Fitting Precipitation Variability in Dobrudja Region

ALINA BĂRBULESCU
Department of Mathematics and Computers Science
Ovidius University of Constanța
124, Mamaia Bd., Constanța
ROMANIA
emma.barbulescu@yahoo.com http://www.math-modeling.ro/

JUDICAEL DEGUENON
Université d'Abomey – Calavi
École Polytechnique d'Abomey - Calavi,
01 BP 2009 Cotonou,
BENIN
tjudy73@yahoo.fr

Abstract: - An analysis of precipitation variability since 1965 has been carried out using annual and monthly data from Dobrudja, a region of situated in the South –East of Romania, between the Danube and The Black Sea. Performing the principal component analyses it has been proved that one of the series has a special behavior (due to its geographical position). Therefore, the annual precipitation evolution in the entire region can be modeled using the data provided by nine series, with small information loss. Finally, the regional precipitation evolution has been described by wavelets models using ten and nine series and the comparison between the results has been done.

Key-Words: - precipitation evolution, principal component analysis, wavelets

1 Introduction

Climate change is a topic of worldwide interest for all scientists. Analysing patterns, building models and testing their validity is a step in understanding and predicting the weather evolution [11][12][13].

The complexity of the problem of modeling meteorological time series derives from their non-linear behavior and to the lack of methods' adaptation. This makes the problem very well suited for the use of heuristic methods, which are more flexible. Therefore, neural networks [9], genetic algorithms [7] or hybrid approaches [1] [17] has been used to predict the precipitation evolution.

If building a good model at a local scale is a difficult problem, to describe the behavior of a meteorological phenomena at a regional scale becomes more complicated. Often, classical methods are not appropriate, the nonparametric approach providing an alternative in these situations [6] [15].

In this context, this article comes to complete the knowledge concerning the weather evolution in Dobrudja region [2] [4] proposing nonparametric models for the annual precipitation evolution, based on the data collected at ten meteorological stations in the period 1965 – 2005. Their geographical position can be seen in Fig. 1, where the two climatic units - one, influenced by the Black – Sea and another, influenced by the moderate continental belt - are also delimitated.

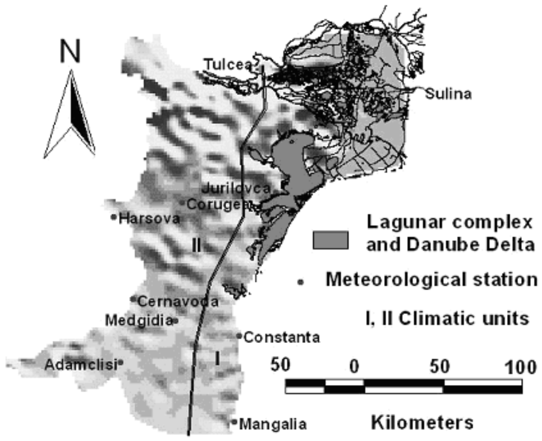

Fig. 1. The region of Dobrudja and the meteorological stations

The chart of mean annual series is represented in Fig. 2.

The interest of this study is not only theoretic, but also a practical one, taking into account the importance of knowledge on weather evolution in irrigations system design in a region where the frequency of droughty years is of 89 % [10].

