Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 77-90, 1998

Geostatistical Classification and Class Kriging

Denis Allard
Unité de Biométrie, Institut National de la Recherche Agronomique, INRA Domaine Saint Paul, Site Agroparc, 84914 Avignon cedex 9, France
allard@avignon.inra.fr

ABSTRACT **A new method is proposed for the classification of data in a spatial context, based on the minimization of a variance-like criterion taking into account the spatial correlation structure of the data. Kriging equations satisfying classification bias conditions are then derived for interpolating the rainfall data while taking into account the classification.**
KEYWORDS : **geostatistics, spatial statistics, classification, kriging, interpolation, rainfall data.**

Contents

1. Introduction

1.1 Exploratory data analysis

In this subsection it will be argued that the data need to be separated into two groups, corresponding to low and high rainfall values. Let us consider a cutoff value of 215 (here, and in the rest of the paper, units are always 0.1 mm). This value seems for the moment somewhat arbitrary, but it will be justified later, in section 2. Table 1 shows elementary

statistics for these two groups as well as for the complete data set. The group of higher values contains only 31 data and its mean is three times the mean of the group of lower values. The coefficient of variation is decreased in both groups. It should also be noted that if the overall correlation between elevation and rain is low (-0.2), it is increased in absolute value in both groups, reaching -0.4 in the group of higher value.

Table 1 Elementary statistics for the two groups and for the complete data set

	N	μ	σ	μ / σ	$r(Z,H)$
Z < 215	69	113.1	48.5	0.43	-0.3
Z > 215	31	329.4	78.7	0.22	-0.4
All Data	100	180.2	116.7	0.65	-0.2

A graphical exploration of the data corresponding to this cutoff is depicted in Figure 1. The histogram (Figure 1a) looks bimodal, which by itself is an indication that the data can be separated into two groups with different probability densities. The cutoff is visible in the 5^{th} bin and is located towards the lower values of the intuitive limit between the two modes. This will be explained in section 2.3.

On Figure 1b high values are organized along what seems to be two fronts oriented SW-NE. Between and around these two fronts, rainfall is low, especially in the SW-NE corridor separating them. The convex hulls of the two groups are also shown on the base map (in dashed lines). In the SE group, the convex hull contains no low values (i.e. below the cutoff 215), and only three of them (79, 107, 194) are present in the NW group. This shows how convex and concentrated these groups are. Interestingly enough, the variograms (isotropic variograms, with a 6.4 km lag and a lag tolerance of 3.2 km) computed within each group and normalized by the corresponding variances can easily be superimposed (Figure 1c) and modeled using a unique isotropic correlation function:

$$r(\bullet) = 0.53\,sph(\bullet, 17 \text{ km}) + 0.47\,sph(\bullet, 100 \text{ km}),$$

where $sph(\bullet, a)$ is the spherical correlation function:

$$sph(h,a) = 1 - 3/2\,(|h|/a) + 1/2\,(|h|/a)^3.$$

This will be of great interest in section 2, when the classification algorithm will be presented.

