Analysis of Rainfall Data by Robust Spatial Statistics using S+SPATIAL STATS

M. G. Genton[1] and R. Furrer[2]

[1]*Department of Mathematics*
Massachusetts Institute of Technology
Cambridge
MA 02139-4307
USA

E-mail: genton@math.mit.edu

[2]*Chair of Applied Statistics*
Swiss Federal Institute of Technology
CH-1015 Lausanne
Switzerland

E-mail: reinhard.furrer@epfl.ch

Abstract This paper discusses the use of robust geostatistical methods on a data set of rainfall measurements in Switzerland. The variables are detrended via non-parametric estimation penalized with a smoothing parameter. The optimal trend is computed with a smoothing parameter based on cross-validation. Then, the variogram is estimated by a highly robust estimator of scale. The parametric variogram model is fitted by generalized least squares, thus taking account of the variance-covariance structure of the variogram estimates. Comparison of kriging with the initial measurements is completed and yields interesting results. All these computations are done with the software S+SPATIALSTATS, extended with new functions in S+ that are made available.

Keywords: Robustness; Trend; Variogram; Generalized least squares; Kriging.

1. INTRODUCTION

Statistical methods widely known under the name *kriging* are intended to predict unobserved values of a variable in a spatial domain, on the basis of observed values. These techniques are based on a function which describes the spatial dependence, the so called *variogram*. Therefore, variogram estimation and variogram fitting are important stages of spatial prediction. Because they determine the kriging weights, they must be carried out carefully, otherwise kriging can produce unreliable maps.

In practical situations, a fraction of outliers is often included in observed data. Experience from a broad spectrum of applied sciences shows that measured data contains as a rule between 10 to 15 percent of outlying values due to gross errors, measurement mistakes, faulty recordings, *etc.* One might argue that any reasonable exploratory data analysis would identify and remove outliers in the data. However, this approach is often subjective and outlier rejection is highly opinion dependent. Thus, in this paper, we advocate the use of robust geostatistical methods, which prevent the negative effects of outlying values. Note that the existence of exploratory techniques does not supersede the utility of robust techniques.

The data set contains N=467 measurements of rainfall in Switzerland, from which only n=100 were made available. A complete description of the data set, as well as the location map of the measurements, are presented in an introduction report. However, we would like to point out the skewness in the first histogram of Figure 1, and the possibly bimodality of the distribution of the data. Therefore, a logarithmic transformation is applied on the data, followed by centering and reducing operations. The resulting histogram is visualized in Figure 1.

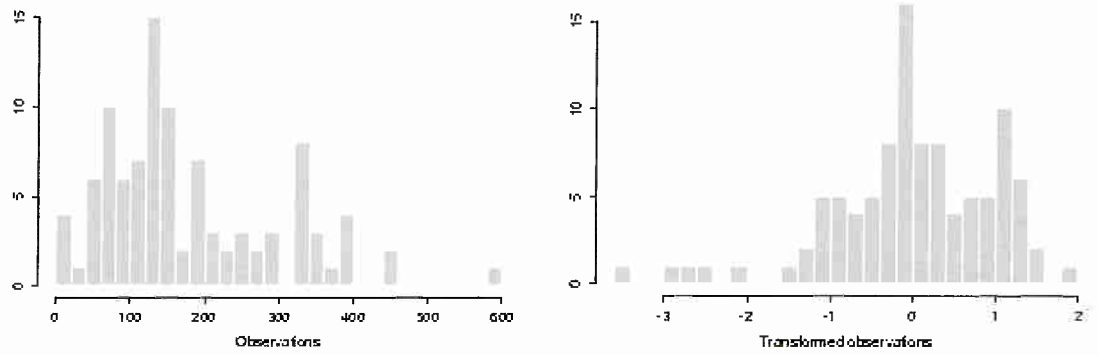

Figure 1: *Histogram of the observations and histogram of the transformed observations (logarithm followed by centering and reducing). Note the skewness and possibly bimodality of the distribution of the data.*

In order to describe the data, the following simple model is used for the rainfall variable

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}), \qquad \mathbf{x} = (x,y)^T, \tag{1}$$

where $m(\mathbf{x})$ is the deterministic part of Z and $\varepsilon(\mathbf{x})$ the stochastic one. Because of the local behavior of the data, the trend $m(\mathbf{x})$ is computed by a non-parametric approach and removed, as shown in the next section. Highly robust variogram estimation is performed in section 3, followed by generalized least squares variogram fitting in section 4. Finally, kriging results are discussed in the last part of this paper.

2. TREND DETECTION

The first step of spatial data analysis consists in detecting the trend of the variables (Cressie, 1991), *i.e.* we determine $m(\mathbf{x})$, the non-stochastic part of (1). Figure 2 shows the trend surfaces of the $n = 100$ rainfall measurements of the data set drawn with the command `symbols` in S+. We note the local behavior of the trend. By comparing the trend with the geographic characteristics of Switzerland, we verify that there is a correspondence between the amount of rainfall and the elevation of the measures (plain and mountains), as well as with particularly sunny counties like Wallis.

