Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 52-64, 1998

Modeling of the Daily Rainfall Values Using Surface Under Tension and Kriging

Anatoly A. Saveliev
Faculty of Ecology, Kazan State University, 18 Kremlevskaja Street, Kazan, Tatarstan 420008, Russia
Anatoly.Saveliev@ksu.ru

Svetlana S. Mucharamova
Faculty of Ecology, Kazan State University, 18 Kremlevskaja Street, Kazan, Tatarstan 420008, Russia

Gennady A. Piliugin
Faculty of Ecology, Kazan State University, 18 Kremlevskaja Street, Kazan, Tatarstan 420008, Russia

ABSTRACT **Various estimation methods are discussed in respect of rainfall estimation. The usefulness of two methods was demonstrated by comparison of estimated daily rainfall values in unknown locations with the results of real observations in the same locations. The accuracy of prediction as well as applied properties of the estimations and methods are discussed.**

KEYWORDS: **surface under tension, variogram, kriging, rainfall estimation.**

Contents

1. Problem Description

The estimation of rainfall level in arbitrary locations based on the results of meteorological observations is of significant importance for environmental sciences. It could be of prime interest for the surface hydrology, modeling of mass transport, the investigations of erosion processes, pollution precipitation, etc. Apart from long-term averaged rainfall levels often used for the analysis of pollution precipitation, more precise forecasts of rainfall level are required, especially when some accident consequences are considered. In this case, besides the requirements to the average accuracy of the method used, it is important to achieve a good accuracy for the largest data values which highly affect the polluted patches on the surface. Different methods of calculation are appropriate for long-term averaged estimation and for short-term exact estimation. In this paper various estimation methods were compared and two of them were selected as more convenient for short-term exact estimation. The usefulness of methods was demonstrated by comparison of estimated daily rainfall values in unknown locations with the results of real observations in the same locations. The accuracy of prediction as well as applied properties of the estimations and methods used are discussed.

2. Source Data

This report is devoted to the application of various methods for the estimation of rainfall values. The primary set of data includes 100 daily rainfall records made in Switzerland on May, 8th, 1986. The collection of the data was carried out by the Air Pollution Group at Imperial College, London, under financial support of JRC-Ispra. These data were supplemented with the 367 other location points where rainfall values were to be estimated and true observed values were also known. Digital Elevation Model (DEM) and the map of Switzerland Country were also used for data processing. The area to be characterized spans approximately the square 360 by 250 km and is shown in Figure 1. Filled points refer to the 100 initial observed data while, non-filled ones refer to the other 367 true observed data. Histograms for the initial 100 observed data and for the full data set are shown in Figures 2 and 3 below.

Figure 1 The distribution of the locations of the measurement and estimation of rainfall values and rough separation into zones. Symbols scaled according to observed values.

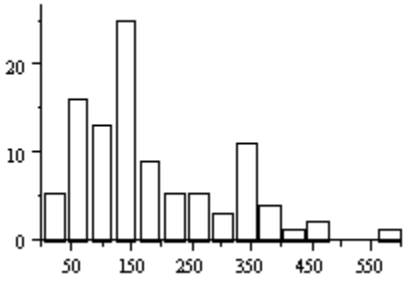

Figure 2 100 Initial Values Histogram

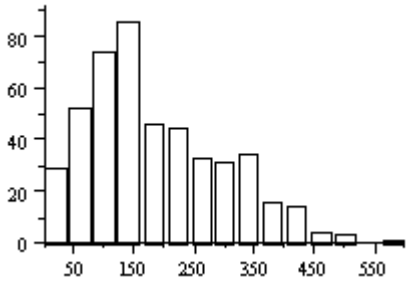

Figure 3 All Observed Values Histogram

3. Exploratory Data Analysis

The initial 100 data of rainfall values were plotted on the map in a manner such that the size of each point was proportional to the appropriate rainfall value (Figure 1). After preliminary analysis of the map, three different zones were differentiated on the bases of the specificity of relief structure with different mean rainfall values. The long axes of the zones proposed are oriented south-west - north-east. The zones were roughly delineated to avoid the occurrence of "ill-formed" pairs from different populations in variogram analysis. It should be also accounted in the following modeling.

