ASPECTS CONCERNING THE OUTLIERS PROBLEM IN THE CONTEXT OF DIGITAL CLIMATIC MAPPING

Cristian Valeriu PATRICHE

Key words: mean annual precipitations, spatial statistical models, outliers, Moldavia.
Cuvinte cheie: precipitaţii medii anuale, modele spaţiale statistice, outliers, Moldova.

ABSTRACT:

When regression analysis is used as a global spatialisation method for climatic variables, one must pay special attention to the presence of values evading the spatial variation rules stated by the model (outliers). The outliers may alter significantly our regression models, therefore leading us to drawing the wrong conclusions. Our study focuses on the outliers problem through a simple example of mean annual precipitations spatialisation in eastern Romania using the altitude as predictor. The identification of the outliers is based on the magnitude of the residuals, on cross-validation and on the comparison of the regression residuals with the deleted residuals (jackknife error). After the identification stage, we construct regression models leaving out the outliers in order to quantify their negative effects. We then present several possible options to avoid these effects, focusing on the one which eliminates the outliers from the regression models but keeps the residual values in the respective points during the kriging stage in a residual kriging approach.

1. Introduction

Our study focuses on the identification of outliers, the assessment of their influence on the regression-based spatial models of climatic parameters and on some possibilities of dealing with this problem.

An outlier is a point value showing a significant deviation from the statistical model (therefore marked by a high residue), corresponding to points (meteorological stations, rain gauges) denoting the presence of spatial anomalies for the analysed parameter's distribution (e.g. föehnization areas, areas of orographic enhancement of precipitations, temperature inversion areas etc.). Such a "rebel" value may be also an error value and this possibility must be checked out. If no error is identified then we should proceed to the assessment of the degree in which this value is altering the statistical models, mainly the regression models. This is happening in the case of the regression analysis because it is used mainly as a global interpolation method and the regression itself is incapable to render spatial anomalies. If such spatial anomalies exist, then the integration within the statistical model of values describing these anomalies may significantly alter the regression equations, which therefore become unreliable (Patriche C.V., 2007).

2. Input data and methods

Our study region lies in eastern Romania (the region of Moldavia) comprising a relief of hilly plains, hills and plateaus, covering a surface of about 30000km^2 (figure no. 1). We analysed the outliers influence on the spatial distribution of mean annual precipitations using a sample of 28 stations.

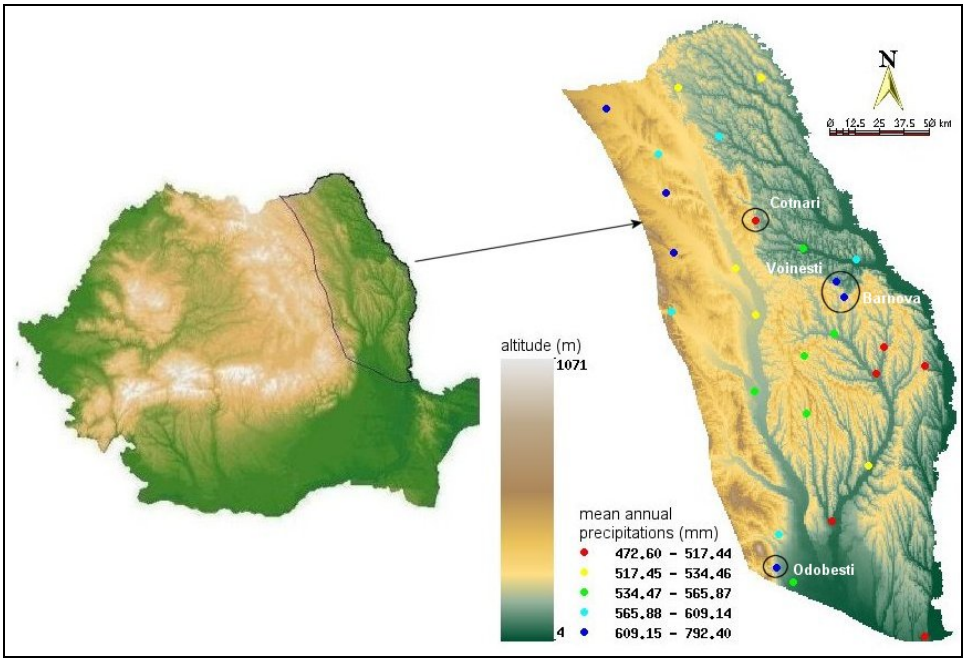


Fig. 1. Location of the study region and of analysed sample points showing the position of 4 possible outliers.

Numerous studies have proved that, among the various spatialisation methods, the regression and kriging generate the best spatial models for climatic variables (Dobesch et al., 2007, Hengl, 2007, Silva et al., 2007, Lhotellier, 2005). We used the multiple linear stepwise regression as a global spatialisation method combined with ordinary kriging of the residuals (residual kriging, regression kriging, detrended kriging) for deriving spatial models of mean annual precipitations. The spatial analysis including the application of the regression equations and kriging interpolations was performed using the TNTmips 6.4 software[1]. We tested the influence of many potential predictors (Lhotellier, Patriche, 2007) derived at a spatial resolution of 90m starting from the SRTM DEM (USGS, 2004), such as latitude, longitude, slope, West-East aspect component, relief local energy, plan and profile curvature, low-pass filtered grids in order to account for scale dependency of precipitations. We also tested the

[1] TNT Map and Image Processing System, Lincoln, Microimages Inc.

influence of some qualitative predictors (CLC[2] land use and indicators of physiographic subregions) through ANOVA analysis.

