ANALYSIS OF THE RELATIONSHIPS AMONG ERROR VALUES AND VALUES OF MORPHOMETRIC PARAMETERS DERIVED FROM THE DEM

Jana, SVOBODOVÁ[1], Lukáš, MAREK[1], Pavel, TUČEK[1]

[1]Department of Geoinformatics, Faculty of Science, Palacky University, tř. Svobody 26, 771 46, Olomouc, Czech Republic

j.svobodova@upol.cz, lukas.marek@upol.cz, pavel.tucek@upol.cz

Abstract

A fundamental tool for the exploration of the relations (dependencies) among two or more variables is a correlation analysis. Main goal of the correlation is to analyse an existence or absence of the relation among chosen variables and also quantification of the strength of this (in)dependence. A dependent variable as well as an independent variable is not determined during the correlation analysis. This determination of dependency is expressed subsequently during the regression analysis, when the proved relations are mathematically expressed.

The research is aimed at the exploration of the relations among values of errors and values of morphometric characteristics or their changes (derived from digital elevation models). Firstly, the correlation analysis is used. Then, correlated variables entered the regression analysis. Correlation analysis as well as regression analysis has been applied on values (or their changes) of morphometric parameters which were derived from so-called "high-quality" and "low-quality" DEMs of modelled areas. The aim is to find out the variability of relations using the most different DEMs according to the non-spatial evaluation of the metric accuracy. Different terrain configurations were taken into account: three rugged or flat highlands, uplands, hilly areas and three flat lands. Using the multiple areas with the same type of relief allows us to observe a presence of the trend in the relationships among values of errors and values of morphometric characteristics. The paper describes not only the calculation of correlation and regression analyses but also data pre-processing.

Keywords: DEM, morphometric parameters, correlation, linear regression, GWR

INTRODUCTION

Correlation and regression analyses serve for a research of dependencies between variables. The logical procedure of statistical investigation is as follows. Firstly, the correlation analysis is carried out in order to find out whether there are any dependencies between variables. If the dependency among variables has been proven, it is possible to express it mathematically by the regression analysis.

Within the research, a correlation has been used to investigate existence and closeness of relationships between error values and morphometric parameter values. The variables where the relationship has been proven then entered into the regression analysis. The correlation and regression analysis were applied to the error values and the values of morphometric parameters derived from the so-called low-quality of the DEM for each sample area. The aim was to determine the trend of changes in values of morphometric parameters using low-quality DEM evaluated by global metric accuracy methods.

STATE OF ART IN RESEARCH OF STUDIED RELATIONSHIPS

At present, very few authors deal with the research of the relationship between error values and values of morphometric parameters using e.g. correlation and regression analysis. As the most important works in this area research by the authors Erdogan (2009, 2010) and Carlisle (2002, 2005) could be mentioned.

Erdogan (2009) used 4 basic morphometric parameters for regression analysis of the dependence of errors: slope, aspect, total curvature and elevation. Evans (1972) replaced the total curvature by planar and profile curvature and together with the other parameters he considered them as adequate for the exhaustive quantification of surface shapes. Carlisle (2002, 2005) is not limited in his work to these basic morphometric parameters, but he complements them by the relative relief, texture, mean extremity, minimum extremity, maximum extremity and vertex distance. For most parameters he tested also the second and third derivatives of these variables, in order to investigate non-linear relationships between error values in the DEM and values of morphometric parameters.

To assess the basic relationship between the error values in the DEM and values of morphometric parameters the calculation of the correlation coefficients were separately performend for each parameter. The above-mentioned authors (e.g. Carlisle 2005, Erdogan 2010), however, consistently state that it is inappropriate to study the strength of the relationships of individual morphometric parameters and errors separately. On the other hand, the scales of the error values cannot be adequately explained only by one parameter alone, it is necessary to use more independent (explanatory) parameters simultaneously. This can be achieved using regression methods.

