USING ORDINARY KRIGING TO MODEL RADIOACTIVE CONTAMINATION DATA

Elena Savelieva

Nuclear Safety Institute (IBRAE), Russian Academy of Sciences, B. Tulskaya 52, 113191, Moscow, Russia
Correspondence to: esav@ibrae.ac.ru

This paper deals with an application of ordinary kriging (OK) for spatial interpolation of data in a completely automatic ("one-click mapping") manner. The important set of kriging parameters (semivariogram model, search strategy, etc.) were tuned based on the prior characteristics of the phenomenon considered. The prior information provided as 10 sets of monitoring observations taken at different days was used to analyse and model the spatial correlation of the phenomenon. Furthermore, the prior information was expected to be consistent within a rather long time range and therefore assumed to reflect the structure of the contamination pattern at any given day. The approach applied here gave satisfactory results for both routine and emergency data sets. The benefits and drawbacks of the kriging model were well illustrated in the study. Ordinary kriging can be considered as a real candidate for the implementation in a decision support system.

1. INTRODUCTION

The present paper describes capabilities of a well-known geostatistical interpolation technique – ordinary kriging (OK) – for fast on-line analysis of radioactively contaminated territories. The main problem with on-line application of the geostatistical approach is the expert work dealt with estimation and modelling of the spatial correlation structure. Several attempts, more or less successful, were made to automate this stage (Cressie 1985; Haas 1990; Zhang et al. 1995). Nevertheless, most geostatisticians agree that modelling a spatial correlation structure requires expert knowledge and effort (Isaaks and Srivastava 1989; Pannatier 1996; Goovaerts 1997; Deutsch and Journel 1998).

In the framework of the SIC2004 exercise (application of a "one-click mapping" method), ordinary kriging was performed with a predefined spatial correlation structure model based on the available prior data sets. These data sets are routine monitoring data and they represent well the first exercise data set (data set I). The second ("joker") data set (II) shows absolutely different behaviour including some anomalous features, which were difficult to predict. In the current work we will discuss advantages of ordinary kriging, e.g. rather fast computation, direct estimation of uncertainty, easiness of an understanding and treatment, etc; and some of its drawbacks, mainly the smoothing effect, which is especially strong when the search radius is too large.

The main purpose of the work is to draw the attention of decision makers, who work with emergency situations caused by pollution, to the capabilities of a relatively simple and well-known classical method, which can be easily implemented in any decision support system for crisis situations. Some improvements can be expected by incorporat-

ing additional information on weather conditions (wind direction) and by using other available pollution patterns in the prediction model.

2. METHODOLOGY

Ordinary kriging is extensively described in the literature, for example Isaaks and Srivastava (1989), Goovaerts (1997), Chiles and Delfiner (1999), etc. Below we outline how we have adapted the kriging method and equations to the current exercise.

Let us suppose that there is a random field $Z(x)$ represented by a set of given values $Z(x_i)$ measured at locations x_i. According to a linear estimator a value at the unmeasured location x_0 ($Z^*(x_0)$) is given by the following equation:

$$\boldsymbol{Z^*(x_0) = \sum_{i=1}^{N(x_0)} \lambda_i Z(x_i)}, \tag{1}$$

where $N(x_0)$ is the number of samples from the neighbourhood of x_0 taken into the account for the estimation, and λ_i are ordinary kriging weights. The neighbourhood depends on a user-defined search rule. It can be defined as a circle with a search radius or as an ellipse with major and minor axis and orientation. The neighbourhood is expected to be small enough to preserve a limitation on a constant mean. For a meaningful estimation the neighbourhood needs to contain at least 3 samples.

