Stochastic Environmental Research and Risk Assessment 15 (2001) 18–32 © Springer-Verlag 2001

Wavelet analysis residual kriging vs. neural network residual kriging

V. Demyanov, S. Soltani, M. Kanevski, S. Canu, M. Maignan, E. Savelieva, V. Timonin, V. Pisarenko

Abstract. This paper deals with the problem of spatial data mapping. A new method based on wavelet interpolation and geostatistical prediction (kriging) is proposed. The method – wavelet analysis residual kriging (WARK) – is developed in order to assess the problems rising for highly variable data in presence of spatial trends. In these cases stationary prediction models have very limited application. Wavelet analysis is used to model large-scale structures and kriging of the remaining residuals focuses on small-scale peculiarities. WARK is able to model spatial pattern which features multiscale structure. In the present work WARK is applied to the rainfall data and the results of validation are compared with the ones obtained from neural network residual kriging (NNRK). NNRK is also a residual-based method, which uses artificial neural network to model large-scale non-linear trends. The comparison of the results demonstrates the high quality performance of WARK in predicting hot spots, reproducing global statistical characteristics of the distribution and spatial correlation structure.

1 Introduction

Problem of spatial data analysis and mapping is essential in many environmental applications. Very often spatial data analysts are facing the problem of multiscale correlation and non-stationarity. The problem of modelling

V. Demyanov (✉), M. Kanevski, E. Savelieva, V. Timonin
Nuclear Safety Institute RAS, 52, B. Tulskaya, Moscow 113191, Russia

S. Soltani
Technological University of Compiegne, France

S. Canu
Institut National des Sciences Appliquees (INSA), France

M. Maignan
University of Lausanne, Switzerland

V. Pisarenko
International Institute of Earthquake Prediction Theory and Mathematical Geophysics (MITPAN), Moscow, Russia

The work was supported in part by grant INTAS 96 1957 and INTAS 97 31726. The authors appreciate the organisers of SIC'97 contest for providing the data for the present case study.

large-scale non-linear trends and predicting the residuals is common in geostatistics and often occurs for spatial data of different kind. This problem is usually solved by separate trend modelling with polynomial functions, moving window estimation, or using higher order statistical moments. An alternative approach is to use learning based algorithms to model large-scale non-linear trend. The modelling can be achieved using, for instance, wavelets or artificial neural networks (ANN). The method based on ANN was introduced by Kanevsky et al. (1996) and is called Neural Network Residual Kriging (NNRK). NNRK proved to be a very powerful algorithm in various applications (Kanevsky et al. 1996, 1997a, b, c).

2
Wavelet analysis residual kriging

Our approach consists in combining a large scale trends modelling based on the wavelet analysis and residual kriging. The trend modelling – described below – is treated as a functional estimation problem.

2.1
Wavelet estimation

The problem is to estimate a functional dependency between an input x and an output y of a system. For this purpose, we use a set of observations (x_i, y_i), $x_i \in \Re^d$, $y_i \in \Re$, $i = 1, \ldots, N$ which follows the probability law $P(X, Y)$. The relationship between x and y is given is of the following form

$$Y = r(X) + \varepsilon \ ,$$

In the general case, the nature of this dependency is unknown, so we use flexible. This kind of estimators has the ability to approximate a large set of functions. The performance of the estimation is measured by the quadratic cost function given by

$$C_{\text{gen}}(f) = E_{XY}(Y - f(X))^2 = \iint (Y - f(X))^2 P(X, Y)\mathrm{d}X\,\mathrm{d}Y \ .$$

The minimum of C_{gen} in the space F is the regression function given by

$$r(x) = E(Y/X = x) \ .$$

Since the probability law $P(X, Y)$ is unknown, a common approach consists in estimating the regression using the empirical risk given by

$$C_{\text{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 \ . \qquad (1)$$

Unfortunately, this minimization problem is ill-posed because the set F contains an infinity of functions for which $C_{\text{emp}} = 0$. Most of these functions learn noise as well as useful information and thus are not considered as valuable estimation of the regression.

A common method that can be used to avoid this problem is to regularize it by adding a high frequency penalty term. Indeed, the presence of high frequencies imply in most cases learning noise. Another approach consists on searching within sets F_m subsets of the space F. The index m corresponds to the estimator size, the set F_m contains all the subspaces (noted G_m^k) which are generated by m functions, the multi-index k includes the parameters of these functions. The estimator related to each of the subspaces is obtained by minimization of the empirical criterion. Our objective is to find the size for which the risk is minimal, i.e.

$$F_{m^*} = \arg\min_{F_m \in F} E_{XY}(Y - f(X))^2 \tag{2}$$

The weights of these functions are estimated by the empirical risk. The estimator used in this paper and presented below is based on wavelet frames decompositions, which means that the some of the multi-index k components contain dilatations and translations.

