MISSING DATA IMPUTATION OF CLIMATE DATASETS: IMPLICATIONS TO MODELING EXTREME DROUGHT EVENTS

GLÁUCIA TATIANA FERRARI E VITOR OZAKI

Universidade de São Paulo Escola Superior de Agricultura Luiz de Queiroz (USP/ESALQ), Piracicaba, São Paulo, Brazil

glautf@usp.br, vitorozaki@gmail.com

Received September 2012 - Accepted July 2013

ABSTRACT

Time series from weather stations in Brazil have several missing data, outliers and spurious zeroes. In order to use this dataset in risk and meteorological studies, one should take into account alternative methodologies to deal with these problems. This article describes the statistical imputation and quality control procedures applied to a database of daily precipitation from meteorological stations located in the State of Parana, Brazil. After imputation, the data went through a process of quality control to identify possible errors, such as: identical precipitation over seven consecutive days and precipitation values that differ significantly from the values in neighboring weather stations. Next, we used the extreme value theory to model agricultural drought, considering the maximum number of consecutive days with precipitation below 7 mm for the period between January and February, in the main soybean agricultural regions in the State of Parana.

Keywords: Imputation, quality control, precipitation, risk

RESUMO: IMPUTAÇÃO DE DADOS FALTANTES E SUA APLICAÇÃO NA MODELAGEM DE EVENTOS EXTREMOS DE SECA AGRÍCOLA

Este artigo relata o procedimento utilizado na reconstrução de um banco de dados contínuo de precipitação diária de estações meteorológicas localizadas no Estado do Paraná, Brasil. Após a imputação, os dados passaram por um processo de controle de qualidade que teve como objetivo identificar possíveis erros como precipitação idêntica em sete dias consecutivos (não aplicados a dados de precipitação zero) e valores de precipitação que diferem significativamente dos valores em estações meteorológicas vizinhas. Com o banco de dados contínuo, o interesse foi utilizar a teoria de valores extremos para modelar a seca agrícola, considerada como sendo o número máximo de dias consecutivos com precipitação abaixo de 7 mm para o período entre janeiro e fevereiro, crítica para a fase de enchimento de grãos da soja nas principais regiões produtoras do Estado do Paraná.

Palavras-Chave: Imputação, controle de qualidade, precipitação, risco

1. INTRODUCTION

Drought is one the most serious environmental factors limiting crop yields worldwide with devastating economic and social consequences (Rivero et al*.,* 2007). Accurate estimates of the risk of drought are important because of its importance in the decision making process for farmers in reference to the best season for planting and harvesting.

The size of the crop loss depends on the period of their cycle and when they are exposed to extreme climatic conditions. Thus, studies of risk of extreme weather conditions should be taken into account and specific features of each weather condition. According to the Brazilian Agricultural Research Corporation (EMBRAPA, 2002), in order to obtain maximum yield, the need for water in the soybean crop throughout its cycle, ranges from 450 to 800 mm depending on the weather conditions, crop management and the duration of the cycle.

The soybean has two well-defined critical periods in relation to the lack of water: germination through to flowering and grain filling. During the flowering and grain filling stages, the plant presents its higher water necessity (7-8mm/day), decreasing thereafter (EMBRAPA, 2002).

The critical precipitation period of the soybean occurs between January, 15 through February, 28 (needs of 7-8 mm in the stage of flowering and grain filling). To quantify the risk of extreme weather conditions and their consequences for the soybean crop, the probability of the occurrence of adverse meteorological phenomena in agriculture, especially drought, is of paramount importance to the rural sectors dealing with insurance, financing and the planning of farming activities.

Extreme drought in the series are analyzed by the maximum annual series adjusting the widespread distribution of extreme values and extreme value generalized distribution. These drought series are obtained by determining the maximum dry season each year from January through February, so that the length of the series is equal to the number of years that the data series is available.

However, the extreme values analysis requires continuous climate series that can generate reliable results in the decision making. In Brazil, climatic data are released from the National Institute of Meteorology (INMET), the Center for Weather Forecasting and Climate Studies (CPTEC/INPE) and the National Water Agency from the automatic and manual weather stations. In general, the series present serious problems of missing data and inconsistencies.

In order to overcome the problems of inconsistency, inaccuracy and measurement errors and to put together a solid database for analysis, a reconstruction process is required which involves a method of imputation and quality control of precipitation data (Feng and Qian, 2004; Vicente-Serrano et al. 2010).