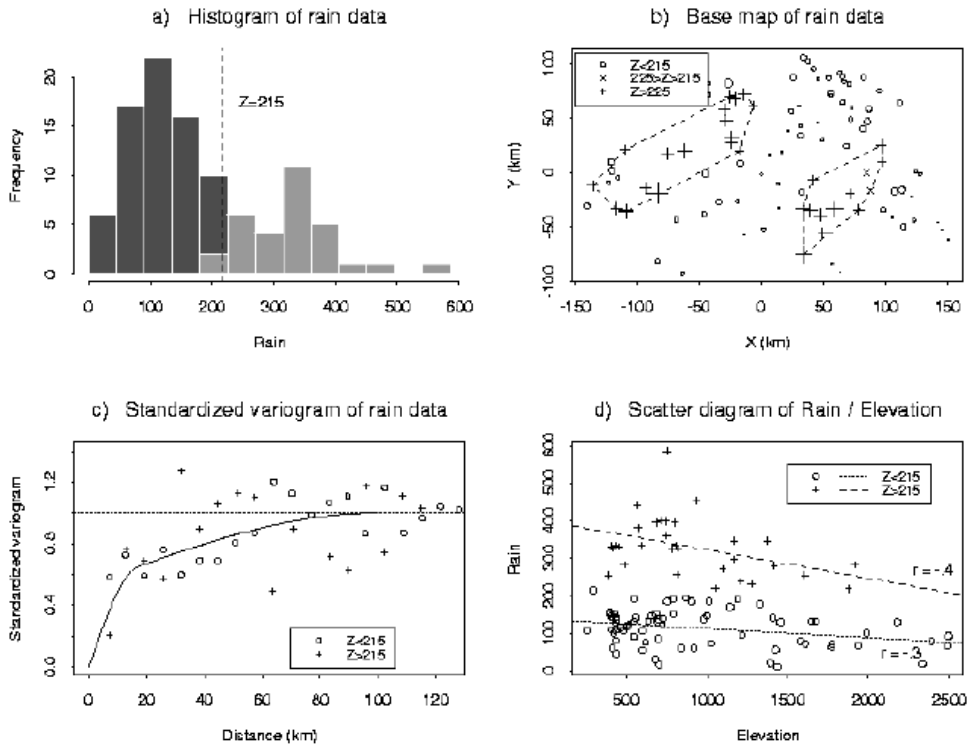

Figure 1 Exploration of the data: a) Histogram of the rainfall, with the cutoff indicated (215); in black, part of the histogram below the cutoff; in gray, part of the histogram above the cutoff. b) Map of the data, with the convex hulls (dashed lines) of the class of higher values. Two data (218 and 220) are highlighted with a '×' sign. c) Experimental variograms of the two groups normalized by their variances and the theoretical model (solid line). d) Scatter plot and regression of rainfall vs. elevation for the two groups, with the regression lines (dashed lines).

Figure 1d shows the relation between rain and elevation (given by the closest elevation on the grid of the digital elevation model provided with the rain data). As already mentioned, the overall correlation (-0.2) is increased in absolute value in each group. It remains however rather weak: -0.3 on the group of lower values (69 data) and -0.4 on the group of higher values (31 data). These values are not significant for independent data at the 95 % confidence level; hence they are not significant for dependent data. Moreover, real elevations are not known, but rather the elevation at the closest grid point. The error can be important for high elevations and in the mountains. The regression lines are also depicted, and the slope is almost negligible (-0.025 10^{th} mm/m) for the group of lower values. For all these reasons, it was decided to not include elevation data in the geostatistical prediction model.

In summary, it seems that there is a case for separating the data into two groups, one corresponding to the higher values (probably rain fronts), and one for the rest of the country. The rest of the paper is organized as follows: the next subsection is a short review of classification methods in a spatial context. Section 2 addresses the question of estimating the cutoff separating the data into two groups. In section 3, kriging equations

in the case of several groups are developed, the so called Class Kriging and Probability Class Kriging. In section 4, the interpolated data are compared to the real data, and the method is discussed.

1.2 Classification procedures: a short review

There is a substantial literature about classification, and only a partial review will be attempted here. There are two main approaches to this problem: a parametric approach usually based on the Gaussian model, and a non-parametric approach based on the minimization of a variance-like criterion.

Celeux and Govaert (1992, 1995) and Banfield and Raftery (1993), among others, extensively studied model-based clustering, possibly in the presence of noise. The model is a mixture of multivariate Gaussian random variables with independent data and the classification procedure is based on a Maximum Likelihood argument using either the EM algorithm (Dempster *et al.*, 1977) after choosing a parameterization of the model in terms of decomposition of the covariance matrices in terms of eigenvalues and eigenvectors, or a Markov Chain Monte-Carlo method in a Bayesian framework (Richardson and Green, 1997). Discriminant analysis (Venables and Ripley, 1994) can be seen as one of the simplest special cases of this approach.

The non parametric approaches are probably the oldest. In the case of several classes, the variance can always be decomposed as the sum of a within-class variance and a between-classes variance. Non parametric methods aim at minimizing the within-class variance, hence maximizing the contrast between classes. This is basically the idea of hierarchical clustering, (Venables and Ripley, 1994). Breiman *et al.* (1984) proposed a general method, called Classification And Regression Trees (CART), for the classification of multivariate data. It is based on the minimization of a deviance criterion, related to (but different from) the within-classes variance. Again, these methods always consider independent data.

In fact, only a few authors have studied classification of spatial data. Switzer *et al.* (1982) gave an algorithm for smoothing discriminant analysis classification maps using a prior probability method. Oliver and Webster (1989) proposed a method for clustering multivariate non-lattice data. They propose modification of the dissimilarity matrix of the data by multiplying it by a variogram. Although this approach leads to a sensible algorithm and to well-behaved maps, the statistical model behind the method is not at all clear. More recently, Ambroise *et al.* (1997) proposed an interesting classification algorithm for Markov random fields on lattices, that is able to take into account spatial dependence. It is based on a Gibbs model and the procedure is an EM algorithm. They give interesting references to earlier work on spatial classification. This work has been extended by the same authors (private communication) to non lattice data using a neighborhood defined by the Delaunay Graph of the data (i.e. the nearest-neighbor graph based on the Voronoï tessellation).

In this paper, a new classification procedure based on a Gaussian geostatistical model is proposed. It can be seen as an extension of model-based clustering to the spatial case.