2.2 Wavelet frames based estimator

The estimator presented in this section is defined by a finite linear combination of wavelets extracted from a frame (Daubechies, 1992). Wavelets and frames are concepts with interesting properties in the context of regression estimation. In order to introduce the details concerning the estimator, some fundamental notions are presented below.

2.3 Wavelet frames

A frame is a set of non-orthogonal and complete functions $\varphi_\lambda(x)$[1] which satisfies some conditions given in (Daubechies, 1992) such that every function $L^2(\Re^d)$ can be written as follows:

$$f(x) = \sum_\lambda c_\lambda \varphi_\lambda(x) \ .$$

The coefficients c_λ are given by $c_\lambda = \langle f \mid \varphi_\lambda^* \rangle$, where functions φ_λ^* constitute a dual frame. For a given frame, there exists an infinity of possible dual frames. It means that the decomposition is not unique (Daubechies, 1992; Candès, 1995). For instance, one can search for the dual frame which provides the decomposition with minimal $\sum_\lambda c_\lambda^2$.

In the context of multidimensional wavelet frames theory, one can define projective or radial frames. We restrain ourselves to the first case. For frames with projective functions $\varphi_\lambda(x)$, called ridgelets in (Candès, 1995), the wavelets are defined as follows

$$\varphi_\lambda(x) = a_0^{k/2}\varphi(a_0^k\langle x|u_{l\theta_k}\rangle - mt_0), \quad \lambda. = (k, l, m) \ .$$

where $a_0 > 1,\ t_0 \in \Re$ are called elementary dilatation and translation, vectors $u_{l\theta_k}$ are directive unitary vectors are the unitary sphere, $l \in Z^d$ and θ_k is the elementary $d-1$-tuple directions angles at scale k.

In both cases (projective and radial), elementary parameters must satisfy conditions given by Caldern–Zygmund theory (Daubechies, 1992; Candès, 1995).

[1] In the general case, λ is a multi-index of integers.

The wavelets must also satisfy the admissibility condition which is an oscillation constraint. Among the set of coefficients c_λ, a lot are close to zero because of spatio-frequency properties[2] of the function $\varphi(x)$. The idea of wavelet frames-based estimator is precisely to use this property for reducing the estimator size.

2.4 Wavelet frames based estimator

The estimator $\hat{f}_W(x)$ is written as a combination of the wavelets which generates a subspace $G^{k^*}_{m^*} \subset F_{m^*}$. Recall that $G^{k^*}_{m^*}$ is a subspace generated by m^* wavelets. The estimator $\hat{f}_W(x)$ is a linear combination of m wavelets

$$\hat{f}_W(x) = \sum_{\lambda \in G^{k^*}_{m^*}} c_\lambda \hat{\varphi}_\lambda(x) \quad , \tag{3}$$

The first step consists in estimating the best size of the estimator using the criterion given in (2) (i.e. m^*). The second step consists in extracting the best m^* wavelets in the frame to be used for the estimation (i.e. space $G^{k^*}_{m^*}$ or simply k^*). In our case, the space F is equal to $\ell^2(\Re^d)$. The minimization of the criterion given in (2) is not practically possible because the space $\ell^2(\Re^d)$ is of infinite dimension. Thus, the space F is restricted to become a finite dimension space within which we can find an acceptable solution. This restriction is based on the information explanation power of each wavelet in the frame. Actually, only the lower dilatations are considered, the higher ones are eliminated because the related wavelets

(i) don't include a sufficient number of observations in their support to estimate the related coefficient,
(ii) correspond to higher frequencies[3] which are considered as noise. This key point can be seen as an assumption on the regularity of the function.

For each of the considered dilatation, we remove wavelets with a few observations in their support. In this way, a library L of finite size is constructed. The space $\ell^2(\Re^d)$ is replaced by this library. This modification leads to the following criterion

$$F_{m^*} = \arg\min_{F_m \in L} E_{XY}(Y - [\arg\min_{\hat{f} \in F_m} C_{\text{emp}}](x))^2$$

The problem can now be solved in finite time, but it is combinatorial because we must consider all possible spaces with elements of L.

One can avoid this problem by using subset selection methods. This method uses a sequential procedure for selecting the best approximating functions in the library L (Zhang, 1995; Breiman, 1996). The obtained solution is of course sub-optimal, but is quite satisfactory. The generalization cost C_{gen} is also unknown because the probability law $P(X, Y)$ is unknown, the resampling methods (e.g. cross-validation or bootstrap) on the raw data are commonly used to estimate it (Efron and Tibshirani, 1993). These methods generate randomly some new data sets from the original one. After selecting the optimal size, the subset selection

[2] The wavelets are rapidly vanishing both in space and frequency domains.

[3] The wavelets are band-pass filters.

method is applied to the original learning set to select the m^* wavelets that fits more the data.

