IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 13–30

Aspects regarding the uncertainty of spatial statistical models of climate parameters

Patriche Cristian Valeriu

Romanian Academy, Department of Iasi, Geography Group,
8 Carol I, 700505, Iasi, Romania; E-mail: pvcristi@yahoo.com

(Manuscript received in final form April 26, 2010)

Abstract — Any transformation of a discrete variable into a continuous one is subject to uncertainty. Consequently, the identification and assessment of errors is essential for avoiding misinterpretations of models describing the spatial distribution of climatic parameters. Our study attempts to identify the main sources of errors affecting the statistical spatial models of climatic parameters and to assess their impact on the accuracy of these models. In particular, we focus on georeference errors, the representativeness of the stations network and the related extrapolation problem, the outliers problem, error propagation from simple to complex variables, the problems aroused by heterogeneous regions.

Key-words: uncertainty, spatial statistical models, climate parameters, georeference errors, stations network representativeness, extrapolation, outliers, heterogeneous regions, error propagation

1. Introduction

Our study derives from previous attempts to model the spatial distribution of various climate parameters, which were based, in most cases, on small samples of meteorological stations/rain gauges (*Patriche*, 2007; *Patriche et al.*, 2008). Therefore, our conclusions are applied especially to outputs achieved from such samples, knowing that the degree of uncertainty rises significantly as the sample size used for statistical modeling decreases.

There are many potential sources of uncertainty, which may be grouped into two broad categories:

- **Errors from data pre-processing stage (data quality)**
  - Data recording errors / data series gaps;
  - Instrumental errors;
  - Changes in measurements standards;

  - Change in the location of the station / changes in land use around the station.
- **Errors from data processing stage**
  - Georeference errors;
  - Errors derived from the spatial representativeness of the stations network;
  - Errors induced by the presence of outliers;
  - Errors derived from the heterogeneity of the region;
  - Statistical errors;
  - Cumulated errors from computation of complex parameters (error propagation).

Our study focuses on the errors from the data processing stage.

2. Georeference errors

Although simple, the georeference stage is very important. Georeference errors refer to errors of the *x*, *y*, *z* coordinates. Misplacements of stations / rain gauges points on the map may induce significant errors, especially in highly fragmented terrain, when predictors' values are extracted from raster layers or when local interpolators, such as kriging, are used for spatial modeling. The former will lead to wrong predictors' values and, therefore, inaccurate regression models, while the latter will generate locally displaced climatic fields.

The correlation between the stations / rain gauges altitudes and the respective DEM (Digital Elevation Model) altitudes may be used for identifying possible georeference errors or errors in recording the stations / rain gauges altitudes. The correlation should be very good, although not perfect for several reasons: the DEM generalizes the altitude information according to its resolution; the stations / rain gauges latitude and longitude values are generally given in degrees and minutes. Following up the latter issue, if we suppose that the seconds are rounded up or down to the closest minute, it actually means that we may have a coordinate error of up to 30 seconds, meaning about 900 m for latitude and 600 m for longitude, for middle latitudes. These errors double if no coordinate rounding was performed and the seconds were just disregarded.

In the example shown in *Fig. 1*, extracted from a study attempting to model the spatial distribution of mean annual precipitations in Vrancea County, Romania (*Patriche et al.*, 2008), we notice one point (Groapa Tufei) situated outside the correlation cloud indicating a possible georeference error. The recorded altitude for this rain gauge is 125 m, while the DEM altitude for this particular location is 355 m. We can see how far the 125 m altitude isoline is, along which the rain gauge should be located. There are two possible explanations for this error: either the horizontal coordinates of Groapa Tufei are wrong, or the recorded altitude is incorrect. Let us now see the potential

negative impact of such a georeference error on spatial statistical models of precipitations. If the real altitude of Groapa Tufei is 125 m, so the recorded altitude is correct, but the horizontal coordinates are wrong, then this point may be used for regression analysis, provided that neither DEM altitude values nor other derived predictors' values are used for models computation. In a geostatistical approach (ordinary kriging, residual kriging, etc.) it is not advisable to include such misplaced points, because they will misplace, in their turn, the precipitation values. Still, if the value of a misplaced point is similar to those of the neighbouring points, as it is in our case, the error induced by the georeference error may be small enough, and the respective point may be kept.

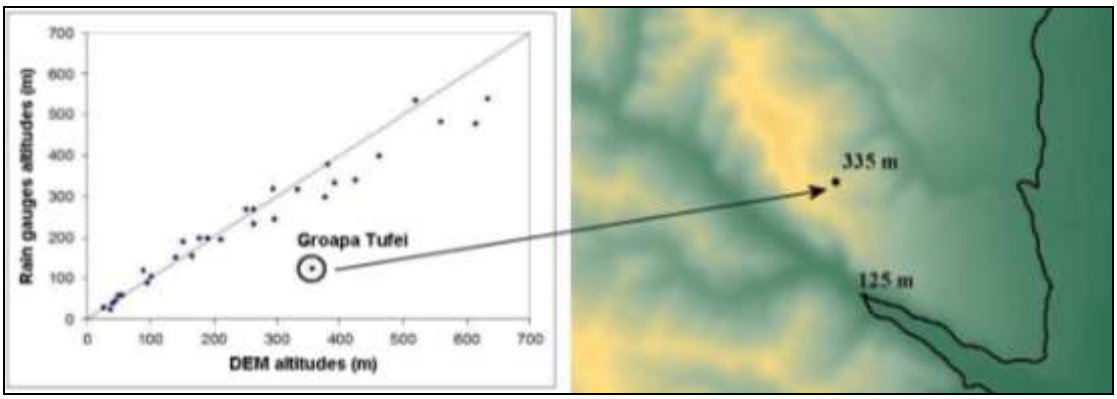

Fig. 1. Revealing two georeference errors for a sample of rain gauges situated in Vrancea County, Romania (*Patriche et al.*, 2008).

3. Spatial representativeness of the stations network and the extrapolation problem

The spatial representativeness of the meteorological stations / rain gauges network is an important issue which needs to be addressed in a preliminary stage, as it constitutes a potential source of errors. Theoretically, the spatial distribution of the meteorological network should be well-balanced, in order to grasp all the meteorological and climatological aspects of a territory. However, in most cases, the spatial representativeness of the stations network is more or less inappropriate, due to both its feeble density and its biased location, mainly in valley bottoms.