In the frame of the study, only six basic values of morphometric parameters (slope, aspect, total, profile or planar curvature and elevation) were used to build the regression model. The aim was to find out to what extent are these parameters, which are according to Evans (1972) sufficient for an exhaustive description of surface shapes, able to predict error values. The use of further explanatory variables (e.g. according to Carlisle 2005) was tested on one selected area. Their inclusion in the regression analysis, however, has brought hardly any improvement of quality of a model if the coefficient of determination is regarded.

METHODS

Model areas and data

The study is focused on the research among the error values (Fig. 1) and the values of morphometric parameters. The research was implemented through the three sample areas (flat highlands): Studenská, Uhřická and Divácká highlands. The aim of this study was to describe the relationships among the error values and the values of morphometric parameters derived from the low-quality DEMs. It is possible to suppose a greater degree of dependence of monitored variables than in the high-quality DEMs. The investigation of the relationships among these variables derived from the high-quality DEM (even for other types of relief) is subject for further research of the authors.

A more detailed review of methods of DEM quality assessment can be found in Svobodová (2011). Svobodova also stated that for the original data with high density of points only non-spatial methods of DEM evaluation together with a visual inspection of DEM can be suitable. The mentioned work provides the summary of interpolation methods and their settings for the creation of quality DEMs. Because a larger set of different DEMs, generated by different interpolation methods, was selected in Svobodova (2011), the work contains also an overview of methods and setting which are inappropriate for selected terrain configurations. These inappropriate methods were then used in this study for creation of low-quality DEMs with pixel size of 10 m. List of employed methods and setting of their parameters is shown in Tab. 1.

Tab. 1. Interpolation methods and setting of their parameters for creation of low-quality DEM of model areas and the resulting values of root mean square error (RMSE), total absolute error (AE) and hammock index (H) (source: Svobodová 2011)

Model area	Interpolation method	Power	No. of input points	RMSE	AE	H
Divácká highlands	IDW	0,5	10	2,91	4072,29	0,26
Studenská highlands	IDW	2	20	3,04	2012,62	0,68
Uhřická highlands	IDW	0,5	20	3,71	4686,53	0,24

Contour lines with 5-m equidistance from the data model DMÚ25 were chosen as the input data for DEM creation. The contour lines were firstly converted to elevation points. According to the principle of split-sample validation, 15 % of points had been removed before the interpolation. These points were set aside points of reference because of the DEM evaluation. When the DEMs were created (based on the Tab. 1) the raster layers of morphometric characteristics could have been derived. Particularly – slope, aspect, plan, profile and overall curvature were extracted, but only the values of characteristics in the reference points were used further for the correlation and regression analysis because the reference points were the only places where real error of certain DEM could have been calculated. Beside morphometric characteristics, the elevation in reference points was also used in the analyses.