3. Results and discussions

In spite of the various potential predictors used, only one variable stands out as a statistically significant explicative factor for the spatial distribution of annual precipitations, respectively the local altitude of the stations. This may have several explanations, besides the simple fact that these predictors do not have any influence upon the dependent variable: the weak spatial representativeness of the station network due to both the its feeble density and its biased location mainly in valley bottoms; the local action of some predictors or the combined effects of some predictors: e.g. West-East aspect component becomes a significant predictor only when the altitude range (the local relief energy) is high, or high slopes determine local enhancements of precipitations when they are exposed towards West and when they are associated with high relief energy values.

From the viewpoint of their influence upon the regression models, we may identify two types of outliers:

- *Type one*: outliers showing high residues but with similar values of the real residues and the deleted residuals (computed without taking into account the anomaly point – jackknife error). Because such outliers do not modify significantly the regression models, they can be therefore included in the analysis.
- *Type two*: outliers showing high residues but with significant differences between the values of the real residues and those of the deleted residuals. Such outliers modify the regression model and must be therefore eliminated if the induced modifications are proved to be significant.

How can we *identify* an outlier? How great should a residue value be in order to regard the corresponding point as an outlier? The simplest way is the *visual inspection of the correlation charts*.

Figure no. 8 shows the correlation between the mean annual precipitations and the altitude. The chart indicates at least 2 suspect points situated outside the correlation cloud, one with a lower precipitation value then expected for the respective altitude (Cotnari station), another with significantly higher precipitation amounts then expected (Bârnova station). These deviations are related to local terrain conditions influencing the pluviometry. Cotnari station is situated in a föehnization area of western air masses descending the eastern slopes of Dealul Mare – Hârlău Hill. Here, the real mean annual precipitation value is 121.3 mm lower than the value predicted by the altitude regression model using all stations. On the contrary, Bârnova station is situated in an area of orographic enhancement of precipitations caused by the presence of a high energy slope (Iasi Cuesta) facing the more humid western air masses and by the shape of the Bârnova-Voineşti depression which causes the convergence of the western air

[2] Corine Land Cover

masses. Another factor is related to the location of Bârnova station within a well forested area. Being the only station from our sample situated within forested areas, it is impossible for us to assess the relative importance of these factors and to state which of them, the local topography or the presence of the forest, is more responsible for the high precipitation values recorded at this location. The real mean annual precipitation value at Bârnova station is 172.7 mm higher than the predicted value.

If the visual inspection of the correlation charts gives us a first guess on the presence of possible outliers, other methods provide more insight. First, we should inspect *the magnitude of the residuals*. Generally, if some value goes out the interval limited by ±2.5 RMSE[3] (equivalent with the standard deviation of the residues), then it is possible that this value is an outlier. From figure no. 2 we notice that the residue from Bârnova station goes beyond the +2.5 RMSE, while the residue from Cotnari station is very close to the – 2.5 RMSE limit. If we eliminate only Bârnova station we find that the residual value at Cotnari goes also beyond the specified limit. So the conclusion is that both stations must be excluded to ensure stability for the regression model. But if we exclude these two stations and rebuild our regression model, we shall find that yet another station (Odobeşti) displays residues greater than the + 2.5 RMSE limit. Furthermore, if we chose to eliminate Odobeşti station, we obtain another high residual value for Voineşti station, situated in the same area of orographic enhancement of precipitations as Bârnova station, only at a lower altitude. Should we eliminate these stations as well?

So far we have established that we have some poor estimated points in our sample, displaying high residual values. So we are certain that we have some points acting like type 1 outliers (referring to the above classification). But is it necessary to eliminate them from the regression model? Would this elimination improve significantly the model?

To answer this question one must test the influence of these outliers on the regression models and find out whether or not we are dealing with outliers of type two.

One way to establish that is to perform *cross-validations*, that is to compare the observed values with the predicted values obtained by successive elimination of the sample points. If the regression models are stable, one should find that the cross-validation charts are similar to the correlation charts between the observed and the predicted values. In our case, we may notice that the differences between the observed vs. predicted correlations and the cross-validation correlations decrease as the outliers are removed from the models, from about 11%, in the case of all stations model, to about 6%, in the case of the regression model obtained by removing all 4 possible outliers (figures no. 2-6). The slight difference is hampering us so far to state that the removal of the 4 stations improves significantly the regression models.