The representativeness of the meteorological network in relation with the potential predictors may be visualized and evaluated by comparing the predictors' histograms with the histograms of the same predictors, which are based on the predictors' values associated to the meteorological stations / rain gauges.

An example is given in *Fig. 2* for the altitudinal representativeness for a sample of meteorological stations situated in eastern Romania. In an ideal situation, the curves of the cumulated histograms, derived from the DEM and the stations' altitudes, should overlap. However, we notice the shortage of

stations between 300 m and 350 m of altitude. Also, we observe the lack of stations at lower altitudes (< 59 m) and especially at higher altitudes, where the highest meteorological station is situated at 391 m of altitude, while the terrain altitudes go as high as 1071 m. As a consequence, we are forced to extrapolate the altitude-based regression models in these areas. As the extrapolation may induce errors, we need to give a special attention to these areas and to consider carefully the reliability of the estimated values.

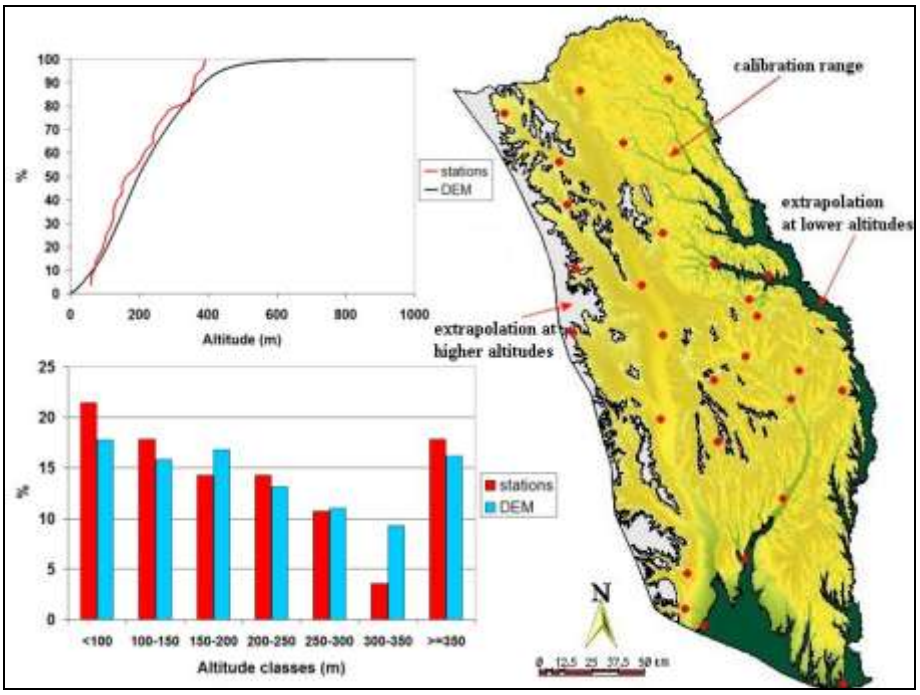

Fig. 2. Assessment of spatial representativeness of stations network by comparing frequencies of predictors' values for station points and for the whole region. Example from eastern Romania (Moldavia) for altitude representativeness for a sample of 28 stations.

Fig. 3 shows an example in which the extrapolation of the regression model should be avoided (*Patriche et al.*, 2008). The mean annual precipitation – altitude regression model, elaborated for Vrancea County (Romania), was based on a sample of 34 rain gauges. The westernmost mountainous part of the region is uncovered by rain gauges, meaning that we must extrapolate our regression model there, if we want to estimate the mean annual precipitation values for this part as well. Performing the extrapolation up to 1770 m of altitude, we estimate precipitation values of up to 1463 mm. Such estimated values are, in our opinion, unrealistic. If the extrapolation is unreliable, then we should confine ourselves with the calibration area of our model. Taking into account that the highest rain gauge altitude is 540 m, we recommend that the study region should not extend over 700 m (*Fig. 3*, black line). Therefore, the entire westernmost part of our region should be excluded from the final map because of extrapolation uncertainty.

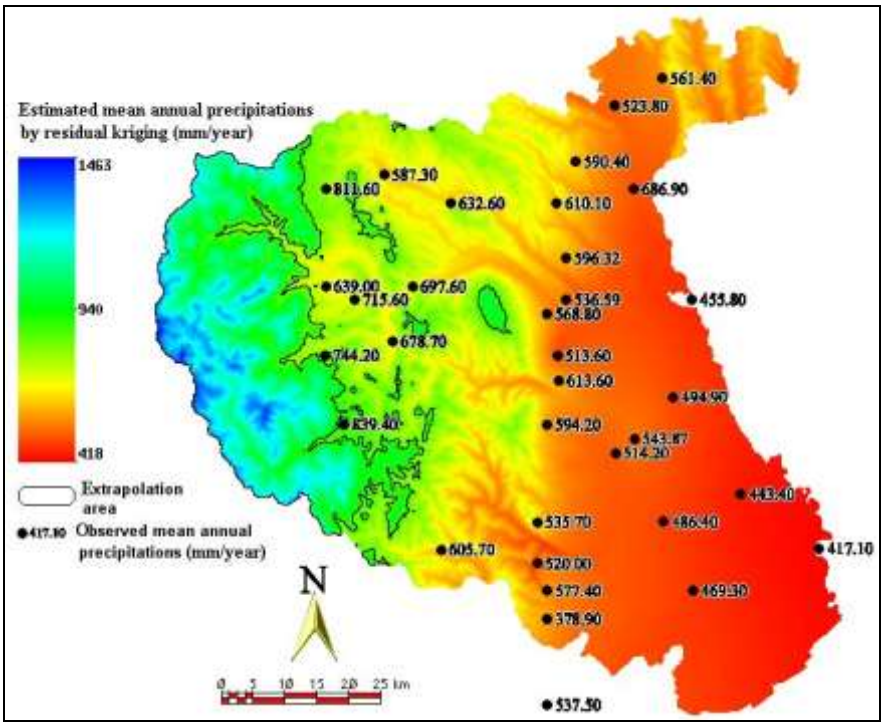

Fig. 3. Avoiding extrapolation. An example from Vrancea County (Romania) for mean annual precipitations.

4. The outliers' problem

An outlier is a point value showing a significant deviation from the statistical model (therefore, marked by a high residual value), corresponding to points (meteorological stations, rain gauges) which mark spatial anomalies for the analyzed parameter's distribution (e.g., foehnization 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 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.

