A new approach in space-time analysis of multivariate hydrological data: Application to Brazil's Nordeste region rainfall

Emeline Sicard

Institut de recherche pour le développement (IRD), Montpellier, France

Robert Sabatier

Laboratoire de Physique Moléculaire et Structurale, UMR 5094, Faculté de Pharmacie, Montpellier, France

Hélène Niel and Eric Cadier

Institut de recherche pour le développement (IRD), Montpellier, France

Received 29 April 2002; revised 16 September 2002; accepted 16 September 2002; published 28 December 2002.

[1] The objective of this paper is to implement an original method for spatial and multivariate data, combining a method of three-way array analysis (STATIS) with geostatistical tools. The variables of interest are the monthly amounts of rainfall in the Nordeste region of Brazil, recorded from 1937 to 1975. The principle of the technique is the calculation of a linear combination of the initial variables, containing a large part of the initial variability and taking into account the spatial dependencies. It is a promising method that is able to analyze triple variability: spatial, seasonal, and interannual. In our case, the first component obtained discriminates a group of rain gauges, corresponding approximately to the Agreste, from all the others. The monthly variables of July and August strongly influence this separation. Furthermore, an annual study brings out the stability of the spatial structure of components calculated for each year. *INDEX TERMS:* 1854 Hydrology: Precipitation (3354); 1833 Hydrology: Hydroclimatology; 3299 Mathematical Geophysics: General or miscellaneous; *KEYWORDS:* Brazil, Nordeste, rainfalls, STATIS method, variogram, SCM

Citation: Sicard, E., R. Sabatier, H. Niel, and E. Cadier, A new approach in space-time analysis of multivariate hydrological data: Application to Brazil's Nordeste region rainfall, *Water Resour. Res.*, *38*(12), 1319, doi:10.1029/2002WR001413, 2002.

1. Introduction

[2] Many situations such as rainfall measurement involve multivariate data and include a spatial or temporal feature: individuals and/or variables are then linked by a relationship of spatial and/or temporal proximity and cannot be considered as independent. In the case of spatial and multivariate data, the aim of statisticians is usually to produce maps of the phenomenon. In the multivariate context, one mean is to produce a separate map per variable, for instance using an estimation by cokrigeage (by a linear estimator that minimizes the estimation variance) [*Wackernagel*, 1998], but the interpretation of numerous maps can be delicate. Another technique involves reducing the dimension of the multivariate space by means of the calculation of a small number of new variables that reflect the spatial and multivariate phenomenon.

[3] The purpose of this paper is to implement an original method for spatial and multivariate data, in order to describe the variability of rainfall in the Nordeste region of Brazil, which is a vast zone that represents 20% of Brazil's surface area. In this region, rainfall is characterized by pronounced spatial and temporal variability, resulting from complex climatic phenomena. More precisely, we focus on the part of the Nordeste region called the drought polygon. It covers 950,000 square kilometers and is characterized by a mean annual rainfall between 400 and 800 mm [*da Cunha*, 1902]. Two physiographic zones are distinguishable in this region: the "Sertão", a semi-arid zone where rainfall is low and irregular, with alternately torrential rains and strong droughts, and the "Agreste", a transitional zone between the Sertão and a narrow and humid coastal strip with a tropical humid climate, which does not belong to the drought polygon [*Cadier*, 1993, 1996]. Annual precipitation variability is unusually high at these latitudes. Hence it is important to characterize the stable features in the rainfall structure using a space-time analysis approach.

[4] The study is based on a set of monthly rainfall records over several years in different meteorological stations. The context is a multivariate, spatial, and temporal situation on a monthly scale.

[5] All calculations are made with S-Plus® 6.0 [*Venables and Ripley*, 1999; *Mathsoft*, 2000]. Calculations of variograms and kriging are made using the specialized add-on module S+SpatialStats [*Kaluzny et al.*, 1997]. Graphs and programs that are not included in this paper are available from the authors on request.

2. Data Description and First Analyses

2.1. Climatic Context

[6] Since *da Cunha* [1902] several authors have dealt with Nordeste rainfall, identified as a decisive factor for

0043-1397/02/2002WR001413

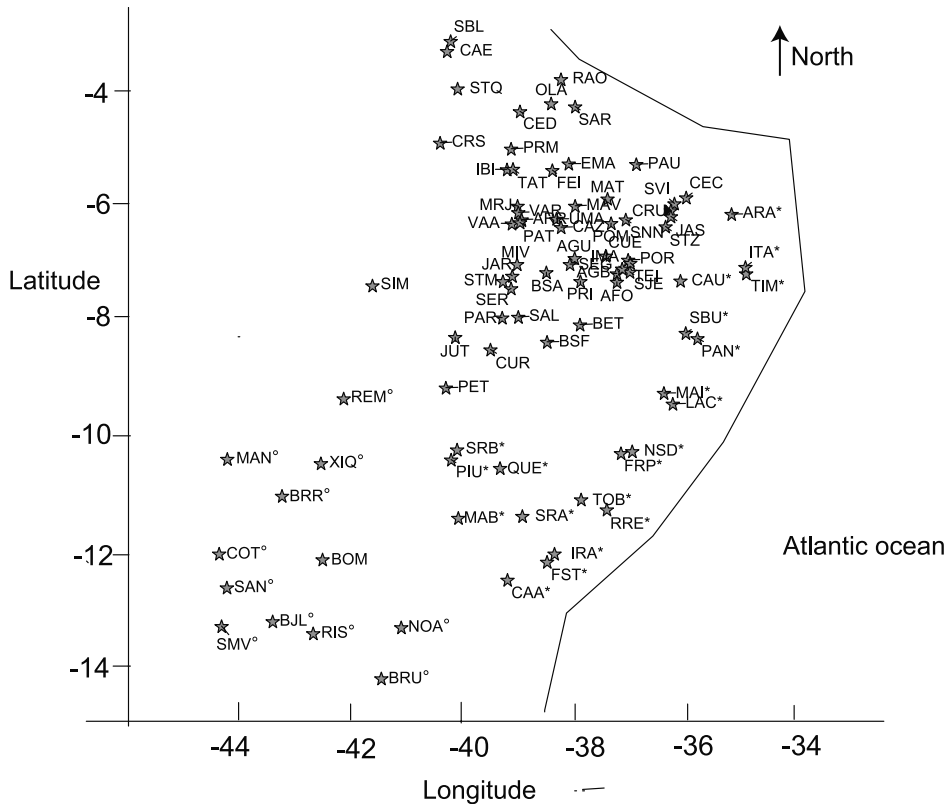

Figure 1. Map of the 82 stations of the Nordeste region retained for our analysis. Stations are mentioned by their three letter abbreviations, listed in the work of *Cadier* [1993]. Longitudes and latitudes are in decimal degrees. XXX* means that the station belongs to the Agreste region and XXX° means that the station belongs to the southwest of the Sertão.

the development and even the survival of this region. The climatic mechanisms are complex [*Nimer*, 1973; *Kouski and Moura*, 1981] but the majority of authors describe the substantial irregularity in occurrence and duration of the dry events, attempting to establish links with global climate indicators such as El Niño-Southern Oscillation (ENSO), Atlantic Ocean oscillations and Sea Surface Temperature Anomalies (SSTA), Intertropical Convergence Zone (ICTZ), or Polar air incursion occurrences, with the final objective to try to find explanations for these anomalies and, if possible, to predict them [*Hastenrath*, 1990; *Hastenrath and Greischar*, 1993; *Folland et al*, 2001; *Chaves and Cavalcanti*, 2001; *Pezzi and Cavalcanti*, 2001]. These researches are presently supported by Climatic Research programs led in Brazil by INPE/CPTEC.

