AUTOMATIC MAPPING OF MONITORING DATA

Søren Lophaven
Informatics and Mathematical Modelling, Technical University of Denmark
Hans Bruun Nielsen
Informatics and Mathematical Modelling, Technical University of Denmark
Correspondence to: hbn@imm.dtu.dk
Jacob Søndergaard
Informatics and Mathematical Modelling, Technical University of Denmark

This paper presents an approach, based on universal kriging, for automatic mapping of monitoring data. The performance of the mapping approach is tested on two datasets containing daily mean gamma dose rates in Germany reported by means of the national automatic monitoring network (IMIS). In the second dataset an accidental release of radioactivity in the environment was simulated in the South-Western corner of the monitored area. The approach has a tendency to smooth the actual data values, and therefore it underestimates extreme values, as seen in the second dataset. However, it is capable of identifying a release of radioactivity provided that the number of sampling locations is sufficiently high. Consequently, we believe that a combination of applying the presented mapping approach and the physical knowledge of the transport processes of radioactivity should be used to predict the extreme values.

INTRODUCTION

In this paper, an approach for automatic mapping of monitoring data is presented. The mapping approach was based on the well-known universal kriging approach (Cressie 1993; Chiles and Delfiner 1999; Wackernagel 2003; Goovaerts 1997), which was implemented by the authors in the Matlab-toolbox DACE (Lophaven et al. 2002a; Lophaven et al. 2002b), primarily developed for Design and Analysis of Computer Experiments (Sacks et al, 1989).

The performance of the mapping approach was tested on two datasets containing daily mean gamma dose rates in Germany reported by means of the national automatic monitoring network (IMIS). In the second dataset an accidental release of radioactivity in the environment was simulated in the South-Western corner of the monitored area. Predictions were computed at 808 locations, based on 200 observations, and the performance was measured by comparing the predictions with the true values.

In order to automate the mapping approach an algorithm, which chooses between combinations of regression model and correlation function, is applied. The development of the algorithm was based on 10 datasets containing daily mean gamma dose rates in Germany at the same 200 locations as in the two test datasets.

Section 2 briefly describes the universal kriging approach, focusing on the elements which deviate from the traditional geostatistics. Further details can be found in the widely available literature on geostatistics (Cressie 1993; Chiles and Delfiner 1999; Wackernagel 2003; Goovaerts 1997). Furthermore, details are given about the use of the 10 training datasets (prior information), and the algorithm, which chooses between combinations of regression model and correlation function. In section 3 results are presented, whereas a discussion of the obtained results is given in section 4, focusing on possible improvements of the applied automatic mapping approach.

METHODOLOGY

The universal kriging predictor is the predictor which minimizes the mean squared error. It can be shown that this is given by

$$\hat{y}(x) = f(x)\beta^* + r(x)^T \psi^*$$

where $\beta^* = (F^T R^{-1} F)^{-1} F^T R^{-1} Y$, $\Psi^* = R^{-1}(Y - F\beta^*)$, $f(x)$ is a vector defining the regression model as a function of spatial location x, F is the design matrix defining the regression model as a function of sampling location s, $r(x)$ is a vector with spatial correlations between x and s, R is a matrix of spatial correlations between sampling locations, whereas Y is a vector of observations. The mean squared error of the predictor is given by

$$\varphi(x) = \sigma^2 (1 + u^T (F^T R^{-1} F)^{-1} u - r(x)^T R^{-1} r(x))$$

where $u = F^T R^{-1} r(x) - f(x)$.

REGRESSION MODELS

In this paper we consider regression models with polynomials of order 0,1 and 2, i.e. of the form $\beta_1 f_1(x), \ldots, \beta_p f_p(x)$. In the following we denote the lth component of x by x_l, and the dimension by m, which in the current application is 2. The applied regression models are:

Constant, $p=1$: $f_1(x)=1$
Linear, $p=m+1$: $f_1(x)=1$, $f_2(x)=x_1, \ldots, f_{m+1}(x)=x_m$
Quadratic, $p=1/2(m+1)(m+2)$:
$f_1(x)=1$,
$f_2(x)=x_1, \ldots, f_{m+1}(x)=x_m$,
$f_{m+2}(x)=x_1^2, \ldots, f_{2m+1}(x)=x_1 x_m$
$f_{2m+2}(x)=x_2^2, \ldots, f_{3m}(x)=x_2 x_m$
$f_p(x)= x_m^2$