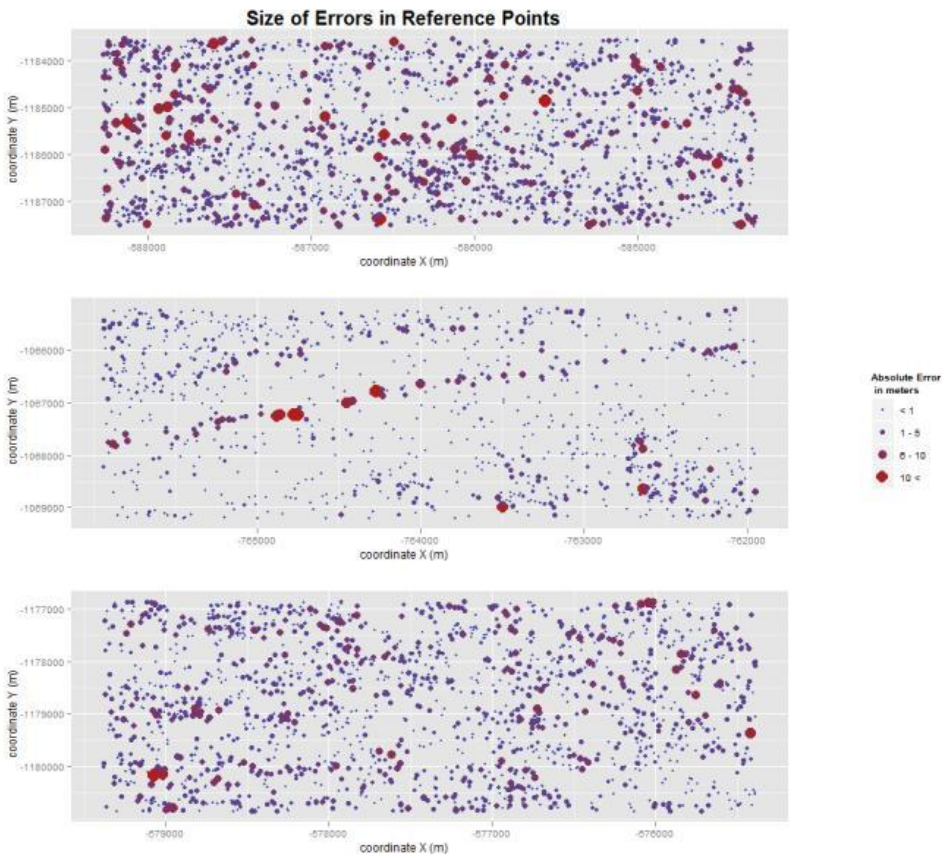

Fig. 1. Spatial distribution of reference points and size of DEM errors. Top position - Divácká highlands (DH), middle position - Studenská highlands (SH), bottom position - Uhřická highlands (UH).

Software

Several softwares were used for analyses of spatial data. ArcMap 10.0 was utilized for creation of DEMs and derivation of morphometric parameters and also for basic operations such as extraction of coordinates from the shapefile, but RStudio – IDE for R, was the main processing software. The R project disposes plenty of packages (libraries) for analyses of geodata. Because we dealt with spatial autocorrelation and geographically weighted regression the most important were packages *spdep* (Spatial dependence: weighting schemes, statistics and models) and *spgwr* (Geographically weighted regression), that provide us the proper analytical capabilities. Alternatively, the software OpenGeoDa that serves as an introduction to spatial data analysis can be used.

Correlation and Spatial Autocorrelation

Initial work by geographers on measuring the strength of a linear relationship between two variables (x and y) relied upon the Pearson product–moment correlation coefficient (r_{xy}), which values lie in the <-1;1> interval (Robinson 2009). A value of -1 represents perfect negative relationship, $r_{xy} = 0$ indicates the absence of any statistical relationship between two variables while $r_{xy}=1$ means perfect positive relationship. If the variables are not normally distributed, than the Pearson product–moment correlation coefficient is replaced with Spearman's rank coefficient (r_s) (Myers a Well 2003).

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(i_x - i_y)^2}{n \cdot (n^2 - 1)} \qquad (1)$$

where expression $(i_x$ - $i_y)$ means differences in ranks of corresponding values of *x* and *y* ; and *n* = numbers of pairs of *x* and *y* values. Pair correlation among more than two variables can be expressed by a correlation matrix. This matrix is squared and its every entry is r_{xy} or r_s. The correlation matrix can be visualized in the form of simple level plot, where intensity of linear relationship is described by the colour on the continuous scale (Sarkar 2008) (Fig. 4a). Second alternative of visualization is a correlation ellipse (Fig. 4b), where correlation is expressed with the angle of ellipse and the size of the minor semiaxis and furthermore with colour (Murdoch and Chow 1996).

