements are available at a given sampling site, these repeated measurements can be used to obtain the (soft) pdf describing the true contamination concentration at that site. Such soft pdfs are incorporated into the methods as probabilistic “soft” data. The main goal of this

work is to investigate whether an analysis based only on "soft" data leads to reasonable results that are useful for decision-makers. For this purpose we performed a special validation procedure using real data on ^{137}Cs radioactive soil contamination caused by the Chernobyl fallout.

The validation procedure is complicated by the absence of data providing the true contamination level with which to compare estimated values. So, the validation process is made of two steps – comparison between the most probable value provided by the methods' posterior pdf with the value that occurred most often in the data set of repeated raw measurements; and comparison between methods' posterior pdf with the pdf of the repeated raw measurements. According to results obtained it is evident that "soft" geostatistics provides promising results and opens an area for future research work.

2 Some theoretical remarks on the methods used

2.1 The BME method

A detailed description of the BME theory is certainly beyond the scope of this paper, and the interested reader can find all computational and theoretical aspects of the method in Serre and Christakos (1999) and in Christakos (2000) and practical recommendations concerning the application together with the BME-based computer software BMELIB library in Christakos *et al.* (2002). In this section we briefly discuss the main features of the BME method that are relevant to the present work.

The spatial distribution of a physical variable (in our case, the radioactive soil contamination by ^{137}Cs) is routinely represented by means of a spatial random field (SRF) $X(\boldsymbol{s})$, where the vector $\boldsymbol{s}$ denotes spatial location. The BME mapping framework integrates various physical knowledge bases, such as the general knowledge base $\mathcal{G}$ (physical laws, empirical relations, statistical moments of any order, scientific theories etc) and the site-specific knowledge base $\mathcal{S}$ (real measurements, uncertain observations, secondary information etc) to construct the posterior pdf of $X(\boldsymbol{s})$ at any mapping point $\boldsymbol{s}_k$. In the Chernobyl fallout case considered in this work, the general knowledge $\mathcal{G}$ was limited by the variogram of SRF (the bar denotes stochastic expectation)

$$\gamma_X(\boldsymbol{s},\boldsymbol{s}') = \tfrac{1}{2}\overline{(X(\boldsymbol{s}) - X(\boldsymbol{s}'))^2}\,, \tag{1}$$

which expresses the spatial correlation between any two points $\boldsymbol{s}$ and $\boldsymbol{s}'$. The site-specific knowledge $\mathcal{S}$ includes only the set of soft data $\boldsymbol{\chi}_{\text{soft}}$ at points $\boldsymbol{s}_{\text{soft}}$. An example of $\boldsymbol{\chi}_{\text{soft}}$ used in this work are soft probabilistic data in the form of the soft pdf $f_S(\boldsymbol{\chi}_{\text{soft}})$, which represent the distribution of repeated measurements at the sampling sites $\boldsymbol{s}_{\text{soft}}$. The BME approach consists of three main stages of synthesizing and processing the general and site-specific knowledge bases, as follows:

1. *Structural stage*: The general knowledge $\mathcal{G}$ is considered and used to derive the structural (or prior) pdf model, f_G, of the radioactive ^{137}Cs contamination at all mapping points $\boldsymbol{s}_{map}=(\boldsymbol{s}_{soft}, \boldsymbol{s}_k)$.
2. *Specificatory stage*: The site-specific knowledge $\mathcal{S}$ is organized in terms of the soft data $\boldsymbol{\chi}_{soft}$ so that, $\boldsymbol{\chi}_{map}=(\boldsymbol{\chi}_{soft},\boldsymbol{\chi}_k)$.
3. *Integration stage*: The total knowledge base, $\mathcal{K}=\mathcal{G}\cup\mathcal{S}$, is assimilated by means of an operational Bayesian conditionalization rule, thus leading to the posterior pdf f_K of ^{137}Cs soil contamination at all mapping points as follows

$$f_K(\boldsymbol{\chi}_{\boldsymbol{\kappa}}) = A^{-1} \int d\boldsymbol{\chi}_{soft} f_S(\boldsymbol{\chi}_{soft}) f_G(\boldsymbol{\chi}_{map}) \tag{2}$$

where A is a normalization parameter. Clearly this posterior pdf is not limited by any specific form (since it is a function of the soft pdf, which may take any arbitrary form), leading to a realistic stochastic description of the radioactive contamination across space.

The posterior pdf, f_K, varies across space offering a complete stochastic description of the ^{137}Cs contamination at each mapping point. BME estimate of ^{137}Cs contamination at each mapping point is the most probable value according to obtained posterior pdf. Uncertainty can be described by confidence intervals, also estimated basing on the posterior pdf.

2.2 "Soft" indicator kriging

"Soft" indicator kriging, mentioned in the work of Saito and Goovaerts (2002), is a generalizing modification of the well-known indicator kriging. The main modification is that the indicator value is not a discrete function with values 0 and 1, but is a continuous function taking a value in an interval (0,1). Such continuous indicator transform is possible only based on some "soft" knowledge presented as local pdf, which is available in our case from the sets of repeated samples.

Again we consider a SRF $X(\boldsymbol{s})$, where the vector $\boldsymbol{s}$ denotes spatial location. The continuous indicator transform for location $\boldsymbol{s}$ and level z_n can be written as

$$i(s; z_n) = \Pr\{X(s) \le z_n\} = F(s; z_n) \tag{3}$$

and it is estimated based on the modeled local pdf for location $\boldsymbol{s}$ $(p(\boldsymbol{s};z))$ by a numerical integration:

$$F(s; z_n) = \int_{-\infty}^{z_n} p(s; z)dz\,. \tag{4}$$

Using kriging for indicators we obtain the estimates of cumulative distributions $F^*(\boldsymbol{s}_k;z|\ \mathcal{S})$ conditional for site specific initial information.

Modeled cdf, $F^*(\boldsymbol{s}_k;z|\ \mathcal{S})$, varies across space allowing probabilistic mapping of the ^{137}Cs contamination to exceed (or not to exceed) a corresponding level. It provides a probabilistic description of the uncertainty.

3 Description of the data

For our analysis we use the data on soil radioactive contamination by ^{137}Cs due to the Chernobyl fallout. The region under study is located approximately 200 km North-East of the Chernobyl Nuclear Power Plant (South-Western part of Briansk region). The territory of the study polygon is approximately 420 km^2 (70 km by 80 km). The real geographic coordinates are transformed to metric Lambert coordinate system projection, with (0,0) coordinate referring to the center of the most contaminated area in Russia. In the current work we focused on the area containing more locations with large number of measurements.

The dataset used in this work consists of 537 sampled locations with more than 3 measurements. Repeated measurements were taken in a close neighborhood (much smaller then the average distance between sample locations) of the position prescribed to a sample set, and the measurements taken at different times where all used to back-calculate the contamination value at the date of Chernobyl accident – 26.04.1986. None of temporal trends were studied, as real time of a measurement was unknown. Thus the distribution of the repeated measurements describes nothing than uncertainty.

At 410 sampled locations of our database we only have a minimum, a maximum and a value deemed as "official". These "official" values are truly the result of some kind of expert's judgment, so their usage as "hard" exact knowledge is not accurate. Here we treat the "official" values as the most probable value for the location and we construct for these locations triangular pdfs as the "soft" data.

The other 127 locations have more than 15 measurements each. Such number of measurements allows to perform an analysis of the local distribution of raw values, which can be used in place of less informative "official" values. Several examples of raw histograms and fitted pdfs of known probability distribution function models are presented in Fig.1. The quality of pdfs fitting was checked by traditional statistical tools (qq-plots).

36 locations from the 127 described above were extracted for the validation purpose, so the final training data set is composed from 501 location described by soft pdf's fitted to the repeat raw measurements.

4 Application

4.1 Data preparation

The BME method was used to process the knowledge provided by the "soft" probabilistic data (pdfs) and the model of spatial correlation structure of the variable under study (^{137}Cs soil contamination). The usage of the soft data to model the spatial correlation structure poses a problem of estimation, which is out of scope of the current work and will be considered in future research. In the current work the correlation structure was modeled using the most probable values. The

experimental variogram appeared to be close to isotropic, especially for small distances (≤10 km). Hence a simple isotropic model was used (Fig. 2).

"Soft" indicator kriging requires additional preliminary analysis. The selection of an adequate set of cut levels is a usual problem for the indicator approaches. In the current work we used a set of 19 cuts, based on 7 quantiles estimated for the most probable values (1/8q – 144.4; 1/4q – 228.85; 3/8q – 327.82; 1/2q – 448.1; 5/8q – 618.35; 3/4q – 829.3; 7/8q – 1179.23) and some additional values from each interval. The final set of cut levels is the following: 30 70 144 160 229 270 328 380 448 520 618 720 829 870 970 1070 1179 2000 3000.

The variogram analysis for the data after the "soft" indicator transform indicated 4 groups of variogram structures (I – first 4 cuts; II – following 4 cuts; III – following 8 cuts; IV – last 3 cuts). Group III is the only group presenting clear isotropic structure. Group IV is the most complex for modeling, because of the high variogram variability. In Fig. 3 we present examples (1 for each group) of experimental indicator variograms and the variogram models selected.

BME and "soft" indicator kriging predictions of the ^{137}Cs radioactive contamination distribution for each mapping point (both from validation data set and from the regular grid) were computed using the "soft" data from the nearest minimum 5 up to 10 samples found within the circular search area with radius of 25 km. No un-estimated locations were detected.

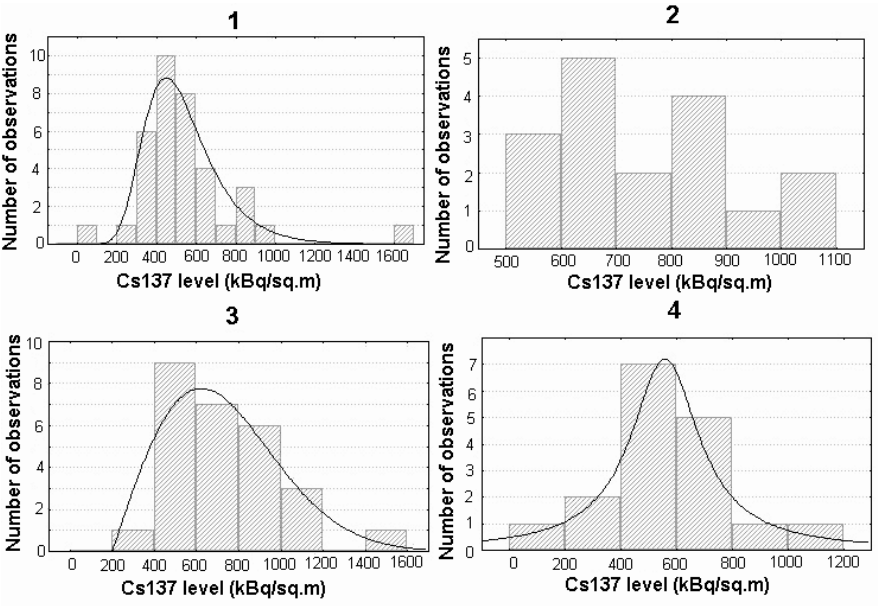


Fig. 1. Examples of raw histograms and fitted models of pdf (scaled to be compared with the histogram): **1** – Extreme distribution; **2** – Uniform distribution; **3** – Rayleigh distribution; **4** – Cauchy distribution

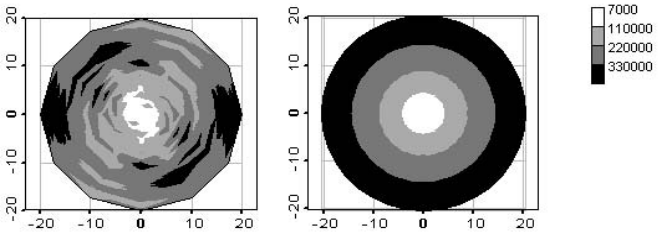


Fig. 2. Experimental variogram and variogram model roses for most probable values of ^{137}Cs soil contamination

4.2 Validation

Validation of the results of our analysis is the most important part of the current work, as it can show the ability of "soft" geostatistical methods. As mentioned earlier, performing validation is a problem in the absence of any exact measurements, so we use the value that occurred most often in the data set of repeated raw measurements (the mode of the raw histogram) for validation purpose. The BME method provides the posterior pdf at the estimation point, from which we can rigorously derive the most probable value (the BME mode estimate). So for the case of BME such type of comparison can easily be performed and yields a set of BME estimation errors calculated as the BME mode estimates minus mode of the raw histogram for each of the 36 validation locations. The mean error obtained for the BME errors is negative (-35.14), which indicates some overestimation. The statistical distribution of BME errors is rather symmetric (the 1/4 and 3/4 quantiles are correspondingly –151.6 and 104.8). The correlation coefficient between BME mode estimates and the mode of the raw histograms for corresponding validation locations is 0.8, which indicates a satisfactory correspondence.

In the case of the "soft" indicator kriging method, we only obtain the probability cumulative distribution function (cdf) estimated for set of selected cut values. The pdf can be crudely calculated as a numerical derivative of the cdf, but the most probable value according to such pdf is always equal to a value of a cut, as the change in the slope of the cdf can only be detected for the cuts. Thus, for "soft" indicator kriging we can only estimate the interval where the most probable value belongs. According to our results for 27 of 36 validation locations the interval containing the highest values of pdf covers the mode of the raw histogram, which indicates a good performance (75%).

The other aspect of comparison is the reproduction of local pdfs. For the comparison we used the approach widely distributed in statistics – qqplots (Fig. 4). In the current case qqplot is constructed as the graph of values of quintiles from the posterior distribution versus the values of same quintiles according to raw samples. The solid line indicates the bisector – the closer the markers line is to a bisector, the better is the reproduction of the local pdf. One can see, that both methods

give rather good reproduction and it is not easy to indicate which one is better. It means that both methods are useful within the "soft" geostatistical approach.

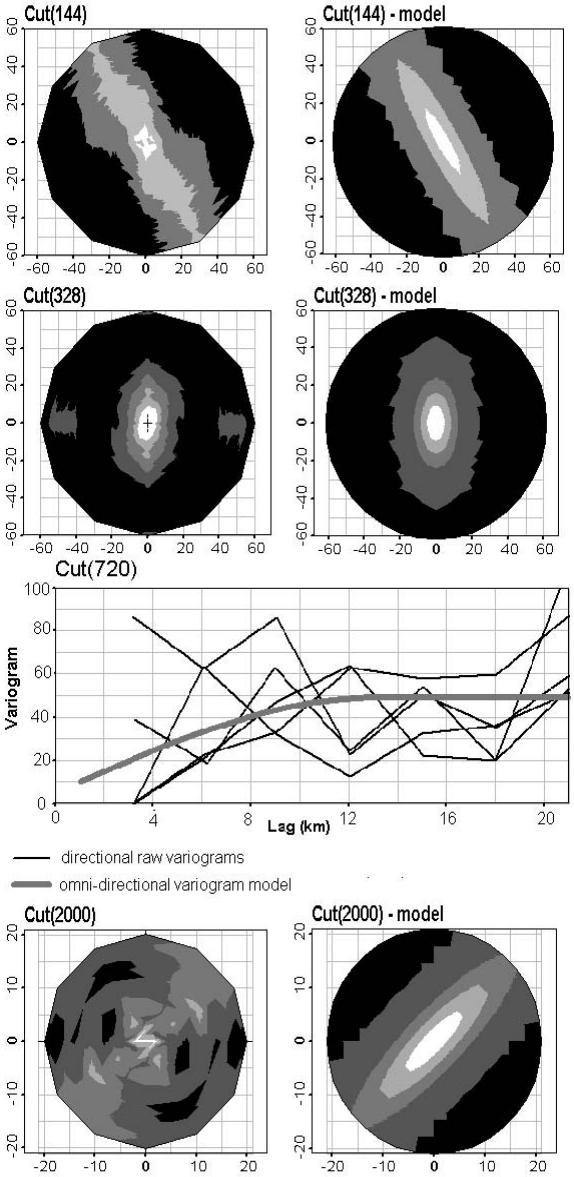


Fig. 3. Examples of raw indicator variograms and their models

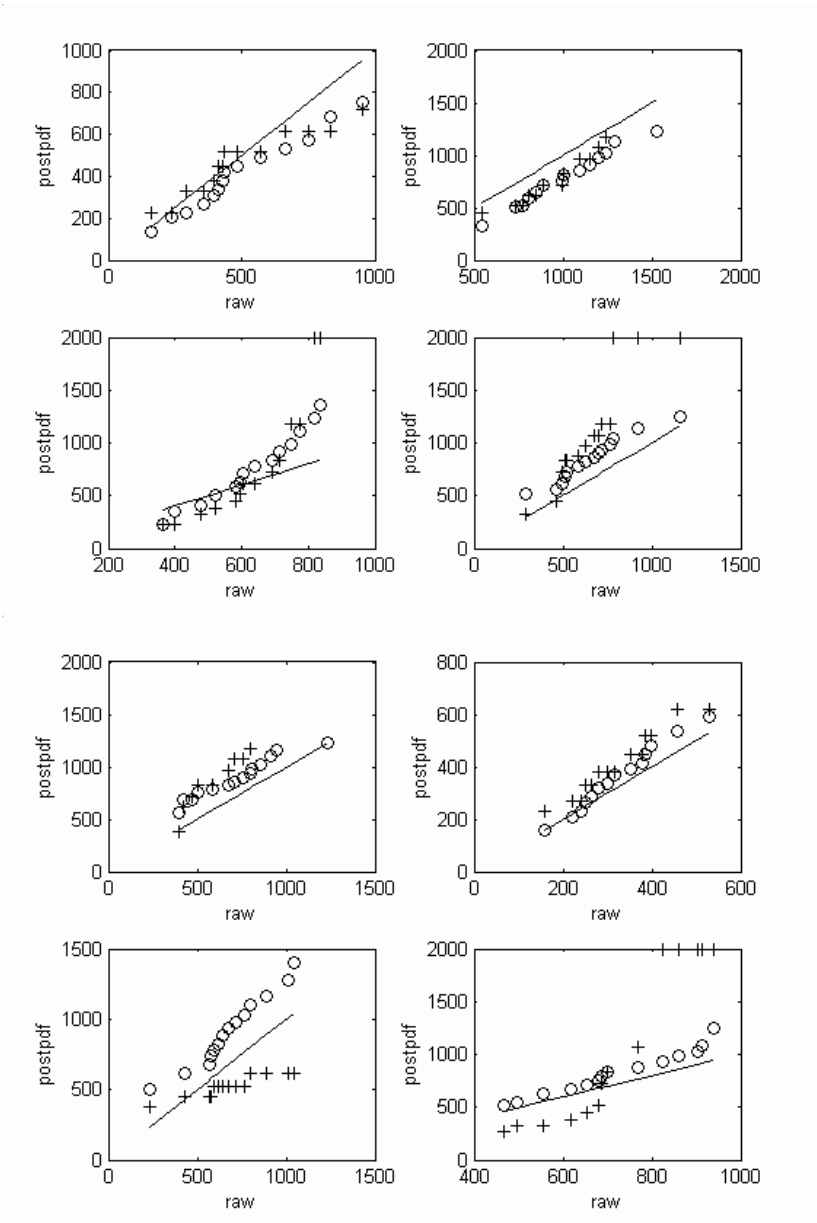


Fig. 4. QQplots for comparing reproduction of posterior pdfs: + correspond to "soft" indicator kriging pdf, O correspond to BME posterior pdf, line corresponds to abisector

4.3 Mapping Results

As a final result decision-makers usually want to obtain the estimation on the dense grid with the corresponding interpretation of uncertainty. "Soft" geostatistics can be used to obtain such information. Results on the dense grid (70x80 cells with a size of 1x1 km^2 – the total number of nodes is 4200) provided by both methods are presented in Fig. 5 and 6.

Fig. 5 presents four probabilistic maps obtained from "soft" indicator kriging. The darkness of a pixel in the figure is in accordance with the probability not to exceed the corresponding level indicated above the map (the darker, the higher probability).

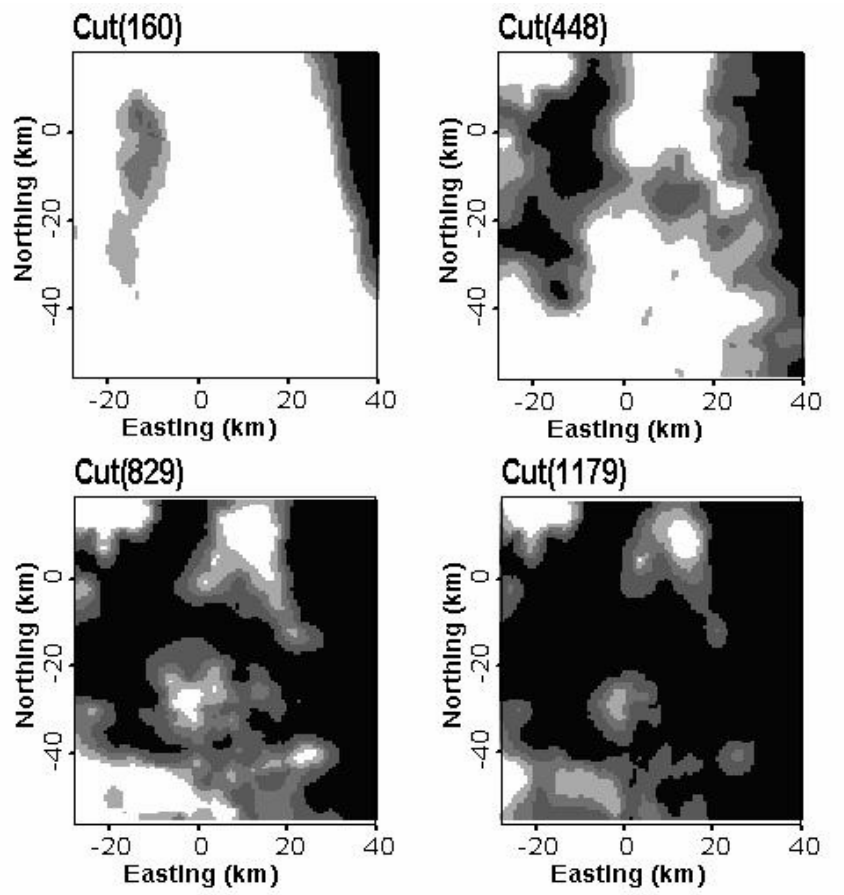


Fig. 5. Probability mapping after "soft" indicator kriging: the darker the higher probability not to exceed the cut level

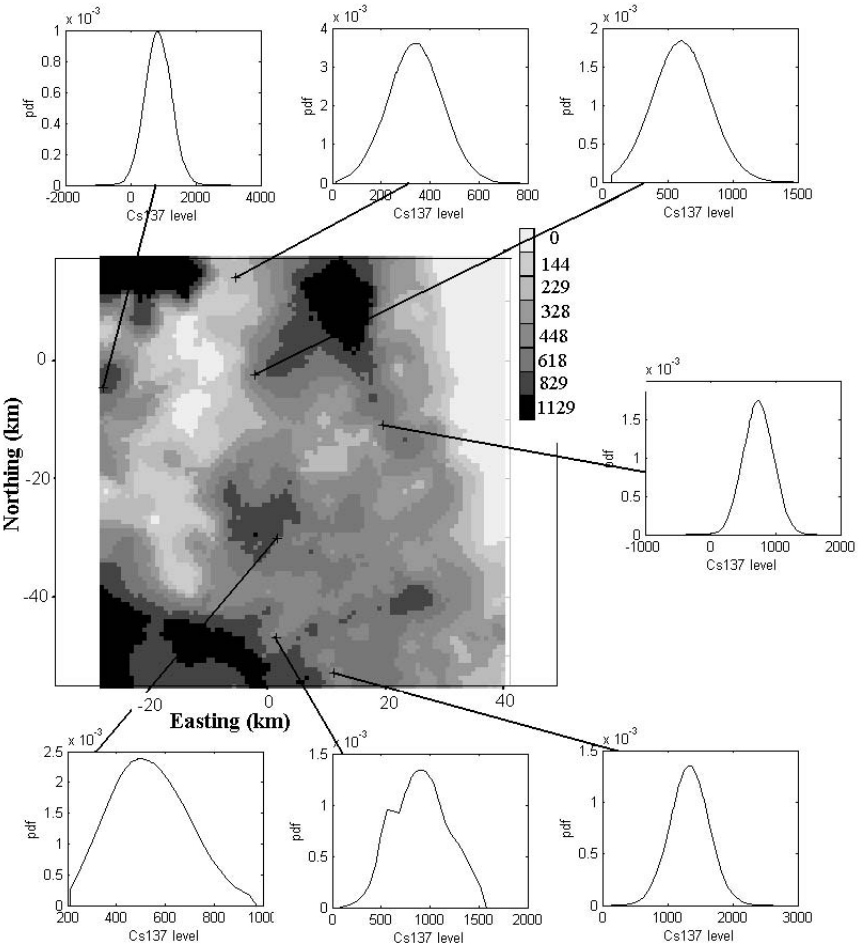


Fig. 6. BME estimates surface with several examples of posterior pdfs

In Fig. 6 we present a map of the BME estimates (the most probable value according to the BME posterior pdf) together with seven examples of BME posterior pdfs. BME pdfs characterize the uncertainty. Not all BME posterior pdfs are symmetric. Also the dependence of pdf's width can be observed: the pdf is wider where the gradient of the variable is higher – hence higher uncertainty is related to the prediction of spatial changes in contamination.

Also we present cumulative distribution functions (cdfs) estimated using "soft" indicator kriging (Fig. 7) and BME (Fig. 8) for 6 locations. These cdfs can be used to compare the local probabilistic characteristics of results by different approaches. The correspondence in the features of "soft" indicator kriging cdfs and BME posterior cdfs is visible.

The general agreement between two different approaches is evident from both spatial presentation of results and their local probabilistic characteristics

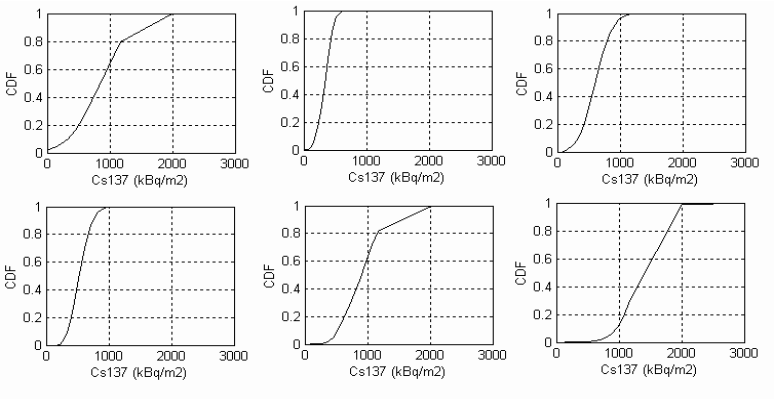


Fig. 7. Examples of cumulative distribution functions after BME

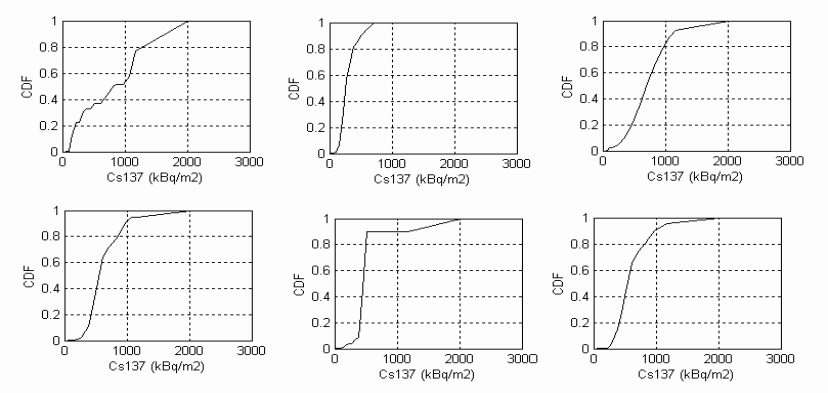


Fig. 8. Examples of cumulative distribution functions after "soft" indicator kriging

5 Concluding remarks

Two different approaches were applied to the spatial analysis of data presented as pure "soft" data, without any exactly known ("hard") information. "Soft" data was interpreted in the probabilistic way – as pdf. The main conclusion of this work is that "soft" geostatistical analysis may provide reasonable results of both spatial distribution of the value and the associated uncertainty. This conclusion is based on both validation procedures presented in the current work.

Both methods (BME and "soft" indicator kriging) provided reasonable results and it is not easy to make a conclusion whether one performs better. Even though

the comparison of these two methods is not the objective of this work, still some comparisons concerning their abilities and simplicity of application can be made:

- "Soft" indicator kriging does not always allow to estimate precisely the most probable value. This is due to the limited number of steps of the cumulative distribution function.
- The BME method can be used for probabilistic mapping, as presented in the current work.
- "Soft" indicator kriging allows to introduce the spatial correlation structure without any problems by a set of indicator variograms, but estimation and modeling of large amount of indicator variograms requires a lot of expert work. The situation allowing to apply the median kriging (Goovaerts 1997) is not a frequent one. Otherwise application of a set of variograms makes "soft" indicator kriging a rich tool for spatial evaluation, perhaps even more rich then BME version using one model as in the current work.

The analysis of "soft" data also leads to a set of interesting problems for future theoretical and practical research. One of the problems is connected with "soft" analysis of spatial correlation structure. Another problem is connected with the validation, which needs to be more directed to the estimation of the difference between probability distributions. Special qualitative measures are required for comparison of numerical presentation of statistical distributions.

Acknowledgements

The work is partly supported by INTAS Aral sea project 00-1072.

References

Christakos G (2000) Modern Spatiotemporal Geostatistics. Oxford University Press, New York

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS. Springer-Verlag, New York

Goovaerts P (1997) Geostatistics for Natural Resources Evalution. Oxford University Press, New York Oxford

Rivoirard J (1994) Introduction to Disjunctive Kriging and Non-linear Geostatistics. Clarendon Press, Oxford

Saito H, Goovaerts P (2002) Accounting for measurement error in uncertainty modeling and decision making using indicator kriging and p-field simulation: Application to a dioxin contaminated site. Environmetrics, 13: 555 – 567

Savelieva E, Demyanov V, Kanevski M, Serre M, Christakos G (2003) BME Application for Uncertainty Assessment of the Chernobyl Fallouts. In: Book of Abstracts Pedometrics 2003, University of Reading, pp. 36-37

Serre ML, Christakos G (1999) Modern geostatistics: Computational BME in the light of uncertain physical knowledge--The Equus Beds Study, Stochastic Environmental Research and Risk Assessment, 13: 1-26

Modelling the spatial distribution of copper in the soils around a metal smelter in northwestern Switzerland

A. Papritz[1], C. Herzig[1], F. Borer[2] and R. Bono[3]

[1]Institut für terrestrische Ökologie, ETH Zürich, Grabenstrasse 3,
CH-8952 Schlieren, e-mail: andreas.papritz@env.ethz.ch
[2]Amt für Umwelt des Kantons Solothurn, Werkhofstrasse 5, CH-4509 Solothurn
[3]Amt für Umweltschutz und Energie des Kantons Basel-Landschaft,
Rheinstrasse 29, CH-4410 Liestal

1 Introduction

Since 1895 a metal smelter operates in the village of Dornach (northwestern Switzerland). The smelter produces copper products and alloys for the manufacturing industries. Until 1972 metal dust, containing mainly copper, zinc and cadmium, was released into the atmosphere without any filtering. During the next decade filters were installed, and this greatly reduced the emissions. Since the end of the 1980s the emissions of the smelter comply with the Swiss air quality standards. Nevertheless, because of the long period during which the smelter was in operation, the soils around the smelter are polluted by heavy metals. A first survey, conducted in 1983-1986, revealed elevated metal concentrations over an area of a few square kilometers (Wirz and Winistörfer 1987). In the vicinity of the smelter concentrations up to a few g Cu per kg of soil were recorded in the topsoil. The Cu content of the soil reaches background levels only at a distance of 2-3 km from the smelter. Additional surveys conducted in the 1990s (cf. Keller *et al*. 1999) confirmed these findings.

In 1997/98 the Swiss environmental protection law was revised. A new ordinance introduced *guide*, *trigger* and *clean-up thresholds* for the various heavy metals. If the concentration exceeds the clean-up value then either restrictions on land use are imposed, or the land owner is requested to clean-up the land. If the content is below the clean-up but exceeds the trigger threshold then the authorities must evaluate the risk arising from the contamination. If there is some unacceptable risk restrictions will again be imposed. Concentrations exceeding only the guide threshold are less severe, here the authorities have to take measures to prevent a further rise of the concentration of the pollutant and to prevent uncontrolled displacement of soil from contaminated building grounds.

There is enough evidence that both clean-up and trigger thresholds are exceeded in Dornach (Fig. 1). The soil protection agencies and the owner of the smelter therefore started a new survey, with the aim to collect sufficient data to

predict for each parcel (contiguous piece of land belonging to same owner[s]) whether its heavy metal content exceeds any of the thresholds of the ordinance. In addition to the 236 sites studied so far, the heavy metal content of the topsoil will be measured at another 450 locations at the most.

In the past surveys observations were recorded at "points" (support of measurements 10-100 m^2). However, to classify the parcels according to their pollution, one must predict the mean metal content of entire parcels of land that may be much larger than the support of the data (mean area 2300 m^2). Thus, we face a non-linear and non-stationary spatial prediction problem with change of support. In this paper, we discuss the choice of a suitable geostatistical approach (section 2), we explain our analysis (section 3) and demonstrate its validity (section 4) by comparing the predictions, computed from past survey data, with the measurements obtained in the first stage of the new survey in summer 2003.

2 Review of modelling approaches for non-linear and non-stationary prediction problems with change of support

2.1 Disjunctive and indicator kriging

Non-linear prediction problems, arising in mining, prompted in the 1970/80s the development of disjunctive (DK) and indicator kriging (IK) (cf. overview in Chilès and Delfiner 1999, chap. 6). Whereas for DK the (Gaussian) discrete model offers a consistent method for change of support, *ad-hoc* variance corrections were suggested for IK (Oz *et al*. 2002). DK and IK further require that the bivariate distributions of a random process, $\{Z(\mathbf{s})\}$, are stationary. There have been attempts to overcome these limitations and to use DK and IK for modelling non-stationary patterns of spatial variation. Kolbjørnsen and Omre (1997) put DK into a Bayesian setting for modelling a non-stationary mean, but this is done at the price that the anamorphosis is no longer possible.

Goovaerts and Journel (1995) and followers tried to generalize IK for the same purpose (simple indicator kriging with varying local means). They suggested to compute the simple kriging predictions of the indicator residuals, $R(z;\mathbf{s}) = I(z,\mathbf{s}) - \hat{F}(z;\mathbf{s})$, where $I(z,\mathbf{s}) = 1$ if $Z(\mathbf{s}) \le z$ and $I(z,\mathbf{s}) = 0$ otherwise, $F(z;\mathbf{s}) = \mathrm{E}[I(z,\mathbf{s})]$ is the cumulative distribution function of $Z(\mathbf{s})$ and $\hat{F}(z;\mathbf{s})$ is some estimate of $F(z;\mathbf{s})$ (called local soft prior probability by Goovaerts). We denote the bivariate cumulative distribution function of a pair of random variables $(Z(\mathbf{s}),Z(\mathbf{s}'))$ by $F(z,z';\mathbf{s},\mathbf{s}') = \mathrm{E}[I(z;\mathbf{s})I(z';\mathbf{s}')]$. For the time being, let us assume that we know $F(z;\mathbf{s})$ for all $\mathbf{s}$ (which is not the case in reality and poses a difficult problem). Then using the definition of the variance of the difference of two correlated random variables we obtain for the variogram of the indicator residuals

$$
\begin{aligned}
2\gamma_R(z,z;\mathbf{s},\mathbf{s}') &= \mathrm{Var}[R(z;\mathbf{s}) - R(z;\mathbf{s}')] \\
&= \mathrm{Var}[R(z;\mathbf{s})] + \mathrm{Var}[R(z;\mathbf{s}')] - 2\mathrm{Cov}[R(z;\mathbf{s}), R(z;\mathbf{s}')] \\
&= \mathrm{Var}[I(z;\mathbf{s})] + \mathrm{Var}[I(z;\mathbf{s}')] - 2\mathrm{Cov}[I(z;\mathbf{s}), I(z;\mathbf{s}')] \\
&= F(z;\mathbf{s}) - [F(z;\mathbf{s})]^2 + F(z;\mathbf{s}') - [F(z;\mathbf{s}')]^2 \\
&\quad -2F(z,z;\mathbf{s},\mathbf{s}') + 2F(z;\mathbf{s})F(z;\mathbf{s}').
\end{aligned} \tag{1}
$$

Now, if $F(z;\mathbf{s})$ varies spatially, $\gamma_R(z,z;\mathbf{s},\mathbf{s}')$ is a function of the absolute positions $\mathbf{s}$ and $\mathbf{s}'$ and not merely of the displacement $\mathbf{s}-\mathbf{s}'$. Thus, we cannot get rid of the non-stationarity by working on the indicator residuals and we require that $\{Z(\mathbf{s})\}$ has stationary bivariate distributions. Only if $F(z;\mathbf{s}) = F(z;\mathbf{s}') = F(z)$ and $F(z,z';\mathbf{s},\mathbf{s}') = F(z,z';\mathbf{s}-\mathbf{s}')$ can we estimate the correlation structure of the indicator (residuals) by the customary method-of-moment estimator. Notwithstanding this difficulty, the advocates of the method estimate γ_R by grouping the centred indicator data into lag classes, and they proceed as if the indicator residuals were stationary (cf. Goovaerts 1997, p. 307-308). But we cannot see any grounds on which one might justify such a procedure and we conclude that the whole approach is flawed.

2.2 Trans-gaussian model and lognormal kriging

The observations, $z(\mathbf{s}_i)$, are modelled as a 1:1-transform, $\phi(\cdot)$, of a weakly stationary Gaussian random process, $\{Y(\mathbf{s})\}$, i.e. $Z(\mathbf{s}) = \phi(Y(\mathbf{s}))$. The Gaussian process has a linear mean function, $\mathrm{E}[Y(\mathbf{s})] = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$ and a stationary covariance $\mathrm{Cov}[Y(\mathbf{s}+\mathbf{h}),\ Y(\mathbf{s})] = C(\mathbf{h})$. Here, $\mathbf{x}(\mathbf{s})$ denotes the vector with the explanatory variables for location $\mathbf{s}$, $\boldsymbol{\beta}$ is the vector of the regression coefficients and ' denotes transpose.

A particular case is the lognormal model with $\phi(\cdot) = \exp(\cdot)$. This model has found some attention in the early 1980s, when attempts were made to model change of support for this approach (Rendu 1979, Journel 1980, Dowd 1982). The former two authors assumed that the distribution of point and block values, $\{Z(B)\}$, is jointly lognormal, although lognormality is not preserved when we average spatially. Dowd did not assume this, but the joint distribution of the predicted and true block means remain unspecified in his approach, and probabilistic statements are impossible. Marcotte and Groleau (1997) proposed yet another block LK-predictor that does not rely on the assumption of joint lognormality of $\{Z(B)\}$ and $\{Z(\mathbf{s})\}$. They suggested to predict a block mean, say $Z(B_0)$, by some linear predictor, $\hat{Z}(B_0)$, (e.g. by universal kriging) from the data. Then they assumed that the joint distribution of the predicted and the true block mean is bivariate lognormal and derived a closed-form expression for the conditional distribution of $Z(B_0) \mid \hat{Z}(B_0)$. For $\phi(\cdot)$ other than $\exp(\cdot)$ no closed from expressions are in use, and it appears that one has to resort to conditional gaussian simulations to

predict $Z(B)$ or any non-linear functional, say $g(Z(B))$, of that quantity. Of course, simulations can also be used when $\phi(\cdot) = \exp(\cdot)$ to avoid the approximation of joint lognormality of point and block values.

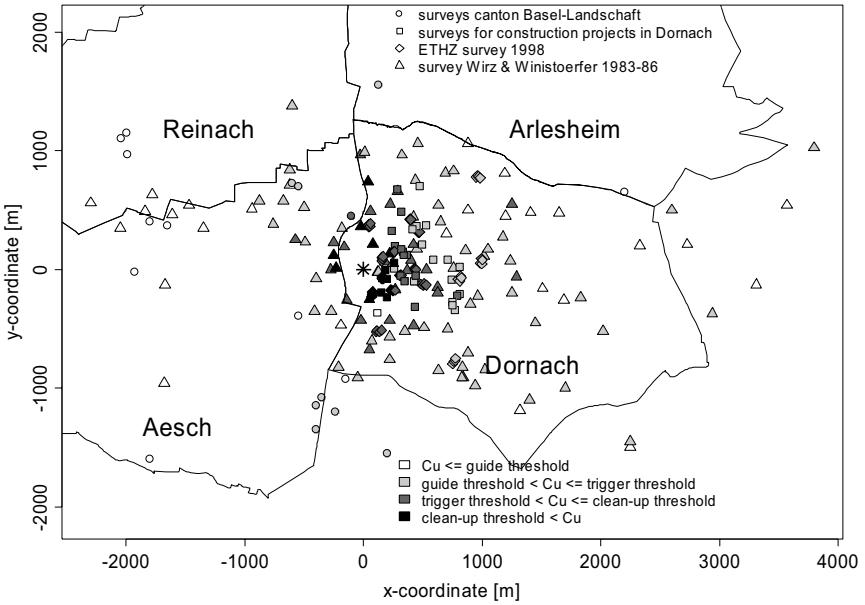


Fig. 1. Sampling locations of 4 soil pollution surveys around the metal smelter in Dornach. The star marks the position of the smelter. The grey level of the symbols codes the Cu content of the topsoil (0-20 cm depth). The coordinates are centred on the smelter.

Two remarks on the trans-gaussian model appear essential: (i) The trans-gaussian model provides a consistent way to parametrize a non-stationarity mean function. The transformed data, $y(\mathbf{s}_i) = \phi^{-1}(z(\mathbf{s}_i))$, are analysed by the well-established model building tools offered by regression analysis. Unlike the traditional paradigm of geostatistics which tends to consider the spatial dependence of the target data as the signal, we try to model the signal by the mean function and consider the remaining unexplained (and autocorrelated) part of the variation as a nuisance quantity. If we avoid an overparametrization of the mean function then we are likely to obtain more precise predictions because the explanatory variables will add some independent information at the prediction locations which we otherwise would not be able to incorporate into our algorithm.

(ii) By choosing the transformation function in the light of the results of the regression analysis we can model heteroscedastic patterns of variation. Fig. 2 (top left) shows that the local variation of the Cu content is not constant: close to the smelter, where the concentration is large, we observe a much larger spread of the data than farther apart. The absolute values of the residuals of the loess-smoother

increase with increasing concentration (Fig. 2, top right). A logarithmic transformation of the Cu content linearizes the relation to the logarithm of the distance (Fig. 2, bottom left) and leads to homoscedastic residuals (Fig. 2, bottom right). It is important to note that the type of transformation function should be chosen in the course of model-building and not *a priori*. In practice, a Box-Cox-transform is likely to stabilize the variance of the residuals in many instances.