The comparison between the observed vs. predicted values and the cross-validation charts only tells us something about the stability of the regression models. In order to investigate the influence of particular values, we may find it

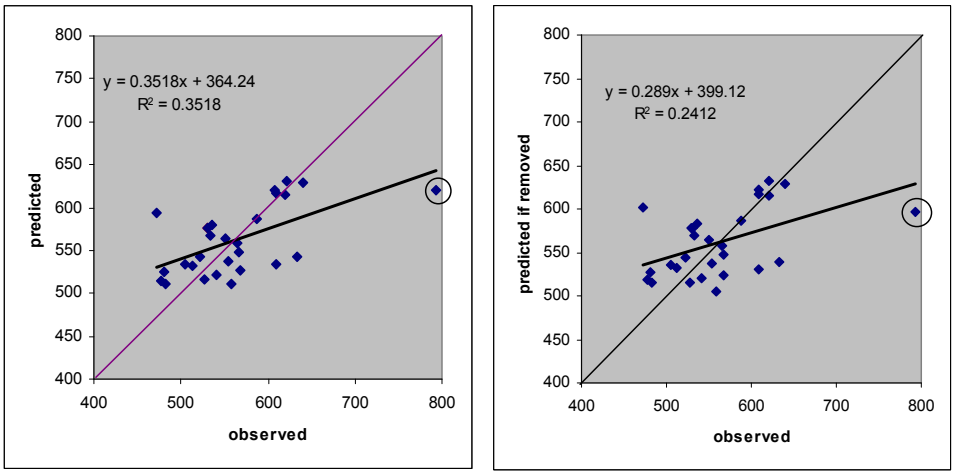
[3] Root Mean Square Error

useful *to compare the regression residues with those obtained by eliminating the suspect point* (named deleted residuals or jackknife error). If the suspect point is not an outlier, then the magnitude of the residues should be very similar. In our case, we notice that the difference between the actual and the deleted residuals is the greatest in the case of Bârnova (22.5mm), which means that its exclusion from the model changes significantly the altitude – precipitation relationship (figures no. 2, 3). The next greatest difference we find in the case of Cotnari station (7.8mm). Even if this is not such an important difference, keeping Cotnari station without Bârnova station generates an even poorer regression model than the one using all stations. This is due to the fact that these 2 points, one above, the other bellow the regression line, have opposite effects, balancing the regression line to the extent that if one point is removed, the other will "attract" the line towards it. This means that if we chose to eliminate Bârnova station, we must eliminate Cotnari station as well.

If we construct our model without these two stations and analyze the residuals, we find that yet 2 other stations display high residuals, going beyond the +2.5 RMSE: Odobeşti and Voineşti stations, the latter being situated within the same area of orographic enhancement of precipitations as Bârnova station (figures no. 4, 5). However, the difference between the actual and the deleted residuals is not very significant. The elimination of all these 4 stations leads to a regression model where no more points display residuals beyond the 2.5 RMSE limits.

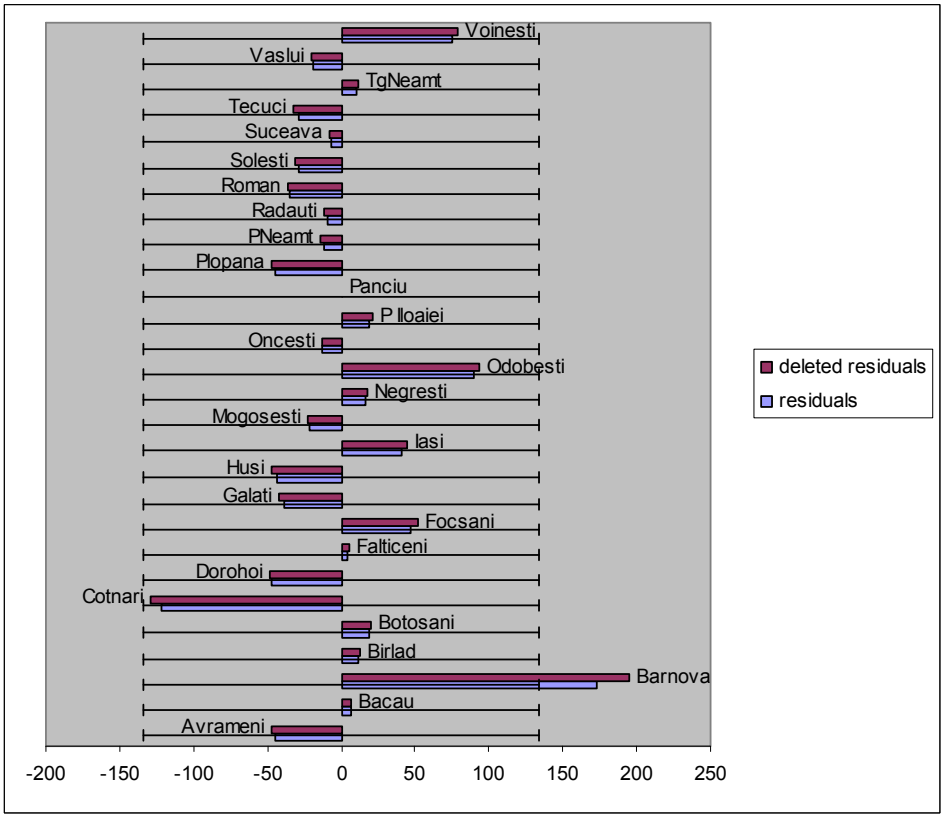
Table 1 shows how significant is the influence of the 4 outliers on the regression models. We notice that the regression quality parameters (correlation coefficients, standard error of estimate) improve by excluding these outliers. However, one should bear in mind that even if there is an overall improvement of the regression models excluding the outliers, these models will still perform poor in the case of the outliers themselves. But is the altitude – precipitations relationship significantly changing? As we stated before, the regression model without Bârnova only is not reliable due to the "attraction" effect of the Cotnari station and we can clearly see that this model is the most different from the others, showing the highest intercept and the lower pluviometric vertical gradient (regression coefficient). The other models display quite similar parameters: intercepts ranging from 485.6mm to 498.9mm and gradients from 30.1mm/100m to 36.2mm/100m. From figure no. 7 we may see that 31% of the station sample display the lowest residuals under the 2^{nd} model (without Bârnova and Cotnari stations). A similar percent (30%) is found for the 4^{th} model (without all 4 outliers).

