○ AUTOMATED MAPPING USING MULTILEVEL B-SPLINES

Anatoly A. Saveliev

*Faculty of Ecology, Kazan State University, Kazan, Russia
Correspondence to: saa@ksu.ru*

Andrey V. Romanov

R&D Center ScanEx, Moscow, Russia

Svetlana S. Mukharamova

Faculty of Ecology, Kazan State University, Kazan, Russia

This paper describes application of multilevel B-Splines approximation (MBA) algorithm to the SIC2004 exercise as a high performance automatic mapping means in emergency situations. MBA method was compared with kriging, and various flexible MBA method's tuning "controls" were proposed and discussed. Outliers automatic detection and delineation techniques considered could be used to process outliers without the loss in performance. Prior data were used to adjust method parameters and discover the pattern of spatial correlation. The interpolation algorithm development did not assume any information about the phenomena in addition to values given.

1. INTRODUCTION

The goal of Spatial Interpolation Comparison exercise (SIC2004) involves estimating variable values (levels of natural radioactivity) at given locations by means of *fast* and *automatic* interpolation algorithm, which does not need human intervention/decision when processing the data. All the locations (1008 points) were divided into two datasets, i.e. the locations with known values (200 locations) and locations with the values to be estimated (808 locations). Additional 10 days of measurements at those 200 locations were given as prior information (Dubois and Galmarini, 2005).

The requirements for the algorithm to be *fast* and *automatic* were considered as most important goals in the development.

Scattered data interpolation is crucial in meteorological and geological modeling, medical imaging, cartography, image processing, 3D modeling etc. Hence, many interpolation methods were suggested and intensively assessed w.r.t. computational speed and accuracy. In general the 2D interpolation problem is formulated as follows: given scattered points $P = \{p_1, \dots, p_N\} \subset R^2$ and values of function f at these points $f(p_1), \dots, f(p_N)$, it is necessary to find interpolation (approximation) function $\hat{f} : R^2 \rightarrow R$, that fit to the values of f on P . Most of interpolation methods present

$\hat{f}$ as a weighted sum of known values $z_k = f(p_k)$: $\hat{f}(x, y) = \sum_{i=1}^N w_i(x, y)z_i$, and differ in the methods for weights calculation.

In one of the simplest interpolation methods – the Inverse Distance Weighted Method, the interpolation function $\hat{f}(x, y)$ at arbitrary point (x, y) is defined as a weighted sum with the weights proportional to the k^{th} power of the inverse distance between the point (x, y) and points with known values. This method has several well known drawbacks: estimates are bounded by the extrema in the sampled values; it uses weights of every scattered point; whis the $k = 1$ the interpolation function is C^0 -continuous and has many cusps and corners; derivatives are zero on P for $k > 1$. Shephard (Shepard, 1968) proposed variation on the inverse power, with two different weighting functions using two separate neighborhoods. Quadratic Shepard method is another variation (Franke and Nielson, 1980).

Another approach is used in linear geostatistic (Cressie, 1993), where random function model of the phenomenon is used to estimate summing weights at each point using the kriging equations (Matheron, 1969).

Delaunay triangulation of P and subsequent linear or cubic interpolation within each triangle (Farin, 1985; Renka, 1996) could be also a choice to construct interpolation surface.