From the viewpoint of their influence over the regression models, we may distinguish two types of outliers:

- Outliers showing high residuals but with similar values of the real residuals and deleted residuals (also known as jackknife error and computed without taking into account the anomaly point). Because such outliers do not modify significantly the regression models, they can be included in the analysis.

- Outliers showing high residuals but with significant differences between the values of the real residuals and those of the deleted residuals. Such outliers modify the regression model and must be, therefore, taken into consideration if the induced modifications are proved to be significant.

There are many statistical procedures aimed towards the identification of outliers. Good syntheses of these procedures are provided by *Maimon* and *Rokach* (2005), and *Wilcox* (2002).

Our approach is a simple one. In order to identify the outliers, we should first inspect the configuration of the correlation cloud between the dependent variable and the predictor, or between the real and predicted values in the case of multiple predictors, looking for points situated significantly outside the cloud. If such points exist, we should further inspect the magnitude of their residual values and see if they are located outside the ± 2.5 RMSE (root mean square error) interval. If such points exist, we should then test their influence on the regression models. The most common way to do this is to perform a cross-validation, the analysis of 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 eliminate the respective points.

Fig. 4 shows the correlation between the mean annual precipitation and the altitude for a sample of 28 meteorological stations situated in eastern Romania (Moldavia). The chart indicates at least 2 suspect points situated outside the correlation cloud, one with a lower precipitation value than expected for the respective altitude (Cotnari station), another with significantly higher precipitation amounts than expected (Barnova station). These deviations are related to local terrain conditions influencing the pluviometry. Cotnari station is situated in a foehnization area of western air masses descending the eastern slopes of Dealul Mare – Harlau 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, Barnova 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 Barnova-Voinesti depression, which causes the convergence of the western air masses. Another factor is related to the location of Barnova 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 Barnova station is 172.7 mm higher than the predicted value.

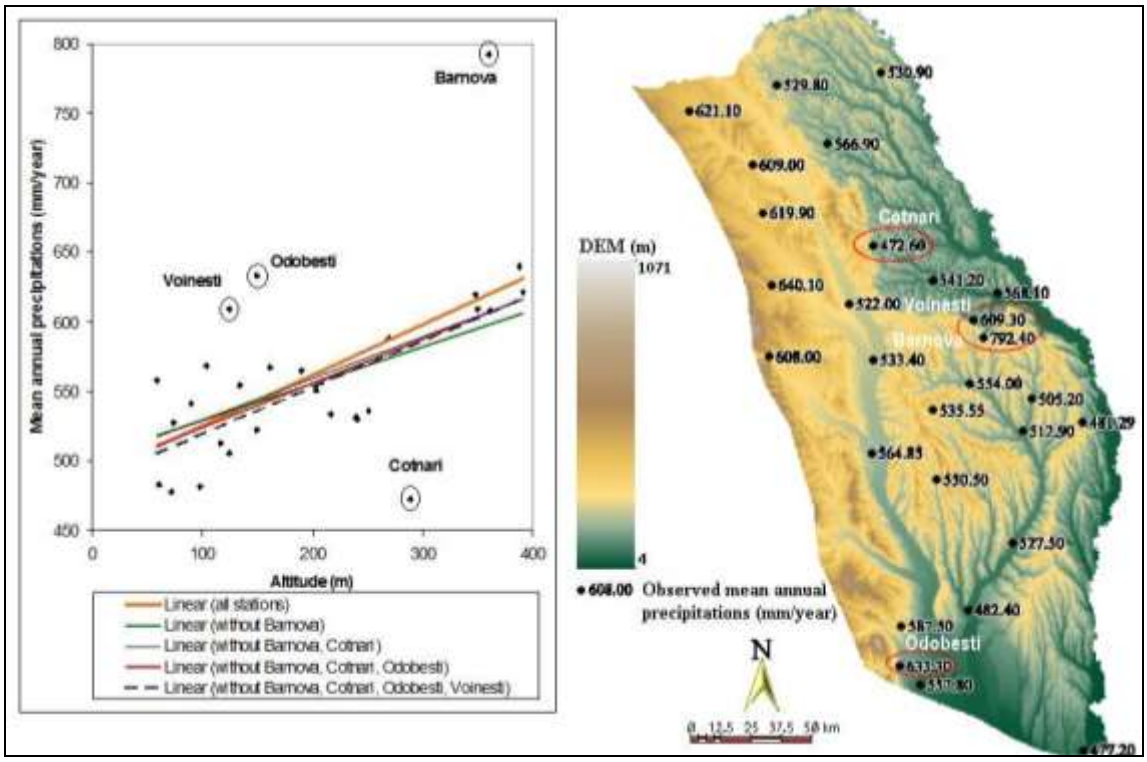

Fig. 4. Correlation chart between altitude and mean annual precipitations for a sample of 28 meteorological stations situated in eastern Romania (Moldavia), indicating the presence of four possible outliers.

If the visual inspection of the correlation charts gives us a first guess on the presence of possible outliers, other methods provide more insight. Our next step is to inspect the magnitude of the residuals. Generally, if some value goes out the interval limited by ± 2.5 RMSE (equivalent with the standard deviation of the residuals for large samples), then it is possible that this value is an outlier. From *Fig. 5c* (left), we notice that the residue from Barnova 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 Barnova station, we find that the residual value at Cotnari goes also beyond the specified limit. Thus, 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 (Odobesti) displays residues greater than the +2.5 RMSE limit. Furthermore, if we chose to eliminate Odobesti station, we obtain another high residual value for Voinesti station, situated in the same area of orographic enhancement of precipitations as Barnova station, only at a lower altitude.

So far we have established that we have some poor estimated points in our sample, displaying high residual values. Thus, we are certain that we have some points acting like the first type of outliers (referring to the above classification). The problem now is to decide whether it is necessary to eliminate them from the regression model that is, if this elimination would significantly improve 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.