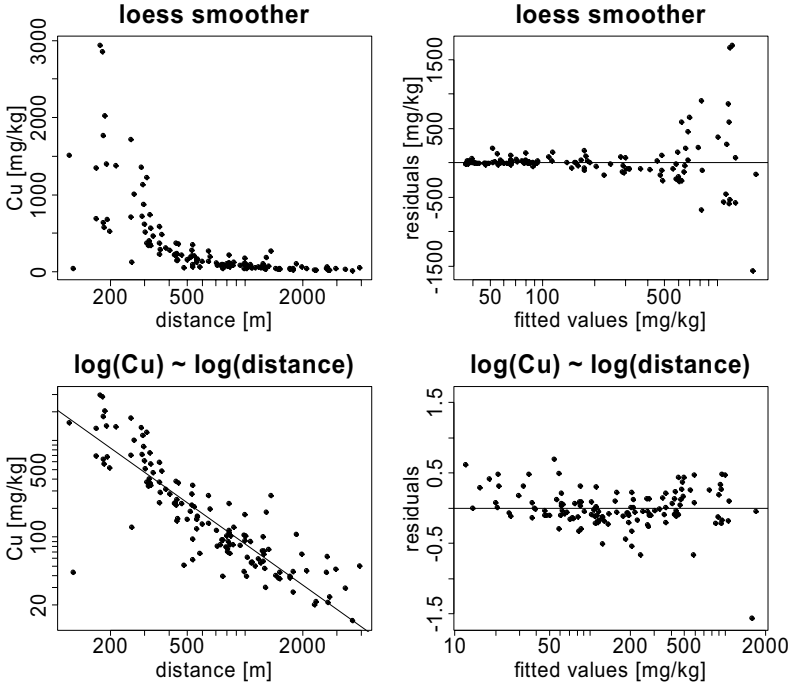


Fig. 2. Scatterplot of Cu content (top left) or logarithm of Cu content (bottom left) of topsoil samples (0-20 cm depth), plotted against the distance to the smelter. We show only the data of the survey locations east of the smelter (sector $\mathrm{NE} \rightarrow \mathrm{E} \rightarrow \mathrm{SE}$). The data are smoothed by a loess function (top left) or by a robust fit of a linear regression model (MM-estimator) where the logarithm of the distance was the only explanatory variable. The plots on the right show the residuals of the two smoothing approaches plotted as a function of the smoothed values (Tukey-Anscombe plots).

2.3 Further approaches: covariance-matching constrained and model-based kriging

Cressie (1993) suggested to predict a nonlinear functional $g(Z(B))$ by $g(\hat{Z}(B)) = g\left(\sum_i \nu_i Z(\mathbf{s}_i)\right)$ where the weights, ν_i, of the linear predictor are chosen such that the mean-squared error is minimized subject to the constraints

$\mathrm{E}\left[\sum_i \nu_i Z(\mathbf{s}_i)\right] = \mathrm{E}[Z(B)]$ and $\mathrm{Var}\left[\sum_i \nu_i Z(\mathbf{s}_i)\right] = \mathrm{Var}[Z(B)]$. For Gaussian $\{Z(\mathbf{s})\}$, $g(\hat{Z}(B))$ is an unbiased predictor for any $g(\cdot)$, and Cressie argues that $g(\hat{Z}(B))$ is approximately unbiased for non-gaussian $\{Z(\mathbf{s})\}$ and "smooth" $g(\cdot)$. Aldworth and Cressie (2003) generalized the constrained kriging predictor by imposing the additional constraints that the covariances of the predictions, $\hat{Z}(B_j)$, must match the covariances of $Z(B_j)$ for a set of blocks, B_j (covariance-matching constrained kriging CM). In the same way as ordinary kriging generalizes to universal (or external drift) kriging, CM can be used with a non-stationary linear mean function.

All the methods mentioned so far are "plug-in" procedures in the sense that they ignore the uncertainty in the estimates of the covariance structure when computing the predictions. The so-called parameter uncertainty can be taken into account by adopting a Bayesian approach to spatial prediction. Diggle *et al*. (1998) proposed model-based kriging, a Bayesian non-linear kriging method, which extends generalized linear mixed models to spatial random processes. However, our own experience (e.g. Moyeed and Papritz 2002) with the appraoch is that the computational burden is large and the implementation of the Markov-Chain-Monte-Carlo methodology requires considerable care.

2.4 Choice of modelling approach for Dornach survey

IK and DK are not valid approaches because neither method allows us to model a non-stationary mean function. Traditional lognormal universal block kriging relies on an inconsistent model for change of support. Both Marcotte and Groleau's LK-predictor and CM seem valid candidates, but rather little is known about their performance. Thus, in view of the legal implications that the results of the geostatistical analysis will have, we therefore decided to use the trans-gaussian model, combined with "plug-in" conditional simulations to predict the mean Cu content of parcels of land around the smelter in Dornach.

3 Geostatistical analysis of Dornach survey data

3.1 Structural analysis of Cu content of topsoil samples

Apart from the heavy metal data from 236 sites, incomplete and inconsistent information about the land use at the sampling locations was available. An exploratory analysis revealed that besides the distance to the smelter, land use had some effect on the Cu content of the topsoil (Fig. 3 left): forest sites had somewhat larger concentration than arable land, grassland and gardens were intermediate. Since we had only maps showing the spatial distribution of forests we could only model the influence of this land use. The orientation of a site relative to the smelter had some effect, too (Fig. 3, right): locations to the west of the smelter had smaller Cu content. However, since there were not many sites in this group (cf. Fig. 1), we

decided that we will model this effect only at the second stage of the new survey, when more data from sites in the west are available. Thus, we used the logarithm of the distance and an indicator variable for forest as the only explanatory variables in the regression model for log(Cu).

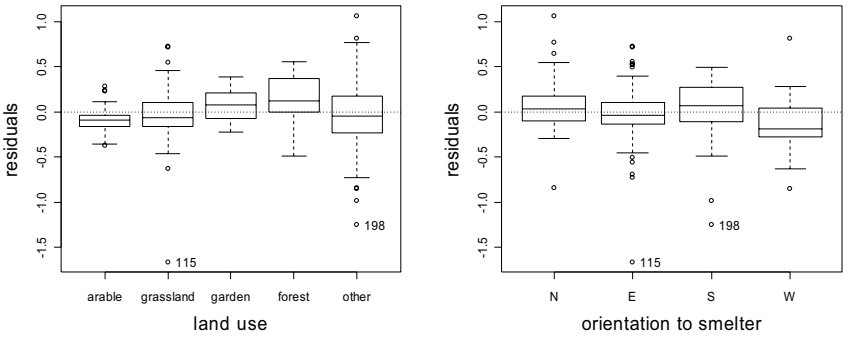


Fig. 3. Boxplots of residuals of a robust fit (MM-estimator) of the linear regression model log(Cu) ~ log(distance). The residuals are grouped according to land use (left) or orientation of the sampling locations relative to the smelter (right).

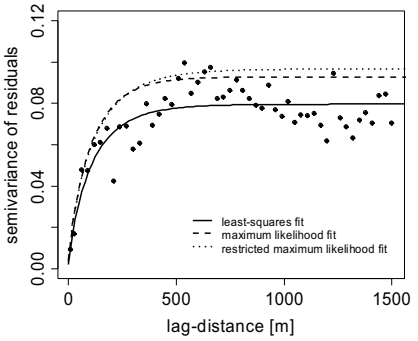


Fig. 4. Sample variogram of residuals of regression model log(Cu) ~ log(distance) + forest and graphs of estimated stable variogram model.

The distribution of the residuals showed heavier tails than the normal distribution. In particular, two sites (no. 115, 198, cf. Fig. 3) had very small residuals. These two observations were excluded when we studied the spatial dependence of the residuals because one observation appeared grossly wrong and the other was from a site where contaminated soil had probably been replaced by clean soil. Figure 4 shows that the residuals were indeed spatially correlated (nugget:sill ratio $\approx$ 0.1, range $\approx$ 200-400 m).

We fitted several model functions to the sample variogram (all combined with a nugget effect) by non-linear least-squares using Cressie's weights. In addition, we estimated the variogram model parameters by (restricted) maximum likelihood. Figure 4 shows the fits we obtained for the stable model (Chilès and Delfiner, 1999, p. 90). The various variants were then compared in a cross-validation using universal kriging (mean function model: $\log(\text{Cu}) \sim \log(\text{distance}) + \text{forest}$) for log(Cu). We used the bias, mean-squared error and MAD of the prediction errors to compare the precision, and we checked the modelling of prediction uncertainty by computing the coverage of one-sided prediction intervals (Papritz and Dubois 1999). The precision of the predictions did not differ markedly for the various models and fitting algorithms, but the coverage probabilities were consistently better when we used the maximum likelihood estimates. On the whole, the maximum likelihood estimates of the stable model seemed best, and we used this parameter set for the conditional simulations.

3.2 Conditional simulations of mean Cu content of parcels

We did not use sequential Gaussian simulations, but we conditioned unconditional realizations of log(Cu) by kriging (e.g. Chilès and Delfiner 1999, sec. 7.3). The unconditional realizations were simulated on a $7\text{x}7\text{m}^2$-grid by the fast circulant embedding algorithm of Chan and Wood (1997). In moderate to large problems such as ours (370'000 grid nodes) sequential algorithms can condition only locally and the order how the nodes are visited may introduce artifacts. Conditioning by kriging is straightforward and fast, provided that the matrix with the kriging weights can be kept in computer memory. The simulated values of log(Cu) were then transformed back to the original scale, and the Cu values of all the nodes that were in the same parcel were averaged. We simulated 2000 realizations of the mean Cu content of 7370 (area 18.101 km^2) out of 7780 parcels (area 18.112 km^2), the remaining parcels were too small to contain a node of the grid. This allowed us to approximate the conditional distribution of the parcel means numerically by a maximum error of 2% (95%-confidence level).

3.3 Choosing the sampling locations for the new survey

To select the new locations for sampling soil, we computed the conditional 5%-quantile, Q05, and 95%-quantile, Q95, of the mean Cu content of the parcels and grouped them into classes by comparing the quantiles with the thresholds of the ordinance and another two *ad-hoc* thresholds (trigger A [300 mg/kg] and trigger B [500 mg/kg]), chosen to subdivide the large range between trigger (150 mg/kg) and clean-up thresholds (1000 mg/kg).

For five parcels Q05 exceeded the clean-up threshold (Table 1), leaving little doubt that these pieces of land were heavily polluted. For another 95 parcels Q95 exceeded the clean-up threshold, but Q05 was either in class 5 (42), in class 4 (26) or even in class 3 (27). For all these parcels there was a risk > 5% that the clean-

up value might be exceeded. For the remaining parcels (Q95 in classes 1-5) the corresponding risk was less than 5%. If we accept a maximum misclassification rate of 5% for false negatives (polluted but not detected) then we can declare a parcel as safe if Q95 is not in class 6. However, by using Q95 as criterion, we incur a considerable number of false positives (erroneously declared polluted).

Table 1. Absolute frequency of parcels of land in 6 contamination classes and number of parcels sampled in first phase of new survey (in parentheses). Q95: 95%-quantile; Q05: 5%-quantile; class I: quantile ≤ guide threshold; class II: guide < quantile ≤ trigger threshold; class III: trigger < quantile ≤ trigger A threshold; class IV: trigger A < quantile ≤ trigger B threshold; class V: trigger B < quantile ≤ clean-up threshold; class VI: clean-up threshold < quantile.

	Q95 ∈ I	Q95 ∈ II	Q95 ∈ III	Q95 ∈ IV	Q95 ∈ V	Q95 ∈ VI
Q05 ∈ I	12 (0)	4713 (38)	1078 (18)	1 (0)		
Q05 ∈ II		82 (2)	644 (21)	415 (37)	134 (14)	
Q05 ∈ III			3 (0)	43 (3)	127 (20)	27 (5)
Q05 ∈ IV					16 (5)	26 (10)
Q05 ∈ V					2 (1)	42 (39)
Q05 ∈ VI						5 (5)

The classification was done in the same way for the other thresholds. Table 1 summarizes the results. Then we used the following rules to allocate the new sites to the contamination classes:

1. All the parcels not yet sampled and with Q95 in class 6 will be selected. By that we avoid any false positives and we ensure that for the clean-up threshold the rate of false negatives is less than 5%.
2. The other sites will be chosen among parcels with Q95 in classes 4 or 5. Thus, the rate of false negatives is also limited to 5% with respect to trigger A and B thresholds.
3. Parcels with 95%-quantile in classes 1 to 3 will not be sampled except for a small number of sites west of the smelter. These data will be used to model the effect of orientation in later stages of the survey.

When selecting the new sites, we tried to fill in gaps in the spatial arrangement of locations sampled previously. When several parcels seemed equally fit then we selected that with the largest conditional coefficient of variation.

4 Validation of geostatistical analysis of Dornach data

In summer 2003 soil was sampled on 217 parcels of land (cf. Table 1) and chemically analysed. An additional datum was available from another project. Figure 5 compares the measured and predicted Cu content of the 218 samples. Since the support of the measurement was not constant we either used the simulation results for sites with Q95 in class 6 or with irregularly shaped sampling area or punctual universal lognormal kriging for sites where a 10x10m^2-square had been sampled.

Except for a few samples with very large Cu content the predictions matched the data fairly well (Fig. 5). There were a number of sites where the predictions were too large. Based on a comparison with data from adjacent locations and on inspection of aeral views, we suspect that polluted soil had been replaced at 13 sites by clean soil. In general, the differences between predicted and measured content increase with increasing concentration (Fig. 5 left). Therefore, it seems more natural to compare the logarithms of the concentration (Fig. 5 right). On the log-scale (results for the linear scale in parentheses) the mean prediction error was -0.078 (-18), the root mean square error was 0.316 (396) and the correlation was equal to 0.84 (0.75). The predictions were somewhat conditionally biased (predictions too large for large content), but this was partly due to the sites with suspected soil exchange.

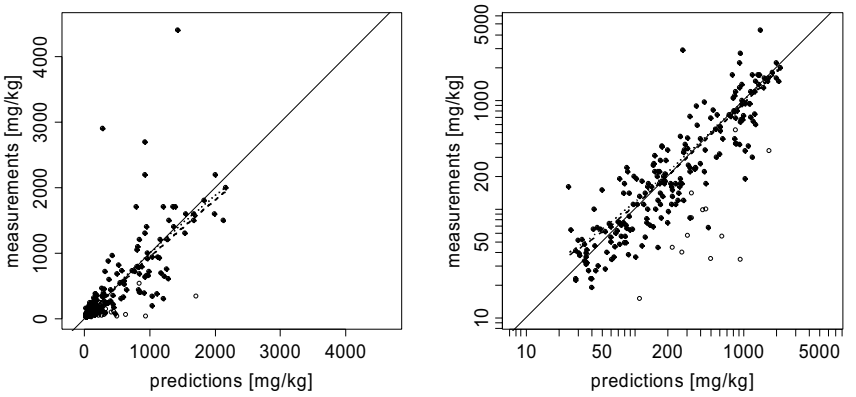


Fig. 5. Scatterplot of measured vs. predicted Cu concent of soil samples, collected in the first phase of the new survey around the metal smelter in Dornach, linear scale (left) and doubly logarithmic scale (right). Open symbol: samples from parcels of land where contaminated soils was probably exchanged by clean soil; dashed line: loess smoother fitted to all the data; dotted line: loess smoother fitted without open symbols.

We assessed the modelling of prediction uncertainty by the coverage of one-sided predictions intervals (Papritz and Dubois 1999). The coverage was too large for small nominal probabilities (Fig. 6 left). This indicates that we underestimated the extent of the lower tails of the conditional distributions, but also this failure was partly due to the sites with suspected soil removal. The coverage was a bit too small for probabilities > 0.9. This shows that also the extent of the upper tails of the predictive distributions were somewhat underestimated.

We further computed the frequency of false negatives and positives when the percentiles of the conditional distributions were used as criteria for exceedance of the clean-up threshold (Fig. 6 right). The observed misclassifcation rate for false negatives was always smaller than the allowed maximum rate. The same was true for false positives. In accordance with theory, the sum of both types of misclassifications was minimized by percentiles close to the median of the conditional

distributions. When the median was used as criterion the misclassification rates were around 5% for both types of error, and this is deemed acceptable by the contracting authorities. In reality, there will be no false positives because the Cu content will be measured for all the parcels with Q95 > clean-up threshold.

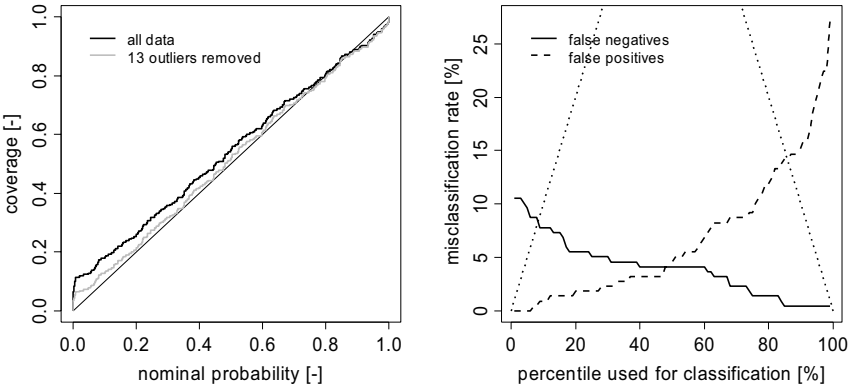


Fig. 6. Coverage probability plot of one-sided prediction intervals (left) and rates of misclassification when the percentiles of the predictive distribution are used as criterion for exceedance of the clean-up threshold (right). The maximum tolerable rates are shown by the dotted lines.

5 Conclusions

The review of non-linear kriging approaches revealed that conditional simulations with the trans-gaussian model are the method of choice to map the non-stationary distribution of pollutants around point sources. Although used in the past for the same purpose, simple indicator kriging with varying local means has no sound scientific basis and its use should be discouraged. With the trans-gaussian model, we can use the well established tools of linear regression analysis to model both non-stationary mean functions and heteroscedastic patterns of variation. Furthermore, change of support is straightforward with this approach.

The results of the validation demonstrate that conditional simulations with the lognormal model were quite successful to adequately describe the uncertainty of the predictions which involved change of support. At some sites, the prediction errors were very large. We suspect that undocumented removal and displacement of contaminated soil are the cause, and we think that this is a major obstacle when we try to precisely map the distribution of soil pollutants in settled areas.

References

Aldworth J, Cressie N (2003) Prediction of nonlinear spatial functionals. Journal of Statistical Planning and Inference 112: 3-41

Chan G, Wood ATA (1997) An algorithm for simulating stationary gaussian random fields. Journal of the Royal Statistical Society Series C 46: 171-181

Chilès JP, Delfiner P (1999) Geostatistics: Modeling spatial uncertainty. Wiley, New York

Cressie N (1993) Aggregation in geostatistical problems. In: Soares A (ed) Geostatistics Troia '92. Kluwer Academic Publishers, Dordrecht, 25-36

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussions). Journal of the Royal Statistical Society Series C 47: 299-350

Dowd PA (1982) Lognormal kriging—the general case. Mathematical Geology 14: 475-499

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Goovaerts P, Journel AG (1995) Integrating soil map information in modelling the spatial variation of continuous soil properties. European Journal of Soil Science 46: 397-414

Journel AG (1980) The lognormal approach to predicting local distributions of selective mining unit grades. Mathematical Geology 12: 285-303

Keller A, Jauslin M, Schulin R (1999) Bodendaten und Stoffffluss-Analyse im Schwermetallbelastungsgebiet Dornach. Bericht, Amt für Umweltschutz, Volkswirtschaftsdepartement des Kantons Solothurn

Kolbjørnsen O, Omre H (1997) Bayesian disjunctive kriging applied to prediction of reservoir volume. In: Baafi EY and Schofield NA (eds) Geostatistics Wollongong '96. Kluwer Academic Publishers, Dordrecht, 609-620

Marcotte D, Groleau P (1997) A simple and robust lognormal estimator. Mathematical Geology 29: 993-1008

Moyeed RA, Papritz A (2002) An empirical comparison of kriging methods for nonlinear spatial point prediction. Mathematical Geology 34: 365-386

Oz B, Deutsch CV, Frykman P (2002) A Visualbasic program for histogram and variogram scaling. Computers and Geosciences 28: 21-31

Papritz A, Dubois JP (1999) Mapping heavy metals in soil by (non-)linear kriging: an empirical validation. In Gómez-Hernández J, Soares A, Froidevaux R (eds) GeoENV II Geostatistics for environmental applications. Kluwer Academic Publishers, Dordrecht, 429-440

Rendu JM (1979) Normal and lognormal estimation. Mathematical Geology 11: 407-422

Wirz E, Winistörfer D (1987) Bericht über die Metallgehalte in Boden- und Vegetationsproben aus dem Raum Dornach. Bericht Kantonales Laboratorium, Solothurn.

Towards a real-time multi-phase sampling strategy optimization

D. D'Or

FSS International R&D, 10, Tienne-de-Mont, B-5140 Sombreffe, Belgique,
e-mail : dimitri.dor@fssintl.com
Département des Sciences du Milieu et de l'Aménagement du Territoire – Environnemétrie, Place Croix-du-Sud, 2 bte 16, B-1348 Louvain-la-Neuve, Belgium.

1 Introduction

Sampling is a crucial issue in many studies related to potentially contaminated sites. At the beginning of the study, the first sampling phase aims at assessing the presence/absence of a contamination. Once a contamination has been identified, the volume V_{rem} of soil to remediate has to be quantified as accurately as possible. To achieve this task, new sampling phases may be necessary. They should be optimized in order to minimize the uncertainty on the estimation and on the position of V_{rem}.

As objective function, existing methods may consider the minimization of the kriging variance (see, e.g., Van Groeningen *et al.* 1999). This function is inadequate in our context since our aim is not to reduce the uncertainty on the estimate of some random variable Z (the contaminant), but rather to minimize the uncertainty of some classification of Z at location $\mathbf{x}_i$.

In the literature, several methods are proposed. They can be classified as one-phase, two-phase or multi-phase. Most of them consider the optimization is made at the beginning of the study or after a first sampling phase. Anyway, they are not interactive since V_{rem} can only be computed once all the data collected during the various sampling phases are available.

In 1994, Englund and Heravi proposed an interactive multi-phase strategy. They compared it to a one- and a two-phase strategy using total remediation costs as criterion. In their conclusions, they recommend the use of the two-phase strategy even if the multi-phase results in lower costs. They conclude that the cost reduction obtained when using the multi-phase strategy is not sufficient to balance the logistical costs, which are not taken into account in their study.

In the last years, several techniques were developed to identify and characterize contaminations on site and nearly in real-time. Among many other possible techniques, let us mention: (i) Photo Ionization Detector (PID) or Flame Ionization Detector (FID) for volatile organic compounds, (ii) colorimetric reactive kit or Laser Induced Fluorescence (LIF) for BTEX, PAH and PCB, and (iii) Field Portable X-Ray Fluorescence for mineral components (e.g., lead, copper or cadmium).

With fast on-site measurement devices now readily available, it becomes possible to develop real-time on-site multi-phase sampling strategies. In such strategies, the most recently collected samples are immediately incorporated to the data set. The estimate of V_{rem} is then updated and the next sampling phase optimized.

In this paper, we first present the specific context in which we place this study. We then expose in detail the multi-phase sampling strategy. After that, we apply it to a simulated case study. Finally, we discuss its efficiency in relation with the choice of some parameters such as the number of samples to be collected in a same phase or the relative variance of the samples.

2 Context

Let us consider a contaminated site. A first sampling phase has been conducted, yielding a given number of accurate measurements about the target contaminant (e.g., from laboratory analyses). The analysis of this first data set has established the presence of a contamination and an investigation study has to be carried out in order to position and quantify the volume V_{rem} of soil to remediate. The available data are then used to classify the soil into one of the three following classes: safe, contaminated or uncertain.

A possible way for obtaining this classification consists in implementing a conditional sequential Gaussian simulation (SGS) procedure following the methodology described by Demougeot-Renard *et al.* (2004). A large number of realizations are produced conditionally to the data, allowing us to estimate at each location $\mathbf{x}_i$ the probability for the contaminant Z to exceed the critical threshold z_c by computing the proportion of realizations for which $Z(\mathbf{x}_i) \geq z_c$. Locations where this probability is higher than a given *Max* threshold (say, 80%) are classified as contaminated. Those with a probability to exceed z_c smaller than a given *Min* threshold (say, 20%) are considered as safe, and the remaining ones (with a probability between *Min* and *Max*) are classified as uncertain.

The choice of the *Min* and *Max* thresholds is a matter of economical efficiency and risk management. Indeed, with a *Max* threshold at 80%, there is potentially 20% chance of classifying safe soil as contaminated. Taking thresholds closer to the 0 and 1 probability bounds will reduce the error risk. But it is done at the cost of an enlargement of the uncertainty area where additional sampling should be conducted and it thus results in higher sampling costs. A detailed discussion of this issue may be found in Demougeot-Renard (2002).

This first simulation stage produces a soil classification map that will further be denoted as the *first intervention map* (Fig. 1c).

In order to delineate with the highest accuracy the soil volume to remediate, additional sampling has to be carried out. The objective of this sampling is to reduce the volume of the uncertain zone.

3 Methods

The development of every sampling strategy begins by defining or choosing: (i) a selection criterion for the sample locations, (ii) a method for updating the classification following the incorporation of new data, and (iii) an objective function. In this section, we review these points and justify our choices.

3.1 Selection of the next locations to sample

As we are interested in minimizing the uncertainty on the soil classification into safe/contaminated, we choose to select preferentially those locations where the uncertainty about soil classification remains the highest.

In order to quantify this uncertainty, several criteria may be considered. Considering that the binary classification into safe or contaminated corresponds to a Binomial situation at each location $\mathbf{x}_i$, uncertainty can be measured using the variance of a Binomial random variable:

$$\sigma^2(\mathbf{x}_i) = p_i.(1 - p_i) \quad (1)$$

Where $p_i = P(Z(\mathbf{x}_i) \leq z_c)$, i.e. the probability for the soil at location $\mathbf{x}_i$ to be safe.

This criterion is appropriate since (i) it reaches its maximum ($\sigma^2(\mathbf{x}_i) = 0.25$) when p_i is equal to 1- p_i (error risk equal to p_i and thus maximum uncertainty) and (ii) $\sigma^2(\mathbf{x}_i) = 0$ when either p_i or 1- p_i is equal to 1 (soil is then classified as safe or contaminated, respectively, with a probability of 1 and there is thus no uncertainty).

Practically, from the simulations described above, the probability for the soil to be safe has been inferred at each location by computing $P(Z(\mathbf{x}_i) \leq z_c)$. Using Eq. (1), it is straightforward to obtain a map of the variance, i.e. of the uncertainty on the soil classification (Fig. 1d).

The candidate locations for further sampling are selected within the set of local maxima on this map. Indeed, it is expected that the reduction in uncertainty will be the largest when sampling at these locations. At each phase of the sampling procedure n local maxima are sampled together. The classification is then updated in the neighborhood of these locations

The choice of the number n of locations to be sampled together may have an influence on the optimization. Its effect is analyzed and discussed further.

3.2 Update of the intervention map

At the n locations selected using the method exposed here above, measurements of the contaminant concentration are made using a fast on-site measurement technique.

However, most of this type of techniques, while being fast, are less accurate than laboratory analyses. They often yield semi-quantitative or qualitative information. That rather imprecise information will further be denoted as "soft" data. In the semi-quantitative case, a measurement error has to be accounted for: the true value may belong to a given interval of values (interval-type soft data), or a probability density function (pdf) for the true value may be inferred (probabilistic-type soft data). In the qualitative case, a distribution of the continuous variable may be estimated conditionally to the qualitative value using those locations where both types of variables have been observed. Again, this yields probabilistic-type soft data. In both cases, we have to deal with imprecise information and the estimation/simulation method that is implemented must be able to take that information into account without loss.

To achieve this task, we propose to use the Bayesian Maximum Entropy (BME) approach (Christakos 2000). Using this approach, the imprecise information delivered by the sampling device is incorporated as soft data and used to update the initial pdf's in the neighborhood.

By lack of place, we refer to Christakos (2000), Christakos *et al.* (2002), D'Or (2003) and to the abundant literature on the subject for a detailed explanation of the BME approach. We here only present the main steps of a BME study.

Consider the vector of random variables $\boldsymbol{Z}_{map} = (\boldsymbol{Z}_{hard}, \boldsymbol{Z}_{soft}, Z_0)$ and let $\boldsymbol{z}_{hard}$, $\boldsymbol{z}_{soft}$ and z_0 respectively denote the values at hard, soft and prediction locations. Let also $f_G(\boldsymbol{z}_{map})$ be the multivariate pdf accounting for the general knowledge $\boldsymbol{K}_G$, before any specific knowledge $\boldsymbol{K}_S$ has been considered.

The **prior step** aims at finding the most general joint pdf $f_G(\boldsymbol{z}_{map})$ while ensuring that all the available information is taken into account. This step is achieved using a maximum entropy procedure where the entropy is computed as:

$$H\left(f_G\left(\boldsymbol{Z}_{map}\right)\right) = -\int_{D_Z} f_G\left(\boldsymbol{z}_{map}\right)\log\left(f_G\left(\boldsymbol{z}_{map}\right)\right)d\boldsymbol{z}_{map} \tag{3}$$

The entropy is maximized under the constraint of respecting the prior available information $\boldsymbol{K}_G$, generally expressed as the global mean and a covariance function.

At the **posterior step**, we seek the posterior pdf for the random variable Z at prediction location $\mathbf{x}_0$, given the hard and soft data at hand:

$$f_K\left(z_0\right) = f_G\left(z_0 \middle| \boldsymbol{z}_{hard}, \boldsymbol{z}_{soft}\right) \tag{4}$$

When the soft data are of probabilistic type $f_S(\boldsymbol{z}_{soft})$, the soft pdf, or equivalently $F_S(\boldsymbol{z}_{soft})$, the soft cumulated density function, the BME solution is given by (Christakos 2000):

$$f_G\left(z_0 \middle| \boldsymbol{z}_{hard}, F_S\left(\boldsymbol{z}_{soft}\right)\right) = \frac{\int f_G\left(z_0, \boldsymbol{z}_{hard}, \boldsymbol{z}_{soft}\right) f_S\left(\boldsymbol{z}_{soft}\right) d\boldsymbol{z}_{soft}}{\int f_G\left(\boldsymbol{z}_{hard}, \boldsymbol{z}_{soft}\right) f_S\left(\boldsymbol{z}_{soft}\right) d\boldsymbol{z}_{soft}} \tag{5}$$

Eq. (4) yields at each location an entire (non-discretized) posterior pdf from which various statistics can be computed, like the mean, the variance, confidence intervals or the probability to exceed some given threshold. In our situation, we are precisely interested in this last result.

In case of probabilistic-type soft data, the shape of the soft pdf may be of any type. There is no specific requirement about the shape of the soft pdf in the BME framework, except that it has to be a valid pdf.

The BME framework is thus used to update the soil classification and redraw the intervention map. Re-estimating the contaminant concentration at every grid node at each phase may be time consuming and partially useless as the pdf's of grid nodes located far from the samples will probably not be modified by those latter. In order to reduce the computation time without loss of efficiency, only the grid nodes located at a distance less than a given threshold distance are updated. The threshold distance may be chosen in accordance with the range of the variogram or after empirical assessment of the pdf changes as a function of the distance from the new soft datum.

3.3 Objective function

The last step in defining a sampling strategy consists in choosing an objective function. The optimal sampling design is attained when this function reaches its minimum.

As the goal of this multi-phase sampling is to reduce the uncertainty on the soil classification, the objective function has to be chosen in accordance. Consequently, we choose to minimize the integral of the variance surface. This integral is approximated by:

$$I = \sum_{i=1}^{n} \sigma^2(\mathbf{x}_i) \tag{6}$$

The multi-phase sampling process is stopped when the last value of I is not at least 1% smaller than at least one of the three previous values. In other words, if four consecutive values of I are almost equal, the process is stopped.

4 A simulated case study

Working on a simulated case study offers some convenient features: (i) there is no limitation in the number of candidate locations for additional sampling. It is only a matter of grid resolution, (ii) the "reality" is known exhaustively and can be used to compute error statistics, and (iii) the variogram model is known and there is thus no interference between variogram inference techniques and the results of the sampling strategy optimization.

4.1 Data generation and production of the first intervention map

A random variable Z is simulated at the nodes of a 50 by 50 nodes grid using a Cholesky decomposition method. The internodes distance is set to unity. The spa-

tial structure is characterized by an exponential variogram model with unit sill and a range equal to 30.

Assuming we have no suspicion about the location of the contaminated area(s), we select a subset grid of 49 nodes out of the 2500 and further considered them as the initial data set (Fig. 1b). The entire grid is further considered as our "reality" or reference (Fig. 1a). On this map, the contaminated area is characterized by Z values greater than a threshold z_c =1.0425.

Using this data set, SGS is implemented to generate 1000 realizations. At each node, the proportion of realizations for which $Z(\mathbf{x}_i) \geq z_c$ is computed. The nodes where this proportion is higher than 80% are classified as contaminated, those where the proportions is found lower than 20% are classified as safe, and the remaining ones are considered to belong to the uncertain zone. The mapping of this classification gives us our first intervention map (Fig. 1c).

To generate soft data as they are collected during the multi-phase sampling strategy, we use the reference realization. Using the "actual" value at the new sample location as mean value, we generate a Gaussian pdf with known standard deviation σ. Then we randomly draw a value z_i from this pdf and use it along with the standard deviation σ to build a new Gaussian pdf that is considered as our probabilistic type soft datum. This procedure guarantees the "actual" value is a member of this soft pdf without knowing this value. Please note that the Gaussian shape of the soft pdf is absolutely not a requirement of the BME approach. The Gaussian shape was chosen here for its easiness of construction as it depends only on two parameters.

Beside the first intervention map, a map of the uncertainty (variance) is also drawn (Fig. 1d). As expected, the area with the largest uncertainty corresponds to the area classified as uncertain in Fig. 1c. Additional samples will be collected in this zone using the local maximum selection procedure described above.

4.2 Influence of the standard deviation of the new samples and of the number of samples per phase

Two main parameters may have an influence on the performance of the multi-phase sampling strategy. The first one is the relative imprecision of the new samples. We are interested here in assessing the relation between the performance of the sampling strategy and the accuracy of the samples. The second parameter is the number n of samples collected at each phase. This number may have an influence on the convergence of the process since simultaneously collected samples may interact for reducing the uncertainty at some locations. It is worth mentioning here that while the latter is subject to a choice of the operator, the former is an intrinsic characteristic of the measurement device.

Practically, we consider a suite of cases by letting vary s_{soft}, the standard deviation of the soft pdf. Let us denote by s_{hard} the standard deviation of the 49 hard data values. In Table 1 are listed the studied values of s_{soft}, expressed as a proportion of s_{hard}. In case n°1, $s_{soft} = 0$, i.e. no measurement error is considered and observed values are considered as accurate or hard. In this situation, using BME,

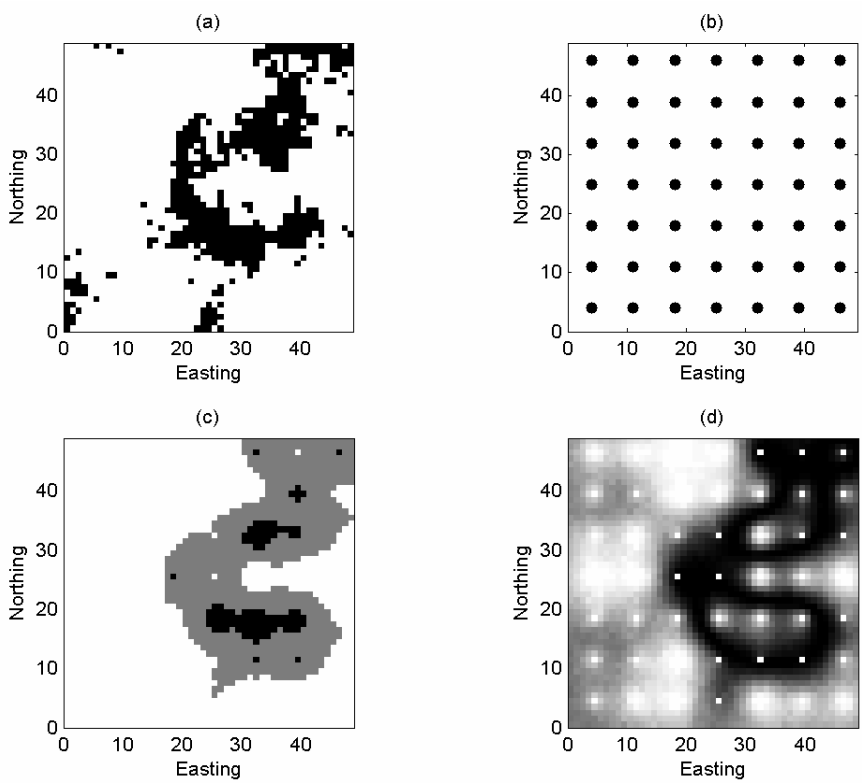


Fig. 1. a) Actual extension of the contamination (black area); **b)** location of the 49 hard data; **c)** First intervention map produced by SGS. The map is subdivided in a safe (white), a contaminated (black) and an uncertain (grey) area; **d)** Variance (uncertainty) map. Dark areas correspond to maximum uncertainty.

Table 1. Values considered for the standard deviation of the soft pdf's. s_{hard} denotes the standard deviation estimated from the 49 hard data values.

Case n°	s_{soft}
1	0
2	$s_{hard}/30$
3	$s_{hard}/20$
4	$s_{hard}/10$
5	$s_{hard}/5$
6	$s_{hard}/3$
7	s_{hard}

kriging or SGS yield the same results as (i) kriging is only a particular case of BME when only hard data are available and all the distributions are Gaussian, and (ii) kriging and SGS theoretically yield equal results for the computation of the probability to exceed a threshold when the distributions are Gaussian. Then, from case n°2 to 7, s_{soft} increases from $s_{hard}/30$ to s_{hard}. In this last case, the standard deviation of the soft pdf's is thus equal to that of the hard data. We thus may suppose the soft data are very uninformative.

To assess the influence of the number n of new samples collected at each phase, we choose for n the values 1, 5 and 30. We so will run the multi-phase sampling strategy for 21 cases, i.e. 7 values for s_{soft} times 3 values for n.

In order to compare the different cases, we use two criteria. Firstly, we plot the relation between the number of samples collected and the safe, contaminated and uncertain volumes. Secondly, we compute the classification errors. Two types of errors may occur: when a node is classified as contaminated when it is actually safe (further referred as Type I error), and the reverse situation (further denoted as Type II error).

5 Results and discussion

In this section, we first analyze the influence of the standard deviation of the new samples and of the number of samples per phase. Then we discuss the classification errors, and finally, we address the question of the financial resources.

5.1 Influence of the standard deviation of the new samples and of the number of samples per phase

The relations between the total number of samples and the safe, contaminated and uncertain volumes are represented in Fig. 2 for n equal to 5 and 30 for the first and second row, respectively. Results for n=1 are very similar to those of n=5 and are thus not shown.

The comparison of the rows in Fig. 2 calls for four comments about the choice of n and the influence of the samples uncertainty.

First, we may think intuitively that collecting one sample at each phase is the most efficient situation because it allows a direct optimization of the positioning of the next sample. The analysis of Fig. 2 shows us the opposite: collecting more samples at each phase allows reaching the minimum of the objective function with a higher total number of samples. The total recovered contaminated volume and the total identified safe volume are also closer to 100%. Note, however, that this remark is valid only if the budget limits are not reached.

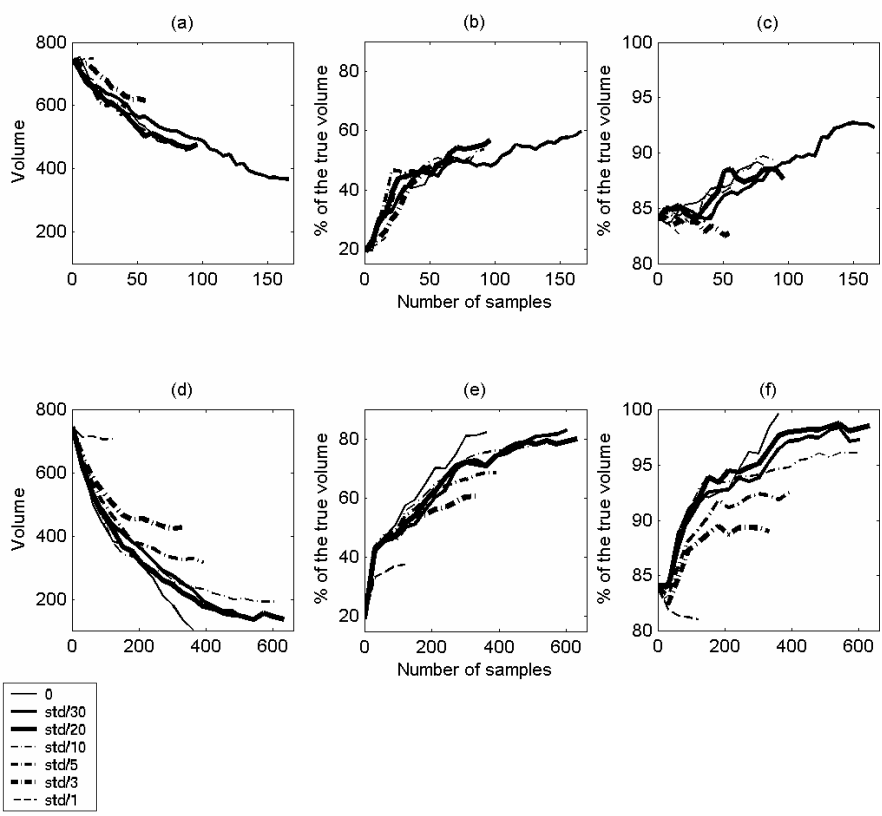


Fig. 2. Evolution of the uncertain (first column), contaminated (second column) and safe (third column) volumes as a function of the number of samples collected. The number n of samples is equal to 5 and 30 for the first and second row, respectively. Each curve corresponds to a given value for s_{soft}. Please note the difference in X scale between rows 1 and 2.

Second, if all graphs are set to the same X scale (Fig. 3a), the curves are almost superposed within each soil class (uncertain, safe or contaminated). This means that the number n of samples collected at each phase has few impact on the evolution of the volumes as a function of n. Consequently, collecting 90 samples with n=1 or with n=30 will approximately give the same results. Again, cost linked to these different strategies may not be equal. Collecting 30 samples at each phase is probably more efficient financially.

Third, the dispersion along the Y axis of the curves seems to increase with the total number of samples. This is particularly clear on the second row graphs of Fig. 2. Beyond 200 samples, using less accurate samples results in significant loss of performance in terms of safe/contaminated volume identification.

Fourth, when $s_{soft} = s_{hard}$, the optimization systematically ends up after only a few phases because the objective function is not decreasing. Using data with such a large standard deviation can be thus considered as of little interest since they are not able to decrease significantly the uncertainty in their neighborhood.