2. Classification

2.1 The model

Let us first consider a sample of size *n*, $\mathbf{Z} = (Z(x_1),\ldots,Z(x_n))^t$, of a Gaussian stationary random function having mean *m*, variance σ^2 and correlation function *r( )*. Then, it is a well known result that provided *r* is known, the Maximum Likelihood Estimators for *m* and σ^2 are

$$\hat{\mu} = \frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} \text{ and } \hat{\sigma}^2 = \frac{(\mathbf{Z}-\hat{\mu}\mathbf{1})^t\mathbf{R}^{-1}(\mathbf{Z}-\hat{\mu}\mathbf{1})}{n} = \frac{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}{n}\left(\frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} - \left(\frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}\right)^2\right), \quad (1)$$

where $\mathbf{1} = (1,\ldots,1)^t$ is the vector of ones of length n and $\mathbf{R}$ is the correlation matrix given by $\mathbf{R}_{ij} = \rho(x_i - x_j)$. These equations can be rewritten in a much more concise way by defining

$$\langle \mathbf{X},\mathbf{Y}\rangle = \frac{\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Y}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} \quad (2)$$

for two vectors $\mathbf{X}$ and $\mathbf{Y}$ of length *n*. Then,

$$\hat{\mu} = \langle \mathbf{Z},\mathbf{1}\rangle \text{ and } \hat{\sigma}^2 = \frac{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}{n}\left(\langle \mathbf{Z},\mathbf{Z}\rangle - \langle \mathbf{Z},\mathbf{1}\rangle^2\right). \quad (3)$$

Remarks:

1) $\hat{\mu}$ is nothing but the kriging of the mean (see Wackernagel 1995). Its variance is $\sigma^2 / \mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}$.

2) If there is no spatial correlation (case of a pure nugget effect), then $\mathbf{R} = \mathbf{I}$, the identity matrix, and all the equations above simplify themselves. In this case $\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1} = \mathbf{1}^t\mathbf{1} = n$, $\hat{\mu}$ is the arithmetic average and $\hat{\sigma}^2 = \mathbf{Z}^t\mathbf{Z}/n - \hat{\mu}^2$ is the usual variance. Equations (3) have clearly the structure of an expectation and a variance. Indeed, the scalar product (2) can be seen as a natural extension of the usual operator $\mathbf{A}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} X_iY_i/n$ to the spatial statistics context.

2.2 A classification algorithm

Let us now consider that the data can be separated into a partition $\{K_1, K_2\}$ with n_1 and n_2 elements respectively ($n_1 + n_2 = n$). Provided that the correlation function is known, means and variances can be calculated according to equations (3) on this partition. The total variation of this partition is defined by $V = n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2$. If the partition is made such that there are two homogeneous groups, the variance in each class will be low, and as a consequence, the total variation will be low as well. The 'best' classification of the data is then defined as the partition minimizing the criterion

$$V = n_1 \hat{\sigma}_1^2 + n_2 \hat{\sigma}_2^2 = \mathbf{1}^t \mathbf{R}_1^{-1} \mathbf{1} \left(\langle \mathbf{Z}_1, \mathbf{Z}_1 \rangle - \langle \mathbf{Z}_1, \mathbf{1} \rangle^2 \right) + \mathbf{1}^t \mathbf{R}_2^{-1} \mathbf{1} \left(\langle \mathbf{Z}_2, \mathbf{Z}_2 \rangle - \langle \mathbf{Z}_2, \mathbf{1} \rangle^2 \right), \quad (4)$$

where Z_i is the vector of data and $\mathbf{R}_i$ is the correlation matrix in group *i*, *i* = 1, 2.

If there is no spatial correlation, this criterion is equivalent to the minimization of the deviance, as introduced by Breiman *et al.* (1984) for the Classification And Regression Tree (CART) method. In this case, the partition minimizing (4) is necessarily reached for one of the n cutoffs separating the data in the sets L of data lower than or equal to the cutoff and H of data larger than the cutoff. Because of the spatial correlation, the optimal partition according to (4) is not necessarily defined by a cutoff, but should rather be found within the set of the 2^n possible partitions. For a large data set ($n \geq 20$, say) a comprehensive search is not possible. A paper, in preparation, proposes a solution to this problem using an iterative method. Since this is not the scope of this paper, this problem will not be addressed here. In this paper, it will be assumed that a reasonable description is achieved for two groups defined by one of the cutoffs. The algorithm for finding the cutoff k_c minimizing (4) is then the following:

- Sort the data, and let $Z_{(1)} \leq, \cdots, \leq Z_{(n)}$ be the order statistics.
- For each k, $1 \leq k \leq n$

  - Define $L_k = \left\{ Z(x_i) : \; Z(x_i) \leq Z_{(k)} \right\}$ and $H_k = \left\{ Z(x_i) : \; Z(x_i) > Z_{(k)} \right\}$;
  - Compute $\hat{A}_l$, $\hat{A}_h$, $\hat{\sigma}_l^2$, $\hat{\sigma}_h^2$ and V_k according to equations (3);
- Find k_c such that V_{k_c} is minimum. The classification is given by the partition $\left\{ L_k, H_k \right\}$.

2.3 Classification of the data

The algorithm was applied using the correlation function fitted on Figure 1d. The result is shown on Figure 2 ('*' signs). The minimum covariance is reached for $k_c = 69$. Figure 2 also shows the total variance criterion ignoring the spatial correlation (o). In this case, the optimal cutoff is $k_0 = 71$. For comparison purposes, Table 2 shows the elementary statistics for three cases: using cutoff k_0, using cutoff k_c and computing the means and variances ignoring and then taking into account the spatial structure.

