Revista de Climatología

Vol. 20 (2020): 1-13 ISSN 1578-8768

Comparison of distance methods for detection of atypical observations in monthly precipitation series

Danny Villegas Rivas[1], Manuel Milla Pino[2], Yary Pérez Pérez[3], Salli Villegas Rivas[4], Oscar Gamarra Torres[5], Víctor Carril Fernández[6], Ricardo Shimabuku Ysa[7]

1. Facultad de Ingeniería Forestal y Ambiental. Universidad Nacional de Jaén, Cajamarca, Perú. *<danny_villegas1@yahoo.com>*

2. Facultad de Ingeniería Civil. Universidad Nacional de Jaén, Cajamarca, Perú.

3. Programa Nacional de Formación Agroalimentaria. Universidad Politécnica Territorial de Portuguesa 'JJ Montilla', Portuguesa, Venezuela.

4. Programa de Ciencias Sociales. Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora, Portuguesa, Venezuela.

5. Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES), Amazonas, Perú.

6. Universidad Nacional de Jaén, Cajamarca, Perú.

7. Facultad de Ingeniería Mecánica y Eléctrica. Universidad Nacional de Jaén, Cajamarca, Perú.

(Received: 29-Ene-2020. Published: 24-Mar-2020)

Abstract

In this paper, methods based on multivariate distances for the detection of atypical observations in monthly precipitation series from a meteorological station and a simulation study were compared with three models of extreme events and a linear model with harmonics and autoregressive AR(1) errors for different periods of time. The precipitation in San Cristobal, Venezuela is not conditioned by times of drought and rain, showing non-perfect symmetry without long tails, normality, seasonality, high variability, atypical observations, an annual cycle with a well-defined maximum in June, autocorrelated residuals, in general typical characteristics of the tropical dry forest, and a tendency to be distributed as a pearson III. The mahalanobis distance reported the best results in relation to the percentage of atypical observations detected for periods of 5 and 10 years and a linear model with two harmonics with real data, while for periods greater than 10 years the Lognormal model showed a trend similar to the series studied that stabilized as the period increased. The euclidean distance showed a behavior similar to that obtained with mahalanobis in a period of 5 years and a lognormal distribution, while for periods greater than 5 years the percentage of atypical observations increased significantly as happened with the other models for periods longer than 5 years. The Manhattan distance showed an increase in the percentage of atypical observations. An overestimation of the amount of atypical observations with the euclidean and manhattan distances was evidenced, presuming a detrimental effect of the serial autocorrelation of the residuals on these two distances. The existence of a potential bogging effect was observed, with fractions of atypical observations greater than $1/(n+1)$, in periods greater than 5 years. The lognormal distribution over a period of 5 years had a favorable effect on the euclidean and manhattan distances, and on that of mahalanobis in periods greater than 10 years, while the linear models with two harmonics for periods less than or equal to 10 years showed a positive effect on the mahalanobis distance.

Key words: Precipitation, outliers, distance, multivariate analysis.

Resumen

En este trabajo se compararon métodos basados en distancias multivariadas para la detección de observaciones atípicas en series de precipitación mensual provenientes de una estación meteorológica y de un estudio de simulación con tres modelos de eventos extremos y un modelo lineal con armónicos y errores autorregresivos AR(1) para distintos períodos de tiempo. Los resultados mostraron que la precipitación en San Cristobal, Venezuela no está condicionada por las épocas de sequía y lluvia, evidenciando simetría no perfecta sin largas colas, normalidad, estacionalidad, alta variabilidad, observaciones atípicas, un ciclo anual con un máximo bien definido en junio, residuos autocorrelacionados, en general características típicas del bosque seco tropical, y una tendencia a distribuirse como una Pearson III. En ese orden, de las tres metodologías multivariadas, la distancia de Mahalanobis reportó los mejores resultados en relación al porcentaje de observaciones atípicas detectadas para periodos de 5 y 10 años y un modelo lineal con dos armónicos con datos reales, mientras que para periodos mayores a 10 años el modelo lognormal mostró una tendencia similar a la serie estudiada que se estabilizó conforme se incrementó el periodo. De igual forma, la distancia Euclídea mostró un comportamiento similar al obtenido con la de Mahalanobis en un periodo de 5 años y una distribución lognormal, mientras que para períodos mayores a 5 años el porcentaje de observaciones atípicas se incrementó significativamente al igual que ocurrió con los demás modelos para periodos mayores a 5 años. De la misma manera, con la distancia Manhattan se observó un comportamiento similar con un incremento del porcentaje de observaciones atípicas. Se evidenció una sobreestimación de la cantidad de observaciones atípicas con las distancias Euclídea y Manhattan, presumiendo un efecto perjudicial de la autocorrelación serial de los residuales sobre estas dos distancias. De igual manera, se observó la existencia de un potencial efecto de empantanamiento, con fracciones de observaciones atípicas mayores que $1/(n+1)$*, en períodos mayores a 5 años. Por otro lado, la distribución lognormal en un periodo de 5 años tuvo un efecto favorable sobre las distancias Euclídea y Manhattan, y sobre la de Mahalanobis en periodos mayores a 10 años, mientras que los modelos lineales con dos armónicos para períodos menores o iguales a 10 años mostró un efecto positivo sobre la distancia de Mahalanobis.*