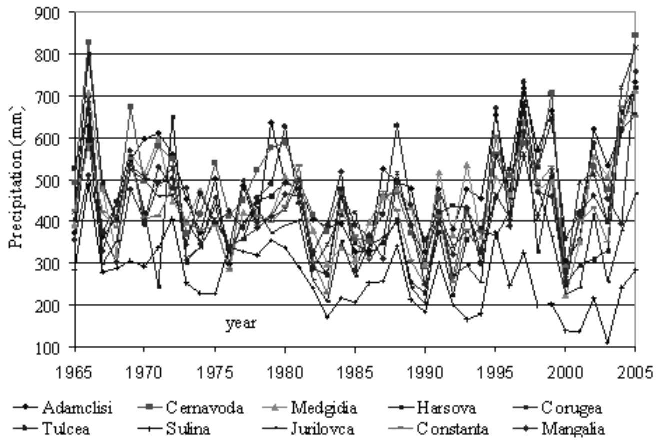

Fig. 2. Annual data series

2 Methodology

Nonparametric methods are statistical techniques that do not require a researcher to specify functional forms for objects being estimated. Such methods are becoming increasingly popular for applied data analysis. These methods are often deployed after common parametric specifications are found to be unsuitable for the problem at hand, particularly when formal rejection of a parametric model based on specification tests yields no clues as to the direction in which to search for an improved parametric model. The appeal of nonparametric methods stems from the fact that they relax the parametric assumptions imposed on the data generating process and let the data determine an appropriate model [14]. It is why we choose to model our data series using wavelets techniques [6].

Regression, of whatever kind, has two main purposes. Firstly, it provides a way of exploring and presenting the relationship between the design variable and the response variable; secondly, it gives predictions of observations yet to be made. A non-parametric method of estimation is desirable, because it does not force the model into a rigidly defined class. An initial non-parametric estimate may well suggest a suitable parametric model but nevertheless will give the data more of a chance to speak for themselves in choosing the model to be fitted [15].

Wavelets are special basis functions with two appealing features: can be computed quickly and the resulting estimators are spatially adaptive. This means we can accommodate local features in the data. The wavelets regression means to determine a model:

$$Y_i = f(x_i) + \sigma\varepsilon_i,$$

where $x_i = i/n$, ε_i is the residual and σ and f must be estimated.

Let $\{\psi_{j,k}\}_{j,k\in Z}$ be an orthonormal wavelets basis for $L^2(R)$ [6]. Any squared integrable function can be represented by:

$$f(x) = \sum_{k=-\infty}^{+\infty} \alpha_{0k}\phi_k(x) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty}\beta_{jk}\psi_{jk}(x),$$

where:

$$\phi_k(x) = 2^{1/2}\phi(2x-k),\ \psi_{jk}(x) = 2^{j/2}\phi(2^j x - k),$$

ϕ is a scaling function, ψ is the mother wavelets and

$$\alpha_{0k} = \int_{-\infty}^{+\infty} f(x)\phi_{0k}(x)dx,\ \beta_{jk} = \int_{-\infty}^{+\infty} f(x)\psi_{jk}(x)dx.$$

The classical nonlinear wavelets regression estimator is defined by:

$$\hat{f}(x) = \sum_{k=1}^{2^J-1}\hat{\alpha}_{0k}\phi_{0k}(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\hat{\beta}_{jk}\psi_{jk}(x),$$

where $\hat{\beta}_{jk}$ denotes the hard (or soft) threshold estimator.

In the case of longitudinal data analysis (as in our situation), the x_i - s are the successive time moments, t_i.

The steps of the wavelets smoothing procedure are:

- Determine preliminary estimate:

$$\tilde{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} y_i\psi_{jk}(t_i),$$

- Shrink: $\hat{\beta}_{jk} \leftarrow \text{shrink}(\tilde{\beta}_{jk})$;
- Reconstruct the function $\hat{f}$.

In practice, the preliminary estimates are computed using the discrete wavelet transform. Two types of shrinkage are used: soft threshold and hard threshold. The last one has been applied in this study.

The method most commonly used in climate data analysis for the dimensionality reduction is *principal component analysis* (PCA) which deals with an eigen decomposition of the input data covariance matrix. PCA is widely applied to transform data into independent PCs to reduce the numbers of variables by several leading PCs that explain a large proportion of the total variance. PCA builds decorrelated components of the data and it finds the spatial patterns that maximize the variance. Hence, second order moments are the foundation of PCA.

In our study PCA was used to reduce the number of data series that participate as input data to the regional model of precipitation evolution [16].

3 Results

The normality tests (Kolmogorov – Smirnov, Shapiro – Wilk, Q-Q plot), the correlation study (by autocorrelation function and Box-Ljung test), the break tests (Pettitt, Buishand, Lee and Heginian tests, Hubert segmentation procedure and CUSUM procedure) and the homoscedasticity test (Levene) have been preliminary performed and the results are presented in [5]. For the series that were not Gaussian noises, different models have been determined, using Box – Jenkins techniques [4], gene expression programming techniques and hybrid methods, AdaGEP - AR [3].