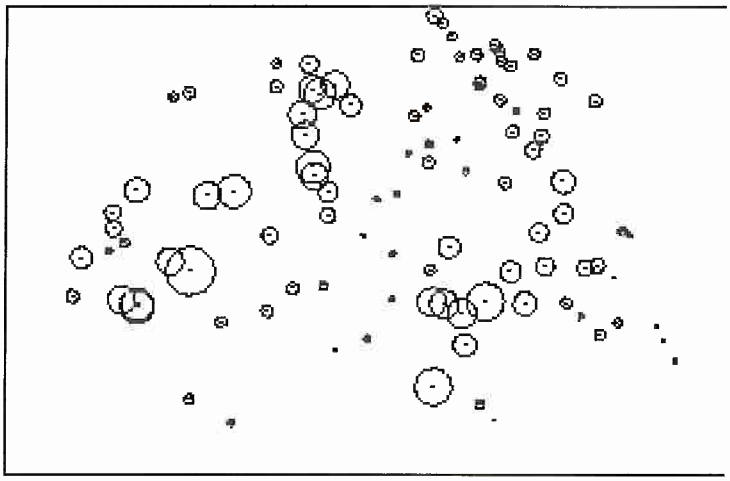

Figure 2: *Trend for the n=100 measurements of the rainfall data set. Bigger circles represent higher amount of rainfall.*

It is not appropriate removing such local drifts by adjusting a polynomial trend surface as shown in Venables and Ripley (1994). A non parametric adjustment is much more sensible to local variations and is therefore more suitable. It is also a simple way of taking account of the elevation of the measurements. The function `loess` fits a local regression model of second degree. To apply a robust fitting, we suppose not a Gaussian but a symmetric distribution of the errors and set therefore the argument `family=symmetric`, as well as `normalize=F`. Full details of `loess` are given by Cleveland *et al.* (1992).

The problem of choosing the smoothing parameter λ is ubiquitous. Often, in geostatistical approaches, exploratory work helps to find a value for a ``good looking" surface. To avoid this ambiguousness, we apply the principle of cross-validation (Green and Silverman, 1994). The basic idea of cross-validation is to choose the smoothing parameter λ which minimizes the criterion

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \hat{g}^{(-i)}(\mathbf{x}_i, \lambda)\right)^2 \qquad (2)$$

where $\hat{g}^{(-i)}(\mathbf{x}_i, \lambda)$ is the nonparametric estimation of $m(\mathbf{x}_i)$ in omitting the observation $\mathbf{x}_i$ from the data set and with smoothing parameter λ . This criterion ensures stability of the fitted surface. In this work, the predictor $\hat{g}(\mathbf{x}, \lambda)$ is the function `loess`. In general, the function $\boldsymbol{CV}(\lambda)$ decreases to a global minimum at λ_{min} close to 0, and converges to a horizontal asymptote $\boldsymbol{CV}(\infty)$. Figure 3 shows the function $\boldsymbol{CV}(\lambda)$ for the transformed rainfall data.

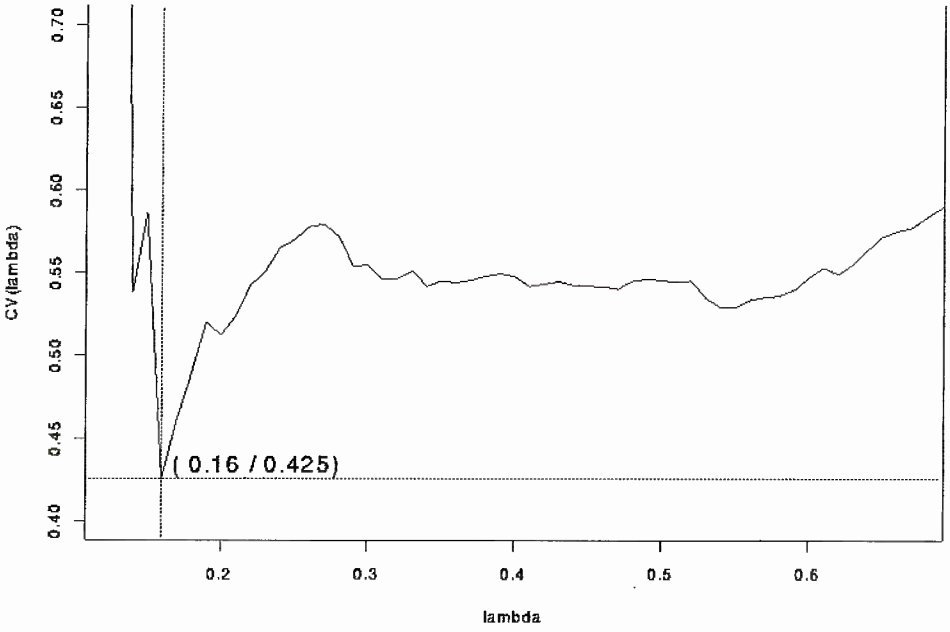

Figure 3: *The cross-validation criterion* $\boldsymbol{CV}(\lambda)$*. It decreases to a global minimum at* $\lambda_{min} = 0.16$*, and converges to a horizontal asymptote* $\boldsymbol{CV}(\infty)$*.*