Because the study dealt with spatial data, we needed to include the spatial extension of correlation. Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional (2-D) surface, introducing a deviation from the independent observations assumption of classical statistics (Griffith 2009). Spatial autocorrelation exists because real-world phenomena are typified by orderliness, (map) pattern, and systematic concentration, rather than randomness. In other words, spatial autocorrelation means a dependency exists between values of a variable in neighbouring or proximal locations, or a systematic pattern in values of a variable across the locations on a map due to underlying common factors (Griffith 2009). Positive spatial autocorrelation refers to the patterns where nearby or neighbouring values are more alike; while negative spatial autocorrelation refers to the patterns where nearby or neighbouring values are dissimilar (Lu and Thill 2009).

Most often used global analysis of spatial autocorrelation are *Moran's I statistics*, Getis-Ord G statistics and Geary's C statistics (Anselin 1995). We focused mainly on Moran's I statistics (2) - an index of spatial autocorrelation, which involves the computation of cross products of mean adjusted values that are geographic neighbours (i.e., covariations). It ranges from roughly (–1, –0.5) to nearly 0 for negative, and nearly 0 to approximately 1 for positive, spatial autocorrelation, with an expected value of *–1/(n – 1)* for zero spatial autocorrelation, where *n* denotes the number of areal units (Griffith 2009). Lu (2009) defined Moran's I as:

$$I_k = \frac{n * \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(k)} * \left(z_i - \bar{z}\right) * \left(z_j - \bar{z}\right)}{\left(\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2\right) * \left(\sum_{i \neq j}\sum w_{ij}^{(k)}\right)} \qquad (2)$$

where I_k is resulting coefficient, $w_{ij}^{(k)}$ is an indication of distance between areas *i* and *j* for step *k*, z_i is observed characteristic and is a mean of features. The local versions of those global indices are grouped under the term LISA, which means Local Indicators of Spatial Association. Graphical output of local version of the Moran's I is called Moran scatterplot (Fig. 2).

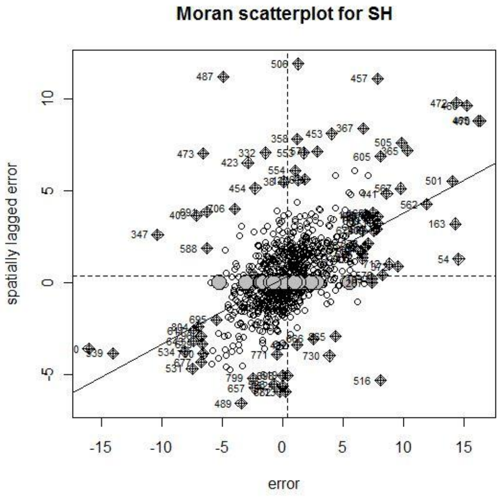

Fig. 2 Moran scatterplot

Multiple Linear regression and Geographically Weighted Regression (GWR)

Regression analysis is a useful tool for analyses where the aim is to examine whether a particular outcome (the dependent variable) is in some way linked to variations in another phenomenon (the independent or explanatory variable). The simplest case deals only with one dependent and one independent variable and it can be thought as extension of correlation. But frequently it is demanded to examine a relationship among one depend variable and more independent (explanatory) variables. This is called multiple or multivariate regression (3). A multiple regression equation has the general form:

$$y = \alpha + \beta_1 x_1 + ... + \beta_n x_n + e \tag{3}$$

In the above equation *y* is dependent variable, *α* is intercept, each value of β_1 to β_n is called the partial regression coefficient, x_1 to x_n are explanatory variables and *e* is residual. Residual is the difference between the value of the dependent variable predicted by the model and the true value of the dependent variable. The method of estimation that we used is ordinary least squares.

There exist several assumptions for the regression model (Pearce 2009):

1. There is a linear relationship between the explanatory variable and the response variable.
2. Residuals are normally distributed.
3. Mean of residuals equals zero.
4. Homoscedasticity of residuals.
5. Autocorrelation: the residuals should be independent of each other.
6. Lack of measurement error.