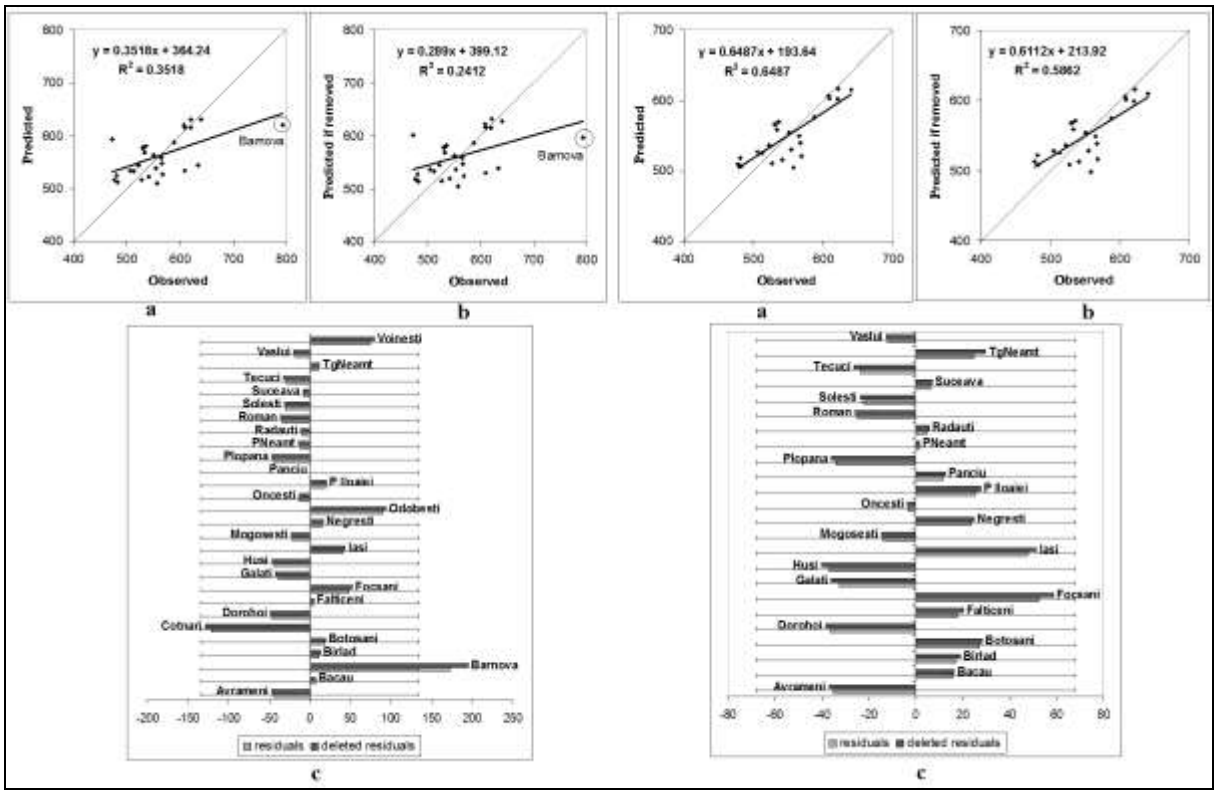

Fig. 5. Correlation between observed and predicted mean annual precipitation (a), cross-validation (b), and comparison of residuals vs. deleted residuals with bars showing the ± 2.5 RMSE (c), using all stations (left) and without four possible outliers (right).

One way to establish the answer is to perform cross-validation, 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 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 of the four possible outliers (*Fig. 5a,b*). The slight difference is hampering us so far to state that the removal of the 4 stations significantly improves the regression models.

The comparison between the observed vs. predicted values and the cross-validation charts tells us only something about the stability of the regression models. In order to investigate the influence of particular values, we may find it useful to compare the regression residuals with those obtained by eliminating the suspect point (deleted residuals, 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 deleted residuals is the greatest in the case of Barnova (22.5 mm), which means that its exclusion from the model significantly changes the altitude – precipitation relationship (*Fig. 5c*). The next greatest difference can be found in the case of Cotnari station (7.8 mm). Even if this is not such an important difference, keeping Cotnari station without Barnova station generates an even poorer regression model than the one using all stations. This is due to the fact, that these two points, one above, the other below 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 Barnova 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: Odobesti and Voinesti stations, the latter being situated within the same area of orographic enhancement of precipitations as Barnova station. However, the difference between the actual and deleted residuals is not very significant. The elimination of all these 4 stations leads to a regression model, where no more points display residuals outside the ±2.5 RMSE interval (*Fig. 5c*, right).

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. It is necessary for us to assess if the altitude – precipitation relationship is significantly changing. As we stated before, the regression model without Barnova 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 485.6 mm to 498.9 mm and gradients from 30.1 mm /100 m to 36.2 mm /100 m.

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

Regression model	**Intercept**	**Regression coefficient**	**R^2**	**Standard error of estimates**
All stations	489.21	0.362	0.352	54.472
Without Barnova	501.82	0.265	0.321	41.678
Without Barnova, Cotnari	498.90	0.301	0.450	36.190
Without Barnova, Cotnari, Odobesti	492.72	0.315	0.547	31.697
Barnova, Cotnari, Odobesti, Voinesti	485.64	0.335	0.649	27.626

From *Fig. 6* we may see that 31% of the station sample displays the lowest residuals under the 2nd model (without Barnova and Cotnari stations). A similar percent (30%) is found for the 4th model (without all of the 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.

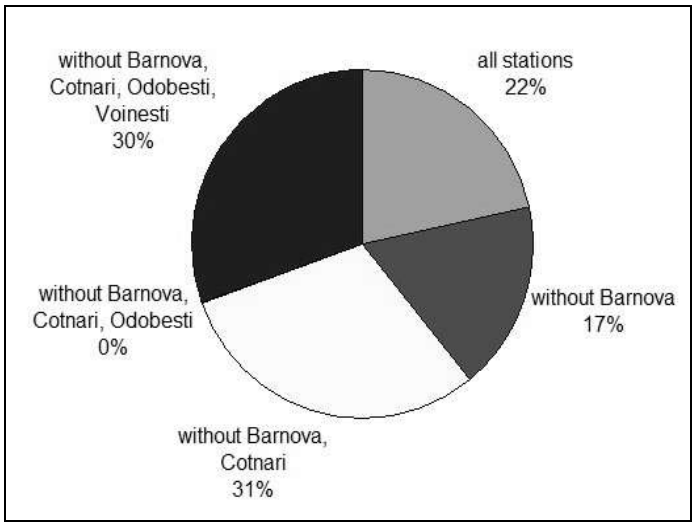

Fig. 6. The optimum altitude regression model (lowest values of actual residuals minus deleted residuals) for each station.

The problem is that we can not 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.

Some of the possible solutions could be:

- data transformation (logarithms);
- derivation of new predictors to account for spatial anomalies;
- application of robust regression methods (*Wilcox*, 2002);
- application of regression as a local interpolator (e.g., geographically weighted regression method);
- application of residual kriging.