5.2 Analysis of the classification errors

Fig. 3b and 3c represent the classification errors at the end of the sampling process for n equal to 5 and 30, respectively, and for $s_{soft} = 0$ (Graph for n=1 not shown as it is very similar to n=5). From these plots, it appears that the Type II error seems to be very constant whatever the values of n and s_{soft}. This can be explained by the fact that no additional samples are collected in the area classified as safe after the simulations. Uncertainty in this area is small and locations belonging to it are not candidate for the local maximum selection criterion. The zones were Type II classification errors are met, are essentially the small contaminated spots at the southern limit and in the left lower corner of the site.

The Type I error, at the opposite, is considerably reduced in comparison with the first intervention map. However, the decrease is less important as the samples become less accurate. This relation appears to have a steeper slope when n is larger.

Collecting samples with n=30 and $s_{soft} = 0$, $s_{soft} = s_{hard}/30$ or $s_{soft} = s_{hard}/20$ yields the same performance in matter of classification errors.

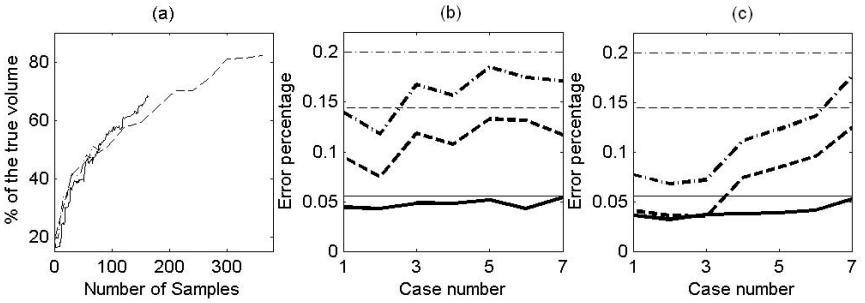


Fig. 3. a) Comparison of the contaminated volumes recovered with n=1 (plain line), 5 (dash-dotted line) and 30 (dashed line) and with $s_{soft} = 0$. **b)** and **c)** Classification errors for n=5 and 30, respectively. The horizontal lines figure out the errors computed from the first investigation map with type I error (dashed line), type II error (plain line) and sum of the errors (dash-dotted line).

5.3 Financial considerations

Financial aspects should also be taken into account. On-site measurement devices are often much cheaper than laboratory analysis. Let us consider a cost ratio of 1/3. This means that 300 samples collected with an on-site measurement technique costs as much as 100 laboratory analyses (Fig. 4). Consider two situations with $s_{soft}=s_{hard}/30$ and $s_{soft}=s_{hard}/3$, respectively. Using these soft samples allows recovering 60% (resp. 71%) of the total contaminated volume instead of 53% with laboratory measurements ($s_{soft}=0$).

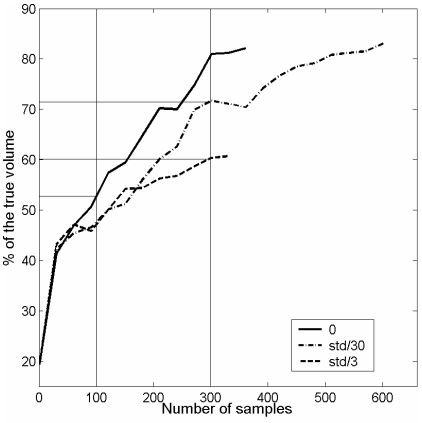


Fig. 4. Comparison in financial terms of the recovered contaminated volumes with n=30 and a cost ratio between hard and soft samples equal to 3.

6 Conclusions

In this paper, we present and illustrate a new multi-phase sampling strategy dedicated to the real-time on-site investigation of contaminated areas. We show that the samples collected using on-site fast and cheap measurement devices can be fully valorized even if they are rather imprecise. In this line, the Bayesian Maximum Entropy approach offers a solution for taking such data into account without loss of information.

Taking more samples at each phase allows recovering a larger contaminated soil volume, identifying a larger safe volume and reducing the uncertain volume to a small fraction of the total volume.

Practical implementation of the multi-phase sampling strategy should be preceded by a thorough comparison of its efficiency with other (single phase) strategies, in particular for the costs and the contaminated soil recovering performance. First results show that cheaper on-site measurement techniques may help recover a larger fraction of the contaminated soil volume at costs comparable to classic laboratory analyses.

Another point for further research concerns the updating of the variogram model using the new samples. The problem, at the moment, is to be able to estimate a variogram from a mixture of hard and soft data.

Finally, such a strategy should be extended to multivariate case and should take advantage of new algorithmic developments for optimizing of the positioning of the samples at each phase.

Acknowledgments

The author acknowledges Hélène Demougeot-Renard and Michel Garcia, his colleagues from FSS International, Patrick Bogaert, from the Université catholique de Louvain, as well as three anonymous reviewers for their useful comments on this paper.

References

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS. Advanced Functions for Field-Based Applications. Springer-Verlag, New York NY

Demougeot-Renard H (2002) De la reconnaissance à la réhabilitation des sols industriels pollués : Estimations géostatistiques pour une optimisation multicritère. Thèse ETHZ n°14615

Demougeot-Renard H, de Fouquet C, Renard Ph (2004) Forecasting the number of soil samples required to reduce remediation cost uncertainty. Journal of Environmental Quality 33:1694-1702

D'Or D (2003) Spatial prediction of soil properties: the Bayesian Maximum Entropy approach. Ph.D. thesis dissertation. Université catholique de Louvain. http://edoc.bib.ucl.ac.be:81/ETD-db/collection/available/BelnUcetd-05012003-113316/

Englund JE, Heravi N (1994) Phased sampling for soil remediation. Environmental and Ecological Statistics 1:247-263

Van Groenigen JW, Siderius W, Stein A (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87 (3-4) : 239-259

Spatio-temporal mapping of sea floor sediment pollution in the North Sea

E. J. Pebesma[1] and R. N. M. Duin[2]

[1]Department of Physical Geography, Geosciences Faculty, Utrecht University, Utrecht, e-mail: e.pebesma@geog.uu.nl
2National Institute for Coastal and Marine Management/RIKZ, The Hague, The Netherlands, e-mail: r.n.m.duin@rikz.rws.minvenw.nl

1 Introduction

Sea floor sediment in the Dutch part of the North Sea is polluted by heavy metals and organic compounds. The origin of this pollution is found mainly in (past) industrial activity, as sediments are mainly contributed by the major rivers Rhine, Meuse and Schelde. The Dutch National Institute for Coastal and Marine Management (RIKZ) started monitoring heavy metals in sea floor sediment in 1981, and added various organic compounds in 1986. Previously, Laane *et al.* (1999) qualitatively described temporal trends in these data but they did not address spatial variability in temporal trends in a quantitative way. This study attempts to quantify spatially distributed estimates of temporal trends (changes over time) in a sediment pollution variable, thereby assessing the errors in the trend estimates.

2 Monitoring network data

The sediment data set is collected by the RIKZ during a monitoring programme, aimed at describing spatial and temporal variability in sea floor sediment. Samples were collected using box core samplers, and only the fraction smaller than 63 μm was analyzed for contaminants. Fig. 1 shows the available observations for an organic compound, PCB138 (one particular polychlorinated biphenyl). The monitoring frequency is 5 years, but for some reason at irregular intervals, additional measurements have been made, and the 2001 monitoring round has been done in 2000. The four "main" monitoring years, 1986, 1991, 1996 and 2000 have 45, 42, 49 and 31 measurements respectively. It can be seen that in the first year more emphasis was put to collect near-coast samples, and that in later years data were collected more in an off-coast directed transect. Table 1 gives distribution summaries of the measurement years. These figures suggest a gradual decrease in PCB138 concentration over time.

Although we will try to use data for all years for identifying a model for spatial and temporal variability, for spatial predicting we will only focus on the four main monitoring years because we do not expect that years with very few measurements contribute much.

Table 1. PCB138 (μg/kg dry matter) summary statistics; years marked with a * are the main monitoring years, other years result from additional sampling programs.

Year	1986*	1987	1989	1991*	1993	1996*	2000*	All
Mean	7.29	8.39	4.08	3.70	1.03	1.58	1.27	4.20
Median	6.90	7.50	2.65	3.05	0.775	1.40	0.90	2.85
Max	21.1	19.7	12.3	13.1	2.70	4.90	3.30	21.1
Min	1.60	2.10	1.00	0.70	0.25	0.20	0.20	0.20
Nr. obs	45	29	14	42	6	49	31	216

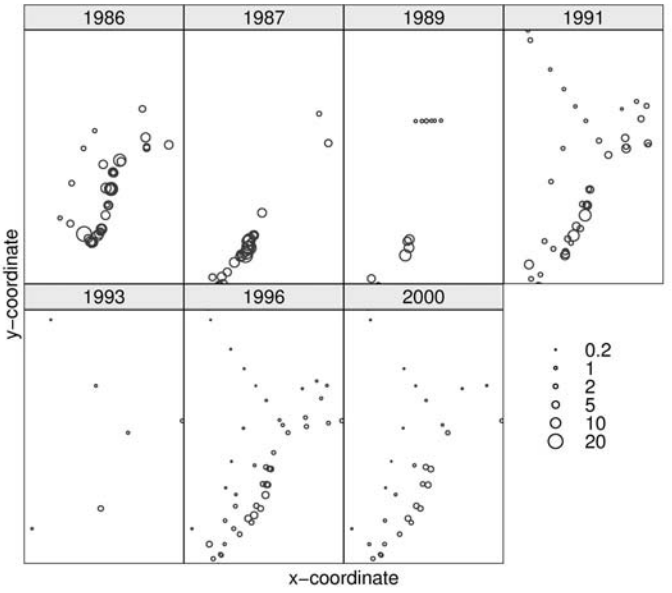


Fig. 1. Maps with PCB138 measurements (μg/kg dry matter) for each monitoring year. The unsampled white area in the south-east corner of the maps is the mainland of the Netherlands. In the area shown, the x-coordinate ranges from 464000 m. to 739000 m., the y-coordinate ranges from 5696500 m. to 6131500 m., projection UTM31

3 Exhaustive information

The spatial pattern of PCB138 measurements (Fig. 1) shows a decreasing PCB138 concentration with increasing distance from the coast. Although the summary statistics of Table 1 indicate that the PCB138 concentrations decrease over time, Fig. 1 suggests that observations that are further off-shore have lower PCB138 concentrations, and that these off-shore observations form a larger part of the samples in later years. Therefore, the temporal trend may be at least *partially* attributed to the increase of the fraction of offshore sampling points over time. To further investigate this, Fig. 2 shows how concentrations depend on water depth. Here, PCB138 was graphed on a log-scale to linearize the relationship. The strong relation does not come as a surprise: most of the polluted sediment originates from the major rivers contributing sediment to the North sea, and the sediment is transported along-coast North-bound by governing sea flows. Sea depth is available as an exhaustive variable, and therefore may help predict PCB138 at unobserved locations.

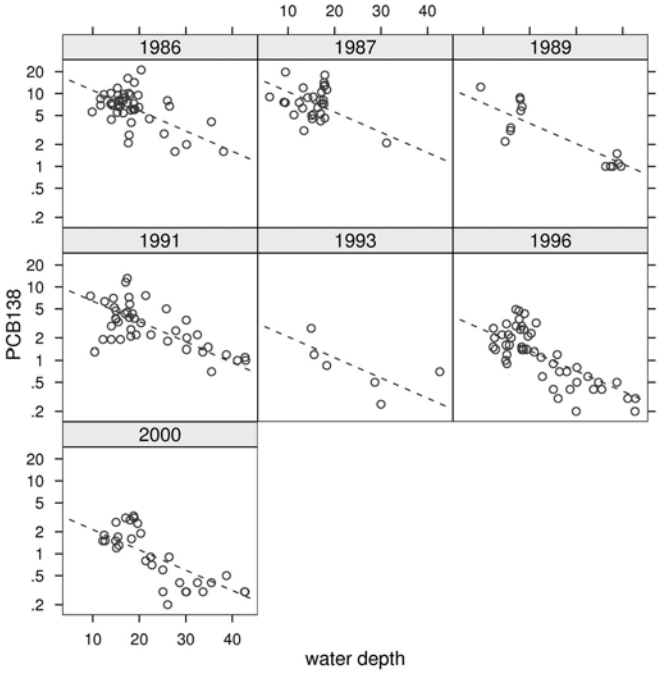


Fig. 2. PCB levels (µg/kg dry matter) as a function of sea depth; fitted models (dashed lines) share a common slope

Fig. 2 further supports the hypothesis that, despite the temporal variation in observation locations, PCB138 levels decrease over time. The overall change of log(PCB138) with water depth does not seem to change over time: under independence assumptions the interaction of time and change in (log) PCB138 level

with water depth tests hardly significant, $p=0.06$, and under more realistic spatial dependence conditions any significance would vanish. As is evident from Fig. 2, the mean level (intercept) for the regression line does gradually decrease over time.

Here, water depth is not *the* variable that causes PCB138 to have certain values; it rather hides a complex of transport processes with dynamic sources, convection and dispersion and complex water flow patterns. In absence of knowledge of these processes, water depth serves as a simple proxy for much of this process, one that explains a fair proportion of the variability. Fig. 2 does not give evidence to remove outliers.

4 A spatio-temporal model

Fig. 2 suggests the model for the data:

$$Z(s,t) = m_t + \beta D(s) + e(s,t) \tag{1}$$

with $Z(s,t)$ the log PCB138 at location s, year t; m_t the intercept for year t, which is the expected log-PCB138 concentration when $D(s)$, sea depth at location s, is zero, and $e(s,t)$ a residual. For $e(s,t)$, we may assume spatial and temporal dependence, and given this dependence we will make predictions for $Z(s,t)$ or make inferences about its change over time.

Spatial correlation of $Z(s,t)$ is hard to infer for each year, as Table 1 tells that sample sizes are small. Fig. 2 however suggests that residual variability does not vary considerably over time. For that reason we assumed that residual spatial correlation is constant over time and we calculated the variogram pooled over years

$$\gamma(\tilde{h}) = \frac{1}{2\sum_{t=1}^{7} N_{h(t)}} \sum_{t=1}^{7} \sum_{h=1}^{N_{h(t)}} (e(s,t) - e(s+h,t))^2 \tag{2}$$

with $t = 1,\ldots,7$ corresponding to the available monitoring years (Table 1), and $N_{h(t)}$ the available number of point pairs with separation distance close to $\tilde{h}$ for year t. This variogram only addresses point pairs taken in the same year. Fig. 3 shows this variogram, along with the total number of point pairs used for each estimate, and a fitted exponential model

$$\gamma_m(h) = 0.08(1-\delta(h)) + 0.224(1-\exp(-h/17247)),$$

with $\delta(h) = 1$ if $h = 0$ and $\delta(h) = 0$ if $h > 0$.

Fig. 4 shows for the four main monitoring years the direct and cross variograms. The noise on all variograms immediately confirms the trouble that one would face when modelling each of them individually, or to fit a full linear model of coregionalization (Goulard and Voltz 1992). The fitted models for all direct variograms were set to the model fitted to the pooled variogram of Fig. 3.

Cross variograms are scaled versions of this model, $r_{i,j}\gamma_m(h)$, where $r_{i,j}$ is the point-wise correlation between years i and j. Given estimates for $r_{i,j}$ this model is the simpler intrinsic correlation or proportional covariances model (Chilès and Delfiner 1999).

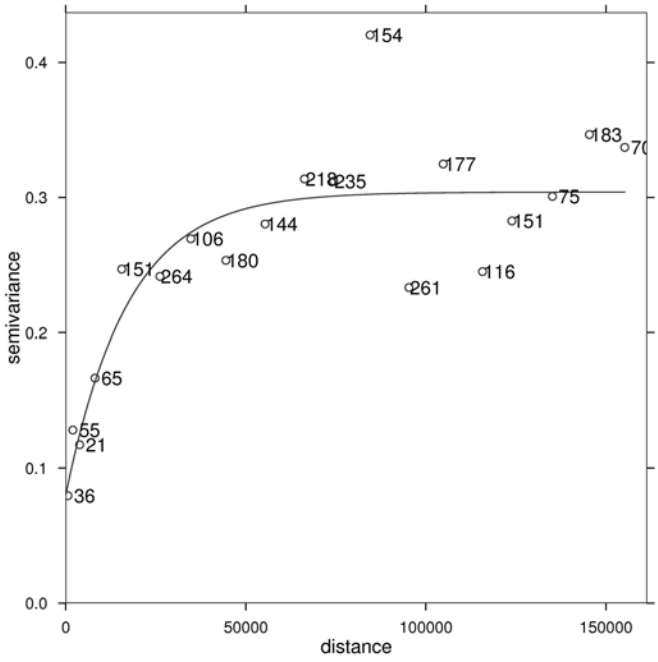


Fig. 3. Pooled, within-year variogram for log(PCB) levels with fitted models (solid line) for all available years; numbers indicate the number of point pairs used

We estimated R, the matrix with $r_{i,j}$ entries, the following way. For each observation point in year i, the spatial nearest neighbour observation in year j is obtained, and correlation is computed. This gave the following, asymmetric correlation matrix estimates: $\widetilde{R}$:

	1986	1991	1996	2000
1986	1.000	0.343	0.651	0.635
1991	0.496	1.000	0.780	0.705
1996	0.615	0.679	1.000	0.920
2000	0.722	0.541	0.925	1.000

Because correlation matrices need to be symmetric, we symmetrized these simple estimates by simply averaging them $R = 0.5(\widetilde{R}^T + \widetilde{R})$ where ^{T} denotes matrix transpose:

	1986	1991	1996	2000
1986	1.000	0.420	0.633	0.678
1991	0.420	1.000	0.730	0.623
1996	0.633	0.730	1.000	0.923
2000	0.678	0.623	0.923	1.000

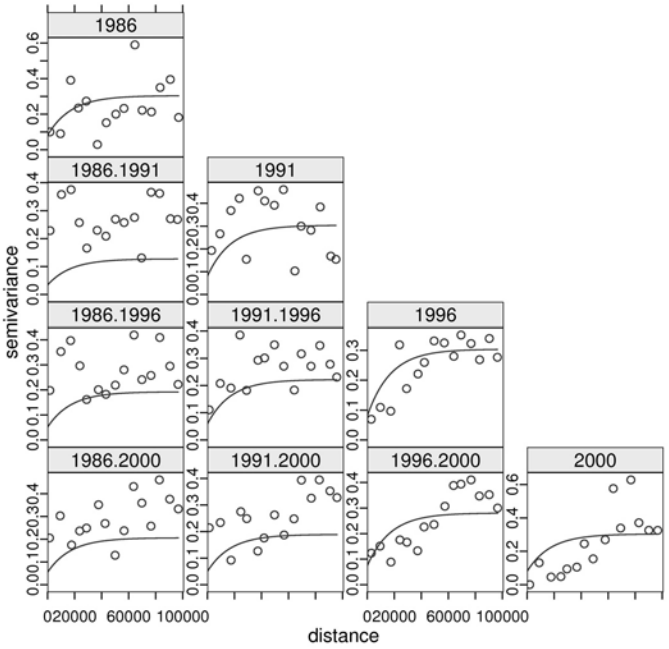


Fig. 4. Direct variograms (diagonal) and cross variograms (off-diagonal) for log(PCB) levels with fitted intrinsic correlation models (solid line); see text for how correlations were estimated

Clearly, Fig. 4 indicates that some cross correlations are not well represented by this model, but the overall agreement is not bad, given the amount of data available. The present approach will underestimate true temporal correlations, and the consequence of this will be discussed later.

5 Spatial prediction

Cokriging predictions for the four main years under model (1), with cross variography shown in Fig. 4, are shown in Fig. 5. The cokriging is a four-variable universal cokriging with sea depth as predictor (or external drift) variable. Each variable has a unique constant (intercept), but the four variables share a common trend coefficient, β in Eq. 1. The sharing of a single coefficient across multiple variables extends the idea of collocated cokriging, by Wackernagel (1998) more precisely coined as collocated ordinary cokriging. Cokriging equations are, amongst others, found in Wackernagel (1998), Pebesma (2004) and Chilès and Delfiner (1999).

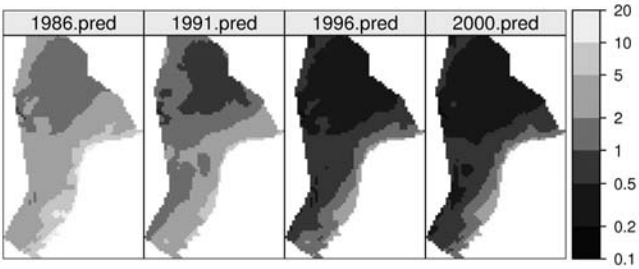


Fig. 5. Cokriging predictions; estimates shown are obtained by taking the exponent of the cokriging prediction on the log-scale

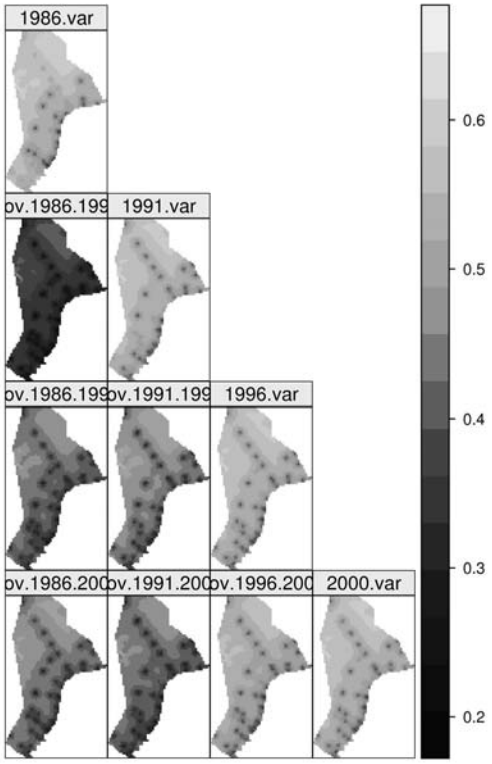


Fig. 6. Cokriging prediction variances (diagonal) and covariances (off-diagonal) for log-PCB138

6 Temporal Change

Cokriging, as formulated e.g. by Ver Hoef and Cressie (1993) or Pebesma (2004) yields not only predictions and prediction errors for each of the four variables, but also prediction error covariances for all pairs of years. Cokriging prediction error variances and covariances are shown in Fig. 6.

Let the prediction vector be $\hat{Z}(s) = (\hat{Z}(s,t_1),\ldots,\hat{Z}(s,t_4))^T$ and let $\Sigma(s)$ be the $4\,\mathrm{x}\,4$ prediction error covariance matrix for $\hat{Z}(s)$, which has prediction error variances on the diagonal and prediction error covariances on the off-diagonal elements. We can now define contrasts as $C(s) = \lambda\hat{Z}(s) = \sum_{i=1}^{4} \lambda_i \hat{Z}(s,t_i)$, and

each contrast has prediction error $\lambda\Sigma(s)\lambda^T$. Simple examples of useful contrasts are e.g.

- a single year, for t_2 we take $\lambda = (0,1,0,0)$
- a difference, for the difference between year 2 and year 4 we take $\lambda = (0,-1,0,1)$
- a mean value, e.g. for the four-year unweighted mean we can take $\lambda = (0.25,0.25,0.25,0.25)$

A simple approach for estimating the gradual change over time is to calculate the contrast that would estimate the regression slope from four years by ordinary least squares. Consider the following regression model:

$$\hat{Z}(s,t) = \alpha_0(s) + \alpha_1(s)t + e(s,t)$$

with $t \in \{1986, 1991, 1996, 2000\}$. The ordinary least squares estimate of $\alpha_1(s)$ is a contrast in $\hat{Z}(s,t)$, with coefficients

1986	1991	1996	2000
-0.0655	-0.0203	0.0248	0.0609

These coefficients are obtained by the usual ordinary least squares equations: if we write the regression model $y = X\beta + e$, with y the response vector and X the design matrix with the predictor variables in its columns, then the coefficient vector is estimated by $(X^T X)^{-1} X^T y$, and the second row of $(X^T X)^{-1} X^T$ contains the contrast coefficients given here.

In Fig. 7 we show the trend estimates, and the trend estimates divided by their standard error $\sqrt{\lambda\Sigma(s)\lambda^T}$.

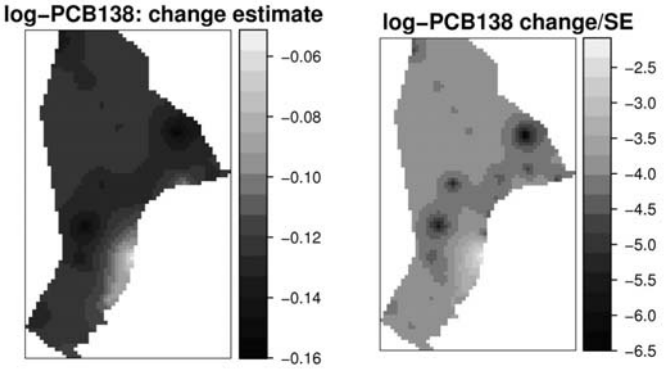


Fig. 7. Left: trend estimates, as yearly change in log-PCB138; **right**: trend estimates divided by their standard errors. The change estimate is for the period 1986-2000

7 Discussion and conclusions

The question posed at the start of this research, "can we estimate spatial time trends from North Sea sediment pollution data", seemed, given the constraints of having less than 170 measurements distributed over the four "main" monitoring years and a strong variation in monitoring design over time, at least to the authors rather unlikely to be positively answered. It turned out however, that very distinct trends could be distinguished (Fig. 7). The main reasons for this must be sought in the data: the temporal change in sediment concentration seems evident (Table 1, Fig. 2) and the spatial variability is largely explained by a strong, temporally persistent linear relation with the proxy variable sea depth, which is exhaustively available (Fig. 2).

The constraint of having less than 50 measurements for each of the "main" monitoring years led to the following constraints:

- the spatial trend (change of log-PCB138 with increasing sea depth) was kept constant over time (see Eq. 1), Fig. 2
- temporal change in the trend was adapted by making the intercept of the trend line time-dependent (see Eq. 1)
- a single, pooled residual variogram (see Eq. 2) was fit to residuals for all years, only addressing residual pairs from identical years
- the intrinsic correlation model was used for cross correlation, leaving only a year-year correlation coefficient to be estimated for each cross correlation
- spatial nearest neighbours were sought to approximate spatial matching observations needed to estimate year-year cross correlations.

The analysis presented here has a number of shortcomings:

- true year-year cross correlations are underestimated by using spatial neighbours instead of spatially matching observations; our approach to approximate these cross correlations does not guarantee positive definite correlation matrices (although we could easily modify the procedure to accomplish this)
- ordinary least squares was used to estimate spatial time trends from correlated predictions, whereas weighted or generalized least squares may have been more appropriate
- uncertainties in variogram coefficients (Diggle *et al.* 1998) were not considered
- cokriging predictions of Fig. 5, estimated on the log-scale, were simply backtranformed by taking exponents; this only yields median-type estimates, and a more elaborate approach (e.g. Diggle *et al.* 1998) could be used to obtain expected values and predicting intervals on the observation scale.

Other approaches to this same data set could possibly address the relation between water depth and sediment pollution in a correlation context, instead of in a regression context. Rivoirard (2002) showed that this leads potentially to a wider class of predictors.

Data, software and acknowledgments

The sea floor sediment data used in this chapter are available from the author's web site. The software used throughout this paper is the R system (Ihaka and Gentleman 1996), an open source implementation of the S language (Becker *et al.* 1988). Within R, we used the gstat package (Pebesma 2004). This package extends the formula/models interface (Chambers and Hastie 1992) of S to multivariable geostatistical models. The models interface takes care of automatic translation of categorical variables into the necessary dummy variables, and allows a simple definition of e.g. interactions or nested effects; in addition gstat provides support for shared (common) trend coefficients across different variables.

The sea floor surface sediment data set and financial support for the development of the gstat S package were gratefully obtained from the Dutch National Institute for Coastal and Marine Management (RIKZ).

References

Becker RA, Chambers JM, Wilks AR (1988) The New S Language. Chapman & Hall, London

Chambers JM, Hastie TJ (1992) Statistical Models in S. Chapman & Hall, London

Chilès JP, Delfiner P (1999) Geostatistics, modeling spatial uncertainty. Wiley, New York

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. Applied Statistics 47(3): 299-350

Goulard M, Voltz M (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Mathematical Geology 24 (3): 269-286

Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5(3): 299-314

Laane RWPM, Sonneveldt HLA, Van der Weyden AJ, Loch JPG, Groeneveld G (1999) Trends in the spatial and temporal distribution of metals (Cd, Cu, Zn and Pb) and organic compounds (PCBs and PAHs) in Dutch coastal zone sediments from 1981 to 1996: a model case study for Cd and PCBs. Journal of Sea Research 41: 1-17

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691

Rivoirard J (2002) On the structural link between variables in kriging with external drift. Mathematical Geology 34: 797-808

Ver Hoef JM, Cressie NAC (1993) Multivariable Spatial Prediction. Mathematical Geology, 25 (2): 219-240

Wackernagel H (1998) Multivariate Geostatistics; an introduction with applications, 2nd edn, Springer, Berlin.

Merging Landsat TM and SPOT-P images with geostatistical stochastic simulation

J. Carvalho[1], J. Delgado-García[2] and H. Caetano[1]

[1] Environmental Group of the Centre for Modelling Petroleum Reservoirs, CMRP/IST, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[2] Dpto. Ingeniería Cartográfica, Geodésica y Fotogrametría. Escuela Politécnica Superior. Universidad de Jaén. c/ Virgen de la Cabeza, 2 – 23071 Jaén, Spain

1 Introduction

In remote sensing, sensor pixels are observed in different portions of the electromagnetic spectrum. They also vary in spatial resolution. Because many bands imply a large volume of data storage, the multispectral sensor does, usually, have a lower-spatial resolution in order to keep an adequate image size. Currently there are many environmental applications (for example, classification, image interpretation, etc.) that require good spectral information (in order to obtain terrain information) with an adequate spatial information (in order to work with medium or large scales). Using appropriate algorithms it is possible to combine these data and produce synthetic imagery with the best characteristics of both – high-spatial and high-spectral resolution. This procedure is a kind of multisensor data merging.

The objective of multi-resolution image merging is to generate synthetic high-spatial resolution multispectral images that attempt to preserve the radiometric characteristics of the original low-spatial resolution multispectral data. Not distorting the radiometric characteristics is important for calibrating purposes and to ensure that targets that are spectrally separable in the original data are still separable in the merged data (Chavez *et al.* 1991). For example, the combination of SPOT panchromatic (SPOT-P) image data, having a spatial resolution of 10m, with Landsat Thematic Mapper (Landsat TM) images, having six spectral bands at 30m resolution, can provide a synthetic image with a good spatial detail, and useful spectral information for identification of small stands of species, which is not possible from either the Landsat or the SPOT images alone.

These merged images have important environmental applications such as in the identification of land cover (global monitoring studies, resource management and planning), in agriculture (health of the crop, extent of infestation or stress damage, or potential yield and soil conditions), forestry (sustainable development, biodiversity, deforestation and reforestation monitoring and managing, biophysical monitoring), cadastre and geology (environmental geology, lithological and structural mapping); applications that need to combine multispectral information with a good spatial resolution that allows one to make maps at adequate scales.

Several methods for spatial enhancement of low-resolution imagery combining high and low-resolution data have been proposed. Some widely used ones are: Intensity-Hue-Saturation (IHS) (Chavez *et al.* 1991), Colour Normalized (CN) (Vrabel 1996), Principal Components Analysis (PCA) (Pohl 1998, Chavez *et al.* 1991) and Brovery transform (Marr 1982). We present a new multisensor image merging technique; to merge low-resolution multispectral images with high-resolution panchromatic images, and to compare the results with classical merging methods. The merging procedures are introduced in section 2. In section 3, an example using the proposed method is presented, and the quality of the images merged with the classical and geostatistical procedures is compared in section 4. The results are discussed in section 5.

2 Merging Procedures

2.1 Classical Merging Procedures

To compare results, two commonly used non-geostatistical image merging procedures were applied, as follows:

2.1.1 IHS Transform

The IHS transform is one of the most common methods of merging images. It consists of two basic steps. In the first step, red, green and blue colour values for three selected TM multispectral bands are converted to hue, saturation and intensity colour components. The intensity component is equivalent to brightness, hue is equivalent to the dominant wavelength of colour, and saturation is colour purity defined as percent whiteness. Mathematical functions are used to convert RGB values to IHS values. The higher-spatial resolution image is constantly stretched in order to adjust the mean and variance to unit intensity. The second step is the substitution of the stretched panchromatic image for the intensity component of IHS and retransformation to RGB.

2.1.2 Colour Normalized

The colour normalized method uses a mathematical combination of the colour image and high-resolution data to merge the higher-spatial and higher-spectral resolution images. Each band in the higher-spectral image is multiplied by a ratio of the higher-resolution data divided by the sum of the colour bands. The function automatically resamples the three-colour bands to the high-resolution pixel size by nearest neighbour, bilinear, or cubic convolution. The output RGB images will have the pixel size of the input high-resolution data.

2.2 Geostatistical Merging Procedure

The objective of the procedure we propose is to create a synthetic image of each Landsat band by a stochastic simulation that integrates the spatial structure present in the high-resolution SPOT-P image and preserves the spectral characteristics of the low-resolution Landsat channel, so that a downscaling of the simulated image produces the original Landsat band.

Geostatistical simulation, in particular, Direct Sequential Cosimulation (Soares 2001), allows us to obtain simulated values of the 10m Landsat image derived from the original 30m Landsat values and the existing correlation between the Landsat and SPOT-P images. It generates several realizations of the original values with a specific pixel size, preserving the basic statistical characteristic of the original images and using information derived from the high-resolution image according to the level of correlation.

Let us consider $TM_i(x_0)$ as the digital value of the original 30mx30m Landsat TM image for the band *i* at the location x_0, $PAN(x_1)$ the value of the original 10mx10m SPOT-P image at the location x_1 and $TM^s_i(x_1)$ the simulated value of Landsat TM image for the band *i* at the 10mx10m grid (SPOT-P grid) at the position x_1 (Fig. 1).

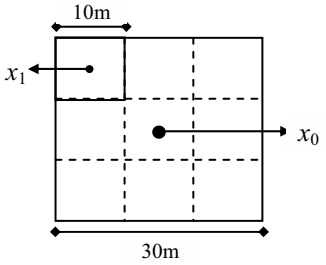


Fig. 1. Landsat and SPOT schematic spatial resolution

With the proposed algorithm the simulated $TM^s_i(x_1)$ image must have the spatial pattern of SPOT-P image, the histogram of band *i* of TM but with a variance corrected for the 10mx10m grid and the same local mean of band *i* of TM; i.e., the mean of 9 pixels $TM^s_i(x_1)$ must be equal to the correspondent value $TM_i(x_0)$.

In short the simulated $TM^s_i(x_1)$ must satisfy the following:

1. For any digital number (DN): $prob\{TM_i'(x)<DN\} = prob\{TM^s_i(x)<DN\}$; where $TM_i'(x)$ is the corrected TM variable for the variance of a 10mx10m grid;
2. $\gamma_{PAN}(h)=\gamma_{TM^s_i}(h)$, where $\gamma_{PAN}(h)$ and $\gamma_{TM^s_i}(h)$ are the variograms of the original SPOT-P and simulated Landsat TM merged image, respectively;
3. Conditioning of the simulated images to the local means of original TM: the mean of the simulated pixels grouped according to the 3x3 pixels scheme must be equal to the 30m Landsat original image values:

$$TM_i(x_0) = \frac{1}{9}\sum_{j=1}^{9} TM_i^s(x_{1,j}) \quad (1)$$

The idea of the proposed algorithm is to use the CoDSS – direct sequential co-simulation – to generate $TM^{s}_{i}(x)$ in a 10m grid using SPOT-P(x) as secondary information. The histogram of band *i* of TM corrected for the variance of 10x10m grid is used for the simulation procedure.

The spatial correlation between primary and secondary variables, Landsat and SPOT-P images (after upscaling to the 30mx30m grid), cannot be considered homogeneous and representative of the entire image. Hence a local model of co-regionalization is applied using the Markov-type approximation (Pereira *et al.* 2000). This means that local correlation coefficients between the two images are calculated (inside local windows), and adopted as the co-regionalization model of the two variables in the cosimulation procedure.

The cosimulation assures the previous conditions 1 and 2, i.e. the corrected histogram of TM and the variogram. But it does not assure condition 3, i.e. the local means of original TM.

To meet condition 3, the proposed merging method is iterative, and can be summarized in the following steps:

1. Generation of high number of images by the direct sequential cosimulation, with Landsat TM as primary information, the high-resolution image (SPOT-P image) as secondary information and the map of local correlation coefficients between Landsat TM and SPOT-P (calculated for local windows, with dimensions dependent on the variogram range);
2. Averaging the simulated images, in cells of 30mx30m (equivalent to 3x3 pixels). For each 30mx30m cell, the mean of the simulated values is compared with the real TM value (condition 3, see Eq. 1);
   Among the several different simulated images, the cell is selected that meets Eq. 1 and, simultaneously, presents the maximum local correlation with the SPOT image, at each location x_0, calculated in the 3x3 pixels window;
3. Rebuild a new secondary image replacing the original SPOT cells by the selected ones – that meet local TM means and have highest correlation with original SPOT values. High local correlation coefficients are assumed for the replaced cells, in order to ”freeze” them in the following simulation steps. Return to step i until all cells meet the local means condition.

Shortcut: if local correlation coefficients between TM and upscaled merged image are sufficiently high, the simulated images are similar to each other. Hence, after step ii, the secondary image practically meets all three above conditions.

The methodology is applied to all the Landsat multispectral bands, except for the TM6, which covers the thermal infrared region of the spectrum.

3 Experimental results

To show the capabilities of the proposed method, the merging procedures presented in section 2 were applied to a test area. The selected area covers a 2400mx2400m area in the Jaén province (South of Spain), with several kinds of land use (urban, olive trees, riverside vegetation, roads, etc).

Fig. 2. Geographical localization of the test area

The data set used for this application comprises a portion of the following images:

1. Landsat TM images. Scene: 20034/95. Date: 08/26/1995. Image size: 80x80 pixels. GSD=30m (TM6 band has not been considered);
2. SPOT-P image. Scene: 35-274/O-P. Date: 06/01/1995. Image size: 240x240 pixels. GSD=10m.

Both images were obtained on similar dates to ensure the merging process quality.

When simulating the Landsat image at a 10mx10m spatial resolution, we treat the Landsat data as the primary variable and the SPOT-P data as the secondary variable. With the CoDSS algorithm the simulated image is conditioned to reproduce the histogram of the Landsat image (corrected for the variance) and the variogram of SPOT, which means that we will obtain an image that has the spectral characteristics of the Landsat and the spatial distribution of the panchromatic SPOT.

The basic statistics of the different images are presented in Table 1.

Table 1. Basic statistics

Image	Mean	Variance	Std. Dev.	Median	Min.	Max.
TM1	99.10	224.68	15.00	97	66	166
TM2	51.65	103.33	10.17	50	26	97
TM3	67.37	214.87	14.66	66	28	131
TM4	74.99	206.18	14.37	74	34	133
TM5	123.56	762.55	27.62	122	45	226
TM7	66.67	281.00	16.77	66	22	139
PAN	140.58	662.78	25.74	137	63	254

The SPOT-P image is considered resampled to 30m pixel size

For all the Landsat bands and the SPOT image the variograms are omnidirectional, and we have fitted to the sample values exponential models. The sills meet the sample variances. The range of the variograms for the Landsat bands is 480m for TM1 and TM4, 420m for TM2, 450m for TM3, and 660m for TM5 and TM7, for the panchromatic SPOT image the variogram presents a range of 330m. The semivariograms and the histograms of the panchromatic SPOT image and, as an example, the Landsat TM4 are shown in Fig. 3.

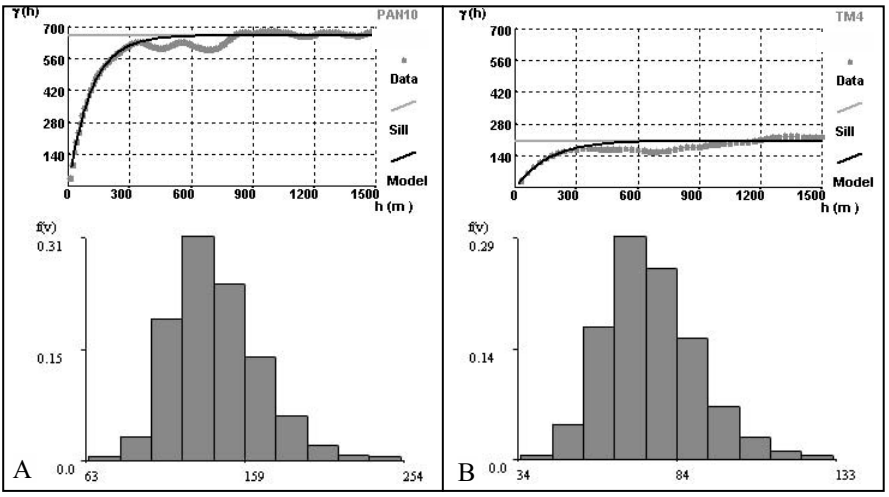


Fig. 3. Semivariograms and histograms of A) SPOT-P and B) Landsat TM4

The information derived from the secondary data, i.e. the high-resolution image, is reflected in the final simulated image according to the correlation between primary and secondary images. In terms of global correlation coefficients, the correlation between the Landsat TM visible and SPOT panchromatic bands is high (0.83), but this value decreases considerably (to about 0.72) for the Landsat infrared bands (see Table 2). This difference between correlation coefficients of a panchromatic image and the visible and non-visible channels of multispectral imagery is always verified.

Table 2. Correlation matrix

	TM1	TM2	TM3	TM4	TM5	TM7	PAN
TM1	1.00	0.96	0.90	0.84	0.85	0.81	0.83
TM2		1.00	0.97	0.90	0.88	0.87	0.83
TM3			1.00	0.91	0.89	0.90	0.82
TM4				1.00	0.88	0.85	0.74
TM5					1.00	0.97	0.72
TM7						1.00	0.70
PAN							1.00

The SPOT-P image is considered resampled to 30m pixel size

Local correlations were computed to account for local differences in Landsat-SPOT data correlation (Fig. 4). To compute the local correlations, a window size of 150mx150m, which is half of the variogram range, was considered the most appropriate. Local correlations range from 0 to 0.981, 0.988, 0.985, 0.954, 0.976 and 0.973 for Landsat TM1, TM2, TM3, TM4, TM5 and TM7, respectively.

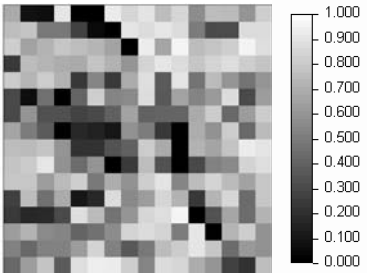


Fig. 4. Local correlation coefficients truncated at 0 between SPOT-P and Landsat TM4 in a 150mx150m radius

Ten thousand simulations were computed for each iteration, until the hybrid Landsat 10mx10m image was complete. The final simulated image was checked for a correct visual appearance.