To sum up, *our conclusion is that, in the particular case of our sample, the elimination of the identified 4 outliers improves the regression model even though the differences among the various models are not very important.*



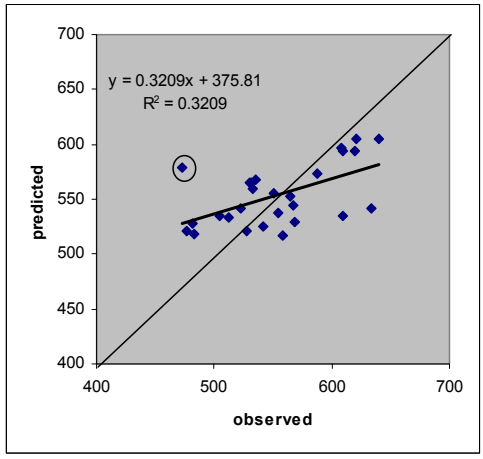
a

b

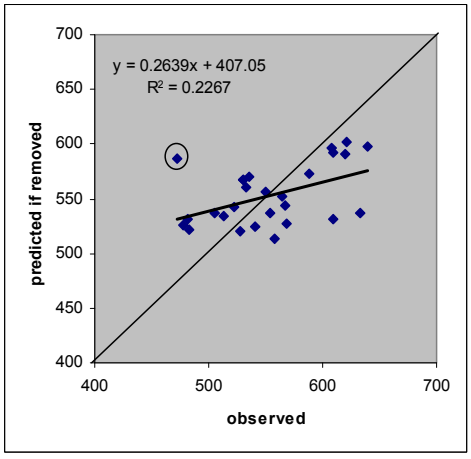


c

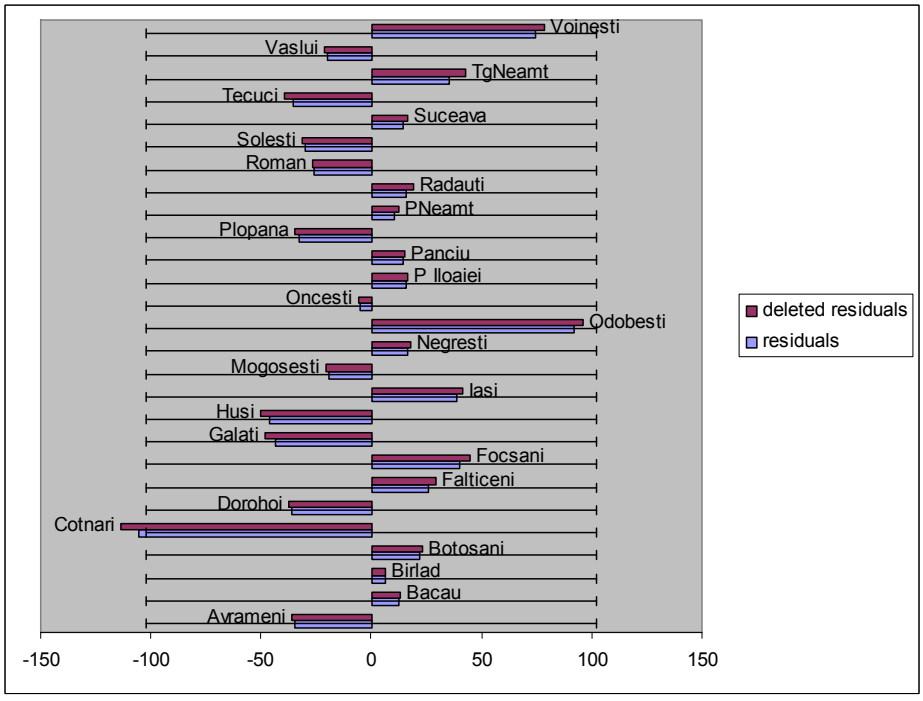
Fig. 2. Correlation between observed and predicted mean annual precipitation values using all stations (a), cross-validation (b) and comparison of the residuals and the deleted residuals with bars showing the ± 2.5 RMSE (c).



