Mkrtchyan, A., 2004: Spatial interpolation of field data on plant abundance. In: Commarmot, B.; Hamor F. D. (eds) Proc. Internat. conf. 13-17 October 2003. –Birmensdorf, Swiss: Federal Research Institute WSL; Rakhiv: Carpathian Biosphere Reserve. –P. 314-321.

Spatial interpolation of field data on plant abundance

Alex Mkrtchyan
L'viv National Ivan Franko University, Doroshenko st. 41, 79000 L'viv, Ukraine.
alemkrt@yahoo.com

Abstract

In order to effectively organize conservation activities it is important to delineate as accurately as possible the actual and potential habitats of rare and endangered plant and animal species. As primary data are mostly collected on plots in the field, they should be interpolated spatially. The present research considers the approaches and methods for such interpolation and is based on abundance data of a plant species in a nemoral forest near L'viv (Western Ukraine). Two methods of interpolation were tested, the first based on geostatistical interpolation and the second on a multiple regression model using land morphology attributes as indicators. The reliability of each method was assessed by test samples. The relative error weighting method was then used to combine the outputs of both primary methods, which allowed to obtain a more accurate final output. When different interpolation methods produce results of comparable accuracy, the proposed method can combine their results to significantly increase the final accuracy.

Keywords: species abundance, habitat, data integration, geostatistical interpolation, multiple regression, ecological factors, error weighting

1 Introduction

In the field, primary data on the presence of plant and animal species and the characteristics of communities, including data on forest structure, are mostly collected in discrete locations. Therefore, such data should be interpolated spatially to obtain a continuous coverage and gain knowledge on the spatial distribution of species and communities characteristics. The range of applications of spatial ecological data is very extensive and includes the delineation of the protection areas, forest inventories, planning of the different conservation activities. In particular, precise mapping of habitats of rare and endangered species could help to more effectively organize their protection, for instance by regulating activities that potentially threaten these species and/or their habitats.

The task of spatial interpolation of point data is very common in ecology, but it is surprisingly that only during the last two decades it became a common research topic. Among the most significant recent contributions containing an overview of the subject there should be mentioned an extensive review paper of Swiss researchers concerning predictive habitat modeling (Guisan and Zimmermann 2000) and the comprehensive study of Scott et. al. (2002).

While scores of methods and techniques were developed for the task, not every of them provide the means to directly assess accuracy and reliability of the interpolation results. Still, this ability is crucial when the obtained information directs the decision making process. When the information provided by a certain interpolation method is not accompanied by the error measure or this measure is not reliable, it is difficult to compare the efficiency of the different interpolation methods or assess the risk entailed by decisions based on such information. In fact, there is a trade-off between the ability to get probably the less accurate information, but supplemented with a measure of its accuracy (as with the parametric predictive functions), and the ability to get presumably more accurate predictions, but without such a measure (using non-parametric functions).

While the spatial distribution of the species is influenced by a number of various factors and processes (ecological and physiological, local and spatial, natural and anthropogenic, necessary and accidental), any single interpolation method tends to predominantly catch the effect of some factors while disregarding the

others. In fact, any feasible method produces some information about the true spatial distribution of species, however usually incomplete and able to be supplemented by the information derived by the other methods. Thereby the techniques which allow to combine the results of the different interpolation methods, thus utilizing their benefits by taking into account different factors and processes, can be very valuable.

2 Material and methods

The data for the study were collected on the study area in the form of 1800m * 750m rectangle, exclusive of its northeast part with cultural vegetation (Fig. 1), situated in the semi-natural beech forest near L'viv (W. Ukraine). The studied species was a perennial rhizomatous plant *Anemone nemorosa L.*, common in beech forests of the region. As the living cycle of this species almost completely falls on early- and midspring, before the full development of the tree canopy, its distribution is fairly independent of the distribution of the trees. This allows to ignore the factor of the tree canopy distribution, which is difficult to map, but which governs the spatial patterns of the majority of herbaceous species present during summer. In April 2001, data were collected on 46 sample sites selected with a random-stratified scheme, evenly representing main terrain elements. Abundance and vitality of plant species could generally be indicated by their total ground cover. This has been estimated using an ordinal scale proposed by Mirkin and Rosenberg (1978). This scale assigns the grade according to the visual estimate of the ground cover in %, from the grade 0 when the species is totally absent on the site to the grade 5 when its ground cover exceeds 50%. This scale has two advantages over the simple percentage estimates. Firstly, it is more adequate to the human perception of ground cover (for instance, we more easily distinguish between 20 and 40% than between 60 and 80%), thus allowing more reliable visual estimates. Secondly, it often leads to a close to normal distribution of values, which is a basic requirement of some common statistical approaches, while the distribution of the ground cover percentage is often significantly skewed towards the low values.