We get $\boldsymbol{CV}(\lambda_{min}) = 0.425$ at $(\lambda_{min}) = 0.16$. Note that the estimation of `loess` with smoothing parameter $\lambda = \infty$ models a parabolic surface which is not sufficiently effective to remove local trends. Thus we removed the non-parametric trend $\hat{m}(x) = \hat{g}(x, \lambda_{min} = 0.16)$ from $Z(\mathbf{x})$, the transformed data set.

3. HIGHLY ROBUST VARIOGRAM ESTIMATION

Variogram estimation is a crucial stage of spatial prediction, because it determines the kriging weights. It is important to have a variogram estimator which remains close to the true underlying variogram, even if outliers (faulty observations) are present in the data. Otherwise kriging can produce non-informative maps. Let $\varepsilon(\mathbf{x}) = Z(\mathbf{x}) - m(\mathbf{x})$, be the detrended spatial stochastic process, which is assumed to be intrinsically stationary. The classical variogram estimator of a sample $\varepsilon(\mathbf{x}_1)$, ..., $\varepsilon(\mathbf{x}_n)$ proposed by Matheron (1962), based on the method-of moments, is

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N_h} \sum_{N(\mathbf{h})} (\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j))^2, \qquad \mathbf{h} \in \Re^d \tag{3}$$

where $N(\mathbf{h}) = \{(\mathbf{x}_i, \mathbf{x}_j): \mathbf{x}_i - \mathbf{x}_j = \mathbf{h}\}$ and N_h is the cardinality of $N(\mathbf{h})$. This estimator is unbiased, but behaves poorly if there are outliers in the data. One single outlier can destroy this estimator completely. However, it is not enough to make simple modifications to formula (3), such as the ones proposed by Cressie and Hawkins (1980), in order to achieve robustness. In this section, we advocate the use of a highly robust variogram estimator (Genton 1996, 1998a)

$$2\hat{\gamma}(\mathbf{h}) = (Q_{N_h})^2, \qquad \mathbf{h} \in \Re^d \tag{4}$$

which takes account of all the available information in the data. It is based on the sample $V_1(\mathbf{h})$,..., $V_N(\mathbf{h})$ from the process of differences $V(\mathbf{h}) = \varepsilon(\mathbf{x} + \mathbf{h}) - \varepsilon(\mathbf{x})$ and the robust scale estimator Q_{N_h}, proposed by Rousseeuw and Croux (1992, 1993)

$$Q_{N_h} = 2.2191 \{|V_i(\mathbf{h}) - V_j(\mathbf{h})|; i < j\}_{(k)} \tag{5}$$

where the factor 2.2191 is for consistency at the Gaussian distribution,

$$k = \binom{[N_h/2]+1}{2},$$

and $[N_h/2]$ denotes the integer part of $N_h/2$. This means that we sort the set of all absolute differences $|V_i(\mathbf{h})-V_j(\mathbf{h})|$ for $i < j$ and then compute its k-th quantile ($k \approx$ ¼ for large N_h). This value is multiplied by the factor 2.2191, thus yielding Q_{N_h}. Note that this estimator computes the k-th order statistic of the $\binom{N_h}{2}$ interpoint distances.

At first sight, the estimator Q_{N_h} appears to need computation time, which would be a disadvantage. However, it can be computed using no more than time storage, by means of the fast algorithm described in Croux and Rousseeuw (1992).

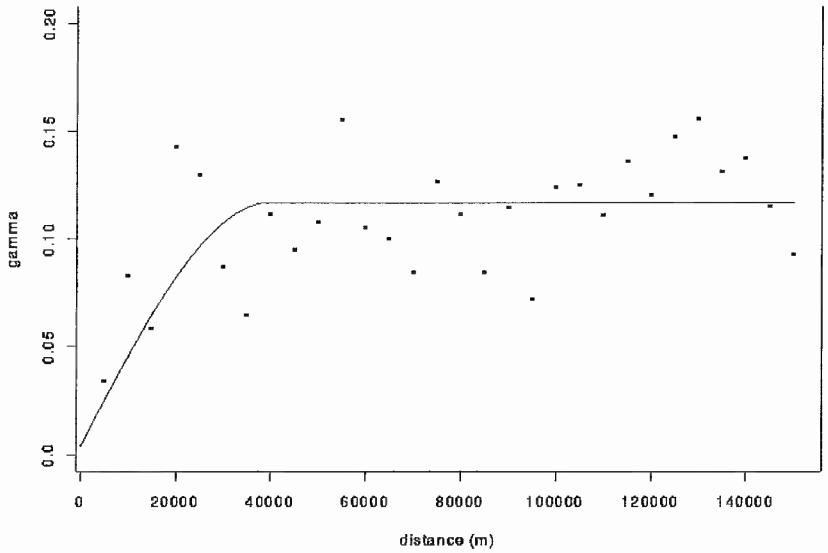

Figure 4: *Omnidirectional variogram estimated by Q_{N_h} and fitted by generalized least squares* (GLSE).