[7] However, few of the above-mentioned authors have tried to delimit zones according to the pluviometric regimes. *Chu* [1983] separates the Northern Nordeste and the Southern Nordeste, more influenced by the Southern Hemisphere's atmospheric circulation. In a similar approach, *Uvo et al.* [1998] propose a more detailed zoning (but varying every month) of the influence of climatic indicators.

[8] As far as we know, none of these authors mentions and characterizes the well-known physiographic zone of Agreste with the pluviometric criteria proposed by *Cadier* [1993] and by the present work. *Frankenberg and Rheker* [1988] use this zonation in their multivariate analysis of rainfall.

2.2. Data

[9] The data consist of monthly rainfall records at 82 stations in the Nordeste region over the period from 1937 to 1975. Data were not recorded for all consecutive years and the maximum gap is 5 years between 1957 and 1963. Stations are represented in Figure 1, where they are mentioned by their three letter abbreviation, listed in the work of *Cadier* [1993, appendix 3].

[10] For each year we have 12 monthly variables Pj ($j = 1 \ldots 12$), which are the total amounts of rainfall during the month j.

[11] The data set (Figure 2) is sorted into $q = 29$ matrices $\mathbf{X}_k$ ($k = 1 \ldots q$) of dimension 82×12, whose lines are the $n = 82$ stations ($i = 1 \ldots n$) and whose columns are the $p = 12$ variables ($j = 1 \ldots p$) described previously. Each matrix corresponds to a year of observation. $\mathbf{D}$ is the diagonal matrix giving the weight of each station. For our study, this matrix is set to $\mathbf{D} = (1/n)\mathbf{Id}_n$. ($\mathbf{Id}_n$ being the $n \times n$ identity matrix). All the variables are then centered and reduced according to $\mathbf{D}$.

[12] Histograms of the variables, all years combined or not, show a heavily skewed distribution due to the large number of low values in comparison with high values. We decided not to make any transformation, because neither PCA nor kriging predictions require any assumption of normality to function, even if this can result in an unstruc-

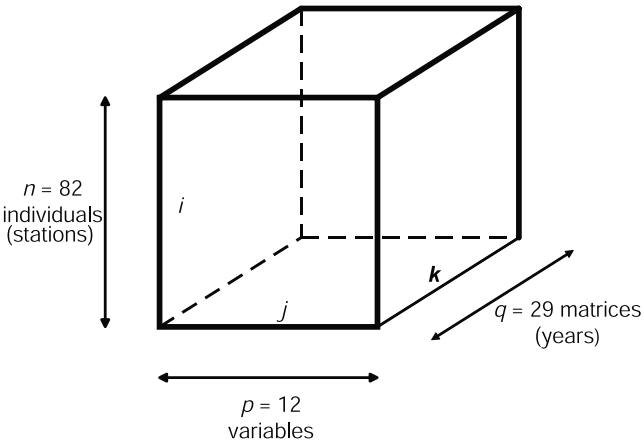

Figure 2. The three-way data array.

tured variogram, with relatively high values for every lag class, even for the short lags near the origin. In our case, the variograms obtained in our further analysis are satisfactory.

[13] A short analysis of temporal stationarity was done on the annual amounts of rainfall per station. The procedures used [*Lee and Heghinian*, 1977; *Pettitt*, 1979; *Hubert et al.*, 1989] aimed to detect trends and breakpoints. It was not carried out monthly, but the results for the annual level show that for most of the stations (51 stations out of 82), the annual amounts are random. For the other stations, no clear tendency can be revealed that would reflect a climatic change on a global scale. Considering these satisfactory results, it is possible to study the temporal stability of the distances between stations.

3. General Methodology in Space-Time Multivariate Analysis

3.1. Overview of Methods for Spatial and Multivariate Data

[14] Principal Component Analysis (PCA) [*Mardia et al.*, 1979] is a very widely used method of multivariate data analysis that makes it possible to transform a set of correlated variables into uncorrelated quantities, extracting a maximal amount of variance from the data. It is used by many authors in a hydrological context; see for example *Domroes et al.* [1998] in the case of precipitation variables. However, this method does not take into account the locations of the individuals.

[15] Local factor analyses are derivatives of this method that introduce a neighborhood matrix between samples [*Meot et al.*, 1993; *Chessel and Sabatier*, 1994; *Sabatier*, 1998]. They are based on the definition of a linear combination of the variables that maximizes the local variance, calculated by means of the neighborhood matrix.

[16] An alternative, called factorial kriging, stems from multivariate geostatistics [*Wackernagel*, 1998; *Arnaud et al.*, 2001; *Goovaerts et al.*, 1993]: it is a generalization of factor analysis in a spatial framework that incorporates the idea of continuity of spatial links. The resulting factors are not correlated spatially and each of them is related to a definite spatial structure.

[17] The STATIS method, for the French expression "Structuration des Tableaux à Trois Indices de la Statistique" [*Lavit*, 1988; *Lavit et al.*, 1994; *Meyners et al.*, 1998], provides a means to deal with three-way arrays. It can then be used when several data arrays have been measured at different points in time in order to deal with the temporal feature, but without taking into account the notion of temporal proximity. It is based on the calculation of a "consensus matrix," that is a synthesis of all the matrices. Many other generalizations of standard multivariate analysis such as principal component analysis (PCA) or canonical correlation Analysis (CA) have been proposed to study three or more sets of variables, i.e., a multi-way table. Most of these methods determine the dimension (or rank) of the model step by step, using linear combinations of each set of variables and optimizing one criterion [*Carroll*, 1968; *Gower*, 1975; *Escofier and Pagès*, 1984]. STATIS is the only simple method that takes into account the study of an interstructure and that introduces the idea of a consensus matrix.