As seen in Figure 2, 100 initial observed values could be considered as a combination of at least two populations with means of approximately 125 and 340, respectively. Calculated means of the rainfall values are 109 for the middle zone, and 253 and 187 for the upper and bottom zones respectively.

Univariate statistics of the observed rainfall values are shown in Table. Spatial distribution of the values presented as directional semivariograms for the directions 0, 45 and 90 degrees are shown in Figures 4 and 6. The character '+' refers to the construction of experimental semivariogram before eliminating "ill-formed" pairs and character 'o' to that after the elimination of mentioned pairs.

VarioWin 2.1 package, developed by Yvan Pannatier, Institute of Mineralogy and Petrography, University of Lausanne, Switzerland, was used for the analysis of the spatial distribution of the data.

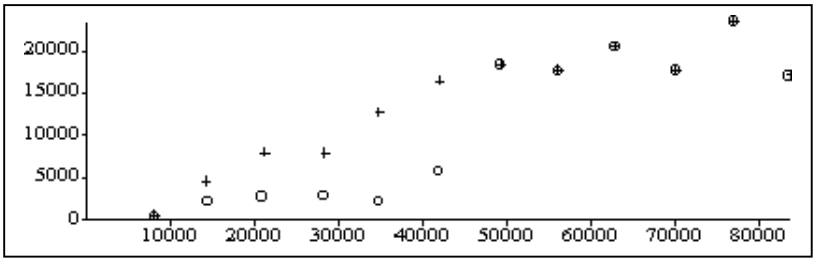

Figure 4 Semivariogram for Direction 0

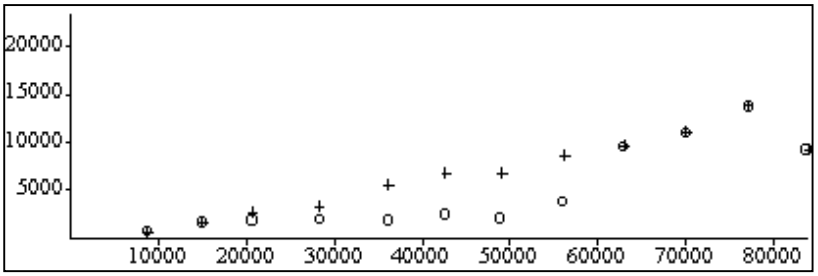

Figure 5 Semivariogram for Direction 45

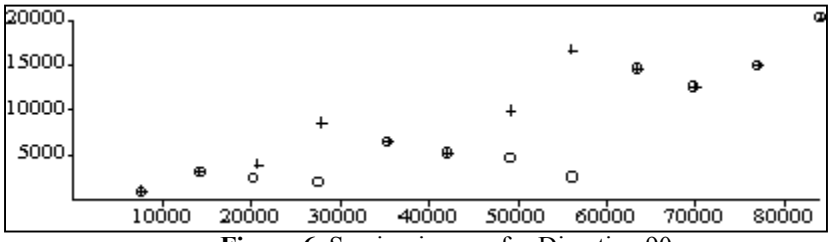

Figure 6 Semivariogram for Direction 90

The statistical analysis of the data should also make clear if any relationship between DEM and rainfall values existed. For this purpose, certain of additional DEM properties were derived at the source data locations (e.g. smoothen DEM elevation, **log(**elevation**)**, slope, **log(**slope**)** and the aspect). After that, the correlation, rank correlation and linear regression of rainfall values and these parameters were estimated. As was established, only one statistically significant non-zero correlation takes place (between the rainfall values and X-coordinate). All the coefficients in linear regression are considered insignificant. It may be concluded that the application of DEM properties

as additional variables for co-kriging or for linear trend estimation cannot result in an essential improvement of modeling results.

4. The Selection of Modeling Method

The following peculiarities of the problem to be solved will affect the selection of the appropriate modeling method:

- The daily rainfall values cannot have random distribution or be a random function realization. These values depend substantially on the relief and weather conditions.
- Daily rainfall value does not refer to the averaged values, which are most often processed in the methods of geostatistics. Thus, kriging method, most typical for the geostatistics, does not seem to be appropriate in this particular case.