2.5
Analysis of the residuals, structural analysis and kriging

Residuals obtained after the learning phase are analyzed with the help of a geostatistical approach. It has been found that, unlike original data, residuals have shown a stationarity property and well behaved associated semivariograms. Spatial continuity is modelled with the help of variogram model, that is a measure of the spatial continuity, i.e. the correlation between the samples in space. Raw variogram can be estimated using the method-of-moment as (Goovaerts, 1997):

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (V(x_i) - V(x_i + \mathbf{h}))^2$$

The raw variogram is modelled using conditionally negative definite functions. This model is then used for the kriging estimation of residuals. Kriging is the best linear unbiased estimator in the sense that it minimizes the variance of the estimation error. The ordinary kriging estimate is given by (Isaaks and Shrivastava, 1989):

$$V^*_{\mathrm{OK}}(x_0) = \sum_{j=1}^{n} w_j \cdot V(x_j)$$

where the weights are obtained by solving the following system of equations:

$$\sum_{j=1}^{n} w_j \gamma_{ij} - \mu = \gamma_{i0}, \quad \forall\ i = 1, \ldots, n$$

$$\sum_{i=1}^{n} w_i = 1$$

Ordinary kriging also provides an estimation variance, which quantifies the quality of the prediction:

$$\sigma^2_{\mathrm{ROK}} = \sum_{i=1}^{n} w_i \gamma_{i0} + \mu$$

It is worth noting that this estimation variance depends only on the sampling locations (monitoring network) and not on the sampling values themselves.

The final WARK predictions are obtained as a sum of the global trend wavelet interpolation estimates and kriging estimates of the residuals:

$$V^*_{\mathrm{WARK}} = V^*_{\mathrm{wavelet}} + V^*_{\mathrm{OK}}$$

3
Neural network residual kriging

Neural network residual kriging model have been widely described (e.g., Kanevski et al., 1996). We will briefly describe here the major outlines of the algorithm.

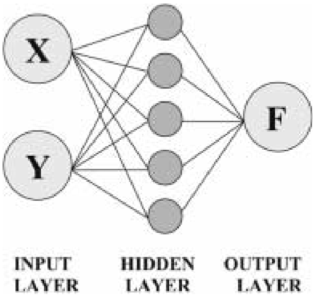

Fig. 1. Structure of multi-layer perceptron used for spatial prediction

A multi-layer perceptron (MLP) neural network (see Fig. 1) was applied for modelling general trends. Basic MLP architecture consists of weighted summation units (neurones) which form input, output and hidden layers. Linear connections (sinapses) link neurones of different layers. Input layer consists of two neurones to which spatial coordinates X and Y of the samples are exposed. Output layer consists of one neurone corresponding to the considered spatial function F. There can be one or two hidden layers, which are essential for modelling non-linearities. The number of hidden neurones can vary and its selection is based on testing and validation of the MLP's performance.

Modification of classical vanilla backpropagation algorithm was implemented for training the MLP. The algorithm proceeds as follows:

- the weights are initialized using genetic optimization or simulated annealing algorithm;
- the weights are modified during the training procedure. MLP training employs gradient algorithms (e.g. conjugate gradients, Levenberg–Marquardt). One of these algorithms is applied to find the minimum of the optimization criterion (mean square error);
- simulated annealing is used in order to avoid local minimum;
- validation and testing is performed after MLP has been trained. For instance, accuracy test is estimating function values at training points.

Residuals remaining at the training points are explored and predicted with geostatistical models as it was described above and ordinary kriging estimates of the residuals are computed. The final NNRK prediction is a sum of general trend modelled with MLP and ordinary kriging estimates of residuals:

$$V^*_{\mathrm{NNRK}} = V^*_{\mathrm{MLP}t} + V^*_{\mathrm{OK}}$$

4
Case study

The data chosen for the case study is the rainfall amount in Switzerland, averaged over periods of 10 days for year 1986. The task was to estimate 367 validation points using only 100 input samples for analysis (see Fig. 2). The data were obtained from the international contest Spatial Interpolation Comparison (SIC'97), organised through AI-GEOSTAT mailing list (http://curie.ei.jrc.it/SIC97).

The spatial pattern of the data is formed as two parallel bumps running in the SW–NE direction. Anisotropic spatial correlation is present at a local scale (50–100 km) within the bumps and at a global scale (over 120 km) between the bumps (see Fig. 3a). Presence of both positive, negative and around zero trends is demonstrated in different directions by the drift function (see Fig. 3b) (Kanevski et al., 1998).