CORRELATION MODELS

Below the correlation models applied in this paper are shortly presented. We restrict our attention to correlations of the form

$$R(\theta, w, x) = \prod_{l=1}^{m} R_l(\theta, w_l - x_l)$$

i.e., to products of one-dimensional correlations. The models in table 1 are considered:

Name	$R_l(\theta, d_l)$
Exponential	$Exp(-\theta_l \lvert d_l \rvert)$
General exponential	$Exp(-\theta_l \lvert d_l \rvert^{\theta_{m+1}}),\ 0<\theta_{m+1}\leq 2$
Spherical	$1-1.5\xi_l+0.5\,\xi_l^3,\ \xi_l=min\{1,\ \theta_l \lvert d_l \rvert\}$
Linear	$max\{0,1-\theta_l \lvert d_l \rvert\}$

Table 1
Correlation models. $|d_l|$ is the distance between two points in the lth dimension

It is seen that distance is measured separately is the individual dimensions. The use of separable correlation functions corresponds to the models usually used in space-time geostatistics (De Cesare et al, 2001). In space-time geostatistics the separability assumption can be checked visually by comparing the sample space-time semivariogram and the semivariogram model corresponding to separability. Something similar to this could be done within the DACE-approach. Furthermore, the correlation functions presented here do not include a nugget effect, which is reasonable for many engineering applications. However, when modelling noisy data, such as environmental monitoring data, the inclusion of a nugget effect would be beneficial (Sasena et al, 2002). A nugget effect is not applied in this paper but could be included as

$$R(\theta, w, x) = (1 - nugget)\prod_{l=1}^{m} R_l(\theta, w_l - x_l)$$

Assuming a Gaussian process, the optimal parameters of the correlation model solves

$$\min\left\{\det(R)^{1/n}\sigma^2\right\}$$

This definition corresponds to maximum likelihood estimation. When minimizing (5) a pattern search method, which is a modified version of the Hooke & Jeeves method (Kowalik and Osborne 1968), was applied. When estimating the parameters in the 12 models presented here, this method used between 10 and 14 function evaluations to solve (5), when applied to the first dataset. We did not see any differences in the number of function evaluations between the different correlation functions. All computations were performed in a standardized space, i.e. the coordinates of sampling locations as well as the actual data values were standardized by subtracting the mean value and dividing by the standard deviation.

USE OF PRIOR INFORMATION

The 10 training datasets (prior information) were used to set initial values for the optimisation, based on (5), of the parameters in the correlation model. Hence, it was not used to estimate the parameters, but just to give starting values for the optimisation algorithm. The initial values were obtained by computing sample semivariograms of the 10 datasets in the standardized space (Figure 1).

Based on Figure 1 it was chosen to set initial values of the correlation parameters in both dimensions to 2, and the process variance to 1. Histograms for the 10 training datasets were also computed (Figure 2), and from these we decided not to do any kind of data transformation prior to applying the mapping approach. However, the current approach could be extended by including automatic transformation of data.

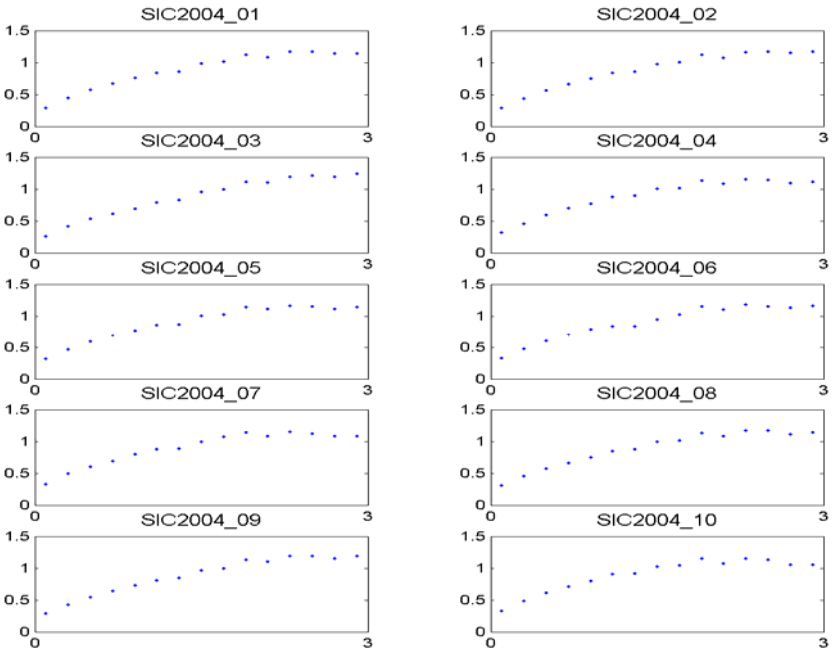

Figure 1
Sample semivariograms for the 10 training datasets in the standardized space.