[18] In our case, where data consist of variables measured at different stations over different years, we introduce here an original methodology associating the STATIS method and the geostatistical kriging method. This idea was initiated by *Cornillon and Sabatier* [1999]. In this methodology the principal components obtained by the STATIS method are slightly modified to fit the theoretical variogram that was chosen to model them, making it possible to improve these components by taking into account spatial dependencies at the same time as their calculation.

3.2. STATIS Method

[19] We expose here the theory of STATIS as described by *Lavit* [1988]. It focuses on the study of the $n \times n$ Escoufier matrix $\mathbf{W}_k$ of scalar products between stations, associated with each matrix $\mathbf{X}_k$, $k = 1 \ldots q$:

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{X}_k^t \tag{1}$$

(where superscript ^{t} implies a transposed matrix), by means of the Hilbert-Schmidt scalar product between the Escoufier matrices:

$$(\mathbf{W}_k \mid \mathbf{W}_{k'})_{HS} = \mathrm{tr}(\mathbf{D}\mathbf{W}_k\mathbf{D}\mathbf{W}_{k'}) \tag{2}$$

where tr stands for the trace of a matrix. All the scalar products between the $\mathbf{W}_k$ are the entries of the $q \times q$ matrix $\mathbf{S}$: $S_{k,k'} = (W_k|W_{k'})_{HS}$, $k,k' = 1 \ldots q$. The main idea is the calculation of a $n \times n$ "consensus matrix" which summarizes the initial ones. It is a weighted average of the initial normed Escoufier matrices:

$$\mathbf{W}_c = \sum_{k=1}^{q} \alpha_k \frac{\mathbf{W}_k}{\| \mathbf{W}_k \|_{HS}} \tag{3}$$

with $\|\mathbf{W}_k\|_{HS}^2 = (\mathbf{W}_k|\mathbf{W}_k)_{HS}$.

[20] The coefficients α_k are sought to maximize $\|\mathbf{W}_c\|_{HS}^2 = (\mathbf{W}_c|\mathbf{W}_c)_{HS}$: they are the terms of the vector α defined as:

$$\alpha = \lambda_1 \mu_1 \tag{4}$$

where μ_1 is the eigenvector of $\mathbf{S}$ associated with the largest eigenvalue λ_1. The meaning of the consensus matrix is

checked by the representation of the ''interstructure'', by associating each matrix $\mathbf{W}_k$ with a point in Euclidian space, whose coordinates on the axis i are the components of the vector $\sqrt{\lambda_i}\mu_i$. This makes it possible to appreciate the similarity between all the arrays, to show if there is actually a common structure, represented by the consensus matrix.

[21] The STATIS method makes it possible to turn a three-way array into a two-way consensus array, which enables the application of a PCA, i.e., the diagonalization of $\mathbf{W}_c$. It can be shown that this diagonalization is equivalent to the PCA of the $n \times pq$ matrix $\mathbf{Y}$, defined by the juxtaposition of the initial matrices $\mathbf{X}_k$ multiplied by the annual coefficients $\sqrt{\alpha_k}$. The Escoufier matrix (equation (1)) corresponding to $\mathbf{Y}$ is $\mathbf{W}_c$. This PCA makes it possible to obtain a consensus Euclidian representation of the individuals.

3.3. Spatialized Component Method (SCM)

[22] Once the consensus matrix is calculated the same context as that of *Cornillon and Sabatier* [1999] is reproduced, which allows us to use their methodology.

[23] We consider the $n \times pq$ matrix $\mathbf{Y}$ whose corresponding Escoufier matrix is $\mathbf{W}_c$. Y_j, the j^{th} column of $\mathbf{Y}$, then corresponds to the initial j^{th} variable of $\mathbf{X}_k$ weighted by $\sqrt{\alpha_k}$, and is a regionalized variable, realization of the random function $Y_j(x)$. Hence we have a pq-valued spatial process $\{\mathbf{Y}(x) = Y_1(x). . . Y_{pq}(x), x \in \mathbb{R}^2\}$ defined over points locations x within the spatial domain $\mathbb{R}^2$.

[24] The main idea of the SCM consists in reducing the dimension of the variable space, by calculating a linear and one-dimensional component that contains a maximum amount of the variance of the data, and that takes into account the spatial dependence between stations. This is done by simultaneously estimating the best component and its fitted variogram so that the residual sum of squares between its theoretical and its experimental variogram is minimized. To simplify, we use the term of variogram instead of the appropriate one semi-variogram [*Matheron*, 1969].

[25] Our goal is to find a linear combination $c(x)$ of these pq variables, whose variogram is as close as possible to the theoretical variogram chosen to fit it, such as:

$$c(x) = \sum_{j=1}^{pq} u_j Y_j(x) = \mathbf{Y}u \tag{5}$$

[26] We therefore must find the parameters u (vector of coefficients for the component $c(x)$) and θ (vector of the coefficients of the theoretical variogram $\gamma(\theta, h)$) that minimize the objective function:

$$\varphi_1(u, \theta) = \sum_h (\hat{\gamma}_u(h) - \gamma(\theta, h))^2 \tag{6}$$

under the constraint:

$$u'u = 1. \tag{7}$$

[27] In the previous equation, $\hat{\gamma}_u(h)$ is the experimental variogram, computed as half the average squared difference between the components of data pairs [*Goovaerts*, 2000]:

$$\hat{\gamma}_u(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [c(x_\alpha) - c(x_\alpha + h)]^2 \tag{7}$$

where $N(h)$ is the number of pairs of data locations a vector h apart. The vectors h are determined empirically by adjusting the azimuth tolerance and the number and the length of the lags, so that each point of the variogram is estimated by a sufficient number of points.

[28] The technique used is iterative and is composed of two steps. In a first step, the starting point is chosen to be the first component $c_1 = Yu_1$ of the PCA of $\mathbf{Y}$. A theoretical variogram $\gamma(\theta, h)$ is then fitted, that is the reference variogram used afterwards.

[29] We then try, in a second step, to bring the reference variogram as close as possible to the experimental variogram stemming from the component, by conjointly modifying the coefficients of the theoretical variogram and those of the component.

[30] The SCM looks like the method of *Bailey and Krzanowski* [2000], but in the context of PCA instead of the context of factor analysis. The difference between the two techniques is that PCA is merely a transformation of data that gives components that are empirically determined aggregates of the variables without presumed theory, whereas factor analysis supposes that the data comes from a well-defined distributional model and gives factors that are theoretically the underlying variables that cause the covariation between observed variables [*Mardia et al.*, 1979].