Table 2 Elementary statistics for two classification methods and for two estimators.
(Note: (1) non spatial cutoff; (2) spatial cutoff, non spatial statistics; (3) spatial cutoff, spatial statistics).

	cutoff	n_l	n_h	$\hat{A}_l$	$\hat{A}_h$	$\hat{\sigma}_l$	$\hat{\sigma}_h$
(1)	225	71	29	116.0	337.1	51.0	75.6
(2)	215	69	31	113.1	329.5	48.5	78.7
(3)	215	69	31	113.4	323.3	50.7	83.9

In the case of the spatial classification algorithm, the group of higher values has two more data than in the case of the usual CART algorithm. These two data ($Z = 218$ and $Z = 220$) are highlighted on Figure 1 with a '×' instead of a '+'. These two extra points increase substantially the continuity and the convexity of the SE front. This is

precisely the reason why a classification algorithm taking into account the spatial correlation should be preferred. As a consequence, however, the cutoff is slightly shifted (here towards the lower values) from the visual point of bimodality (on the histogram, the eye ignores spatial correlation). A second interesting feature of the method is in the estimation of the mean and variance. In the second line, means and variances are computed using the usual formulas, whereas equations (3) were used in the third line. If the mean is not much affected, the variance is substantially increased (+13%) for the group of higher values. This is a well known result among geostatisticians, and it can be illustrated using a very simple example. Let Z_1 and Z_2 be two data. Without correlation, the estimator of the variance is $\hat{\sigma}_0^2 = (Z_1 - Z_2)^2 / 4$. In the presence of a correlation r, the estimate is $\hat{\sigma}_\rho^2 = \hat{\sigma}_0^2 / (1 - \rho)$.

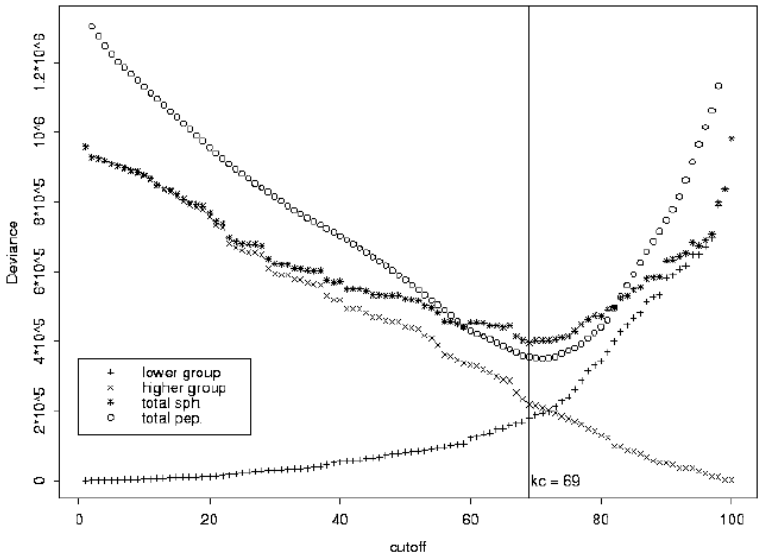

Figure 2 Deviance as a function of the cutoff. The deviance in each group is shown. As the cutoff increases, the deviance in the lower group (+) increases and the deviance of the higher group ($\times$) decreases. The minimum of the total deviance ($*$) is reached for $k_c = 69$. Deviance without spatial correlation (o) has a different minimum.

3. Interpolation

The aim is to build a predictor which is able to account for the separation of the data into two classes. Therefore, we need three things: first, a general model for deriving the interpolation equations; second, an algorithm for deciding which class the point to interpolate belongs to, and third, an interpolation procedure once this information is known. In the following sections a kriging approach able to integrate the class information is developed.

3.1 The model

According to the classification procedure described above, the model is the following: consider a random field $Y(x)$ with mean 0 and correlation function $r(\cdot)$. The random field $Z(x)$ describing the rain is then obtained in each class by multiplication and addition of the corresponding standard deviation and mean: $Z^{(i)}(x) = \mu_i + \sigma_i Y(x)$ (the superscript recalls the class at point x). Section 2 has shown how to estimate μ_i and σ_i. The next subsection shows how to interpolate Z at point x_0 if its class is known.