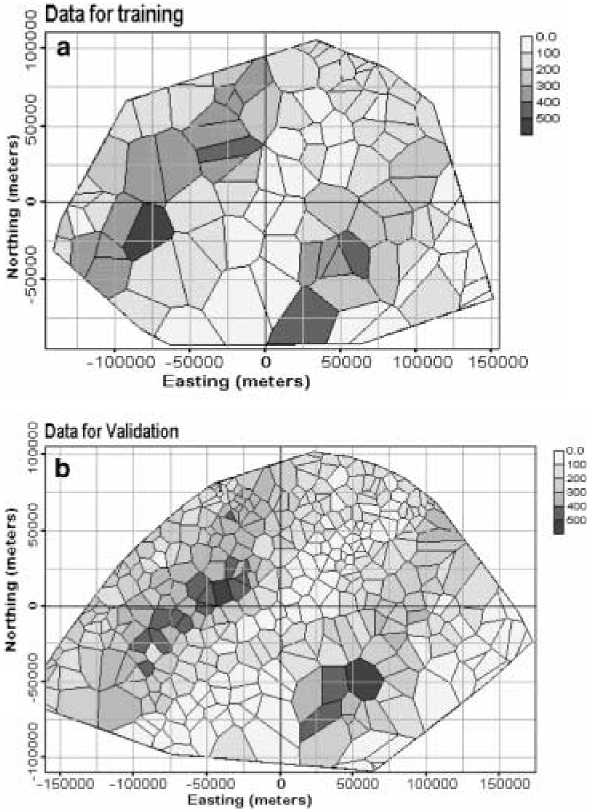

Fig. 2a, b. Voronoi polygons for training **a** and validation **b** monitoring networks

Sequential application of wavelet algorithms and kriging gave very good result. Wavelet interpolation method was able to model large-scale spatial structure and capture global characteristics of the pattern (see Fig. 6c). The number of chosen wavelets was $m^* = 12$ in (2). The residuals, remained at the sampling locations, feature spatial correlation at small scale (20–25 km), which is shown by omni-directional variogram for residuals (see Fig. 4) and variogram rose for the residuals (see Fig. 5a). This fact means the presence of some correlation that has not been modelled by wavelet analysis. The variogram structure of the residuals features slight anisotropy illustrated by variogram rose (see Fig. 5a). The variogram model for residuals was fitted with a nested structure of spherical models (see Fig. 5). Stationary behaviour of residuals allowed to apply kriging model which significantly improved wavelet interpolation results. The final result is the sum of wavelet and kriging estimates.

The final WARK estimates are validated with the help of the validation set (the 367 samples, which have not been used in the previous analysis). WARK prediction at validation points is presented in Fig. 6d as triangulation contours of the data. Along with the contours of validation measurements there are contours for Wavelet estimates, ANN estimates and NNRK estimates at validation points.

The analysis of the contour maps in Fig. 6 shows that ANN gives the smoothest estimation pattern which represents only two major large scale spots in the west and in the east. Wavelet estimate is more variable than one of ANN. WARK estimate in comparison to pure wavelet method detects hot spots and small scale peculiarities. WARK estimate is similar to NNRK estimate. There can be pointed out some areas where one method shows better performance. Thus, NNRK is

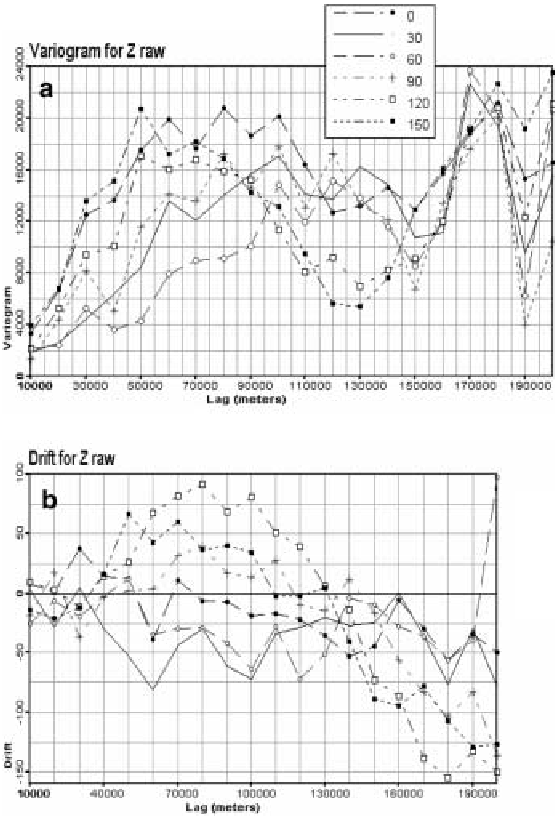

Fig. 3a, b. Directional raw variograms **a** and directional drift **b**

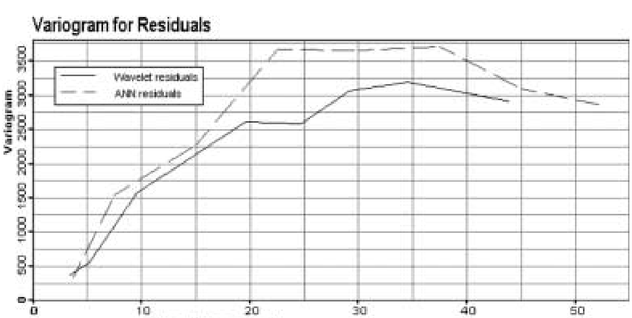

Fig. 4. Omnidirectional variograms for residuals at training points after wavelet interpolation (dashed line) and for ANN residuals (solid line)