After the values of the ground cover grade were assessed for all 46 sample sites, 10 randomly selected values were set aside and excluded from the analysis to be used as an independent validation (test) data set to evaluate the predictions (Fig. 1).

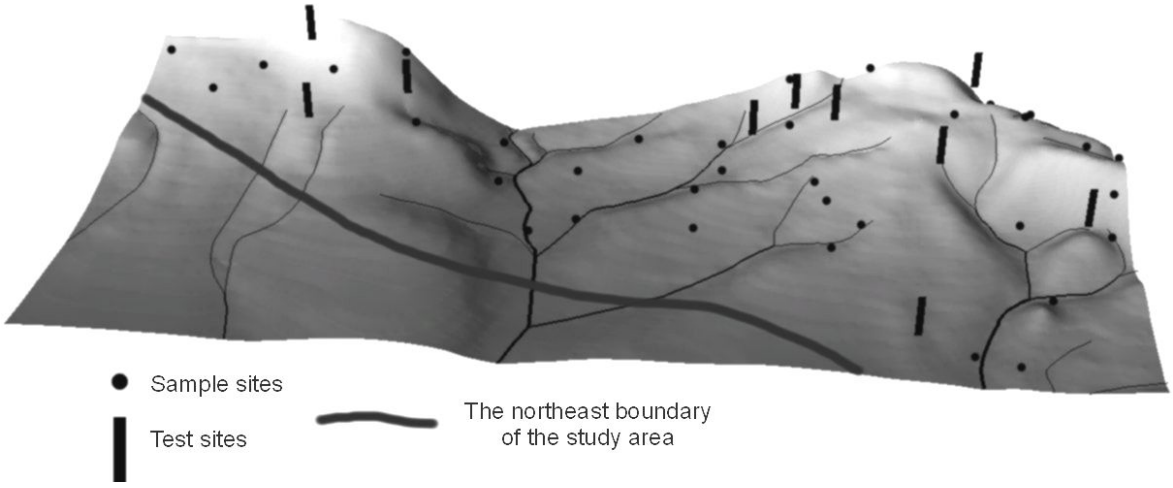

Figure 1. The terrain of the study area and the distribution of the sample and test points (view from the south).

The first interpolation method applied to the data was (ordinary and universal) kriging. The decision to use this technique entails the choices (1) of the number of the nearest points to be used in the calculation of the sample semi-variogram, and (2) of the mathematical function to be used to model the semi-variance. These choices can significantly affect the outcome. In our case, the best parameters were selected by the trial-and-error method, comparing the predictions with the values of the independent data set. The universal linear kriging with the window of 6 nearest points turned out to be the best choice. Here and later the predicted real values are rounded to an integer scale for the purpose of presentation. An older inverse distance weighting (IDW) method was also tested but gave unsatisfactory results.

The major advantage of kriging is that it can produce the spatial representation of the supposed error of interpolation (the minimized estimation kriging variance), which supplements the map of the interpolated values and allows judging their reliability at any point of the area.

The second approach employed for the task was predictive modeling of the species habitat. It predicts the distribution of species by defining the relationships between their presence/abundance and some direct or indirect ecological gradients – factors influencing the species distribution. These relationships are assessed on the sample points and expanded throughout the area. The most accurate available maps of the ecological factors are usually derived from the digital elevation model (DEM) while the common geological, soil and climatic maps often lack precision and reliability (Guisan and Zimmermann 2000).