a

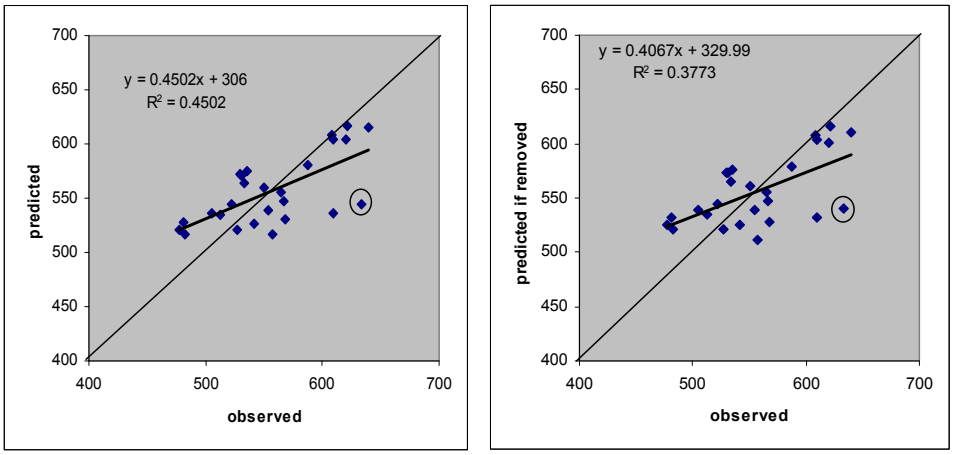


b

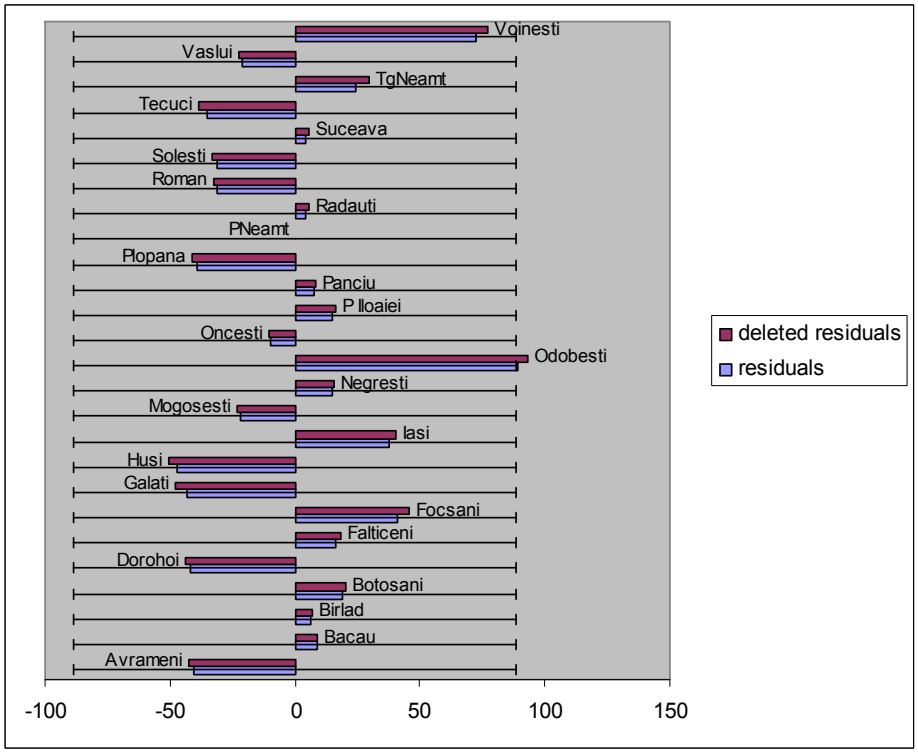


c

Fig. 3. Correlation between observed and predicted mean annual precipitation values obtained by removing Bârnova station (a), cross-validation (b) and comparison of the residuals and the deleted residuals with bars showing the ± 2.5 RMSE (right).



a b



c

Fig. 4. Correlation between observed and predicted mean annual precipitation values obtained by removing Bârnova and Cotnari stations (a), cross-validation (b) and comparison of the residuals and the deleted residuals with bars showing the ± 2.5 RMSE (c).