better in predicting the low value region between the bumps and the low value spot within the bumps themselves. WARK demonstrated better precision in reproducing both location and value of the hot spots in the bumps.

5
Discussion

WARK performance has been compared with Neural Network Residual Kriging (NNRK). NNRK is also a sequential method of modelling large-scale trend with

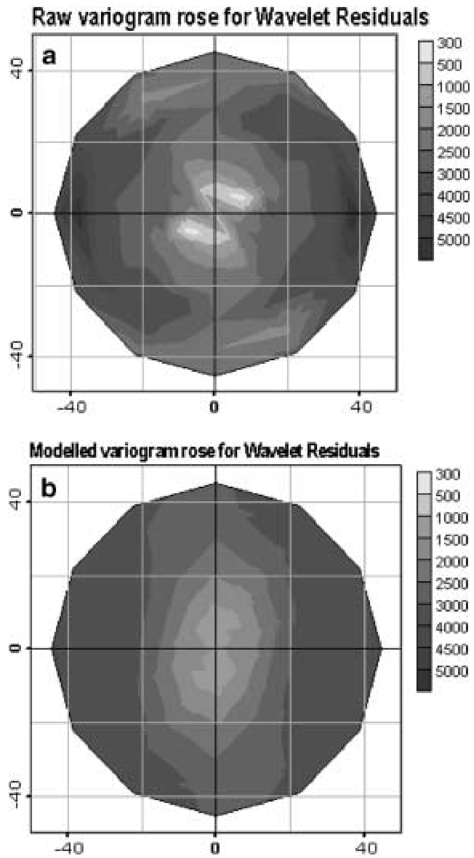

Fig. 5a, b. Raw **a** and modelled **b** variogram surface contours for the residuals after wavelet interpolation

Table 1. Validation: summary statistics on raw samples and estimates with different methods

Statistics	Raw data	ANN	Wavelet	NNRK	WARK
Mean	185	186	191	181	187
Median	162	178	174	160	167
Minimum	0	0	0	0	0
Maximum	517	426	450	514	513
Lower quartile	100	101	114	105	109
Upper quartile	264	262	270	256	267
SD	111	101	110	105	112
Skewness	0.58	0.37	0.24	0.66	0.43
Kurtosis	−0.37	−0.58	−0.70	−0.08	−0.36

artificial neural networks and kriging applied to the residuals. NNRK provided very good results for different kinds of highly variable data featuring spatial correlation at different scales.

Summary statistics of the measured and estimated distributions is given in Table 1. It can be seen that both WARK and NNRK provide good mean and median estimates, as well as maximum and quartile values. WARK seems to give better dispersion of the estimated distribution and kurtosis of WARK estimates is very close to the one of the raw measurements.