If the model complies with assumptions mentioned above, then one can discuss its reliability. The most common measure of how well future outcomes are likely to be predicted by the model is called the coefficient of determination R^2. It is the proportion of variability in a data set that is accounted for by the statistical model (Steel and Torrie 1960). The R^2 value provides us with details about the overall model fit and its values range from 0 to 1 (or 0 – 100 %), in the case of perfectly fitted model.

In the case of multivariate model that contains bigger amount of explanatory variables, it is very likely that dropping-off some of these variables would not cause significant loss of the model's precision. The automatic procedure of the choice of explanatory variables, which remain in the model, is called stepwise regression. Usually, this takes the form of a sequence of F-tests, but other techniques are possible, such as t-tests, R^2, Akaike information criterion (*AIC*), Bayesian information criterion (*BIC*), Mallows' *Cp*, or false discovery rate (Hocking 1976, Draper and Smith 1981).

Geodata can be processed using spatial extension of (multiple) linear regression - Geographically Weighted Regression (GWR). While multiple regression model is rather global method, GMR is local regression technique that allows the model parameters to vary across the space. Although the local technique does not allow extrapolation beyond the region in which the model was established, it allows the parameters to vary locally within the study area, and may provide a more appropriate and accurate basis for descriptive and predictive purpose (Propastin et al. 2008).

The local estimation of the parameters with GWR is given by the equation (for two independent variables) (Fortheringham 2002):

$$y = \beta_0(\mu,v) + \beta_1(\mu,v)x_1 + \dots + \beta_n(\mu,v)x_n + e \qquad (4)$$

This regression equation orders the regression parameters to be estimated at a location for which the spatial coordinates are provided by the variables μ and v and parameters can also be estimated at locations where there are no data (Propastin et al. 2008). In GWR, the regression and its parameters in each point of space is quantified separately and independently from other points. The regression model is calibrated on all data that are positioned within the region described around a regression point and the process is repeated for all regression points (Páez and Wheeler 2009). GWR works in the way that each data point is weighted by its distance from the regression point. It means that the closer a data point is to the regression point, the more weight it receives (Propastin et al. 2008). In fact, there exist two main approaches for estimation of regression parameters on individual points. The first approach is application of kernel with a *fixed* bandwidth; the second is then *adaptive* bandwidth (Fig. 3). A fixed bandwidth (e.g. based on Gaussian function) is a suitable choice for modelling if the sample points are reasonably regularly spaced in the study area, otherwise adaptive bandwidth is recommended (Charlton and Fortheringham 2009).

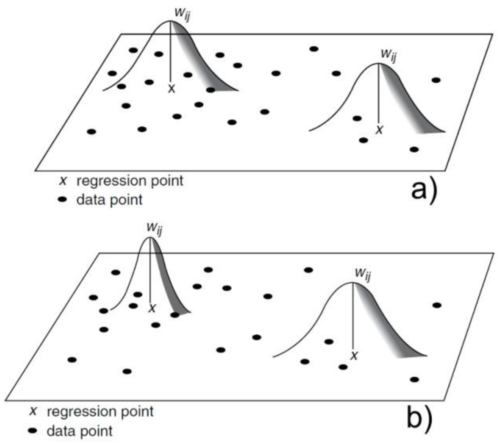

Fig. 3. GWR with fixed kernels (a) and adaptive kernels (b) (Fortheringham et al. 2002)

RESULTS AND DISCUSSION

Correlation and Spatial Autocorrelation

Firstly, correlation coefficients between all pairs of characteristics were computed. Results of correlation are shown in the Fig. 4, where two possible ways of the visualization of correlation are used. One can easily find where the relationship between two characteristics exists. The strongest relations were discovered between all types of terrain curvatures. Strong positive correlation ($r_{s\,=}$ 0.9) was found between overall and planimetric curvature (dark red square in the Fig. 4a and narrow ellipses in the Fig. 4b). Strong negative correlations were found between profile and planimetric curvature ($r_{s\,=}$ 0.7) and also profile and overall curvature ($r_{s\,=}$ 0.9) (dark blue color in Fig. 4a). These tight relations between characteristics assume possible removing of some characteristics from the subsequent linear model. Values of correlation coefficients (Fig. 4) are results for the area of Divácká highlands but results for the other two highlands are very similar.