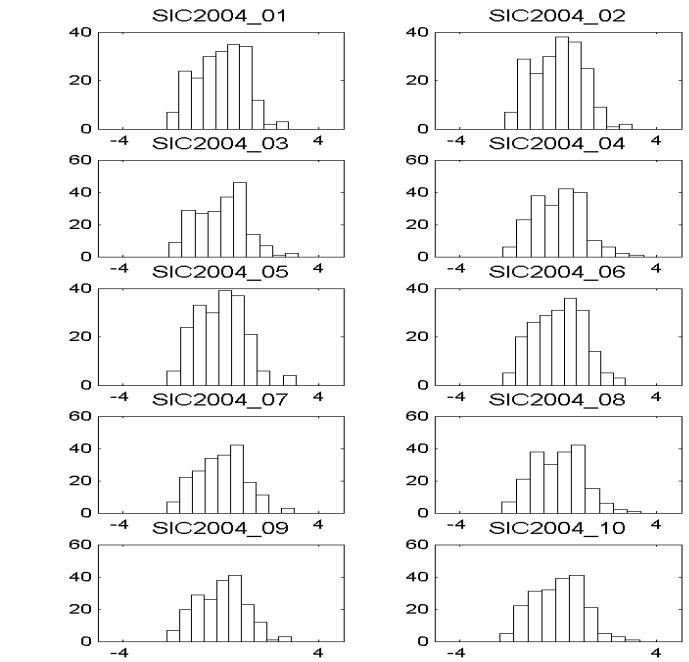

Figure 2
Histograms for the 10 training datasets in the standardized space

TUNING THE ALGORITHMS

The three regression models and four correlation models can be combined into 12 potential model approaches. The automatic mapping approach chooses between these 12 combinations by means of leave-one-out cross validation, i.e. one observation is removed from the dataset at a time. For each observation removed the model is reestimated from the remaining data values, and this model is used to predict the left out observation by each of the 12 combinations. The mean squared error between left out observations and predictions was used to decide which combination to apply for predicting values in the 808 locations.

RESULTS

When applying the automatic mapping approach to the two datasets, it was found that the combination of a first order polynomial and the general exponential correlation model performed best when applied to the first dataset, whereas a polynomial of zero order and the general exponential correlation model was the best combination for computing predictions in the second dataset.

Table 2 shows descriptive statistics for the 200 observations in the two datasets, as well as for the 808 predicted values based on each of the datasets.

N = 808	Min.	Max.	mean	Median	std. dev.
Observed (first data set)	57.0	180.0	98.0	98.8	20.0
Predictions (first data set)	66.7	132.5	96.8	99.9	14.4
Observed (second data set)	57.0	1528.2	105.4	99.0	83.7
Predictions (second data set)	80.0	494.9	109.5	103.4	36.4

Table 2
Comparison of the predicted and measured values (nSv/h)

It is seen that the mapping approach tends to smooth the observations, i.e. the range of the predictions is smaller than for the original dataset, with is also indicated by smaller standard deviations. In particular the high concentrations in the second dataset, due to the simulation of an accidental release of radioactivity, are severely underestimated.

In order to assess the efficiency of the automatic mapping approach, the mean absolute error (MAE), the bias (or mean error ME), and the root mean squared error (RMSE) of the predictions at the n = 808 locations are reported in Table 3. These are computed as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^* - y_i\right|$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - y_i\right)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^* - y_i)^2}$$

where y_i^* is the predicted value at location i and where y_i is the true value. Furthermore, Pearson's r coefficient of correlation between the predicted and true values is shown in Table 3.

Data sets:	MAE	ME	Pearson's *r*	RMSE
First data set	9.7	1.2	0.76	13.1
Second data set	22.2	-4.1	0.54	71.2

Table 3
Comparison of the errors

It is seen that the suggested approach performs much better when applied to the first dataset, thus it is not that well suited for handling extreme values.

Below, maps showing the predicted values (Figure 3) and the associated uncertainties (Figure 4), computed by (1) and (2), respectively, are presented. Furthermore, a 3D map of the second dataset is shown in Figure 5. In Figure 3 the simulated release of radioactivity is clearly seen, even though the actual values are underestimated. In Figure 4 it is seen that the uncertainty given by (2) is highly dependent on the distances to the 200 locations of observations, i.e. the uncertainty is small close to observation points, and vice versa.