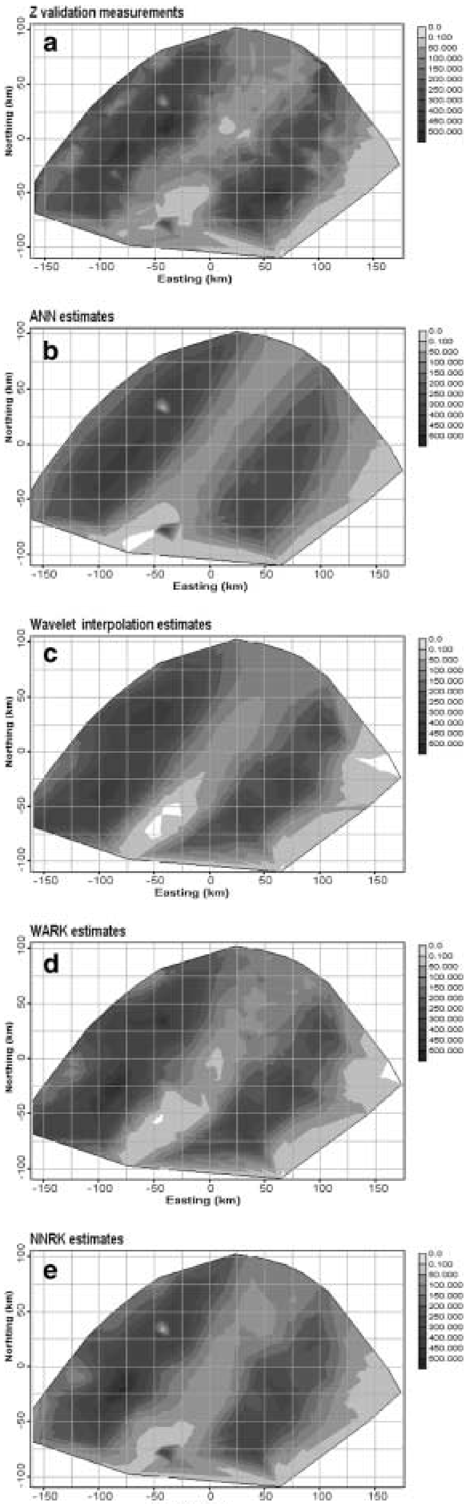

Fig. 6a–e. Contours of measurements **a** and estimates at validation points: MLP **b**, wavelet interpolation **c**, WARK **d** and NNRK **e**

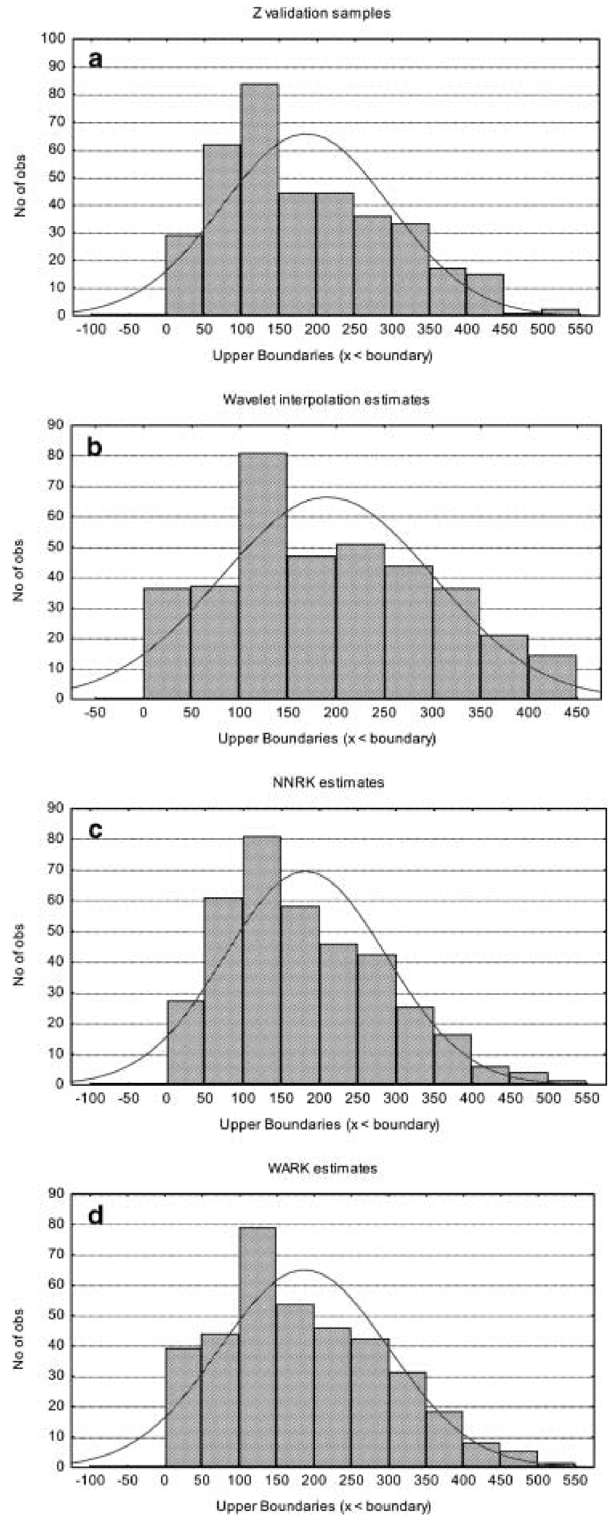

Fig. 7a–d. Histograms of measured and estimated distributions at validation locations: raw data **a**, wavelet interpolation **b**, NNRK **c**, WARK **d**

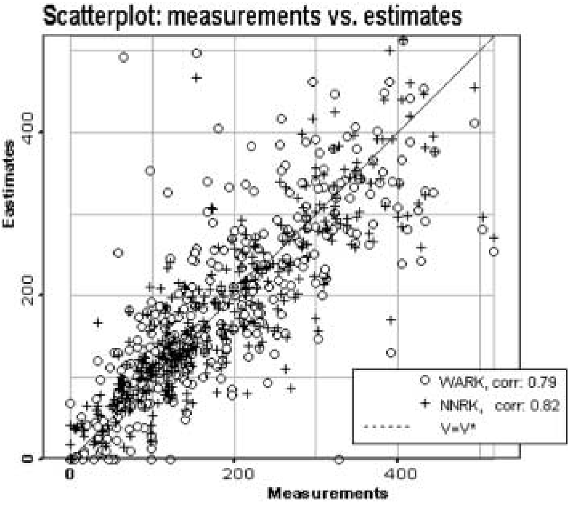

Fig. 8. Scattergrams of measurements vs. WARK and NNRK estimates at validation points