This variogram estimator possesses several interesting properties of robustness. For instance, its influence function, which describes the effect on the estimator of an infinitesimal contamination, is bounded. This means that the worst influence that a small amount of contamination can have on the value of the estimator is finite, in opposition to Matheron's classical variogram estimator. Another important robustness property is the breakdown point ε^* of a variogram estimator, which indicates, how many data points need to be replaced to make the estimator explode (tend to infinity) or implode (tend to zero). The highly robust variogram estimator has an $\varepsilon^* = 50\%$ breakdown point on the differences $V(\mathbf{h})$, the highest possible value, whereas Matheron's classical variogram estimator has only an $\varepsilon^* = 0\%$ breakdown point, the lowest possible value. More details about the use and properties of this estimator, including some simulation studies, are presented in Genton (1998a).

We set the lag unit to $u = 5000$m, which is a reasonable approximation for the nearest neighbor distance mean in east-west (E-W) and north-south (N-S) direction. We estimate the variogram at lags $h_i = iu$, $i = 1,...,30$, with the common tolerance of a half unit to achieve higher robustness. We compute the directional variogram for the N-S and the E-W direction, as well as the omnidirectional variogram with (5) by using `variogram.qn`, a new function in S+. The N-S and E-W directional variograms show similar behavior, suggesting an underlying isotropic process. Therefore we decide to fit the omnidirectional variogram (Figure 4).

4. VARIOGRAM FITTING BY GENERALIZED LEAST SQUARES

Variogram fitting is another crucial stage of spatial prediction, because it also determines the kriging weights. Careful fitting implies on one hand the use of a highly robust variogram estimator (Genton, 1998a). On the other hand, variogram estimates at different spatial lags are correlated, for the same observation is used for different lags. As a consequence, variogram fitting by ordinary least squares is not satisfactory. This problem is addressed by Genton (1998b), who suggests the use of a generalized least squares method with an explicit formula for the covariance structure (GLSE). A good approximation of the covariance structure is achieved by taking account of the explicit formula for the correlation in the independent case. Simulations were carried out with several types of underlying variograms, as well as with outliers in the data. Results showed that the (GLSE) technique, combined with a robust estimator of the variogram, improves the fit significantly.

Consider a omnidirectional variogram estimator for a given set of lags $h_1, \ldots, h_k$, where $1 \leq k \leq K$ and K is the maximal possible distance between data. Denote further by $2\hat{\gamma} = (2\hat{\gamma}(h_1),\ldots,2\hat{\gamma}(h_k))^T \in \Re^k$ the random vector with variance-covariance matrix $\mathrm{Var}(2\hat{\gamma}) = \tau^2\Omega$ where τ^2 is a real positive constant. Suppose that one wants to fit a valid parametric variogram $2\gamma(h,\theta)$ to the estimated points $2\hat{\gamma}$. The method of generalized least squares consists to determine the which minimizes

$$G(\theta) = (2\hat{\gamma} - 2\gamma(\theta))^T \Omega^{-1} (2\hat{\gamma} - 2\gamma(\theta)) \tag{6}$$

where $2\gamma(\theta) = (2\gamma(h_1,\theta),\ldots,2\gamma(h_1,\theta))^T \in \Re^k$ is the vector of the valid parametric variogram, and $\theta \in \mathrm{R}^p$ is the parameter to be estimated. Note that $2\gamma(h,\theta)$ is generally a nonlinear function of the parameter. Journel and Huijbregts (1978) suggest to use only lag vectors h_i such that $N_{h_i} > 30$ and $0 < i \leq K/2$. This empirical rule is often met in practice, and is used in this work. The `GLSE` algorithm is the following:

[1] Determine the matrix $\Omega = \Omega(\theta)$ with element Ω_{ij} given by

$$Corr(2\hat{\gamma}(\mathrm{h}_i),2\hat{\gamma}(\mathrm{h}_j))\gamma(\mathrm{h}_i,\theta)\gamma(\mathrm{h}_j,\theta)/\sqrt{\mathrm{N}_{h_i}\mathrm{N}_{h_j}} \tag{7}$$

[2] Choose $\theta^{(0)}$ and let $l = 0$.

[3] Compute the matrix $\Omega = \Omega(\theta^{(l)})$ and determine $\theta^{(l+1)}$ which minimizes

$$G(\theta) = (2\hat{\gamma} - 2\gamma(\theta))^T \Omega(\theta^{(l)})^{-1} (2\hat{\gamma} - 2\gamma(\theta)) \tag{8}$$

[4] Repeat [3] until convergence to obtain $\hat{\theta}$.