3.2 Class kriging

Let us suppose for the moment that the class at point x_0 is known, and denote $i_0 \in I = \{l, h\}$. The following linear interpolator is defined:

$$\hat{Z}^{(i_0)} = \sum_{\alpha \in L} \lambda_\alpha^{i_0} Z_\alpha + \sum_{\gamma \in H} \lambda_\gamma^{i_0} Z_\gamma ,$$

where L (resp. H) is the set of lower (resp. higher) observed values. The non bias condition $E\left[\hat{Z}^{(i_0)}\right] = \mu_{i_0}$ leads to the following conditions:

$$\sum_{a \in L} l_a^{(i_0)} = d_{a\, i_0} \quad \text{and} \quad \sum_{g \in H} l_g^{(i_0)} = d_{g\, i_0} \qquad (5)$$

where δ_{ij} is the delta of Kronecker: $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. The usual argument for minimizing the variance of error using the Lagrange multipliers μ and ν leads to the following equations, referred to as the Class Kriging equations:

$$\begin{pmatrix} \sigma_l^2 \mathbf{R}_{[\alpha\beta]} & \sigma_l \sigma_k \mathbf{R}_{[\alpha\gamma]} & \mathbf{1} & \mathbf{0} \\ \sigma_l \sigma_k \mathbf{R}^t_{[\alpha\gamma]} & \sigma_k^2 \mathbf{R}_{[\gamma\varepsilon]} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}^t & \mathbf{0}^t & 0 & 0 \\ \mathbf{0}^t & \mathbf{1}^t & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_\alpha^{(l)} \\ \boldsymbol{\lambda}_\gamma^{(k)} \\ \mu \\ \nu \end{pmatrix} = \begin{pmatrix} \sigma_l \sigma_{i_0} \mathbf{R}_{\alpha 0} \\ \sigma_k \sigma_{i_0} \mathbf{R}_{\gamma 0} \\ \delta_{l i_0} \\ \delta_{k i_0} \end{pmatrix}, \qquad (6)$$

where α, β and γ, ε are indices in L and H, respectively. Accordingly, $\mathbf{R}_{[\alpha\beta]}$ denotes the $n_l \times n_l$ correlation matrix of points in L, $\mathbf{R}_{[\alpha\gamma]}$ the $n_l \times n_k$ correlation matrix between the points of L and H, and so on. $\mathbf{R}_{\alpha 0}$ and $\mathbf{R}_{\gamma 0}$ are the correlation vectors between the point x_0 and the sets of points L and H, respectively, and the $\mathbf{0}$'s are vectors of zeros of correct length. In practice, estimators of the variance will replace the true values. It is also necessary to know the value of the class i_0, which can only be estimated. This is the subject of the next subsection.

3.3 Estimating the class

The first and maybe the most intuitive way of estimating the class at a new point is the nearest-neighbor approach: the estimated class at a new point x_0 is the class of the nearest data point. This is equivalent to estimating the limits of the fronts using Voronoï polygons. The final estimator is then $\hat{Z}(x) = \hat{Z}^{(i)}(x)$ where i is the class of the nearest data point of x_0.

A second approach is possible. Instead of estimating the class, only the probability of belonging to each class is estimated. This is done by kriging the indicator function of the set *H*. The variogram of its indicator function has a strong elliptic anisotropy in the 20°; NE direction. An exponential structure with the following parameters is fitted: 300 km practical range in the 20°; NE direction, 100 km practical range in the 70°; NW direction; and a sill equal to 0.33. The kriging of the indicator function, denoted p^*, can be seen as the (best linear) estimation of the probability of belonging to *H*. According to the equality $\hat{Z}(x) = \hat{Z}^{(i)}(x)P^*(x \in L) + \hat{Z}^{(k)}(x)P^*(x \in H)$, the new kriging system, referred to as the Probability Class Kriging (PCK) is now:

$$\begin{pmatrix} \sigma_i^2\mathbf{R}_{[\alpha\beta]} & \sigma_i\sigma_k\mathbf{R}_{[\alpha\gamma]} & \mathbf{1} & \mathbf{0} \\ \sigma_i\sigma_k\mathbf{R}^t_{[\alpha\gamma]} & \sigma_k^2\mathbf{R}_{[\gamma\gamma]} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}^t & \mathbf{0}^t & 0 & 0 \\ \mathbf{0}^t & \mathbf{1}^t & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_\alpha^{(i)} \\ \lambda_\gamma^{(k)} \\ \mu \\ \nu \end{pmatrix} = \begin{pmatrix} \sigma_i\mathbf{R}_{\alpha 0}\left(p^*\sigma_k + (1-p^*)\sigma_l\right) \\ \sigma_k\mathbf{R}_{\gamma 0}\left(p^*\sigma_k + (1-p^*)\sigma_l\right) \\ 1-p^* \\ p^* \end{pmatrix}. \quad (7)$$

The kriging system (6) is merely a restriction of (7) in which p^* is either 0 or 1.

The two methods of estimating the class are now compared using a cross-validation procedure on rainfall prediction: each point in turn is re-estimated and compared to its true value. The model and its parameters are not re-estimated for each point, but rather are considered fixed. Hence, it is not procedures which are compared, but fully specified models. The model for the correlation function $r(\bullet)$ was defined section 1.1 as:

$r(\bullet) = = 0.53sph(\bullet, 17 \text{ km}) + 0.47sph(\bullet, 100 \text{ km})$.