Palabras clave: *Precipitaciones, observaciones atípicas, distancias, análisis multivariado.*

1. Introduction

Precipitation is a climatic variable of great importance for hydrological, agricultural, industrial and energy systems. The understanding of their temporal and spatial behavior is of great interest, especially in climate risk studies, where the availability of high-resolution and good quality information is essential. In that sense, much of the success of the statistical analysis of data, especially those of precipitation underlies the collection of information. However, no matter how careful you are, you will not be free of sampling errors and anomalous values (outliers, discrepants, unusual, strange, among others). These values are far from the general behavior of the rest of the data set and cannot be considered as a manifestation of the process under study (Pérez, 1987, Rousseeuw and Van Zomeren, 1990). The anomalous values can generate erroneous results as a result of the statistical analysis and, consequently, it is unlikely to obtain precise responses that allow characterizing the process under study; because of this, it is essential to detect these values in the data set, either to eliminate them or to mitigate their effects in the analysis. That is why, both in statistical inference and in the analysis of experimental data, whether univariate or multivariate, it is essential to assess the quality of the data subject to study, which is why it is necessary to explore and build methods that help the detection of certain observations that may affect the location and scale measures (means and covariances) in addition to those of orientation (correlations) in the multivariate case. These observations are known in the statistical literature as 'outliers' (Agudelo, 1991). Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnett and Lewis (1994) indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs, similarly, Johnson (1992) defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. Other case-specific definitions are given below. Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks (Hawkins, 1980; Barnett and Lewis, 1994; Ruts and Rousseeuw, 1996; Fawcett

and Provost, 1997; Johnson et al., 1998; Penny and Jolliffe, 2001; Acuna and Rodriguez, 2004; Lu et al., 2003).

Hence, in the univariate case many informal and formal works have been developed, such is the case of an extensive literature review presented by Barnett and Lewis (1994) and by Beckman and Cook (1983). However, in the multivariate case the situation is different, the detection of outliers requires a much more detailed exploration, since these observations are not so easy to detect visually by the problem of the dimension. The present investigation has the purpose of comparing methods based on multivariate distances for the detection of atypical observations in monthly precipitation series from simulation of extreme event models.

2. Material and methods

2.1. Data

Data were obtained from the San Cristóbal meteorological station for the period between January 1951 and December 2000. The monthly precipitation series were constructed by simulating data considering extreme event models (log-normal, pearson III and gumbel I) as indicated in Table 1, and the parameters for each model were estimated using the maximum likelihood estimation method. However, it was considered to include a linear model with two harmonics, such as the one described by Guenni *et al.* (2008), in which case the parameters were estimated using generalized least squares.

Table 1: Extreme event models and linear model with two harmonics.

Model	Density function
Log-Normal	$f(x) = \frac{1}{x\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{log(x)-\mu_x}{\sigma_x})^2}$
Pearson III	$f(x) = \frac{1}{\lvert\alpha\rvert\Gamma(\beta)}(\frac{x-x_0}{\alpha})^{\beta-1} e^{(\frac{-x-x_0}{\alpha})}$
Gumbel I	$f(x) = \frac{1}{\alpha} e^{(\frac{-x-\beta}{\alpha} - e^{(\frac{-x-\beta}{\alpha})})}$
Linear model with two harmonics	$Y_t = \alpha + \gamma_1 cos(2\pi t/12) + \delta_1 sin(2\pi t/12) + \gamma_2 cos(4\pi t/12) + \delta_2 sin(4\pi t/12)$

2.2. Descriptive analysis of a time series

Any analysis of a univariate time series begins with the presentation of a graph showing the evolution of the variable over time. In this regard, based on the monthly rainfall data from the San Cristóbal meteorological station in the period between 1951 and 2000, a time series graph was made in order to observe the behavior of the monthly rainfall over the entire period of study, visualizing certain characteristics, including the presence or not of cyclic periods (seasonal cycles) or of possible changes in the series trend. Likewise, a precipitation histogram with its adjusted density is necessary to describe the monthly reduction in terms of symmetry and distribution. In this way, a statistical description was made of the monthly estimate of measures of central tendency and dispersion or variability.

2.3. Atypical observations (outliers) detection in the monthly precipitation series

Atypical observations were detected by multivariate methods, which often indicate whether the observations are relatively far from the center of the data distribution, specifically the euclidean, manhattan and mahalanobis distances, as well as through graphic devices, such as the case of 100% (1-α) 100% confidence ellipses, built in the R Software environment, through a function called 'Confelli', developed by Roger Koenker (Department of Economics, University of Illinois), with which they plot confidence ellipses based on the F statistic. This function allows an ellipse with covariance matrix C to be plotted,

centered at b (vector of means), and which by default contains 95% of the observations based on the F distribution $(2, df)$, where $df = n - 2$.