From these considerations the following interpolation methods were chosen for the application: (1) C-1 "thin plate" interpolation based on the Delaunay triangulation and on the construction of a surface under tension, and (2) kriging with linear semivariogram model.

The original idea included the use of an artificial neural network for the estimation of the spatial correlation using training patterns, consisted of the rainfall location and direction and distance from this location to another one. The estimation obtained in this way was then assumed to use in kriging at the locations where the rainfall values were estimated. A similar approach could provide satisfactory results even though a large amount of very noisy data would processed. However, for a modest number of input data with minor noise as in our case more usual methods offer advantages.

5. Surface under Tension Estimation

The method selected for the "thin plate" interpolation is based on the exact interpolation algorithm from SRFPACK (Renka, 1996) and using Delaunay triangulation for the TIN construction. This algorithm makes it possible to estimate normal gradient at each node and use Hermite interpolatory tension spline. Along the triangle arcs, the interpolatory function is the Hermite interpolatory tension spline defined by the values and tangential gradient components at the nodes, and the derivative in the normal direction to the arc varies linearly between the normal gradient components at the endpoints. A first-order C-1 blending method is used to extend the function to the interior of the triangle. Thus, since values and gradients on an arc depend only on the vertex data, the method results in C-1 continuity when used for interpolation over the triangulation. As usual for triangulation algorithms there is a problem of extrapolation. These algorithms show sufficient results only inside the convex hull of all known locations. To avoid this problem, the buffer was created along the border of Switzerland at 15 km distance, and new locations were added with 15 km spacing along the buffer line. Rainfall values at new locations were estimated using the kriging method described below. Additional information, derived from the relief structure and expert knowledge could be used for taking in account the zone anisotropy by forcing edges in triangulation along the zones borders, but it was not done.

6. Variogram Modeling

VarioWin 2.1 package was used for the modeling semivariograms. To take into account the zone anisotropy, all the source locations were splitted into three subsets in accordance with the delineation of the zones described above, and univariate statistics for each zone were calculated. As shown, the univariate distribution can be considered as normal. Other tests were not performed because of the scarce population of data subsets (approximately 30 locations in each zone). After eliminating some of the most "ill-formed" pairs from different populations, the full data set was used for semivariogram estimation. Experimental directional semivariograms (before clearing) for three directions are shown above (Figure 4-6). The semivariograms for the same directions after the clearing procedure are also shown in Figure 4-6. To take into account the zone anisotropy, the directional anisotropy was used.

We assume that the phenomenon to estimate is "smooth", (i.e., daily rainfall values change gradually with the distance). In this respect the linear semivariogram without nugget effect was selected as the base model for calculations. This sort of semivariograms provides continuous and smoothen estimation near the known location, whereas semivariogram models with nugget effect give, in some cases, estimations which differ significantly from the known value even at short distances. The estimated semivariogram parameters were as follows:

Gamma (h) = 0.0723 * h

Direction: 55.4 degree, Anisotropy: 0.53, Search Radius: 50 000m

The same semivariogram and search parameters were used for estimation of the values along the buffer line in the method of surface under tension.

7. Kriging Estimation

The GSLIB package (Deutsch and Journel, 1992) was used to implement the kriging and the novel driver program based on the KTB3D has been developed to work with the evaluation locations in simplified GeoEAS format. For the data search the isotropic search algorithm was selected with search radius from 50 000m to 70 000m, and maximum number of data values for kriging was limited to 16. Three implementations of kriging were performed: the first one for the whole area with the semivariogram described above (search radius 50 000 m); the second one with separate kriging for each zone with the same semivariogram and search radius 70000m, and the last one with separate kriging for each zone with the same semivariogram without anisotropy and search radius 70 000m. The results of the calculations are shown below:

Gamma (h) = 0.0723 * h

Direction: 0 degree, Anisotropy: 1, Search Range: 70 000m

Certain changes in the search radius in the last two implementations were introduced to avoid non-estimated locations.