In step [1], the correlation $Corr(2\hat{\gamma}(\mathrm{h}_i),2\hat{\gamma}(\mathrm{h}_j))$ can be approximated by the one in the independent case. An explicit formula can be found in Genton (1998b), which depends only on the lags h_i and h_j, as well as on the size $n = n_1 n_2$ of a spatial rectangular data set. In step [2], the choice of $\theta^{(0)}$ can be carried out randomly, or with the result of a fit by ordinary least squares (OLS).

A spherical variogram

$$\gamma(\mathrm{h},\theta) = \begin{cases} 0 & \text{if} \quad \mathrm{h} = 0 \\ \theta_1 + \theta_2\left(\frac{3}{2}\left(\frac{\mathrm{h}}{\theta_3}\right) - \frac{1}{2}\left(\frac{\mathrm{h}}{\theta_3}\right)^3\right) & \text{if} \quad 0 < \mathrm{h} \leq \theta_3 \\ \theta_1 + \theta_2 & \text{if} \quad \mathrm{h} > \theta_3 \end{cases}$$

has been fitted to the empirical omnidirectional variograms by GLSE using `glse.fitting`, another new function in S+. The starting value $\theta^{(0)}$ was set as the solution of a fit by OLS. The estimated parameter is $\hat{\theta}$ = (0.003, 0.113, 7.938). To calculate (7), we neglected the irregularities of the grid and set $n_1 = 60$ and $n_2 = 40$, which is a crude approximation of the grid.

5. KRIGING AND DISCUSSION OF THE RESULTS

Epitomizing, kriging is a linear interpolation method that allows predictions of unknown values of a random function from observations at known locations. For further details see Cressie (1991). `S+SPATIALSTATS` performs 2-dimensional kriging by using the `krige` and

`predict.krige` functions. The kriging results are easily visualized with the functions `contour` or `persp`. Further details and examples are given by Kaluzny et al. (1996). Figure 5 shows the kriging maps for the rainfall data. The ten lowest values are represented by circles and the ten highest by squares.

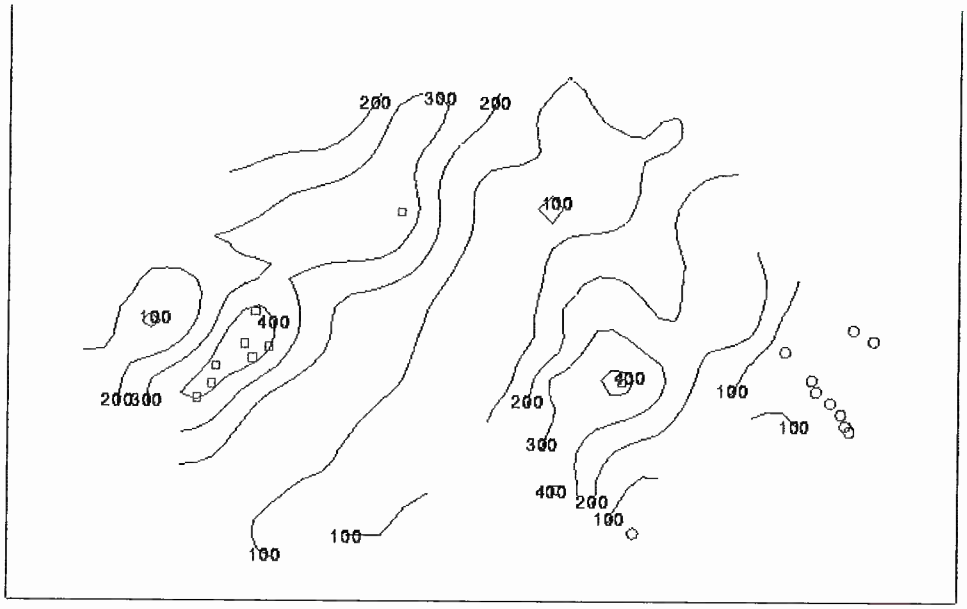

Figure 5: *Kriging map of the rainfall data. The ten lowest values are represented by blue circles and the ten highest by red squares.*

Figure 6 visualizes the corresponding errors and kriging variances respectively. The overall performance of our method in predicting the remaining 367 rainfall data is summarized in Table 1. We consider the true values $Z(\mathbf{x})$, the estimated values (by kriging), the errors, the absolute errors $|e(\mathbf{x})|$and the relative errors $|e(\mathbf{x})|/Z(\mathbf{x})$. For each of these quantities, the minimum, the maximum, the mean, the median and the variance is computed. The distribution of the estimated values by kriging is in close agreement with the distribution of the true values. This is confirmed by a plot of estimated values (horizontal) against true values (vertical) in Figure 7.

	min	max	mean	median	std. dev.
true values	0	517	185	162	111
estimated values	14	433	171	151	96
errors	-255	230	-14	-10	61
absolute errors	0	254	32	44	43
relative errors	0	3.31	0.32	0.21	0.42