2.4. Modeling and estimation of parameters in monthly precipitation series

Monthly rainfall was plotted in order to fit extreme event models (log-normal, pearson III, and gumbel I) and the parameters for each models were estimated using the maximum likelihood estimation method proposed by Sir RA Fisher, in addition to the parameters of the linear model with two harmonics and autoregressive errors AR (1) estimated by generalized least squares.

2.5. Monte Carlo simulation

A Montecarlo simulation study was carried out in order to generate monthly precipitation series, which are known stationary stochastic processes every moment of the probability distribution (mean, variance and self-covariance), specifically a process First-order AR (1) self-regressive based on three models of extreme events (pearson III, gumbel and log-normal) and a linear model with two harmonics and autoregressive AR (1) errors, taking as reference monthly recorded precipitation series at San Cristóbal station in the period between 1951 and 2000. In that order, the statistical analyzes and the simulation was carried out with the help of scripts in the free software programming environment R 3.3.1 (see appendix 1).

3. Results

Table 2 shows a statistical description of the series of monthly rainfall from San Cristóbal meteorological station between 1951 and 2000, which are necessary for the estimation of the parameters; there is a monthly average rainfall of 133, 3 mm, with minimum monthly rainfall of 0 mm and monthly maximum of 829.4 mm. This range of monthly precipitation values is reflected in the high variability present throughout the series (Coefficient of Variation=75.75%), resulting in characteristics of a stochastic process, such as hydrological processes specifically the monthly precipitation series.

Table 2: Statistics for monthly precipitation from San Cristóbal meteorological station between 1951 and 2000.

Statistic	Value
N	552
Mínimum (mm)	0
Maximum (mm)	829.4
First quartile Q1 (mm)	53.7
Median (mm)	116.5
Third quartile Q3 (mm)	196.5
Mean (mm)	133.3
Variance (mm^2)	10202.28
Standard deviation (mm)	101.01
Coefficient of variation (%)	75.75

Figure 1 shows the monthly rainfall from the San Cristóbal meteorological station between 1951 and 2000, where there is a discontinuity in the trend of the series in the period between 1973 and 1983, which affects the high variability of the monthly rainfall (coefficient of variation=75.75%), with the subsequent effect observed in their variance. This behavior in the series trend for that period is associated with the presence of outlier observations, caused by errors in the records, for which it is recommended in addition to the detection of outlier by multivariate methods, and adjustment and estimation of parameters through methodologies that consider the presence of irregular likelihoods typical of hydrological processes and

asymmetric probabilistic distributions, as is the case of extreme event models (log-normal, pearson III, log-pearson III, gumbel I), it is also suggested to cut the series and eliminate the period that presents the discontinuity in the trend (1973-1983) in order to reconstruct the series of monthly rainfall throughout the period. It is important to note that discontinuities in the series are due to lack of information in the records.

Figure 2 shows the distribution of the monthly rainfall recorded at the San Cristóbal weather station in the period 1951-2000, by means of a histogram and the adjustment of a density function for this series, showing the asymmetry of the frequency distribution of rainfall reported in the literature.

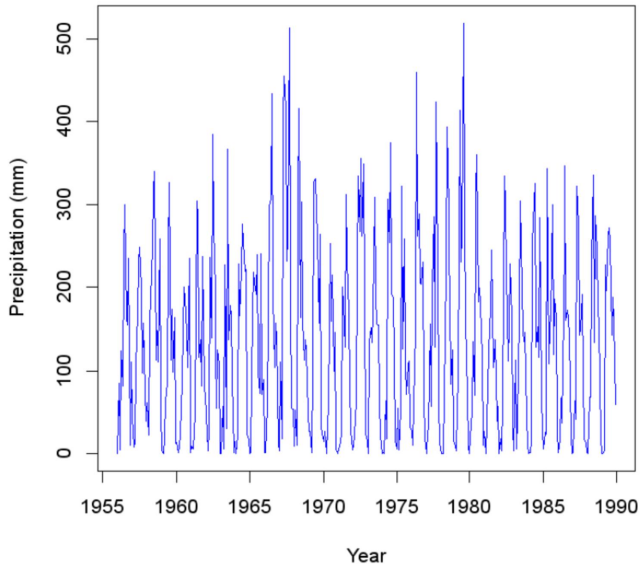

Figure 1: Monthly rainfall from the San Cristóbal meteorological station between 1951 and 2000.