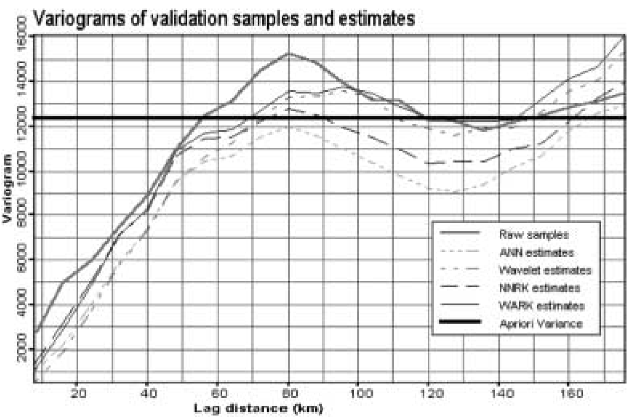

Fig. 9. Directional variograms of validation data: raw (thick solid) vs. ANN estimates, wavelet estimates, NNRK estimates, WARK estimates

Histograms of the measured and estimated values are given in Fig. 7. Histograms clearly show that adding residual kriging estimates to the large scale trend estimation with wavelet interpolation improves the shape of the estimated distribution in the upper and lower tails area.

Measured and estimated values at validation points for the both models are plotted on a scatter graph (see Fig. 8). It clearly shows that NNRK estimates are less scattered and demonstrate higher correlation with the sample values than WARK estimates.

Variograms for raw data and estimates at validation points are computed and presented in Fig. 9. Raw variogram illustrates variability of the spatial distribution. The estimates obtained by all the models give smoother patterns, which can be illustrated but variograms of estimates levelling off at the lower level than the raw one (see Fig. 9). Variogram for NNRK estimates has the most similar shape to the raw variogram, although it levels off at a lower level (about the apriori variance). Variogram for WARK estimates is almost identical to the one for NNRK at the small scale (0–50 km), whereas at larger scales it demonstrates higher variability than the one for NNRK, with values close to the raw variogram.

However, the range of variogram for WARK estimates is slightly longer (100 km) than the range for the raw variogram (80 km).

Residuals (estimated-measured) at validation points can help to assess the quality of the applied model by means of unbiasedness, absence of spatial correlation and help to compare WARK and NNRK. Summary statistics of validation residuals is given in Table 2. WARK residuals have lower absolute mean, whereas NNRK residuals have lower range and inter-quartile range. Both distributions of NNRK and WARK residuals are centred around 0 (NNRK residuals are slightly more symmetric in terms of median and skewness coefficient, see Table 2).

Geostatistical predictors, as mentioned above, provide estimation error along with the estimate itself. Error distribution for WARK and NNRK estimates are visualised as normal probability plot (see Fig. 10). It is clear that WARK estimation error is closer to normal distribution than the one of NNRK.

Table 2. Summary statistics for validation residuals

Statistics	NNRK	WARK
Mean	3.91	−1.26
Median	0	0.67
Minimum	−313.24	−426.83
Maximum	245.00	328
Lower quartile	−25.53	−41.19
Upper quartile	31.30	37.21
SD	56.28	74.18
Skewness	0.154	−0.497
Kurtosis	4.37	5.67

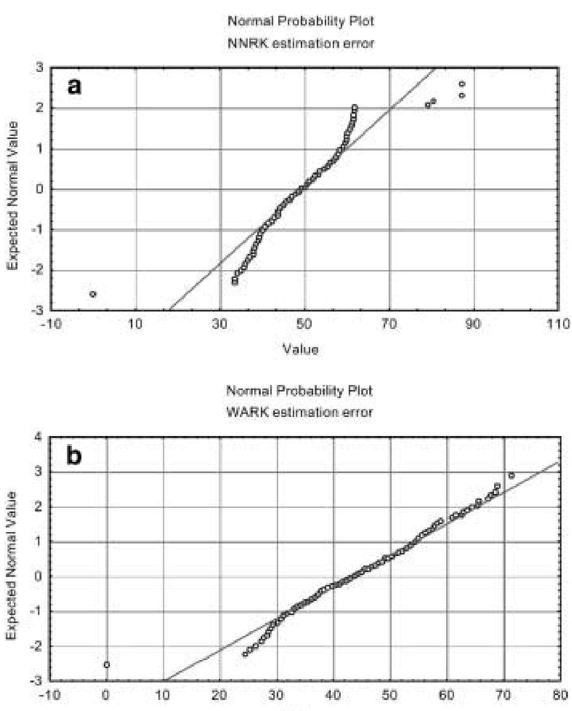

Fig. 10a, b. Normal probability plots for NNRK **a** and WARK **b** estimation error