The set of weights (λ) is determined by minimizing of the estimation variance under the constraint of unbiasedness:

$$\min_{\lambda} \quad \boldsymbol{Var}\left(\left(Z^*(x_0;\lambda) - Z(x_0)\right) - \mu\left(\sum_{i=1}^{N(x_0)} \lambda_i \right) \right), \tag{2}$$

by solving the so-called ordinary kriging equations:

$$\begin{cases} \sum_{i=1}^{N(x_0)} \lambda_i \gamma_{ij} - \mu = \gamma_{j0}, \quad j = 1 \ldots N(x_0) \\ \sum_{i=1}^{N(x_0)} \lambda_i = 1 \end{cases}, \tag{3}$$

where μ is the Lagrangian multiplier. It is a linear system of $N(x_0)+1$ equations with $N(x_0)+1$ unknowns. γ_{ij} correspond to a spatial correlation structure of the random field described by a semivariogram:

$$\gamma_{ij} = \gamma(\vec{h}_{ij}) = \tfrac{1}{2}\boldsymbol{Var}\left(Z(x+\vec{h}_{ij}) - Z(x)\right), \tag{4}$$

where $\vec{h}_{ij}$ is a vector separating locations x_i and x_j. The ordinary kriging equations have a unique solution if they are well-defined (not singular). The kriging estimation error is described by the kriging variance

$$\sigma^2 = \sum_{i=1}^{N(x_0)} \lambda_i \gamma_{i0} + \mu, \tag{5}$$

directly estimated once the kriging equations (3) are solved. Under the assumption of a local Gaussian distribution a 95% confidence interval around the kriging estimate is $[Z^*(x_0)-2\sigma; Z^*(x_0)+2\sigma]$.

Therefore, to use ordinary kriging it is necessary to define the search rule and the semivariogram values for all possible separating vectors (a semivariogram model). These parameters are usually determined from the available raw observation data. The rules of the SIC2004 exercise required defining of all method's parameters in advance prior to obtaining the estimation data set to be interpolated. In such a situation ordinary kriging parameters can be evaluated based on some general expert experience or from any available prior knowledge about the data.

2.1 USE OF PRIOR INFORMATION

Prior information represents the spatial behaviour of the monitoring data for 10 different past days, with unknown different conditions (perhaps different seasons, weather conditions, etc). It was assumed that these data sets reflect the key spatial features of the pattern distribution in ordinary (not emergency) situations. Also, it was expected that based on these features and as new raw measurements became available at the target time, reasonable predictions could be worked out. As for the mysterious "joker" data (Dubois and Galmarini 2005), no reasonable assumptions can be made a priori as the emergency situation may be quite different from the usual monitoring situation. One plausible thing to check is whether the same approach would work with the data featuring anomalies in the spatial pattern.

The following analysis was performed using the prior data sets:

- Statistical analysis of the data series available in each of the provided locations (10 observations from each location): mean, median, minimum, maximum, variance, range, etc;
- Spatial correlation analysis: experimental semivariograms computed for each available time point (day) and one averaged over all 10 semivariograms analyzed;
- Analysis of differences and/or similarities in the semivariogram structures;
- Selection of the semivariogram model to fit spatial correlation for observation at each of the 10 days.

It was found that the data sets are not exactly the same for each day, but they are very similar. For example, there are only 6 locations with range of observations higher than 30 units, while the global minimum over all data sets is 50.

Experimental semivariograms for different days also look rather similar (Figure 1): they present no anisotropy at a short range (<60 km) and indicate some anisotropy in East-West direction at medium range (from 60 to 200 km). The variogram "sill" (a priori variance) is the same for all data sets, and it is reached at the same range. The semivariogram averaged over 10 semivariograms presents similar features (Figure 2, left). Consequently, the semivariogram model fitted to the averaged semivariogram could be good representation of the spatial structure at any day of observation (Figure 2, right). Finally, a spherical semivariogram model with a nugget of 33.59, sill of 267.0 and two ranges of 306,300 m and 230,400 m with the larger one oriented in East-West direction was selected.