A common solution is to derive one or more predictors (*Lhotellier* and *Patriche*, 2007) capable to explain the anomaly associated to the outlier point (e.g., the west-east aspect component combined with terrain local altitudinal range could theoretically explain the precipitations anomaly identified at Cotnari, Barnova, and Voinesti stations from the previous example). Practically, we are often hampered in our analysis by the poor spatial representativeness of the stations network, especially when we have to work with small stations samples, which is, in most cases, unable to fully account for all terrain aspects relevant for the spatial distribution of the analyzed climatic parameter.

The application of residual kriging is also a common approach (*Lhotellier*, 2005; *Dobesch et al*., 2007; *Hengl*, 2007; *Silva et al.*, 2007). Thus, what regression is unable to explain (the residuals), is interpolated using ordinary kriging, then the spatial trend, derived by regression, is added to the spatial anomalies, resulting in the final spatial model of the climate parameter. The output of this approach is still influenced by the quality of the regression model. If the model is significantly influenced by the outliers, then we can not attempt to interpret the predictors-predictand relations.

An alternative solution could be the elaboration of the regression model without the values identified as outliers, the spatialization 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 spatialization. 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 (*Fig. 7*).

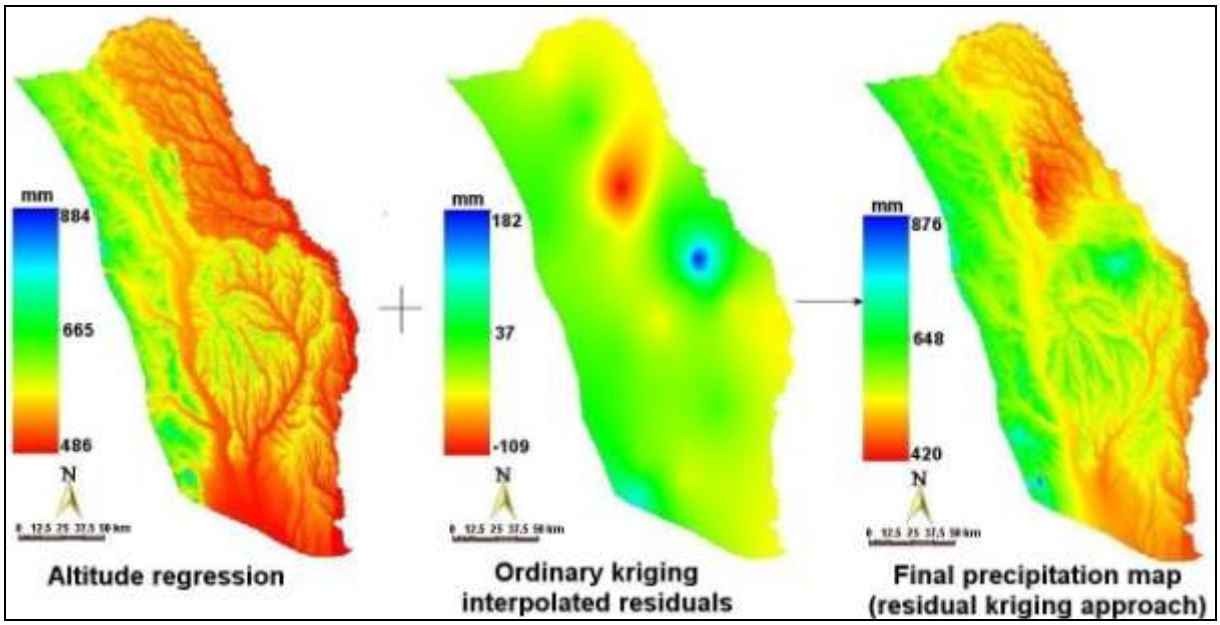

Fig. 7. Mapping the optimum solution: residual kriging approach leaving out the outliers during the regression stage, but taking the outliers' residuals into account during the kriging stage.

A better approach consists in the application of regression as local interpolator (e.g., geographically weighted regression, *Fotheringham et al.*, 2002), taking into account the spatial anomalies (*Engen-Skaugen* and *Tveito*, 2007; *Maracchi et al.*, 2007). The local regression can be further included into a residual kriging approach in order to improve the quality of the output. The main drawback to this approach is the need of a sufficiently large stations sample in order to be capable to derive local regression models.

Let us now see a situation, in which the outliers may indicate possible data errors or different recording intervals. The example is extracted from a study attempting to model the spatial distribution of mean annual precipitations in Vrancea County (Romania) on the basis of 34 rain gauges (*Patriche et al.*, 2008).

Figs. 8 and *9* show 2 points situated significantly outside the altitude – precipitation correlation cloud, namely Pufesti (686.9 mm) and Slobozia Bradului (378.9 mm), therefore, indicating the presence of two possible outliers. In the case of Pufesti rain gauge, the mean annual precipitation regime is characterized by a secondary maximum in August. Taking into account, that all other rain gauges display a single maximum in June, we are inclined to believe that either the August data is incorrect or the Pufesti data represent a shorter time frame, corresponding to a more humid period. On the other hand, the mean annual value recorded at Slobozia Bradului rain gauge is obviously too small for the climatic conditions of our region. Because the monthly values display a normal annual distribution, we are inclined to believe, as before, that the data correspond to a shorter time frame from a drier period.