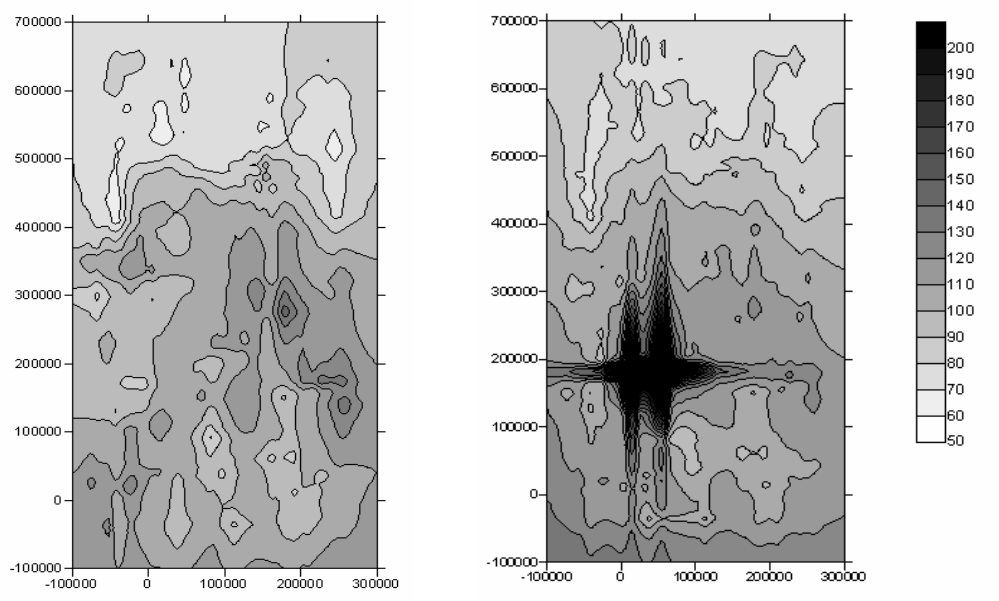

Figure 3
Isoline levels (nSv/h) for the 1st set (left) and the 2nd set (right) computed based on 200 observations

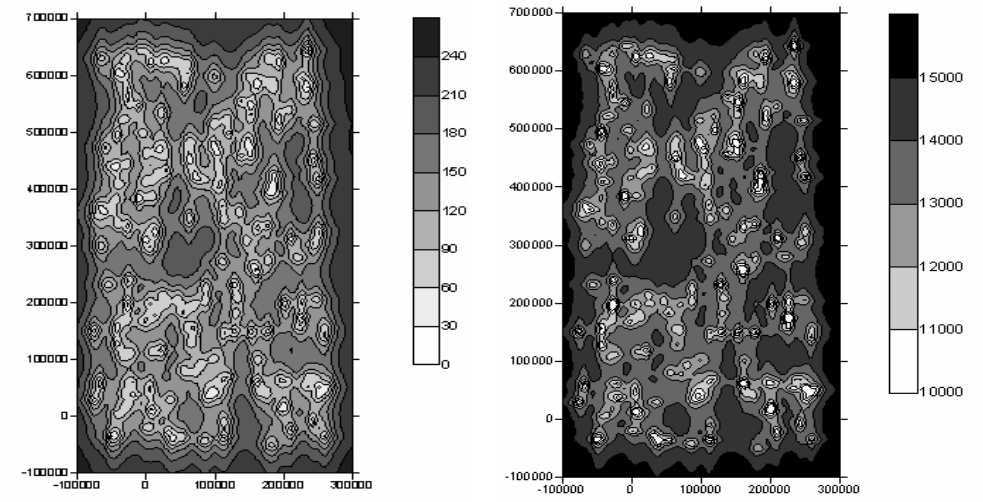

Figure 4
Isoline levels showing the uncertainty, given by (2), associated to the predictions obtained for the 1st set (left) and the 2nd set (right). Computations were based on 200 observations