Inherent to the procedure, the merged images respect the value of the low-resolution data at their locations when they are subjected to a 30mx30m up scaling. The final simulation reproduces the histogram of the Landsat band and the variogram of the SPOT-P (Fig. 5).

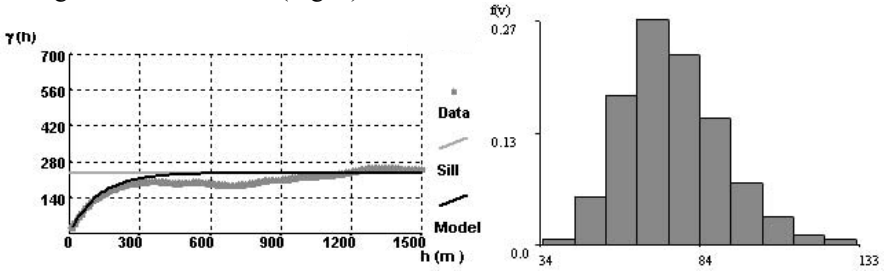


Fig. 5. Variogram and histogram of the TM4 simulated in a 10mx10m grid

4 Comparison between geostatistical method and classical approaches

To demonstrate the potential of the proposed methodology and compare differences, some Landsat bands and the results of the application of the proposed

method and the classical methods to the Landsat TM and SPOT-P images are presented in Fig. 6.

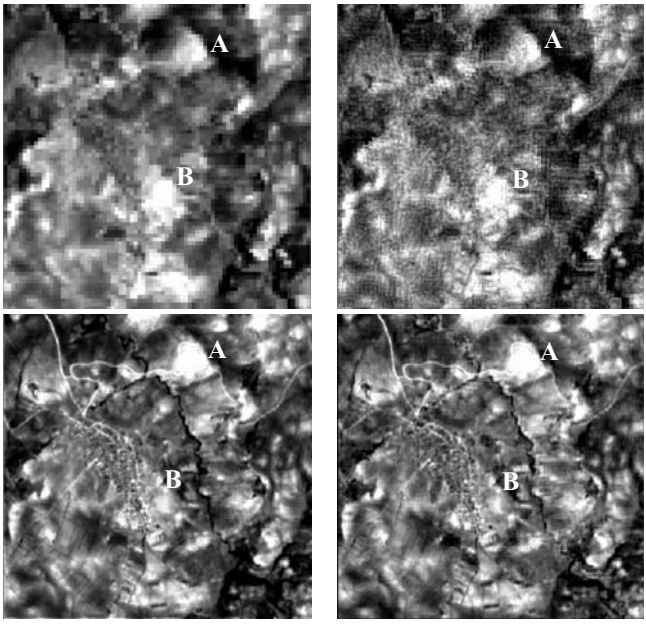


Fig. 6. Landsat TM4; Simulated TM4; IHS 4; CN 4. Linear expansion 2%

First of all, we can evaluate the visual appearance of the merged images (due to restrictions to colour images printing, it was not possible to present colour composites images of the merged bands, with which the results would be more evident).

The main characteristic of the images obtained from the classical methods is that they have a close resemblance with the SPOT-P. This could make the photo interpretation easier, but is an illusory advantage, once the spatial features are all reproduced in the original SPOT-P. Furthermore, these images have a final aspect of softly coloured SPOT-P images, in which the colour tones have been obtained from the Landsat TM ones. This is a clear drawback, because a thematic classification can hardly be done with those digital values.

The geostatistically merged image is more similar to the Landsat TM original images, but with better spatial feature details (derived from SPOT-P). For example, several linear features (roads) that are difficult to distinguish in the Landsat images are visible in these merged images. But more importantly, colours of the different features have been preserved by this method. For example, to highlight the improvement obtained with the geostatistical method, when compared with the classical, notice the white features left of the labels A and B: the dimension of these features is different in TM4 and on the images obtained by IHS and CN, the only method that reproduces their dimensions is the Geostatistical one.

Also interesting is the comparison of the statistical characteristics in Table 3. Here we can see that some basic statistics of Landsat TM are honoured only by the geostatistical merging procedure: the merged bands have an equal mean and variance. We emphasize that this failure of the traditional methods is due to the necessary transformation that is applied previously to the merging process.

The geostatistical merging procedure reproduces the spatial pattern of SPOT-P as they are revealed by the variograms.

Table 3. Basic statistics of Landsat TM and merged images

Image	Band	Mean	Std. Dev.	Min.	Max.
Landsat images	TM3	67.37	14.66	28	131
	TM4	74.99	14.37	34	133
	TM5	123.56	27.62	45	226
GEOSTATS merged images	Geo3	67.37	15.56	28	131
	Geo4	74.99	15.35	34	133
	Geo5	123.59	29.44	45	226
IHS merged images	IHS3	75.14	45.83	0	245
	IHS4	65.41	36.38	0	202
	IHS5	108.27	60.09	0	255
CN merged images	CN3	42.89	9.42	17	91
	CN4	39.42	7.19	20	79
	CN5	64.93	12.20	27	123

The IHS and CN merging methods reduce the mean values (reaching a half of the original values for the CN method). The IHS method increases the variance (up to three times); giving final values higher than the corresponding bands that are merged. In contrast, the CN method produces a decrease of the variance in opposition to the reduction in pixel size. Another important aspect concerns the global correlation coefficients between the different images (bands) used in the merging process. The quality of the spatially enhanced images can also be measured, for each band, by the correlation coefficient between the pixel values of the SPOT-P and the corresponding values of the spatially enhanced images. Column 1 of Table 4 shows the correlation coefficient between SPOT-P and the real Landsat TM image for each band, and columns 2, 3 and 4 the equivalent statistics for the geostatistical procedure, IHS and CN, respectively.

IHS and CN methods produce a very significant increase in the correlation coefficients between the merged bands and the panchromatic one. These coefficients that are around 0.82 (for visible bands) and 0.73 (for infrared bands) in the original images, as mentioned before in section 3, reach values higher than 0.98 for the IHS merging method and 0.91 for the CN method, both methods presenting very similar values of correlation for visible and infrared bands. On the other hand, the proposed geostatistical method preserves the original correlation values (with an increase of about 0.04-0.05 due to the influence of the SPOT-P image in the final merged images) and reproduces the differences in the correlation values for the visible and non-visible channels.

Table 4. Global correlation between SPOT-P and Landsat TM and the merged images

	PAN		PAN		PAN		PAN
TM1	0.83	Geo1	0.84	IHS1	0.99	CN1	0.97
TM2	0.83	Geo2	0.86	IHS2	0.99	CN2	0.99
TM3	0.82	Geo3	0.84	IHS3	0.98	CN3	0.97
TM4	0.74	Geo4	0.78	IHS4	0.98	CN4	0.91
TM5	0.72	Geo5	0.76	IHS5	0.99	CN5	0.99
TM7	0.70	Geo7	0.76	IHS7	0.99	CN7	0.96

The conservation of the correlation coefficients is produced at both global and local levels. Local correlation coefficients can be calculated inside a 150mx150m moving window. In Fig. 7, TM4 vs. SPOT-P local correlation coefficients distributions are shown.

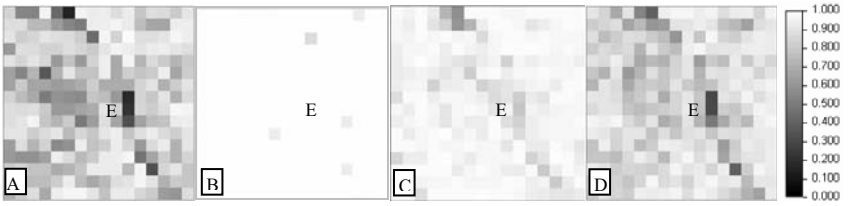


Fig. 7. Local correlation values (considering a 150mx150m window). A: Landsat TM4/SPOT-P; B: IHS4/SPOT-P; C: CN4/SPOT-P; D: Geostat4/SPOT-P

In the original TM/SPOT-P minimum correlation values are around -0.69. This value is related to the presence of riverside vegetation (label E in Fig. 7), which produces large reflectance values in TM4 and small on the visible (panchromatic) bands (see Table 5). This behaviour is preserved only in the geostatistical method that has a minimum correlation coefficient of -0.41, while the other methods always produce positive correlation coefficients.

Table 5. Local correlation statistics

	Mean	Std. Dev.	Min.	Max.	RMS	Abs.max.error
TM4/PAN	0.5663	0.2903	-0.6179	0.9210	–	–
IHS4/PAN	0.9755	0.0284	0.6749	0.9969	0.4975	1.5898
CN4/PAN	0.8963	0.1088	0.1251	0.9939	0.4023	1.2602
Geostat4/PAN	0.6901	0.2178	-0.4061	0.9497	0.1598	0.5063

RMS and Absolute maximum error consider differences between local correlation coefficients of original Landsat TM4/SPOT-P images and the merged images/SPOT-P

5 Discussion and conclusions

This study proposes geostatistical multi-sensor image merging based on the stochastic Direct Sequential Cosimulation (CoDSS) procedure and on local corre-

gionalization models. It shows that this algorithm produces images that, unlike those from other classical merging procedures, preserve the spectral characteristics of the higher-spectral resolution images.

Visual and statistical evaluation of the merged images indicates that IHS and CN change the DN of the images, which means that the spectral features are distorted.

The visual aspect of the geostatistically merged images is different from that of the images obtained with classical methods (these images produce a relevant spatial resolution improvement that makes their interpretation easier), but reveals pertinent spatial features of SPOT-P, honouring the variogram and the statistics of each band.

The geostatistical method takes into account the global and local correlation coefficients between the images in the integration, and those coefficients are preserved in the merged image. This is important when working with non-visible spectral bands, which are poorly correlated with higher spatial resolution images that are usually panchromatic.

Multiscale image merging is usually a trade-off between the spectral information extracted from multispectral images and the spatial information extracted from the high spatial resolution images. Classical merging images have a rich spatial quality (same as SPOT-P) but a poor spectral quality, which makes these transformed images suitable only for visual interpretation, and useless for thematic classification, since the spectral characteristics are distorted. The geostatistical method produces an image with improved spatial resolution (compared with the original Landsat) and, in addition, it preserves the radiometric characteristics of the original high-spectral resolution image.

Most classical methods are not considered really merging methods but substitution methods. They consist of simple substitution of the high-spectral images with a high-spatial resolution image based on the correlation coefficient between the two data sets. The geostatistical method can be considered a true integration of the multisensor data, producing an image that can be upscaled back to the spatial resolution of the lower spatial resolution image with exactly the same radiometric characteristics.

The main drawback of the geostatistical approach is its complexity, requiring understanding and suitable software. The latter must be designed and optimised for processing large sets of data.

References

Chavez PS, Sides SC, Anderson JA (1991) Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT Panchromatic. Photogrammetic Engineering and Remote Sensing, vol 57, n 3, 295-303

Marr D (1982) Vision. Freeman and Company, New York

Pereira MJ, Soares A, Rosario L (2000) Characterization of forest resources with satellite SPOT images by using local models of co-regionalization. Kleingeld WJ and Krige DG (eds), Geostatistics 2000 Cape Town, vol 2, 581-590

Pohl C (1998) Multisensor image fusion in remote sensing, review article. Int. Remote Sensing, vol 19, no 5, 823-854

Soares A (2001) Direct Sequential Simulation and Cosimulation. Mathematical Geology, vol 33, n 8, 911-926

Vrabel J (1996) Multispectral imagery band sharpening study. Photogrammetic Engineering and Remote Sensing, vol 62, n 9, 1075-1083

Characterising local spatial variation in land cover using geostatistical functions and the discrete wavelet transform

C. Lloyd[1], P. Atkinson[2], and P. Aplin[3]

[1]School of Geography, Queen's University, Belfast, BT7 1NN, UK
[2]School of Geography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK
[3]School of Geography, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

1 Introduction

In this paper, texture in an airborne multispectral image from south eastern England is characterised using the local variance, local semivariance and local variogram range. Local variation in these texture measures is then expressed as images and using the global histogram. The images and histogram are then used to question approaches for defining a single spatial resolution based on the mean of statistics such as the local variance. As a second component of the analysis, the discrete wavelet transform (DWT) is applied to the image and the amount of energy in each sub-image is quantified.

Several authors realised in the late 1980s that choice of spatial resolution for remotely sensed imagery should be based on the scale of spatial variation in the property of interest or, more generally, scene of interest. Specifically, Woodcock and Strahler (1987) identified two general classes of interaction between spatial resolution and scale of spatial variation: (i) the L-resolution case in which the variation or objects are not resolved and (ii) the H-resolution case in which the continua or objects are resolved. Depending on the objective (e.g., to produce a thematic map of land cover) and the method of analysis (e.g., hard classification, area proportions prediction) a suitable spatial resolution could be chosen based on the scale(s) of spatial variation in the scene.

The average local variance has been used previously to help select a suitable spatial resolution (Woodcock and Strahler 1987; Jupp *et al*. 1988, 1989). The average local variance $\bar{\sigma}_{vw}^{\,2}$ may be estimated from a moving (3 by 3) window w applied to an image of L rows by M columns of pixels with support v using:

$$\bar{\sigma}_{vw}^{\,2} = \frac{1}{L \cdot M} \sum_{l=1}^{L} \sum_{m=1}^{M} \frac{1}{9} \sum_{j=-1}^{+1} \sum_{k=-1}^{+1} [\bar{z}_v(l+j, m+k) - z_v(l+j, m+k)]^2 \qquad (1)$$

assuming that there is a buffer of one pixel surrounding the image to be analysed. The average local variance $\overline{\sigma}_{vw}^{2}$ is calculated for a range of integer multiples of the original pixel size $|v|$ and expressed as a function of pixel size. The plot usually rises to a peak and thereafter decreases with increasing pixel size. The peak is supposed to help identify the predominant scale of spatial variation in the image. From this information a suitable spatial resolution can be chosen (e.g., a spatial resolution considerably finer than that at which the peak occurs should be sufficient to resolve the variation of interest).

Atkinson and Aplin (2004) suggested that local variation within a remotely sensed image can make choosing a spatial resolution on the basis of an average (see eq. 1) problematic. This paper challenges the assumption that the plot of average local variance against spatial resolution provides meaningful information on which to base a choice of spatial resolution. The hypothesis is that texture measures such as the local variance, local variogram and wavelet coefficients vary so much across a typical remotely sensed image that it is dangerous to choose a single spatial resolution based on the average and that a more sophisticated approach based on the distribution of the statistics is required.

2 Study site and data

The study focuses on the town of St Albans in Hertfordshire, England. Compact airborne spectrographic imager (CASI) imagery with a spatial resolution of 4 m was obtained of the study site (Fig. 1). The study site was chosen as it has a mixture of urban and agricultural landcovers. The spectral wavebands selected match those of the IKONOS satellite sensor (Aplin *et al*. 1997). They are: 0·45–0·52 μm, 0·52–0·6 μm, 0·63–0·69 μm and 0·76–0·9 μm. Here, only two wavebands were analysed: the red (0·63–0·69 μm) and near-infrared (0·76–0·9 μm). More details about the imagery are provided by Atkinson and Aplin (2004).

Fig. 1. Image of the study area.

3 Methods

Three techniques were used: the local variance, local variogram and the DWT. These techniques were used to (i) characterise local variation in image texture and (ii) inform the choice of spatial resolution locally. These are described below.

3.1 Local variance

Equation 1 describes computation of the average local variance. Here, the local variance is computed per pixel, but not averaged. Rather the image of local variances is retained for analysis. Further, the computation is structured such that (i) all spatial resolutions are included (not just integer multiples of the original) and (ii) all possible starting positions of the degraded grid are included. In practice, only odd spatial resolutions (1, 3, 5, …) are included such that an original pixel is always centred on a new larger pixel. Thus, the local variance v. spatial resolution plot can be obtained per original pixel (i.e., an image of plots is obtained). From each plot it is possible to extract the spatial resolution at which the maximum occurs in the discrete sense. This image of spatial resolutions is a new concept proposed in this paper and has not been produced previously. It is the basis for testing the hypothesis set out in the introduction.

3.2 Variogram

Since the local variance is related to, and can be derived from, the variogram (Atkinson 1997, Atkinson and Curran 1997) we also explore local variation in the image using the local variogram. The semivariance is defined as half the expected squared difference between paired Random Functions (RFs). The variogram (or semivariogram) $\gamma(\mathbf{h})$ relates semivariance to lag $\mathbf{h}$, the distance and direction between paired RFs. The experimental variogram can be estimated for the $p(\mathbf{h})$ paired observations, $z(\mathbf{u}_\alpha)$, $z(\mathbf{u}_\alpha + \mathbf{h})$, $\alpha = 1, 2, \ldots p(\mathbf{h})$ with a support v:

$$\hat{\gamma}_v(\mathbf{h}) = \frac{1}{2p(\mathbf{h})} \sum_{\alpha=1}^{p(\mathbf{h})} \{z_v(\mathbf{u}_\alpha) - z_v(\mathbf{u}_\alpha + \mathbf{h})\}^2 \quad (2)$$

Webster and Oliver (1992) state that to estimate a variogram at least 100 observations are needed. According to Webster and Oliver, a variogram based on 150 observations might be satisfactory, while using 225 observations is usually reliable. Attempts to use smaller data sets, in some cases as few as 30 observations, may produce reasonable results (e.g., Goovaerts, 1999). In this paper, the variogram is estimated for a moving window.

The range, a, of a global variogram model may be used to identify the 'characteristic scale' of variation in an image. In this paper, the variogram is estimated for a moving window and the range is estimated for each window position. The method of Ramstein and Raffy (1989) was used to estimate the range of the exponential model locally. Given the theoretical value of the sill variance, C:

$$C = \lim_{h\to\infty} \gamma(h) = \lim_{h\to\infty} \frac{1}{2}\left[\underset{\mathbf{x}\in U}{\text{mean}}(z(\mathbf{x}+\mathbf{h}) - z(\mathbf{x}))^2\right] = \underset{\mathbf{x}\in U}{\text{mean}}\, z^2(\mathbf{x}) \tag{3}$$

a can be estimated with:

$$\hat{a} = -1/[\log(1-\gamma(1)/\underset{\mathbf{x}\in U}{\text{mean}}\, z(\mathbf{x}))^2)] \tag{4}$$

where *U* refers to the window. In addition, a weighted least squares (WLS) procedure was employed to fit variogram models locally. With this procedure, initial values of the coefficients of a spherical model are provided and the routine of Boggs et al. (1989) was used to fit the model locally.

3.3 Wavelet analysis

Impressive results have been obtained in recent years with wavelet analysis (see e.g., Hubbard, 1998) and several papers have appeared in which wavelets are compared to geostatistical procedures (e.g., Chen and Blong 2003). Therefore, in this paper we extend our analysis to include local variation in wavelet coefficients. Wavelets are mathematical functions used to split data into different frequency components and each component is analysed with a spatial resolution matched to its scale (Graps 1995). There are many introductions to wavelet transforms in image analysis (e.g., Stollnitz et al. 1996) and remote sensing (e.g. Chan and Peng 2003). The focus here is on the multiresolution analysis approach of Mallat (1989).

3.4 Multiresolution analysis

In multiresolution analysis the wavelet transform of a signal can be conducted using simple digital filters. The wavelet coefficients, c_j (where *j* is the scale index), may be thought of as a filter (Graps 1995). The coefficients are placed in a transformation matrix that is applied to the raw data vector. The coefficients are ordered using two dominant patterns: one works as a smoothing filter while the other extracts information on local variation (detail). The DWT is achieved by successive low-pass filtering and high-pass filtering of the data vector. The outputs of the low-pass (scaling) filter, *h*, are the "smooth" components, c, and the outputs of the high-pass (wavelet) filter, *g*, are the "detail" components, w. The set $c_{j+1}(k)$ can be computed from the set $c_j(k)$:

$$c_{j+1}(k) = \sum_i h(i-2k)c_j(i) \tag{5}$$

$$w_{j+1}(k) = \sum_i g(i-2k)c_j(i) \tag{6}$$

where i specifies the translation and k is the index for the input data (here pixels). These equations state that the wavelet and scaling coefficients at level j+1 are a weighted combination of the coefficients at level j. That is, we start with the finest spatial resolution and the recursion continues until the coarsest level is reached.

3.5 2D wavelet transforms

There are various approaches to applying the DWT to two (or higher) dimensional (2D) data. This section outlines the horizontal and vertical analyses of Mallat (1989). Starck *et al.* (1998) provide a summary. With this approach, the 2D algorithm entails the application of several 1D filters. The steps followed are: 1. convolve the rows of the image with a 1D filter, 2. discard the odd numbered columns (where the left most column is numbered zero), 3. convolve the columns of the resulting signals with another 1D filter and 4. discard the odd numbered rows (where the top row is numbered zero) (Mallat 1989; Castleman 1996). This process is conducted with both the h filter and the g filter. The result is four images; three of these images (gg, gh and hg) represent "detail" components. Image hh is a smoothed representation of the original image and the filters can be applied to hh in the same way as to the original image leading to four new images: $gg(hh)$, $gh(hh)$, $hg(hh)$ and $hh(hh)$. The filters are then applied to the twice-smoothed image $hh(hh)$ and so on. The scaling function at spatial resolution j+1 is obtained from that at spatial resolution j with:

$$c_{j+1}(k_x, k_y) = \sum_{l_x} \sum_{l_y} h(l_x - 2k_x) h(l_y - 2k_y) c_j(l_x, l_y) \tag{7}$$

and the "detail" components are obtained with (Starck et al., 1998):

$$w^1_{j+1}(k_x, k_y) = \sum_{l_x} \sum_{l_y} g(l_x - 2k_x) h(l_y - 2k_y) c_j(l_x, l_y) \tag{8}$$

$$w^2_{j+1}(k_x, k_y) = \sum_{l_x} \sum_{l_y} h(l_x - 2k_x) g(l_y - 2k_y) c_j(l_x, l_y) \tag{9}$$

$$w^3_{j+1}(k_x, k_y) = \sum_{l_x} \sum_{l_y} g(l_x - 2k_x) g(l_y - 2k_y) c_j(l_x, l_y) \tag{10}$$

Where l_x, l_y are the scaling coefficient indices and k_x, k_y are the location indices for scale $j+1$.

3.6 Wavelet families

The wide range of existing basis functions enables the user to select one that is well suited to the task in hand. Choice of a basis function represents a trade-off between how compactly the basis functions are localised in space and the degree of

smoothness (Graps 1995). Widely used basis functions (that is, wavelet families) include the Daubechies (Daubechies 1988) Mother functions.

3.7 Quantifying energy

The sub-images obtained through applying a DWT can be used to explore spatial variation in the input image. One summary of these sub-images is the wavelet energy signature — the sum of squares of the three directional sub-images at a given spatial resolution (Van de Wouwer *et al.* 1999; Chen and Blong 2003). By assessing changes in energy with change in spatial resolution, an appropriate spatial resolution can be identified. The amount of variation resolved by the wavelet at different scales can be quantified and this information used to determine the amount of variation lost as the image is degraded. In this paper, the amount of energy contained in the sub-images at various spatial resolutions is explored.

4 Analysis

The local variance and semivariance were estimated from the data with a 3 by 3 pixel moving window. Only the images for the red waveband are shown (Fig. 2) due to limitations of space.

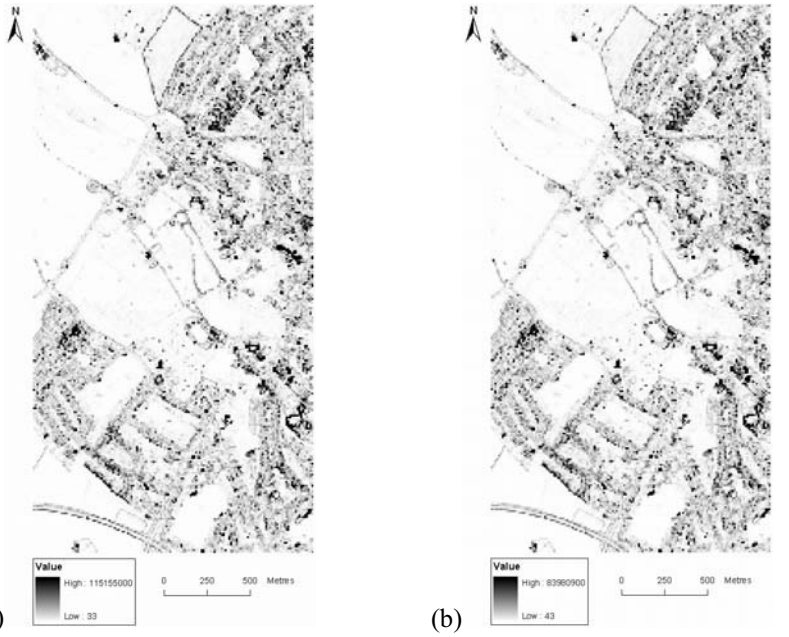
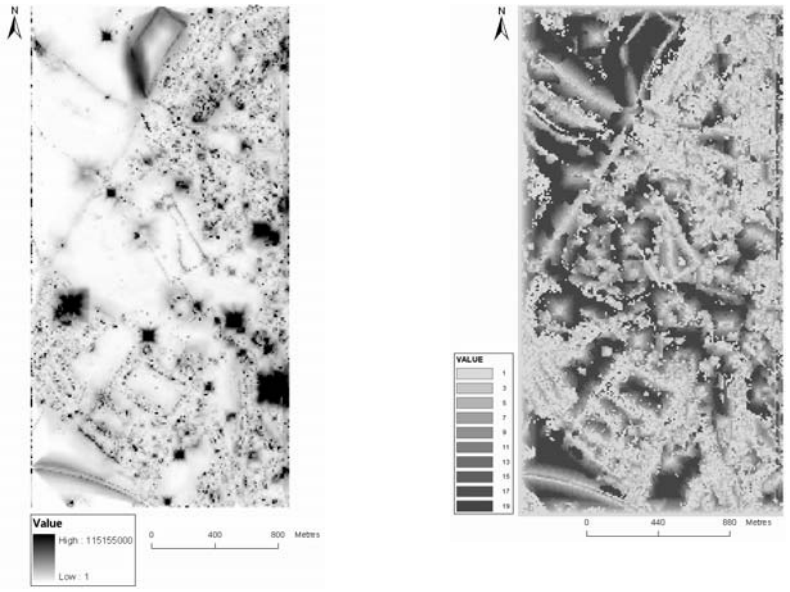


Fig. 2. (a) Local variance and (b) semivariance (lag of 1 pixel) for red waveband.

The original image was smoothed using square filters of 3, 5, 7, 9, 11, 13, 15, 17 and 19 pixels on a side. This created 10 images, each at a different spatial resolution, but each with the same number of pixels as the original image. Each "degraded" image was then filtered to obtain an image of local variance at that spatial resolution. From the set of 10 images of local variance, each at a different spatial resolution, it was possible to obtain a plot of local variance against spatial resolution *per-original pixel*. Subsequently, the maximum local variance and the spatial resolution at which that maximum occurred were extracted and are shown as images in Figure 3. The histogram of all spatial resolutions is shown in Figure 4. Both the images (Fig. 3a: maximum variance; Fig. 3b: corresponding spatial resolution) and histogram (Fig. 4) demonstrate clearly that a single spatial resolution chosen based on the mean local variance across an image (Eq. 1) will be appropriate for only a very small proportion of the original scene. In particular, note that the histogram is multi-modal. Following this, the variogram was estimated within an 11 by 11 pixel moving window. This window size was considered large enough to estimate the variogram robustly, although comparison of larger window sizes would be a useful component of future work. The mapped range estimated using the approximation of Ramstein and Raffy (1989) is given in Fig. 5 while the range estimated using WLS (spherical model) is shown in Fig. 6.



(a) (b)

Fig. 3. (a) Maximum variance and (b) spatial resolution corresponding to maximum variance.

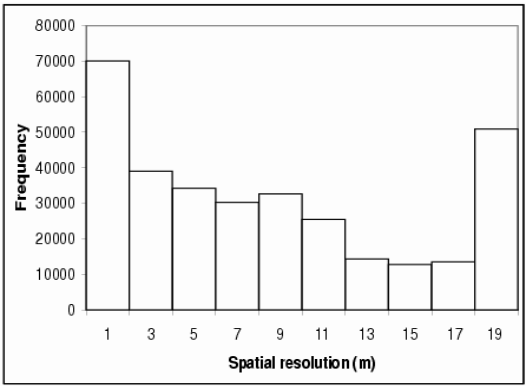


Fig. 4. Histogram of spatial resolutions corresponding to maximum local variance.

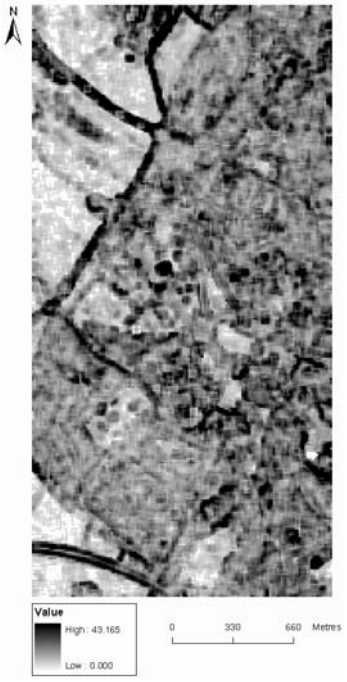


Fig. 5. Variogram range (Ramstein and Raffy) for an 11 by 11 pixel moving window for red waveband.

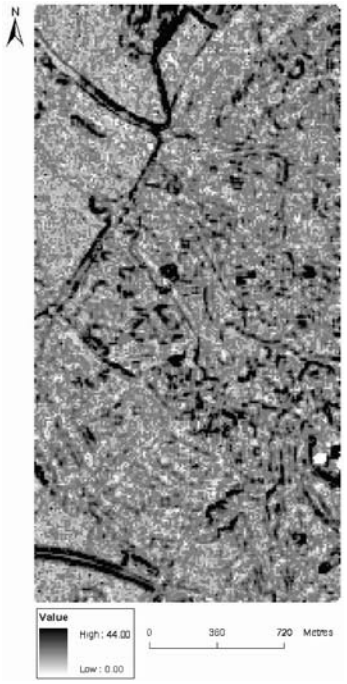


Fig. 6. Variogram range (WLS) for an 11 by 11 pixel moving window for red waveband.

The DWT was applied to the imagery to nine levels. That is, nine sets of three sub-images were generated. The S+ Wavelet software (Bruce and Gao, 1996) was employed for the analysis. The Daubechies 4 (D4) wavelet (Daubechies, 1992) was applied in this case. Firstly, the energy contained in each of the subimages

was explored. Fig. 6 indicates the proportion of energy lost as the image was coarsened.

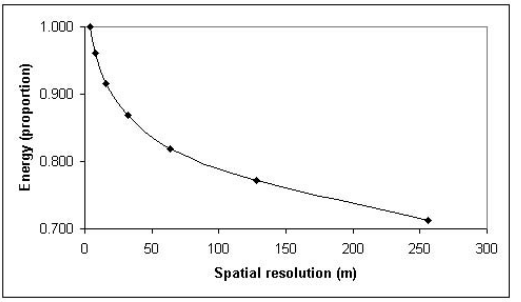


Fig. 7. Reduction in energy as the spatial resolution is coarsened; based on DWT for 7 levels.

One way of visualizing spatial variation in wavelet coefficients is to map significant coefficients only. In fig. 7 the locations of the top 5% of coefficients are indicated for decomposition to one level.

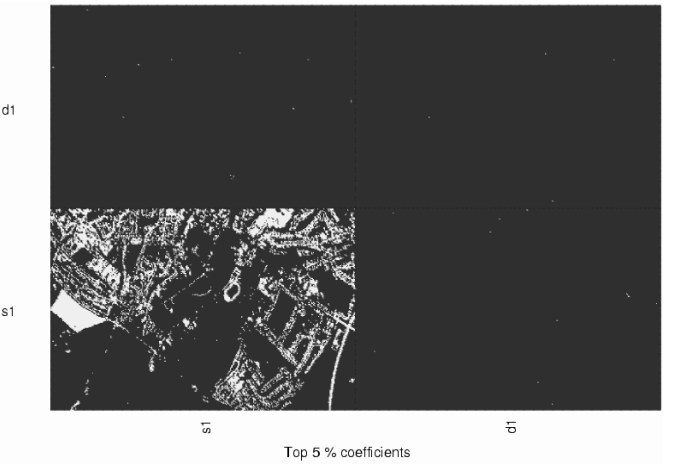


Fig. 8. DWT decompositions to one level: top 5% of coefficients (red waveband).

5 Discussion

As the images of local variance, local semivariance and variogram range indicate, there is marked spatial variation in the values of these statistics. Edges between areas representing different land cover types are demarcated clearly but also different land cover types are represented internally by clearly different local variances and semivariances. The obvious conclusion, as Atkinson and Aplin (2004)

indicate, is that as the frequency of spatial variation varies with land cover type then a range of spatial resolutions may provide as much information as a single fine spatial resolution but with less redundancy. As such, plots of mean local variance against spatial resolution (Woodcock and Strahler 1987) are likely to be of limited value for areas with a range of land cover types.

Interestingly, Figure 3 (especially part b) reveals much that would be hidden in the plot of mean local variance against spatial resolution. Clearly, areas of urban land use are associated with a peak in local variance at a very fine spatial resolution (1 by 1 original pixels). In contrast, agricultural fields have a maximum local variance at a spatial resolution of 19 by 19 original pixels. Some linear features have local variance peaks at a range of spatial resolutions in-between these extremes. Therefore, the image (Fig. 3b) can be used to identify the optimal spatial resolution corresponding to the particular land cover or land use that is of interest. The histogram of spatial resolutions corresponding to maximum local variances (Fig. 4) is multimodal and it illustrates further that a single spatial resolution is unlikely to be appropriate for representing a real world scene. The histogram acts in a similar way to the plot of local variance against spatial resolution, but in the presence of a range of land cover types (corresponding to different spatial frequencies) it provides information about spatial variation in a scene that is not available when only the average local variance is computed. The mapped variogram ranges (Fig. 5) illustrate how the scale of variation changes markedly from place to place and, as such, any approach based on the global variogram may be of limited value if the scene contains a variety of land cover types and, therefore, spatial frequencies.

The plot in Fig. 6 indicates (for DWT decomposition to seven levels) reductions in the proportion of energy as the spatial resolution is coarsened. Such an approach provides a global summary and to be useful, like plots of local variance against spatial resolution, it is necessary to assume that a single spatial resolution is appropriate. In fig. 7, the top 5 % of coefficients are mapped for decomposition to one level. The three detail images (s1-d1, d1-s1, d1-d1) appear almost entirely black: the significant coefficients are visible in the smooth image (s1-s1), but are sparse in the detail images (they appear in the image as white specks). These images indicate further that a single fine spatial variation will result in redundancy. Future work should focus on the pros and cons of methods such as the local variance and variogram range (which could potentially be used to help inform sensor design) and the DWT (which could be applied to compress imagery once it has been acquired). All of the approaches applied in this paper provide information about spatial variation in imagery and more in-depth research may help to reveal how far these approaches duplicate information or compliment one another.

6 Conclusions

This paper supports previous work (for example, Atkinson and Aplin 2004) that argues selection of an optimal spatial resolution for remotely sensed imagery can-

not sensibly be based on plots of average local variance against spatial resolution. Instead, use of images representing local variance and variogram range is recommended.

Identification of an appropriate spatial resolution is based upon assessing how much information is lost before the image becomes unacceptably degraded. The local variance, local semivariance and local variogram range provide useful approaches for characterising texture in successively degraded images and therefore identifying how much variation is lost locally at each stage. The histogram of spatial resolutions corresponding to maximum local variances is a useful tool that could be used to inform any exercise where identifying appropriate spatial resolutions is the objective. As well as assessing information loss with degradation of imagery, the images of local maximum variances and corresponding spatial resolutions or the local range can be used as a basis for segmentation. An additional objective may be to compress imagery rather than to degrade it – possible approaches include retaining only the largest wavelet coefficients (and thus reducing storage space) or the use of quadtrees (Burrough and McDonnell 1998).

In summary, the local variance, local semivariance, local variogram range and the DWT all demonstrate the inherent limitations of approaches for selecting a spatial resolution based on mean statistics or plots representing the entire image. To assess transferability of these methods in different situations, future work should be focused on applying these methods to other data sets.

References

Aplin P, Atkinson PM and Curran PJ (1997) High spatial resolution satellite sensors for the next decade. International Journal of Remote Sensing 18: 3873–3881

Atkinson PM (1997) Selecting the spatial resolution of airborne MSS imagery. International Journal of Remote Sensing 18: 1903–1917

Atkinson PM and Aplin P (2004) Spatial variation in land cover and choice of spatial resolution for remote sensing. International Journal of Remote Sensing 18: 3687–3702

Atkinson PM and Curran PJ (1997) Choosing an appropriate spatial resolution for remote sensing investigations. Photogrammetric Engineering and Remote Sensing 63: 1345–1351

Boggs PT, Donaldson JR, Byrd RH and Schnabel RB (1989) Algorithm 676 ORDPACK: Software for weighted orthogonal distance regression. ACM Transactions on Mathematical Software 15: 348-364

Bruce A and Gao H-Y (1996) Applied Wavelet Analysis with S-Plus. Springer, New York

Burrough PA and McDonnell RA (1998) Principles of Geographical Information Systems. Oxford University Press, Oxford

Castleman KR (1996) Digital Image Processing. Prentice Hall, Upper Saddle River, New Jersey

Chen K and Blong R (2003) Identifying the characteristic scale of scene variation in fine spatial scale resolution imagery with wavelet transform-based sub-image statistics. International Journal of Remote Sensing 24: 1983–1989

Chan AK and Peng C (2003) Wavelets for Sensing Technologies. Artech House: Boston

Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics 41: 909–996

Daubechies I (1992) Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics No. 61. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania

Goovaerts P (1999) Using elevation to aid the geostatistical mapping of rainfall erosivity. Catena 34: 227–242

Graps A (1995) An Introduction to Wavelets. IEEE Computational Sciences and Engineering 2: 50–61

Hubbard BB (1998) The World According to Wavelets. Second Edition. A. K. Peters, Natick, Massachusetts

Jupp DLB, Strahler AH, and Woodcock CE (1988) Autocorrelation and regularization in digital images I. Basic theory. IEEE Transactions on Geoscience and Remote Sensing 26: 463–473

Jupp DLB, Strahler AH, and Woodcock CE (1989) Autocorrelation and regularization in digital images II. Simple image models. IEEE Transactions on Geoscience and Remote Sensing 27: 247–258

Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11: 674–693

Ramstein G and Raffy M (1989) Analysis of the structure of radiometric remotely-sensed images. International Journal of Remote Sensing 10: 1049–1073

Starck JL, Murtagh F and Bijaoui A (1998) Image Processing and Data Analysis: The Multiscale Approach. Cambridge University Press, Cambridge

Stollnitz EJ, DeRose TD and Salesin DH (1996) Wavelets for Computer Graphics: Theory and Applications. Morgan Kaufmann Publishers, San Francisco, California

Van de Wouwer G, Scheunders P and Van Dyck D (1999) Statistical texture characterisation from discrete wavelet representations. IEEE Transactions on Image Processing 8: 592–598

Webster R and Oliver MA (1992) Sample adequately to estimate variograms of soil properties, Journal of Soil Science 43: 177–192

Woodcock CE and AH Strahler (1987) The factor of scale in remote sensing. Remote Sensing of Environment 21: 311–322

Distinguishing features from outliers in automatic Kriging-based filtering of MBES data: a comparative study

P. Bottelier[1], C. Briese[2], N. Hennis[3], R. Lindenbergh[3] and N. Pfeifer[3]

[1] Fugro Intersite BV
[2] Institute of Photogrammetry, Vienna UT
[3] Delft Institute of Earth Observation and Space Systems, Delft UT

1 Introduction

Multi beam echo sounding (MBES) is the state of the art technique of surveying sea floors. A set of sound signals, a ping, is emitted simultaneously, at distinct angles, towards the sea floor. The time it takes for a signal to travel to the sea floor and back is used, together with the angle of emittance, to determine the position and depth of the point of reflection on the sea floor. MBES surveys produce large data sets. Typically, several measurements are available for every square meter in coastal waters. In case of offshore engineering, often real-time processing of the MBES data is required, for example to verify if a pipeline construction turned out successfully. The processing of MBES consists basically of two steps: outliers should be removed and the data density should be decreased, while maintaining a realistic model of the sea bottom. For this purpose a first automatic filtering and thinning algorithm was designed, based on Kriging. Unfortunately, this algorithm had an important drawback: not only blunders, also points representing pipelines were removed by the algorithm.

As the removal of features like pipelines is highly unwanted, methods for improvement were considered. In this paper we discuss these methods and test them on four different data sets of MBES data containing different configurations of pipelines.

The first new method is an extension of the original algorithm. In the original algorithm, soundings are cross-validated in one specific direction, the so-called ping direction. A 1D covariance function is used by the interpolation method Kriging to predict a depth value that is compared with the measured value. If the difference exceeds a certain test value, the measured depth is considered an outlier. Measurements from pipelines perpendicular to the ping direction are easily considered outliers. We show that by using 2D cross-validation this problem can be partially solved. It should be noted that some of the definitions used in the two approaches discussed so far are not considered standard. But as these definitions are used in the implementation causing the problems that initiated this research, we have chosen to present the original definitions rather than standard methods.

An alternative method, also using the Kriging paradigm, was originally designed for filtering laser altimetry data. This type of data, where the time is measured that an emitted light pulse needs for traveling from the laser sensor, mounted on an aircraft, towards the earth and back, is often used to determine a Digital Elevation Model of the bare earth. But laser points are not only reflected by the bare earth but by trees as well. The `tree points' are filtered away as follows. By using a covariance function with a smoothing effect, an average elevation is determined of all available laser points. Now this average elevation divides the measurements in two groups: points below the average are probably bare earth points, points above it are probably tree points. An iterative version of this algorithm turns out to be very effective. We apply this method on our pipeline data. In the multi beam setting we want to filter the `real' outliers/spikes from the sea bottom data including the pipes.

Finally the results of the methods are compared on the data sets, giving satisfying results in most cases.

2 Multi Beam Echo Sounding

Echo sounding is based on the principle that water is an excellent medium for the transmission of sound waves and that part of a sound pulse will return to its source as an echo. If a pulse is emitted from the bottom of the ship at an angle ψ with the vertical line through the emitter, the depth d and the position y of the sea floor hit by this pulse are determined from $(d, y) = ct(\cos\psi, \sin\psi)/2$ where t denotes the time it takes between the initiation of the sound pulse, traveling with velocity c, and reception of the echo, see also Fig. 1.

As illustrated, a swath MBES system (De Jong *et al.* 2002) transmits an acoustic pulse that resembles a fan. Per pulse transmission a high number of depths is thus generated. A ping contains, by definition, all soundings from one pulse transmission. Transversal to the pings are the beams: a beam consists of all soundings with the same emittance angle ψ, therefore a beam contains exactly one sounding of every ping. By combining the depth d and the position y with the position of the ship, determined real-time by GPS (Global Positioning System) and INS (Inertial Navigation System), one obtains coordinate system referenced xyz-data of the sea floor.