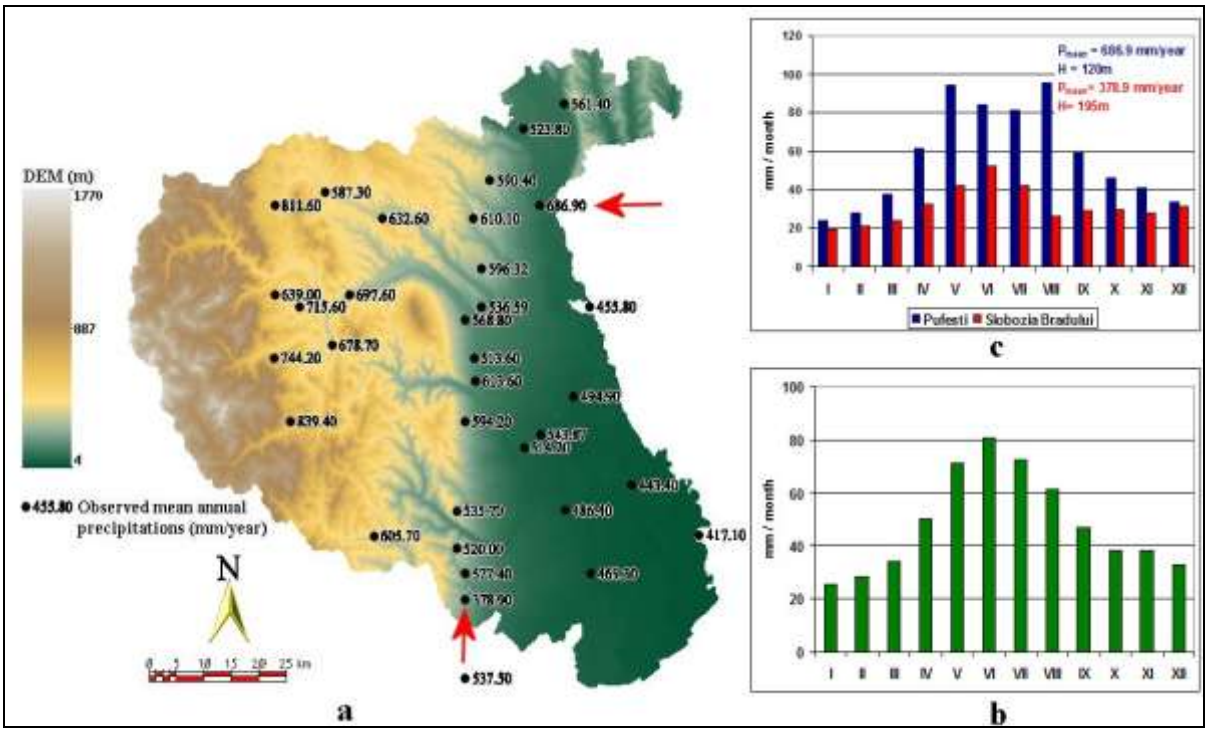

Fig. 8. Observed mean annual precipitations in Vrancea County, Romania (a), mean annual precipitations regime for all stations (b), and for the two suspect points (c).

From *Fig. 9b*, we notice that even though these two points are associated with the highest residuals, the difference between the actual and deleted residuals (jackknife error) is small, meaning that their removal from analysis does not significantly change the altitude regression model. This is happening because the points are situated on opposite sides as compared to the regression

line (*Fig. 9a*) and, therefore, they have opposite effects, balancing the regression line. Their removal increases the correlation coefficient but does not significantly change the direction of the regression line, meaning that the regression equations are very similar with or without these points. This can also be grasped, if one notices that the altitude – precipitation correlation coefficient (0.66) is quite similar with the cross-validation correlation coefficient (0.62), meaning that the one by one removal of all sample points does not significantly change the altitude – precipitation relationship (*Fig. 9c*).

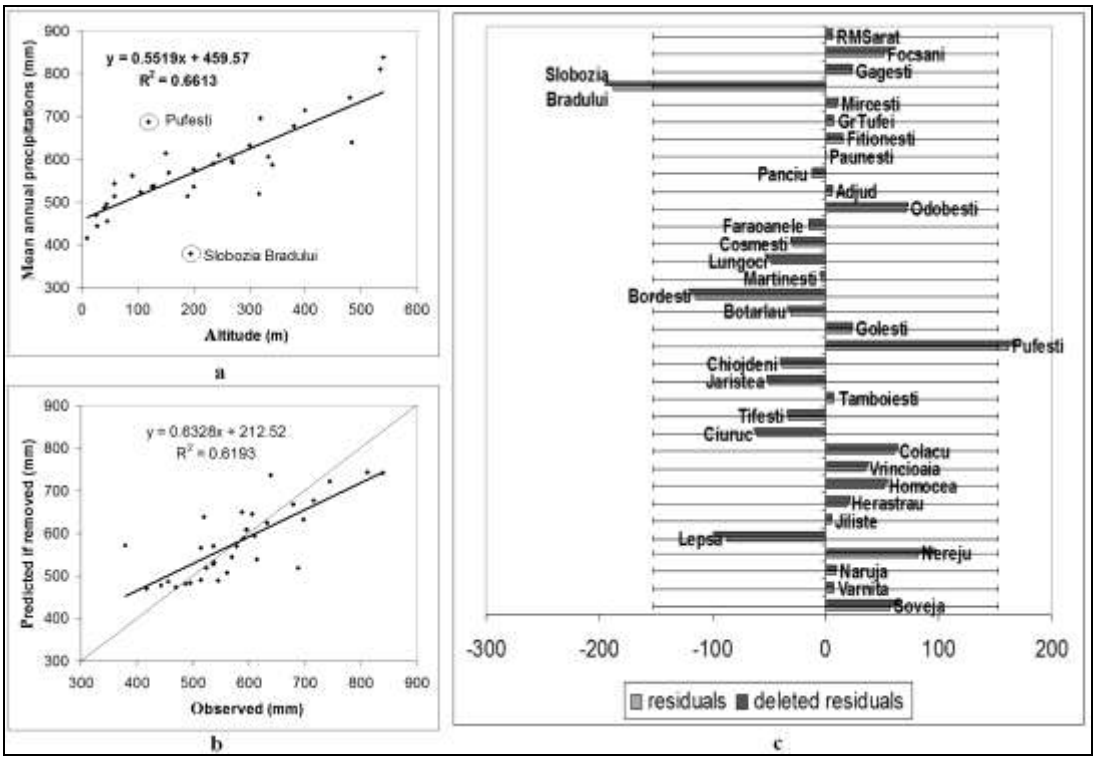

Fig. 9. The altitude – mean annual precipitation relationship (a) and the comparison between actual and deleted residuals (c) showing the presence of two possible outliers. Cross-validation of the altitude model using all stations (b).

Let us see the effects on other predictors. We must mention that, apart from altitude, we also used latitude and longitude as predictors, and at first we obtained a good regression model using both altitude and latitude. Looking further into details, we noticed that the latitude – precipitation correlation is a false correlation, induced by the presence of the two outliers (*Fig. 10*), one with a higher precipitation value situated in the northern part of our region (Pufesti), the other one with a lower precipitation value situated in the south (Slobozia Bradului). If one eliminates these two points, the latitudinal correlation is no longer statistically significant.

For this reason and because of our intention of using also kriging for spatialization, in which case the great residual values of the two suspect points would be represented on the map, we decided to eliminate them from analysis.