In the present work, the DEM for the study area was created by digitizing several layers of the detailed topographic map and interpolating the elevation data using ANUDEM – the interpolation method specifically designed for the creation of the hydrologically correct DEMs (Hutchinson and Dowling 1991). To predict the distribution of the species, two topologically related indirect ecological factors were used. The slope values were derived from the elevation data by the simple algorithm; they show a significant relationship with the plant ground cover grades because they are themselves strongly associated with the soil depth, soil moisture content and some other powerful ecological factors. The second factor used to predict the ground cover grades was the illumination index calculated as a direct solar radiation incidence on inclined surfaces on April 1st at noon relatively to the horizontal plane. The Student's t was used for calculating the significance of the influence of these two factor, showing the values t = 3.8 for the slope and t = 3.45 for the illumination index. Some other presumed topologically related gradients (like the surface curvature) were also tested but proved not to be significant. The layers of the two selected factors were input into a linear least-squares multiple regression model. As the scale used to assign the ground cover grades can be relegated as ordinal and unbounded, this model could not be fully justified statistically; still, as it is rather simple, common and widely used approach to model different sorts of relations, it is applied here to illustrate the potential of the proposed output combination method. Despite its strong requirements for the quality of the input data, this method is quite robust and can be safely applied in many cases if an independent data set is available to validate the results. The model produces a map of interpolated values supplemented with a single value of RMS error.

While both applied methods have produced different outputs of comparable average accuracy (see Results), a way could be explored to integrate the information contained in both outputs to obtain a single final result, more precise and reliable than any of the maps produced by the above-described techniques. At the heart of the proposed data integration approach is the idea that the amount of information contained in any map for a single point is inversely proportional to the supposed error (more precisely, its variance) at that point.

When GIS is used to estimate some value (e.g., the land suitability), the common situation is when the value to be estimated is influenced by several factors represented in a GIS database by continuous surfaces. In such cases, the common procedure to combine the information from several criteria is the weighted linear combination (Voogd 1983). It implies combining factors by applying a weight to each followed by a summation of the results. In the case of two factors, represented by two surfaces, the expression is:

$$C = XA + YB \tag{1}$$

where A and B are the values at some point of either of the surfaces, X and Y are the weights of the surfaces, which sum to 1, and C is the value of the resultant surface at that point. The similar approach can be used to combine the different sources of information on the plant abundance. In this way, the errors of these sources could partially be cancelled out. The extent of this canceling out will depend on the weights assigned to each primary surface. To select the best system of weights, let's first take the linear approximation of the error of the function of two variables, given by:

$$\tau^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\left\{\rho_{ij}\sigma_i\sigma_j g'_i g'_j\right\} \tag{2}$$

where σ_i, σ_j are the errors of the primary surfaces, ρ_{ij} is the correlation of the errors, g'_i, g'_j are partial derivatives (Heuvelink 1998). As in our case the function is a simple addition of the two surfaces multiplied by their weights, an expression for the squared error of (1) can be given as:

$$\tau^2 = \sigma_A{}^2 X^2 + \sigma_B{}^2 Y^2 + 2\sigma_A\sigma_B XYr \tag{3}$$

where σ_A, σ_B are the errors of the input surfaces at some point, r is the correlation of these errors. Then by taking $X + Y = 1$ and minimizing τ^2 we get the formula for the best weights of (1):

$$X = \frac{\sigma_B^{\,2} - r\sigma_A\sigma_B}{\sigma_A^{\,2} + \sigma_B^{\,2} - 2r\sigma_A\sigma_B} \tag{4}$$

and similarly for *Y*. In the case of the uncorrelated errors of the input layers it reduces to:

$$X = \frac{\sigma_B^{\,2}}{\sigma_A^{\,2} + \sigma_B^{\,2}}. \tag{5}$$

The error of the outcome in the latter case is calculated by:

$$\tau = \sqrt{\sigma_A^{\,2} X^2 + \sigma_B^{\,2} Y^2}\,. \tag{6}$$

The weights of the input layers calculated with (4) or (5) are supplemented in (1) to calculate the values of the outcome surface (map). As some of the interpolation methods (kriging among them) allow to derive the spatial distribution of the prediction errors, the weights for the input surfaces produced by such methods will differ from point to point. Thus, the surface obtained by kriging is the main contributor to the final outcome near the sample points, whereas far from them where the error of kriging grows bigger than that of the multiple regression interpolation (which is uniform throughout the area), the latter becomes the main contributor. Eventually the rule of thumb here is "the more reliable the primary surface is at some location of the study area, the more weight is assigned to it at that location and the more influence it has on the final output". And vice versa, more error means less weight and less influence on the final output. When the errors of both surfaces are close to equal, averaging the surfaces cancels out their errors to some degree, adding statistical reliability to the output. The proposed method of the information combination could be termed the relative error weighting (REW).