The search neighborhood was chosen relevant to the averaged spatial correlation structure: an ellipse with the main axes 310,000 m and 235,000 m, the larger one

oriented in East-West direction. The ordinary kriging estimation was computed where the neighborhood contained at least 3 samples, otherwise the location was left unestimated. The maximum number of samples used for estimation at each location x_0 was 20 (the closest to x_0).

2.2 TUNING THE ALGORITHM

Cross-validation leave-one-out technique was used to check the suitability of the selected parameters. Cross-validation was carried out for all 10 prior data sets. The results were compared using the root-mean-square error (RMSE) and Pearson's r correlation coefficient between the real values and the obtained estimates. The results of the cross-validation summarised in Table 1 appeared to be rather similar for all 10 prior data sets and can be considered as satisfactory.

Finally, a prediction grid for all 10 days was estimated. These predictions are characterised by the reasonable fluctuations due to the day-to-day variations of the initial data. So it was decided that the method could be used to predict similar patterns at different times. Also, it was considered to check how the model would perform for the data with unknown structure, which could be anomalous and contain outliers, as indicated about the second data set (II) in the SIC2004 announcement.

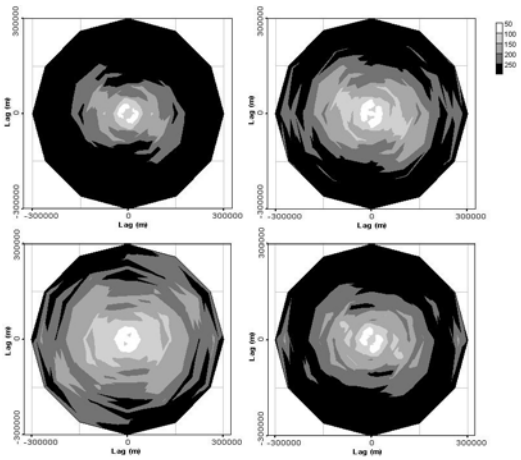

Figure 1
Examples of experimental semivariogram roses for prior data

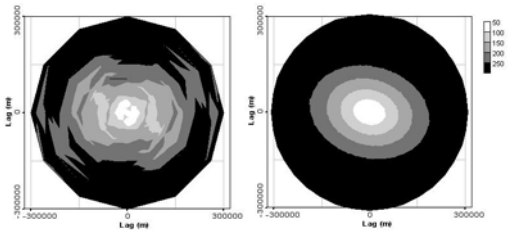

Figure 2

Experimental semivariogram rose averaged over 10 days (left) and fitted semivariogram model (right)

Day	Pearson's *r*	RMSE		Day	Pearson's *r*	RMSE
1	0.78	17.02		**6**	0.75	14.90
2	0.78	16.96		**7**	0.73	15.63
3	0.79	15.98		**8**	0.75	15.41
4	0.75	15.70		**9**	0.77	16.34
5	0.76	15.33		**10**	0.74	16.27

Table 1
Cross-validation error parameters

3. RESULTS

Calculations performed on the two data sets (I and II) including downloading them from the internet, setting a new connection with the SIC 2004 website and uploading the result took 3 minutes. Calculations were performed separately for each data set, every time solving all the ordinary kriging equations. Generally, once the kriging weights had been calculated, they could be reused for another data set as they depended only on the mutual position of the estimated locations and the semivariogram model.

3.1. GENERAL RESULTS

The main statistics for the observed and estimated data are given in Table 2. It is evident that means and medians are reproduced rather well for both data sets. Evidence for the smoothing effect lies in the underestimated standard deviation. The maximum value is underestimated as well due to this smoothing effect. The smoothing is much stronger for the second data set II, where the actual correlation range is smaller than the one used for estimation, according to the source data (Dubois and Galmarini, 2005). An interesting observation is the underestimation of the minimum value in the second data set. On the contrary, overestimation of the minimum was expected due to the smoothing effect. This minimum, together with 6 other data points estimated to be lower than the minimum of the raw data, is located just North of the hot spot (white spot in Figure 3, right). All of the underestimated points are located in the area of the strongest pattern gradient, where it is even correct to apply local kriging because the stationarity assumption is violated (constant mean can't be expected even close to these locations). It looks as if the estimates at these points compensate the high hot spot values so as to keep the local mean constant.