An ordinary kriging method (directly on *Z*, without any classification) is used as a benchmark (linear model; elliptic anisotropy: long axis is 25°; NE, scale factor = 10 m; short axis is 65°; NW, scale factor = 3.3 m; unique neighborhood). The results of the cross-validations are summarized Table 3.

Table 3 Comparison of cross-validation result

Method	$E[Z^* - Z]$	$\sigma[Z^* - Z]$	$E_{\text{Rob.}}[Z^* - Z]$	$\sigma_{\text{Rob.}}[Z^* - Z]$
NN	5.5	75.0	13.6	50.1
PCK	3.0	65.5	8.0	47.6
OK	1.7	61.1	8.2	46.1

Columns 3 and 4 of Table 3 are robust error means and variances; i.e., statistics computed after excluding the 5 highest errors (in absolute value). The nearest neighbor method (NN) for estimating the class does not lead to very good results. When a class is

wrongly estimated by NN the error of estimation is automatically to the order of the difference of the means. Kriging the probability of belonging to a class smoothes the transition from one class to another and the estimation is improved, especially in terms of the raw variance (column 2). Results of PCK are close to an ordinary kriging procedure with a linear variogram, especially in terms of the robust variance. In theory, ordinary kriging should have been retained as the best candidate for the interpolation algorithm. However, the classification based model is a more appealing model for this particular dataset, as the classification procedure gives a dichotomy of the data directly, which ordinary kriging is not able to do.

4. Comparison and discussion

4.1 Comparison of estimated vs. true data

Overall performance The classification algorithm followed by the Probability Class Kriging equations was run to estimate rainfall on the 367 locations where true data were known. A comparison of estimated vs. true data is summarized in Table 4 and Figures 3 and 4. Since kriging is an exact interpolator, comparison statistics are computed on the 367 data to be estimated. The bias is low (-4.6), indicating a slight overall underestimation and the root mean square error is equal to 57.4. Mean absolute and relative errors are equal to 42.1 and 0.68 respectively. The histogram of Z^* is depicted (along with the histogram of Z) in Figure 3a. It can be seen that Z^* has a peak of distribution in the third bin, around the mean of the group of lower data. The histogram of errors $(Z^* - Z)$ (Figure 3b) is not symmetrical and has a quite important tail towards the lower values, illustrating a problem of under-estimation on a few points.

Table 4 Comparison of estimated vs. true rain data.
(Note: statistics on true values are computed on the 367 values to be estimated, except line 5, where data with Z=0 were excluded).

	Min	Max	Mean	Median	Std. Dev.
Z	0	517	185.2	162.0	111.2
Z^*	21.2	438.6	180.7	160.5	85.9
$(Z^* - Z)$	-290.0	215.7	-4.6	2.6	57.4
$\lvert Z^* - Z \rvert$	0.3	290.0	42.2	31.6	39.0
$\lvert Z^* - Z \rvert / Z$	0.0	91.6	0.68	0.18	2.21

As is always the case with linear interpolation procedures, estimated values are under-dispersed ($\sigma^2(Z^*) = 7393$, whereas $\sigma^2(Z) = 12371$). Hence, low values are consistently over-estimated ($E[Z^* - Z] = 84$ for the 10 lower values *Z*) and high values are consistently under-estimated ($E[Z^* - Z] = -158$ for the 10 higher values Z). This is clearly visible on Figure 3c and 3d: the regression line of Z given Z^* (dashed line) has a negative intercept with a slope larger than one and the correlation between $Z^* - Z$ and Z^* is equal to -0.65. Figure 4a is a map of errors, where positive errors ($Z^* > Z$) are proportionally depicted with a cross (+), whereas negative errors ($Z^* < Z$) are depicted

with a circle (o). On the same picture, the convex hulls computed according to the classes (see section 1.1) are also shown. Important errors are generally located along the border of the convex hulls, except for a couple of underestimates in the NE corner of the image. This illustrates the principal weakness of the method: if the indicator of the classes is wrongly estimated, the estimation error on the data will be important (to the order of the mean difference). The southern limit of the NW front is a typical example of this problem: in the neighborhood of the points to be estimated, there were significantly more points belonging to the estimated class of lower values, while these points had in reality high rainfalls.