3 Results

Figures below show the results of the two primary interpolation methods as well as the presumed error distribution for kriging. The result of the multiple regression interpolation (Fig. 4) shows high values of plant ground cover for the shallow slopes with deep soils, and low values for the steep slopes facing north, which are often shaded and have stony soils. The result of the kriging interpolation (Fig. 2) is reliable only in the vicinity of the sample points and thus fails to show any significant general pattern.

For both of the applied interpolation methods an independent measure of the accuracy of prediction was acquired by comparing values calculated for validation data points withdrawn from the analysis with the actual values; it corresponded well with the data on the calculated error map for kriging (Fig. 3) and the calculated error of the regression model. The squared differences between predicted and actual values for the 10 validation points averaged 0.94 grades for the kriging and 0.97 grades for the multiple regression estimate. While the error of the latter is presumed to be uniform throughout the area, the error of kriging interpolation is minimal in the vicinity of the sample points and quickly increases with the distance from them (Fig. 3). Beyond the range of the kriging semi-variogram (~400m) interpolation results become meaningless.

While the average errors of both methods calculated theoretically and on the independent validation points were roughly equal, these methods have actually produced very different outputs, as can be seen in Figs. 2 and 4. The application of the proposed REW method allowed to combine the outputs, significantly increasing the accuracy of the prediction. As the primary methods are conceptually different, their errors may be considered as uncorrelated, and formulas (5) and (6) may be used for the purpose of their integration. Figs. 5 and 6 show the output maps of the interpolated ground cover grades and their variance, calculated by these formulas, and Table 1 shows some measures of resulting accuracies of both primary methods and the REW, calculated across the test sites.

Table 1. Mean value, root mean square (RMS) error and bias of the ground cover grades for the 10 test points, calculated with different interpolation methods.

Measure \ Method	Kriging	Regression	REW of the primary outputs
Mean	2.61	2.52	2.55
RMS error	0.97	0.99	0.87
Bias (mean error)	0.25	0.17	0.19

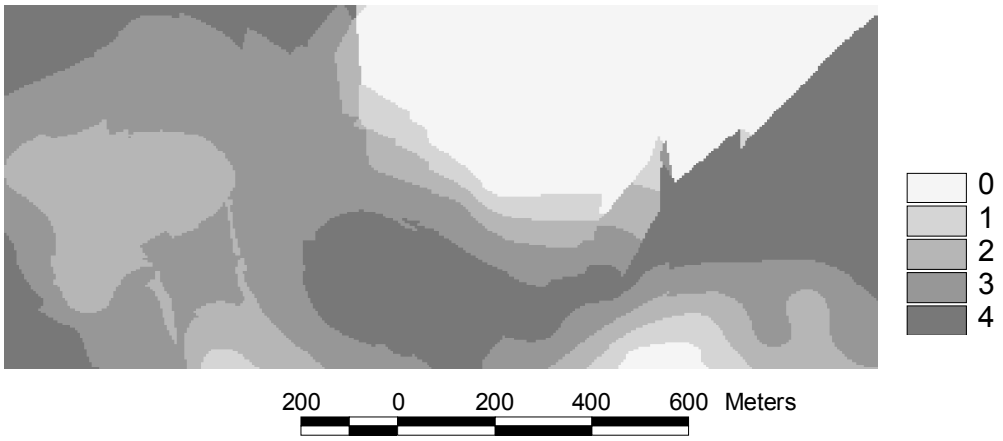

Figure 2. An interpolated map of the ground cover grades obtained by the universal linear kriging with a window of 6 points.