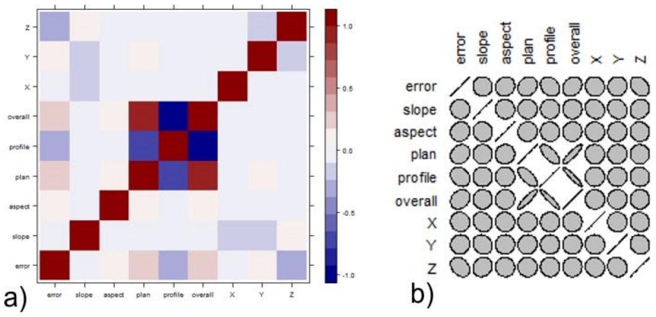

Fig. 4. Level plot of correlations (a) and correlation ellipses (b)

Then the spatial autocorrelation of DEMs errors were investigated. Because measures of the spatial autocorrelation are based on spatial weights in the neighbourhood, earlier than their computation, spatial weights needed to be defined. Points in the range from 0 to 150 m were chosen as the neighbourhood (Fig. 5a). Moran's I was then computed using not only the spatial weights but also the Monte Carlo simulations. Values of Moran's I statistics were I_k = 0.21 for Divácká highlands, I_k = 0.37 for Studenská higlands and I_k = 0.35 for Uhřická highlands. These values proved an existence of some at least weak spatial autocorrelation. Moran's plot for Divácká highlands is shown in the Fig. 5b.

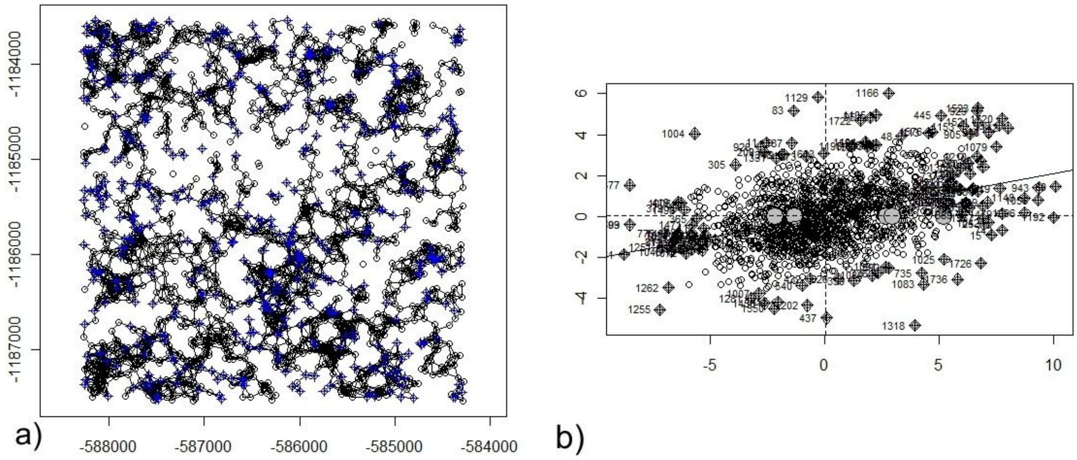

Fig. 5. Visualization of neighbours for each points (a) and Moran scatterplot for DH (b)