The main goal of this study is to obtain a database of reconstructed rainfall for the State of Parana (Southern Region of Brazil). As a secondary goal, the generalized extreme value distribution was fitted to this reconstructed series to model the number of consecutive days with daily precipitation less than 7 mm (critical water necessity) in the regions of Parana.The article is outlined as follows: Section 3 presents the study region and data. Section 2 introduces the methods of imputation, quality control of data and the theory of extreme values. Section 4 presents results and Section 5 concludes the article.

2. THE DATASET

The State of Parana is located in the Southern of Brazil covering a total area of 199,314 km^2, which corresponds to 2.3% total area of Brazil and has 399 municipalities divided into 10 regions which several municipalities with similar economic and social situations.

The data used in this study were obtained from the National Institute of Meteorology (INMET), the Center for Weather Forecasting and Climate Studies (CPTEC/INPE) and the National Water Agency (ANA). The observations refer to the daily precipitation in millimeters (mm) of 484 weather stations located in the State of Parana. Each series has 35 years between January 1975 and December 2009. The spatial density is a station of 411.8 km^2. The Figure 1 shows the spatial distribution of the stations.

3. METHODOLOGY

The imputation and quality control involves two basic steps described by Vicente-Serrano et al. (2010). The first step involves the imputation of missing data through methods that use auxiliary information from neighboring weather stations to obtain continuous data series. The second step evaluates the quality control of the series reconstructed to identify and replace unverified records in the database (negative precipitation, some zero values and the records that differ significantly from

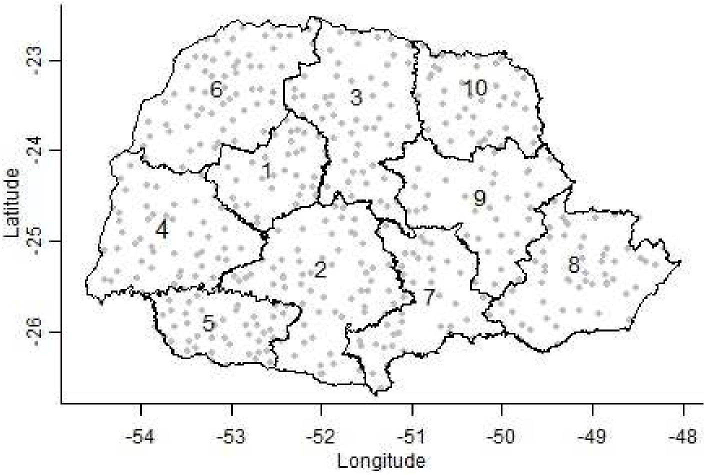

Figure 1 - Geographic distribution of the weather stations.

the amounts recorded in the neighboring stations). After the imputation and quality control of the series, the next step consists in using the extreme value theory to model the drought period and its return period in the selected regions.

3.1 Imputation methodology

Several imputation methods have been proposed in the literature. The adequacy of each method depends on the missing data mechanism (Rubin, 1996; Schafer, 1997; Schneider, 2001; Junninena et al., 2004; Coulibaly, 2007; Göktürk, 2008; Ramos-Calzado et al., 2008). In this study three methods are tested: the nearest neighbor method, the inverse distance weighting method and the linear regression method.

The nearest neighbor method is widely used due to its conceptual simplicity that leads to a straightforward implementation. If and   are two time series, then this method considers the smallest distance between station and station , in other words . Thus, the missing data are imputed directly in the observed data time series from the nearest weather station.

The inverse distance weighting method has the advantage to be ease implement computationally. In order to predict the missing data from one station, this method uses the values measured in the neighbors of this station. In this method, the values measured in the nearest stations will have a greater influence on the forecast than those measured further away. The method can be defined as:

$$z(x_j) = \frac{\sum_{i=1}^{n} z(x_i) d_{ij}^{-r}}{\sum_{i=1}^{n} d_{ij}^{-r}}, \qquad (1)$$

where $z(x_j)$ is the expected predicted value of the station according to the weighted average of the observed stations $z(x_1)$, $z(x_2)$,..., $z(x_n)$ and d_{ij}^{-r} is the weighting factor, defined as the Euclidean distance between the observation $z(x_i)$ and the value to be estimated $z(x_j)$ and r a positive real number (usually $r = 2$).