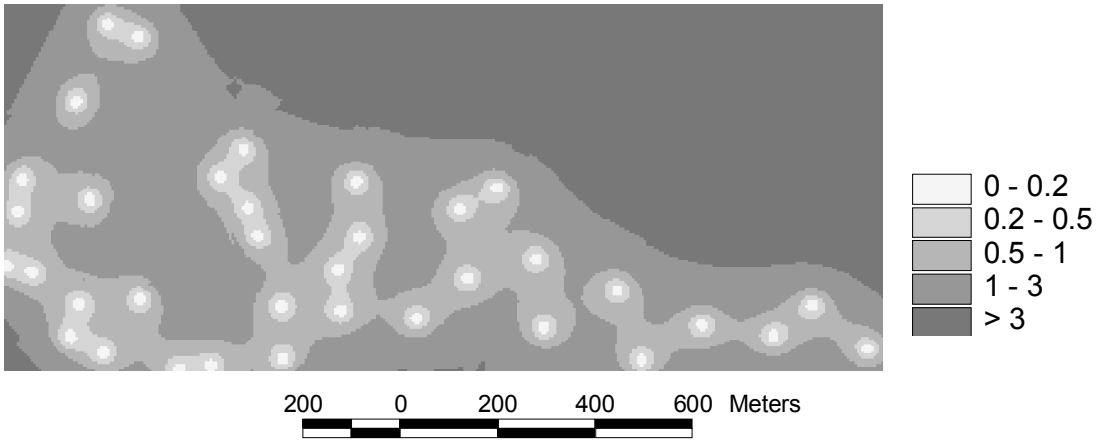

Figure 3. Estimate of the variance of kriging predictions (Fig. 2).

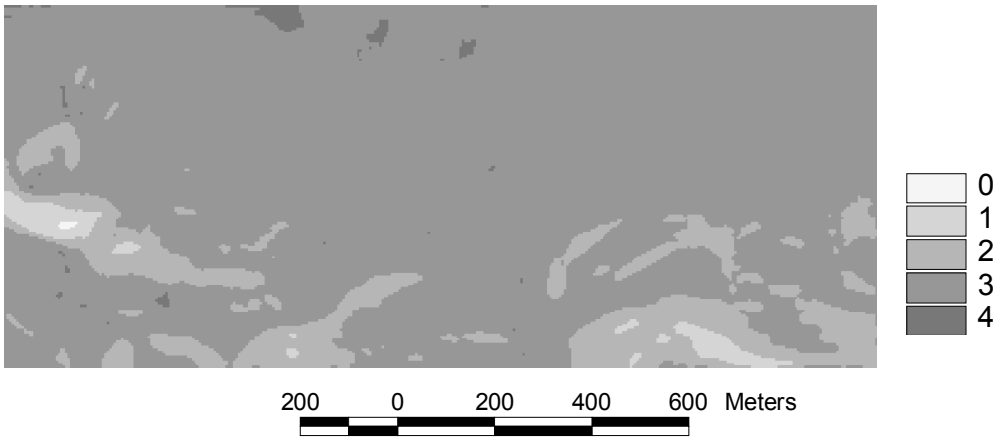

Figure 4. Prediction of the ground cover grades based on slope and illumination index, using multipe linear regression.

As can be noticed, kriging produced slightly smaller RMS error, while the predictions from the regression model have smaller bias. The latter model gave generally more uniform distribution of values (Fig. 4) than kriging (Fig. 2) and has better caught larger-scale pattern controlled by the abiotic factors, while kriging better reproduced small local variations.

The application of the proposed data integration method allowed to reduce noticeably the error of the output (Figs. 4 and 5), producing RMS error values lower than that for either of the primary surfaces. The bias of the REW output was close to that of the primary output with the lowest bias.