Next method originally suggested in (Hardy, 1971) defines the interpolation function as weighted sum of radially symmetric basis functions (RBF): $f(x, y) = \sum_{i=1}^N w_i \phi_i(x, y)$, where $\phi_i(x, y) = \phi(r_i)$ and r_i is Euclidian distance between the point (x, y) and p_i . Common choices for $\phi(r_i)$ include Hardy's Multiquadratic function $\phi(r_i) = \sqrt{r_i^2 + c^2}$, Gaussian $\phi(r_i) = e^{-cr_i^2}$, thin plate spline $\phi(r_i) = r^2 \log(r)$, Inverse Multiquadratic function $\phi(r_i) = 1/\sqrt{r_i^2 + c^2}$ (Franke and Nielson, 1991). Weights w are derived from a solution of a linear algebraic system of N equations. When N is high, the solution of such a system requires some special approach, e.g. Iterative Techniques (Faul, 2000), Compactly Supported RBF (Wendland, 1995) or Hierarchical Schema (Floater and Iske, 1998; Iske, 2000). The last method decomposes the given set of the points into a hierarchy $P_L \subset P_{L-1} \subset \dots \subset P_0 = P$ of $L+1$ nested subsets by means of thinning algorithm that generates a nested sequence of subsets from the scattered data set P in a way that the points in each subset are distributed as evenly as possible and their density increases smoothly. Then, for each level $j = L, \dots, 0$ the RBF interpolation function $\hat{f}_j$ defined on the set P_j is constructed and used as a trend for the next level.

This approach is most similar to MBA (multilevel B-Splines approximation) (Lee, *et al*, 1997) method used in this study. Instead of hierarchy of nested subsets, MBA uses the

hierarchy of nested control lattices with interpolation weights as described in detail below.

2. METHODOLOGY

In this section, the results of the prior data analysis followed by the selection of interpolation method described. The prior information (10 days of measurements at 200 locations) was used for two purposes, i.e. (i) to find the model of the phenomenon and (ii) to define the interpolation method, to estimate its parameters and uncertainty of the interpolation results.

2.1 USE OF PRIOR INFORMATION

The model of the phenomenon is based on the prior information analysis of the local variability in spatial law (correlation). We have constructed a model of the phenomenon under consideration as a sum of two components. The first one is the trend surface that is considered constant or very stable in time and represents large-scale spatial pattern of the natural radioactivity. Spatial Gaussian (possibly non-stationary) random function that represents stochastic variation of the spatial pattern at the fine scale is the second component.

a) Temporal stability of the large-scale spatial pattern was justified as follows. In the context of the model suggested, data values at each point are random variables with normal distribution. To estimate the sample distribution homogeneity at each location, Student's t-statistic was used. It allowed testing the belongings of each sample of prior dataset at given location to the normal distribution with parameters estimated with other 9 samples. The distribution of the t-statistics, collected over all locations during these tests, is shown on the Fig.1 below. The null hypothesis that samples belongs to the distribution of other data at the same location, is accepted for 95% of 2000 tests (5% critical value for the Student's test is near 2.3).

Figure 1
Student's t-statistic distribution for prior data

All the 10 samples at each location were also used to test if new dataset sample belongs to the same distribution (only the 1st “regular” dataset was used; the 2nd “joker” dataset was different in few locations that could be considered as outliers, and therefore wasn't used). The null hypothesis stating that new sample belongs to the distribution of prior data at the same location is accepted for all the 200 tests at 5% level (Fig.2).

Figure 2
Student's t-statistic distribution for new data

Distributions on Fig.1 and Fig.2 confirm our assumption that natural radioactivity mean at each location could be considered as stable.

If prior data are sequential in time (no information were available on that matter), some minor temporal drift at each location could be seen: correlation coefficient between the value at each location and the time step were calculated, and correlation coefficients distribution is show on Fig.3 below. Thus, we will consider the location means as stable and use them to estimate and tune the method parameters.

Figure 3
Temporal correlation coefficients distribution

b) Ten repeated measurements at each location of the prior dataset were considered sufficient to estimate the linear correlation coefficient between the values in neighbour locations. To select adjacency locations, the Delaunay triangulation of the given 200 known locations was used, and each correlation value between two locations of the triangulation was attributed to the edge of Voronoi diagram, corresponding the Delaunay triangulation. Correlation coefficient distribution is given on Fig.4, and fitted linear regression of the correlation on the distance between locations on Fig.5.

Figure 4
Neighbour correlation histogram

Figure 5
Correlation-distance dependence

2.2 SURFACE APPROXIMATION METHOD

In the context of the phenomenon model, we decompose the interpolation task into two subtasks. First, we need trend surface estimation, and second, we need some means to approximate the stochastic component (residuals from the trend surface).