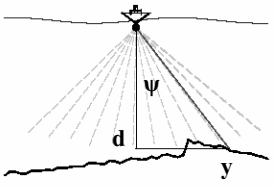


Fig. 1. The multi beam geometry. Shown is one ping of signals

The total list of MBES error sources is extensive. A major error source is wave-induced movement of the ship that can be divided in pitch, roll, heave and heading errors. However, these errors should be eliminated immediately by the INS motion sensors mounted at the ship. Another important error source is the positioning of the ship, done by GPS. The errors we consider however are not the systematic ones, but the real blunders, caused by reflection of the signal on fish or debris, or by occasional electronic errors.

3 Filter methods

All methods we consider are based on the geostatistical interpolation method Kriging (Chilès and Delfiner 1999 and Cressie 1991). Therefore we first recall its basics. The first two filter methods apply Kriging-based cross-validation. In the first case 1D cross-validation is used, in the second 2D. The last, iterative, method uses Kriging to define a smoothed approximation of the sea floor.

3.1 Kriging

Kriging determines weights w_i for the prediction of a depth $\hat{z}_0 = w_1 z_1 + \ldots w_n z_n$ at location p_0, given depth observations $z_1,\ldots,z_n$ at locations $p_1,\ldots,p_n$ and given a covariance function that returns a covariance value as a function of horizontal distance between the observations (isotropic case). First we discuss theoretical and empirical covariance functions, then we show how Kriging uses a covariance function to determine the weights in an optimal way.

The covariance function. The theoretical covariance function or second moment of a stationary random function Z(x) is defined as Cov(s) = E{Z(x)-m}{Z(x+s)-m}, where m=E{Z(x)} denotes the mean or first moment of Z(x). Given some observations a discrete experimental covariance function can be determined by computing experimental covariances between any two observations and by grouping the obtained outcomes according to some distance interval. A continuous covariance function is obtained from the experimental values by fitting them into a covariance model. One can also take a covariance function that is suited to perform some special task. For example, a Gaussian covariance model without nugget effect but with a long range drops relatively slow and therefore has a smoothing effect on the data interpolation.

Ordinary Kriging. Suppose that, as above, we are given height measurements $z_1, \ldots, z_n$ and want to predict a height $\hat{z}_0 = w_1 z_1 + \ldots w_n z_n$ at position (x_0, y_0). Assume moreover that we are given a covariance function cov(.) producing a covariance value C_{ij} between two positions (x_i, y_i) and (x_j, y_j). The ordinary Kriging system consists of $n+1$ equations:

$$w_1 C_{i1} + w_2 C_{i2} + \ldots + w_n C_{in} + \mu = C_{i0} \quad \text{for all } i = 1, \ldots, n$$
$$w_1 + w_2 + \cdots w_n = 1. \tag{3.1}$$

This implies that the weights can be found by:

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} C_{10} \\ \vdots \\ C_{n0} \\ 1 \end{pmatrix}. \tag{3.2}$$

The μ is a so-called Lagrange multiplier and is an extra variable added to make the system solvable. The ordinary Kriging system is obtained within a *random function model*. This means that with every position a random variable is associated. In the case of ordinary Kriging it is assumed that the expected height is independent of the location and that the mean of the heights is unknown. Ordinary Kriging aims at optimizing two parameters and this optimization results in Equations (3.1) and (3.2).

First of all the expected error $r_0 = \hat{z}_0 - z_0$ in the height prediction should be *unbiased*. It can be shown that this condition $E\{r_0\} = 0$ leads to the equation $w_1 + \ldots + w_n = 1$. The other aim is to minimize the error variance $\mathrm{Var}\{r_0\}$. Looking for the *best* solution for the weights under this condition gives the other Ordinary Kriging equations. Moreover, one obtains a formula for the error variance $\sigma^2_{\text{prediction}} = \mathrm{Var}\{r_0\}$:

$$\sigma^2_{\text{prediction}} = \sigma^2 - \sum_{i=1}^{n} w_i C_{i0} - \mu \tag{3.3}$$

As the predicted height $\hat{z}_0 = w_1 z_1 + \cdots + w_n z_n$ forms a *linear* combination of the measured heights, it is by now clear why Ordinary Kriging is often called BLUP, Best Linear Unbiased Prediction.

3.2 1D cross validation

The first method we discuss was developed, see (Bottelier *et al.* 2000), as part of a data thinning algorithm. Given one ping of MBES data, outliers are eliminated in two steps. In a first step, all blunders, defined as soundings above a certain minimal depth and below some maximal depth, are eliminated. These two threshold values are based on a priori depth information on the surveyed area.

The second step is a cross validation step: a depth value $\hat{z}_{(x,y)}$ is predicted for every sounding location (x, y). This prediction $\hat{z}$ is compared to the actual measurement $z_{(x,y)}$. If the difference between the predicted and measured depth, relative to the standard deviation, is tested too big, the sounding is rejected.

Predicting the depth value, 1D case. The prediction of the depth values is done ping-wise by means of Kriging using a covariance function based on the soundings in the ping. For this purpose first an experimental, discrete covariance function dcov*(B_k)* is determined for every ping. A bin width $B = \sum_{i=1}^{n-1} \|p_i - p_{i+1}\| /(n+1)$ is defined as the average horizontal separation distance between consecutive soundings in the ping. The bin B_k consist of all pairs of soundings $\{p_i, p_j\}$ s.t.

$$(k-\tfrac{1}{2})B < \|p_i - p_j\| \le (k+\tfrac{1}{2})B \tag{3.4}$$

The experimental covariance function used in this approach is defined by the following rarely used expression.

$$\begin{aligned} C_0 &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 \\ \mathrm{dcov}(\mathrm{B_k}) &= \frac{1}{2\,|\,B_k\,|}\sum_{\{i,j\}\in B_k}\frac{(z_i-\bar{z})(z_j-\bar{z})}{(z_i-\bar{z})^2+(z_j-\bar{z})^2}\,C_0, \quad k=1..K \end{aligned} \tag{3.5}$$

To decrease the irregular tendency, this empirical covariance function is smoothed with a moving average of five points. From this smoothed function the distance *d* of the first zero crossing, and the correlation length ξ, that is, the distance at which the covariance value is dropped, for the first time, to half of the value at distance zero, is determined. The curvature κ at the origin is defined as κ = (log 0.3149)/ log(ξ/*d*). These parameters, see also (Moritz, 1980), are used to fix the following rarely used analytical hole-effect model, (Kitanidis 1997):

$$\mathrm{cov}(s) = C_0\,(1-f)^{\kappa}\exp(-f^{\kappa}) \quad \text{where} \quad f = s/d \quad \text{and} \quad s,d > 0. \tag{3.6}$$

Using this covariance function, a predicted value $\hat{z}_{(x,y)}$ is determined for each measurement position *(x, y)* in the ping. Here, only the two neighboring soundings on the left and on the right of the sounding to cross validate are used in the interpolation. Note that from the Kriging we obtain $\sigma_{\mathrm{prediction}}$ as well.

Testing the prediction. The last step is to compare the predicted value $\hat{z}_{(x,y)}$ to the measured value $z_{(x,y)}$. This is not done directly, but, again, the variability of the measurements is taken into account. This variability is split in two components. One component is the measurement noise σ_{noise}, that is included to prevent that an observation is marked an outlier just because of random errors. As an indication of the measurement noise, a percentage of 90% of the root of the difference between the first two experimental covariance values is taken, that is $\sigma_{noise} = 0.90\sqrt{C_0 - \mathrm{dcov}(B_1)}$ .The other component is the prediction standard error $\sigma_{\mathrm{prediction}}$ This leads to the following test:

$$\frac{\left|z_{(x,y)} - \hat{z}_{(x,y)}\right|}{\sqrt{\sigma^2_{\mathrm{noise}} + \sigma^2_{\mathrm{prediction}}}} > C_1 \tag{3.7}$$

If the test value C_1 is exceeded, the measurement is considered an outlier and removed. Here it is assumed that the depth data are normally distributed. The test value C_1 is based on a 5% confidence level, yielding a critical level of $C_1 = 1.96$. For normally distributed data this confidence level implies that 5% of the measurements are expected to be tested as outlier.

3.3 2D cross validation

The main difference with the 1D cross validation method is that in the 2D method not only soundings in the current ping are used for determining a covariance function and for the actual cross validation. Instead of one ping, a data set of at least three pings and three beams is considered. Again, first the gross blunders are eliminated by thresholding.

Selecting Neighboring points: The experimental covariance function is determined in basically the same way. All data in the data set, consisting of different beams and pings, are used for determining the experimental covariance function. The same analytic covariance function model is fitted on the experimental covariance data as above.

Since the purpose of this algorithm is to cross validate in two directions, mandatory neighbors in both different pings and different beams are included in the Kriging prediction. Including the mandatory neighbors, a fixed number of e.g. eight closest neighbors is used in the cross validation.

The soundings are tested in a specific order, following the ping and beam indices. If a sounding is considered an outlier it is removed from the list of observations and it will not participate in the testing of the following soundings. It did however participate in the testing of some soundings previous to its own testing. Therefore the tests can not considered to be independent.

3.4 Robust Interpolation

Like the cross validation techniques described in the previous sections, the robust interpolation is a method to filter outliers from a point set describing a surface, see also (Kraus and Pfeifer 1998, Pfeifer *et al.* 2001). One problem with the techniques used so far is that the residuals have quadratic impact on the error variance and ordinary Kriging aims at minimizing the squares of this variance. One way of decreasing the impact of outliers is to minimize another sum of discrepancies function, e.g. a function that is more close to the L_1- norm as in that case the influence of residuals only would increase linearly with their size. Such minimization can be performed by changing the weights of the observations in an iterative way, by giving suspicious observations less and less influence during the iterations. This technique is known in the literature as the `robust approach', (Kraus 1997, Rottensteiner 2001). The residuals analyzed in our case are the differences between the observations and a surface. This surface changes during every iteration

and is obtained by a slightly adapted version of ordinary Kriging that incorporates the residual weights returned by the residual weight function.

The residual weight function. The basic idea of this approach is to give less influence to soundings that are likely to be erroneous. How `good' the *i*-th sounding is considered after the *k*-th iteration, depends on the residual $r_i^k = z_i - \hat{z}_i^k$ between the observation z_i and the estimation $\hat{z}_i^k$ after the *k*-th iteration. That is, the residuals r_i^k are the input of the residual weight function *q(r)* given by

$$\mathrm{q}: \mathbf{R} \rightarrow [0,1], \qquad \mathrm{q}(\mathrm{r}) = \frac{1}{1+(a\,|\,r\,|)^b}\,. \tag{3.8}$$

So, the individual residual weight q_i^{k+1} for the next, the *(k+1)*-th, iteration is determined as $q_i^{k+1} = q(r_i^{k+1})$. If, however, a residual weight q_i^{k+1} is smaller then a certain threshold value $0 \le \varepsilon < 1$, the observation z_i is marked as erroneous and removed from the set of observations. The shape of the residual weight function is described by the half-width *1/a* and the slope *4/ab* at the half-width point *(1/a, q(1/a))*. Before the first estimation step all residual weights are initialized to *1*, that is, $r_i^0 = 1$ for all *i*.

The covariance model. In order for the residual weight function to work properly, the predictions should be on a rather smooth surface that runs on an averaging way between the `real' observations and the `erroneous' observations. Such surface can be obtained by combining two effects. One is to omit the nugget effect in the proximity vector $\boldsymbol{c}$ of covariances with the interpolation position, compare Equation 3.2. The other is to use a covariance model with an almost horizontal slope near the origin and with a not too small range. The covariance function for the area under study follows the Gaussian model, (Wackernagel, 1998), with the same correlation length ξ, see Subsection 3.2 everywhere:

$$\mathrm{cov}(s) = C_0 \exp\!\left(-\left(\frac{s \ln 2}{\xi}\right)^2\right);\ s > 0 \tag{3.9}$$

Here, C_0 denotes the maximum of the covariance function, defined by $C_0 = \sum_i (z_i - z)^2 - \sigma^2$, and *s* the horizontal distance between observations. The measurement accuracy is denoted by σ. To assume isotropy in the data sets considered seems justified by the satisfying results obtained by this method. Note that cov*(s)* is the same in every iteration, for *s>0*. So, the off-diagonal elements of the variance-covariance matrix are given by $C_{ij} = \mathrm{cov}(\,\|\, p_i - p_j \,\|\,)$, where p_i and p_j denote the positions of the observations z_i and z_j.

The interpolation step. The diagonal elements of the variance-covariance matrix however are iteration step, *k*, dependent and are given by $C_{ii}^k = C_0 + \sigma^2 / q_i^k$. They contain a nugget effect σ^2 / q_i^k that incorporates the individual residual weight q_i^k. This ensures that suspicious observations have less influence in the interpolation. The nugget effect is omitted in the proximity vector $\boldsymbol{c} = C_{i0}$ of covariances with the interpolation position, so $C_{i0} = \text{cov}(\,||\, p_i - p_0 \,||\,)$ if $p_i \neq p_0$ and $C_{i0} = C_0$ if $p_i = p_0$. This is a well-known technique to filter short-scale signals and is known as filtering, (Goovaerts 1997). By now all entries of the Ordinary Kriging system of Equation 3.2 are given and the Kriging weights for the next iteration can be determined, resulting in new residuals between the observations and new estimations and thereby in new residual weights. Finally, at every iteration, observations with a residual smaller than the threshold value ε are marked as outlier and removed. This can simply be done by wiping out the row and column corresponding to the outlying observation. The algorithm terminates after a fixed number of iterations, or after all residuals drop below some critical value. If, ideally, all residual weights are close to one, the distances from the observations to the interpolated surface will be in the order of the measurement accuracy σ.

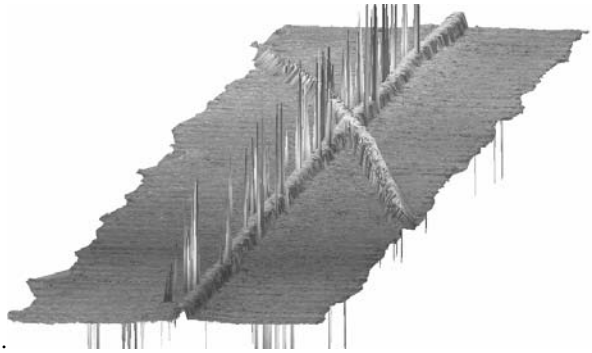


Fig. 2. Data set 1 before processing.

4 Filtering the data sets

For testing the different methods, four data sets are used. We will concentrate on the first however, see Fig. 2., obtained by a multi beam system mounted on a ROV, a Remotely Operated Vehicle, operating at about 15 m above the sea floor. This data set has been acquired using a Reson Seabat 9001 multi beam system. This system has 60 beams and a ping rate of 7 pings a second. The average depth is around 145 meter and the distance between consecutive pings is approximately 0.03 m and the approximate distance between two adjoining beams is 0.1 m.

The second data set has an average depth of approximately 13 meter in the first half and then 23 meter in the second half. The approximate distance between consecutive pings is 0.13 m and between two adjoining beams 0.45 m. This data set does not cover pipelines but represents a short steep sloop.

The third and fourth data set have an average depth of around 300 meters and approximate distance of 0.12-0.14 m between consecutive pings and 0.19-0.24 m between adjoining beams. In both data sets a pipe and some templates are present.

Parameter choices for the different methods: In the 2D method it was possible to vary the number of pings processed at once and to select the number of neighbors. The data presented here were obtained by processing 10 pings at a time while 8 neighbors were used for the cross validation. Although the exact numbers change, similar results were obtained with different parameter choices.

The robust method was run with a maximum of seven iterations. The measurement accuracy was σ=10 cm, the half-width was set on $1/a=2\sigma$, while b=2.

4.1 The results

Table 1. Numbers and percentages of removed outliers by the different methods.

	1D method	2D method	Robust method
MBES 1	104 040	111 476	111 476
# outliers	3 149	1 394	193
% outliers	3.03	1.25	0.17
MBES 2	125 245	130 641	130 641
# outliers	3 853	1 746	1 407
% outliers	3.08	1.34	1.08
MBES 3	297 958	303 014	303 014
# outliers	4 364	3 084	256
% outliers	1.46	1.02	0.08
MBES 4	215 557	219 217	219 217
# outliers	1 955	3 479	1 154
% outliers	0.91	1.59	0.53

Unfortunately it is not very clear what distinguishes good from bad soundings, or, which soundings should be removed. The methods discussed above all divide the soundings objectively in one of these two categories. In most cases it is, subjectively, clear, by a simple visualization, if a wrong decision is made. Often (Teunissen 2000) the following two types of wrong decisions are distinguished: on one hand a Type I error: a sounding is rejected, although it is correct and on the other hand a Type II error: a sounding is accepted, although it is wrong

In Table 1 the numbers of outliers removed are given. Clearly, the 1D method finds a lot of `outliers'. As the 1D method is designed as a data thinning procedure this is an essential part of the algorithm. In the case of data set 1 however, a lot of type I errors are made, due to the presence of the pipelines. This was the reason to

consider alternative methods. In general the 2D method removes fewer soundings, but as we will see later, still type I errors are made. Based on visual inspection, we conclude that the robust method removes the smallest number of points and seems to perform the best, by minimizing both the number of type I and type II errors.

Results of 1D method. In Fig. 3 the results of the 1D algorithm applied on the first data set are shown. On the left a Digital Elevation Model (DEM) of the accepted soundings is given while on the right a top view is given of the accepted soundings in gray and the rejected soundings, in black. The DEM still contains spikes, while the top view shows that a lot of `outliers' were found on the diagonal pipe. The DEM of the data set before processing however, see Fig. 2, shows no outliers on the diagonal pipe. From this visual inspection we conclude that some outliers were not found while many `good' points were rejected by the algorithm.

Results of 2D method. Fig. 4 shows the results of the 2D algorithm on the first data set. In this case no more clear spikes are present in the Digital Elevation Model on the left. The number of soundings marked as outliers is much lower than in the 1D case as can be seen in the top view image on the right. Still many soundings are rejected as outlier situated near the diagonal pipe. We conclude that most of the reported outliers for the 1D and 2D approach are removed unwantedly and should be considered type I errors.

Results of the robust method. In Fig. 5 an overview is given of the accepted soundings, again in gray, and the rejected ones, in black. The digital elevation model of the accepted soundings is not given in this case, as it is similar to the one shown in Fig. 4: no obvious spikes are left after applying the robust method. The robust method only reports a small number of outliers, about one tenth of the number returned by the 2D method, while all spikes seem to be removed. Moreover, only a few `outliers' were found on the diagonal pipe. Still these `outliers' seems to be type I errors, indicating that even the robust method has some problems with the pipes.

5 Conclusions

Both the 1D and 2D method are forced to filter away good points by the 5% confidence level. If the confidence level drops, the number of Type II errors will however increase. Before applying these methods one should have a rough idea on the expected number of outliers. Here, the analysis of Receiver Operating Characteristic Curves could help in future to visualize how the proportion of false alarms increases as the confidence level increases. It should also be considered what the local impact (near features) of a globally set confidence level is.

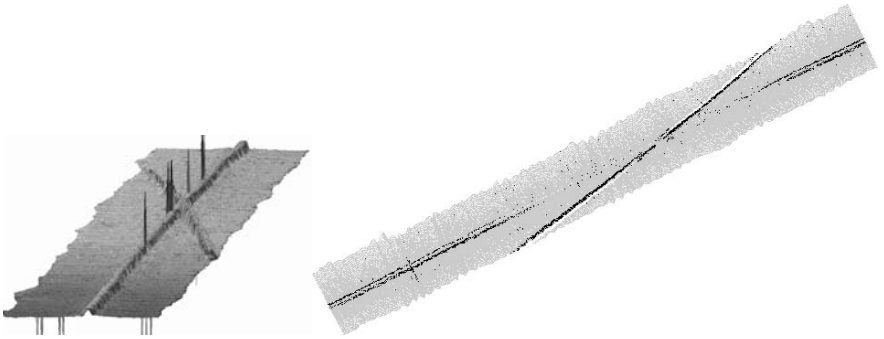


Fig. 3. Data set 1 processed by the 1D method: on the left the DEM of the accepted soundings, on the right a top view of the accepted soundings, gray, and the rejected, black.

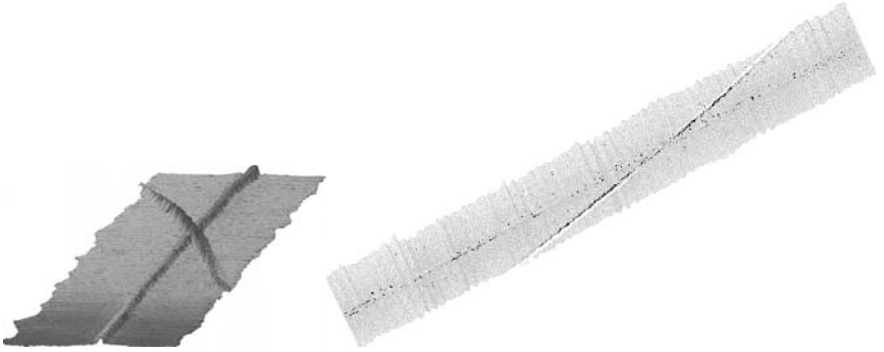


Fig. 4. Data set 1 after processing with the 2D method.

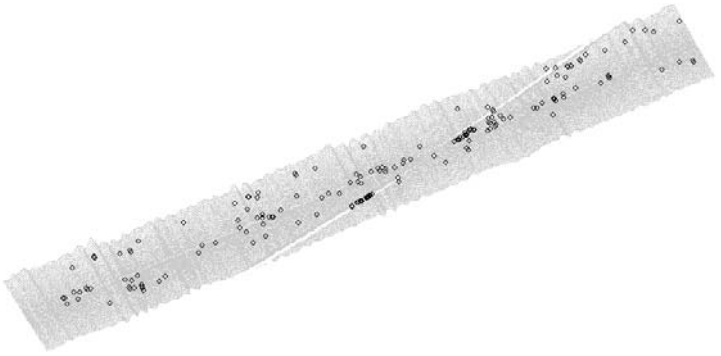


Fig. 5. Accepted and rejected soundings of Data set 1, as found by the Robust method.

In the 2D method, forced including of soundings of different pings is applied. But these included soundings do not get automatically a relevant interpolation weight: for the first two data sets, the ratio of ping width versus beam width is 1:3. Therefore the 2D method is to some extent still a `1D method'.

Both the 2D method and the robust method can be very time inefficient if implemented without consideration. The parameters of the covariance function should preferably be determined in a heuristic way while the number of soundings processed at once should not become too big. The number of soundings included in the Kriging system should be strongly limited while in selecting small number of neighbors one could try to use the structure of the data file.

Analyzing and comparing the results of the different methods is difficult, as it is uncertain which points are truth `ground points' and which points are outliers. A similar comparison on small simulated datasets could be helpful. This would also make it possible to analyze what kind and relative size of features will cause problems for the different methods.

Acknowledgments

The Delft UT Research Theme Earth is thanked for financial support enabling the useful visit of C. Briese to the Delft University of Technology. Fugro Intersite BV and namely C. de Jong are thanked for raising the problem, hosting Natasha Hennis and for the data sets. Finally the reviewers are thanked for their many helpful comments and suggestions.

References

Bottelier P, Haagmans R, Kinneging N (2000) Fast reduction of high density multibeam echosounder data for near real-time applications. The Hydrographic Journal, 98:23-28

Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York

Cressie NAC (1991). Statistics for Spatial Data, Wiley, New York

de Jong CD, Lachapelle G, Skone S, and Elema IA (2002) Hydrography. Delft University Press, Delft

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, Oxford

Kitanidis P (1997) Introduction to Geostatistics. Cambridge University Press. Cambridge.

Kraus K (1997) Photogrammetry - Advanced Methods and Applications. Dümmler, Bonn

Kraus K, Pfeifer N (1998) Determination of terrain models in wooded areas with airborne laser scanner data. ISPRS Journal of Photogrammetry and Remote Sensing 53:193-203

Moritz, H (1980) Advanced Physical Geodesy. Abacus Press, Tunbridge Wells

Pfeifer N, Stadler P, Briese C (2001) Derivation of digital terrain models in the SCOP++ environment. OEEPE Workshop on Airborne Laserscanning and Interferometric SAR for Digital Elevation Models, OEEPE Publication 40, Stockholm

Rottensteiner F (2001) Semi-automatic extraction of buildings based on hybrid adjustment using 3D surface models and management of building data in a TIS. PhD Thesis, Vienna University of Technology

Teunissen PJG (2000) Testing theory; an introduction. Delft University Press, Delft

Wackernagel H (1998) Multivariate Geostatistics, 2nd edition, Springer, Berlin

Forecasting volcanic eruptions using geo-statistical methods

O. Jaquet[1], R. Carniel[2], R.Namar[1] and M. Di Cecca[2]

[1] Colenco Power Engineering Ltd, Täfernstrasse 26, CH-5405 Baden
[2] Dipartimento di Georisorse e Territorio, Università di Udine, Italy

1 Introduction

Active volcanoes can generate severe natural hazards with consequences which are likely to be catastrophic on society. The permanent or temporary increase in population leads to the occupation of additional areas exposed to potential volcanic hazards. Therefore, in these threatened zones, the assessment and mitigation of volcanic hazards require the application of methods allowing forecasts of eruptive events. For many volcanoes, the forecast of volcanic activity can be attempted by monitoring measurable changes in geophysical and geochemical parameters prior to eruptions.

Optimal monitoring at active volcanoes requires automatic recording of many geophysical parameters. The resulting multi-parametric time series are usually data sets which are prone to data losses due to geological hazards likely to occur on active volcanoes. Most statistical techniques, such as Fourier methods, require time series with fairly long data sets without gaps in the observations. In addition, non-oscillatory behaviours are often observed for volcanic time series sampled at active volcanoes. Therefore, the use of analysis methods in the time domain (Jaquet and Carniel 2001 2003), capable of handling incomplete time series, are needed for the interpretation of multi-parametric data sets.

Several authors have proposed forecasting approaches using rises in seismicity (Kilburn and Voight 1998, Kilburn 2003, Ortiz *et al.* 2003) and ground deformation (Voight *et al.* 1998) as well as integrated approaches (Voight *et al.* 2000). These models enable forecasts of eruptive events at short term (days to weeks). Although some successful predictions of eruptions were carried out, these deterministically based models fail to integrate aleatory and epistemic uncertainty (Woo 1999) when forecasting volcanic eruptions. Uncertainty is mainly related to imperfect knowledge of non-linear physical process inherent to volcanic activity and to limited amount of monitoring information. Therefore, a probabilistic formalism is required for the forecasting of volcanic eruptions (Sparks 2003).

In order to account for uncertainty and small data sets (with gaps), a stochastic approach, named DEVIN (Deducing Eruptions of Volcanoes In Near Future), aiming at forecasting volcanic activity was developed within the framework of the EU-project MULTIMO (multi-disciplinary monitoring, modelling and forecasting

of volcanic hazard). DEVIN is a multivariate approach, based on geostatistical concepts (Chilès and Delfiner 1999), which enables characterisation in the time domain of the behaviour for multi-parametric (incomplete) time series sampled at active volcanoes.

Volcanic processes develop at quite a number of different time scales, from degassing regimes alternating on the time scales of minutes (Ripepe *et al.* 2002) to dynamical transitions that separate days-to-months long periods of rather stable activity (Carniel and Di Cecca 1999, Carniel *et al.* 2003). By estimating time scales at which these processes are likely to occur (see also Carniel *et al.* 2004), the DEVIN approach can provide insight into natural processes involved in volcanic eruptions.

The DEVIN approach includes four steps: (1) detection of time correlation by (cross) variogram analysis, (2) parameterisation of time series behaviour, (3) identification of precursors by parameter monitoring and (4) forecasting - with uncertainty - of volcanic activity using stochastic simulation methods. An application of the DEVIN approach using data from the Soufrière Hills volcano, located on the island of Montserrat (West Indies) is given.

2 Detection of time correlation

Occurrences of volcanic activity are often clustered in time; i.e., volcanic events seem not to occur at random, but rather suggest behaviour correlated in time. The variogram is a statistical tool allowing the detection and quantification of time correlation. The variogram, popularized in geostatistics by Matheron (1962), was mainly applied to spatial problems. Jaquet and Carniel (2001; 2003) have shown the capabilities of variogram analysis for time series sampled at active volcanoes.

Variogram and cross variograms allows quantifying the scale at which correlation and cross correlation occurs in the time domain. When this time scale becomes significant, the behaviour of the time series tends to remain similar for that amount of time. This persistent behaviour for a time series expresses the memory of its past activity. For forecasting purposes, characteristics of persistence are needed for time series in order to be considered as potential precursor.

3 Parameterisation of time series behaviour

Once the sample variogram is computed from the data, a variogram model is fitted to the sample variogram in order to parameterise the observed behaviour. Mathematical properties must be fulfilled in order to consider functions as variogram model (Chilès and Delfiner 1999). Among the available models, the following one enables the description of the behaviour for time series sampled at active volcanoes (Jaquet and Carniel 2001):

$$\gamma_M(\tau) = b_0 + b_1\left[\frac{3}{2}\frac{\tau}{a} - \frac{1}{2}\frac{\tau^3}{a^3}\right] \qquad \tau \le a \tag{1}$$

$$\gamma_M(\tau) = b_0 + b_1 \qquad \tau > a$$

where $\gamma_M(\tau)$ is a model composed of a spherical variogram with a discontinuity at the origin. The parameter, b_0, represents the intensity of the random component of the time series. This component is mainly related to variability occurring below the sampling scale and to measurement errors. The parameter, b_1, corresponds to the intensity of the stochastic component for the time series. Finally the parameter, a, is the time scale which quantifies the persistence of the time series.

4 Identification of precursors

Time series with persistent behaviour represent potential precursors. The evaluation of the forecasting capabilities of these time series can be achieved by parameter monitoring. It consists in identifying variogram parameters which time behaviour is likely to be precursory in relation to eruptive events. For a time series, $V(t_\alpha)$, sampled at time t_α ($\alpha = 1, \ldots, N$), parameters presenting potential as precursors of volcanic activity can be estimated using a moving window approach as follows:

$$B^{w_\alpha} = \frac{b_1^{w_\alpha}}{b_0^{w_\alpha} + b_1^{w_\alpha}} \qquad \text{with } w_\alpha = t_\alpha + \frac{L}{2}, \quad \alpha = 1,\ldots,N-L \tag{2}$$

$$G^{w_\alpha} = \int_{L\cdot\Delta t} \gamma^{w_\alpha}(\tau)\, d\tau$$

where w_α is the time for the moving window and $L\cdot\Delta t$ its size. The parameter, B, specifies the relative intensity of the stochastic component. This parameter varies between 0 (random behaviour without memory) and 1 (persistent behaviour with memory). The parameter, G, integrating the total intensity (random and stochastic) and the persistence for the time series delivers a measure of the overall variability at the window scale.

5 Forecasting by stochastic simulation

Parameter monitoring can be applied for forecasting, but no uncertainty can be associated with such forecasts. Therefore, on the basis of precursory behaviour identification, the likelihood for the evolution of the time series is desired at short- to medium term. The realisation of such forecasts (with uncertainty) requires the use of stochastic simulation on basis of potential evolution scenarios. The chosen stochastic simulation method starts from the following decomposition (Chilès and Delfiner 1999):

$$V(t_0) = V^*(t_0) + \left[V(t_0) - V^*(t_0)\right] \quad (3)$$

where $V^*(t_0)$ is the kriging estimator (Wackernagel 1995) at time t_0 using the data $V(t_\alpha)$; and the term $[V(t_0) - V^*(t_0)]$ is the kriging error. Since the true value, $V(t_0)$, is unknown, one considers the same equation expressed in terms of simulation:

$$V^s(t_0) = V^{s*}(t_0) + \left[V^s(t_0) - V^{s*}(t_0)\right] \quad (4)$$

where $Vs(t_0)$ is the simulation of $V(t_0)$ and $Vs^*(t_0)$ is the kriging estimator using only the simulated values at the points t_α and then the kriging error is replaced by its simulation:

$$V^{cs}(t_0) = V^*(t_0) + \left[V^s(t_0) - V^{s*}(t_0)\right] \quad (5)$$

where $V^{cs}(t_0)$ is the conditional simulation. This method allows generating simulations that honour the data points of the time series. This conditioning property is important when performing simulation of the future behaviour of the time series; i.e., the simulation, starting off at the last data point available for the time series, allows integration of the latest characteristics of the data. A dilution method (Lantuéjoul 2002) is applied for the (non conditional) simulation, $V^s(t_0)$. Based on this method, simulations of a Gaussian stochastic process are produced with a spherical variogram. Since the data are usually not Gaussian, simulations matching the observed histogram are obtained using a gaussian (bijective) transformation (Chilès and Delfiner 1999) applied to the time series.

Potential scenarios for the evolution of volcanic activity can be considered using the following expansion:

$$V^{cs}(t_0) = m(t) + V_r^*(t_0) + \left[V_r^s(t_0) - V_r^{s*}(t_0)\right] \quad (6)$$

where $m(t)=E[V(t)]$ is the deterministic drift which form is assumed (arbitrary) polynomial and the V_r correspond to residuals with zero mean. The drift becomes time dependent solely when performing forecasts. This possibility allows the introduction of external knowledge for performing sensitivity studies with respect to various scenarios of volcanic activity.

The likelihood of volcanic activity is assessed by analysing the tendency of time series to exceed given thresholds. Using a Monte Carlo approach allows performing a large number of stochastic simulations in order to estimate probability of threshold exceedance for a given period. These estimates are interpreted as forecasts (with uncertainty) of eruptive scenarios for a given period. These forecasts constitute valuable input as needed for probabilistic risk assessments.

6 Case study

Montserrat (island) is a dependency of the United Kingdom, one of the Leeward Islands of the Lesser Antilles, in the Caribbean Sea. The Lesser Antilles is a volcanic island arc formed along the junction where the Atlantic tectonic plate is sub-

ducted beneath the Caribbean plate. Montserrat is only 16 km long (north - south) and 10 km wide (east - west), and is built almost exclusively of volcanic rocks. The island comprises several volcanic centres or massifs of differing age, among them the active volcano of the Soufrière Hills in the south.

The current eruption of the Soufrière Hills Volcano began in July 1995 and continues to the time of writing (June 2004). The eruption followed a three-year period of precursor seismic activity. The eruption has been characterised by the growth of an andesite lava dome with associated pyroclastic flows, vulcanian explosion and debris flows.

The Soufrière Hills eruption can be divided into several stages (Robertson *et al.* 2000, Sparks and Young 2002). The first stage involved phreatic explosive activity between July and November 1995. The andesite dome appeared in mid-November 1995 and growth continued nearly continuously until March 1998 in the first major stage of dome growth. Our dataset includes part of this dome growth stage from 1 November 1996 to 9 March 1998 (Fig. 1). A stage of dome dormancy occurred between 10 March 1998 and 26 November 1999. A second stage of dome growth then started and then finished on 13 July 2003 when the volcano moved into a second period of dome dormancy accompanied by minor unrest. Our seismic dataset (Fig. 1) also captures most of the second dome growth stage.

Among the complex sequences of events that have occurred during the course of the Soufrière Hills eruption, we have decided to choose the onset of the dome growth in November 1999 as event for forecasting. The choice of this important event was motivated by: (a) correlated effects likely to be expressed in terms of seismic events and (b) the implications of forecasting the onset of the dome growth in relation to probabilistic risk assessments.

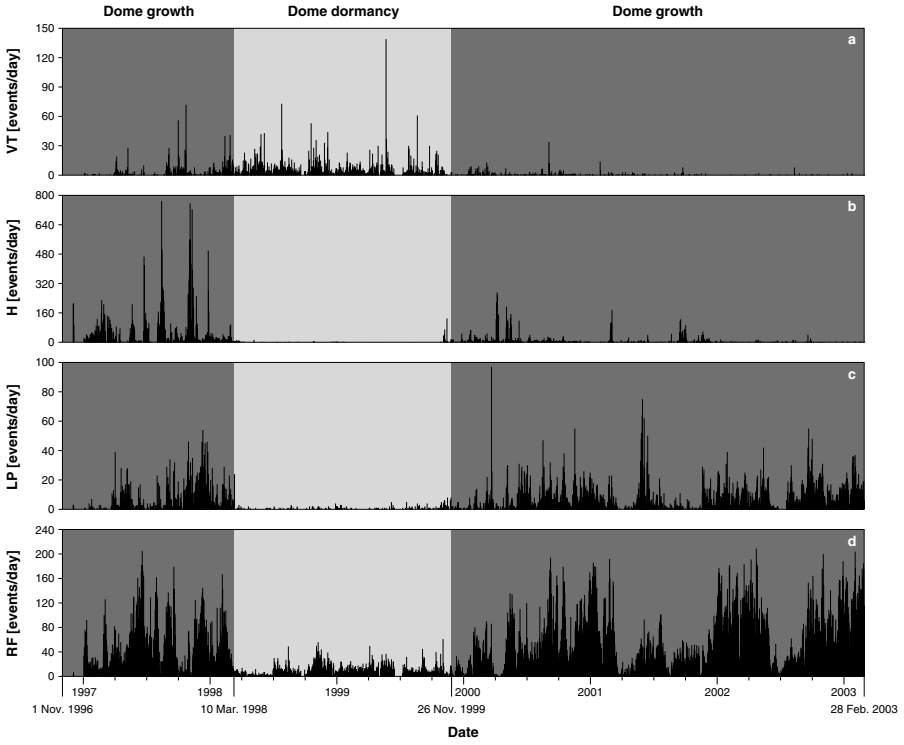


Fig. 1 a) Time series of volcano-tectonic, **b)** hybrid and **c)** long-period earthquakes, and **d)** rockfall signals.

6.1 Seismic data

The MVO (Montserrat Volcano Observatory) digital seismic network was installed in October 1996, and uses a mixture of broadband and short-period seismometers. Using these monitoring instruments, the number and type of earthquakes being produced underneath the volcano during the eruption were recorded. These seismic events were manually classified into the following categories: volcano-tectonic earthquake, long-period earthquake, hybrid earthquake or rockfall signal (Miller *et al.* 1998).

The daily activity was summarized by the total number of events (of each type), and the cumulative energy of those events, expressed as a cumulative magnitude. For this case study, the following time series were selected for the period from 1 November 1996 to 28 February 2003 (Fig. 1):

- Number of volcano-tectonic earthquakes per day (VT).
- Number of hybrid earthquakes per day (H).
- Number of long-period earthquakes per day (LP).

- Number of rockfall signals per day (RF).

as well as:

- Daily cumulative magnitude for volcano-tectonic earthquakes (VT_cm).
- Daily cumulative magnitude for hybrid earthquakes (H_cm).
- Daily cumulative magnitude for long-period earthquakes (LP_cm).

6.2 Variogram analysis

The analysis of correlation and cross correlation in time was performed using respectively the variogram for individual times series and the cross variogram for pairs of time series.

Differences in variability of event occurrences between stages of dome growth and dome dormancy were observed over the six years period (Fig. 1). Therefore, the variograms were calculated by periods of dome growth and dome dormancy for the time series VT, H, LP and RF events (Fig. 2).

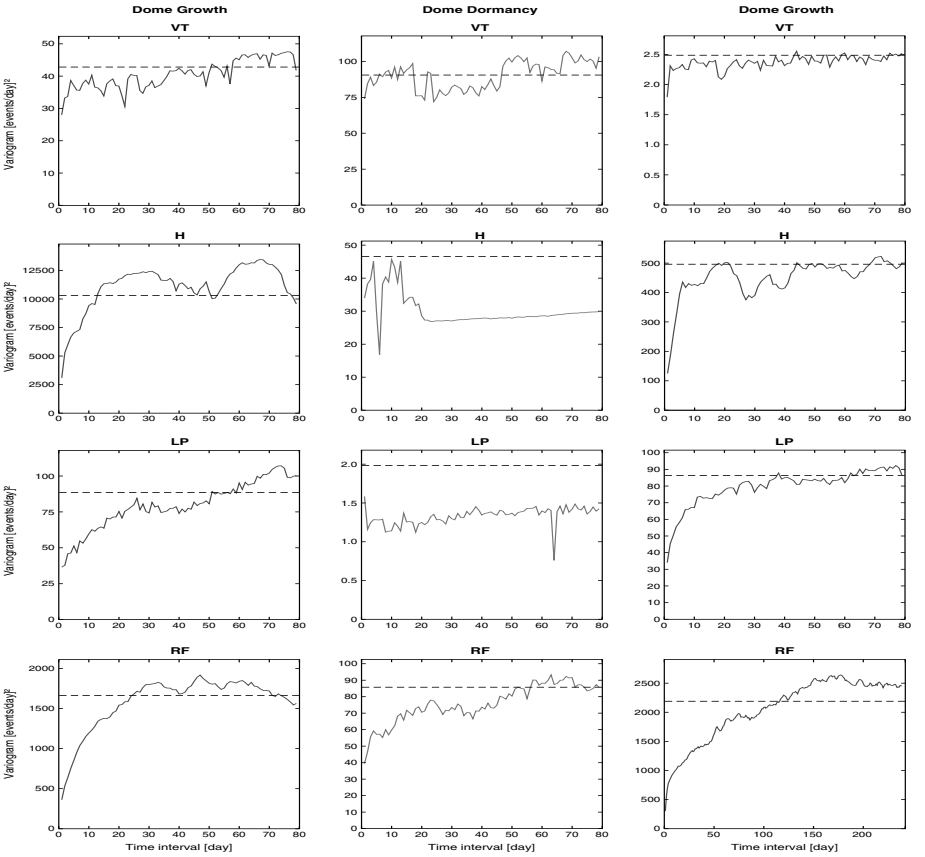


Fig. 2. Variograms for time series VT, H, LP and RF during episodes of dome growth Nov. 96 - Mar. 98 and Nov. 99 - Feb. 03) and dome dormancy (Mar. 98 - Nov. 99).

During stage of dome growth, except for the VT events, the growing behaviour of the variograms for H, LP and RF events were followed by a more stable part fluctuating around the variance of the data (dashed line). For the first stage of dome growth (Nov. 96 to Mar. 98), the time scale corresponding to the first stabilisation level of the variogram (around the variance) was estimated to be between 20 and 40 days for H, LP and RF events.

These time series exhibit a persistent behaviour; i.e., the activity occurring today presents some similarity with the activity of the 3-5 weeks before. Such behaviour was not observed for VT events; its variogram was not showing a significant growing behaviour. For the second stage of dome growth (Nov. 99 to Feb. 03), the variographic analysis has lead to similar results, except for the time scale associated to the RF events which displayed a significant increase and reached a value equal to about 150 days.

For the stage of dome dormancy, no structured behaviour was observed for VT, H and LP events. The only persistent behaviour was displayed by the RF events with a time scale between 30 and 50 days.

For the different dome stages, similar types of persistent behaviours were also detected for the time series of daily cumulative magnitude. The absence of persistent behaviour during dormancy could be related to classification problems for the VT events and to the low level of seismic activity for the H and LP events during that period.