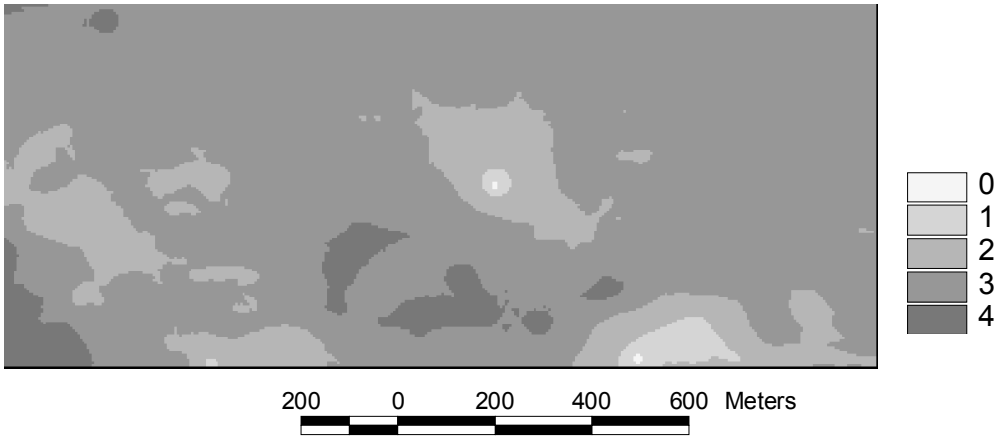

Figure 5. Prediction of the ground cover grades obtained by combining kriging and multipe linear regression with the proposed REW method.

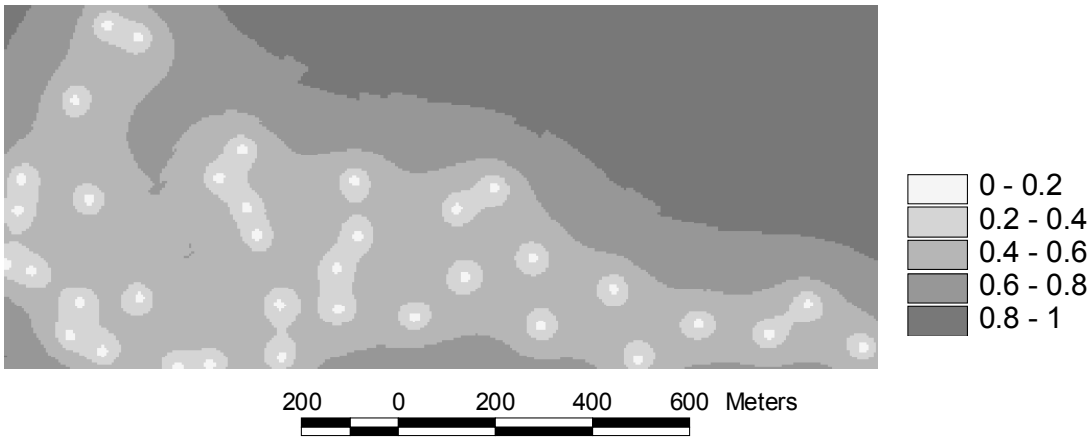

Figure 6. Estimated variance of the prediction of the ground cover grades obtained by using the REW method to combine the results of kriging and multipe linear regression (Fig. 5).

4 Discussion

An extensive and constantly increasing arsenal of methods and techniques exists for solving the problem of point data interpolation. Recent explosion in the development of the interpolation methods has been mainly due to the rapidly increased capabilities of modern computers to quickly process large amounts of data. The different methods usually produce different interpolation results depending on the quality of data and the character of spatial variation of the modeled phenomenon, and there is no single method or technique that always gives the best results regardless of the attendant circumstances. So there is the problem of choice of

the certain interpolation technique to be used in any particular case.

The methods for point data interpolation can roughly be put into the two main groups. The first base their predictions on direct spatial relations between the values in different locations, assuming that adjacent locations are more similar than more remote ones. This can be explained by species propagation and reproduction (e. g., cloning) or by the gradual character of the changes in the ecological gradients. Methods of the second group are based on the direct specification of the relations between species presence/abundance or communities' characteristics, and ecological factors, i.e. on the parameterization of their ecological niches. These methods generally ignore spatial correlations between locations, however they can be used to interpolate point data and map the vegetation characteristics if maps of the relevant ecological factors exist.

Among the first group of methods, the most used presently are methods based on geostatistics (kriging and cokriging), which succeeded extensively used earlier, computationally less demanding but at the same time less precise methods like IDW and splines. Kriging and its multivariate extension, cokriging, are based on the regionalized variable theory that assumes that the spatial variation in the phenomenon is statistically homogeneous throughout the area (Cresie 1993).

The second group represents an extensive family of models; among the most frequently used and recommended are regression and classification trees, proportional odds regression models for ordinal data, general linear models (GLM or GLIM) and their nonparametric extension, general additive models (GAM) (Guisan and Zimmermann 2000). An advantage of a GLM over GAM is that GLM provides an explicit formula for the model function, while GAM employs a non-parametric smoothing. On the other hand, GAM provides more flexibility compared with GLM (Hastie and Tibshirani 1990; Yee and Mitchell 1991). Most of these models require special software and a significant amount of expertise on the part of the user.