The MBA algorithm (Lee, *et al*, 1997) was chosen for both subtasks because it is very fast and provides control over the degree of the surface generalization (smoothness), the “area of influence” for each known value and the (local) anisotropy. Short description of the method is given below.

Given a set of scattered data points

$$P = \{(x_i, y_i) | 0 \leq x_i < m, 0 \leq y_i < n, i = 1, \dots, N\} \text{ and } z_i = f(x_i, y_i),$$

the algorithm find approximating surface $\hat{f}(x, y)$ defined on a rectangular domain $\Omega = \{(x, y) | 0 \leq x < m, 0 \leq y < n\}$. A reconstructed surface is formulated to be uniform bicubic B-spline surface passing through P . The B-spline surface is defined by control lattice Φ overlaid on domain Ω , and containing weights, used for approximation. Without loss of generality, we assume that Φ is an $(m+3) \times (n+3)$ lattice that

spans the integer grid in Ω . Let ϕ_{ij} be the value of the ij -th control point (weight assigned to node located at position (i, j) of lattice Φ) for $i = -1, 0, \dots, m+1$ and $j = -1, 0, \dots, n+1$. Reconstructed function (approximated surface) $\hat{f}$ is defined in terms of these (for a while unknown) control points by

$$\hat{f}(x, y) = \sum_{k=0}^3 \sum_{l=0}^3 B_k(s) B_l(t) \phi_{(i+k)(j+l)}, \quad (1)$$

where $i = \lfloor x \rfloor - 1$, $j = \lfloor y \rfloor - 1$, $s = x - \lfloor x \rfloor$, and $t = y - \lfloor y \rfloor$. B_k and B_l are uniform cubic B-spline basis functions defined as

$$\begin{aligned} B_0(t) &= (1-t)^3 / 6, & B_1(t) &= (3t^3 + 6t^2 + 4) / 6, \\ B_2(t) &= (-3t^3 + 6t^2 + 3t + 1) / 6, & B_3(t) &= t^3 / 6, \end{aligned} \quad 0 \leq t < 1$$

The values of control points ϕ_{ij} are computed along the following lines, giving the least square solution to weights ϕ_{ij} (Hsu, *et al*, 1992):

$$\phi_{ij} = \left(\sum_{p \in P} \omega_p^2 \phi_p \right) / \left(\sum_{p \in P} \omega_p^2 \right), \quad \phi_p = \frac{\omega_p z_p}{\sum_{a=0}^3 \sum_{b=0}^3 \omega_{ab}^2},$$

where

$$\begin{aligned} \omega_p &= \omega_{kl} = B_k(s) B_l(t), \quad k = (i+1) - \lfloor x_p \rfloor, \quad l = (j+1) - \lfloor y_p \rfloor, \quad s = x_p - \lfloor x_p \rfloor, \\ t &= y_p - \lfloor y_p \rfloor. \end{aligned}$$

Having the control lattice weights ϕ_{ij} estimated (1) provides calculation of the approximation $\hat{f}(x, y)$ in any location of the domain Ω . Smoothness and accuracy of the approximation and the contribution of each data point to the value of $\hat{f}(x, y)$ at given location depends primarily on the control lattice size and spacing, so computationally efficient hierarchical schema is used.

2.3 MULTILEVEL SURFACE APPROXIMATION

A tradeoff exists between the shape smoothness and accuracy of the approximation function generated by the MBA algorithm, and a multiresolution algorithm is used to circumvent this tradeoff (Lee, *et al*, 1997). The algorithm makes use of a hierarchy of control lattices to generate a sequence of functions $\hat{f}_k$ by formula (1), whose sum approaches the desired approximation function. In the sequence, a function from a coarse lattice provides a rough approximation that is further refined in accuracy by the functions derived from inner lattices. The approximation function $\hat{F}$ is defined as

$$\hat{F}(x, y) = \sum_{i=1}^K \hat{f}_i(x, y),$$

where K is number of hierarchy levels, $\hat{f}_i(x, y)$ are functions computed with formula (1) for the control lattices Φ_i of size $(n^i + 3) \times (m^i + 3)$ (corresponding to domain Ω_i of size $n^i \times m^i$) in accordance with the following steps:

1. Let $i = 1$, $z_p^{(1)} = z_p$ for all $p \in P$ and domain Ω_1 have size $n \times m$.
2. Compute control lattice Φ_i for domain Ω_i of size $n^i \times m^i$ using data points $z_p^{(i)}$, and $\hat{f}_i(x, y)$.
3. Compute residuals $z_p^{(i+1)} = z_p^{(i)} - f_i(x_p, y_p)$, $p \in P$.
4. Let $i = i + 1$.
5. Repeat the process from step 2 for $i = 2 \dots K$.

If we start with control lattice Φ_1 covering the whole domain Ω_1 of size, say, 2×3 (any other domain needed could be linearly mapped onto the Ω_1 , preserving the control lattice size proportions fit to the interpolation area proportions to avoid the approximation anisotropy), then the control lattice Φ_1 have size 5×6 at first and most smooth approximation, and almost each data point contribute to the approximation value (4×4 sublattice of the Φ_1 is used to calculate approximation at each location at step 2). Subtracting the approximation from each data value, then we map our data locations (with new, residual, values) on the new domain Ω_2 of size $2^2 \times 3^2 = 4 \times 9$, and use control lattice Φ_2 of size 7×12 to calculate new approximation at step 2, etc. After the control lattice achieve size for which only one data point fall in any 4×4 sublattice, formula (1) gives exact interpolation, and sum of all functions, constructed at step 2, give exact interpolation to the initial data set.

This method gives better results if data values have zero mean and no spatial drift, so in our implementation of MBA we subtract linear drift from the data before the MBA been applied and add it to the MBA result.

By variation of the initial domain Ω size, the influence area of each data point, resulting in approximation smoothness and generalization, can be controlled. Other “control” is the mapping on the domain Ω method, which could be used to control (local) anisotropy.

2.4 MBA METHOD TUNING “CONTROLS”

Method parameters are denoted as $MBA(n, m, K)$ or simply (n, m, K) if appropriate, where $n \times m$ is initial domain Ω size (giving $(n + 3) \times (m + 3)$ control lattice Φ size), and K is a number of control lattices hierarchy levels. The interpolation area is always expanded up to square box to avoid undesirable anisotropy. For our case study, we symmetrically expand the area in X direction to the same size, as it has in Y direction).

Main “controls” for the MBA approximation are the control lattice size, the number of the levels in hierarchy and (non-)linear mapping of real-world coordinates on control lattice coordinates. The size of the coarsest lattice Φ_1 controls the “area of influence” for known data; the number of hierarchy levels controls the approximation smoothness and accuracy (up to exact interpolation); coordinate mapping controls directional anisotropy in case that coordinates are linearly scaled in certain direction. More general, it controls pattern of the spatial similarity for non-linear mapping.

The control lattice size and number of hierarchy levels influence are shown on Fig.6 below. One can see more smoothed surface for the MBA(2,2,4) comparing with MBA(2,2,10). Despite the fact that MBA(2,2,10) and MBA(48,48,5) give the same exact values at data locations, the picture in other locations is quite different. More dense initial control lattice for MBA(48,48,5) reveals linear trend used in our MBA implementation, masked by the data influence in MBA(2,2,10).

Figure 6
Control lattice influence on the interpolation results for 1st “regular” dataset.
From left to right: MBA(2,2,4), (2,2,10) and (48,48,5) parameters used.

The coordinate mapping influence on the approximation results is shown on Fig.7 below.

Figure 7
Anisotropy caused by the coordinate mapping (from left to right): 2nd “joker” dataset MBA(2,2,4) approximation, 2nd “joker” dataset MBA(2,2,4) approximation with 1:2 anisotropic mapping in X direction. Voronoi diagram of outliers is overlaid in white.