Table 1: *This table presents the minimum, the maximum, the mean, the median, and the standard deviation for the 367 true rainfall values, the estimated values, the errors, the absolute errors, and the relative errors.*

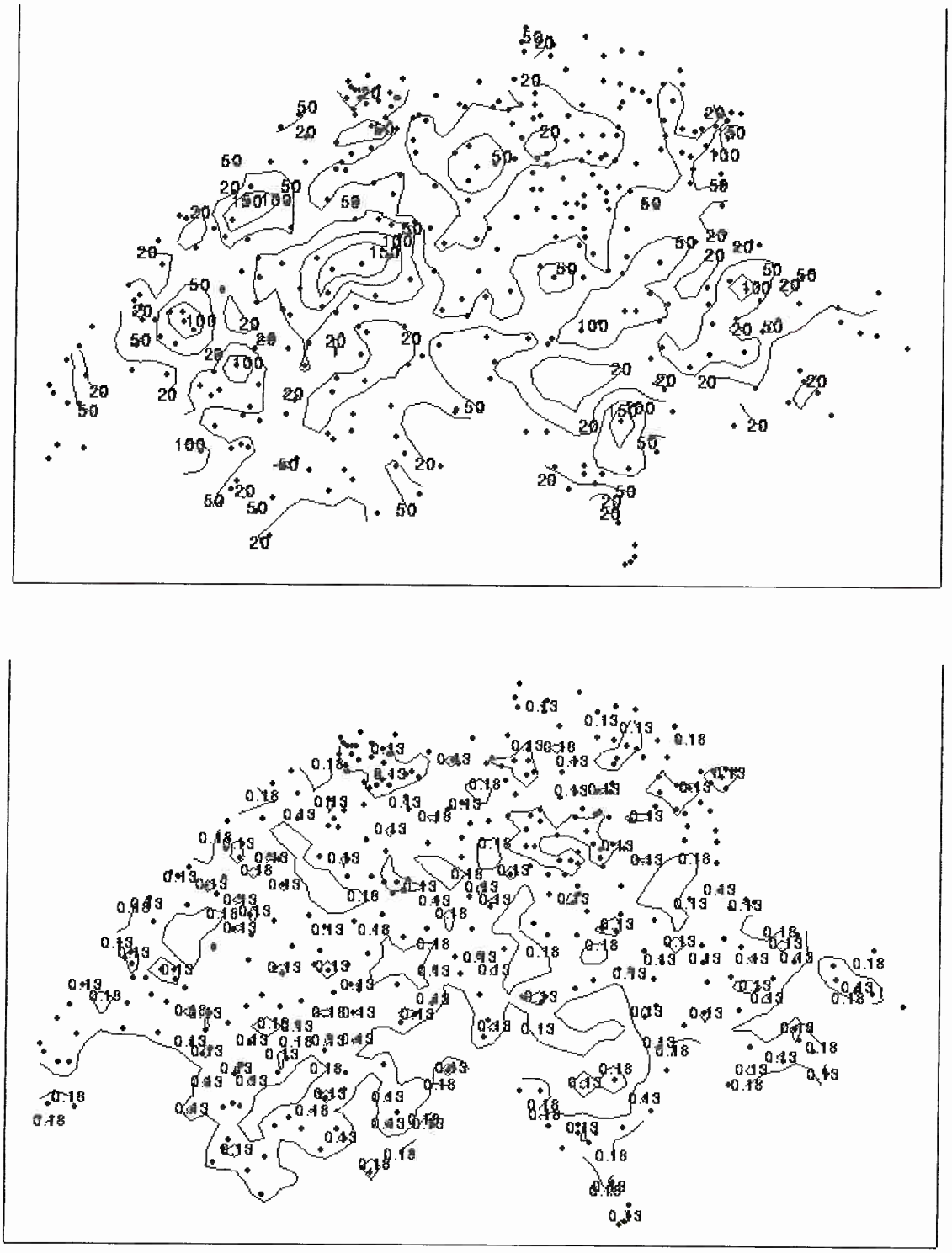

Figure 6: *Absolute errors and kriging variances corresponding to the kriging map of the rainfall data.*

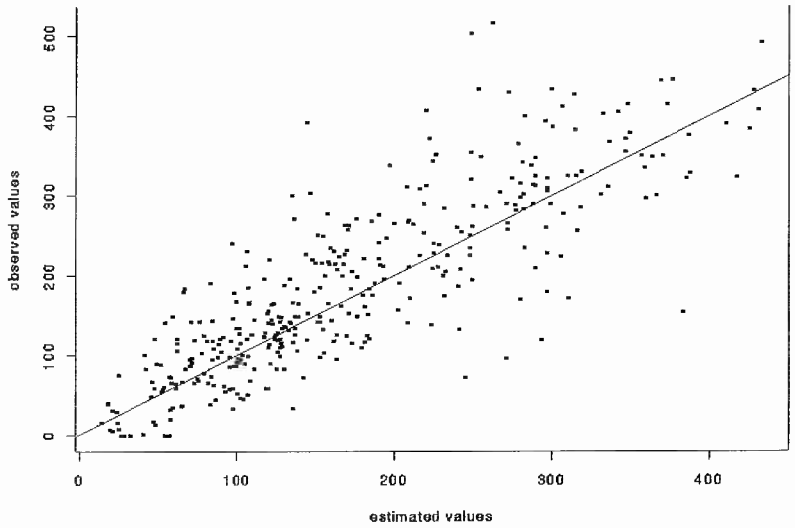

Figure 7: *Plot of estimated values (horizontal) against true values (vertical).*