8. Estimation Results

The results of estimation are marked as $\mathbf{Z_1}$ (Surface under Tension Estimation), $\mathbf{Z_2}$ (Kriging Estimation), $\mathbf{Z_3}$ (Zone Kriging with Anisotropy Semivariogram Estimation), $\mathbf{Z_4}$ (Zone Kriging with Isotropy Semivariogram Estimation), $\mathbf{Z_0}$ refer to the true values. **SD** denotes standard deviation of values. **RMSE** denote square root of mean squared

residuals ($\mathbf{Z_i}$-$\mathbf{Z_0}$), **Bias** - mean of residuals, **Mean Relative Error (MRE)** - mean of **Relative Error abs(**100*($\mathbf{Z_i}$-$\mathbf{Z_0}$)/$\mathbf{Z_0}$**)** calculated for all $\mathbf{Z_0}$ values above or equal to 10, and **Mean Absolute Error (MAE)** - the mean of **Absolute Errorabs(**$\mathbf{Z_i}$-$\mathbf{Z_0}$**)**. Statistics for $\mathbf{Z_0}$-$\mathbf{Z_4}$ and errors are summarized in the following Table 1.

Table 1 True Values, Estimated Values and Estimation Error Statistics.

	Min	Max	Mean	Median	SD	RMSE	Bias	MRE	MAE
$\mathbf{Z_0}$	0.0	517.0	185.4	162.0	111.2				
$\mathbf{Z_1}$	-40.4	666.7	185.7	165.2	112.0	71.0	0.3	33.2%	47.9
$\mathbf{Z_2}$	15.4	485.7	182.8	164.6	99.3	54.6	-2.5	27.8%	37.5
$\mathbf{Z_3}$	16.0	527.6	179.8	154.0	102.1	61.4	-5.6	29.0%	41.8
$\mathbf{Z_4}$	11.0	548.1	180.7	150.1	101.1	64.0	-4.7	30.1%	44.0

Histograms for $\mathbf{Z_1}$ and $\mathbf{Z_2}$ estimations (including initial 100 observed values) are shown below in Figure 7-8.

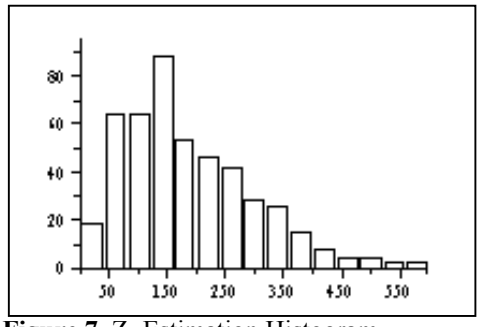

Figure 7 Z_1 Estimation Histogram

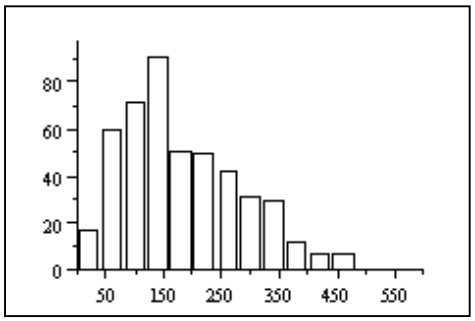

Figure 8 Z_2 Estimation Histogram

Other statistics of errors (**Min**, **Max**, **Mean**, **Median** and **Standard Deviation** (**SD**) for **Residuals** ($\mathbf{Z_i}$-$\mathbf{Z_0}$) and **Relative Residuals** 100*($\mathbf{Z_i}$-$\mathbf{Z_0}$)/$\mathbf{Z_0}$, and **Mean**, **Median** and **Standard Deviation** for **Absolute Error** and **Relative Error**) are presented in the Tables 2-4. **High10** and **Low10** denote subsets of 10 biggest and the 10 smallest true values to be estimated. Relative error was not calculated for smallest 10 values because their magnitudes were compatible with the error of the measurements.

Table 2 Residuals Statistics

Estimation	Residuals Statistic				
	Min	Max	Mean	Median	SD
Z_1	-309.1	306.4	0.3	2.2	71.0
Z_2	-267.9	318.9	-2.5	2.3	54.5
Z_3	-284.5	362.7	-5.6	1.8	61.3
Z_4	-284.1	342.6	-4.7	1.9	64.0
Z_1 for High10	-309.0	61.1	-125.4	-133.3	121.6
Z_2 for High10	-267.9	21.2	-113.9	-107.6	92.2
Z_1 for Low10	9.0	54.6	25.3	23.0	14.2
Z_2 for Low10	10.4	44.7	24.8	23.5	11.5