[31] We wrote a minimization algorithm using the S-Plus® function *nlminb*. The S-Plus® function, which is minimized by *nlminb*, integrates the constraint (7):

$$\varphi_2(u, \theta) = \sum_h (\hat{\gamma}_{u/\|u\|}(h) - \gamma(\theta, h))^2 \tag{7}$$

[32] This is in fact the objective function (6) calculated at the normed vector u. The *nlminb* minimization is included in a loop that stops when the criterion is reached. The criterion concerns the relative variations between the steps (n) and $(n-1)$ of the objective function φ_2 and of the parameters of minimization u and θ:

$$(\varphi_2^{(n)} - \varphi_2^{(n-1)})/\varphi_2^{(n-1)} \leq \varepsilon$$

$$\sum_{i=1}^{pq} \left| \left(u_i^{(n)} - u_i^{(n-1)}\right)/u_i^{(n-1)} \right| + \sum_{i=1}^{2} \left| \left(\theta_i^{(n)} - \theta_i^{(n-1)}\right)/\theta_i^{(n-1)} \right| \leq \varepsilon$$

where superscript $^{(n)}$ implies the object evaluated at step (n), and $\varphi_2{}^{(n)}(u^{(n)}, \theta^{(n)})$ is noted $\varphi_2{}^{(n)}$ to simplify the expression. The parameter ε is set after several attempts to 10^{-10}.

4. Results

4.1. STATIS Method

[33] We perform the STATIS method on the 29 matrices described above. Plotting of the first Euclidian plane of the interstructure (Figure 3) shows a quite good similarity between the annual matrices, the first axis representing 99.05% of the total variance (this percentage being calculated by means of the eigenvalues λ of $\mathbf{S}$). Hence all the

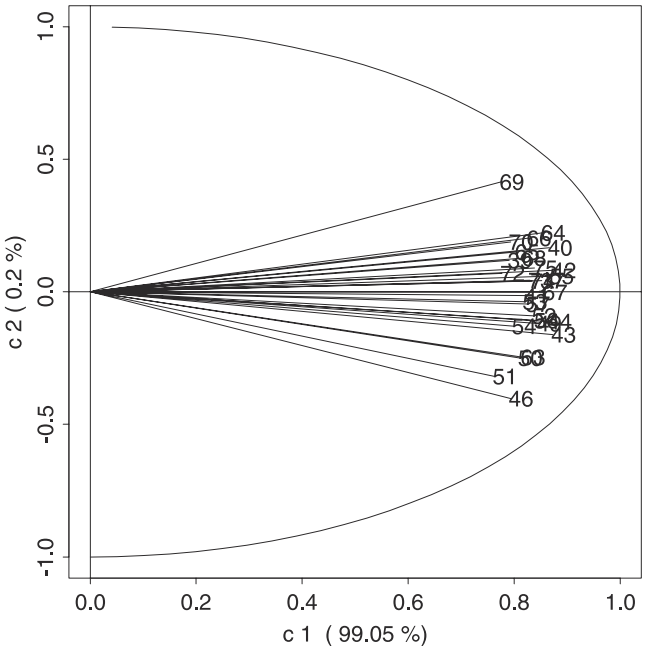

Figure 3. First plane of the interstructure obtained with the STATIS method. This is an approximated Euclidian image of the matrices $\mathbf{W}_k$, $k = 1\ldots29$, which are indicated by the two last digits of the corresponding year.

years seem sufficiently homogeneous to consider that they can be represented in a satisfactory way by a consensus matrix, even if some years (1946, 1950, 1951, 1963, 1969) can nonetheless be distinguished from the others, as they are quite isolated from the main group of years. This means that they have a slightly different structure.

[34] We can then interpret the diagonalization of **Y**. We consider only the first two components c_1 and c_2 which explain 50.16% of the total variance (a level of variance that is usual in this method), with the corresponding plane represented in Figure 4. The first component accounts for 28.38% of the variability and discriminates the stations corresponding approximately to the Agreste, which have negative coordinates, from all the others, characterized by positive coordinates. The second component accounts for 21.78% of the variability and discriminates the southwest stations from the northeast ones. The third component accounts for only 6.29% of the variability, and since we found no relevant interpretation, we decided not to study it.

[35] Interpretation of the correlation circle is rather difficult due to the large number of variables (12×29). We then choose to divide it by plotting one circle by type of variable, thus representing the correlations between axes and the considered variable over all years. Figure 5 shows the circles for three characteristic months: January, August, and November.

[36] The first axis seems to be strongly and negatively correlated with the monthly variables of months 7 (July: for 72% of the years the correlation is higher then 0.8 in absolute value) and 8 (August: 89%). It contrasts the stations of the Nordeste region, characterized by rainy months in July and August, with the other stations where these months are drier. The second axis is harder to interpret, because the correlations are not as stable with the years. It seems to be correlated with the variables $P11$ (November: for 58% of the years the correlation is higher then 0.8 in absolute value) and $P12$ (December: 41%) for many of the years. This means that there is a gradient of precipitation for the months November and December, from the southwest (high values) to the northeast (low values). In Figure 5, the isolated position of the year 1951 for November means that in November 1951 the gradient Agreste/Sertão was much larger than during the other years because of low rainfall in the Sertão, and that the gradient southwest/northeast was lower because of low rainfall in the southwest.

[37] The results of the STATIS method therefore show that in spite of the pronounced annual variability of precipitation in the considered region, some structure can be extracted so that we can work on an "average array" that summarizes all the years, without losing too much information.

[38] It is then possible to see in more detail why each initial matrix is different from the consensus matrix, by showing, in the consensus Euclidian representation, the coordinates of each individual of each year on c_1 and c_2. For instance, such a representation (not presented here) shows that for 1969, the coordinates on c_1 of most stations in the Agreste are far lower than the corresponding consensus coordinates (difference higher than 0.5 in absolute value). This explains the difference in structure pointed out in Figure 3. The amount of rain in August for this year for these stations was indeed low in comparison with the same stations for the other years.

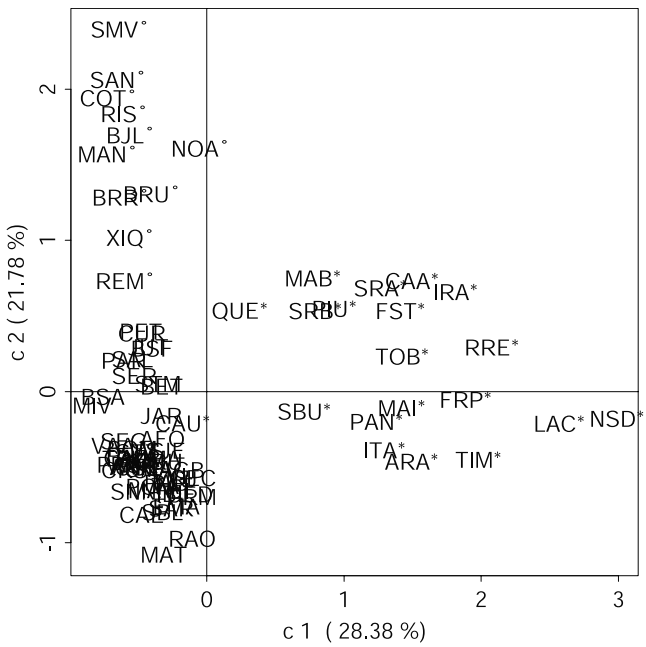

Figure 4. The first plane of the PCA of the consensus matrix **Y**, a "consensus" approximated Euclidian image of the stations, which are indicated by their three letter abbreviations listed in the work of *Cadier* [1993]. XXX* means that the station belongs to the Agreste region, and XXX° means that the station belongs to the southwest of the Sertão, as noted in Figure 1.