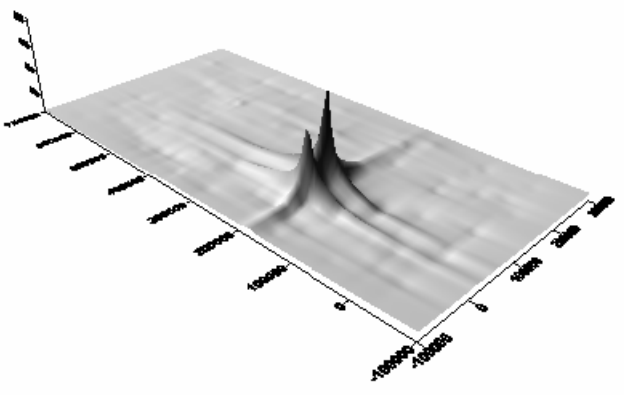

Figure 5
3D map showing extreme values found in the 2nd set (vertical scale in nSv/h)

DISCUSSION

In this paper we have presented an approach for automatic mapping of monitoring data. This is based on universal kriging, and chooses between 12 different kriging models by means of cross validation. The approach has a tendency to smooth the actual data values, and it therefore underestimates extreme values, as it is seen when applied to the second dataset. On the other hand it does identify the radioactivity release simulated in the second dataset (Figure 5). Therefore, we believe that even though the prediction of extreme values is unreliable, the mapping approach could be used as a warning system, as it is known from classical process control (Montgomery 2000). Such a system is activated by extreme predictions, and an approximate location of release can be estimated by the mapping approach. To obtain reliable predictions the physical knowledge about the

transport processes of radioactivity should somehow be applied. We have also tried to apply the mapping approach to the log-transformed data values in the second dataset. This did not improve the performance significantly.

The average computational time for each dataset was approximately 45 minutes. This time is primarily used to compare the 12 potential kriging models by means of cross validation. As described in section 2 the model parameters are reestimated each time an observation is left out from the dataset in the cross validation algorithm. We also applied the described algorithm without reestimation of the model parameters to the first dataset. This reduced the computational time to only four minutes, i.e. it is the parameter estimation part of the procedure that takes a lot of time. The values of ME, MAE and RMSE computed without reestimation of the parameters were similar to those given in the first row of table 3.

CONCLUSIONS

This paper presents an approach, based on universal kriging, for automatic mapping of monitoring data. The approach has a tendency to smooth the actual data values, and therefore it underestimates extreme values, as seen in the second dataset. However, it is capable of identifying a release of radioactivity provided that the number of sampling locations is sufficiently high. Consequently, we believe that a combination of applying the presented mapping approach and the physical knowledge of the transport processes of radioactivity should be used to predict the extreme values.

CODES

The Matlab kriging toolbox DACE and documentation can be downloaded from http://www.imm.dtu.dk/~hbn/DACE.

A link to the toolbox can be found at http://www.ai-geostats.org.

Matlab code for cross-validation, estimation of sample semivariograms etc. can be obtained by contacting the first author.

REFERENCES LIST

Chiles, JP; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley; 1999.

Cressie N. *Statistics for spatial data*. New York: Wiley; 1993.

De Cesare, L; Myers, DE; Posa, D. 'Product-sum covariance for space-time modelling: an environmental application'. *Environmetrics* 2001; 12: 11-23.

Goovaerts, P. *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press; 1997.

Kowalik, J; Osborne, MR. *Methods for unconstrained optimization problems*. New York: Elsevier; 1968.

Lophaven, SN; Nielsen, HB; Søndergaard, J. 'DACE - A Matlab kriging toolbox'. Technical report. Informatics and Mathematical Modelling, Technical University of Denmark; 2002a.

Lophaven, SN; Nielsen, HB; Søndergaard, J. 'Aspects of the Matlab toolbox DACE'. Technical report. Informatics and Mathematical Modelling, Technical University of Denmark; 2002b.

Montgomery, DG. *Introduction to Statistical Quality Control*. New York: Wiley; 2000.

Sacks, J; Welch, WJ; Mitchell, TJ; Wynn, HP. 'Design and analysis of computer experiments'. *Statistical Science* 1989; 4 (4): 409-435.

Sasena, M; Parkinson, M; Goovaerts, P; Papalambros, P; Reed, M. 'Adaptive experimental design applied to an ergonomics testing procedure'. *Proceedings of DETC'02*. Montreal. Canada; 2002.

Wackernagel, H. *Multivariate Geostatistics: An Introduction with Applications*. Berlin: Springer; 2003.

Cite this article as: Lophaven, Søren; Nielsen, Hans Bruun; Søndergaard, Jacob. 'Automatic mapping of monitoring data'. *Applied GIS*, Vol 1, No 2, 2005. pp. 13-01 to 13-09. DOI: 10.2104/ag050013