Table 3 Relative Error Statistics

	Relative Residuals				
Estimation	Min	Max	Mean	Median	SD
$\mathbf{Z_1}$	-352.3%	630.7%	33.2%	0.4%	59.1
$\mathbf{Z_2}$	-64.8%	315.1%	27.8%	0.7%	46.4
$\mathbf{Z_3}$	-68.5%	235.5%	29.0%	0.4%	45.7
$\mathbf{Z_4}$	-68.5%	274.7%	30.1%	0.1%	47.2
$\mathbf{Z_1}$ **for High10**	-59.8%	14.1%	-27.0%	-30.6%	25.3
$\mathbf{Z_2}$ **for High10**	-51.9%	4.9%	-24.4%	-24.5%	18.7

Table 4 Absolute Error and Relative Error Statistics

	Absolute Error			Relative Error		
Estimation	Mean	Median	SD	Mean	Median	SD
$\mathbf{Z_1}$	47.9	30.8	52.3	33.2%	20.7%	49.7
$\mathbf{Z_2}$	37.5	25.8	39.6	27.8%	15.5%	38.4
$\mathbf{Z_3}$	41.8	30.3	45.2	29.0%	18.1%	36.1
$\mathbf{Z_4}$	44.0	30.5	46.6	30.1%	19.7%	37.3
$\mathbf{Z_1}$ **for High10**	141.6	133.3	99.9	30.7%	30.6%	20.2
$\mathbf{Z_2}$ **for High10**	118.1	107.6	86.1	25.4%	24.5%	17.2
$\mathbf{Z_1}$ **for Low10**	25.3	23.0	14.2	-	-	-
$\mathbf{Z_2}$ **for Low10**	24.8	23.5	11.5	-	-	-

Based on the results presented in Tables 1-4, it can be seen, that the use of zone border ($\mathbf{Z_3}$ and $\mathbf{Z_4}$ estimations) does not contribute to the improvement of the kriging estimation ($\mathbf{Z_2}$), which provides the best results of estimation among the other approaches. Error map analysis has revealed that a bigger error for $\mathbf{Z_3}$ and $\mathbf{Z_4}$ results from the insufficient delineation of zones. The further refinement of the borders can be established with the assistance of the experts in meteorology. Another way is overlapping the zones or the use of the fuzzy membership function. Consequently, only $\mathbf{Z_1}$ and $\mathbf{Z_2}$ estimations are discussed further. Two scatter plots in Figure 9 show the relationship between the real values and estimation results (lines refer to the regression of true value on estimator and vice versa):

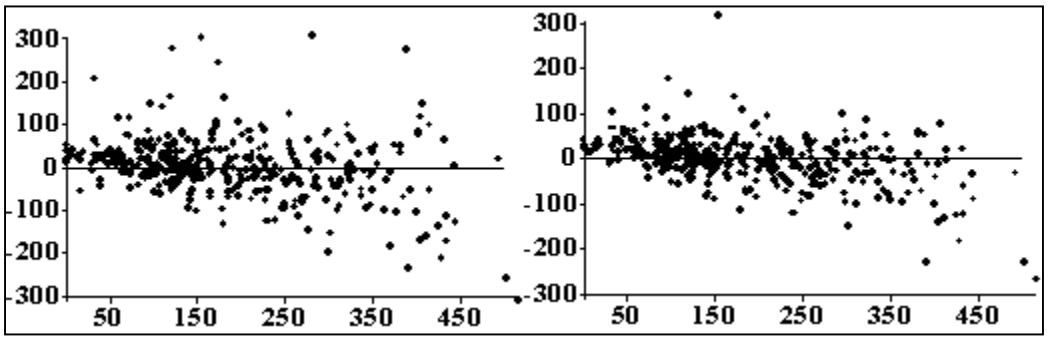

Figure 9 Scatter plots for Surface under Tension and Kriging Estimations

Correlation of absolute error and residuals with observed values was also estimated (see Table 5). The scatter plot of residuals against observed values is shown in Figure 10. These results show that errors are not correlated with the observed values, but the error variance increases with true values. Both methods tend also to underestimate the high values and overestimate small values (negative correlation of residuals and observed values was found).