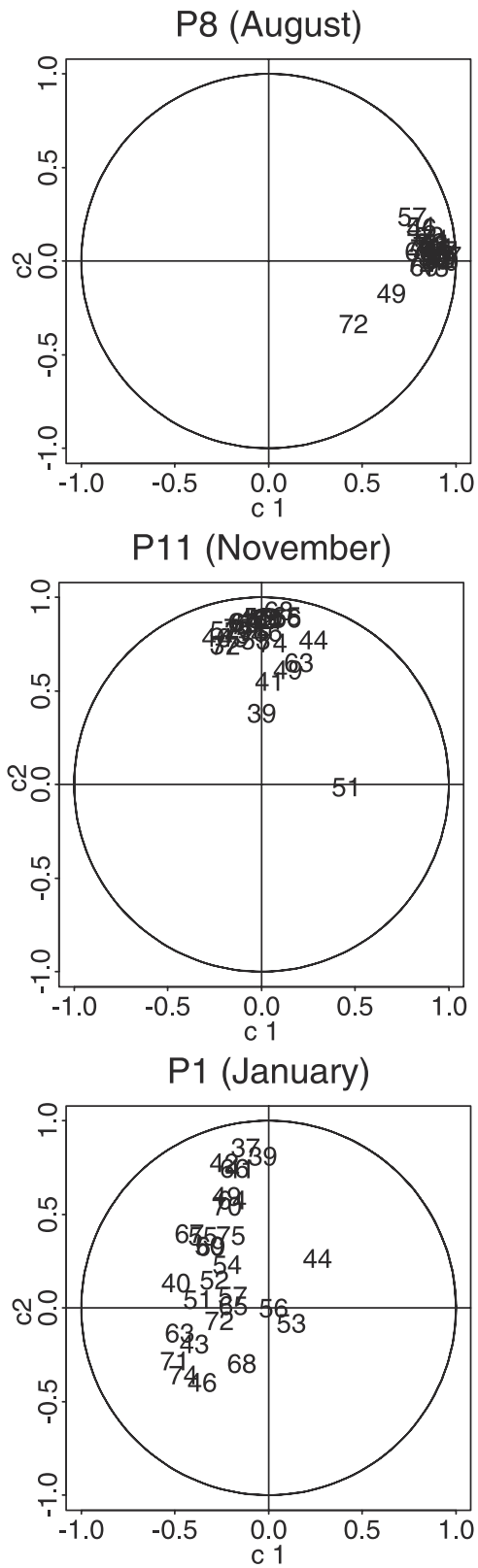

Figure 5. Correlation circles of the PCA of the consensus matrix **Y** for the variables of January (*P*1), August (*P*8) and November (*P*11). Years are mentioned by their two last digits. The first circle shows, for instance, the correlations between each axis of the PCA and the variable P8 of each year.

[39] Other records of daily rainfall had previously been used by *Cadier* [1993] who made a pluviometric zonation of daily rain events of short duration by fitting different laws of probability quantiles. Comparing the obtained zones and the results of the PCA of the consensus matrix shows that the stations contrasted by the first axis correspond exactly to three zones contrasted by *Cadier* [1993]. It then shows that our results with monthly variables are coherent with analysis of a different nature, using only daily variables. Therefore the STATIS method makes it possible to objectively find results using monthly variables. Moreover, the second component establishes a zonation that goes further than a simple distinction Agreste/Sertão.

[40] Thus the first components characterize different levels of variability and reveal a spatial zonation, even if no spatial information has been introduced. It therefore seems appropriate to take into account the spatial dependencies in the analysis in order to see what improvements would be made. This zonation will be made more precisely using kriging techniques.

4.2. Variogram Study

[41] The spatial structure of the first components resulting from the PCA of the consensus matrix is studied by calculating the corresponding experimental variograms. Directional variograms are estimated for the first two components (see Figure 6; only the first component is represented to simplify). Both first and second components present a zonal anisotropy, as in some directions the variogram is higher than in others, and has no sill, corresponding only to an intrinsically stationary random function.

[42] In the case of the first principal component that will be presented here, Figure 6 shows that the direction of main variability is the direction 135°, perpendicular to the direction of the lowest variogram, 45°. Angles are measured relative to the horizontal direction.

[43] It is therefore fitted by a nested variogram, defined as the sum of an isotropic structure, depending only on the

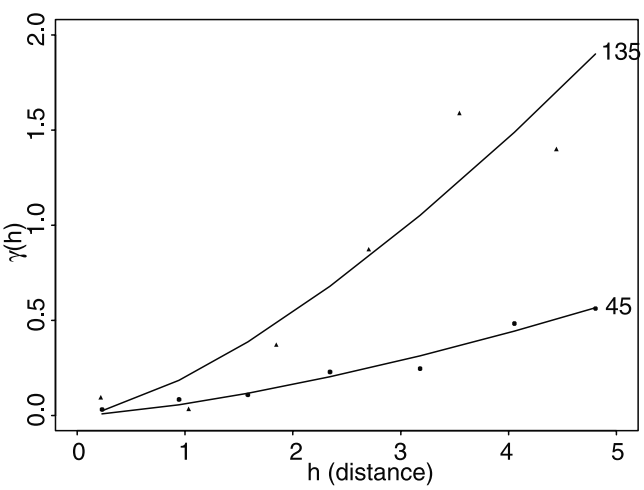

Figure 6. Experimental and fitted variograms for the first component of the consensus matrix PCA, in the directions 45° and 135° relative to the horizontal direction.