In the linear regression method the missing data are obtained using the most correlated series. Since there are no negative precipitation values, the line of regression is forced through the origin, providing a model with only the slope coefficient:

$$Y_i = \beta X_i + \epsilon_i, i = 1,2,\cdots,n, \qquad (2)$$

where Y is the series with missing data to be estimated, X_i is the series most correlated with Y_i, β is the slope parameter and ε_i is the error term, with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. The parameter β is estimated by ordinary least squares and is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}. \qquad (3)$$

For the comparison of imputation methods, following the description of Vicente-Serrano et al. (2010), 1% of the observations were selected (excluding missing data) for each weather station. After the selection of these data, we assume that these observations were missing and the three methods were applied. We use the root mean square error (RMSE) as a criteria to choose the best method of imputation. The RMSE is used to measure forecast error and is found by calculating the square root of the sum of squared prediction errors divided by the number of observations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}}, \qquad (4)$$

where x_i is the amount of precipitation observed, $\hat{x}_i$ is the expected predicted value of precipitation (in this case it is the imputed value of precipitation).

3.2 Quality control of the data

The objective of the quality control is to identify incorrect data or unverified records of the data. Due to the large amount of observation in our dataset, we use the approach adopted by Vicente-Serrano et al. (2010), which compares the rank of each data classification and the average rank of the data recorded in adjacent observatories.

The rank of the original series of daily precipitation after clearing the zero values is converted into percentiles and each value of precipitation is replaced by its corresponding share according to the rank. After processing, the values zero were assigned a zero percentile. For each station we choose stations located within a radius of 55 km[1] and a minimum of four stations as a condition for the test (at that distance, all stations have at least four neighbors). In the first phase, only observations above 99th percentile were observed.

According to Vicente-Serrano et al. (2010), the maximum difference allowed between an observation of the corresponding station and the average percentile of the neighboring stations is fixed at the 60 percentage units. If the difference is greater, the observation is replaced by the observation of the closest series.

In the second step, observations below the 99th percentile are compared with the average of neighboring stations. In this case, a difference of 70 percentile units are defined as the threshold for identifying unverified data, and the higher values are flagged and replaced with the data from the nearest station (Vicente-Serrano et al., 2010). The values zero coinciding with substantial precipitation in the nearby stations are also replaced

[1] In order to avoid problems with missing data in the neighboring stations we selected stations located within a radius of 55 km with an average correlation of 0.56. Otherwise, it would not be possible to have a reasonable quantity of stations to perform the imputation and the quality control of the series (in this study,, if we consider a distance of 15 km, the average number of neighbor stations is equal to one).

with data from the closest station to the average percentage in the neighboring stations which is greater than 50. Following is checked at each occurrence series of identical values (excluding zero) in at least 7 consecutive days.

These data are replaced by precipitation values from the nearest station. As Vicente-Serrano et al. (2010), this methodology can affect the probability of distribution of the most extreme observations in a time series. Thus, a test is applied using standard methods for analysis of extreme values. We calculated the coefficients of L-skewness and L-kurtosis of the data series before and after the quality control process. Series of partial length or numbers of peaks above a threshold was taken from each station in order to capture the extreme values.

Given the number of precipitation in an weather station, $X = (x_1, x_2, ..., x_n)$, where x_n is the observation of a given day, the series of partial length $Y = (y_1, y_2, ..., y_j)$ is in excess of the original series along a predetermined threshold, x_0

$$y_j = x_i - x_0, \forall\, x_i > x_0 \,. \tag{5}$$

Therefore, the size of the series depends on the threshold value, . For each series, the precipitation values corresponding to percentiles 90th and 95th, before and after the process of quality control, are used as limits for the construction of a series of partial length. The L-asymmetry coefficient L-coefficients of skewness (τ_3) and L-kurtosis coefficient L-coefficients of kutosis (τ_4) are calculated as follows:

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \text{ e } \tau_4 = \frac{\lambda_4}{\lambda_2}\,, \tag{6}$$

where λ_2, λ_3 and λ_4 are the L-moments of the series of partial length (Y), given by:

$$\lambda_1 = \beta_0 \tag{7}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{8}$$

$$\lambda_3 = \beta_0 - 6\beta_1 + 6\beta_2 \tag{9}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0, \tag{10}$$

where β_s (s= 1, 2, 3, 4) are estimated through the probability-weighted moments calculated from the data of the series of partial length (Y) arranged in ascending order, given by:

$$b_0 = \frac{1}{n}\sum_{j=1}^{n} y_{(j)} \tag{11}$$

$$b_r = \frac{1}{n}\sum_{j=1}^{n} \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} y_{(j)}\,,\ r \geq 1, \tag{12}$$

where n is the sample size. If the relationship between τ_3 and τ_4 is approximately linear before and after the quality control process indicates that the process did not affect significantly the statistical characteristics of the observations.