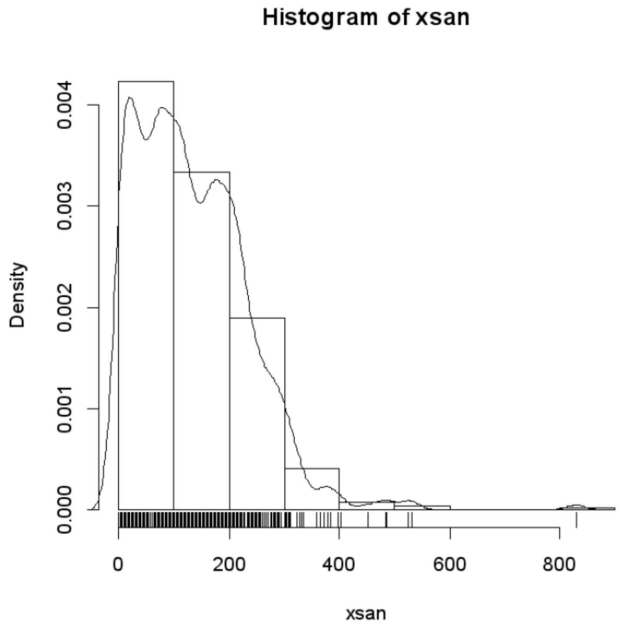

Figure 2: Distribution of monthly rainfall from San Cristóbal meteorological station between 1951 and 2000.

Figure 3 shows the results of the multivariate outlier detection using the mahalanobis distance (D2) vs quantiles of the chi-square distribution from San Cristóbal meteorological station between 1951 and 2000, it is observed that there are four observations that could be considered atypical observations, given that they move considerably away from the center of mass (centroid or multivariate mean), which are described in detail in Table 3, where it is observed that these atypical observations are associated with mahalanobis distances relatively large (D2 $>$ 10), and are distributed in the rainy season, with high rainfall occurring in August 1960 (378.7 mm), June 1984 (484 mm), July 1985 (531 mm) and July 1989 (451.8 mm).

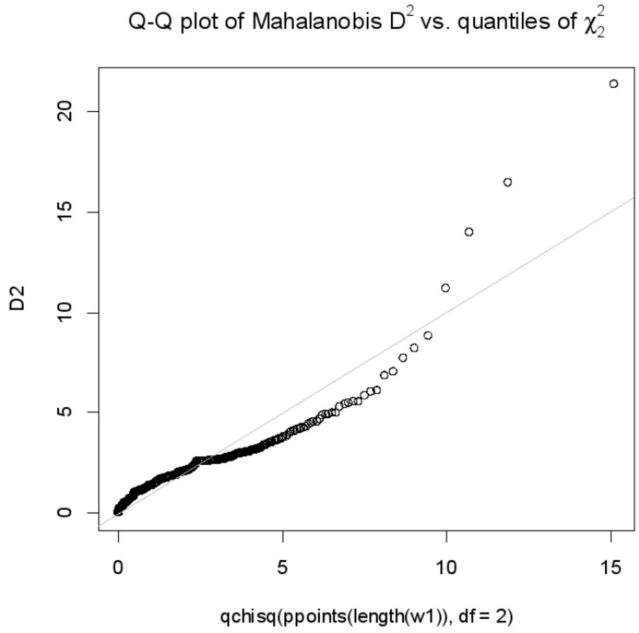

Figure 3: Mahalanobis distance (D2) vs quantiles of the chi-square distribution from San Cristóbal meteorological station between 1951 and 2000.

Table 3: Description of atypical observations in series of monthly rainfall from San Cristóbal meteorological station between 1951 and 2000.

Year	Month	Precipitation (mm)	Mahalanobis distance (D2)
1960	August	378.7	11.4
1984	June	484	16.39
1985	July	531	21.29
1989	July	451.8	13.91

Table 4 shows the results of the fit and estimation of parameters of extreme event models in monthly precipitation series from San Cristóbal meteorological station between 1951 and 2000. The Kolmogorov-Smirnoff test suggests that the model that shows the best fit to the monthly precipitation data set is the pearson III, results verified by observing Figure 4, where the densities for the four extreme event models are shown. There it is observed that the model pearson III is the one that best fits the monthly rainfall histogram of San Cristóbal station. These results verify those indicated by some authors (Guenni *et al.*, 2008; Paredes *et al.*, 2014), who affirm that the pearson III model is the one that best fits the distribution of monthly rainfall.

Tables 5, 6 and 7 show the results of the detection of atypical observations using three distances (mahalanobis, manhattan and euclidean) in data from monthly precipitation series from San Cristóbal meteorological station between 1951 and 2000, as well as those obtained by simulation of three models of

extreme events (pearson III, gumbel and lognormal), in addition to a linear model with two harmonics and AR(1) autoregressive errors for different periods (n = 5, 10, 15, 20, 25, 30 and 35 years). In that sense, in the first place, when comparing the methodologies based on multivariate distances, it is evident that the one that reported the best results in relation to the percentage of atypical observations detected in the sample was the mahalanobis distance (D2). Likewise, when comparing the obtained with this distance in each simulated scenario (see Table 5), it is observed that for short periods (5 and 10 years) the linear model with two harmonics shows a similar behavior in relation to the percentage of atypical observations (0% and 1.66%, respectively), when compared with the percentages of atypical observations (0% and 1.67%, respectively), when actual data were used (monthly San Cristóbal precipitation series between 1951 and 2000). While for periods of time greater than 10 years, the lognormal model shows a trend similar to the series studied that stabilizes as the study period increases. Secondly, when observing Table 6, it is observed that in the case of the euclidean distance the only simulated scenario where a percentage of atypical observations detected (0.06%) similar to that obtained with the San Cristóbal precipitation series is evidenced when mahalanobis distance was used, it is one where a short period of time (5 years) was considered and monthly precipitation was modeled using a lognormal distribution.