Multiple Linear Regression

The main aim of this study was to prove whether the linear relationship between the value of error and morphometric parameters exists. The multiple linear regression was used for the first assessment of the data. The value of error of created DEM was used as dependent variable and morphometric parameters and elevation as independent (explanatory) variables. The intercept and partial regression coefficients of linear models were calculated using function for fitting linear models in R. Because some correlations between explanatory variables exist, the reduction of the model can be expected. The simplifications of linear models were performed using the stepwise algorithm with combined forward and backward direction. Finally, the formula with generally usable variables (regression coefficient depends on the particular situation) was chosen. This formula has form: *Error ~ Intercept + Slope + Ascent + Profile Curvature + Elevation.* Resulting coefficients, *AIC* and R^2 of the computed models are shown in the Tab. 2. The coefficients of determination (R^2*)*, which expresses the quality of the model, are very low R^2 = 0.06 – 0.17. That means that these linear models are not suitable for the modelling of errors using morphometric characteristics and elevation.

Tab. 2. Values of intercept, partial regression coefficients and AIC and coefficient of determination (R^2). DH is original multiple regression model for Divácká highlands containing all explanatory variables, DH step is same model after stepwise regression procedure and General DH is finally chosen general model which uses only four explanatory variables. The same is for Studenská highlands (SH) and Uhřická highlands (UH).

		Partial regression coefficients								
	Intercept	**Slope**	**Ascent**	**Planimetric Curvature**	**Profile Curvature**	**Overall Curvature**	**Elevation**	***AIC***	***R*²**
DH	6.492	-0.016	0.004	55.558	-56.026	-55.560	-0.024	8128.873	0.17
DH step	6.392	-	0.004	-	-0.455	-	-0.024	8125.068	0.17
General DH	6.491	-0.016	0.004	-	-0.455	-	-0.024	8125.245	0.17
SH	3.374	0.071	0.003	-243.273	242.902	243.421	-0.008	4720.827	0.06
SH step	3.374	0.071	0.003	-243.273	242.902	243.421	-0.008	4720.827	0.06
General SH	3.425	0.071	0.003	-	-0.631	-	-0.008	4720.834	0.06
UH	11.431	-0.024	0.003	434.896	-435.033	-434.935	-0.042	7999.138	0.16
UH step	11.277	-	0.003	436.557	-436.686	-436.591	-0.042	7998.318	0.16
General UH	11.527	-0.024	0.003	-	0.045	-	-0.042	8006.681	0.16

Geographically Weighted Regression

It was proved earlier, that a spatial autocorrelation exists in the spatial distribution of the error. Although the autocorrelation is not very strong, certain spatial trend can be assumed. Using spatial extension of linear regression – GWR we tried to improve the general multiple linear models. GWR needed to define spatial weights for each point. Both methods of spatial kernels (fixed bandwidth and adaptive bandwidth) were tested and the adaptive kernel was chosen. Sizes of the adaptive kernels were chosen using the automated procedure *gwr.sel* in the R package *spgwr*, which was used also for the computation of GWRs. Parameters of adaptive kernels for GWR as well as their *quasi-global* R^2 and *AIC* are shown in the Tab. 3. We can roughly compare resulting values of GWR with matching multiple linear regression and see the improvement of the model quality. Tab. 3 also shows that coefficient of determination increased significantly in GWR in comparison with linear models and by contrast the value of *AIC* decreased which also proved enhancement of model by using GWR.

Tab. 3. Size of adaptive kernels for GWR, *AIC* and coefficient of determination (R^2).

	Adaptive kernel (quantile)	**Average no. of points in adaptive kernel**	**AIC**	**R²**
DV	0.006	10	7412.95	0.54
SH	0.013	12	4309.20	0.48
UV	0.005	8	7224.43	0.59

CONCLUSIONS

Firstly we tried to prove that it is possible to model and also predict errors of DEM using correlation and multiple linear regression. But after the computing of both we discovered that it is not suitable to use these non-spatial methods. We also computed the Moran's I statistic of spatial autocorrelation and proved existence of certain type of spatial trend in data sets. That is why we used local method of linear regression – GWR. GWR significantly improved our linear models (R^2 increased about 0.3). Notwithstanding these improvements of models, we cannot recommend linear regression and even GWR for the modelling of DEMs errors spatial distribution (at least for the highlands). While using the multiple linear regression we computed particular formula for the prediction of error, using the GWR the situation and also the interpretation is more difficult, because of the nature of GWR, which is a local method. It means that a basic multiple linear regression is computed for all points using the defined kernel and also partial regression parameters are calculated for this particular point.