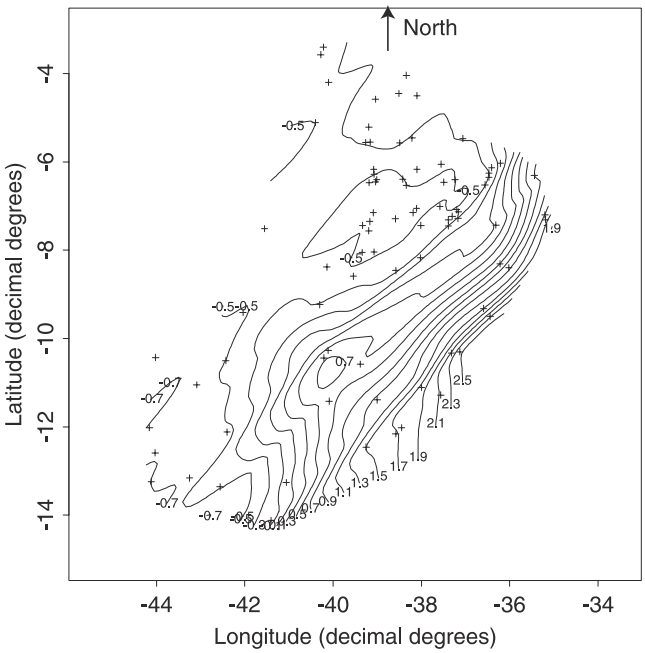

Figure 7. Map of the first component c_1 of the consensus matrix PCA (anisotropic model). Contour lines represent the values of c_1. Stations are represented by crosses.

distance $|h|$ between stations, and of an anisotropic structure, depending on the direction of the vector h in relation to the direction of zonality, here 135° [*Journel and Huijbregts*, 1978]. For both structures, we choose a power variogram. We add a nugget effect of 0.03.

[44] The variogram model for the first component can therefore be written as follows:

$$\gamma(\theta, h) = 0.03 + \gamma_{45}(\theta_{45}, h) + \gamma_{135}(\theta_{135}, h_{135}) \qquad (10)$$

where

$$\gamma_{45}(\theta_{45}, h) = a_{45} h^{b_{45}}$$

$$\gamma_{135}(\theta_{135}, h_{135}) = a_{135} h_{135}^{b_{135}}$$

$$\theta = (a_{45} \quad b_{45} \quad a_{135} \quad b_{135})$$

and where h_{135} is the coordinate of the vector h along the direction 135°, a_{45} and a_{135} are, respectively, the sills in the directions 45° and 135°, and b_{45} and b_{135} are the ranges in the directions 45° and 135° ($0 < b < 2$). For $b \geq 2$, $-\gamma(h)$ is no longer a conditional positive definite function [*Journel and Huijbregts*, 1978].

[45] All coefficients are fitted by nonlinear least squares regression (function *nls* of S-Plus®): θ = (0.041 1.642 0.125 1.491) (see the corresponding fitted variogram in Figure 6).

[46] We perform a cross-validation as described by *Wackernagel* [1998]. It consists in an estimation of each measure point by means of the fitted model, using the $(n - 1)$ other points. The distribution of cross-validation errors looks like a normal distribution, with mean 0 and standard deviation 1, and the absolute cross-validation errors are all below the threshold of 1.96 except for 2 stations: Araruna (ARA) and Nossa Senhora das Dores (NSD). Both are stations of the Agreste. This can be explained for ARA by the relatively isolated location of this station, and for NSD by its extreme value, as it corresponds to the maximum for c_1.

[47] The corresponding map is obtained by ordinary kriging [*Wackernagel*, 1998], represented in Figure 7. As the module S+SpatialStats® cannot compute kriging predictions with nested variograms, we wrote a new S-Plus® program in order to take this aspect into account, using the method of resolution of kriging systems given by *Cressie* [1993].

[48] The link between the STATIS method and ordinary kriging produces a map for the component that accounts for most of the variance from the data. Other maps (not included) for the other components have been made in the same way to describe the main rainfall features.

4.3. Application of SCM

[49] In order to avoid complexities associated with the anisotropy, we choose to work with an omnidirectional variogram and so to use as the reference variogram the theoretical variogram $\gamma(\theta, h)$ determined without taking anisotropy into account. We choose a power-variogram without nugget effect:

$$\gamma(\theta, h) = a \mid h \mid^b \qquad (11)$$

with $\theta = (ab)$, a and b being fitted by *nls* minimization: θ_0 = (0.115 1.777). The initial objective is φ_1 (u_1, θ_0) = 0.046.

[50] The "isotropic" model (equation (11)) is a little worse in terms of cross-validation, because with the criterion used before, six stations instead of two in the anisotropic case (equation (10)) can be distinguished from the others, all of them belonging to the Agreste and therefore having high values for c_1. For our purposes, we consider

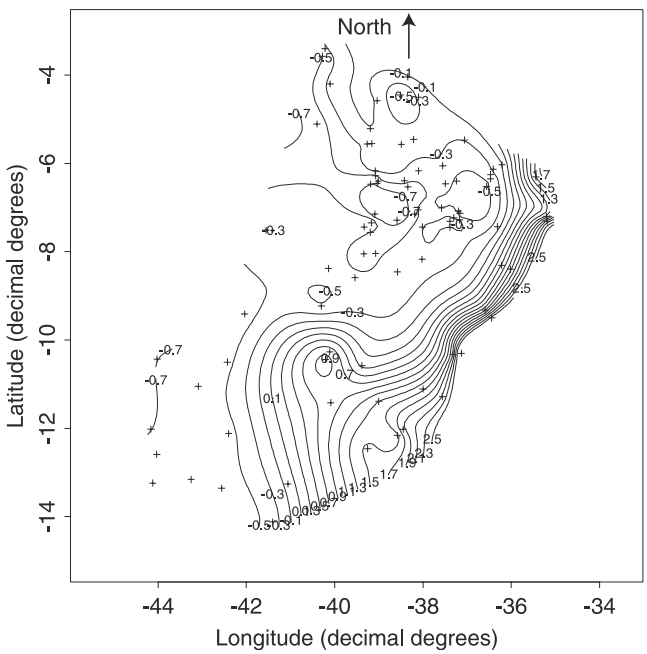

Figure 8. Map of the first component c_1 of the consensus matrix PCA (isotropic model). Contour lines represent the values of c_1. Stations are represented by crosses.

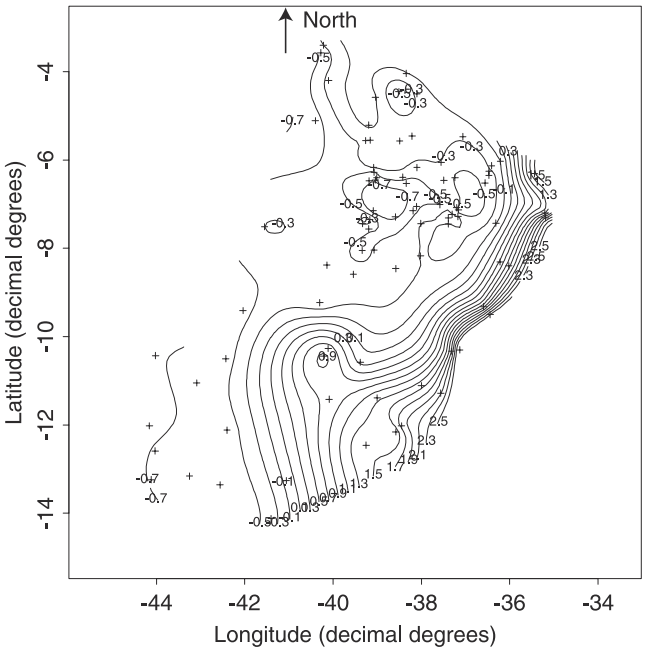

Figure 9. Map of the fitted component c_f (isotropic model). Contour lines represent the values of c_f. Stations are represented by crosses.

that this isotropic model is satisfactory. The corresponding map is represented in Figure 8.