N = 808	Min.	Max.	Mean	Median	Std. dev.
Observed (first data set I)	57.00	180.00	98.02	98.80	20.02
Estimates (first data set II)	66.80	130.35	96.63	98.50	15.19
Observed (second data set I)	57.00	1528.20	105.42	98.95	83.71
Estimates (second data set II)	37.53	651.18	103.23	97.89	53.17

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

The mean absolute error (MAE), the bias (or mean error, ME), the root-mean-squared error (RMSE) and Pearson's *r* coefficient of correlation between the estimated and true values were computed based on the 808 validation locations. The results are presented in Table 3.

In general, results for the first data set (I) are better, which is not surprising because of its similarity with the 10 prior spatial distributions. The RMSE and Pearson's *r* correlation coefficient for the first data set are similar to those obtained through the cross-validation procedure (Table 1). The same parameters for the second data set (II) are worse, but they are not meaningless. The general underestimation appears in the results for both data sets.

Data sets:	MAE	ME	Pearson's *r*	RMSE
First data set I	9.11	-1.39	0.78	12.49
Second data set II	19.68	-2.18	0.56	69.08

Table 3
Comparison of the validation errors

Two sets of contour maps were produced: the map of the OK estimates (Figure 3) and the map of the associated estimation uncertainties, OK variance (Figure 4). OK variance is exactly the same for both cases (data sets I and II), as it depends only on the semivariogram model and on the spatial location of the samples. That is why there is only one plot associated with the kriging errors. In all of these figures crosses indicate the locations of the estimated values and boxes indicate the locations of the input data.

The results of interpolation for the two data sets differ from each other (Figure 3), as they depend on the input values. The detected hot spot corresponds to the flux in the second data set II and is observed in Figure 3 (right). But other parts of the region (especially the northern part) look unaffected by the pollution anomaly, so they are very similar to the estimates based on both data sets.

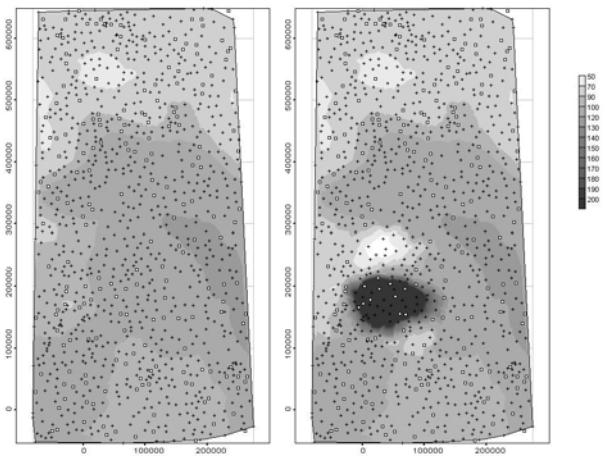

Figure 3
Levels of OK estimates (nSv/h) for the 1st (left) and the 2nd data set (right)

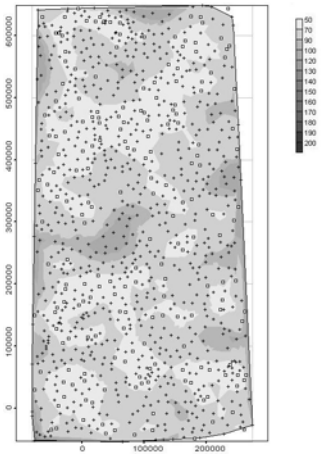

Figure 4
Levels of OK variance represent the uncertainty associated with the estimates obtained for the 1st and the 2nd data sets