Spatial law model (usually variogram) is traditionally reduced in geostatistics to reference isotropy model using distance transform. Model is considered as a function of some “reference” distance between the locations, and the actual value of that function is calculated from the transformed distance. This approach requires stationarity. The spatial law variation depends only on the direction but not on the location of the data values in the geographic space (Isaaks, Srivastava, 1989; Pannatier, 1996; Chiles, Delfiner, 1999).

In case of non-stationarity, there are two choices, i.e. to calculate location-dependence coordinate transform function or to transform the coordinates themselves in the way that the distance between the transformed data locations is exactly equal to the “reference” distance. To calculate an appropriate spatial law model Delaunay triangulation can be used. Since proper distance between the nearest points connected by edges in triangulation is most important, all points are moved to make the transformed coordinate distance

between them equal to the “reference” one. We used relaxation method that minimised the sum of squared triangulation edges distances residuals from “reference” distances.

The reason to use the coordinate transform described above for the MBA is as follows. In accordance with the duality of kriging equations (Goovaerts, 1997), kriging is considered as functional approximation of the data values using specific basis (adapted covariance functions). Therefore, the same idea to use the coordinate transform commonly uses in geostatistics for calculating “reference” distance can be applied with other bases including MBA.

Voronoi diagram with SIC2004 data visualizes correlation between the points of prior dataset. This provides estimation of “discontinuity lines”, possible anisotropy and other heterogeneities in exploratory data analysis. The correlation dependence on the direction and distance between the points is shown on Fig.8a. The thicker lines correspond to lower correlation between the points separated by the line. Fig.8b shows discontinuity based on the difference in Student’s t-statistic (see section 2.1) for the 2nd “joker” dataset. Area with outliers was separated from the other “regular” values. Fig.8c shows artificial regional variation and each data location shift in a manner, that data correlation versus new distance fit the reference correlogram.

Figure 8
 (a) 1st “Regular” dataset correlation “discontinuity lines”;
 (b) 2nd “Joker” dataset t-statistic “discontinuity lines”;
 (c) Artificial regional variation and points shift needed to fit “reference” model

The coordinate transform of arbitrary location could be generated with exact interpolation by means of MBA or other techniques.

2.5 MBA COMPARISON WITH KRIGING IN TREND SURFACE APPROXIMATION

To ascertain the MBA applicability for the first subtask (trend surface estimation), it was compared with kriging for the sample mean at each of 200 locations as input data. To set kriging parameters, linear variogram with nugget effect was fitted (Fig.9). Experts found

no significant anisotropy and linear drift. Ordinary kriging of the sample means (of all the data) was used to model the trend surface.

Figure 9
Omnidirectional sample means variogram and variogram surface

Quality of the results was estimated with “leave-one-out” method. Ordinary kriging cross-validation residuals distribution is shown on Fig.10.

Figure 10
Ordinary kriging cross-validation residuals for the trend surface approximation

The same sample means were also used to estimate the trend surface by means of MBA with the parameters, automatically selected as best ones - MBA(2,2,4), (Fig.11).

Figure 11
MBA(2,2,4) cross-validation residuals for the trend surface approximation

Both methods show similar behavior, although the MBA is much faster and gives some less residuals variance. Ppossibly kriging residuals variance could be decreased with the variogram model tuning.

2.6 MBA COMPARISON WITH KRIGING IN STOCHASTIC COMPONENT APPROXIMATION

The MBA comparison with kriging was also performed for the second subtask, i.e. stochastic component estimation. To prepare the stochastic component value at each location, location sample mean was subtracted from all location values (prior and new; sample mean was calculated from the prior data).

Simple kriging weights at given location were calculated with zero-mean data prepared and kriging equations in Delaunay triangulation neighbours were solved using covariances estimated from the prior dataset. The weights were then used in estimation of the stochastic component at each location for the 1st “regular” dataset. Quality of the results estimated by “leave-one-out” cross-validation method is shown on Fig.12.

Figure 12
Simple kriging cross-validation residuals distribution for the stochastic component approximation

MBA aproximation quality for the 1st “regular” dataset (sample means subtracted) is shown on Fig. 13.