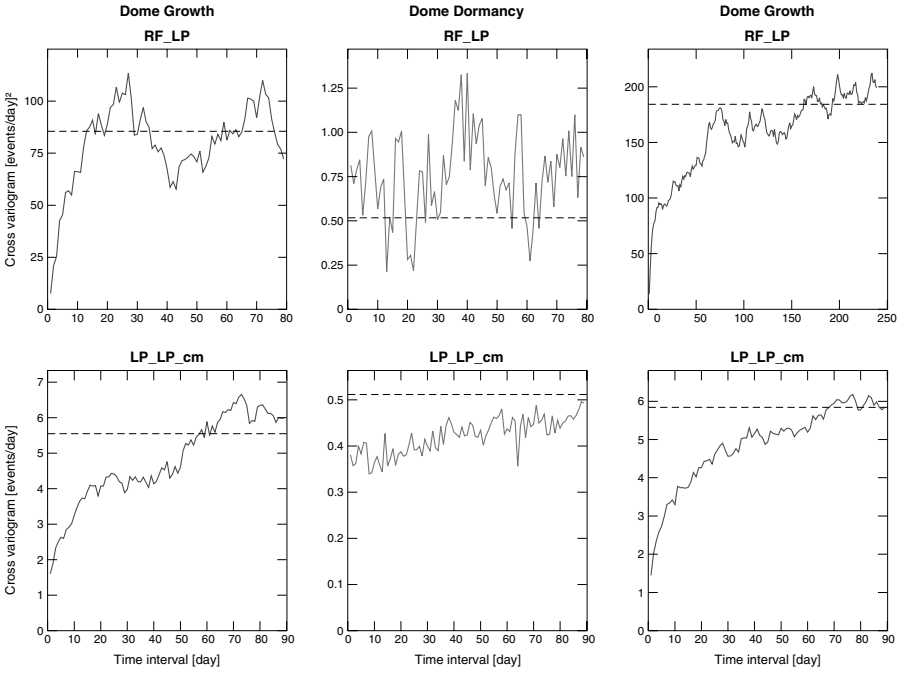


Fig. 3. Cross variograms for time series RF – LP and LP – LP_cm during episodes of dome growth Nov. 96 - Mar. 98 and Nov. 99 - Feb. 03) and dome dormancy (Mar. 98 - Nov. 99).

The calculation of cross variograms by period was performed on the basis of the largest correlation coefficients. The pairs RF – LP and LP – LP_cm events were selected due to their persistent behaviour during periods of dome growth (Fig. 3). The time scales for the first period of dome growth were equal to approximately 20 days for the pairs RF – LP and LP – LP_cm and then these time scales increased to about 70 to 100 days for the second period of dome growth.

6.3 Precursor identification

Time series with persistent behaviour represent potential precursors. Their capability needs however to be evaluated in relation to forecasting specific eruptive

events. Precursor identification was achieved by parameter monitoring of time series prior to the onset of dome growth in November 1999.

Monitoring changes in parameters was performed by (cross) variogram calculation using a moving window approach. Instead of calculating the (cross) variogram for the entire period of dome growth (or dormancy), it was computed by applying a moving window of 100 days along the entire time series. By estimating parameters for each (cross) variogram, the evolution of the behaviour for the time series can be monitored in terms of persistence and intensity.

This approach was applied for the RF events and the time series LP and LP_cm. In both cases, the monitored (cross) variogram parameters display behaviours with higher parameters values for period of dome growths in comparison to the period of dome dormancy. For the RF events, the variogram parameter, B, specifying the relative intensity of the stochastic component varies between 0 and 1. For the time series LP and LP_cm, the cross variogram parameter, G, is obtained by integration of the variogram at the window scale. This parameter delivers a measure of the overall time variability at the window scale between two time series.

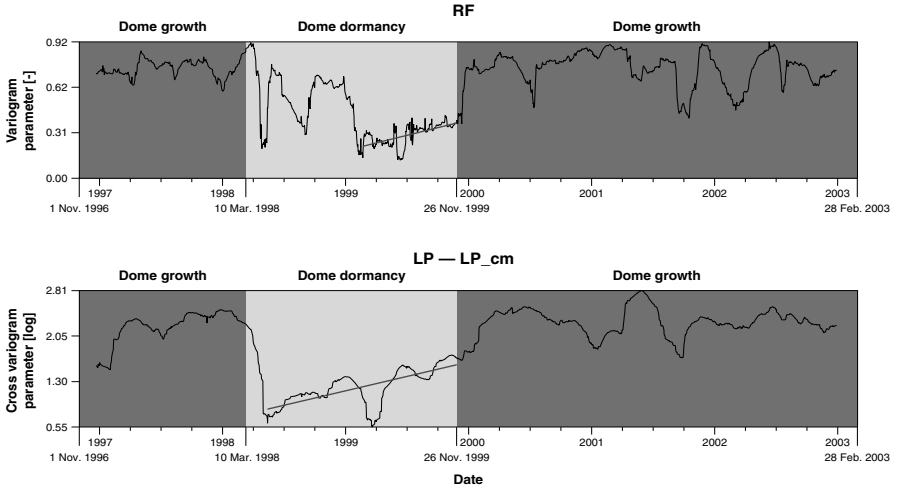


Fig. 4. Forecasting the onset of dome growth using univariate parameter monitoring estimated from rockfall events (above) and multivariate parameter monitoring estimated from long-period events and their cumulative magnitude (below).

Both monitored parameters exhibit an upward trend occurring months before the onset of the dome growth in November 1999 (Fig. 4). These trends can be considered as precursory behaviours indicating changes in the volcanic edifice in terms of persistent behaviour for RF events as well as in terms of correlation between the time series LP and LP_cm.

The monitoring of delay effects offers precursory potential. This parameter corresponding to a shift in time of the maximal correlation between two time series can be estimated using the asymmetrical behaviour with respect to the origin of

the cross covariance (Wackernagel 1995, Jaquet and Carniel 2001). The forecasting capability of delay effects is currently under investigation.

The application of parameter monitoring for forecasting remains problematic, since no uncertainty can be evaluated, the making of any decision becomes difficult. The realisation of forecasts with uncertainty requires the use of stochastic simulation on basis of potential scenarios. Such approach was attempted by Jaquet *et al.* (2004). The probability of occurrence for the onset of the dome growth over a period of several months was estimated on the basis of potential scenarios of seismic activity. Such forecasts constitute valuable input data as required by probabilistic risk assessments.

7 Conclusions

A geostatistical framework is provided for analysis and forecasting of volcanic activity using multi-parametric (incomplete) time series sampled at active volcanoes. This stochastic approach, applied in the time domain, allows detection and quantification of time (cross) correlation using (cross) variograms. The identification of precursors is achieved by parameter monitoring of specific parameters estimated from multi-parametric time series. Using precursory behaviour, forecasts are produced by stochastic simulation and their associated uncertainty is estimated within the framework of Monte Carlo approach.

The DEVIN approach provides forecasts with uncertainty as required by probabilistic risk assessments. In particular, such valuable input can be integrated to the formalism of generalised Bayesian Belief Networks (BBN) as applied by Aspinall *et al.* (2003). The BBN principle constitutes an increasingly accepted approach for performing decision-making under uncertainty. Using such approach, capable of accommodating any forecasting results (see, e.g., Aspinall *et al.* 2004), should enable to constraint the range of forecast uncertainty when performing decision-making during volcanic crises.

Further developments will consist in evaluating other types of multivariate parameters for monitoring time series in relation to specific volcanic events. In terms of forecasting, multivariate methods for stochastic simulations will be investigated in order to produce estimate of probability accounting of cross correlation between time series with potential precursory behaviour.

In terms of perspectives, these developments can be applied to other domains of the earth sciences, in particular for the analysis of incomplete and short time series monitoring geophysical parameters in relation to the forecasting of geological and hydrological hazards such as landslides, rock instabilities and floods.

References

Aspinall WP, Woo G, Voight B, Baxter P (2003) Evidence-based volcanology: application to eruption crises. Journal of Volcanology and Geothermal Research, 128(1-3), 273-285

Aspinall WP, Carniel R, Jaquet O, Woo G, Hincks T (2004) Using Hidden Multistate Markov models with multi-parameter volcanic data to provide empirical evidence for alert level decision-support. Journal of Volcanology and Geothermal Research, submitted to Journal of Volcanology and Geothermal Research

Carniel R, Di Cecca M (1999) Dynamical tools for the analysis of long term evolution of volcanic tremor at Stromboli. Annali di Geofisica, 42, 3: 483-495

Carniel R, Di Cecca M, Rouland D (2003) Ambrym, Vanuatu (July-August 2000): Spectral and dynamical transitions on the hours-to-days timescale. Journal of Volcanology and Geothermal Research, 128, 1-3: 1-13

Carniel R, Ortiz R, Di Cecca M (2004) Spectral and dynamical hints on the timescale of preparation of the 5 April 2003 explosion at Stromboli volcano. Journal of Volcanology and Geothermal Research, submitted to Journal of Volcanology and Geothermal Research

Chilès JP, Delfiner P (1999) Geostatistics: modelling spatial uncertainty. Wiley Series in Probability and Mathematical Statistics, p. 695

Jaquet O, Carniel R (2001) Stochastic modelling at Stromboli: a volcano with remarkable memory. Journal of Volcanology and Geothermal Research, 105, 249-262

Jaquet O, Carniel R (2003) Multivariate stochastic modelling: towards forecasts of paroxysmal phases at Stromboli. Journal of Volcanology and Geothermal Research, 128, 261-271

Jaquet O, Carniel R, Sparks RSJ, Thompson G (2004) DEVIN: a forecasting approach using stochastic methods applied to the Soufrière Hills Volcano, submitted to Journal of Volcanology and Geothermal Research

Kilburn CRJ, Voight B (1998) Slow fractures as eruption precursors at Soufrière Hills volcano, Montserrat. Geophysical Research Letters, Vol. 25, No. 19, 3665-3668

Kilburn CRJ (2003) Multiscale fracturing as a key to forecasting volcanic eruptions. Journal of Volcanology and Geothermal Research, 125, 271-289

Lantuéjoul C (2002) Geostatistical simulation: models and algorithms. Springer, p. 256

Matheron G (1962) Traité de géostatistique appliquée. Tome 1, Editions Technip, Paris, 334

Miller AD, Stewart RC, White RA, Luckett R, Baptie BJ, Aspinall WP, Latchman JL, Lynch LL, Voight B (1998) Seismicity associated with dome growth and collapse at the Soufrière Hills volcano, Montserrat. Geophysical Research Letters, 25, 3401-3404

Ortiz R, Moreno H, Garcia A, Fuentealba G, Astiz M, Peña P, Sanchez N, Tárraga M (2003) Villarica Volcano (Chile): Characteristics of volcanic tremor and forecasting of small eruptions by means of means of material failure method. Journal of Volcanology and Geothermal Research, 128, 1-3, 247-259

Ripepe M, Harris AJL, Carniel R (2002) Thermal, seismic and infrasonic evidences of variable degassing rates at Stromboli volcano. Journal of Volcanology and Geothermal Research, 118, 285-297

Robertson REA, Aspinall WP, Herd RA, Norton GE, Sparks RSJ, Young SR (2000) The 1995-98 eruption of the Soufrière Hills volcano, Montserrat. Philosophical Transactions of the Royal Society 358, 1619-1637

Sparks RSJ (2003) Forecasting volcanic eruptions. Earth and Planetary Sciences Letters, 210, 1-15

Sparks RSJ, Young SR (2002) The eruption of Soufrière Hills Volcano, Montserrat: overview of scientific results. In: Druitt, T.H. and Kokelaar, B.P. (eds) The eruption of Soufrière Hills volcano, Montserrat, from 1995 to 1999, Geological Society, London, Memoirs 21, 45-69

Voight B, Hoblitt RP, Clarke AB, Lockhart AB, Miller AD, Lynch L, McMahon J (1998) Remarkable cyclic ground deformation monitored in real time on Montserrat, and its use in eruption forecasting. Geophysical Research Letters, Vol. 25, No. 18, 3405-3408

Voight B, Young KD, Hidayat D, Subandrio, Purbawinata MA, Ratdomopurbo A, Suharna, Panut, Sayudi DS, LaHusen R, Marso J, Murray TL, Dejean M, Iguchi M, Ishihara K (2000) Deformation and seismic precursors to dome-collapse and fountain-collapse nuées ardentes at Merapi Volcano, Java, Indonesia, 1994-1998. Journal of Volcanology and Geothermal Research, 100, 261-287

Wackernagel H (1995) Multivariate geostatistics. Springer-Verlag, Berlin, 1-256

Woo G (1999) The mathematics of natural catastrophes. Imperial College Press, London, p. 292

Delineation of estuarine management units: Evaluation of an automatic procedure

F. Bação[1], S. Caeiro[2], M. Painho[1], P. Goovaerts[3] and M. H. Costa[4]

[1]ISEGI/CEGI, Institute for Statistics and Information Management of the New University of Lisbon, Portugal, e-mail: bacao@isegi.unl.pt, painho@isegi.unl.pt
[2]IMAR/Department of Exact and Technological Sciences of the Portuguese Distance Learning University, Lisbon, Portugal, e-mail: scaeiro@univ-ab.pt
[3]BioMedware, Inc. 516 North State Street, Ann Arbor, USA,
e-mail: goovaerts@biomedware.com
[4]IMAR, Faculty of Science and Technology of the New University of Lisbon, Caparica, Portugal, e-mail: mhcosta@fct.unl.pt

1 Introduction

A costal zone management (CZM) program should have well defined zones that can be managed based on their specific characteristics and needs. Once established, these management units constitute the backbone of the whole management strategy. The objective of a zoning plan is to delineate smaller areas that can be managed in a more flexible way (Cicin-Sain and Knecht 1998). Over the last few decades, there has been a move towards developing ways to identifying these units (McGlashan and Duck 2002).

The definition of the transition zone between the ocean and terrestrial environment, ocean and coastal zones, and zones (or units) within the coastal areas is sometimes not an easy task. Physical criteria, political boundaries, administrative boundaries, arbitrary distances or selected environmental units can and are often used (Clark 1996).

Most CZM projects use administrative boundaries instead of adopting an ecosystem approach looking at impacts coming from outside the area considered (Belfiore 2000). Coastal management units are evolving by becoming more inclusive, relying more on processes than on administrative boundaries and by incorporating a wider range of expertise in defining relevant areas (McGlashan and Duck 2002). The correct way to delineate estuarine management units is based on the development of robust and ecological representative processes. Eventually, these processes can be implemented using automatic procedures capable of providing promptly answers to complex problems. The automatic procedure proposed here draws inspiration on previous work developed in geographic zone design. In fact, the task of developing estuary management units can be viewed as a special case of the more general problem of geographical zone design (Martin 2000). Zone design is a long-standing geographical problem that is present in a number of geo-

graphical tasks; the best known example probably is electoral districting (Williams 1995, Macmillan and Pierce 1994). Zone design algorithms have also been used in a variety of tasks; such as school districting (Ferland and Guénette 1990), the design of zones with appropriate characteristics for posterior socio-economic and epidemiological analysis (Haining *et al.* 1994, Openshaw and Rao 1995, Openshaw and Alvanides 1999), the design of sales territories (Fleischmann and Paraschis 1988) and census output geography (Martin 1997, Martin 1998).

The aim of this paper is to develop an automatic optimization procedure, based on genetic algorithms, to delineate sediment estuarine management units, and to compare these results with a well-established multivariate geostatistical method. Both approaches will be illustrated using the Sado Estuary. The resulting management units will represent the support infrastructure of an environmental data management framework to monitor this ecosystem.

1.1 Zone design using multivariate geostatistical tools

Geostatistical techniques like kriging allow estimation of attribute values at unsampled locations taking into account the spatial continuity of the data (Soares 2000). Since kriging is preceded by an analysis of the spatial structure of the data, the average spatial variability of the data is already integrated into the estimation/interpolation process (Wackernagel 1995).

Multivariate methods like principal component analysis, cluster analysis and discriminant analysis can be coupled with the different types of kriging (Oliver and Webster 1989, Reed *et al.* 2001, Goovaerts 2002) allowing one to group sampling sites that both have similar properties and are geographically close. With these multivariate geostatistical techniques interpolation is improved, small occurrences of one kind of land within others of fairly similar kind are disregarded and undesirable fragmentation avoided (Goovaerts 1997, Reed *et al.* 2001).

For example (MacDonald *et al.* 2000) developed an ecosystem-based framework for assessing and managing sediment quality conditions in Tampa Bay previously defined management areas. Those areas were delineated using interpolated contour lines based on sediment chemistry data and guidelines of potential adverse effects. Picollo *et al.* (2003) used homogenous units for the coastal zone management of the Ligurian region. These subdivisions of the coast corresponded to physiographic units (topographic elements).

1.2 Zone design using genetic algorithms

The constraints of the zone design problem are similar to the ones that characterize the clustering problem. Let the set of initial areal units be $X = \{x_1, x_2..., x_n\}$, where x_i is the ith areal unit. The number of zones is denoted k, and Z_i is the set of all the areal units that belong to zone Z_i. Then:

$$\mathbf{Z}_i \neq \varnothing, \text{ for } i = 1,...,k, \tag{1}$$

$$\mathbf{Z}_i \cap \mathbf{Z}_j = \varnothing, \text{ for } i \neq j, \tag{2}$$

$$\bigcup_{i=1}^{k} \mathbf{Z}_i = \mathbf{X} \tag{3}$$

These constitute the set of constraints that can be applied equally in clustering and in zone design. Nevertheless, the typical zone design task usually presents an additional constraint, which accounts for contiguity and creates a more complex problem (Macmillan and Pierce 1994). We will not address the issue of contiguity here, as it is irrelevant in the context of this research. However the automatic procedure described below guaranties that the zones created will be contiguous.

To deal with the zone design problem a number of different algorithms have been proposed (for a thorough review in the context of electoral districting, see Williams 1995). Nevertheless, there are two major problems in applying the existing automatic zone design algorithms to the development of estuary management units. First, most of the algorithms described in the literature are not available as they result from research efforts and most of them were never implemented as software packages. Second, most algorithms are based on an areal perspective of the zone design problem, i.e. they use areas as the basic units for zone design, which conflicts with the point supports used in the context of this research.

The zone design algorithms proposed in the literature use different optimization strategies ranging from hill climbing procedures (Horn 1995) to simulated annealing (Macmillan and Pierce 1994, Openshaw and Rao 1995), tabu search (Openshaw and Rao 1995) and linear programming associated with branch-and-bound (Mehrotra *et al.* 1998). Genetic Algorithms (GA) remain largely unexplored in this field. To our knowledge the only reference is Altman (1998) and no details are provided on how GA's were applied to this particular problem. However, GA's have been used extensively as search procedures in related fields such as the P-Median Problem (Correa *et al.* 2001) and Cluster Analysis (Murthy and Chowdhury 1996). Other fields facing complex optimization problems, such as Pattern Recognition, Image Processing and Machine Learning (Ankenbrandt *et al.* 1990, Belew and Booker 1991, Back *et al.*, 1997) have also benefited from the use of GA's.

One of the reasons why the zone design problem is especially difficult is the size of the solution space. The dimension of a "usual" real world problem makes unfeasible any attempt to enumerate all the possible solutions explicitly (Cliff and Hagget 1970). The total number of possible solutions, *S*, for a zone design problem, is similar to the clustering problem and is calculated as the Stirling number of the second kind (Anderberg 1973, Keane 1975):

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^i\left(\frac{k!}{(k-i)!i!}\right)(k-i)^n \tag{4}$$

This means that for a medium size problem like the one addressed in the results section of this paper, S (153, 19) yields $3.65*10^{178}$ possible solutions. Additionally, in terms of computational complexity the zone design problem has been shown to be NP-Complete (Altman 1997). Thus, heuristic techniques seem to be the best way available to produce solutions in reasonable computational time.

2 Methods

2.1 Sampling design

Sediment samples were collected at 153 sites using a Global Positioning System, according to a systematic unaligned sampling design (500 × 750 m) to provide a uniform coverage of the area as well as pairs of close observations required for modeling the short-scale variability. In each site 3 sediment characterization attributes were determined: Fine Fraction content (FF), Total Organic Matter (OM) and Redox Potential (Eh) (Caeiro *et al.* 2003a). The central problem consisted in building a small number of regions (areas) based on the original 153 sample points. This will enable the periodic monitoring and adequate management of the estuary. Boundaries of spatially contiguous and homogeneous regions of sediment structure were derived through two alternative approaches described below.

2.2 Multivariate Geostatistical approach

This method starts with a principal component analysis (PCA) of original data (FF, TOM and Eh), followed by the computation and fitting of a spherical model to the experimental semivariogram of principal components (PC) scores. Following Oliver and Webster (1989), the dissimilarity between any two sampling sites i and j is then computed as:

$$d_{ij}^{*} = d_{ij} \times \frac{c}{c_0 + c} \times \left[1.5 \times \frac{|u_{ij}|}{a} - 0.5 \times \left(\frac{|u_{ij}|}{a} \right)^3 \right] + d_{ij} \times \frac{c_0}{c_0 + c} \quad \text{for } 0 < u_{ij} \leq a \tag{5}$$

$$d_{ij}^{*} = d_{ij} \quad \text{for } u_{ij} > a$$

Where:

d_{ij} - Distance in the attribute space between *i* and *j*
c - Sill of the spherical semivariogram model
c_0 - Nugget variance
a - Range of the spherical semivariogram model
u_{ij} - Euclidean geographic distance between *i* and *j*.

These values are assembled into a dissimilarity matrix that undergoes hierarchical clustering using the complete linkage rule (Everitt and Dunn 2001). The spatial continuity of each cluster is characterized using semivariograms computed on indicators of occurrence of these clusters. Indicator kriging is then used to interpolate the probability of occurrence of the clusters at unsampled locations. Finally, each grid node is assigned to the cluster with the highest probability of occurrence (maximum likelihood classification).

This method generates relatively smooth maps showing locally dominant classes, uncluttered by outliers. This procedure fulfills the purpose of computing contiguous sediment regions for management and monitoring purposes. A detailed description of this method is available in Caeiro *et al.* (2003a). The area corre-

sponding to the sampling points was further clipped with the study area boundary including the coast line (Caeiro *et al.* 2003b) using ArcGIS 8.0 software.

2.3 Genetic algorithms approach

Genetic algorithms (GA) are a subset of a broader and rapidly expanding area known as Evolutionary Computing (Fogel 2000). As the name indicates, these algorithms drew inspiration on Darwin's theory of evolution, and have been used to solve hard optimization and machine learning problems (Goldberg 1989). The basic idea is that each solution to the problem is coded as a bit string, a chromosome, possibly with a number of sub-strings that act as genes. At any given point in time (or generation), a number of such chromosomes are kept, each representing a different "individual" or solution to the problem. Natural selection is simulated by evaluating the fitness (or goodness) of each solution, measured by how well it solves the problem at hand, and giving the best individuals a higher probability of being chosen to breed (crossover) and thus passing their characteristics into the next generation. To obtain new solutions, two operators are used: crossover, and mutation. Crossover is implemented by combining bits of two different chromosomes (possibly divided along genes) to form a new solution, while mutation amounts at randomly changing some bits or chromosomes. The details of how this basic idea is implemented may vary considerably.

Given enough time, a conveniently coded GA will always find an optimal solution. However, to obtain reasonable solutions within reasonable time, care must be taken in the encoding of the problem into chromosomes, and in the choice of the fitness function that will be optimized.

The application of genetic algorithms to the development of estuary management units first requires a strategy for encoding a solution to the problem. In other words, a specific partition of the sampling points into a smaller set of management units needs to be encoded in such a way that genetic operators may be used. This could be performed using a number of different ways (Bação *et al.* 2004). Bearing in mind that in the specific case of estuary management units compactness is not a relevant constraint, the encoding used here enables the seed of each management unit to be placed anywhere within a rectangle comprising the study region.

Each string represents a possible plan configuration, and the fitness of each specific configuration is evaluated using the following expression:

$$min \sum_{i=1}^{k} \sum_{z=1}^{n_z} \sum_{v=1}^{m} \left| x_{izv} - \overline{x_{zv}} \right| \tag{6}$$

where x_{izv} represents the value of sampling point i from management unit z for variable v, and $\overline{x_{zv}}$ represents the mean value of variable v in management unit z. In order to assess the fitness of each solution we have to calculate the sum of the distances between each sampling point and the mean of the management unit to which it belong along all variables. Finally, the quality of the solution is assessed by the sum of the distances within all the management units.

The genetic algorithm is initialized with a random population of size *p*. For the encoding described above, *p* strings of length 2**k* are initialized. This is done by forcing all elements of the strings to be located within a rectangle comprising the study region. GA literature does not provide guidelines for choosing the size of the initial population. In this study 10 parallel populations were used, with different string numbers. Migration of strings between populations can occur with a probability of 0.001. Identical strings are not allowed, so there are no twins in the population.

The type of selection used is tournament (Goldberg 1989). Sensitivity analysis led to the choice of a uniform crossover, with a probability ranging from 0.95 to 1. Mutation rate was 0.001 and 0.002, and an elitist strategy was adopted, assuring that the best individual of the population would always be carried to the next generation. Different stopping criteria were also used.

Thus, the algorithm proceeds as follows:

1. Generate p sets of k points, according to the selected encoding.
2. For each of the *n* sampling points find the closest seed, and assign the sampling point to the seed.
3. Evaluate the fitness of each string, based on equation 6.
4. Apply selection, crossover and mutation operators, creating a new population;
5. Return to step 2 until the stopping criterion is met.

The final result is a classification of the sampling points into *n* zones. To generate areas (management units) an allocation function was computed using the Spatial Analyst extension of ArcGIS 8.0 software.

3 Results

In this section we compare the results provided by the two approaches. The major goal is to identify the differences rather than trying to prove the superiority of one technique over the other. The elusiveness of a strict objective criterion to the delineation of estuary management units forces a somewhat subjective analysis of the results. The use of expert knowledge was one of the ways of circumvent this difficulty.

3.1 Multivariate Geostatistical approach

The hierarchical classification yielded four clusters that are reasonably distinct, with a decline in organic load from Cluster 1 to 4: HO – High Organic; MHO – Medium High Organic; MO – Medium Organic and LO – Low Organic loads (Fig. 1). For each cluster, the indicator semivariogram was computed along four directions and a geometric anisotropy model was fitted visually. All semivariograms display longer ranges in the direction of azimuth 120º, which corresponds to the water flow and is in agreement with other studies (Martins *et al.*

2001). Maximum likelihood classification performed on estimated probabilities generated 70 areas after clipping with the estuary costal line. The areas smaller than the sampling grid size were assigned to the neighboring area with the longest common border, resulting in a final set of 19 management units. A detailed description of the methodology is available in Caeiro *et al.* (2003a). Results of cluster classification of the 153 sampling points and the corresponding 19 management units are displayed in Fig. 1.

The results of this method are generally in agreement with earlier work performed in the estuary (Rodrigues and Quintino 1993). Low organic load sediments correspond essentially to the estuarine entrance and tend to extend to the inside of the estuary, basically through the southern channel (see at the estuary entrance, a large homogenous area of low organic load, LO1 - Fig. 1). In the middle of the estuary bay the gradient splits into two major components, one directed towards the North Channel and the other towards the South Channel in accordance with the water flow. Since high organic load areas are associated with low hydrodynamics and rich organic discharges, they are more common in the North Channel near industrialized zones and the city of Setubal. They are also distributed in small homogenous areas (Fig 1). These results were compared to the classification provided by two other multivariate approaches using map similarity measurements (Caeiro *et al.* 2003a, Caeiro *et al.* 2004). These previous studies demonstrated the robustness of this multivariate approach, indicating that the different methods yield similar results and thus are of equal value to delineate management units in the estuary.

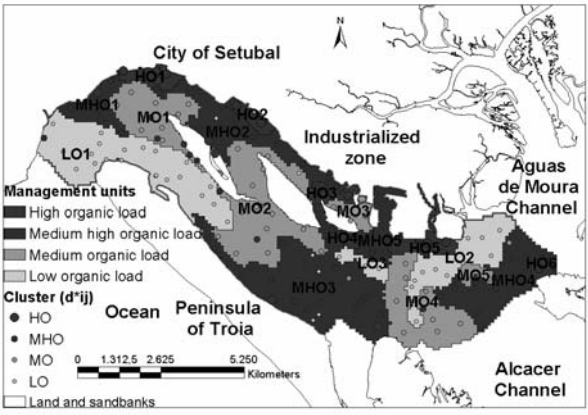


Fig. 1. Classification of the 153 sampling point produced using hierarchical classification (four clusters), and the final set of 19 management units obtained with the multivariate geostatistical approach.

3.2 Genetic algorithms approach

Genetic algorithm was run using five different sets of parameters (see Table 1) and under the constraint of creating 19 areas. Solution S1 achieves the worst result

in terms of within cluster variability, nevertheless it is important to note that only 50000 individuals (10 populations of 10 individuals over 500 generations) were evaluated, which constitutes a very small fraction of the solution space ($3.65*10^{178}$). This result indicates that genetic algorithms are quite robust, as they need to search only a small amount of this space in order to find good solutions. In the case of S5 the algorithm stops after 1000 generations without improvements in the objective function, yielding a total of 2907 generations. Only solution S1 did not lower the within cluster variability produced by the geostatistical method (134.17).

Table 1. - Results and parameter specification of the 5 runs performed in the genetic algorithm.

	Sum of Within management unit variability	Population	Mutation	Stopping Criterion	Crossover probability
S1	143.92	10*10	0.001	500 generations	1
S2	123.30	10*15	0.001	1000 generations	1
S3	123.05	10*25	0.001	800 generations	1
S4	126.16	10*25	0.002	800 generations	0.95
S5	117.19	10*25	0.001	2907 generations	0.95

Because of the page limitations, only the best (S5) and worst results (S1) are displayed (see Fig. 2 and 3). The analysis of the maps shows that S5 is able to isolate 6 sampling points into 4 areas, which creates smaller areas while improving the objective function. The two solutions are, however, characterized by similar spatial pattern of the management units.

The genetic algorithm yields 19 management units without clustering them in 4 groups, which makes the comparison with the geostatistical method more difficult. The final result is the classification of each sampling point into one of the 19 management units, unlike the geostatistical approach which produces surface areas. Despite these differences the analysis of the resulting maps shows that similar patterns of small areas are found in the North Channel. In fact, the general results are quite similar in terms of the "macro" trends. The genetic algorithm identifies small management areas (like area nº 19 in Fig. 2 or nº 8 in Fig. 3), which were also detected by the multivariate geostatistical method (see Fig. 1 cluster results), but as the interpolation procedure only yielded an area smaller than the sampling grid size, that was discarded. The split gradients in the North and South channels displayed by the hierarchical classification are less noticeable in the two solutions of the genetic algorithm. This can be explained by the fact that the genetic algorithm does not take into account the anisotropic pattern of variability of the attributes, which can be modeled and incorporated using the geostatistical approach. The estuarine sedimentary environment is strongly controlled by the water flow, which supports the existence of the long areas identified by the geostatistical method, particularly in the south channel where the hydrodynamics is higher, representing the natural physical behavior. Due to distance restrictions imposed by the encoding the genetic algorithm is unable to create elongated areas (see areas nº 7 and 11

in Fig. 2 or nº 7 and 18 in Fig. 3, that correspond to one single area – LO1 in Fig. 1).

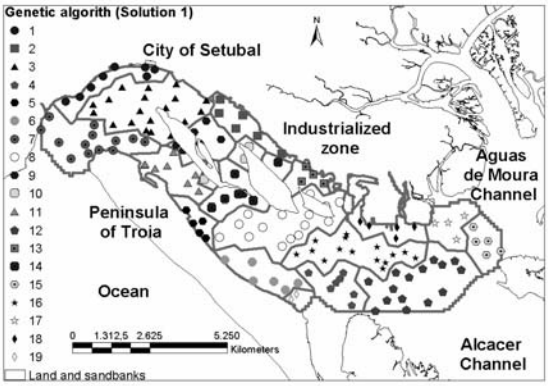


Fig. 2. Classification of the 153 points produced using the genetic algorithm approach – S1.

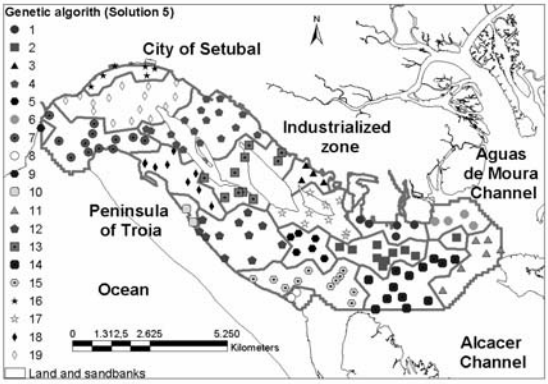


Fig. 3. Classification of the 153 points produced using genetic algorithm approach – S5.

About 60 % of the total number of the GA's areas of both solutions 1 and 5 are mainly formed by one single cluster (i.e. proportion is higher then 80 %). However, in solution 5, only 31.6 % of the areas are 100 percent classified as belonging to one single type of cluster, compared to 15.8 % for Solution 1. In spite of this S1 seems to represent better the estuary hydrodynamic behavior, and is more in accordance with the geostatistical approach. In addition, 47.4 % of the 19 centroids of Solution 1 are associated with one single centroid of the management units of geostatistical approach compared to only 36.8 % in solution 5.

In order to organize the major findings of this study, the quality of each solution was assessed using four point expert evaluation criteria, see Table 2. As expected, the quality of the ecological representation is superior in the geostatistical method, as it accommodates specific needs related to underlying phenomena, such as the hydrodynamic behavior. Nevertheless, it is noteworthy that the genetic algorithm

is not inadequate, since for instance the expert assessment is quite encouraging (does not need to have an in-depth understanding of the methods).

Table 2. Assessment scores of the results of the multivariate geostatistical approach (MG) and solutions 1 and 5 provided by GA. Best score is 3.

Method	Sum of within management unit variability	Ecological interpretation (N-S differences)	Ecological interpretation (Hydrodynamic behavior)	Required knowledge about the method from the user*	Expert Assessment	Total
MG	134.17	3	3	1	3	10
S1	143.92	2	2	3	2	9
S5	117.19	2	1	3	2	8

* 1 means that the user needs to have an in-depth understanding of the methods.

4 Conclusions

In this paper, we implemented a genetic algorithm approach for the delineation of management units in estuary and the results were compared to a multivariate geostatistical approach based on a hierarchical clustering of a spatial dissimilarity matrix followed by indicator kriging. Although the task of comparing such different approaches is difficult, due to different types of outputs and the absence of a "reference truth", this study indicates that the results of both methods are relatively similar. The automatic procedure presented here has the potential to become a valuable option in the delineation of estuary management units. The current genetic algorithm can provide a benchmark for other approaches; enabling the possibility of critical assessment of theoretical based approaches.

The first solution S1, although with higher within area variability, seems to represent better the ecological behavior of the estuary (see Table 2). This fact may indicate that using the within management unit variability as the optimization criteria might be misleading, or at least it can be improved. Finally, future work should investigate alternate encoding options which would enable the algorithm to produce more elongated, and still contiguous, areas. This improvement should lead to a better representation of the hydrodynamic effects, which play a relevant role in the definition of the management units.

In this study the issue of defining the number of regions was not addressed. Clearly, in the future, the genetic algorithm should incorporate new strategies for the automatic computation of the number of regions to be built. This should capitalize on the work developed within cluster analysis, for instances the application of a pseudo F statistic developed by Calinski and Harabasz (1974) or the cubic clustering criterion proposed by Sarle (1983).

References

Anderberg MR (1973) Cluster Analysis for Applications. Academic Press

Ankenbrandt CA, Buckles BP, Petry FE (1990) Scene recognition using genetic algorithms with semantic nets. Pattern Recognition Letters 11(4):285-293

Altman M (1997) The Computational Complexity of Automated Redistricting: Is Automation the Answer? Rutgers Computer and Technology Law Journal 23 (1):81-142

Altman M (1998) Modeling the Effect of Mandatory District Compactness on Partisan Gerrymanders. Political Geography 17 (8): 989-1012

Bação F, Lobo V, Painho M (2004) Applying Genetic Algorithms to Zone Design. Accepted for publication in Soft Computing: A Fusion of Foundations, Methodologies and Applications

Back T, Fogel DB, Michalewicz Z (eds.) (1997) Handbook of Evolutionary Computation. Oxford University Press NY

Belew RK, Booker LB (eds.) (1991) Proceedings of the Fourth International Conference on Genetic Algorithms. La Jolla, CA: Morgan Kaufmann

Belfiore S (2000) Recent developments in coastal management in the European Union. Ocean & Coastal Management 43:123-135

Caeiro S, Goovaerts P, Painho M, Costa MH (2003a) Delineation of Estuarine management areas using multivariate geostatistics: the case of Sado estuary. Environmental Science and Technology 37:4052-4059

Caeiro S, Goovaerts P, Painho M, Costa MH, Sousa S (2003b) Spatial sampling design for sediment quality assessment in estuaries. Environmental Modeling and Software 18(10): 853 – 859

Caeiro S, Sousa S, Painho H (2004) Map Similarity Measurements and its application to the Sado Estuary. Accepted for publication in Finisterra

Calinski T, Harabasz J (1974) A Dendrite Method for Cluster Analysis, Communications in Statistics 3:1 -27

Cicin-Sain B, Knech RW (1998) Integrated Coastal and Ocean Management. Concepts and Practices. Washington, D. C, Covelo, California, Island Press

Clark JR (1996) Coastal Zone Management Handbook. Boca Raton, Florida, Lewis Publishers

Cliff AD, Hagget P (1970) On the efficiency of alternative aggregations in region-building problems. Environment and Planning 2: 285-294

Correa ES, Steiner MTA, Freitas AA, Carnieri C, (2001) A genetic algorithm for the P-median problem. In: Proc. 2001 Genetic and Evolutionary Computation Conf. (GECCO-2001),Morgan Kaufmann, 1268-1275

Everitt B, Dunn G (2001) Apllied Multivariate Data Analysis. New York, Arnold

Ferland JA, Guénette G (1990) Decision support system for a school districting problem. Operations Research. 38:15-21

Fleischmann BJ, Paraschis N (1988) Solving a large scale districting problem: a case report. Computers and Operations Research. 15:521-533

Fogel DB (2000) Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. 2nd edition, IEEE Press, Piscataway, NJ

Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Mass

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. New Work, Oxford University Press

Goovaerts P (2002) Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data. Journal of Geographical Systems 4: 99-111.

Haining R, Wise S, Blake M (1994) Constructing regions for small-area analysis – material deprivation and colorectal cancer. Journal of Public Health Medicine, 16: 457-469.

Horn M (1995) Solution techniques for large regional partitioning problems. Geographical Analysis 27: 230-248.

Keane M (1975) The size of the region-building problem. Environment and Planning A 7: 575-577.

MacDonald DD, Lindskoog RA, Smorong DE, Greening H, Pribble R, Janicki T, Janicki S. Grabe S, Sloane G, Ingersoll CG, Eckenrod S and Long ER (2000) Development of an Ecosystem-Based Framework for Assessing and Managing Sediment Quality Conditions in Tampa Bay, Florida. Florida, Tampa Bay Estuary Program.

Macmillan B, Pierce T, (1994) Optimization Modelling in a GIS Framework: The Problem of Political Redistricting. In: Fotheringham S, Rogerson P (eds.) Spatial analysis and GIS. Taylor & Francis Inc, Bristol, 221-246.

Martin D (1997) From enumeration districts to output areas: experiments in the automated creation of a census output geography. Population Trends, 88:36-42

Martin D (1998) 2001 Census output areas: from concept to prototype. Population Trends, 94:19-24

Martin D (2000) Automated zone design in GIS. In: Atkinson, P. and Martin, D. (eds.) GIS and Geocomputation, Innovations in GIS 7: Taylor and Francis, London, pp. 103-113

Martins F, Leitão P, Silva A and Neves R (2001). 3D modelling in the Sado estuary using a new generic vertical discretization approach. Oceanologica Acta 24:S51 - S62.

McGlashan D. and Duck RW (2002) The Evolution of Coastal Management Units: Towards the PDMU. In: F. Veloso-Gomes, F. Taveira-Pinto, and L. Neves (eds.). Littoral 2002, The Changing Coast. Porto, European Coastal Zone Association for Science and Technology (EUROCOAST)- Portugal Association /EUCC - The Coastal Union, 29 – 33.

Mehrotra A, Johnson EL, Nemhauser GL (1998) An optimization based heuristic for political districting. Management Science, 44:1100-1114.

Murthy CA, Chowdhury N (1996) In search of optimal clusters using genetic algorithms. Pattern Recognition Letters, 17:825-832

Oliver MA and R. Webster, R (1989) A Geostatistical Basis for Spatial Weighting in Multivariate Classification. Mathematical Geology 21:15-35.

Openshaw S, Rao L, (1995) Algorithms for reengineering 1991 Census Geography. Environment and Planning A 27:425-446

Openshaw S, Alvanides S (1999) Applying geocomputation to the analysis of spatial distributions, In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW, (eds.) Geographical Information Systems Principles, Techniques, Management and Applications, Wiley, Chichester, 267-282.

Picollo A, Albertelli G, Bava S and Cappo S. (2003) The Role of Geographic Information Systems (GIS) and of DPSIR Model in Ligurian Coastal Zone Management.In: Proceedings of 5th International Symposium on GIS and Computer Cartography for Coastal Zone Management Italy, Geographical Information Systems International Group (GISIG/International Center for Coastal and Ocean Policy Studies(ICOOPS). 1 – 5

Reed J, Chappel, A, French JR and Oliver MA (2001) A geostatistical analysis of PCB-contaminated sediment in a commercial dock, Swansea, UK. P. In: Monestiez, P., Allard, D.,. Froidevaux, R. (eds.) *geoENV III* - Geostatistics for environmental applications. Kluwer Academic Publishers, Dordrecht. 487 – 498.

Rodrigues AM and Quintino VMS (1993) Horizontal Biosedimentary Gradients Across the Sado Estuary, W. Portugal. Netherlands Journal of Aquatic Ecology 27:449-464.

Sarle WS (1983) Cubic Clustering Criterion, SAS Technical Report A-108, Cary, NC: SAS Institute Inc.

Soares A (2000) Geostatística para as Ciências da Terra e do Ambiente. IST Press Lisboa.

Williams JC (1995) Political Redistricting - A Review. Papers in Regional Science 74(1): 13-39

Wackernagel H (1995) Multivariate Geostatistics: an Introduction with Applications. Springer-Verlag.

Estimating indicators of river quality by geostatistics

C. Bernard-Michel and C. de Fouquet

Ecole des Mines de Paris, 35 rue Saint Honoré, 77305 Fontainebleau, France.

1 Introduction

In order to assess river water quality, nitrate concentrations are measured in different monitoring stations. The information contained in these measurements is summarized in a few synthetic quantitative indicators such as the 90% quantile of yearly concentrations or the annual mean making it possible to compare water quality in different stations, and its yearly evolution. The current French recommendations are based on the French water quality's evaluation system (SEQ EAU, http://www.rnde.tm.fr/) and the water framework directive in Europe, which aims at achieving good water status for all waters by 2015.