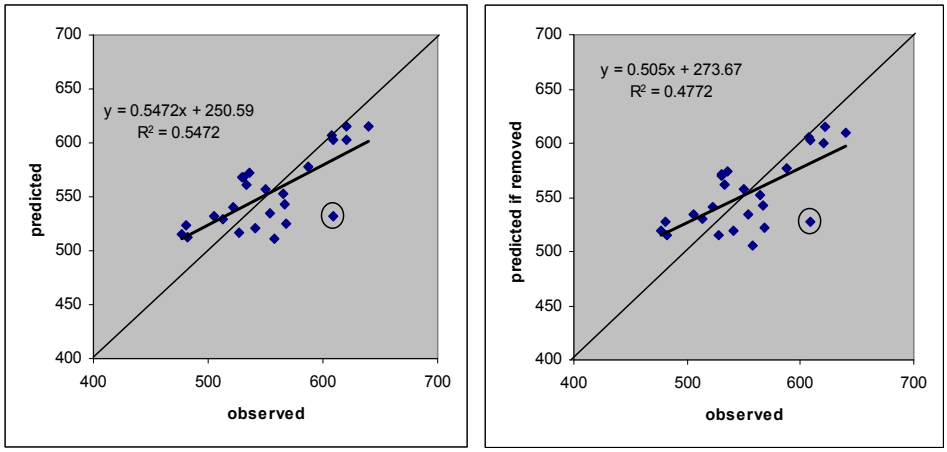
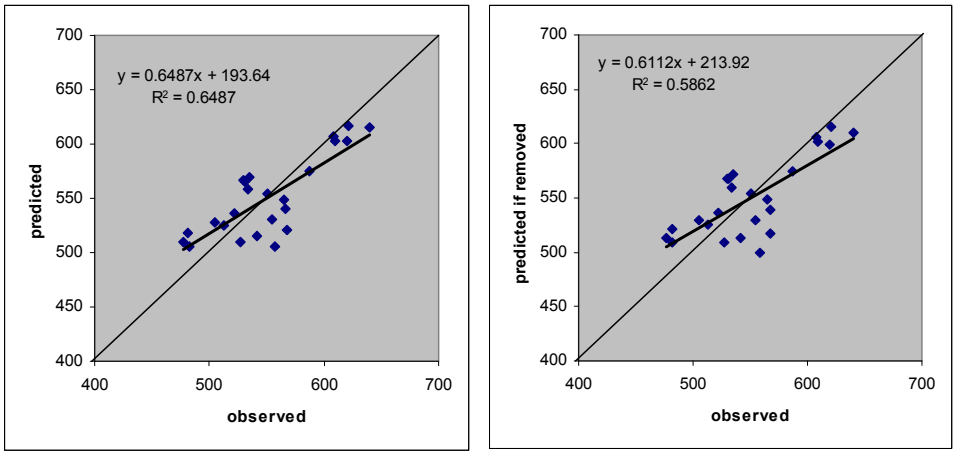
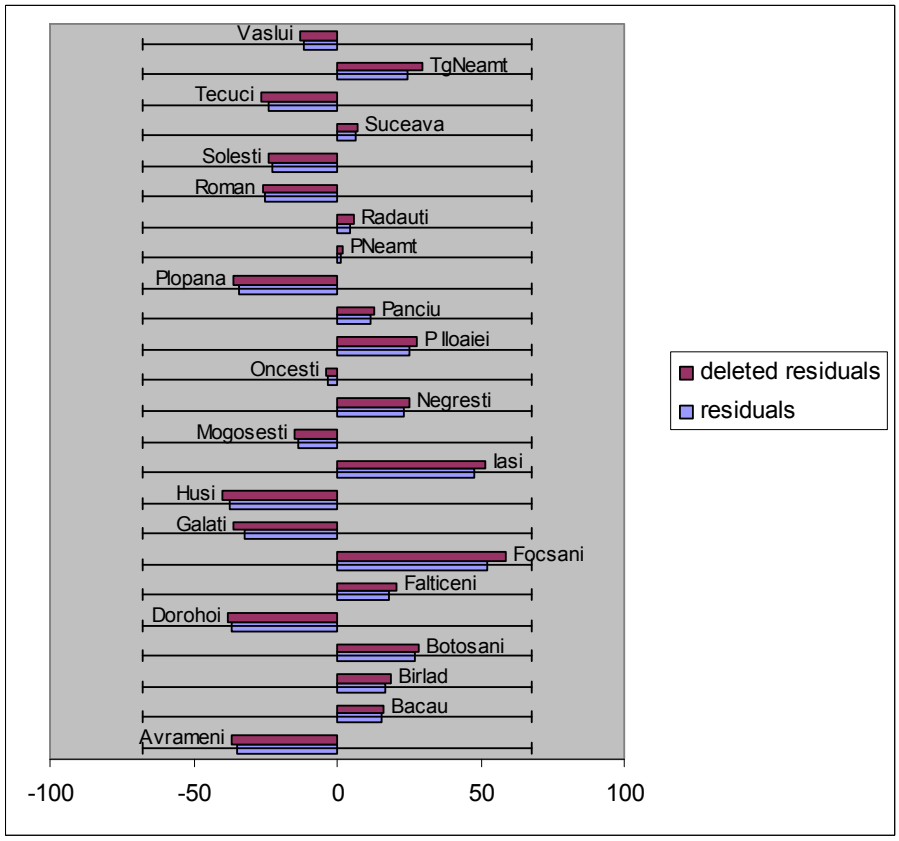


Fig. 5. Correlation between observed and predicted mean annual precipitation values obtained by removing Bârnova, Cotnari and Odobeşti stations (a), cross-validation (b) and comparison of the residuals and the deleted residuals with bars showing the ± 2.5 RMSE (c).



a

b



c

Fig. 6. Correlation between observed and predicted mean annual precipitation values obtained by removing Bârnova, Cotnari, Odobeşti and Voineşti stations (a), cross-validation (b) and comparison of the residuals and the deleted residuals with bars showing the ± 2.5 RMSE (c).

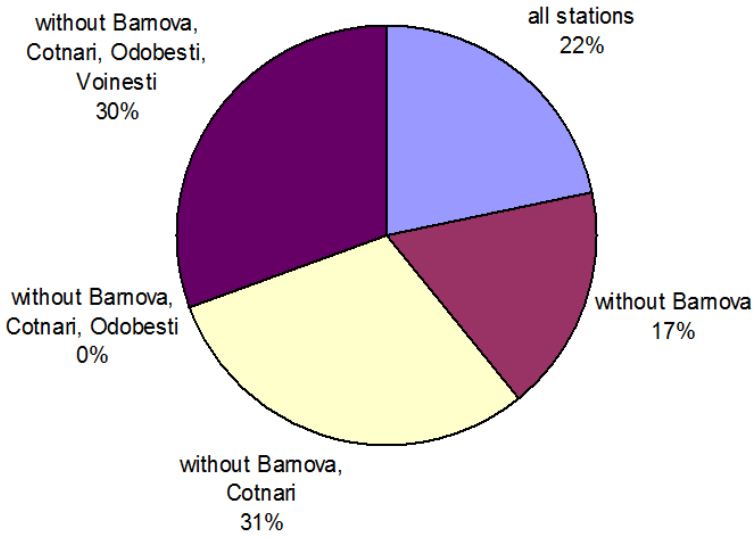


Fig. 7. The optimum altitude regression model (actual residuals minus deleted residuals) for each station