Table 4: Fit and estimation of parameters of extreme event models in monthly precipitation series from San Cristóbal meteorological station between 1951 and 2000.

Model	Estimated parameter	Goodness of fit (Kolmogorov-Smirnoff test)	
		Statistic (D)	Signification (P value)
Log-normal	$\mu_x = 4.459281$ $\sigma_x = 1.275493$	0.15533	1.019e-7
Pearson III	$\alpha = 1.230967$ $x_0 = 0.0090068$	0.089845	0.007263
Log-pearson III	$\alpha = 10.75065$ $x_0 = 2.339255$	0.19097	1.894e-11
Gumbel I	$\alpha = 91.01674$ $\beta = 4.360523$	0.53869	2.2e-16

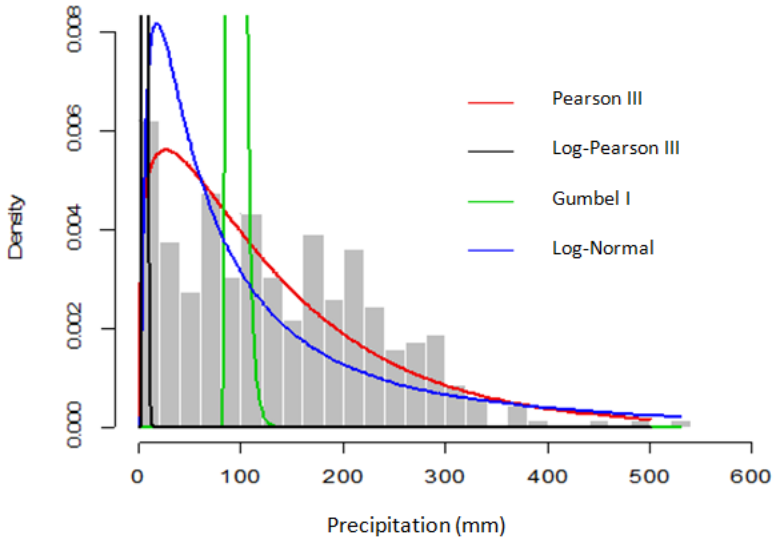

Figure 4: Fit of extreme event models in monthly precipitation series from San Cristóbal meteorological station between 1951 and 2000.

Table 5: Diagnosis of atypical observations by mahalanobis distance (D2) in simulated monthly precipitation series with AR(1) autoregressive errors.

Period (year)	n	1/(n+1)	Outliers (%) Monthly precipitation distribution				
			Real data	Pearson III	Gumbel	Lognormal	Linear with harmonics
5	60	1.64	0	4.01	3.17	2.54	0
10	120	0.83	1.67	4.02	3.30	2.75	1.66
15	180	0.55	3.33	4.05	3.35	2.846	1.11
20	240	0.41	2.5	4.05	3.34	2.855	0.83
25	300	0.33	2.66	4.08	3.35	2.895	1.33
30	360	0.28	2.22	4.07	3.38	2.891	1.11
35	420	0.24	2.2	4.06	3.33	2.939	0.98

Table 6: Diagnosis of atypical observations by euclidean distance in simulated monthly precipitation series with AR(1) autoregressive errors.

Period (year)	n	1/(n+1)	Outliers (%) Monthly precipitation distribution				
			Real data	Pearson III	Gumbel	Lognormal	Linear with harmonics
5	60	1,64	95,98	94,48	32,25	0,06	55,35
10	120	0,83	96,86	95,86	49,14	20,6	66,57
15	180	0,55	97,79	96,98	62,32	40.59	75,25
20	240	0,41	98,18	97,62	70,35	53,03	80,51
25	300	0,33	98,52	98,04	75,61	61,27	83,96
30	360	0,28	98,73	98,34	79,29	67,08	86,39
35	420	0,24	98,44	98,56	82,04	71,39	88,18

However, for periods of time longer than 5 years, the percentage of atypical observations detected with this distribution (lognormal) increases significantly in the same way as in the other simulated scenarios, including in the San Cristóbal precipitation series for all sizes of sample (n$\geq$5 years) and the other models used (pearson III, gumbel I and the linear with two harmonics). Likewise, in the case of the manhattan distance (see Table 7) a similar behavior was observed, characterized by an increase in the percentage of atypical observations detected in comparison with that obtained by the euclidean distance. All of the above suggests a potential bogging effect that could increase mahalanobis distance from non-atypical observations as reported by Quaglino and Merello (2012), who point out that bogging occurs when a group of extreme values it biases the estimates of the mean and covariance towards it and the resulting distance from these cases to the average is large, making them appear as atypical. Likewise, the aforementioned authors suggest that observations that are not atypical, will fully determine the estimate of the form and position of the data, in which case, many of the estimation methods fail if the fraction of atypical observations is greater than $1/(n+1)$ where n is the dimension of the data set, indicating that in large dimensions, a small amount of outliers may produce poor estimates. This may explain the fact that in the case of the mahalanobis distance considering the four simulated models and in the case of the manhattan and euclidean distances when a lognormal distribution was used only for short time periods (5 years), where the proportion of atypical observations is less than $1/(n+1)$ a similar behavior is observed in the detection of atypical observations.