Table 3. Summary statistics for estimation error at validation points

Statistics	NNRK error	WARK error
Mean	49.47	43.38
Median	49.42	43.36
Maximum	87.27	71.34
Lower quartile	43.23	35.92
Upper quartile	56.08	51.0
SD	9.91	10.80
Skewness	−0.23	−0.44
Kurtosis	5.88	1.57

Summary statistics of estimation errors is given in Table 3. Estimation errors of WARK have lower mean, maximum and quartiles, which is an advantage over NNRK.

6 Conclusions

A new method – wavelet analysis residual kriging (WARK) – has been developed for spatial prediction of highly variable spatial data with multiscale spatial structure. It is a combination of wavelet interpolation and kriging of the residuals. Wavelet interpolation is used to determine large-scale trend and geostatistical tools (variography and kriging) are used to model the remaining residuals, which feature small-scale correlation structure. This method solves the problem of non-stationarity that occurs often in the spatial analysis. Stationary behaviour of residuals allowed to apply kriging model which significantly improved wavelet interpolation results. The final results reproduce well the statistical characteristics of the original distribution (mean, median, maximum, and minimum), along with the reproduction of anisotropic and periodic spatial correlation structure (variogram).

WARK is compared to another methods for prediction in non-stationary case – neural network residual kriging (NNRK). Their results are equivalent in term of reproduction of statistical parameters of the validation distribution and spatial pattern. WARK however increases dramatically the performance of anisotropic spatial correlation structure reproduction, it has proved to be a very powerful method for estimating highly variable data, featuring spatial correlation on different scales.

References

Bishop CM (1995) Neural Networks for pattern recognition. Oxford University Press, New York

Breiman L (1996) Heuristics of instability and stabilization in model selection. The annals of Statistics, 24(6): 2350–2383

Candès EJ (1995) Harmonic Analysis of Neural Networks. Technical report, Department of Statistics, University of Stanford

Daubechies I (1992) Ten lectures on Wavelets, CBMS-NSF Regional Conference on Applied Mathmatics, No 61, SIAM

Deutsch C, Journel AG (1992) GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press, New York 340 p

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap, volume 57 of Monographs in Statistics and Applied Probability, Chapman & Hall

Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. Neural Computation, 7(2): 219–269

Goovaerts P (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, New York
Härdle W (1990) Applied nonparametric regression, volume 19 of Econometric Society Monographs, Cambridge University Press
Isaaks EH, Shrivastava RM (1989) An Introduction to Applied Geostatistics. Oxford University press, Oxford
Kanevsky M, Arutyunyan R, Bolshov L, Demyanov V, Maignan M (1996) Artificial neural networks and spatial estimations of Chernobyl fallout. Geoinformatics, 7(1–2): 5–11
Kanevski M, Demyanov V, Maignan M (1997a) Mapping of Soil Contamination by Using Artificial Neural Networks and Multivariate Geostatistics. Artificial Neural Networks ICANN '97. 7th International Conference Proceedings. In: Gerstner W, Germond A, Hasler M, Nicould J-D (eds.). Lecture Notes in Computer Science, Springer, 1997, pp. 1125–1130

Kanevski M, Demyanov V, Maignan M (1997b) Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks. IAMG'97 Proceedings of The Third Annual Conference of the International Association of Mathematical Geology. In: Pawlowsky Glan V (ed.) Barcelona, Spain, CIMNE, vol. 2, pp. 533–538
Kanevsky M, Arutyunyan R, Bolshov L, Chernov S, Demyanov V, Linge I, Koptelova N, Savelieva E, Haas T, Maignan M (1997c) Chernobyl Fallouts: Review of Advanced Spatial Data Analysis, geoENV I Geostatistics for Environmental Applications. In: Soares A, Gomez-Hernandes J, Froidvaux R (eds.) Kluwer Academic Publishers, pp. 389–400
Kanevski M, Demyanov V, Chernov S, Savelieva E, Timonin V (1998) Neural Network Residual Kriging Application For Climatic Data. The Journal of Geographic Information and Decision Analysis (GIDA), vol. 2, No. 2, ISSN 1480–8943. (http://www.geog.uwo.ca/gimda/journal/journal.htm)
Vapnik V (1995) The nature of Statistical Learning theory, Springer-Verlag
Zhang Q (1995) Wavelet in Regression Analysis. In: Anestis Antoniadis George, Oppenheim (eds.) Wavelets and Statistics, vol. 103 of Lectures in Statistics, pp. 397–407, Springer-Verlag