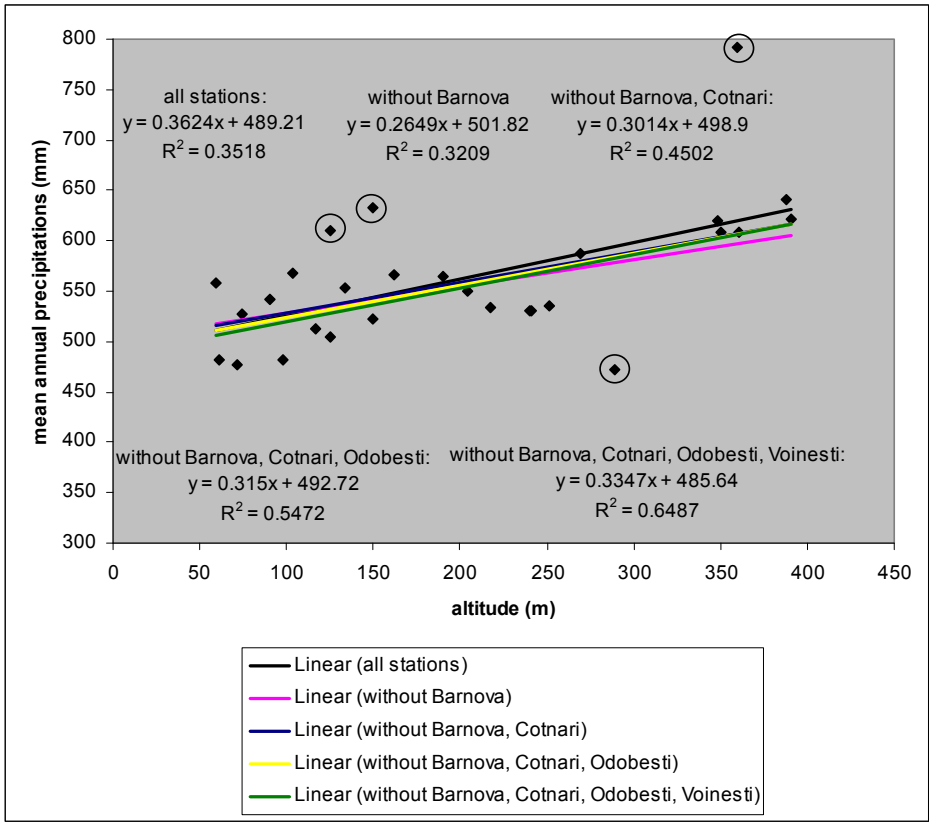


Fig. 8. The altitude – mean annual precipitations relationship showing the presence of 4 possible outliers and the regression lines derived by the successive elimination of these outliers.

Table 1. Comparison of the regression models using and excluding the outliers

Regression model	Intercept	Regression coefficient	R^2	Standard error of estimate
All stations	489.21	0.362	0.352	54.472
Without Bârnova	501.82	0.265	0.321	41.678
Without Bârnova, Cotnari	498.90	0.301	0.450	36.190
Without Bârnova, Cotnari, Odobeşti	492.72	0.315	0.547	31.697
Without Bârnova, Cotnari, Odobeşti, Voineşti	485.64	0.335	0.649	27.626

But is this a right approach? Is it correct to eliminate certain stations from our sample? The problem is that we cannot just exclude some real values from the analysis because then we would obtain an incomplete image of the spatial distribution of the analyzed climatic parameter. The solution may be the elaboration of the regression model without the values identified as outliers, the spatialisation of the residuals by ordinary kriging, including the residuals associated with the anomaly points, followed by the addition of the spatial trend with the interpolated residuals so as to obtain the final spatialisation. We notice that this is a *residual kriging approach which eliminates the outliers during the regression stage*, if these belong to the type two mentioned above, *but includes the residuals from these points during the kriging interpolation stage* (figure no. 9).

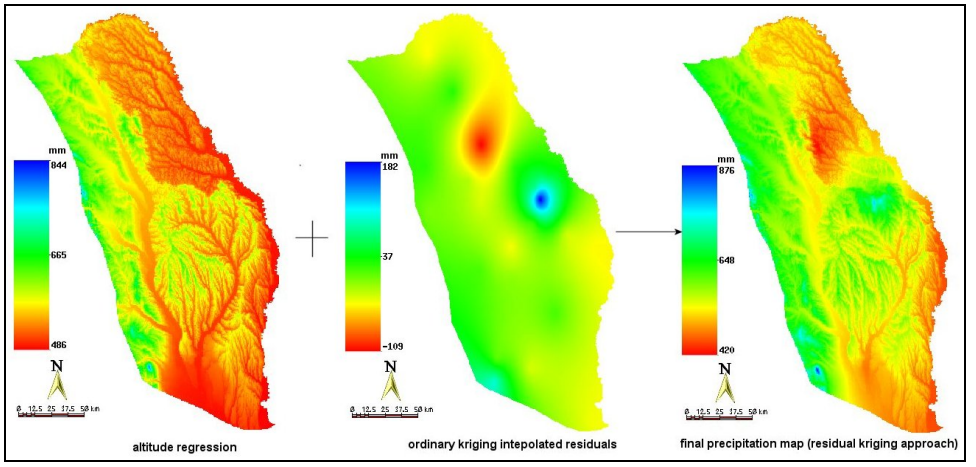


Fig. 9. Mapping the optimum solution: residual kriging approach leaving out the outliers during the regression stage but taking the outliers' residuals into account during the kriging stage.