Table 7: Diagnosis of atypical observations by Manhattan distance in simulated monthly precipitation series with AR(1) autoregressive errors.

Period (year)	n	1/(n+1)	Outliers (%) Monthly precipitation distribution				
			Real data	Pearson III	Gumbel	Lognormal	Linear with harmonics
5	60	1,64	96,83	95,64	44,51	2,88	64,36
10	120	0,83	97,85	97,14	73,11	28,97	75,49
15	180	0,55	98,55	97,93	73,11	47,7	82,18
20	240	0,41	98,83	98,4	79,071	58,93	86,07
25	300	0,33	99,07	98,69	82,88	66,26	88,59
30	360	0,28	99,20	98,89	85,51	71,39	90,34
35	420	0,24	99,14	99,04	87,47	75,18	91,63

Hence, in relation to mahalanobis distances, these must be estimated by a robust procedure in order to provide reliable measures for the recognition of extreme values, among these procedures are the multivariate M estimator, the bicuadratic S estimator, the minimum determinant covariance estimator (GCF), among others, which by definition are less affected by atypical observations. Similarly, given the effect that the proportion of atypical observations has on the estimation of location and scale parameters, the results show that the distribution of the data is related to the amount of atypical observations in the sample. Finally, as is known, the euclidean distance is severely affected when correlated variables are used, in which case the dissimilarity or divergence between the observations will be inflated, hence, the serial autocorrelation of the errors over time may have a detrimental effect on the use of the Euclidean distance and that of manhattan, which is evidenced by an overestimation of the amount of atypical observations in the monthly precipitation series, in contradiction to what was observed in the confidence ellipses of (1-α)100% for the monthly precipitation series of the San Cristóbal precipitation series (see Figure 5), where few observations (four points) are observed outside the trusted regions, and a favorable effect on the mahalanobis distance, given that this distance differs from the euclidean distance, manhattan and others that take into account the correlations of the data set. However, the lognormal distribution of precipitation over a short period of time (5 years) has a favorable effect on the ability to detect atypical observations in the case of Euclidean and Manhattan distances and for periods greater than 10 years in the case of the mahalanobis distance, which is directly related to that indicated by Poblete *et al*. (2002), who identify the log-normal function, among other functions, such as the one that presents better goodness of fit to monthly precipitation series and annual flows, in addition to the property of stabilizing the variances, while the use of linear models with two harmonics for periods less than or equal to 10 years has a positive effect on the ability to detect atypical observations of the mahalanobis distance.

4. Final considerations

It was observed the monthly precipitation series is not conditioned or biased by the presence of times of drought and rain, with typical characteristics of the life zone classified as pre-montane dry forest. Precipitation evidenced the presence of an annual cycle with a well-defined maximum in the month of June and autocorrelated residuals characteristic of a AR(1) autoregressive model. The observations considered atypical were detected, distributed in the rainy season, specifically between May and September. The monthly precipitation was fitted to a pearson III model.

Mahalanobis distance reported the best results in relation to the percentage of atypical observations detected. An overestimation of the number of atypical observations was evidenced in the case of Euclidean and Manhattan distances.

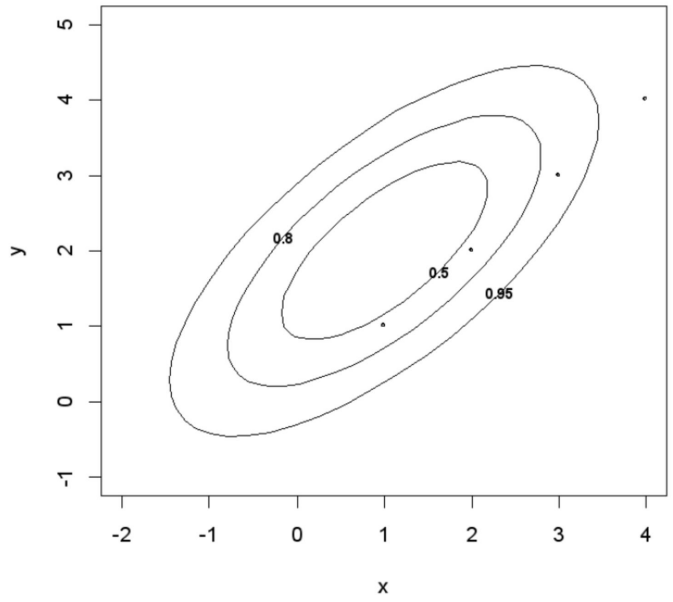

Figure 5: Confidence ellipses of 100%(1-α)100% for the monthly precipitation series from San Cristóbal meteorological station.