Extreme Value Theory

The generalized extreme value distribution (GEV), developed by Jenkinson (1955), can be considered as a generalized distribution which includes the three possible types of asymptotic distributions of extreme weather conditions, known as the Gumbel (type I), Fréchet (type II) and Weibull (type III). The cumulative distribution function of the GEV distribution is expressed by

$$F(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{13}$$

set at $-\infty < x < \mu - \sigma/\xi$ to $\xi < 0$, $-\infty < x < +\infty$ to $\xi \to 0$, , $\mu - \sigma/\xi < x < +\infty$ to $\xi > 0$, in which μ, σ and ξ are the parameters of location, scale and form, respectively, with $\mu \in \mathbb{R}$ and $\sigma > 0$. Extreme values distributions Gumbel, Fréchet and Weibull correspond respectively to particular cases of the GEV distribution. Deriving Equation 9 with respect to x obtain the probability density function of GEV distribution, given by:

$$f(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \tag{14}$$

The logarithm of the likelihood function for the GEV distribution given by Equation 10, for $\xi \neq 0$, is given as follows

$$l(\mu,\sigma,\xi;x) = \sum_{i=1}^{n}\left\{-\ln(\sigma) - \left(\frac{1+\xi}{\xi}\right)\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{15}$$

where n is the sample size.

For the particular case of the Gumbel distribution, in which $\xi = 0$, the cumulative distribution function is given by:

$$F(x) = \exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right], \tag{16}$$

defined in $-\infty < x < +\infty$, that μ and σ are the location and scale parameters, respectively, with $\mu \in \mathbb{R}$ e $\sigma > 0$.

Deriving from Equation 11 in relation to x obtaining the distribution of Gumbel probability,

$$f(x) = \frac{1}{\sigma}\left\{\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]\right\}, \tag{17}$$

defined in $-\infty < x < +\infty$ to $\mu \in \mathbb{R}$ e $\sigma > 0$.

The function of the logarithm likelihood of Equation 12 is given by:

$$l(\mu,\sigma,\xi;x) = \sum_{i=1}^{n}\left\{-\ln(\sigma) - \left(\frac{x_i-\mu}{\sigma}\right) - \exp\left(\frac{x_i-\mu}{\sigma}\right)\right\}. \tag{18}$$

According to Coles (2004), the maximum likelihood method can be used to obtain parameter estimates. Coles (2004) shows in detail the estimation using this method. One can test the null hypothesis that the extremes follow a

Gumbel distribution using the likelihood ratio test as described by Hosking (1984):

$$\Lambda^* = \left(1 - \frac{2.8}{\mathrm{n}}\right)\left(-2\left[l\left(\hat{\boldsymbol{\theta}}_{Gumbel}\right) - l\left(\hat{\boldsymbol{\theta}}_{GEV}\right)\right]\right), \qquad (19)$$

where n is the sample size and $\hat{\theta}^{\mathrm{T}} = \left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

Thus, to test the hypothesis H_0: $\xi = 0$ versus H_A: $\xi \neq 0$, the value of the statistic Λ* should be compared with the tabulated value of chi-square distribution with one degree of freedom (χ_1^2) and a predetermined level of significance α (5%). Is rejected if $\Lambda^* \geq. \chi^2_{1,\alpha}$.

The probability that there will be a drought period longer than a certain value x is given by:

$$P(X > x) = 1 - F(x) = 1 - \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \qquad (20)$$

The return period is calculated by $\tau = 1/((1 - F(x))$, where τ is usually expressed in years. The level of return (x_p), associated with the return period τ, is obtained from the solution:

$$x_p = \mu - \frac{\sigma}{\xi}\left\{1 - \left[-\ln(1-p)\right]^{-\frac{1}{\xi}}\right\}. \qquad (21)$$

The confidence intervals for levels of return are calculated using the Delta method. Rao and Toutenberg (1999) provide details of the method.