Figure 13
MBA(2,2,4) cross-validation residuals distribution for the stochastic component approximation

Although the MBA not uses prior information and only new data were used for the cross-validation, the results obtained are similar to those for kriging, that used specific location correlation pattern.

The statistics above show that main source of approximation error for MBA is an improper trend surface estimation, therefore MBA was applied to the prior dataset without its decomposition into two components (Fig. 14).

Figure 14
MBA(2,2,4) cross-validation residuals distribution for the prior dataset

The results obtained are very similar to those obtained with MBA for the trend surface approximation. The variance in this case is near the sum of the trend surface residuals variance plus stochastic component residuals variance ($126.22 \approx 121.59 + 5.17$). Thus, direct approximation could be done with no significant reduction of the result quality. From our experience of the MBA algorithm application to geological modelling, this could be due to the wide spacing sampling network that prevents estimation of the trend surface with the required accuracy.

2.7 TUNING THE ALGORITHM

The coordinate transform was used in this case to fit the reference model (linear regression is shown on Fig.2). No significant improvement in the model fit was achieved, hence it was decided by experts there was no significant variation in the spatial law to be corrected and untransformed data coordinates were used later on.

To obtain final interpolation results for the 1st “regular” and 2nd “joker” datasets, MBA(2,2,4) approximation of the datasets was used instead of the trend surface estimation from the prior dataset followed by residuals approximation.

Taking into account the prior information volume and the MBA method speed, it was decided to select optimal MBA parameters automatically by means of “leave-one-out” cross-validation method on all the prior data. For each one-day subset of known data values (i) one data value was eliminated, (ii) MBA was applied to other data, and (iii) approximation value at the eliminated location was calculated and compared with that original one. Exhausting search was used to select parameters giving best mean squared error (MSE) for prior data.

The MBA automatically selected parameters give very smooth approximation. Initial domain size was chosen to be 2×2 , and only 4 hierarchical levels were used.

3. RESULTS

3.1 OVERALL RESULTS

The table below gives the minimum, maximum, mean, median and standard deviation of the 808 estimated values and those observed for two datasets investigated. To estimate

the quality of the model suggested, trend surface model was also constructed that did not depend on the new data by MBA(2,2,4) interpolating of prior data means. Ten MBA(2,2,4) approximations of prior data were made and averaged as another estimation of the trend surface.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (first data set)	57.0	180.0	98.0	98.8	20.0
Estimates (first data set)	66.2	127.2	96.4	98.3	15.4
Observed (second data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (second data set)	66.9	341.8	104.8	98.6	38.5
Ten-day models average	54.4	157.0	94.6	95.3	17.7
Trend surface model	62.2	136.6	95.1	97.3	15.4

Table 1
Comparison of the estimated and measured values (nSv/h)

Mean absolute error (MAE), the bias or mean error (ME) and the root mean squared error (RMSE) of the predictions are defined as follows ($n = 808$ locations)

$$MAE = \frac{1}{n} \sum_{i=1}^n |a_i^* - a_i|,$$

$$ME = \frac{1}{n} \sum_{i=1}^n (a_i^* - a_i),$$

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

where a_i^* is the estimated value at location i and a_i is the true value. Pearson's correlation coefficient r and Spearman's coefficient r_s of rank correlation between the estimated and true values are also given.

Data sets:	MAE	ME	Pearson's r	Spearman's r_s	RMSE
MBA estimation					
First data set	9.3	1.6	0.78	0.78	12.6
Second data set	22.2	0.6	0.41	0.74	76.4
Trend surface model					
First data set	9.6	3.0	0.77	0.77	13.0
Second data set	17.0	10.4	0.23	0.77	82.2

Table 2
Comparison of the errors

1) ESTIMATIONS:

Two maps below present filled contour lines obtained for the two datasets with colour scale intervals of 10 nSv/h. Two different symbols indicate the locations of the samples, i.e. crosses for the locations of estimated values and empty squares for the locations of the input values.