These calculations, however, use the classical statistical inference, essentially based on a hypothesis proved to be incorrect for nutrients (Bernard-Michel and de Fouquet 2003): time correlations are not taken into account. Moreover, the seasonal variations of concentrations and the monitoring strategy are ignored. For example, because of the run-off or the leaching of fertilizer, nitrate concentrations of Loire Bretagne basin are often high in winter and low in summer (Payne 1993). Thus, if sampling is increased in winter out of precaution, the annual mean and the quantile can be falsely increased. It is therefore necessary to develop methods that take into account both time correlations and sampling dates, especially in case of preferential sampling. We propose to assign kriging weights or segment of influence weights (Chilès 1999) to measurements for both indicators and to use a linear interpolation of the empirical quantile for the estimation of the 90^{th} percentile.
In this paper, methods are presented and compared for nutriments on simulations.

2 Example

Fig. 1 (left) shows an example of real nitrate concentrations measurements from the Loire River in 1985. The indicators have been estimated first with the totality of measurements (6 in summer, 12 in winter), then with an extracted sample of one regular measurement a month.

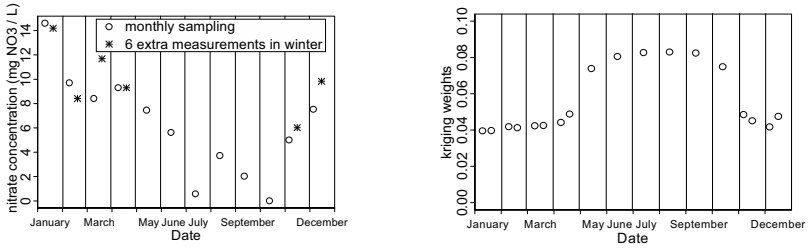


Fig. 1. Preferential sampling of nitrates concentration during one year at one monitoring station. Left: the measurements frequency is doubled in winter; Right: associated kriging weights.

Table 1. Statistical annual mean and quantile of nitrates concentrations.

Sample size	Sample mean	90% empirical quantile
12	6.16	9.70
18	7.41	14.19

Classical estimations are presented in Table 1: usual indicators are obviously increased when sampling is reinforced in winter. It is a consequence of the preferential sampling and of the presence of time correlation showed for many nutriments (Fig. 4). It therefore appears necessary to develop methods to better assess yearly temporal mean.

3 The annual mean: statistical parameter or time average?

3.1 Methods

Experimental temporal variograms calculated on nitrates concentrations show for most of the monitoring stations the evidence of a time correlation. The sample mean (i.e. the arithmetic mean of experimental data) is an unbiased estimator for independent data or regularly spaced correlated data (with certain exception). In presence of time correlation, it is no longer the case when sampling is irregular or preferential. To correct this bias, two methods were studied:

- Kriging with unknown mean (OK) which takes into account correlation in the estimation of the annual mean and in the calculation of the estimation variance;
- A geometrical declustering whose objective is only to correct the irregularity of sampling.

These methods are presented below and compared later on simulations (2.2):

1. Classical statistical method (Saporta 1990): sample values $z_1, z_2, ..., z_n$ are interpreted as realizations of independent random variables $Z_1, Z_2, ...Z_n$ which all

have the same distribution, with expectation m. The yearly mean corresponds to the estimation of this expectation, using the sample mean (1) denoted m^*. The estimation variance (2) is deduced from the experimental variance σ^2 :

$$m^* = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{1}$$

$$Var\left(\overline{Z}_n - m\right) = Var\left(\overline{Z}_n\right) = \frac{\sigma^2}{n} \text{ with } \sigma^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(Z_i - m^*\right)^2 \tag{2}$$

2. Temporal kriging (Chilès 1999): sample values are interpreted as a realization of a random correlated function $Z(t)$ at dates $t_1, t_2, ..., t_n$. We don't estimate anymore the parameter of a distribution, but the temporal mean $Z_T = \frac{1}{T}\int_T Z(t)dt$, still defined even in absence of stationarity. This quantity is estimated using ordinary "block" kriging, with constant but unknown mean :

$$Z_T^* = \sum_{i=1}^{n} \lambda_i Z_i \text{ where } \lambda_i \text{ are kriging weights} \tag{3}$$

$$Var(Z_T^* - Z_T) = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C\left(t_\alpha - t_\beta\right) + \overline{C}(T,T) - 2\sum_{\alpha=1}^{n} \lambda_\alpha \int_T C\left(t_\alpha - t\right)dt \tag{4}$$

Analytical expressions are easy to calculate at 1D, without any discretization (Matheron 1970; Journel 1977). Fig. 1 (right) gives an example of kriging weights, assigning lower weights to winter values, which avoids an estimation bias. The estimation variance and confidence interval, overestimated by classical statistics, are reduced by kriging taking into account the temporal correlation and the annual periodicity of the concentration.

3. Geometrical method (Chilès and Delfiner 1999) by segment declustering, corresponding to 1D polygonal declustering. This technique consists in weighting each data by the relative length of its segment of influence, in the linear combination (5). An example for 4 measurements is given in Fig. 2. Calculating the estimation variance necessitates the variogram.

$$Z_T^* = \sum_{i=1}^{n} \lambda_i Z_i \text{ where } \lambda_i \text{ are segment of influence weights} \tag{5}$$

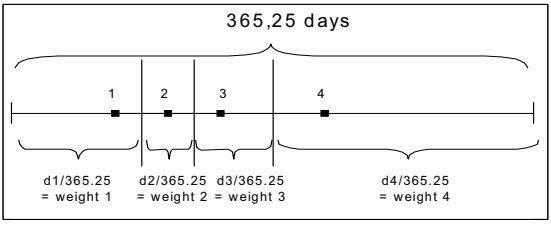


Fig. 2. Example of segment declustering for 4 measurements

3.2 Testing methods on simulations

3.2.1 Choice of the monitoring station

In order to quantify the improvements of the new proposed methods, it would be necessary to have examples in which the annual mean is known. Because it is impossible, we propose to simulate 365 days of measurements based on a real dataset. We then admit that one measurement a day exactly determines the yearly mean.

The best sampled monitoring station available was on the Loire river in Orléans: in 1985, 1 measurement was taken every 2 days, in 1986, 1 measurement a week, and for other years 3 measurements a month. A Gaussian sequential conditional simulation of "daily" concentrations over ten years with respect to experimental data in Orleans and the fitted experimental variogram (Fig. 4) was constructed (Fig. 3, only the 1985 simulated values are presented). Then, samples were extracted to compare the annual means exactly known to the three estimation methods.

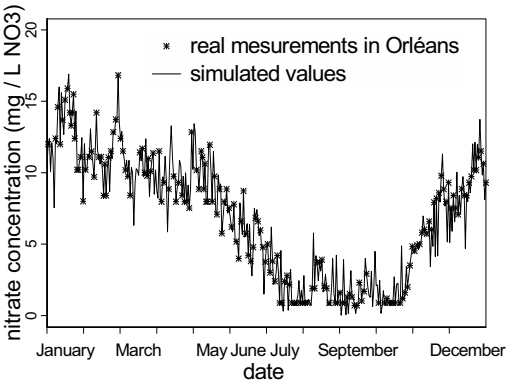


Fig. 3. Conditional simulation based on the real measurements. Loire river in Orléans.

3.2.2 Variogram model

The experimental variogram calculated over several years for Orléans (Fig. 4, right), reflects the annual periodicity of nitrate concentrations. The variogram calculated over one year with a lag of 2 days (Fig. 4) show the predominance of this periodical component. In these mean temporal variograms, winter or summer values are not distinguished. Variograms calculated for each season would differ, but as we are interested in the global annual statistics, the averaged variogram on one year is here sufficient (Matheron 1970).

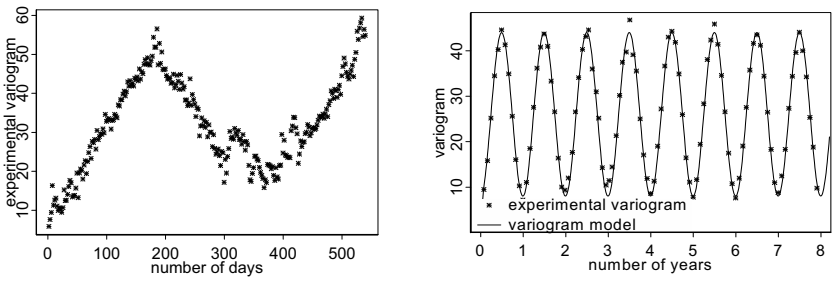


Fig. 4. Experimental variogram for nitrate at Orleans. Left, calculation for one year, lags of 2 days. Right, calculation for 8 years, lag of 30 days. This variogram has been fitted manually by the sum of a nugget effect, a spherical model and a cosinus model.

3.2.3 Simulations and results

In Fig. 5, a comparison between the estimation of the annual mean by statistics and geostatistics over 10 years is presented. The estimations are given with their 95% confidence intervals and compared to the real value of the annual mean estimated with 365 measurements. Samples are preferential (6 values in summer, 12 in winter) and have been extracted from the simulation. Fig. 5 (left) confirms that the corresponding sample mean is often higher than yearly mean, and moreover it leads to a correspondingly large 95% confidence interval. The bias is well corrected by the geostatistical and geometrical methods for which weights are equivalent (Fig. 5, right). Nevertheless, kriging directly gives the estimation variance.

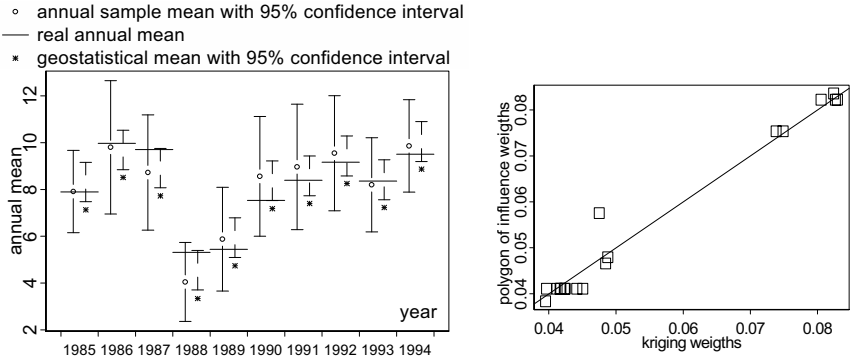


Fig. 5. On the left, estimations by sample mean and kriging are presented with their associated 95% confidence interval. They're compared to real annual mean. On the right, scatter diagram between weights, for kriging (abscissa) and for segments of influence (ordinates).

For most of the years, kriging gives better estimations than statistics and moreover 95% confidence intervals are about twice smaller than the ones given by sta-

tistics, and always include the true yearly mean. These results are confirmed on 1000 simulations (Table 2). Other examples for different stations, parameters and with different sampling strategy can be found in Bernard-Michel and de Fouquet, 2003.

As first conclusion, kriging corrected the bias in case of preferential sampling, better assessed yearly mean, and better predicted the precision of this estimation. However, if we are only interested in the value of annual mean, segment declustering can be used because of its simplicity. If the precision is needed, then kriging should be preferred.

Table 2. 1000 simulations : comparison of statistics and geostatistics estimations in average for a preferential sampling (12 measurements in winter, 6 in summer)-

Average of the 1000 annual mean estimated with 365 measurements	6.72	
Preferential sampling	Statistics	Geostatistics
Average of the annual means	7.62	6.72
Average of the predicted standard deviations of estimation errors	0.97	0.42
Experimental standard deviation of error	0.93	0.31
Experimental 95% confidence interval	[7.10;8.12]	[6.07;7.37]

4 Estimation of the 90% quantile

The 90% quantile is used by water agencies to characterize high concentrations, potentially the most dangerous for human health. However, today's recommendation to approach the 90% quantile is based on the empirical quantile. This statistical method is proved to be problematic for the following reasons:

- It is a biased estimator (Gaudoin 2002).
- As the sample mean, it does not take into account time correlation, and sample irregularity.

We first evaluated the bias of the empirical quantile in the case of independent variables, and proposed three methods to remedy. Then, we took into account the time correlation and the sampling irregularity by weighting the measurements.

4.1 Bias of the empirical 90% quantile of independent data

Generally, the estimation of percentiles is a part of extreme values theory (Coles 2001). However, this theory is based on asymptotic theorems which require many measurements. As we will only dispose of an average of 12 measurements a year, we propose to use a classical non-parametric estimator: the empirical quantile (Saporta 1990, Gaudoin 2002). But this estimator is proved to be biased (Gaudoin 2002). Moreover, this bias is a function of the sample size. This faces with a real

problem when tracing yearly quality or comparing stations with different sampling sizes.

Several methods were studied to reduce the bias:

- Linear interpolation of the empirical quantile : when all the n experimental data are different, the quantile of order i/n is $Z_{(i)} + \frac{Z_{(i+1)} - Z_{(i)}}{2}$, the one of order 0 is the half of the minimum value, and the one of order 1 is the maximum experimental value. Linear interpolation is applied between these quantile values.
- Use of a Gaussian anamorphosis linearly interpolated (Rivoirard 1994);
- Use of a Gaussian anamorphosis fitted with an Hermite polynomial function (Rivoirard 1994).

The biases of the 90% quantile estimated by these methods are not theoretically calculable because the distribution of the concentrations is unknown. That's why we propose to use simulations to evaluate them.

In case of a usual distribution, the expression of the bias is known theoretically but sometimes hard to calculate. We've calculated it for a uniform distribution in order to compare it with simulations results. Because of the similarity of results, we deduced that simulations are a good method to evaluate the bias.

Here we present results for 1000 realizations of an exponential distribution with expectation 1, samples sizes varying from 4 to 36. Because the variables are independent, it is not necessary to construct all the 365 daily values of a year to extract the samples of different sizes. Results are given in average for each different sample size. The evolution of the 90% quantile, the experimental estimation variance, the 95% confidence interval and the distribution of errors are presented Fig. 6.

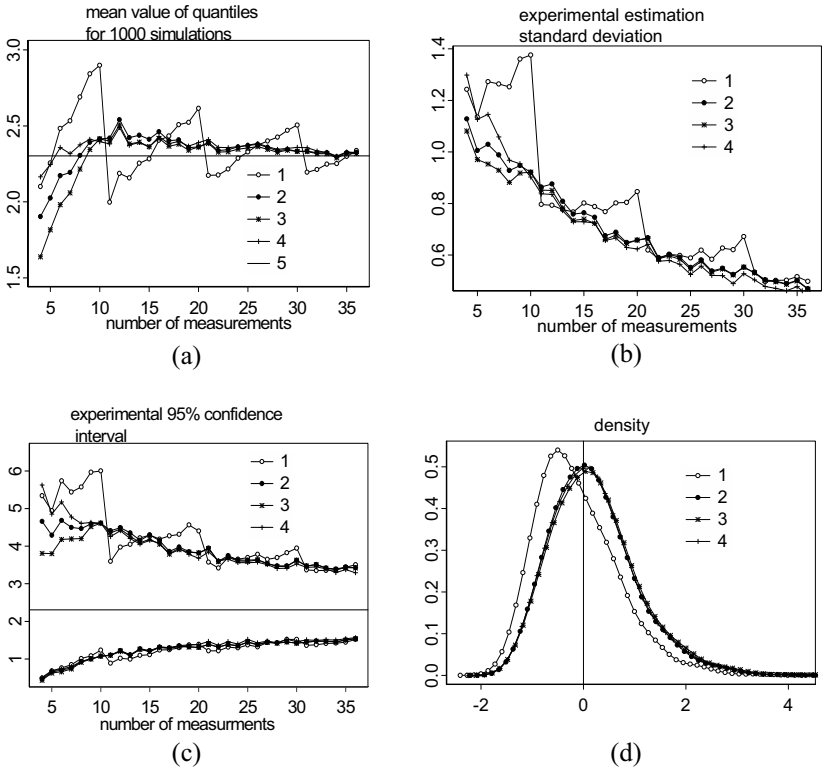


Fig. 6. Quantile estimation for independent variable, compared with theoretical value; results for 1000 simulations, as a function of sample size (a, b, c). Upper left figure **a)**: average of quantiles estimation. Upper right figure **b)**: experimental estimation standard error. Lower left figure **c)**: experimental 95% confidence interval. Lower right figure **d)**: histogram of quantile errors for samples of size 12. *Legend* : 1 represents the empirical quantile, 2 the linear interpolation of anamorphosis, 3 the hermitian interpolation of empirical quantile, 4 the hermitian interpolation of anamorphosis, 5 the real quantile.

The empirical quantile (Fig. 6a) presents a bias, strongly reduced by the other methods. Moreover, strong discontinuities for sample sizes proportional to 10 make difficult the comparison between monitoring stations with different sampling strategy. The three proposed methods are quite similar for samples whose sizes are greater than 10. They don‘t show any more discontinuities, but converge quite regularly toward the theoretical value. For this distribution, with 12 measurements, the 90% quantile is overestimated using the three interpolations functions, and clearly underestimated using the empirical quantile.

Fig. 6b makes possible to evaluate committed errors in the quantile estimation. It gives the following experimental estimation standard error as a decreasing func-

tion of the sample size. However, precision is not really satisfying even with 36 measurements because it is still representing 20% of the real quantile.

Fig. 6c presents for each sample size the interval containing 95% of the 1000 quantile estimations, approximately symmetrical around the theoretical value.

For sample of size 12, Fig. 6d shows that the distribution of the estimation errors is nearly Gaussian for the 3 interpolation functions, but not for the empirical quantile. In this last case, the errors are not centered and not symmetrical.

Other distribution examples (normal, lognormal, gamma and uniform distribution) have been tested leading to the same conclusions. Because of its equivalence to others methods, and its simplicity, the linear interpolation of quantile is advised for the estimation of 90% quantile and will always be used from now on.

4.2 Case of temporal correlation: weighted data

In presence of temporal correlation and in a limited field, we do not try any more to estimate the histogram or the quantile of the a priori distribution; this one corresponds, for an ergodic model, to the distribution of a realization in an infinite field. For a fixed realisation, the distribution to calculate is the one of a random point in the field. Because of the limited number of data per station for one year, we examine an approximate calculation of this "global" distribution.

4.2.1 Irregular sampling

The bias of empirical quantile methods on independent variables can be resolved in practice with a linear interpolation of quantiles. As for the estimation of yearly mean, in presence of temporal correlation and irregular or preferential sampling, the weighting becomes necessary to avoid bias. The weights calculated for annual mean estimation (by kriging or segment of influence) are now used in the estimation of the experimental histogram. Then, the estimated quantiles are calculated are compared below on simulations for irregular but not preferential sampling.

The following example is based on real nitrate data of the Indre River. We have proceeded with 1000 conditional simulations of 365 days respecting real measurements, using the fitted variogram presented on Fig. 7. From each simulation, samples of different sizes have been extracted, from 4 to 36 measurements a year, irregularly spaced in time. Thus we obtain 1000 samples of size 4, 1000 of size 5 etc…. We estimate the 90% quantile for each sample and for each method. Results are given in average for each different sample size and shown in Fig. 8. They are compared to quantiles calculated in average on the 1000 simulations of 365 days. That means we consider that a 90% quantile is well determined with one measurement a day.

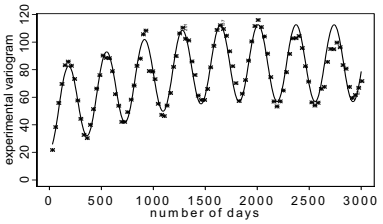


Fig. 7. Mean experimental variogram (mg^2 / L^2 NO3) calculated with monthly sampling and fitted model for the monitoring station on the Indre River. The model is composed of nugget effect (21), cosinus model (period 365.25, amplitude 56) and spherical model (range 1795, sill 35)

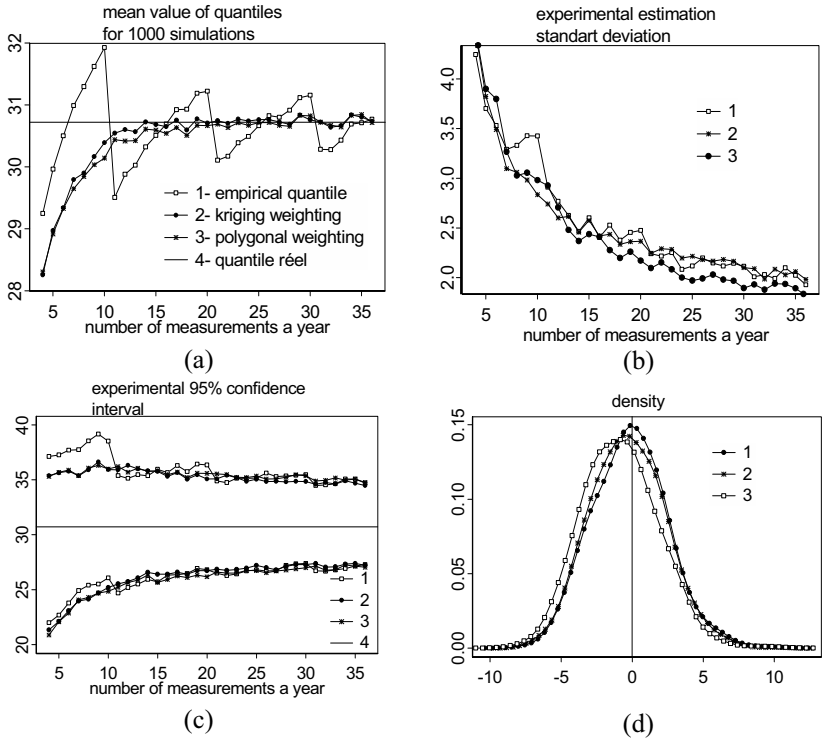


Fig. 8. Quantile estimation for temporal correlated variables, compared with the empirical quantity, for 1000 simulations. This empirical quantity corresponds to the mean, calculated on all the simulations, of the 90% quantiles of 365 values. All calculations are made by linear interpolation of quantiles. In abscissa for **a)**, **b)**, **c)**, the sample size. Upper left figure **a)**: average of quantiles estimation. Upper right figure **b)**: experimental estimation standard error. Lower left figure **c)**: experimental 95% confidence interval. Lower right figure **d)**: histogram of quantile errors for samples of size 12.

On Fig. 8a, the important bias of the empirical quantile is very well corrected by both kriging and segment of influence weighting. However, the estimation variance (Fig. 8b) is not clearly improved for the new proposed methods and is still quite important for a sample size of less than 36 measurements a year. Actually, for 36 measurements a year, errors still represent approximately 6% of the real quantile which gives an approximate 95% confidence interval (Fig. 8c) of $\pm 12\%$ around the real quantile because of the quasi normal distribution of errors (Fig. 8d). For 4 measurements a year, it reaches 14% of the real quantile, and 9% for 12 measurements a year. By simulating data respecting variability and real measurements, we can determine the necessary sample size to reach a desired precision. The theoretical estimator of a confidence interval taking into account temporal correlation would be difficult to construct. Even when random variables are independent, the theoretical announced interval (Gaudoin 2002) is not satisfying because it is limited by the higher order statistic. Simulations can be a solution to evaluate errors committed on estimations. Because results are similar with kriging and segment declustering, and because the estimation variance is difficult to assess for both methods, we propose in the future to weight measurements with segment influence segments, which is easier to automate.

4.2.2 Preferential sampling

Just as in paragraph 2.2.3, we compare statistical and kriging estimations on preferential sampling in Orléans over 10 years. In Fig. 9, on the left we present a scatter diagram between estimations on monthly sampling and preferential sampling in winter.

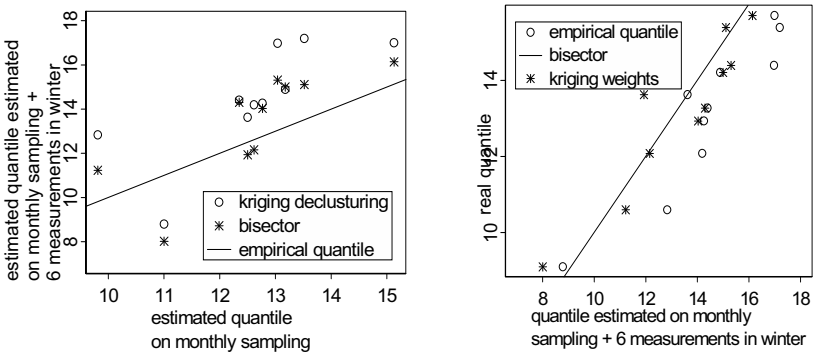


Fig. 9. On the left a scatter diagram presents statistical and geostatistical estimations of quantile calculated with 12 or 18 measurements a year. On the right, the scatter diagram compares 18 measurements estimations to real quantile value

Most of points are upper the bisector because of the bias created by the preferential sampling in winter. But kriging correct this bias (points are closer to the bisec-

tor and always lower than statistics estimations. In Fig. 9, on the right, estimations are compared to the real 90% quantile. It shows a better precision of kriging.

Conclusion

Kriging the annual mean allows to correct the bias induced by a preferential sampling of high concentration periods. The kriging variance is lower than the predicted statistical variance of the mean of independent variables, namely because of the yearly periodic component of the variogram. Associated with a linear interpolation of the experimental quantile function, the kriging weights give an empirical estimation of quantiles practically unbiased.

The segment of influence weighting can be used to simplify the calculations.

In all cases, one or two measurements a month are not sufficient for a precise estimation of the yearly 90% quantile.

Acknowledgements

We would like to thank the French minister of environment (MEDD) for financing the study and L.- C. Oudin and D. Maupas from the Agence de l'Eau Loire-Bretagne for their help.

References

Bernard-Michel C, de Fouquet C (2003) Calculs statistiques et géostatistiques pour l'évaluation de la qualité de l'eau. Rapport N-13/03/G. Ecole des Mines de Paris, Centre de géostatistique

Chilès J-P, Delfiner P (1999) Geostatistics. Modeling spatial uncertainty. Wiley series in probability and statistics

Coles S (2001) An introduction to statistical modeling of extreme values. Springer

Gaudoin O (2002) Statistiques non paramétriques, notes de cours deuxième année. ENSIMAG, http://www-lmc.imag.fr/lmc-sms/Olivier.Gaudoin/

Journel A G (1977) Geostatistique minière. Tome 1. Ecole des Mines de Paris. Centre de Géostatistique

Matheron G (1970) La théorie des variables régionalisées, et des applications. Les cahiers du Centre de Morphologie Mathématique. Ecole des Mines de Paris. Centre de géostatistique

Payne M R Farm (1993) Waste and nitrate pollution. Agriculture and the environment. John Gareth Jones. Ellis Horwood series in environmental management

Rivoirard J (1994) Introduction to Disjunctive Kriging and Non-linear Geostatistics. Clarendon Press, Oxford

Saporta G (1990) Probabilités, analyse de données et statistiques. Technip

Stochastic simulation of rainfall using a space-time geostatistical algorithm

J. A. Almeida and M. Lopes

CIGA/FCT-UNL, Monte da Caparica, 2829-516 Caparica, Portugal,
e-mail : ja@fct.unl.pt

1 Introduction

Erosion of soil by water is a complex weathering phenomenon through which surface soils are disaggregated and transported. Current human activity including construction, deforestation and reshaping of the surface accelerates erosion of the soil by rain and water streams and constitutes a growing problem, leading for instance to desertification. It is increasingly considered a research priority, particularly in sensitive areas with deep slopes, heavy rains and accelerated deforestation, natural or anthropogenic.

Rainfall is one of the extrinsic factors that strongly affects the degree of erosion of a given site. The erosive capability of rain depends on several factors, but these can be summarised as kinetic energy and intensity. An approach involving kinetic energy is beyond the scope of this work but it is strongly related to intensity of the rain. Intensity, as expressed by the amount that falls for a unit of area and time, constitutes an important measure that affects the aggressiveness of heavy rain. It is measured by weather stations networks.

All indices of soil erosion by water must take rainfall intensity into account (Loureiro and Coutinho 2001). For instance, Wischmeier (1959) proposed an index of aggressiveness that depends exclusively on rainfall intensity, measured over a period of 30 minutes.

The main objective of this study is to present a space-time simulation methodology to enable the construction of sets of rainfall images, auto-correlated by global correlation measures. Unlike estimation methods whose main objective is the construction of an average map based on spatial continuity measures, stochastic simulation methods consist of a set of methods able to generate numerical models or realisations of the spatial distribution of a categorical or numerical variable. The set of outputs consists of equiprobable images in the sense that they have the same probability of occurrence (Journel and Alabert 1989, Goovaerts 1997, 2000). These images obtained by simulation constitute an essential tool for the spatial analysis of a specific phenomenon, such as the probability of occurrence of extreme scenarios in each location and evaluation of the space of uncertainty.

The set of simulated rainfall equiprobable images, which gives the space-time uncertainty, fed a deterministic surface dispersion model in a Geographical Information System (GIS) to forecast the surface runoff of the selected area. Results from this space-time model - stochastic images of spatial dispersion of rainfall - can be used to visualise extreme situations of the hydrological behaviour of the watershed and local critical areas.

2 Methodology

Using climatic information from one year (monthly accumulated rainfall) based on measurements from 36 weather stations in the Algarve, southern Portugal, this paper presents a combined methodology to simulate the hydrological behaviour of watersheds. A stochastic simulation technique is applied to characterise the space-time dispersion of rainfall for a given period of time and to characterise the uncertainty summarised for a specific watershed.

Considering rainfall measurements $z(t_j, x_i)$ from n weather stations (x_i, i=1,…n) and N time steps (t_j, j=1,…N), a direct sequential cosimulation (CoDSS) algorithm is used (Soares 2001), which calls for the local estimates of rainfall in month t_j at location x_u, $z(t_j, x_u)$ based on the rainfall of n neighbourhood values for the same month $z(t_j, x_\alpha)$ – the primary variable – and on the digital elevation model (DEM) at location x_u, $z_1(x_u)$, and the previously simulated rainfall images ($z_2(t_k, x_u)$, k=1,…t_j -1) – the secondary variables. This is a collocated cokriging procedure with a multiple set of secondary variables. In order to avoid a large number of redundant secondary images, principal component analysis (PCA) was used, which reduces the number of dimensions in the data for a maximum one or two principal components (PC) images, keeping most of the variance (N_{PC}, with $N_{PC} << t_j$-1).

This space-time geostatistical methodology for the stochastic simulation of rainfall can be summarised in the following sequence of steps:

1. Calculation of basic statistics and correlation analysis for all measurements: elevation of each weather station $z_1(x_i)$ and 12 monthly rainfall measurements $z_2(t_j, x_i)$, with i=1,n and j=1,N;
2. Use of PCA to calculate the PC corresponding to 11 datasets of monthly rainfall: 1) months t_1 and t_2; 2) months t_1, t_2 and t_3, … and finally 11) months t_1, t_2, t_3, …, t_{11}. For each run, selection of the PC with the highest eigenvalues above one (PC representing the amount of variance above the original variables, previously standardised).
3. For the 11 datasets considered in the previous step, calculation of the global measures of correlation between each selected PC and the elevation of the weather station;
4. Calculation of variograms and fitting theoretical models to monthly rainfall measurements and the selected PC as calculated with the 11 datasets of monthly rainfall;

5. Simulation of monthly rainfall images for the whole area by using the CoDSS algorithm for all time steps ($t_{i,}$ i=1, …12), taking into account the following information:
   e.1) Month t_1: rainfall measurements of this specific month ($z(t_1,x_i)$,i=1,n) and the DEM ($z_1(x_u)$) as a secondary information image.
   e.2) Month t_2: rainfall measurements of this specific month ($z(t_2,x_i)$,i=1,n) and two images as secondary information, the DEM $z_1(x_u)$ and the simulated images for month t_1: ($z^s(t_1, x_u)$).
   e.3) Month t_3: rainfall measurements of this specific month ($z(t_3,x_i)$,i=1,n) and two sets of images as secondary information, the DEM $z_2(x_u)$ and the selected PC images $z_{PC_n}(x_u)$, built with data from previous months (t_1 and t_2).
   …
   e.12) Month t_{12}: rainfall measurements of this specific month ($z(t_{12},x_i)$,i=1,n) and two sets of images as secondary information, the DEM $z_2(x_u)$ and the selected PC images $z_{PC_n}(x_u)$, built with data from previous months (t_1 through t_{11}).
6. Using the DEM map, derivation of the flow direction and flow accumulation map for the entire area and identification of the Arade watershed;
7. For each of the simulated monthly rainfall images $z_l^s(t_j, x_u), j = 1,...12$, construction of the corresponding accumulation rainfall maps (one for each simulated scenario l);
8. Evaluation of the uncertainty of the rainfall model. Calculation of the maximum accumulation for each simulated scenario and displaying of comparative monthly results using box-plots.

3 Direct sequential cosimulation

To obtain the required simulation images, the CoDSS algorithm was used. This simulation method with a set of secondary variables is an extension of the algorithm proposed by Soares (2001) and can be summarised as follows (Almeida *et al.* 2002):

1. Define a random path visiting each node of a regular grid of nodes.
2. At each node x_u, simulate the value $z^s(t_j,x_u)$ using the CoDSS algorithm:

   - Identify the local mean and variance of $z(x)$, $z(t_j,x_u)$* and $\sigma^2_{sk}(t_j,x_u)$, using the simple collocated kriging estimator with a multiple set of secondary variables:

$$z(t_j,x_u) = \sum_{\alpha=1}^{n} \lambda_\alpha^{t_j} z(t_j,x_\alpha) + \lambda^{DEM} z_1(x_u) + \sum_{i=1}^{N_{PC}} \lambda^{PC_i} z_{PC_i}(x_u)$$

$$N_{PC} = 1, number\ of\ selected\ PC$$

Using the matrix formalism, the simple collocated kriging system with a multiple set of N secondary variables ($N = N_{PC}+1$) is defined as follows:

$$\begin{bmatrix} 1 & C_{12}^{t_j} & \cdots & C_{1n}^{t_j} & C_{1u}^{t_j DEM} & C_{1u}^{t_j DEM} & \cdots & C_{1u}^{t_j DEM} \\ C_{21}^{t_j} & 1 & \cdots & C_{2n}^{t_j} & C_{2u}^{t_j PC_1} & C_{2u}^{t_j PC_1} & \cdots & C_{2u}^{t_j PC_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{n1}^{t_j} & C_{n2}^{t_j} & \cdots & 1 & C_{nu}^{t_j PC_{N_{PC}}} & C_{nu}^{t_j PC_{N_{PC}}} & \cdots & C_{nu}^{t_j PC_{N_{PC}}} \\ \hline C_{u1}^{t_j DEM} & C_{u2}^{t_j DEM} & \cdots & C_{un}^{t_j DEM} & 1 & C_u^{\mathrm{DEM}\ PC_1} & \cdots & C_u^{\mathrm{DEM}\ PC_{N_{PC}}} \\ C_{u1}^{t_j PC_1} & C_{u2}^{t_j PC_1} & \cdots & C_{un}^{t_j PC_1} & C_u^{PC_1\ \mathrm{DEM}} & 1 & \cdots & C_u^{PC_1\ PC_{N_{PC}}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{u1}^{t_j PC_{N_{PC}}} & C_{u2}^{t_j PC_{N_{PC}}} & \cdots & C_{un}^{t_j PC_{N_{PC}}} & C_u^{PC_{N_{PC}} DEM} & C_u^{PC_{N_{PC}} PC_1} & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1^{t_j} \\ \lambda_2^{t_j} \\ \vdots \\ \lambda_n^{t_j} \\ \hline \lambda^{DEM} \\ \lambda^{PC_1} \\ \vdots \\ \lambda^{PC_{N_{PC}}} \end{bmatrix} = \begin{bmatrix} C_{1u}^{t_j} \\ C_{2u}^{t_j} \\ \vdots \\ C_{nu}^{t_j} \\ \hline C_u^{t_j DEM} \\ C_u^{t_j PC_1} \\ \vdots \\ C_u^{t_j PC_{N_{PC}}} \end{bmatrix}$$

where:

$C_{\alpha\beta}^{t_j}$ - Covariance of rainfall between samples at locations x_α and x_β in time period t_j

$C_{u\alpha}^{t_j DEM}$ - Cross-covariance between DEM at location x_α and rainfall in time period t_j at location to estimate rainfall x_u

$C_{u\alpha}^{t_j PC_i}$ - Cross-covariance between PC_i at location x_α and rainfall in time period t_j at location to estimate rainfall x_u

$C_u^{PC_i PC_j}$ - Cross-covariance between PC_i and PC_j at location to estimate rainfall x_u (equal to zero)

$C_u^{PC_i DEM}$ - Cross-covariance between PC_i and DEM at location to estimate rainfall x_u

$\lambda_\alpha^{t_j}$, λ_u^{DEM} and $\lambda_u^{PC_i}$ - Weights of primary information, DEM and PC_i

$C_{\alpha u}^{t_j}$ - Covariance of rainfall between samples at locations x_α and location to estimate rainfall x_u

$C_u^{t_j DEM}$ - Cross-covariance between DEM and rainfall in time period t_j at location to estimate rainfall

with $\alpha = 1\ldots n$; $\beta = 1\ldots n$;

- Locally resample the histogram of $z(x_u)$, for instance using a normal score transform (φ) of the primary variable $z(x)$, and calculate $y(x_u)^* = \varphi(z(t_j, x_u)^*)$;
- Draw a value p from a uniform distribution $U(0,1)$;
- Generate a value y^s from $G(y(x_u)^*, \sigma^2_{sk}(x_u))$: $y^s = G^{-1}(y(x_u)^*, \sigma^2_{sk}(x_u), p)$;
- Return the simulated value $z_1^s(x_u) = \varphi^{-1}(y^s)$ of the primary variable.

3. Loop until all nodes are simulated.

Assuming a Markov-type approximation, the cross-covariance function can be calculated using the following relation in terms of covariance or correlograms (Almeida and Journel 1994), which calls only for inference of the primary variable covariance function and the correlation index $\rho_{12}(0)$ between the primary and secondary variable.

The set of simulated rainfall images obtained for the entire area in time period t_j enables calculation of extreme scenarios and uncertainty assessment.

Regarding the application of this extended kriging system, using more than one PC, all cross-covariances between PC are equal to zero, due to normality of all coordinates. This constitutes a significant simplification, transforming cross-covariances to zero.

4 Case study

In this case study the simulation of rainfall maps using the proposed methodology is presented. Data were collected from 36 weather stations in the Algarve area (southern Portugal) and show 12 months (from October to September, 2000). All rainfall values consist of total monthly quantity in mm.

4.1 Basic statistics and correlation between measures

The studied area of the Algarve is located in the southernmost part of Portugal, which includes the Arade watershed. For topographic information, a DEM of the entire area in GIS raster format is used. Rainfall is recorded by monthly measurements in a network of 36 weather stations (Fig. 1). The 12 months of data from October to September 2000 were used.

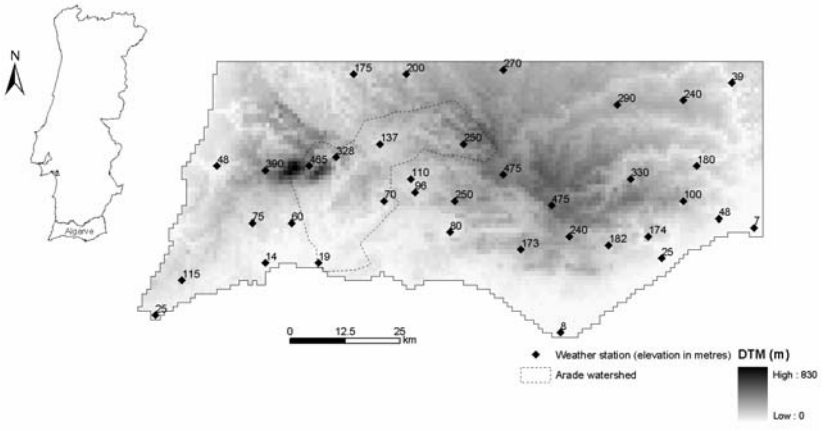


Fig. 1. DEM of the Algarve area in raster format with 1-kilometre spatial resolution and location of the 36 weather stations

Univariate statistics and correlation coefficients were calculated for all data considered. The results are summarised in table 1 and 2.

With the exception of the months of July and August, a strong correlation is observed between rainfall measurements and the elevation of the station (always higher than 0.69 and with a maximum of 0.83 in June). Correlation between consecutive months is usually very high, with the exception of July and August, which are confirmed as atypical.

Table 1. Basic statistics of collected data: monthly rainfall (mm) and elevation *Z* (m)

	Z	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Mean	171,2	59,6	79,3	96,3	69,9	58,0	32,7	42,1	21,0	8,1	1,0	2,0	12,1
Var.	18511	261	485	1133	609	663	130	300	97	14	1	1	17
Min.	7	32.2	39.4	44.4	37.6	27.4	17.2	17.9	7.2	3.2	0	0	5.6
Max.	475	111	148,7	183	137	146,6	73,9	105,9	54,6	16,8	3,2	5,3	22,8

Table 2. Correlation statistics between collected data

	Z	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Z	1,000	0,758	0,724	0,710	0,693	0,753	0,798	0,741	0,830	0,830	0,388	0,328	0,747
Oct		1,000	0,894	0,814	0,880	0,918	0,887	0,908	0,898	0,844	0,123	0,371	0,752
Nov			1,000	0,913	0,935	0,956	0,886	0,888	0,847	0,793	0,244	0,346	0,651
Dec				1,000	0,938	0,927	0,812	0,780	0,780	0,717	0,338	0,310	0,502
Jan					1,000	0,971	0,857	0,855	0,837	0,780	0,227	0,304	0,576
Feb						1,000	0,918	0,907	0,874	0,806	0,218	0,418	0,644
Mar							1,000	0,958	0,892	0,802	0,202	0,466	0,710
Apr								1,000	0,911	0,820	0,204	0,417	0,708
May									1,000	0,885	0,330	0,362	0,747
Jun										1,000	0,380	0,337	0,771
Jul											1,000	-0,038	0,259
Aug												1,000	0,456
Sep													1,000

5 Calculation of PC, correlation analysis and variogram models

Following the proposed methodology, the PCA algorithm was used to synthesise the number of variables (past monthly measurements) in each step. For instance, to simulate rainfall in May, monthly measurements from October to April were synthesised with PCA and only one PC was selected; to simulate rainfall in August (atypical month), measurements from October to July were synthesised with PCA, and in this case two PC were selected. All PC with a corresponding eigenvalue equal to or greater than one were selected. Table 3 represents the main results from PCA. Table 4 shows global correlations between selected PC and the monthly rainfall and elevation.

Table 3. Summary of PCA results

Months	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Number of PC	-	1	1	1	1	1	1	1	1	2	2	2
Cumulative percentage of total variance		94,7	91,6	92,2	93,2	91,7	91,1	90,2	88,4	90,3	84,6	82,2

Table 4. Correlations between selected PC and monthly rainfall and elevation Z

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Z vs rainfall	0,76	0,72	0,71	0,69	0,75	0,80	0,74	0,83	0,83	0,39	0,33	0,75
PC1 vs rainfall	-	-	0,89	0,96	0,98	0,90	0,92	0,90	0,85	0,27	0,39	0,73
PC2 vs rainfall	-	-	-	-	-	-	-	-	-	-	0,16	-0,04
Z vs PC1	-	-	0,76	0,76	0,75	0,75	0,77	0,77	0,79	0,81	0,81	0,81
Z vs PC2	-	-	-	-	-	-	-	-	-	-	-0,17	0,15

Spatial continuity structures of rainfall and PC are measured through spatial variograms calculated for each monthly period. Isotropic spherical models fit all experimental variograms, with ranges between 17 and 30 km (table 5). In Fig. 2 variograms for October and November are presented. Differences in behaviour between some months, notably July and August, are not shown in the variograms.