4. RESULTS AND DISCUSSION

The inverse distance weighting method provided better results, with an average RMSE of 7.195 mm (with a range between 2.164 mm to 15.189 mm) of all stations. The nearest neighbor method provided an average RMSE of 9.273 mm (range 2.368 mm to 22.294 mm) and the linear regression method achieved RMSE average of 7.834 mm (with an interval of 2.456 mm to 16.584 mm). Thus, the missing precipitation data from 484 weather stations were imputed by the inverse distance weighting method and the series are consistent continuous as there is no missing data.

Once the imputation process is concluded, the next step consists in using the complete dataset to the control quality process. The percentiles above the 99th and equal to zero, 0.33% and 0.86% were replaced, respectively, while for the percentile between 0 and 99th, the replacement was 0.02%. On average, the proportion of the substituted data, using the criteria described in Section 3, was 1.21% in each weather station and the lowest proportion of substituted data was 0.094% and the highest was 5.257%. Only 21 stations had more than 3% of the data represented.

Most substitutions (70.93%) corresponded to zero values. For identical values (excluding zeros) in at least 7 consecutive days the proportion of overwritten data was 0.0025% (corresponding to 138 data). The L-coefficients of skewness and kurtosis of the data series before and after the process of quality were calculated and the relationship between the values τ_3 and τ_4 was approximately linear (Figure 2). This provides evidence that the quality control process did not significantly affect the statistical characteristics of the extremes.

The period from January 15 to February 28 was chosen because it represents the period in which most of the soybean is in the stage of flowering-grain filling (as seen earlier, this is the critical period in which it needs from 7-8 mm of precipitation) and considering that in the five regions, the planting of soybeans occur between late October and early November (EMBRAPA, 2002). More importantly, for a more detailed study, it is of interest to analyze separately each municipality, taking into consideration its climate and agricultural calendar. Comparing the statistic Λ*, presented in Table 1, with the tabulated value of $\chi_1^2 = 3.84$, we concluded that the Gumbel distribution is more appropriate to model the data under study, since $\Lambda^* < \chi^2_{1,0.05}$.

Table 1 shows the estimates of the Gumbel distribution parameters. Vicente-Serrano and Begues-Portuguese (2003) also obtained better results by adjusting the Gumbel distribution to the drought data in the Northeastern Region of Spain.

The probability of a drought period greater than 10, 20, 30 and 40 can be seen in Table 2. It is noted that regions 2, 3 and 4 are the regions with the highest probability of occurrence of these values.

The payback return period for the drought period of 45 consecutive days with precipitation below 7 mm occurs once every 60, 57, 53 and 124 years for regions 2, 3, 4 and 5, respectively, while for region 1, the highest recorded drought period (41 days) will occur once every 55 years. The levels of return for 5, 10, 15 and 20 years are shown in Table 3.

The results show that regions 2, 3 and 4 of the state have the greatest probability of occurrence of the maximum number of consecutive days with precipitation below 7 mm and longer drought periods are more likely to occur. The results show that these three regions have the highest agricultural risk in the state. In other words, year after year these regions would present a high prevalence of oscillation with drastic reductions in agricultural yield.

A metric commonly used by the market to check the relative risk of certain regions is the coefficient of variation (CV). The advantage of using the CV is based on the fact that this metric is dimensionless, allowing comparisons between different regions. On the other hand, its usefulness is reduced considerably when the mean value is close to zero. In this case, the CV becomes relatively sensitive to small changes in standard