[51] The optimization program stops after 9 iterations when θ_f = (0.108 1.771) and $\varphi_1(u_f, \theta_f)$ = 0.038. It corresponds to a decrease of 17.7% in the objective function. u_f is the final vector u making it possible to calculate c_f: $c_f = \mathbf{Y}u_f$ (equation (5)). The variation of the vector u is measured by the linear correlation coefficient between u_1 and u_f (cor(u_1, u_f) = 0.965) and between c_1 and c_f (cor(c_1, c_f) = 0.999). It has therefore deviated little from the PCA results.

[52] The corresponding map is represented in Figure 9. It is quite similar to the initial one (Figure 8). The advantage is that this map has been calculated using a variogram model that is much better than the models used before.

5. Return to the Annual Matrices

[53] We decide to return to the initial annual matrices in order to see if the component calculated by means of the consensus matrix is spatially consistent with the components that could have been extracted from each year. On a single graph (Figure 10), we plot the experimental omnidirectional variograms of the first component resulting from the Principal Component Analysis of each initial matrix (each one corresponding to a year). All of them are quite similar. This means that there is some stability in the spatial structure. This should be compared with the plotting of the experimental variograms of the columns of **Y** (not shown here, but similar to Figure 10), after a correction to scale them to the same variance as c_f, grouped together by month: the variograms corresponding to months 5 (May) to 9 (September) are rather homogeneous, unlike those of the other months. Moreover, the annual variograms corresponding to these months are quite similar to the consensus variogram of the first component. As the first component of each year and of the consensus matrix are mainly correlated with these monthly variables, and particularly to months 7 (July) and 8 (August), the observed homogeneity can be accounted for. PCA homogenized the spatial structure of the variables.

6. Conclusions

[54] The use of the STATIS method in our context made it possible to synthesize a large number of variables, leading to components that characterize the climatic regime. These synthetic components characterize spatial zonations in an objective way, using monthly variables. The first one was transformed in order to obtain a component rich in terms of

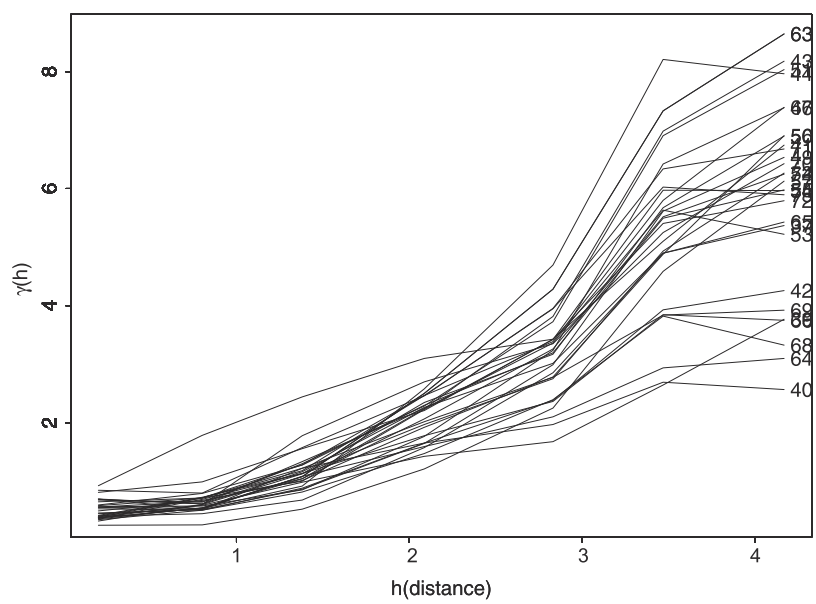

Figure 10. Experimental variograms for the first component resulting from the PCA of each yearly matrix. Years are mentioned by their two last digits.

explicative variance and with a well-determined spatial structure.

[55] The SCM presented here differs from that of *Bailey and Krzanowski* [2000] in that they do not specify using an initial point for their optimization. The initial point and the reference variogram used in our study corresponding to the first component of a PCA, we obtain a component with a variance close to the variance of this first component. Moreover, we introduce the notion of time, involving a three-way array. Of course, the method requires further improvements in order to study the convergence behavior of our algorithm, to establish links between this method and others such as factorial kriging, and to include the calculation of other components in order to describe a larger part of the variability.

[56] In spite of the zonal anisotropy discussed above, applying this method forced us to simplify our problem, since this anisotropy required a nested model that could not be easily handled by our method. It would then be of interest to incorporate this feature, in order to improve the final map.

[57] This paper shows then that using only monthly amounts of rainfall, irregularly located in the Nordeste region and low in number, it is possible to characterize a spatial, seasonal and annual variability. Spatial variability was brought out by geographical zoning that makes it possible to choose a small number of stations to represent a large part of the main pluviometric regimes. We were able to identify and characterize the pluviometric regimes of the Sertão, the Agreste and the southwestern zone. Then months could be grouped by season to characterize seasonal variations. We were able to identify the more relevant months to bring out spatial variability. Finally, plotting the interstructure made it possible to distinguish years that have a different spatial and seasonal structure in a region where water resources are very irregular and often dramatically insufficient.

[58] We thus have an original method that is able to analyze triple variability: spatial, seasonal and interannual. The results are coherent with what is already known, but at the same time reveal new aspects: it makes it possible to draw maps, to study more precisely correlations and interannual variations. They are also coherent with other attempts, not published here, using other types of variables such as the number of rainy days or the length of rainy sequences. It is a promising method that is to be used on variables other than pluviometric variables such as, for example, Ocean SST. The components obtained could be used in a further step to establish correlation links with the explicative variables from the Atlantic or Pacific Ocean.

References

Arnaud, M., X. Emery, C. Fouquet, M. Brouwers, and M. Fortier, L'analyse krigeante pour le classement d'observations spatiales et multivariées, *Rev. Stat. Appl.*, *49*(2), 45–67, 2001.

Bailey, T. C., and W. J. Krzanowski, Extensions to spatial methods with an illustration in geochemistry, *Math. Geol.*, *32*(6), 657–682, 2000.

Cadier, E., Hydrologie des petits bassins du Nordeste brésilien semi-aride: transposition à des bassins non étudiés, études et thèses, 414 pp., Institut Français de Recherche Scientifique pour le Développement en Coopération (ORSTOM), Paris, 1993.