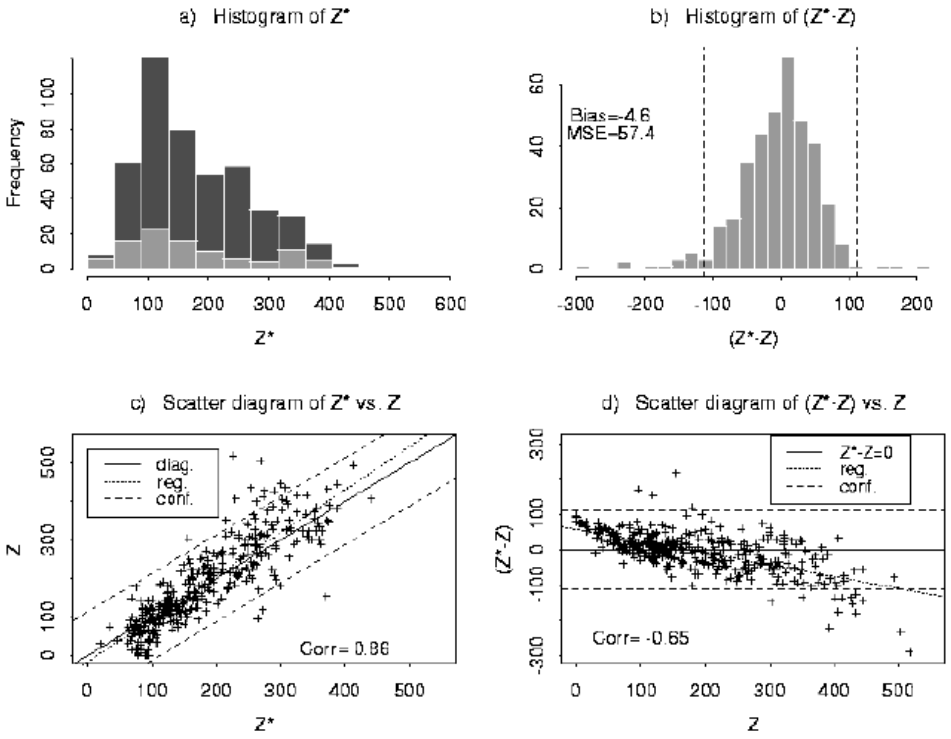

Figure 3 Comparison of estimated vs. true values. a) Histogram of Z^* (in black) and Z (in gray). b) histogram of the error of estimation. c) and d), scatter plots of Z^* and $Z^* - Z$ vs. Z : diagonal and $k_0 = 71$ are shown with a solid line; regression is shown with a dashed line; the envelope $Z^* = Z \pm 1.96\,\sigma(Z^* - Z)$ is also shown with dashed lines.

Estimation of extreme values The performances in predicting the 10 lowest and highest rainfall measurements are summarized in Figure 4b. Since kriging is an exact interpolator, the values included in the sample were always correctly estimated. Among the 10 highest kriging estimates, 7 belong to the sample, 5 belong to the set of 10 highest true values and 1 is wrongly estimated, but not badly located (very close to the upper convex hull). None of true lowest values, there are not included in the sample are

correctly estimated. This is easily understandable as 9 of the lowest values to be estimated (from 0 to 8 10^{th} of mm) are less than the lowest value in the sample (1 mm). It is well known that linear interpolators rarely estimate outside the bounds of the data, hence it was unlikely to be able to locate the minimum values correctly. Except one point in the sample, the set of true and estimated low values are disjoint, and not very close geographically. In summary, if probability class kriging based on geostatistical classification is able to locate the general area of the high values (the fronts, described by the two convex hulls), it does not estimate accurately the exact location of the 10 highest values. As for the low values, the algorithm is not able to identify the locations correctly, because of the sample scheme.

Accuracy The criterion chosen for describing accuracy is the relative error of estimation $(Z^* - Z)/\sigma^*$. On Figure 4c, relative errors larger than 2 (resp. smaller than -2) are highlighted with a cross (+) (resp. with a circle (o)); relative errors with an absolute value less than 2 are depicted with a dot (·). There are more underestimates (23) than overestimates (6). Overall, there are 29 relative errors outside the interval Z; i.e., 7.9% of the errors are close to the theoretical rate of 5%. Again, points with large relative errors are concentrated along the convex hull borders of the class of high values. Hence, the accuracy of the estimate is correctly approximated by twice the kriging standard deviation except along the borders of the classes, where misclassifications (and higher errors than expected) are possible.

4.2 Discussion

A new method for both estimating the classes of bimodal data in a spatial context and interpolating the data was presented. Estimation of the classes is based on a minimum deviance argument, with a maximum likelihood estimator of the deviance. For the moment, the partition is defined by the cutoffs; i.e., it is sought within a restricted subset of all possible partitions. The current classification is thus sub-optimal and a broader search is likely to decrease the number of misclassified points and hence decrease the MSE of the probability class kriging. The maximum likelihood argument is based on the assumption of Gaussian densities in each class, but this assumption is not really necessary. As it is often the case with the Gaussian model, the maximum likelihood solution is also the least square solution (it is for example the case for kriging). Hence, equations (3) and the algorithm of subsection 2.2 can be applied to non-Gaussian data which still are optimal in terms of least squares.

In terms of MSE, it leads to results comparable to ordinary kriging (OK), which is known for being a very robust interpolation method. Its interest lies in the contouring of a region of higher values, which OK is not able to do (it is well known that the contouring of a kriging map is a biased estimator of the set above the threshold). It can be expected that prediction is at least as good as (probably better than) OK inside the classes, but that it can be worst than OK at the borders. Points located at the border of the classes should probably be estimated separately. The final kriging map is depicted in Figure 4d. It shows a strong anisotropy in the SW-NE direction, in accordance with the variogram model. Zones of high values are clearly visible.