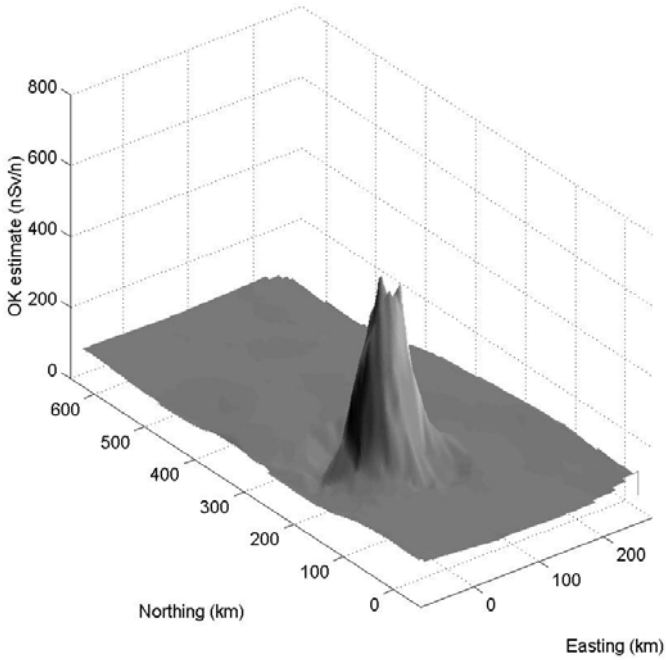

Figure 5
3D map showing extreme values found in the 2nd data set

3.2. **DETECTING ANOMALIES AND OUTLIERS**

Figure 5 represents the locations of the extreme values obtained from OK for the second data set II. Although the extremes were detected, the significant smoothing effect makes the hot spot lower and wider. Still the result is not bad even when using not suitable spatial correlation model and a search neighborhood larger than the actual correlation radius.

4. DISCUSSION

The results on the first data set are fair, as was expected. The data set I exhibits features similar to any one of 10 prior data sets. So the routine monitoring data can be satisfactorily interpolated using kriging weights calculated once and then stored without tuning the model additionally. Perhaps taking into account some seasonal variations can give additional improvement, if there is reason to suspect any seasonal influence on the behaviour of the monitoring data.

The other situation was observed with the second ("joker") data set II. Because of the flux (outliers) the experimental semivariogram features very short correlation range – more than two times shorter than the one for the routine prior data (Figure 6). There is no real correlation between the flux values and the parts of the region not yet affected by the flux. Also, since the theoretical limitations of the OK (constant mean) are violated in the area around the flux, estimation in the area around the anomaly becomes biased (strong underestimation). The unaffected parts of the region are predicted well by the OK with the pre-selected parameters, as they display pattern similar to that of the prior data. Still the flux location was detected well, despite the strong smoothing effect.

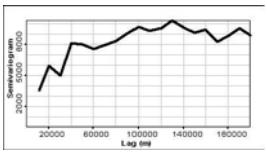

Figure 6
Omnidirectional experimental semivariogram of 2nd data set II

The good correspondence of the local statistics (as indicated by Figure 7) can be considered an advantage of the method. Local means were calculated using a moving average algorithm – averaging the samples over a cell from a regular rectangular grid, which covers the study region. The grid is 7 by 14, with a cell size of approximately 5 km by 5 km. The mean number of validation locations within a cell is 8, and only three border cells contain fewer than 3 validation samples. In Figure 7 each cell is coloured according to the difference between the local mean for the validation data set II and the corresponding OK estimates. Crosses mark validation locations in Figure 7. Good correspondence between local means indicates the possibility of using OK with a site-specific (not event-specific) semivariogram. Also, OK can be used as a model for understanding the average effect of a crisis accidental pollution.