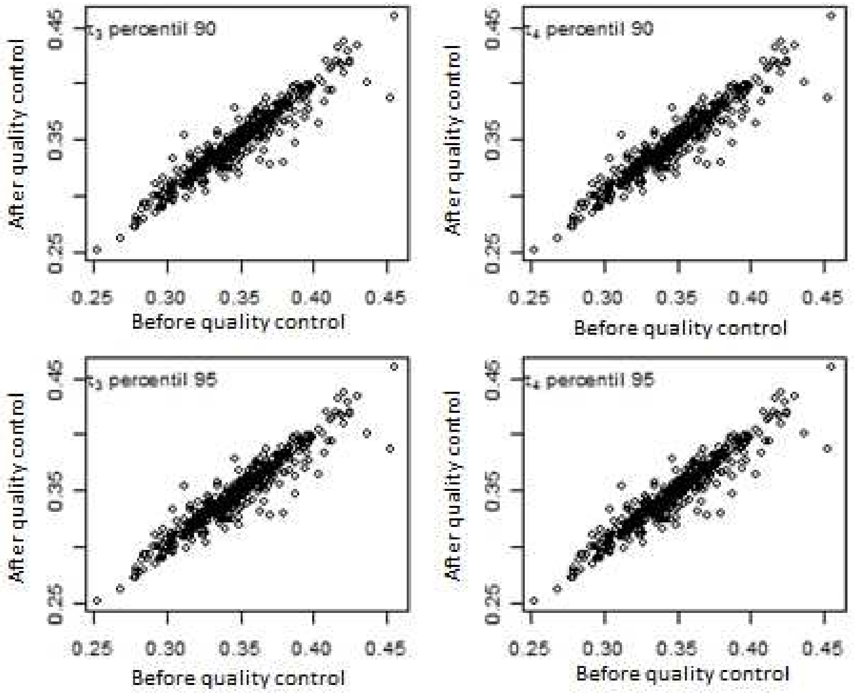

Figure 2 - Relationship between L-coefficient of skewness and kurtosis for the series of partial length percentile 90th and 95th, before and after quality control.

Table 1 - Λ* Statistics and estimates of location parameters (μ) and scale (σ) of the Gumbel distribution.

Region	Λ*	$\hat{\mu}$	$\hat{\sigma}$
1	0.0025	18.34	5.68
2	0.0037	19.62	6.21
3	0.2576	20.03	6.20
4	0.7679	20.36	6.21
5	0.4439	18.72	5.46

Table 2 - Probability of occurrence of the maximum number of consecutive days with precipitation below 7 mm.

Region	>10	>20	>30	>40
1	0.9870	0.5260	0.1205	0.0218
2	0.9910	0.6096	0.1714	0.0369
3	0.9935	0.6339	0.1815	0.0391
4	0.9950	0.6534	0.1908	0.0414
5	0.9928	0.5466	0.1190	0.0201

deviation. This is not the case in this article. Figure 3 compares the risk of all regions.

It is worth noting that region 5 is the most risky and is ranked fourth in the probability of a prolonged drought. On the other hand region 3 has a lower relative risk and is ranked second in the probability of a long drought. Possibly this is because these two regions (5 and 3) have a large number of irrigated farms covering an extensive land area. In fact, analysing the data from the 2006 Agricultural Census Bureau, it shows that the two regions account for nearly 45% of all irrigated properties in the state, and almost 25% of the total irrigated area.

5. CONCLUSION

This article proposes an alternative way to investigate the consistence of precipitation data, traditionally problematic in Brazil. The imputation and quality control data have proven useful in obtaining a continuous daily series and since this procedure did not significantly affect the characteristics of extreme weather conditions.

The generalized extreme value distribution with parameter $\xi \rightarrow 0$ which corresponds to the distribution of type I or Gumbel, proved adequate in studying the behavior of the

Table 3 - Levels of return ($\hat{x}_p$ - in days) and estimated lower limits (LI) and superior (LS) of their respective 95% intervals of confidence for the return periods 5, 10, 15 and 20 years obtained using Delta.

	Period of return (years)											
Region	5 years			10 years			15 years			20 years		
	LI	$\hat{x}_p$	LS	LI	$\hat{x}_p$	LS	LI	$\hat{x}_p$	LS	LI	$\hat{x}_p$	LS
1	24	27	30	27	31	35	29	34	39	30	35	40
2	25	29	33	29	34	39	31	36	41	32	38	44
3	25	29	33	29	34	39	32	37	42	32	38	44
4	26	30	34	29	34	39	31	37	43	33	39	45
5	24	27	30	27	31	35	28	33	38	30	35	40