The use of the lognormal distribution to model rainfall over a period of 5 years had a favorable effect on the ability to detect atypical observations in the case of the Euclidean and Manhattan distances and for periods greater than 10 years in the case of the Mahalanobis Distance.

A more exhaustive study is recommended in relation to the use of the euclidean and manhattan distances in periods of time less than or equal to 10 years considering different autocorrelation structures serial errors, and a more exhaustive study of the mahalanobis distance with periods greater than 10 years and a linear model with harmonics associated with periods less than or equal to 10 years, in addition to the use of robust procedures, among these; the multivariate M estimator, the bi-quadratic S estimator and the minimum determining covariance estimator (MCD), in order to provide reliable measures for the recognition of extreme values. From all the recommendations above, the potential of the use of logarithmic transformations in the study of monthly precipitation is evidenced. Finally, if it must be noted that if the presence of an atypical observation is not due to an error in the records of a series of monthly precipitation, eliminating them is not the solution, since this can modify inferences made from the series, because a bias is introduced and it can affect both distribution and variances.

References

Acuna E, Rodriguez C (2004): Meta analysis study of outlier detection methods in classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, In proceedings IPSI 2004, Venice. <http://academic.uprm.edu/eacuna/paperout.pdf>

Agudelo G (1991): Estimación robusta y detección de outliers. Universidad de Antioquía. Facultad de Ciencias Exactas y Naturales. Departamento de Matemáticas. Segundo Coloquio Regional de Matemáticas y Estadística. Antioquia. Colombia.

Barnett V, Lewis T (1994): *Outliers in statistical data*. 3ed. John Wiley & Sons: Chichester. 584 pp.

Beckman R, Cook RD (1983): Outliers. *Technometrics*, 25:119-158.

Fawcett T, Provost F (1997): Adaptive fraud detection. *Data-mining and Knowledge Discovery*, 1:291-316.

Guenni L, Degryze E, Alvarado K (2008): Análisis de la tendencia y la estacionalidad de la precipitación mensual en Venezuela. *Revista Colombiana de Estadística*, 31:41-65.

Hawkins D (1980): *Identification of Outliers*. London, Chapman & Hall.

Johnson R (1992): *Applied Multivariate Statistical Analysis*. Prentice Hall.

Johnson T, Kwok I, Ng R (1998): Fast Computation of 2-Dimensional Depth Contours. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 224-228. AAAI Press.

Lu C, Chen D, Kou Y (2003): Algorithms for spatial outlier detection, In Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM'03), Melbourne, FL.

Paredes F, La Cruz F, Guevara E (2014): Análisis regional de frecuencia de las sequías meteorológicas en la principal región cerealera de Venezuela. Bioagro, 26:21-28.

Penny KI, Jolliffe I (2001): A comparison of multivariate outlier detection methods for clinical laboratory safety data, *RThe Statistician*, 50:295-308.

Pérez J(1987): Identificación de Outliers en Muestras Multivariantes. Tesis Doctoral. Universidad de Sevilla, España.

Poblete A, Aguiar L, Sánchez G (2002): Estructuras estadísticas de los derrames del río San Juan y el Jáchal y sus relaciones. *Revista geográfica* N°6 del Instituto y Departamento de Geografía de la Universidad Nacional de San Juan. Argentina.

Quaglino M, Merello J (2012): Métodos multivariados en estudios de vulnerabilidad social en la Provincia de Santa Fe. XVII Jornadas de investigación en la Facultad de Ciencias Económicas y Estadística. Argentina.

Ruts I, Rousseeuw P (1996): Computing Depth Contours of Bivariate Point Clouds. *Computational Statistics and Data Analysis*, 23:153-168.

Rousseeuw P,Van Zomeren B (1990): Unmasking Multivariate Outliers and Leverage Points. *J. Am. Stat. Assoc,*, 85:633-651.

Apendix: R code

Confidence ellipses

```
confelli <- function(b, C, df, level = 0.95, xlab = '', ylab = '', add=T, prec=51)
{
d <- sqrt(diag(C))
dfvec <- c(2, df)
phase <- acos(C[1, 2]/(d[1] * d[2]))
angles <- seq( - (pi), pi, len = prec)
mult <- sqrt(dfvec[1] * qf(level, dfvec[1], dfvec[2]))
xpts <- b[1] + d[1] * mult * cos(angles)
ypts <- b[2] + d[2] * mult * cos(angles + phase)
if(add) lines(xpts, ypts)
else plot(xpts, ypts, type = 'l', xlab = xlab, ylab = ylab)
a<-round(runif(1,1,51))
text(xpts[a], ypts[a],paste(level),adj=c(0.5,0.5),font=2,cex=0.7)
}
library(RODBC)
canalexcel<-odbcConnectExcel2007('D:Datos climatológicos Guanare-Aeropuerto1.xlsx')
sqlTables(canalexcel)
x<-sqlFetch(canalexcel,'n=34')
```

```
x1<-c(x[,3])
año<-x[,1]
Datos <- data.frame(x=año,y=x1)
C<-matrix(c(1,0.7,0.7,1),ncol=2)
b<-c(1,2)
x<-Datos[,1]
y<-Datos[,2]
plot(x,y,xlim=c(-2,4),ylim=c(-1,5),cex=0.5)
confelli(b,C,df=406)
confelli(b,C,df=406,level=0.8)
confelli(b,C,df=406,level=0.5)
```