Table 5. Models of variograms fitted to rainfall measures and PC

Months		Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Main variable	C1	261	485	1133	609	663	130	300	97	14	0,7	1,4	17
	a1 (km)	30	23	27	22	22	17	17	21	22	19	30	20
	Model	Sph											
PC1 variable	C1	-	0,947	0,916	0,922	0,932	0,917	0,911	0,902	0,884	0,804	0,746	0,730
	a1 (km)	-	30	30	30	30	30	25	25	25	25	25	20
	Model	-	Sph										
PC2 variable	C1	-	-	-	-	-	-	-	-	-	0,1	0,1	0,1
	a1 (km)	-	-	-	-	-	-	-	-	-	22	22	22
	Model	-	-	-	-	-	-	-	-	-	Sph		

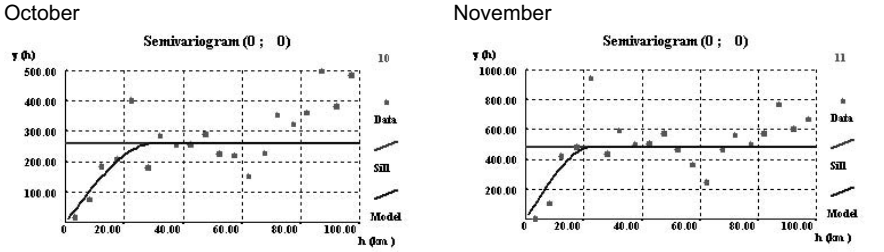


Fig. 2. Experimental variograms and theoretical models fitted to this sequence of winter months

6 Simulation of monthly rainfall images

CoDSS algorithm was applied to obtain the set of 10 stochastic rainfall images. Each unit cell represents a 1x1 km square. The total area of the Algarve is 6682 km^2, including all the Arade watershed (795 km^2).

The earliest month, October, is the first to be simulated: for this month, only rainfall measurements from the same month (hard data) and the DEM are used (soft data). The next month to be simulated is November, conditioned to the October results: for this month rainfall measurements from this month are considered as hard data and the DEM and each of the 10 simulated images from October as soft data.

Each of the remaining months must be conditioned successively to the data of the previous months. For instance, the simulation of July takes into account data

from the same month as hard data and the DTM and simulated images of the PC1, built with data from the previous months.

At the end, each sequence of 12 simulated images constitutes an equiprobable scenario conditioned to all rainfall measurements, corresponding histograms, spatial continuity measures and correlation coefficients, among months and between months and the DEM. It is important to note that in this type of simulation model integrating hard data (rainfall measures) and soft data (elevation and previous months data), the influence of the samples prevails over the soft data in the proximity of the samples and the soft information prevails outside the influence of the samples.

In Fig. 3 and 4 sets of 4 images are shown for the months of December (high rainfall) and August (low rainfall). Images corresponding to high rainfall months show that the results are strongly conditioned by elevation, and as a consequence the differences between them are small – the model carries very low uncertainty. On the other hand, in August, the correlation with the DEM is fair. This is a typical situation of localised and random rainfall (summer thunderstorms). The images obtained are mostly conditioned by the rainfall measurements of this month, and so the results show greater uncertainty.

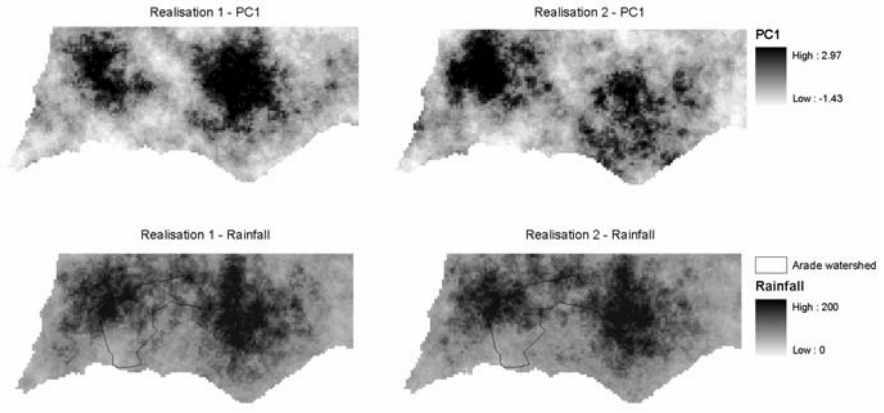


Fig. 3. Two simulated images of PC1 and corresponding rainfall images for December

7 Accumulation uncertainty within the Arade watershed

The final stage of this work is the overlapping of all simulated rainfall images with the flow direction map in order to evaluate the monthly uncertainty of the total accumulation in the Arade watershed. For illustrative purposes, Fig. 5 represents the local accumulation assuming constant rainfall, which equals one unit.

A flow direction map enables monthly calculations of global and local accumulated watershed rainfall. Based on the simulated images, 10 equiprobable out-

comes of total watershed accumulation for each month were obtained; this set of values defines the space of uncertainty and was graphically plotted in a set of 12 box-plots (Fig. 6). Table 6 represents maximum and minimum accumulation, average, difference (maximum-minimum) and difference in percentage of the average.

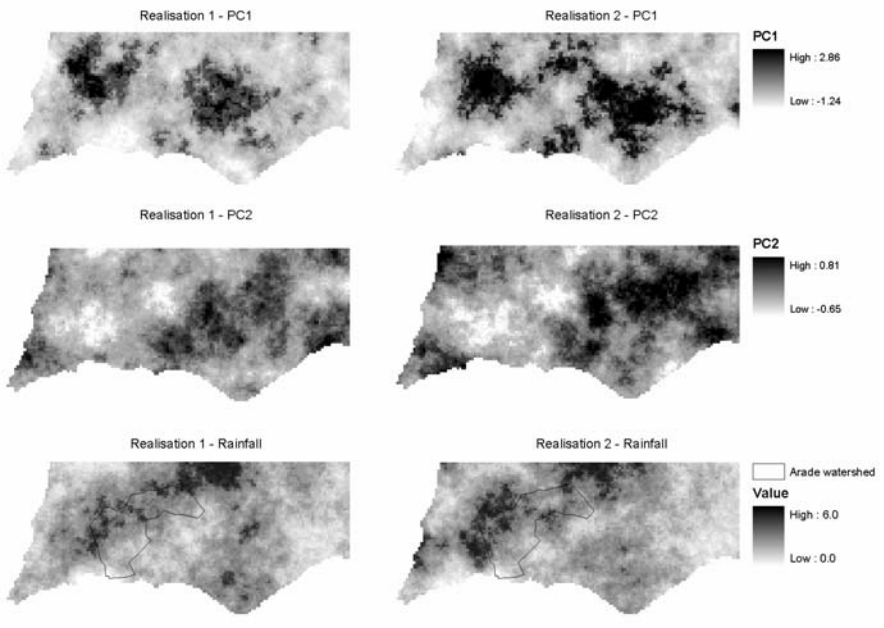


Fig. 4. Two simulated images of PC1 and PC2 and corresponding rainfall images for August

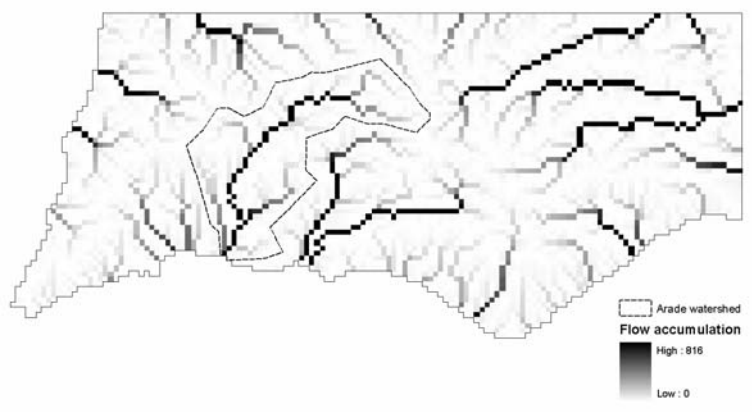


Fig. 5. Example of an accumulation map assuming constant rainfall

Based on these results it is possible to confirm that December is the month with highest rainfall; April is the rainiest month in spring, and the summer months of July and August are the least rainy and the little rain that falls is very dispersed, leading to uncertainty. For instance, the rainiest months between October and February show comparatively low differences in percentage terms; February presents a difference between the total accumulation counted in the Arade watershed of 1,9%. In contrast, July exhibits huge variability (uncertainty measure), higher than 90%, and August with 28,1% appears in second place. Months with regimes of localised thunderstorms give the model an enormous uncertainty due to the small conditioning data effect, namely the digital terrain model and the previous months.

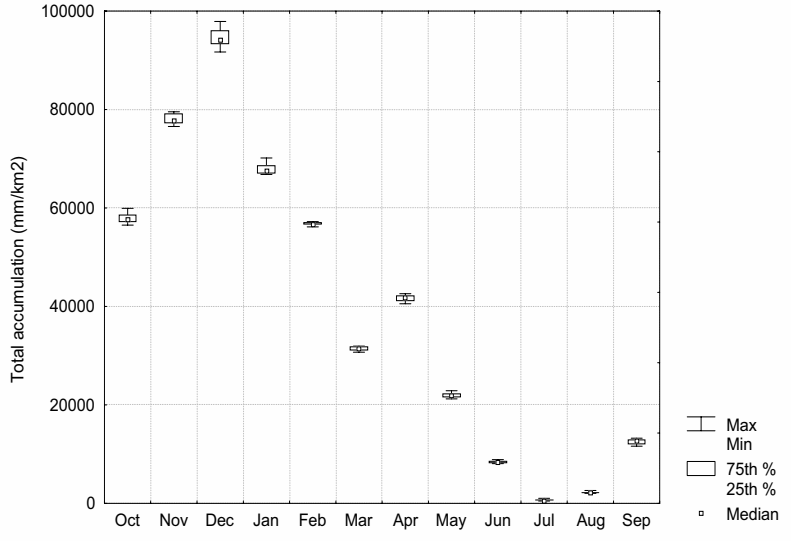


Fig. 6. Representation of the global uncertainty of the simulated rainfall images, converted into total accumulation of the Arade river hydrographical basin

Table 6. Summary of global accumulation amounts (mm/km^2) in the Arade watershed

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Max.	59902	79527	97871	70142	57228	31920	42577	22875	8856	1003	2595	13211
Min.	56518	76539	91685	66779	56162	30693	40523	21205	8040	446	1985	11578
Average	57941	78073	94439	67916	56772	31417	41590	21904	8370	588	2174	12528
Range	3383	2988	6186	3362	1065	1227	2054	1670	816	557	610	1633
Diff. %	5,8	3,8	6,6	5,0	1,9	3,9	4,9	7,6	9,7	94,6	28,1	13,0

8 Final remarks

This paper presents a methodological sequence to create a space-time model of rainfall applied to describe the hydrological behaviour of the Algarve area. Based on the stochastic imaging of monthly rainfall, an important tool to evaluate erosion hazard can be constructed. In fact, after overlapping with the local flow direction, local distribution and extreme scenarios of surface runoff can be predicted for particular meteorological conditions.

In the construction of a spatial-temporal rainfall model, it is essential to impose in the model all of the correlations given by experimental data, namely among months and between these and elevation measures. The construction of rainfall simulated scenarios presents great potential, particularly as input data for erosion maps. As is known, greater or lesser erosion of soil depends on the occurrence of extreme situations of rainfall in which these kinds of models, reproducing extreme scenarios, are the most appropriate.

It is possible to verify that geostatistical simulation tools and GIS functions complement each other in analysing data and deriving new spatial information. The result is a reliable model able to make an important contribution to a regional erosion map, even one using a smaller temporal unit.

References

Almeida A, Journel A (1994) Joint simulation of multiple variables with a Markov-type corregionalization model. Mathematical Geology 26(5): 565-588

Almeida J, Santos E, Bio A (2002) Use of geostatistical methods to characterize population and recovery of Iberian hare in Portugal. In Soares A, Gomez-Hernandez J, Froidevaux R (eds), geoENV IV. Kluwer Academic Publishers, Dordrecht, 127-138

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, New York

Goovaerts P (2000) Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology 228: 113-129

Journel AG, Alabert FG (1989) Non Gaussian data expansion in the earth sciences. Terra Nova 1: 123-134

Loureiro NS, Coutinho MA (2001) A new procedure to estimate the RUSLE El30 index, based on monthly rainfall data and applied to the Algarve region, Portugal. Journal of Hydrology 250: 12-18

Soares A (2001) Direct Sequential Simulation and Cosimulation. Mathematical Geology 33(8): 911-926

Wischmeier W H (1959) A rainfall erosivity index for a universal soil loss equation. Soil Sci. Soc. Amer. Proc. 23: 246-249.

Inferring the lateral subsurface correlation structure from georadar data: Methodological background and experimental evidence

B. Dafflon, J. Tronicke and K. Holliger

Institute of Geophysics, Swiss Federal Institute of Technology (ETH),
ETH-Hoenggerberg, CH-8093 Zurich, Switzerland

1 Introduction

Knowledge of the spatial correlation structure of the hydraulic properties of the shallow subsurface is a key prerequisite for the detailed characterization of aquifers in general and for the realistic simulation of flow and transport phenomena in particular (e.g., Rubin 2003). Whereas the vertical component of the correlation structure is often well constrained from borehole information (e.g., Ritzi *et al.* 1994), the nature of the lateral component of the correlation structure is generally largely unknown. The primary reasons for this are that the spacings between individual boreholes tend to be too large to allow for a reliable interpolation and that traditional hydrological methods, such as tracer and pumping tests, tend to capture the gross average properties of the entire probed region.

Georadar data are highly sensitive to variations in the water-saturated porosity structure (e.g., Davis and Annan 1989), which in alluvial aquifers can be regarded as a proxy for the hydraulic conductivity structure (e.g., Oldenborger *et al.* 2003). This opens the perspective to extract the lateral correlation structure of the shallow subsurface from densely sampled and adequately processed georadar data acquired along the earth's surface. Rea and Knight (1998) were the first to pursue such an approach by comparing the overall correlation structure of a pertinent outcrop image with that of corresponding georadar data. The assumption that the lateral correlation of surface georadar data is directly related to that of a photographic image of an outcrop (Fig. 1) is based on the observation that discernible lithological variations tend to be primarily related to variations in grain size and thus to variations in porosity and water content (Heinz *et al.* 2003).

Here, we complement and extend the seminal work of Rea and Knight (1998) by (i) clarifying the methodological foundations, (ii) using a realistic and highly versatile autocovariance model, whose parameters can be readily interpreted and used for the generation of corresponding synthetic models of the subsurface structure and (iii) assessing the robustness of this correlation analysis with regard to the processing of the georadar data and the corresponding outcrop images.

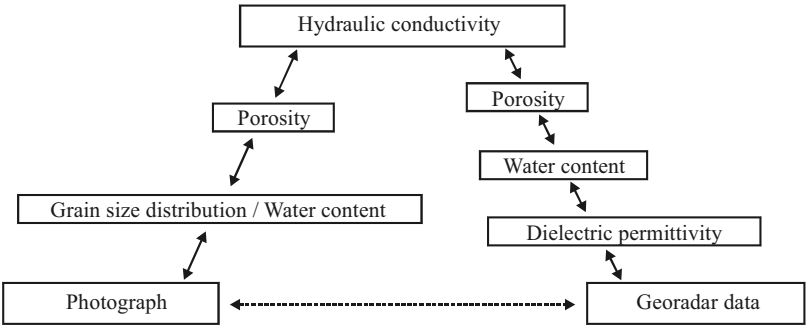


Fig. 1. The assumption that the lateral correlation structures of a pertinent outcrop photograph and corresponding surface georadar data are interrelated is primarily based on the mutual sensitivity of these measurements to variations in the water content.

2 Methodological Foundations

In analogy to perfectly imaged zero-offset seismic reflection data, surface georadar data can be approximated as (Claerbout 1985)

$$s(x,t) = w(t) * r(x,t) \tilde{\propto} w(t) * \Delta\sqrt{\varepsilon_r(x,t)} \tilde{\propto} w(t) * \Delta\Theta(x,t)\,, \tag{1}$$

where x denotes the lateral distance along the profile, t the two-way travel time, w the wavelet emitted by the transmitter antenna, r the distribution of the reflection coefficients in the subsurface, ε_r the relative dielectric permittivity, Θ the water content, Δ the relative change and the asterisks the convolution over the time axis (Davis and Annan 1989; Neil 2004). The above relation between the dielectric permittivity and the water content is based on the common assumption of a two-component mixing model (Wharton *et al.* 1980). The lateral autocovariance function of the surface georadar section s is thus given by

$$C_{ss}(r_x,t) = w(t) * C_{rr}(r_x,t) \tilde{\propto} w(t) * C_{ww}(r_x,t)\,, \tag{2}$$

where r_x is the lag along the horizontal direction and C_{ss}, C_{rr} and C_{ww} denote the lateral autocovariance functions of the georadar section s, the reflection coefficient distribution r and changes in the water content $\Delta\Theta$, respectively.

Eq. 2 thus illustrates that the lateral correlation of surface georadar data is directly related to changes in the square-root of the dielectric permittivity structure, which in turn can be directly related to changes in the water content (Wharton et al 1980; Knight 2001). Eq. 1-2 also indicate that the vertical correlation of the georadar data is dominated by the source signal $w(t)$ emitted by the transmitter antenna and hence cannot be directly related to the vertical correlation of the subsurface structure. The band-limited nature of the source signal $w(t)$ may also influence our estimates of the lateral correlation structure C_{ss} in that a vertical off-

set within a horizontal reflector only comes into effect, if the magnitude of this offset is within the vertical resolution of $w(t)$ (i.e., larger than approximately one eighth to one quarter of the dominant wavelength). This implies that correlation estimates from surface georadar data C_{ss} tend to overestimate the actual lateral correlation of the subsurface structure and that this bias increases with decreasing dominant frequency and thus decreasing vertical resolution, of the emitted source signal $w(t)$.

3 Data Acquisition and Processing

The outcrop images and the corresponding georadar data have been collected at a gravel quarry in Hüntwangen in northern Switzerland operated by Holcim AG (Fig. 2). The sedimentary inventory consists of gravel- and sand-dominated braided stream deposits. At the site considered in this study, these glacio-fluvial sediments are unconsolidated and predominantly horizontally layered (Fig. 2a). The local geology and its expression in georadar images have been described by Huggenberger (1993) and Beres *et al.* (1999).

3.1 Digital Outcrop Photograph

The quarry face was photographed from a distance of ~135 m using a digital camera with a formal resolution of 2 million pixels. Several laterally overlapping images were taken and no zoom was applied in order to minimize distortions in the final composite image of the well-exposed central part of the cliff (Fig. 2a and 2b). Surveyed markers placed along the upper edge of the cliff as well as within the quarry face allowed for an accurate positioning and scaling of the photographs in reference to the georadar profile, for assembling individual photographs into a composite image, for constraining the spatial resolution and for the detecting and analyzing any remaining distortions in the final image. Using this approach, we found the resulting image to have a resolution of ~3 cm and a distortion of ~2%.

Prior to a geostatistical analysis, the original 3-D image matrix of the digital outcrop photograph in conventional color or RGB ("red-green-blue") format had to be transformed into a corresponding 2-D matrix. The goal was to perform this transformation with as little loss of information as possible. To this end, we have compared the average horizontal and vertical autocovariance functions of all three color channels (red, green and blue) of the original image with those of the transformed 2-D image matrices. This comparison was facilitated by the fact that, after normalization, the autocovariance functions of the three channels turned out to be quite similar. Several methods have been tested to convert the original color image to a grayscale image. All of these approaches were based on taking some form of weighted averages of the three color channels and provided rather similar results. The corresponding autocovariance functions compared favorably with those of the individual color channels.

Primarily for reasons of robustness and consistency, we chose to convert the original color image to a grayscale image through the so-called RGB-mean method, which simply replaces the values of the color channels by their arithmetic average. In addition, to this grayscale conversion, we also explored the binary black-and-white approach pursued by Rea and Knight (1998). We found this approach to be less robust than the conversion to RGB-mean format. The resulting autocovariance functions systematically exhibited significant deviations from those of the original color channels and also depended notably on the rather arbitrary choice of the threshold value governing the binary conversion to black-and-white. An additional advantage of the grayscale approach compared to the binary-and-white approach was that the corresponding probability density function was closer to a normal distribution and hence that the corresponding images were more adequately characterized by first and second statistical moments (mean and autocovariance function).

3.2 Surface Georadar Data

The surface georadar data were acquired along the upper edge of the quarry face at a largely uniform lateral distance of ~5 m from the cliff. We used commercial 200 MHz antennas with a constant spacing between the transmitter and receiver antennas of 1 m. Using a sampling interval of 0.2 ns and 64 vertical stacks, this configuration was moved at 0.2 m increments along a straight and topographically even profile with a total length of ~75 m.

Surface georadar data acquired in this bi-static mode can be regarded as the electromagnetic version of "echo-sound" measurements: the transmitter emits a compact pulsed signal, which travels into the subsurface, is reflected from changes in the material properties and is then recorded by the quasi-coincident receiver antenna. To a first approximation, the thus recorded signals can be regarded to represent an image of the variation of the dielectric permittivity structure in the shallow subsurface, which in turn is primarily governed by variations in the water content. Comprehensive reviews of the methodological background and the applications of surface georadar measurements are provided by Davis and Annan (1989) and Neil (2004).

The basic processing of the georadar comprised a "dewow" filter to remove any dc-current components in the data, a time-zero adjustment, a compensation of the decay of the amplitudes with increasing travel time due to the geometrical spreading of the wave front and attenuation effects and the application of a 350/500 MHz low-pass filter. The recovery of the amplitudes was achieved by dividing each georadar trace by the average trace envelope of the entire profile. We also acquired two common-midpoint (CMP) gathers along the profile to obtain an estimate of the large-scale electromagnetic velocity structure based on the hyperbolic curvature of the reflections as a function of antenna separation. This analysis indicated that the velocity in the probed region was remarkably uniform. For this reason, the inferred average velocity of ~0.11 m/ns was used to convert georadar data from two-way travel time to depth (Fig. 2c).

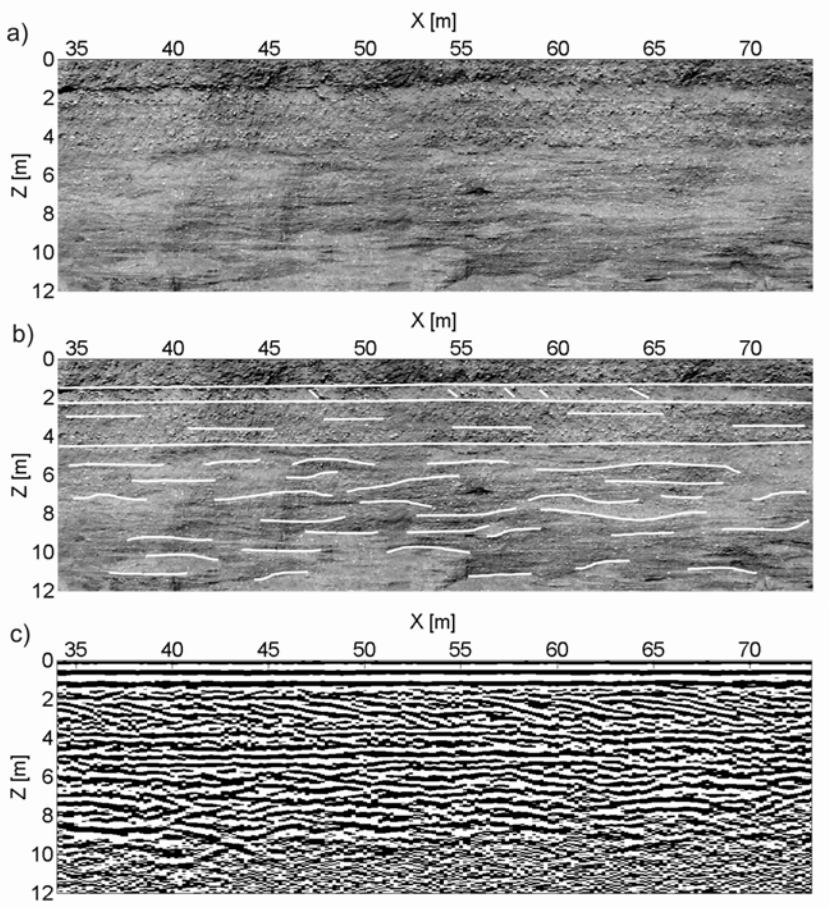


Fig. 2. a) Digital photograph of the well-exposed central part of the gravel quarry face in RGB-mean format. **b)** Same as a) with the most prominent lithological features marked by white lines. **c)** Corresponding depth-converted georadar profile acquired along the upper edge of the cliff using 200 MHz antennas. Note the far-reaching consistency between the photographic image of the outcrop and the georadar section.

4 Lateral Correlation Analysis

In the following, the lateral correlation structures of the digital outcrop images in RGB-mean format (Fig. 2a) and surface georadar data (Fig. 2c) are interpreted based on the band-limited scale-invariant or "fractal" von Kármán autocovariance model (von Kármán 1948):

$$C(r_x) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{r_x}{a_x}\right)^{\nu} K_{\nu}\left(\frac{r_x}{a_x}\right) \tag{3}$$

where r_x is the lag vector in the lateral direction, a_x is the horizontal correlation length, σ is the standard deviation, Γ is the gamma function and K_ν is the modified Bessel function of the second kind of order $0 \leq \nu \leq 1$. ν is related to the "Hausdorff" fractal dimension D through $D = E + 1 - \nu$ with E denoting the underlying Euclidean dimension (Goff and Jordan 1988). The correlation length corresponds approximately to the outer range of scale-invariance. The von Kármán autocovariance model is highly versatile and has been successfully used to characterize a wide variety of scientific data. A detailed description of this autocovariance model is given in Sidler and Holliger (this volume).

For the analysis of the lateral correlation structure, we have subdivided the database (Fig. 2a and 2c) into several subsets. The main reason for this approach was to reduce the lateral geological variations within the data to be analyzed and thus to enhance their statistical stationarity in the lateral direction. The results obtained for the various subsets were found to be internally consistent and hence shall be illustrated and discussed using a typical subset of the outcrop image shown in Fig. 3a. Please note that the uppermost 2 m have been excluded from all analyses. The reason for this is that in the georadar data this depth range is dominated by the direct air and ground waves, which do not contain any information about the lateral correlation structure of the subsurface.

Fig. 3b shows a grayscale image of the lateral autocovariance function of the considered part of the outcrop image in Fig. 3a as a function of depth. This image is obtained by evaluating the experimental autocovariance function for each row of the 2-D image matrix under the assumption of local stationarity. It nicely quantifies the dominant sedimentary stratification discernible in the photograph and thus illustrates that the various sedimentary units exhibit notable variations in the decay of their lateral autocovariance functions. The corresponding normalized average lateral autocovariance function together with its best-fitting parametric model (Eq. 4) is shown Fig. 3c. We used a Monte Carlo optimization approach to minimize the error between the observed and modeled autocovariance functions. Clearly, the von Kármán autocovariance model provides an excellent match to the observed data. The inferred ν-value of 0.08 is indicative of so-called "flicker noise", probably the most common statistical characteristic of observed data throughout the sciences (West and Shlesinger 1990). This finding is consistent with the increasing evidence that the distributions of virtually all petrophysical parameters in sedimentary rocks, in particular also the porosity distribution, exhibit flicker noise character (Hardy and Beier 1994; Kelkar and Perez 2002). The inferred correlation length of 2.18 m is qualitatively consistent with the average lateral extent of the dominant lithological features discernible in Fig. 3a.

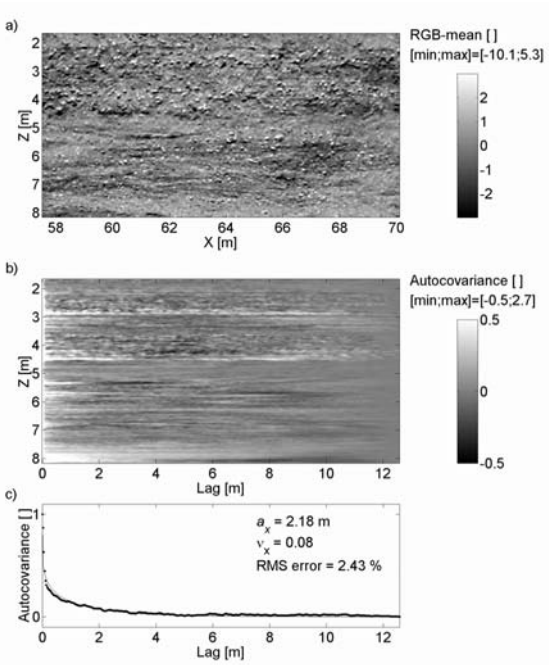


Fig. 3. a) Subset of the outcrop photograph in RGB-mean format shown in Fig. 2a, **b)** image of horizontal autocovariance (evaluated for each row of the image matrix under the assumption of local stationarity) displayed as a function of depth and **c)** average normalized horizontal autocovariance function (dots) with best-fitting von Kármán model superimposed (solid line). The root-mean-square (RMS) error quantifies the mismatch between the observed and modeled autocovariance functions.

This analysis has been repeated for the corresponding subset of the georadar data shown in Fig. 4a. Overall, the structure of the corresponding lateral autocovariance image (Fig. 4b) agrees very well with that of the outcrop photograph (Fig. 3b). This is a valuable and exciting result in its own right, as it indicates that, even in the absence of outcrop or borehole information, this type of analysis can be used for quantitative stratigraphic analysis/zonation of surface georadar data. The corresponding average autocovariance function is again well explained by the used parametric model (Fig. 4c). Both the inferred ν-value and the correlation length are in good agreement with those obtained for the outcrop image (Fig. 3c). A subtle, but interesting difference in the correlation structures of the outcrop image and the georadar data is the response of the foreset structures present in the depth range of ~2-3 m. These structures are clearly imaged by the georadar data (Fig. 4a) and result in a quasi-cyclical response in the lateral correlation image for the corresponding depth range (Fig. 4b) as well as in a corresponding oscillation of the average autocovariance function at larger lags. Conversely, these foreset structures are barely discernible in the outcrop photograph and hence find little or no expression in its lateral correlation structure.

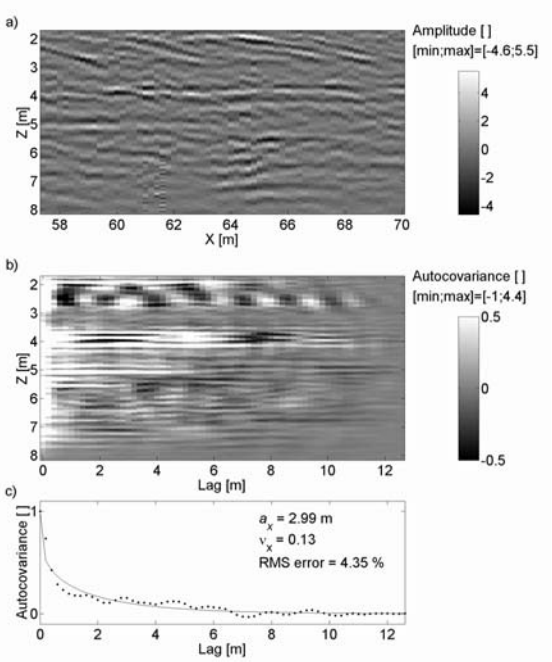


Fig. 4. a) Subset of the georadar data shown in Fig. 2c, **b)** image of horizontal autocovariance (evaluated for each row of the image matrix under the assumption of local stationarity) displayed as a function of depth and **c)** average normalized horizontal autocovariance function (dots) with best-fitting von Kármán model superimposed (solid line). The root-mean-square (RMS) error quantifies the mismatch between the observed and modeled autocovariance functions.

5 Discussion

The results described above (Fig. 3 and 4) are representative for analyzed subsets of the outcrop image and the corresponding georadar data. The inferred ν-values are uniformly close to zero and thus indicative of the seemingly ubiquitous flicker noise character of scientific data. In particular, this finding is consistent with the growing empirical evidence for the universal flicker noise nature of petrophysical properties of sedimentary deposits in general and porosity distributions in particular (Hardy and Beier 1994; Kelkar and Perez 2002). The universality of this phenomenon indicates that it could/should be used as *a priori* or conditional information in a variety of geostatistical analyses. A unique characteristic of von Kármán autocovariance functions with small ν-values is their initially rapid decay at small lags followed by a gradual leveling off at larger lags (e.g., Sidler and Holliger this volume). It is quite likely for this reason that Rea and Knight (1998) used a com-

bination of two spherical autocovariance models with short and long correlation lengths to explain some of their observations. Clearly, the use of such "nested" autocovariance models makes the generation of corresponding synthetic stochastic data fields rather awkward and can also be ambivalent to interpret in geological terms. Together with the ubiquitous and universal nature of flicker noise phenomena in scientific data in general and petrophysical data in particular (Hardy and Beier 1994), this clearly favors the use of the von Kármán autocovariance model (Eq. 4) for such analyses over the more commonly used geostatistical autocovariance models, such as Gaussian, exponential or spherical models.

In contrast to the remarkable uniformity of the ν-values, the average horizontal correlation lengths were found to vary considerably (approximately by a factor of 2 to 3) between the various subsets. However, the correlation lengths inferred from subsets of the outcrop image and the corresponding georadar data were found to be inherently consistent. In agreement with the theoretical considerations outlined above, the correlation length of georadar data tend to be longer (~10-30% on average) than the corresponding values interpreted from the outcrop image. The range of the lateral correlation lengths obtained in this study is also consistent with those inferred by Rea and Knight (1998) for a similar geological environment. The absolute values of the correlation lengths should, however, be considered with some caution, as they to tend be influenced by the scale of the experiment (Gelhar 1993; Western and Blöschl 1999).

We explored the effects of applying different processing strategies to the georadar data and found the estimates of lateral correlation structure to be remarkably robust in this regard. The so-called cosine-of-phase approach illustrated in Fig. 5 can be considered as a representative example in this regard. This processing method can be directly applied to the raw data and largely eliminates the need for further pre-processing of the data. It replaces the original amplitude values by the cosine of their local/instantaneous phases and thus enhances the continuity of events with similar phase relations at the expense of all other events and reduces the overall dynamic range of the data (Yilmaz 1988). We found the application of the cosine-of-phase to be very robust and reliable, particularly also in regions of sub-optimal data quality. Due to its robustness and ease of use, we warmly recommend the use of the cosine-of-phase, possibly assisted by the application of a band-pass filter, as the preferred processing strategy of surface georadar prior to lateral correlation analysis. Finally, we found that the lateral correlation structure of the georadar data is not fundamentally altered by migration. Migration is an image restoration process based on a solution of the acoustic wave equation, which refocuses the data in such a way that diffractions are collapsed and dipping reflectors are moved to their correct subsurface positions (Yilmaz 1988). In analogy to an optical focusing process, which has little or no effects for objects that are close to the lens, the migration georadar data is often ineffective due to their inherently surficial nature and may indeed even be detrimental due to the introduction of numerical artifacts. Moreover, the georadar data considered in this study exhibit a predominantly horizontal stratification and are largely devoid of diffractions (Fig. 2c).

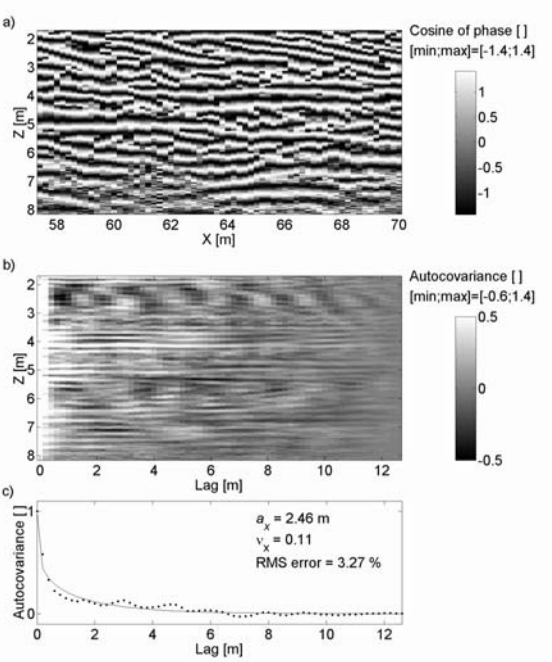


Fig. 5. a) Cosine-of-phase of the raw version of a subset of the georadar data shown in Fig. 2c, **b)** image of horizontal autocovariance (evaluated for each row of image matrix under the assumption of local stationarity) displayed as a function of depth and **c)** average normalized horizontal autocovariance function (dots) with best-fitting von Kármán model superimposed (solid line). The root-mean-square (RMS) error quantifies the mismatch between the observed and modeled autocovariance functions.

In this study, we have concentrated on extracting the normalized horizontal autocovariance function from surface georadar data. In many practical situations, this is indeed the key information that is missing for a complete stochastic characterization of the shallow subsurface as the vertical correlation structure and the variance can be inferred from nearby borehole information. It is, however, important to note that, at least in principle, both the variance and the vertical correlation structure can also be inferred from surface georadar data. Eq. 1-2 indicate that the variance of the distribution of the reflection coefficients in the subsurface can be estimated using the approach described in this paper for non-normalized autocovariance models provided that the "true" absolute amplitudes of the georadar data have been restored prior to correlation analysis. This would imply that all amplitude distortions due to geometric spreading, attenuation, scattering, and instrument effects were accurately accounted and compensated for. In practice, it is unlikely that all the necessary information will ever be available to perform this task with the desired accuracy. Eq. 1-2 also indicate that the vertical correlation length of the subsurface structure can be inferred from the georadar data provided that we can remove the effect of the source signal $w(t)$ (i.e., perfectly deconvolve the geo-

radar data) prior to correlation analysis in the vertical direction. The emitted source signal is considered to be unknown but generally not minimum phase and hence corresponding stochastic deconvolution approaches used in the reflection seismology (Yilmaz 1988) tend to be ineffective or even detrimental. This fundamental problem could be alleviated through a deterministic deconvolution approach, which in turn would require that we could estimate the emitted source waveform $w(t)$ (Eq. 1-2) with reasonable accuracy. To date, very little research has been carried out on this topic. Based on theoretical considerations with regard to the radiative properties of electric dipoles located on or near a dielectric halfspace it is, however, conceivable that $w(t)$ can be inferred from the direct waves traveling through the air and the ground. Interestingly, the air and ground waves are so far considered to represent undesired noise and are often removed from the data.

6 Conclusions

Based on a well-controlled field experiment as well as theoretical considerations, we demonstrate that the lateral correlation structure of surface georadar data can be directly related to that of probed shallow subsurface. For the georadar data as well as for the photographs, the observed lateral correlation functions are well approximated by the band-limited scale-invariant von Kármán autocovariance model. The thus inferred autocovariance models invariably exhibit flicker noise character (i.e., ν-values close to 0), but differ significantly in terms of their horizontal correlation lengths. Overall, we found a far-reaching consistency between the lateral correlation structures extracted from the surface georadar data with those inferred for the corresponding regions of the digital photograph. We also found that the horizontal correlation lengths inferred from the surface georadar data are quite sensitive to geological/lithological variations, but rather robust with regard to details of the processing applied to the data. Our results therefore confirm that the geostatistical analysis of surface georadar data offers an easy, robust and reliable way to estimate the average lateral correlation structure of the shallow subsurface as well as vertical variations thereof. Given the sensitivity of the georadar method to variations in water content, the thus inferred lateral correlation structure may serve as a proxy for the lateral correlation of variations in porosity and hydraulic conductivity (Fig. 1). This study has focused on the normalized lateral autocovariance structure. Based on the methodological foundations and today's understanding of the generation, propagation and recording of georadar waves, we believe that it is unlikely that the variance of the subsurface correlation structure can be inferred from surface georadar data with the desired accuracy. For the same reasons, we are, however, cautiously optimistic that it could be possible to remove the effects of the emitted source waveform and thus to also estimate the vertical component of the subsurface correlation structure directly from surface georadar data.

Acknowledgments

We thank Holcim AG for the permission to acquire the georadar data and the outcrop photographs on their premises and Rosemary Knight and Fritz Stauffer for helpful comments and suggestions. ETH-Geophysics Contribution No. 1377.

References

Beres M, Huggenberger P, Green AG, Horstmeyer H (1999) Using 2- and 3-dimensional georadar methods to characterize glaciofluvial architecture. Sed Geol 129: 1-24

Claerbout JF (1985) Imaging the earth's interior. Blackwell, Oxford

Davis JL, Annan AP (1989) Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy. Geophys Prospect 37: 531-551

Gelhar LW (1993) Stochastic subsurface hydrology. Prentice-Hall, Englewood Cliffs

Goff JA, Jordan TH (1988) Stochastic modeling of seafloor morphology: inversion of sea beam data for second-order statistics. J Geophys Res 93:13589–13608

Hardy HH, Beier RA (1994) Fractals in reservoir engineering. World Scientific, Singapore

Heinz J, Kleineidam S, Teutsch G, Aigner T (2003) Heterogeneity patterns of Quaternary glaciofluvial gravel bodies (SW-Germany): application to hydrology. Sediment Geol 158: 1-23

Huggenberger P (1993) Radar facies: recognition of facies patterns and heterogeneities within Pleistocene Rhine gravels, NE Switzerland. In: Best JL, Bristow CS (eds) Braided rivers, Geol Soc, Spec Publ 75: 163-176

Kelkar M, Perez G (2002) Applied geostatistics for reservoir characterization. Society of Petroleum Engineers, Richardson

Neil A (2004) Ground-penetrating radar and its use in sedimentology: principles, problems and progress. Earth-Sci Rev 66: 261-330

Oldenborger GA, Schincariol RA, Mansinha L (2003) Radar determination of the spatial structure of hydraulic conductivity. Ground Water 41: 24-32

Rea J, Knight R (1998) Geostatistical analysis of ground-penetrating radar data: a means of describing spatial variation in the subsurface. Water Resour Res 34: 329-339

Ritzi RW, Jayne DF, Zahradnik AJ, Field AA, Fogg GE (1994) Geostatistical modeling of heterogeneity in glaciofluvial, buried-valley aquifers. Ground Water: 32, 666–674

Rubin Y (2003) Applied stochastic hydrology. Oxford University Press, Oxford.

Sidler R, Holliger K (this volume) Kriging of scale-invariant data: optimal parameterization of the autocovariance model

von Kármán T (1948) Progress in the statistical theory of turbulence. J Marit Res 7: 252-264

West BJ, Shlesinger M (1990) The noise in natural phenomena. Am Sci 78: 40-45

Western AW, Blöschl G (1999) On the spatial scaling of soil moisture. J Hydrol 217: 203-224

Wharton RP, Hazen GA, Rau RN, Best DL (1980) Advancements in electromagnetic propagation logging: SPE 9267, American Institute of Mining, Metallurgical and Petroleum Engineers

Yilmaz Ö (1988) Seismic data processing. Society of Exploration Geophysicists, Tulsa.

Author index