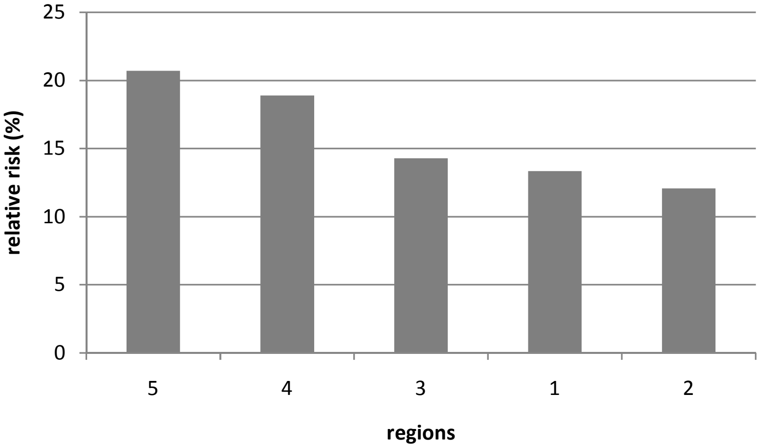

Figure 3 - Relative risk in each Region.

drought periods of the five observed regions in Parana State. Taking this fact into account, farmers are able to use the results to control cash flow so as to adapt to potential losses when drought is expected. This enables producers to create a catastrophe fund to manage drought periods. In addition if farmers are aware of the risk of drought in a particular area they can consider pay a premium and receive an indemnity when drought occurs. In other words, farmers might be protected by using government crop insurance programs and to avoid great economic losses.

6. REFERENCE

COLES, S. **An introduction to statistical modeling of extreme values**. London: Springer, 2004. 208 p.

COULIBALY P. Comparison of neural network methods for in filling missing daily weather records. **Journal of Hydrology**. v. 341, p. 27-41, 2007.

EMBRAPA. Tecnologias de produção de soja região central do Brasil 2003, Embrapa Soja Londrina, 2002, p. 199. Disponível em http://sistemasdeproducao.cnptia.embrapa.br/Fontes HTML/Soja/SojaCentralBrasil2003/exigencias.htm . Acess 4 April 2012.

FENG, S.; QIAN,W. Quality control of daily meteorological data in China, 1951-2000: a new dataset. **International Journal of Climatology**, Chichester, v. 24, p. 853-870, 2004.

GÖKTÜRK, O.M.; BOZKURT, D.; SEN, L.; KARACA, M. Quality control and homogeneity of Turkish precipitation data. **Hydrological Processes**, Londres, v. 22, p. 3210-3218, 2008.

HOSKING, J.R.M. Testing whether the shape parameter is zero in the generalized extreme-value distribution. **Biometrika**, Cambridge, v. 17, p. 301-310, 1984.

JENKINSON, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements.

Quartely Journal of the Royal Meteorological Society, Oxford, v.81, p. 158-171, 1955.

JUNNINENA, H.; NISKAA, H. TUPPURAINENC, K.; RUUSKANENA, J.; KOLEHMAINENA, M. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38 p. 2895–2907, 2004.

RAMOS-CALZADO, P.; GOMEZ-CAMACHO J.; PEREZ-BERNAL F.; PITA-LOPEZ M.. A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. **International Journal of Climatology**. v. 28, p. 1525-1534, 2008.

RAO, C.R. TOUTENBURG, H. **Linear models**. 2nd. ed. New York: Springer-Verlag, 1999. 443p.

RIVERO, R.M., KOJIMA, M., SKAKIBARA, H., MITTLER, R., GEPSTEIN, S.; BLUMWALD E. Delayed leaf senescence induces extreme drought tolerance in a flowering plant**. PNAS**, v. 49, p. 19631-19636, 2007.

RUBIN, D.B. Multiple imputation after 18+ years. J**ournal of the American Statistical Association**, v. 91, p. 473–489, 1996.

SCHAFER, J.L.. **Analysis of incomplete multivariate data**. London: Chapman and Hall, 1997

SCHNEIDER, T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. **Journal of Climate**. v. 14, p. 853-871, 2001.

VICENTE-SERRANO, S.M.; BEGUERÍA-PORTUGUÉS, S. Estimating extreme dry-spell risk in the middle Ebro Valley (northeastern Spain): a comparative analysis of partial duration series with a general Pareto distribution and annual maxima series with Gumbel distribution, **International Journal of Climatology**, v. 23 p. 1103-1118, 2003.

VICENTE-SERRANO, S.M., SANTIAGO, B., LÓPEZ-MORENO, J.I., GARCÍA-VERA, M.A.; STEPANÉK, P. A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. **International Journal of Climatology**, v. 8, p. 1146-1163, 2010.