In any particularly case several conceptually different methods may by equally valid, but revealing different aspects of the spatial pattern of species distribution. For this reason, some hybrid approaches were developed, for example the aforementioned cokriging and regression models which include a spatial autocorrelation term (Miller and Franklin 2002; Leathwick 1998). It would be valuable to provide a method to integrate the outputs of the different primary methods into a single result, which would inherit the merits of the primary methods. Bayesian Probability theory and its generalization, Dempster-Shafer theory, are mostly used in various fields to integrate the evidence from the different sources (Shafer 1976). These approaches are most suited for the integration of evidence in the form of conditional probabilities, and less suited for the spatially distributed quantitative data supplemented with an error estimate. Such data can be combined using the proposed REW method, which is based on an easily comprehensible logical chain: "The smaller the supposed error – the higher the reliability – the higher the relative weight – the higher the influence on the output, and vice versa". When combining two primary predictions, the output from REW for the single location can theoretically be up to 30% more precise then the most accurate of the input layers, in the case the errors of the inputs are equal and totally uncorrelated, as could be seen from the formula (6). However, since the spatial distribution of the errors of different methods can differ significantly, the gain in accuracy for the total area can be quite large. At the same time, if the input errors are strongly correlated (e. g., inputs are obtained by conceptually similar methods) or if they differ greatly between the input surfaces (one input is significantly better than another) the gain in accuracy will be only marginal.

Another advantage of the proposed approach is its openness to explication and interpretation. Given the inputs and the output, it is easy to explain and understand why the output looks the way it does. An explanation may sound as follows: "Here the value is low because the slope is steep or facing north" or "Here the value is high because a short way off is a sample point that has shown a high value". This way, the sources of disagreements between the predicted and the actual values can be easily identified and, possibly, corrected.

5 Conclusions

The rapid development of computer technology and the concomitant progress in quantitative methods in ecology raise new challenges. One of them is the task of the cooperative interplay between the "qualitative" ecological knowledge of practitioners and the modern computational capabilities. Another challenge lies in the ability to obtain reliable measures of the accuracy and error, indispensable for the effective decision-making. One major and common task in ecological research is the spatial interpolation of the point data. While a number of methods and techniques exist to accomplish this task, the majority of them effectively reflect the effects of only some factors and processes while ignoring or underestimating others. The proposed method allows to combine and integrate spatial data obtained by different primary methods, producing more precise final output. It is based on the intuitive, easily comprehensible idea, which allows to explain the

results, detect the sources of errors, and assess the reliability of data, utilizing field ecological knowledge. While experimental material was small and was used to illustrate the ideas rather than to verify them, the field deserves further investigation.

6 References

Brown, J.H.; Mehlman, D.W.; Stevens, G.C., 1995: Spatial variation in abundance. Ecology 76: 2028–2043.

Cressie, N. A. C., 1993: Statistics for spatial data (revised edn). John Wiley & Sons Inc, New York.

Guisan A.; Zimmermann N. E., 2000: Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186.

Hastie, T. J.; Tibshirani, R. J., 1990: Generalized additive models. Chapman and Hall, London, England.

Heuvelink, G. B. M., 1998: Error Propagation in Environmental Modelling. Taylor & Francis.

Hutchinson, M. F.; Dowling, T. I., 1991: A continental hydrological assessment of a new grid-based digital elevation model of Australia. Hydrological Processes 5: 45-58.

Mirkin, B. M.; Rozenberg, G. S., 1978: Phytocenology: principles and methods. Nauka, Moskow (In Russian).

Scott, J.M.; Heglund, P. J.; Morrison, M. L.; Haufler, J. B.; Raphael, M. G.; Wall, W. A.; Samson, F. B., 2002: Predicting species occurrences: Issues of accuracy and scale. Island Press, Washington, DC.

Shafer, G., 1976: A mathematical theory of evidence. Princeton university press, Princeton.

Voogd, H., 1983: Multicriteria evaluation for urban and regional planning. Pion, Ltd, London.

Yee, T. W.; Mitchell, N. D., 1991: Generalized additive models in plant ecology. Journal of Vegetation Science 2: 587-602.