Cadier, E., Small watershed hydrology in semi-arid north-eastern Brazil: Basin typology and transposition of annual runoff data, *J. Hydrol.*, *182*, 117–141, 1996.

Carroll, J. D., Generalization of canonical correlation analysis to three or more sets of variables, *Proc. 76th Conv. Am. Psychol. Assoc.*, *76*(3), 227–228, 1968.

Chaves, R. R., and I. F. A. Cavalcanti, Atmospheric circulation features associated with rainfall variability over Southern Northeast Brazil, *Mon. Weather. Rev*, *129*(10), 2614–2626, 2001.

Chessel, D. and R. Sabatier, Couplage de triplets statistiques et graphes de voisinage, in *Biométrie et Analyse des Données Spatio-temporelles*, edited by J. D. Lebreton and B. Asselain, pp. 58–37, Ecole Natl. Supérieure d'Agron., Rennes, France, 1994.

Chu, P.-S., Diagnostic studies of rainfall anomalies in northeast Brazil, *Mon. Weather Rev.*, *111*(8), 1655–1664, 1983.

Cornillon, P. A. and R. Sabatier, Analyse sur composante spatialisée, XXXIème jour. de stat., pp. 957–960, Soc. Française de Stat., Grenoble, France, 1999.

Cressie, N. A. C., *Statistics for Spatial Data*, revised ed., John Wiley, New York, 1993.

da Cunha, E., *Os Sertões*, Laemmert and C. Editores, Rio de Janeiro, Brazil, 1902.

Domroes, M., M. Kaviani, and D. Schaefer, An analysis of regional and intra-annual precipitation variability over Iran using multivariate statistical methods, *Theor. Appl. Clim.*, *61*(3–4), 151–159, 1998.

Escofier, B. and J. Pagès, L'analyse factorielle multiple: Une méthode de comparaison de groupes de variables, in *Data Analysis and Informatics III*, edited by E. Diday, pp. 41–55, Elsevier Sci., New York, 1984.

Folland, C. K., A. W. Colman, D. P. Rowell, and M. K. Davey, Predictability of northeast Brazil rainfall and real-time forecast skill, 1987–1998, *J. Clim.*, *14*(9), 1937–1958, 2001.

Frankenberg, P., and J. R. Rheker, Zum Niederschlagsregime in Nordostbrasilien, insbesondere in Pernambuco, *Jahrb. Geogr. Gesellschaft Hannover*, 65–96, 1988.

Goovaerts, P., Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *J. Hydrol.*, *228*, 113–129, 2000.

Goovaerts, P., P. Sonnet, and A. Navarre, Factorial kriging analysis of springwater contents in the Dyle River Basin, Belgium, *Water Resour. Res.*, *29*(7), 2115–2125, 1993.

Gower, J., Generalized procrustes analysis, *Psychometrika*, *40*(1), 33–51, 1975.

Hastenrath, S., Prediction of northeast Brazil rainfall anomalies, *J. Clim.*, *3*(8), 893–904, 1990.

Hastenrath, S., and L. Greischar, Circulation mechanisms related to northeast Brazil rainfall anomalies, *J. Geophys. Res.*, *98*, 5093–5102, 1993.

Hubert, P., J. P. Carbonnel, and A. Chaouche, Segmentation des séries hydrologiques: Application à des séries de précipitations et de débits de l'Afrique de l'Ouest, *J. Hydrol.*, *110*, 349–367, 1989.

Journel, A. G. and C. J. Huijbregts, *Mining Geostatistics*, Academic, San Diego, Calif., 1978.

Kaluzny, S. P., S. C. Vega, T. P. Cardoso, and A. A. Shelly, *S+SpatialStats*, Springer-Verlag, New York, 1997.

Kouski, V. E. and A. D. Moura, Previsão de precipitação no Nordeste do Brazil: o aspecto dinâmico, *2244, PRE/029*, Natl. Inst. for Space Res. (INPE), São José dos Campos, Brazil, 1981.

Lavit, C., *Analyse Conjointe de Tableaux Quantitatifs*, Masson, Paris, 1988.

Lavit, C., Y. Escoufier, R. Sabatier, and P. Traissac, The ACT (STATIS) method, *Comput. Stat. Data Anal.*, *18*, 97–119, 1994.

Lee, A. F. S., and S. M. Heghinian, A shift of the mean level in a sequence of independent normal random variables—A Bayesian approach, *Technometrics*, *19*(4), 503–506, 1977.

Mardia, K. V., J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic, San Diego, Calif., 1979.

Matheron, G., Le krigeage universel, *Les Cahiers 1*, 83 pp., Cent. de Morphol. Math., Fontainebleau, France, 1969.

MathSoft, S-PLUS 6.0 for UNIX programmer's guide, Data Anal. Div., Seattle, Wash., 2000.

Meot, A., D. Chessel, and R. Sabatier, Opérateurs de voisinage et analyse des données spatio-temporelles, in *Biométrie et Environnement*, edited by J. D. Lebreton and D. Asselain, pp. 45–71, Masson, Paris, 1993.

Meyners, M., J. Kunert, E. M. Qannari, and P. B. Brockhoff, Comparing generalized procustes analysis and STATIS, *Food Qual. Pref.*, *11*(1–2), 77–83, 1998.

Nimer, E., Climatologia da região Nordeste do Brasil, *Rev. Brasil. Geogr.*, *34*(2), 3–51, 1973.

Pettitt, A. N., A non-parametric approach to the change-point problem, *Appl. Stat.*, *28*(2), 126–135, 1979.

Pezzi, L. P., and I. F. Cavalcanti, The relative importance of ENSO and tropical Atlantic sea surface temperature anomalies for seasonal precipi-

tation over South America: A numerical study, *Clim. Dyn.*, *17*(2/3), 205–212, 2001.

Sabatier, R., Analyse en composantes principales d'observations spatialisées, *Oceanis*, *34*(3), 37–53, 1998.

Uvo, C. B., C. A. Repelli, S. E. Zebiac, and Y. Kushnir, The relationships between tropical Pacific and Atlantic SST and northeast Brazil monthly precipitation, *J. Clim.*, *11*(4), 551–562, 1998.

Venables, W. N. and B. D. Ripley, *Modern Applied Statistics With S-Plus*, 3rd ed., Springer-Verlag, New York, 1999.

Wackernagel, H., *Multivariate Geostatistics*, 2nd ed., Springer-Verlag, New York, 1998.

E. Cadier, H. Niel, and E. Sicard, (IRD), B.P. 64501, 34394 Montpellier cedex 5, France. (emeline.sicard@msem.univ-montp2.fr)

R. Sabatier, Laboratoire de Physique Moléculaire et Structurale, UMR 5094, Faculté de Pharmacie, 15 av. Ch. Flahault, B.P. 14491, 34093, Montpellier cedex 5, France.