A small positive bias is however revealed. A plot of observed (true) values against residuals in Figure 8 indicates that small values are generally overestimated whereas large values are underestimated.

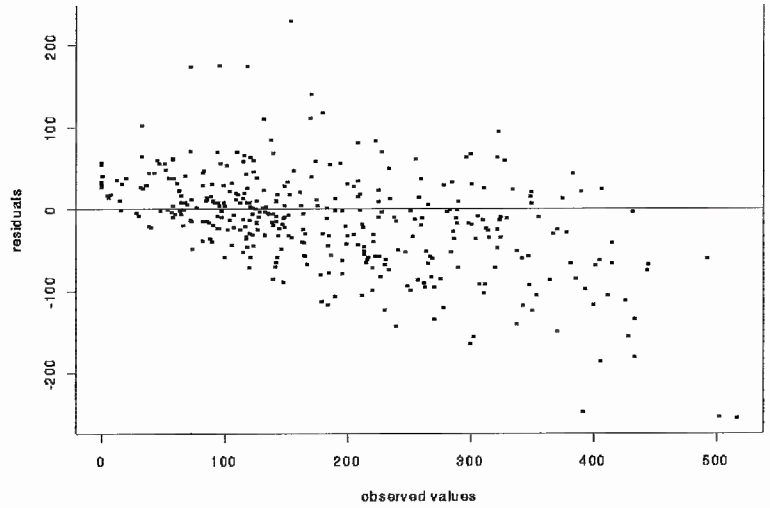

Figure 8: *Plot of observed (true) values (horizontal) against residuals (vertical).*

Proportional plots of the absolute errors and relative errors in Figure 9 indicates the locations of the smaller or higher errors. It seems to be correlated with the smaller or higher rainfall measurements. The root mean squared error is $RMSE = 62$ and should be compared with other predicting methods. Table 2 compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values. This method identified four respectively three locations of the ten highest respectively ten lowest values of the initial data set. The performance in predicting the lowest and the highest 10 rainfall measurements can also be summarized by the root mean squared error $RMSE_{min} = 15$ and $RMSE_{max} = 19$ respectively. It seems that lower values are more accurately predicted than higher ones.

This method of rainfall prediction can be useful for the monitoring of accidental releases of radioactivity in the environment, because it doesn't require huge computations, nor subjective modeling decisions. The procedure is straightforward and almost automatized with the new available S+ functions. As rainfall is strongly correlated with radioactive fallout of accidental

releases, this method can easily be used for emergency situations.