Figure 15
Isoline levels (nSv/h) for the 1st set (left) and the 2nd set (right)

2) UNCERTAINTY:

Figure 16
Isoline levels showing the uncertainty. From left to right: absolute error, associated to the estimations obtained from the prior data cross-validation; 1st dataset interpolated absolute error; 2nd dataset interpolated absolute error (positive residuals are cross-hatched)

The averaged cross-validation absolute error calculated on the prior dataset was used as a measure of the uncertainty for both datasets. The map below presents the levels of uncertainty associated with the estimations shown on the previous figures. Besides, we used the same 200 known locations for “leave-one-out” cross-validation of 1st “regular” and 2nd “joker” datasets to estimate the cross-validation residuals. To produce the resulting uncertainties map, residuals spatial interpolation was used by means of MBA(8,8,8) that gave exact interpolation at 200 known points.

Although the first map on Fig.16 holds statistics based on ten cross-validation errors and hence produces statistically valuable estimation of average error at 200 known locations, it did not allow efficient estimation of the uncertainty over the hole domain. The comparison with the interpolation error at 808 validation points shows 0.25 correlation coefficient for the 1st “regular” dataset, and zero correlation for the 2nd “joker” dataset. Second map shows uncertainty estimations, similar to the first map, and third map reveals outliers area.

3.2. DETECTING ANOMALIES AND OUTLIERS

The model of the phenomenon that assumes the decomposition into the trend surface and random function allows detecting anomalies and outliers. Student’s t-statistic provides selecting potential outliers, and the Voronoi diagram of the data permits to delineate them in geographic space (Fig. 8b) and process data with outliers. First, we have truncated the outliers values at 0.1% significance level, and used those truncated data for MBA approximation, producing Model A. Then we took original untruncated data for MBA approximation and produce Model B. After that, we had “glued up” Model B into the Voronoi diagram area (delineating outliers) of Model A. The example of this approach to the 2nd “joker” dataset is shown on Fig.17 below.

Figure 17

Outliers modelling (see text): Model A, Model B, models composition, models seam zoom (with white seam line)

Figure 18
3D map showing extreme values found in the 2nd set (vertical scale in nSv/h)

Figure 19
3D "seamed" map showing extreme values found in the 2nd set (vertical scale in nSv/h)

Although the resulting picture seems in this case unnatural, the seam is smooth enough because smooth weighting schema was applied with weights depend on the distance to seam line.

The estimates obtained using MBA(2,2,4) for the 2nd set are presented in 3D on Fig.18 in black and white palette to emphasize the extreme values. Seamed model is shown on Fig.19.

4. DISCUSSION

MBA interpolation method proposed in the paper could be used as an automatic mapping means in emergency situations. It shows excellent performance and quite good interpolation quality while “regular” data, free from outliers, are used. It has also enough flexible tuning “controls” to be adapted automatically to specific data given. Outliers automatic detection and delineation techniques proposed in this paper make it possible to overcome problems without losses in performance.

Special computer program for Windows 2000 was developed for SIC2004 exercise and used for all the calculations mentioned. The program developed by authors is based on MBA, variography, kriging and grid and vector libraries. Graphical presentation of the results is given with SURFER v8.0 program and statistical calculations and presentations were made with Statgraphics Plus v5.1 program.

It could be concluded from the comparison of MBA and kriging in stochastic component interpolation, that MBA outperforms kriging in approximation speed, and produces the results of the same quality without preliminary data analysis (kriging requires exact covariance model to be known at each location, that is impossible for non-stationary spatial law).

Comparison of the MBA interpolation results with the trend surface model for the validation part of the 1st “regular” dataset shows their similarity. Trend surface estimation was the main source of the result errors. MBA failed to restore this surface with the acceptable quality. In “joker” case, the MBA shows good enough delineation of the high-values area, but does not show the direction of the plume.

The average computational time for each of two sets was below 1 second using standard PC (1.6GHz P4).

5. CONCLUSIONS

The results presented show that the MBA interpolation has average quality both for the 1st “regular” dataset and for the 2nd “joker” dataset, although additional means could in later case improve the situation and automatically detect, delineate and interpolate outliers. The main advantage of MBA approach is the possibility for automated selection of parameters, high speed of computation and its ability to process huge data sets.