Other approaches are possible, such as the *insertion of one or several predictors* within the regression model that would account for the spatial anomalies. This option is however debatable because if regression is applied as a global interpolator then it is unable to render spatial anomalies, no matter what predictors we use. It only gives us a trend surface, a spatial pattern specific to the analyzed parameter. It is also true that a spatial climatic anomaly can be regarded as a more intense manifestation of a general process. For example, the fact that westward facing slopes are generally wetter than eastward facing slopes may be regarded as a general rule, for middle latitudes at least, but the orographic enhancement of precipitation or föehnization areas occur locally where the relative altitude of terrain is higher. So theoretically, a combination of predictors such as the west-east exposition and relative altitude should be able to depict the spatial anomalies mentioned above. Practically, we are often hampered in our analysis by the poor spatial representativeness of the stations network which is, in most cases, unable to fully account for all terrain aspects relevant for the spatial distribution of the analyzed climatic parameter.

Another possibility to deal with the outliers problem would be to apply *the regression analysis as a local interpolator* (*Engen-Skaugen, Tveito*, 2007) or to apply *a weighted regression* in which the local regression model depends mainly on the neighboring points which have higher weights and less on further points having lower weights (*Maracchi et al.*, 2007) This approach however is also hampered by the scarcity of the stations network.

Conclusions

- When applying the regression as a global interpolation method for the purpose of deriving digital spatial models of climatic variables one must take great care in identifying and assessing the sources of uncertainty.
- One of these sources is the presence of values evading the spatial variation rules stated by the models (outliers) which can negatively influence the regression models, leading the researcher to wrong conclusions.
- In order to identify the outliers, one should first inspect the configuration of the correlation cloud between the dependent variable and the predictor, or between the real and the predicted values, in the case of multiple predictors, looking for points situated significantly outside the cloud. If such points exist, we should further inspect their residual values and see if they are located outside the ± 2.5 RMSE interval. If such points exist, we should then test their influence on the regression models, analysing the differences between the actual residual values and the deleted residuals (jackknife error). If these differences are important then the exclusion of the respective points significantly changes the regression model which is therefore unstable. Next we should actually see these changes by elaborating the models with and without the outliers and finally decide whether to keep or to eliminate the respective points.
- Nevertheless because the exclusion of real values form analysis is not correct we should derive our final spatial model by a residual kriging approach,

eliminating the outliers from the regression stage but keeping the residuals for these points within the kriging stage.

REFERENCES

Dobesch H., Dumolard P., Dyras I. (editors, 2007), *Spatial Interpolation for Climate Data. The Use of GIS in Climatology and Meterology*, ISTE, 320 pp.

Engen-Skaugen T., Tveito O.E. (2007), *Spatially distributed temperature lapse rate in Fennoscandia*, in COST action 719: Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, 25-29 October 2004, Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M. (editors), Luxembourg: Office for Official Publications of the European Communities, 2007, EUR 22596, p. 93-100.

Hengl T. (2007), *A Practical Guide to Geostatistical Mapping of Environmental Variables,* JRC Scientific and Technical Research series, Office for Official Publications of the European Comunities, Luxembourg, EUR 22904 EN, 143 pp.

Lhotellier R., Patriche C.V. (2007), *Dérivation des paramètres topographiques et influence sur la spatialisation statistique de la température*, Actes du XXème Colloque de l'Association Internationale de Climatologie, 3-8 septembre 2007, Carthage, Tunisie, p. 357-362.

Lhotellier, R. (2005), *Spatialisation des températures en zone de montagne alpine*, thèse de doctorat, SEIGAD, IGA, Univ. J. Fourier, Grenoble, France, 350 p.

Maracchi G., Ferrari R., Magno R., Bottai L., Crisci A., Genesio L. (2007), *Agrometorological GIS products through meteorological data spatialization*, in COST action 719: Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, 25-29 October 2004, Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M. (editors), Luxembourg: Office for Official Publications of the European Communities, 2007, EUR 22596, p. 9-16.

Patriche C.V. (2007), *About the influence of space scale on the spatialisation of meteo-climatic variables*, Geographia Technica, Nr. 1 / 2007, Cluj University Press.

Patriche C.V. (2009), *Metode statistice aplicate în climatologie*, Edit. Terra Nostra, Iaşi.

Silva Á. P., Sousa A. J., Espírito Santo F. (2007), *Mean air temperature estimation in mainland Portugal: test and comparison of spatial interpolation methods in Geographical Information Systems*, in COST action 719: Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, 25-29 October 2004, Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M. (editors), Luxembourg: Office for Official Publications of the European Communities, 2007, EUR 22596, p. 37-44.

USGS (2004), *Shuttle Radar Topography Mission*, 3 ArcSecond (90m), scenes: SRTM_u2_n182e027, SRTM_u2_n182e028, SRTM_u2_n183e026, SRTM_u2_n183e027, filled finished-B, Global Land Cover Facility (www.landcover.org), University of Maryland, College Park, Maryland, 2000.

* * * (2000), *Reference manual for the TNT products* V6.4, Lincoln, Microimages Inc.

Cristian Valeriu PATRICHE
Academia Română, Filiala Iaşi, Colectivul de Geografie
E-mail: pvcristi@yahoo.com