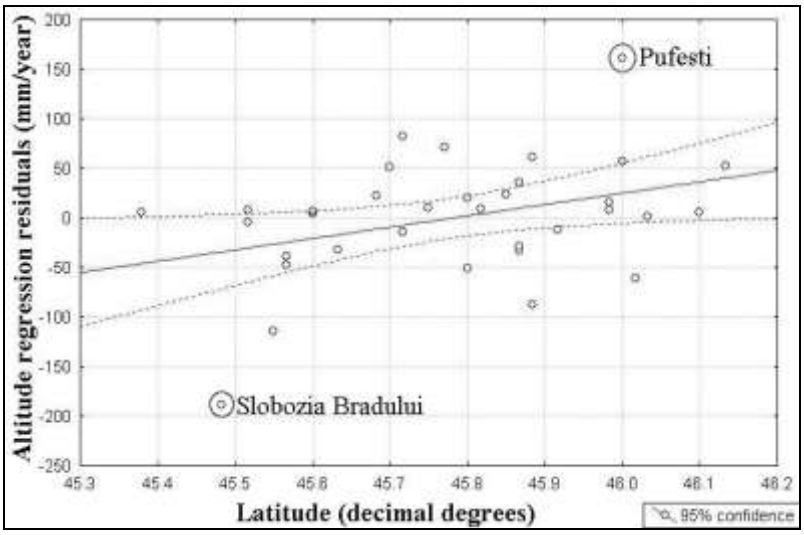

Fig. 10. An unwanted effect of outliers: false precipitation – latitude correlation.

5. Error propagation

Statistically based spatial models are usually computed for elementary variables, such as temperature or precipitations. In order to describe the climate of a region, we also need to compute complex variables, derived from the elementary ones, such as the de Martonne index, potential and real evapotraspiration, etc.

Spatial models of complex variables may be achieved either by computing the complex variable at stations' locations and then interpolating the results or by integrating the spatial models of the elementary variables in order to obtain the complex one. Using the first approach, we are able to quickly compute the errors as well. In this case, we cannot speak of error propagation. Still, in our opinion, this approach is conceptually less appropriate, because the computation of the complex variable is deterministic, according to a physical model. For instance, computing the potential evapotranspiration according to Penman-Monteith approach involves the computation of the net shortwave radiation, which depends on terrain slopes and expositions and on land use. If one computes this parameter at stations' locations and then interpolates the results, neither of these control factors will be taken into account.

The second approach, namely the integration of elementary variables, each of them displaying certain errors, has the disadvantage of inducing invariably in the propagation of these errors to the derived, complex variable. Knowing these errors is important for the assessment of the accuracy of the derived variable's spatial distribution.

A simple example is presented in *Table 2*. The example refers to the derivation of the de Martonne aridity index, for the territory of Moldavia (eastern Romania), on the basis of the mean annual temperatures and precipitations statistically modeled by regression. The mean annual temperature

model uses altitude and latitude as predictors, the computed standard error of estimate is $\pm 0.215\,^{o}$C, meaning that the real temperature differs from the estimated one with $\pm 0.215\,^{o}$C in about 68% of the cases. If we consider, for exemplification, an estimated mean annual temperature of $10\,^{o}$C, then the real temperature will most probably be found within the interval of $9.8-10.2\,^{o}$C. On the other hand, the mean annual precipitation model uses altitude as predictor and has a standard error of estimate of ± 54.472 mm/year, which means that, for an estimated value of 500 mm, the real precipitation values will most probably lie within the interval of 445–554 mm/year. Considering the two estimated temperature ($10\,^{o}$C) and precipitation (500 mm/year) values, it results an aridity index of 25. Taking into account the possible errors for the estimated input parameters, it results that the real value of the aridity index will be most likely found between 22 and 28.

Table 2. Exemplification of error propagation

Statistical parameters		**Mean annual precipitation**	**Mean annual temperature**	**Aridity index**
		Real values		
Exemplification values		500	10	25
Mean		561.95	8.90	29.21
Standard deviation		66.395	0.734	3.136
Standard error		54.472	0.215	–
Confidence interval	Lower limit	445.528	9.785	22.039
	Upper limit	554.472	10.215	28.025
	Range	108.944	0.431	5.986
		Standardized values		
Standardized standard error		0.820	0.294	–
Confidence interval	Lower limit	–1.754	1.206	–2.287
	Upper limit	–0.113	1.793	–0.378
	Range	1.641	0.587	1.909

We are, however, unable to compare these confidence intervals, because the 3 parameters are expressed in different measurement units. One solution is to compute the regression models using the standardized values of the input parameters. We find that the size of the confidence intervals is 1.641 for mean annual precipitation and 0.587 for mean annual temperature. The resulting aridity index distribution for the stations sample is characterized by a mean value of 29.21 and a standard deviation of 3.136. Therefore, the lower limit of the confidence interval (22) corresponds to a standardized value of –2.287 and the upper limit (28) corresponds to a standardized value of -0.378, resulting a range of 1.909. This value is greater than the ones of the input parameters, indicating the propagation and enhancement of the errors, from the elementary variables to the derived, complex variables.

6. Homogeneous vs. heterogeneous regions

Another issue we address in our study is that of heterogeneous regions. Generally, the greater a region, the more heterogeneous it is. A certain level of heterogeneity is necessary for the spatialization of climate parameters. For instance, within a small region, in which the altitudinal range does not exceed, for example, 100 –200 m, the spatial variation of the climate fields may be too feeble for us to correctly infer the spatial variation rules. On the other hand, within a large region, the climatic heterogeneity may be too high for a single statistical model to explain it.