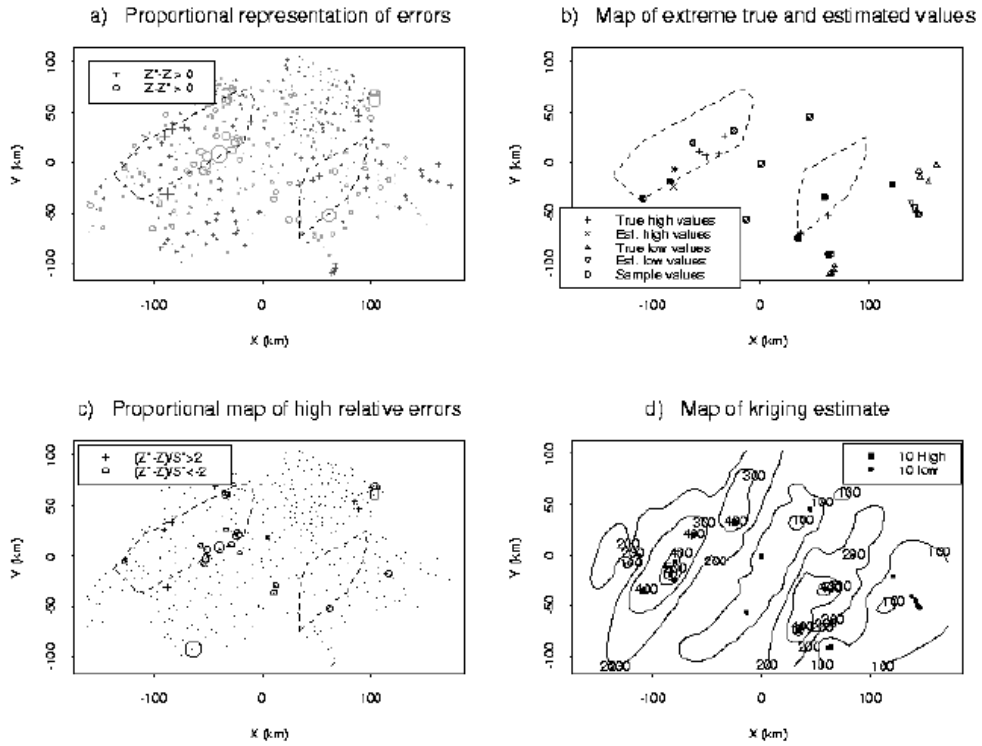

Figure 4 a) and c) Proportional map of errors (a) and relative errors (c); underestimates (o) and overestimates (+) are concentrated along the convex hulls of the classes of higher values (dashed lines). b) Map of the 10 estimated (x) and true (+) highest values, and the 10 estimated (∇) and true (Δ) lowest values; sample values are highlighted with a circle (o). c) Proportional map of high relative errors (in absolute values). d) Kriging map, with the 10 highest (squares) and lowest (circles) estimated values.

Since PCK is a linear interpolator (hence a low-pass filter), it tends to create smooth maps. If the classification part is able to locate the general area of low and high values, it cannot estimate the exact location of extremum values. This is a problem faced by every linear interpolator, and different approaches must be developed for estimating extreme values and middle-range values.

Geostatistical classification and PCK cannot be applied as a black-box for the moment. There is a need for an exploration of the data (are the data bimodal?; are two, or more classes, necessary?) prior to the classification algorithm and a need for pre-modeling (what is the common correlation structure?; what is the variogram of the indicators?) for the kriging. It is a useful tool for long-term management studies: geostatistical classification is able to separate low and high value classes and PCK performs well both in locating high values and in overall estimation, but as a linear interpolator, it has some difficulties in estimating extreme values. This study was performed using the S-plus software and necessary routines were written in this language or in FORTRAN.

References

Ambroise, C., Dang, M. and Govaert, G. (1997) Clustering of spatial data by the EM algorithm. In: *geoENV I - Geostatistics for Environmental Applications*, edited by Soares, A. *et al*. pp. 493-504. Dordrecht: Kluwer Academics Publishers.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering, *Biometrics*, 49, 803-821.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Monterey: Wadsworth and Brooks/Cole.

Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions, *Computational statistics and data analysis*, 14, 315-332.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models, *Pattern Recognition*, 28, 781-793.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Oliver, M. and Webster, R. (1989) A geostatistical basis for spatial weigthing in multivariate classification, *Mathematical Geology*, 21, 15-35.

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, Series B*, 59, 731-792.

Switzer, P., Kowalik, W. S. and Lyon, R. J. P. (1982) A prior probability method for smoothing discriminant analysis classification maps, *Mathematical Geology*, 14, 433-444.

Venables, W. N. and Ripley, B. D. (1994) *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Wackernagel, H. (1995) *Multivariate Geostatistics*. Berlin: Springer-Verlag.

JGIDA vol. 2, no. 2

JGIDA Home