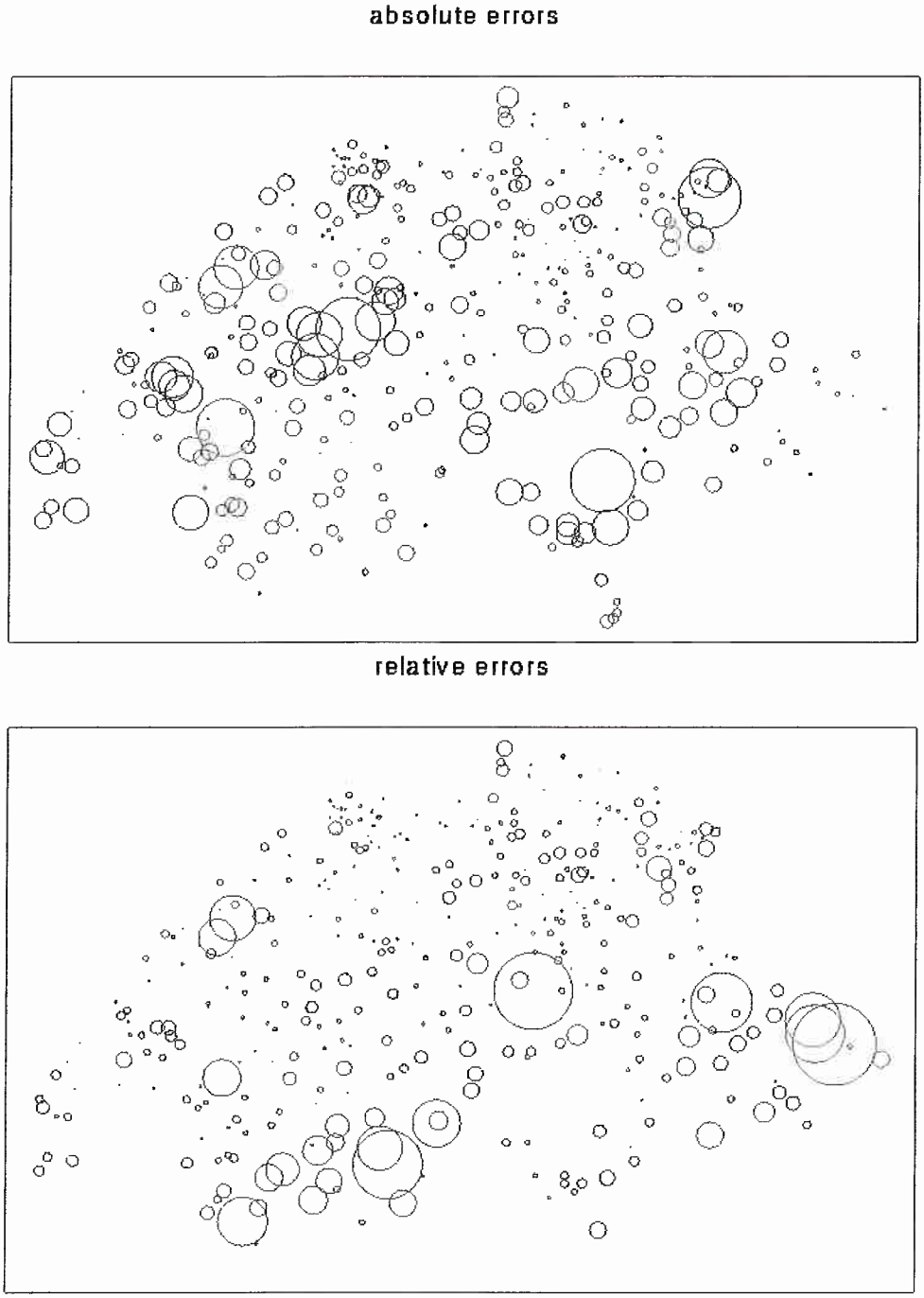

Figure 9: *Proportional plots of the absolute and relative errors for the kriging of the rainfall measurements. Positive errors are represented by red circles, negative errors by blue circles.*

ten lowest values		ten highest values	
true values	estimated values	true values	estimated values
0	57	434	300
0	55	434	253
0	33	441	441
0	27	444	369
0	29	445	377
1	41	452	452
5	22	493	433
6	20	503	249
8	26	517	262
10	10	585	585

Table 2: *This table compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values.*

6. CONCLUSIONS

In this paper we have studied a data set of rainfall measurements in Switzerland. As local drifts are typically present in the data set due to geographical characteristics and elevations, the observations have been detrended by a non-parametric surface, based on a cross-validation criterion. Then, robust methods have been applied for variography with the software `S+SpatialStats`. First, the variogram was estimated by a highly robust estimator. Second, the fit of the variogram estimates was done by generalized least squares thus taking account of their statistical properties. Kriging has been performed and the overall performance analyzed by various criterion. Results are of course different from the true values, but however in good agreements. Thus, our model is a simple way of studying such data sets, without the need of huge computations. As rainfall is strongly correlated with radioactive fallout of accidental releases, this method can easily be used for emergency situations.

The S+ functions used are made available at http://dmawww.epfl.ch/~furrer/SIC97/

References

1. Cleveland, W.S., Grosse, E., Shyu, W. M. (1992). *Local Regression Models.* Chapter 8 in: Statistical Models in S. Edited by Chambers, J. M. and Hastie, T. J. New York: Chapman & Hall.
2. Cressie, N. (1991). *Statistics for Spatial Data*, Wiley.
3. Croux, C., Rousseeuw, P.J. (1992). Time-efficient algorithms for two highly robust estimators of scale, *Comp. Stat.*, 1, 411-428.
4. Genton, M. G. (1996). "*Robustness in Variogram Estimation and Fitting in Geostatistics*", Ph.D. Thesis #1595, Department of Mathematics, Swiss Federal Institute of Technology.
5. Genton, M. G., (1998): Highly Robust Variogram Estimation, *Mathematical Geology*, **30**, 213-221.
6. Genton, M. G., (1998): Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure, *Mathematical Geology*, **30**, 323-345.

7. Green, P. J, Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.
8. Journel, A. G., Huijbregts CH. J. (1978). *Mining Geostatistics*, Academic Press.
9. Kaluzny, S. P., Vega, S. C., Cardoso, T. P., Shelly A. A. (1996). *S+SpatialStats, User's Manual*, MathSoft.
10. Rousseeuw, P. J. and Croux, C. (1992). Explicit scale estimators with high breakdown point. *L_1, Statistical Analysis*, 77-92.
11. Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the Median Absolute Deviation, *Journal of the American Statistical Association*, **88**, 1273-1283.
12. Venables, W. N., Ripley B. D. (1994). *Modern Applied Statistics with S-Plus*, Springer.