Simulated autoregressive series

```
Pearson.III<-ts(c(0.8*rgamma(480,1.1273654022,0.0093966210)+rnorm(1)), start=c(1956, 1),
  frequency=12)
library(MASS)
library(stats4)
library(VGAM)
Gumbel.I<-ts(c(0.8*rgumbel(480, 78.873994, 4.230078)+rnorm(1)), start=c(1956,1), frequency=12)
Lognormal<-ts(c(0.8*rnorm(480,4.282190,1.386791)+rnorm(1)), start=c(1956, 1), frequency=12)
t<-c(seq(1,12))
armon1<-c(rep(cos(360*t/12),40) )
armon2<-c(rep(sin(360*t/12),40))
armon3<-c(rep(cos(720*t/12),40))
armon4<-c(rep(sin(720*t/12),40))
Precipitación<-x1
Precipitación
Mes<-x[,2]
Mes
Año<-x[,1]
Año
Datos<-data.frame(Año,Mes,Precipitación,armon1,armon2,armon3,armon4)
Datos
res<-residuals(lm(Precipitación~armon1+armon2+armon3+armon4,data=Datos))
res
sigma2<-res^2
pesos<-1/sigma2
fit<-lm(Precipitación~armon1+armon2+armon3+armon4,weights=pesos)
summary(fit)
Yt<-119.3199+4.7945*armon1-4.3039*armon2+6.7047*armon3+0.7372*armon4+rnorm(1)
Lineal.armonico<-ts(c(Yt),start=c(1956,1),frequency=12)
z<-cbind(Pearson.III,Gumbel.I,Lognormal,Lineal.armonico)
Datos<-data.frame(z)
Datosz1<-ts(Datos,start=c(1956,1),frequency=12)
z1<-ts(Datos,start=c(1956,1),frequency=12)
plot(z1,main='Cuatro series de precipitación mensual a partir de un modelo AR(1)',
  xlab='Año',cex.main=0.9)
```

Mahalanobis distance power

```
t<-sapply(1:1000,function(x){
no<-ni
n<-ni*1
p<-po
Dist<-ts(c(p*rdist(n,)+rnorm(1)), start=c(inic, ), frequency=12)
```

```
año<-c()
Datos<-data.frame(x=año,y=Gamma)
w<-cbind(Gamma,año)
Sx<-cov(w)
D<-mahalanobis(w, colMeans(w), Sx)
})
z<-t$>$=qchisq(0.95,2)
outlier<-sum(z)
poutlier<-outlier/length(t)*100
```

Manhattan distance power

```
t<-sapply(1:1000,function(x){
no<-ni
n<-no*12
p=po
Dist<-ts(c(p*rdist(n,)+rnorm(1)), start=c(inic, 1), frequency=12)
año<-c()
Datos<-data.frame(x=año,y=Gamma)
D<- dist(Datos,method='manhattan')
})
z<-D$>$=qchisq(0.95,2)
outlier<-sum(z)
poutlier<-outlier/length(t)*100
```

Euclidean distance power

```
t<-sapply(1:1000,function(x){
no<-ni
n<-no*12
p=po
Dist<-ts(c(p*rdist(n,)+rnorm(1)),start=c(inic, 1),frequency=12)
año<-c()
Datos<-data.frame(x=año,y=Gamma)
D<-dist(Datos,method='euclidean')
})
z<-t$>$=qchisq(0.95,2)
outlier<-sum(z)
poutlier<-outlier/length(t)*100
Pearson.III<-ts(c(0.8*rgamma(480,1.1273654022,0.0093966210)+rnorm(1)), start=c(1970, 1),
  frequency=12)
library(MASS)
library(stats4)
library(VGAM)
Gumbel.I<-ts(c(0.8*rgumbel(480,78.873994,4.230078)+rnorm(1)), start=c(1970,1), frequency=12)
Lognormal<-ts(c(0.8*rnorm(480,4.282190,1.386791)+rnorm(1)), start=c(1970, 1), frequency=12)
t<-c(seq(1,12))
armon1<-c(rep(cos(360*t/12),40))
armon2<-c(rep(sin(360*t/12),40))
armon3<-c(rep(cos(720*t/12),40))
armon4<-c(rep(sin(720*t/12),40))
Yt<-119.3199+4.7945*armon1-4.3039*armon2+6.7047*armon3+0.7372*armon4+rnorm(1)
Lineal.armonico<-ts(c(Yt),start=c(1970,1),frequency=12)
hist(Gumbel.I,prob=FALSE,main='',xlab='Precipitación (mm)',ylab='Frecuencia')
hist(Gumbel.I,prob=FALSE,main='',xlab='Precipitación (mm)',ylab='Frecuencia')
```