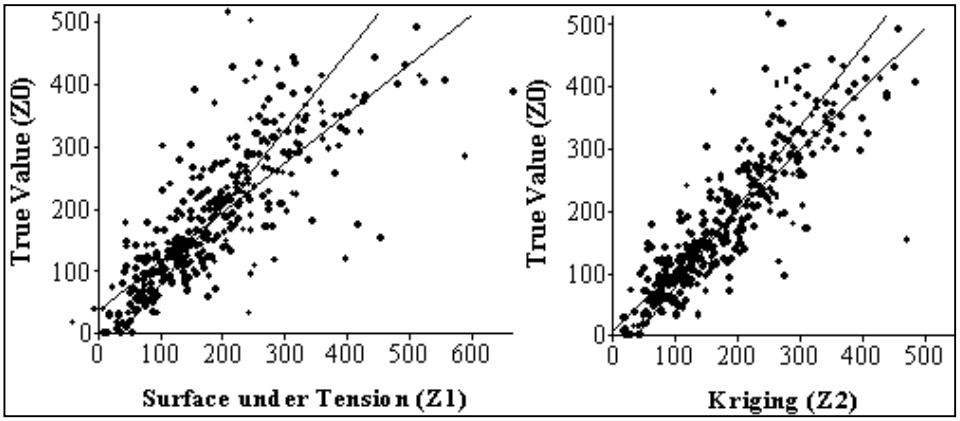

Figure 10 Scatter plot of residuals (in Y), left - Z_1 estimation, right - Z_2 estimation, against observed values (in X).

Table 5 Correlation of Errors

Estimation	Residuals Correlation	Absolute Error Correlation
$\mathbf{Z_1}$	-0.3075	0.3696
$\mathbf{Z_2}$	-0.4519	0.3060

To compare visually the results of estimation with real values of rainfall, one more kriging estimation was performed. It was based on all 467 observed values and the semivariogram and search parameters chosen for $\mathbf{Z_4}$ estimation, but without zone delineation. The map obtained in this procedure is presented in Figure 11. Maps for

$\mathbf{Z_1}$and $\mathbf{Z_2}$ estimations in Figures 12 and Figure 14, respectively, and maps for the residuals for $\mathbf{Z_1}$ and $\mathbf{Z_2}$ estimations in Figure 13 and 15, respectively.

Figure 11 Isoline Map of Kriging Estimation Based on Full Data Set with 10 maximum (squares) and minimum (circles) observed values.

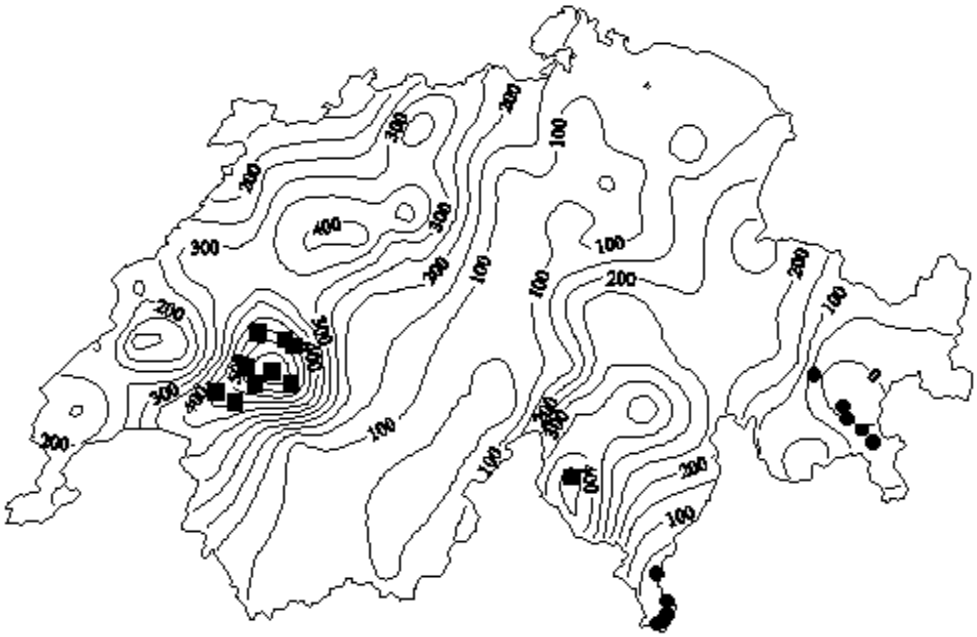

Figure 12 Isoline Map of the Surface under Tension Estimation (Z_1) with 10 maximum (squares) and minimum (circles) estimated values.