REFERENCES

Anselin, L. (1995) Local indicators of spatial association – LISA, Geographical Analysis, 27, pp. 93-115.

Anselin L. (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fisher M, Scholten HJ, Unwin D (eds) Spatial analytical perspectives on GIS. Taylor & Francis, London.

Carlisle, B. H. (2002) Digital elevation model quality and uncertainty in DEM-based spatial modelling [online]. [cit. 2010-01-25]. Greenwich, United Kingdom: University of Greenwich. Ph.D. thesis, URL: http://www.numyspace.co.uk/~unn_szbc1/PhD/index.htm

Carlisle, B. H. (2005) Modelling the Spatial Distribution of DEM Error. Transaction in GIS, 9, 4, pp. 521-540.

Charlton, M., Fotheringham, A. S. (2009) Geographically Weighted Regression: White Paper, 14 p., http://ncg.nuim.ie/ncg/GWR/GWR_WhitePaper.pdf

Draper, N. and Smith, H. (1981) Applied Regression Analysis, 2d Edition, New York: John Wiley & Sons, Inc.

Erdogan S. (2009) A comparison of interpolation methods for producing digital elevation models at the field scale. Earth surface processes and landforms, 34, pp. 366-376.

Erdogan S. (2010) Modelling the spatial distribution of DEM error with geographically weighted regression: An experimental study. Computer & Geosciences, 36, pp. 34-43.

Evans, I. S. (1972) General geomorphometry, derivatives of altitude, and descriptive statistics. In: Spatial Analysis in Geomorphology. Londýn: Methuen, pp. 17-90.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002) Geographically weighted regression: the analysis of spatially varying relationships. Chichester, Willey.

Griffith, D. A. (2009) Spatial Autocorrelation, In: Kitchin, R. and Thrift, N. (eds), International Encyclopaedia of Human Geography, pp. 308-316.

Hocking, R. R. (1976) The Analysis and Selection of Variables in Linear Regression, Biometrics, 32.

Lu, Y., Thill, J.-C. (2009) Assessing the cluster correspondence between paired point locations. Geographical Analysis, 35, pp. 290-309.

Murdoch, D.J., Chow, E.D. (1996) A graphical display of large correlation matrices. The American Statistician 50, 178-180.

Páez, A., Wheeler, D.C. (2009) Geographically Weighted Regression, In: Kitchin, R. and Thrift, N. (eds), International Encyclopaedia of Human Geography, pp. 407-414.

Pearce, J. (2009) Regression, Linear and Nonlinear, In: Kitchin, R. and Thrift, N. (eds), International Encyclopaedia of Human Geography, pp. 302-308.

Propastin, P., Kappas, M., & Erasmi, S. (2008) Application of geographically weighted regression to investigate the impact of scale on prediction uncertainty by modelling relationship between vegetation and climate. Journal of Spatial Data Infrastructures Research, 3, 73–94.

Robinson, G. M. (2009) Statistics - Overview, In: Kitchin, R. and Thrift, N. (eds), International Encyclopaedia of Human Geography, pp. 436-451.

Sarkar, D. (2008) Lattice: Multivariate Data Visualization with R, Springer. http://lmdvr.r-forge.r-project.org/

Steel, R. G. D., Torrie, J. H. (1960) Principles and Procedures of Statistics, New York: McGraw-Hill, pp. 187-287.

Svobodová J. (2011) Hodnocení kvality digitálních výškových modelů pro environmentální aplikace, Katedra fyzické geografie a geoekologie, Přírodovědecká fakulta, Ostravská univerzita v Ostravě. Ostrava . Ph.D. thesis.