OK estimates together with the kriging variance are usually considered a measure of uncertainty. To check the goodness of this measure, the 'prediction errors' (the OK standard deviations) and 'observed errors' (differences between the OK prediction and the true value) were compared at the 808 validation locations. Let the Z score be defined by the ratio of the observed to the predicted errors. The standard deviation of the Z score values was computed as well. In the ideal prediction case, the Z score standard deviation should be equal to one. In this case, it appeared to be 0.96 for the first data set I and 7.87 for the second data set II. Thus, for the first data set I the uncertainty is described fairly well. The same uncertainty model did not appear to be adequate for the second data set II, but still it is not totally irrelevant. The application of the irrelevant semivariogram

model certainly brings some negative consequences, but still the OK estimates are not meaningless.

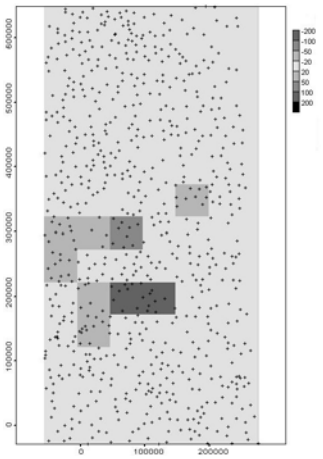

Figure 7
Difference between local means of the original data and the OK prediction for the 2nddata set II

5. CONCLUSIONS

The following conclusions can be drawn based on the study performed:

- A model for spatial correlation structure can be worked out based on the prior monitoring observations from a site if there is a temporal persistence in the data. Otherwise a different spatial correlation structure model can be built and applied under different conditions (season, weather, pollution composition, etc.). It may be interesting to investigate whether using a spatial correlation model fitted to emergency data from other sites will provide good results.
- Ordinary Kriging using the semivariogram model fitted to the prior data provides good prediction for the routine case (data set I) and some information in the emergency case (data set II) – smoothed flux position. OK estimates are always accompanied by the estimation variance usually treated as a measure of the uncertainty, which appears to be adequate for routine case (data set I) and satisfactory for the emergency case (data set II).
- Being a rather fast and simple method, OK can easily be incorporated into a decision support system with semi-automatic semivariogram tuning. A site-specific prior spatial correlation model may be prepared in advance. More detailed expert analysis may be performed later, when more detailed information from the contaminated site has been collected.

SOFTWARE CODE

All calculations and illustrations were made using GSO software. The educational (limited) version of GSO is distributed on CD with Kanevski and Maignan 2004.

REFERENCES

Chiles, J.-P., Delfiner, P. Geostatistics. *Modeling Spatial Uncertainty*, John Wiley & Sons INC; 1999.

Cressie, N. 'Fitting models by weighted least squares'. *Mathematical Geology* 1985; 17(5): pp. 563–586.

Deutsch, C.V., Journel, A.G. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press; 1998.

Dubois, G., Galmarini, S. 'Introduction to the Spatial Interpolation Comparison (SIC) 2004 exercise and presentation of the data sets'. *Applied GIS* 2005; 1(2): pp. 9-01 to 9-10.

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

Haas, T.C. 'Kriging and automated variogram modelling with a moving window'. *Atmospheric Environment* 1990; 24A: pp. 1759–1769.

Isaaks, E. H., Shrivastava R. M. *An introduction to applied geostatistics*. Oxford University Press; 1989.

Kanevski, M., Maignan, M. *Analysis and Modelling of Spatial Environmental Data*. EPFL Press; 2004.

Pannatier, Y. VARIOWIN. *Software for Spatial Data Analysis in 2D*. Springer Verlag; 1996.

Zhang, X.F., Van Eijkeren, J.C.H., Heemink, A.W. 'On the weighted least-square method for fitting a semivariogram model'. *Computers & Geosciences* 1995, 21(4): pp. 605–608.

Cite this article as: Elena Savelieva. 'Using ordinary Kriging to model radioactive contamination data'. *Applied GIS*, Vol 1, No 2, 2005. pp. 10-01 to 10-10. DOI: 10.2104/ag050010