Figure 13 Proportional Symbols Map of the Surface under Tension Estimation (Z_1) Residuals (symbols refer to underestimated locations are filled)

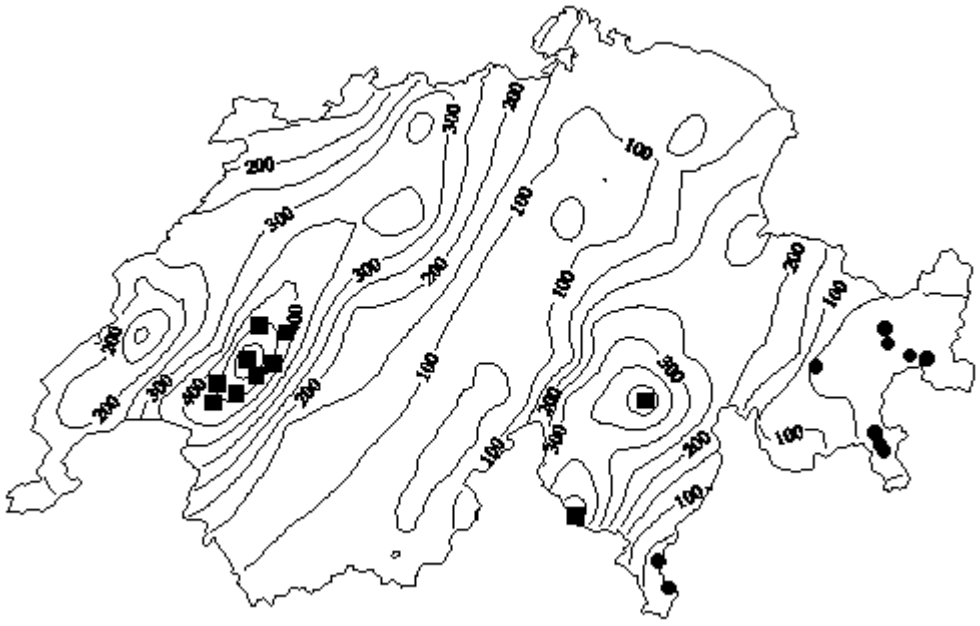

Figure 14 Isoline Map of the Kriging Estimation (Z_2) with 10 maximum (squares) and minimum (circles) estimated values.