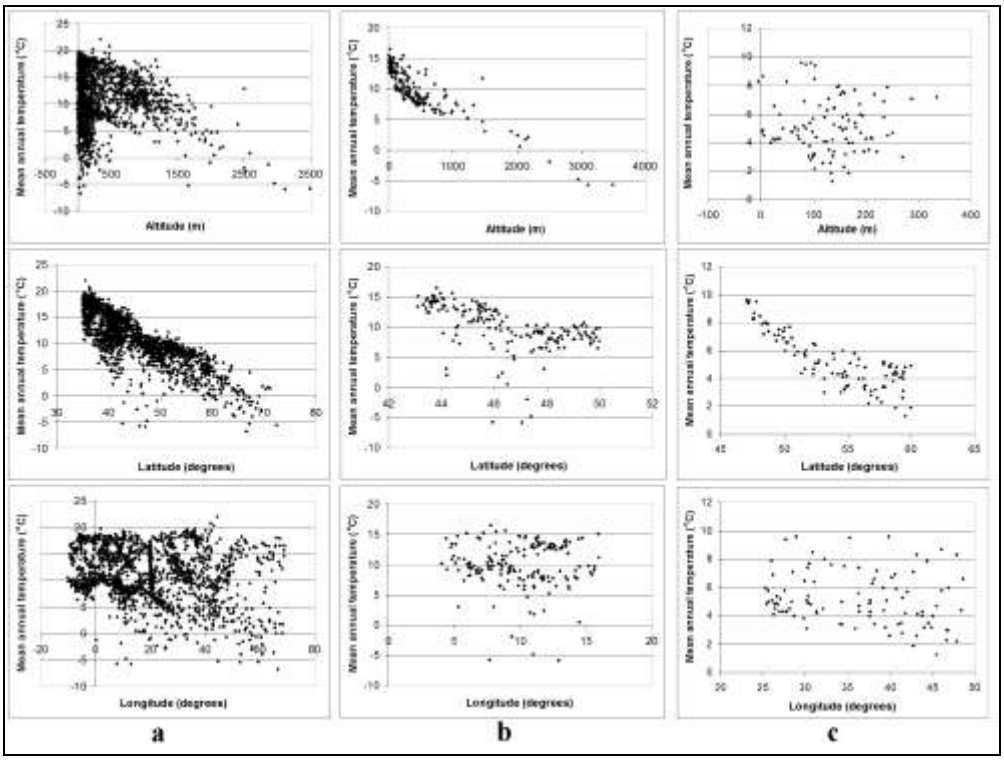

Fig. 11. Changes of the relationships between the mean annual temperatures and the altitude, latitude, and longitude for Europe (a) and for two different subregions: the Alps (b) and the Russian Plain (c). Source of data: *FAO*, 2003.

An example is shown in *Fig. 11* for the relationship between the mean annual temperature (*FAO*, 2003) and 3 predictors: altitude, latitude, and longitude. At continental scale, the territory of Europe is very heterogeneous. We may notice, that the altitude – temperature relationship changes form one region to another to such an extent that a single regression equation for the whole European territory cannot be constructed. A region like the Alps displays a very good altitude – temperature correlation, while the temperature variation within the flat relief of the Russian Plain is statistically independent of the altitude, as temperature inversions are frequent. Here, the latitude comes forward to explain a good part of the temperature spatial distribution.

In such situations, when we deal with large heterogeneous regions, it becomes necessary to divide it into smaller, more homogeneous sub-regions, for which the predictors-predictand relationships do not change. A possible approach could consist in the examination of regression parameters and residuals as we extend or reduce the area of our region and establish the sub-regions limits according to the most stable regression model (maximum correlation, minimum residuals). Another possible approach could be the application of regression as a local interpolator.

7. Conclusions

When applying statistical methods for deriving digital spatial models of climatic variables, one must take great care in identifying and assessing the sources of uncertainty, especially in the case of small stations samples. There are many such sources of different nature, which can easily mislead us towards wrong unrealistic conclusions. Consequently, a good knowledge of data quality, statistical methods, and, needless to say, climatology is imperative for the achievement of sound results. Although simple, the georeference stage is very important. The misplacement of one or more meteorological stations on the map may generate an unwanted chain of errors, because the predictors' values are automatically drawn from the raster maps in GIS environment. The representativeness of the stations network is another important issue, which needs to be analyzed in a preliminary stage of climate parameters spatialization. Theoretically, the spatial distribution of the stations network should be in agreement with terrain complexity, so as to be able to account for all climatic aspects. The extrapolation problem is tightly related to this issue. Unfortunately, in most cases, the stations network is biased, therefore, not sufficiently representative for the terrain. The extrapolation of the spatial models is correct as far as the predictors-predictand relationships do not significantly change outside the calibration area. The outliers problem, meaning the problem of values evading a certain spatial variation rule, is another aspect we analyzed in our study. This is another aspect of the representativeness of the stations network in respect to predictors, which needs to be preliminary addressed in order to minimize the potential errors. Statistical modeling is generally performed on simple, elementary variables, such as temperature or precipitation. For a more thorough investigation of a region's climate, we need to dispose of complex variables, derived from the elementary ones, such as the de Martonne aridity index, potential evapotranspiration, etc. The integration of elementary variables, each having its own statistical errors, into complex variables leads to error propagation. Knowing these errors is very important in order to assess the accuracy of the modeled spatial distribution of the complex variable. Another issue we address in our study is that of the heterogeneous regions. Generally, the greater a region, the

more heterogeneous it is. A certain level of heterogeneity is necessary for the spatialization of climate parameters. On the other hand, within a large region, the climatic heterogeneity may be too high for a single statistical model to explain it. In such a situation, it becomes necessary to divide our large region into smaller, more homogeneous sub-regions, for which the predictors-predictand relationships do not change.

Acknowledgment — This study was carried out with support from project POSDRU/89/1.5/S/49944, coordinated by "Alexandru Ioan Cuza" University of Iaşi (Romania).

References

Dobesch, H., Dumolard, P., and *Dyras, I.* (eds.), 2007: *Spatial Interpolation for Climate Data. The Use of GIS in Climatology and Meterology*. Geographical Information Systems Series, ISTE, London and Newport Beach.

Engen-Skaugen, T., and *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* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, EUR 22596, 93-100.

FAO, 2003: *FAOCLIM-2. World-wide agroclimatic database v.2.02*, FAO/SDRN.

Fotheringham, S., Brunsdon, C., and *Charlton, M.*, 2002: *Geographically Weighted Regression. The Analysis of Spatially Varying Relationships*. Wiley.

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.

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

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

Maimon, O.Z., Rokach, L. (eds.), 2005: *Data Mining and Knowledge Discovery Handbook*, Chapter 7. Outlier detection, Ben-Gal, I. Springer.

Maracchi, G., Ferrari, R., Magno, R., Bottai, L., Crisci, A., and *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* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, 2007, EUR 22596, 9-16.

Patriche, C.V., 2007: About the influence of space scale on the spatialisation of meteo-climatic variables. *Geographia Technica*, no. 1, Cluj University Press, 70-76.

Patriche, C.V., Sfică, L., and *Roşca, B.,* 2008: About the problem of digital precipitations mapping using (geo)statistical methods in GIS. *Geographia Technica*, no. 1, Cluj University Press, 82-91.

Silva, Á.,P., Sousa, A.J., and *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* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, EUR 22596, 37-44.

Wilcox, R., 2002: *Applying Contemporary Statistical Techniques*. Academic Press.