REFERENCES

- Chiles, J.P., Delfiner, P. (1999). *Geostatistics. Modeling Spatial Uncertainty*. J. Wiley&Sons, 695 p.
- Cressie, N. (1993) *Statistics for Spatial Data: Revised Edition*. New York: John Wiley and Sons, 928 p.
- Dubois, G., Galmarini, S. (2005) “Introduction to the Spatial Interpolation Comparison (SIC) 2004 exercise and presentation of the data sets”. *Applied GIS* ; 1 (2) : 09-01 to 09-11.
- Farin, G. (1985) “A modified Clough-Tocher interpolant”, *Comput. Aided Des.*, no 2, pp.19–27.

- Faul, A. C. (2000) Iterative Techniques For Radial Basis Function Interpolation. Ph.D. thesis, University of Cambridge.
- Floater, M.S., Iske, A. (1998) "Multistep Scattered Data Interpolation using Compactly Supported Radial basis Functions", BIT 38, pp.705-720.
- Franke, R., Nielson, G. M. (1980) "Scattered data interpolation of Large Sets of Scattered Data", Intl. J. Numerical Methods in Eng., vol 15, pp.1,691-1,704.
- Franke R., Nielson, G. M. (1991) "Scattered data interpolation and applications: a tutorial and survey" in Geometric Modelling Methods and Applications, H. Hagen and D. Roller, Eds., pp.131-160, Springer, Berlin.
- Goovaerts, P. (1997). Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New-York, 483 p.
- Hardy, R. L. (1971) "Multiquadric equations of topography and other irregular surfaces", J. Geophys. Res., no 76(8), pp.1905-1915.
- Hsu, W.M., Hughes, J.F., Kaufman, H. (1992). "Direct manipulation of Free-Form Deformations", Computer Graphics (Proc. SIGGRAPH '92), vol.26, no.2, pp.177-184.
- Isaaks, E.H., Srivastava, R.M. (1989). An Introduction to Applied Geostatistics, Oxford: Oxford University press, 803 p.
- Iske, A. (2000) "Hierarchical Scattered Data Filtering for Multilevel Interpolation Schemes", Mathematical Methods for Curves and Surfaces, pp.211-220.
- Lee, S., Wolberg, G., Shin, S.Y. (1997) "Scattered Data Interpolation With Multilevel B-Splines". IEEE Trans. Visualization and Computer Graphics, vol. 3, no 3, pp.228-244.
- Matheron, G. (1969) Le Krigeage universe. Cahiers du Centre de Morphologie Mathematique de Fontainebleau, Vol.1. - Fontainebleau, 59 p.
- Pannatier, Y. (1996) Variowin. Software for Spatial Data Analysis in 2D. Springer, 91 p.
- Renka, R. J. (1996) "Algorithm 752; SRFPACK: software for scattered data fitting with a constrained surface under tension", ACM Transactions on Mathematical Software, Vol.22, No. 1, March 1996, pp.9-17.
- Shepard, D. (1968) "A two-dimensional interpolation function for irregularly spaced data", Proceedings of the 23rd ACM National Conference, pp.517-524, ACM, NY.
- Wendland, H. (1995) "Piecewise Polynomials, Positive Definite and Compactly Supported Radial Basis Functions of Minimal Degree", Advances in Computational Mathematics, no.4, pp.389-396.
- Wolberg, G. (1997) "Nonuniform image reconstruction using multilevel surface interpolation". ICIP(1), pp.906-916.

Cite this article as: Saveliev, Anatoly A; Romanov, Andrey V; Mukharamova Svetlana S.. 'Automated mapping using multilevel B-Splines'. *Applied GIS*, Vol 1, No 2, 2005. pp. 17-01 to 17-19. DOI: 10.2104/ag050017

Copyright © 2005 Anatoly A. Saveliev, Andrey V. Romanov, and Svetlana S. Mukharamova