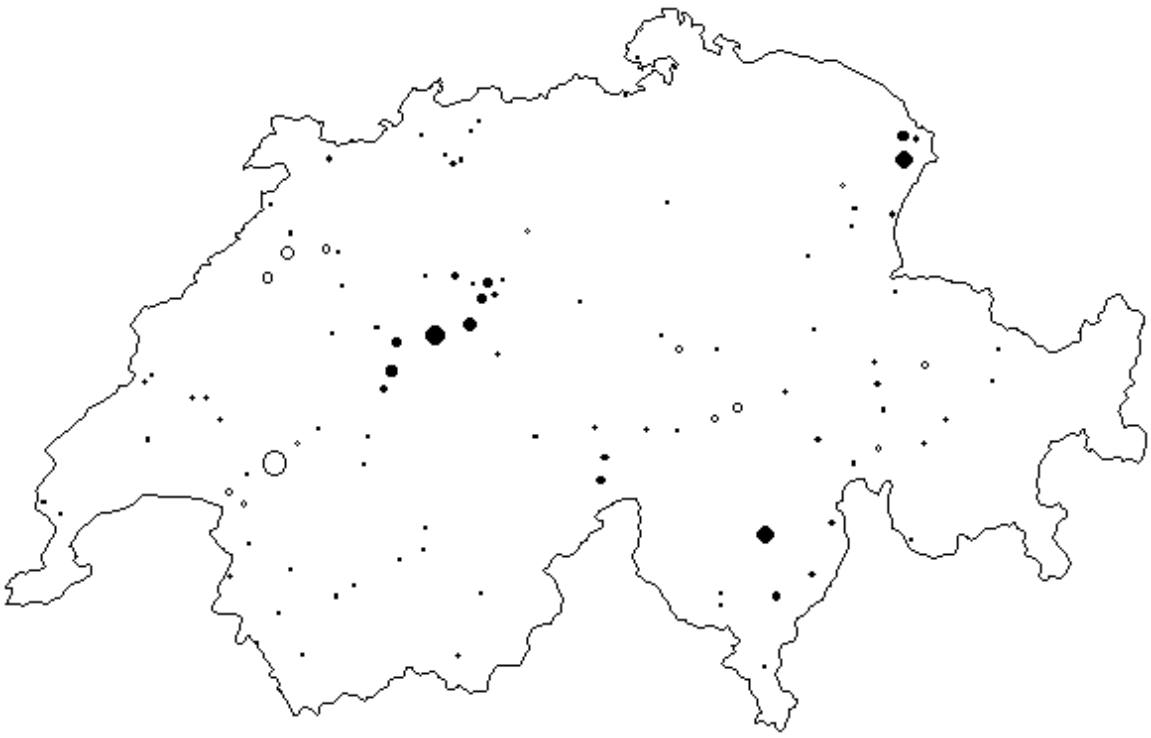

Figure 15 Proportional Symbols Map of the Kriging Estimation (Z_2) Residuals (symbols refer to underestimated locations are filled).

9. Results Discussion

We denote surface under tension estimation ($\mathbf{Z_1}$) as a first method, and kriging estimation ($\mathbf{Z_2}$) as a second method. From Table 1 we could see that the behavior of the methods differs from each other: the magnitude of estimated values obtained by the first method is wider than that of the source values, whereas the second method, as common for kriging methods, is more "conservative" and tends to "average" estimated values. It can be explained by the nature of the methods. For the first method, if the location referring to the rainfall value to be estimated is inside a big triangle with the same trend at its edges (for example, uphill in triangle center direction), the estimated value can be far above the triangle's vertices values. This means that in addition to "ill-formed" triangles, the very big triangles should be excluded from the consideration, and the first method requires some triangulation analysis. To avoid these big triangles, the following procedure can be proposed. After the analysis, the triangulation mesh is modified with new locations, which are added inside all the big or "ill-formed" triangles. The values at these locations can be estimated using kriging procedure. The properties of the first method mentioned above cause more difficulties for its automated implementation. Contrary to that, kriging methods, especially those based on the application of transition semivariogram model,

tend to "shrink" magnitude of the estimated values (when the distances to the known locations are compatible with the semivariogram range). This results in the calculation of the values, which are lower than the maximum and higher than the minimum of source values. If the semivariogram model is estimated, such methods are easier to use in automation mode. Meanwhile, the estimation of the semivariogram model is not assumed as being performed in automated mode and requires human assistance. In our case, the second method gives a big error only for high real values, surrounded by the low values used for estimation, and vice versa.

Both methods showed the same behavior for the biggest 10 and lowest 10 values. The first method found only two maximum values (including one observed), and the second method four values (including three observed). For the lowest values, both methods found five values including one observed. Almost all the biggest values were underestimated, and the lowest values overestimated. Meanwhile the second method was found more accurate for these ranges of values.

For both methods the **Absolute Error** was analyzed. Calculated statistics, except that for kurtosis, could be considered as typical for a normally distributed population. Too big kurtosis could be explained by the high variance, caused by the few sever errors. Spatial distribution of the **Absolute Error** was also analyzed using the VarioWin 2.1 package. No essential anisotropy was established in the spatial distribution for both methods. Also the experimental semivariogram surface obtained for the second method is more smoothen than that of the first one. Semivariogram range was found about 20 000 m for both methods, but the sill for the second one was twofold lower.

References

Deutsch, C.V. and Journel, A. G. (1992) - *GSLIB Geostatistical Software Library and User's Guide*, Oxford: Oxford University Press.

Renka R. J. (1996). Algorithm 752; SRFPACK: software for scattered data fitting with a constrained surface under tension. *ACM Transactions on Mathematical Software*, 22(1); 9-17.