Since our final goal is to obtain a model for the evolution of precipitation in the entire region of Dobrudja, based on the ten data series, the results obtained till now were analyzed, in order to determine a convenient modeling technique. The Box-Jenkins methodology wasn't found appropriate and the generalized additive models didn't give good results. Therefore, we decided to use a nonparametric approach - wavelets and smoothing splines.

After performing the variance analysis, we found enough evidence to reject the hypothesis that there is no difference between the means of the precipitation series. To emphasize the station whose mean is statistically different from the other stations' means, the Tukey HSD and Scheffé' s tests have been performed at the level of significance of 1%. In Table 1, we present the results of Scheffé's test.

Table 1. Results of the Scheffé test

Series	Subsets for α = 0.01	
	1	2
Sulina	261.6341	
Jurilovca	378.3927	378.3927
Hârșova		408.8220
Constanța		423.0390
Mangalia		427.7366
Corugea		434.6659
Tulcea		434.6659
Medgidia		449.9244
Adamclisi		484.5415
Cernavodă		487.5951
p-value	0.012	0.029

Removing Sulina series, and performing again the Scheffé test, we didn't find enough evidence to reject the hypothesis that the means of the nine series are different.

Therefore, the PCA was applied to reduce the number of data series used to determine the regional model of annual precipitation evolution.

The fact that Sulina series forms a distinct group is in concordance to the particular position of the hydro-meteorological station. Here, we mention only that it is situated 13 km offshore, in the Danube Delta, so its climate is influenced by the Danube and the Black Sea, differing from those of the other meteorological stations (situated inside the Dobrudja region).

From the eigenvalues scree (Fig. 3) we deduce that two principal components are necessary to extract the essential information from the data series, the biggest part of the explained variance being ascribed to Adamclisi series (70.52%) and only 1.75%, respectively 1.54% to Sulina and Jurilovca.

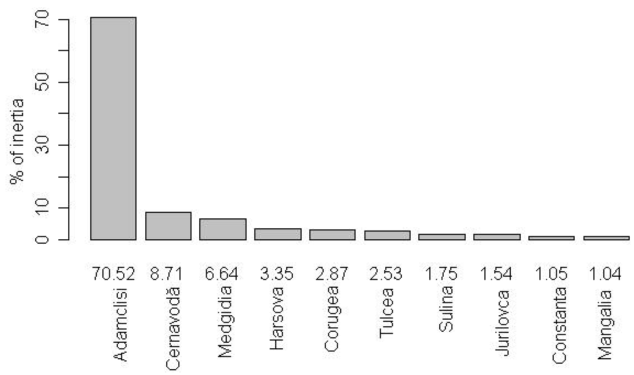

Fig. 3. The eigenvalues scree (%) for annual series

The results of PCA are represented in the plan of principal components of variables - stations, in our case - (Fig. 4), in that of the individuals – years, in our case - (Fig. 5), and simultaneous (Fig. 6).

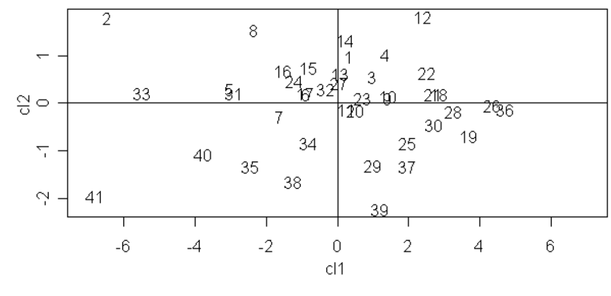

Fig. 4. The distribution of the *i* - th year in the period 1965 – 2005 on the second principal component

Interpreting Fig. 5, we conclude that Sulina series has the smallest contribution on precipitation explanation on the first component. Therefore, the model for general precipitation evolution in Dobrudja region can be designed after the elimination of this series, with small loss of information.

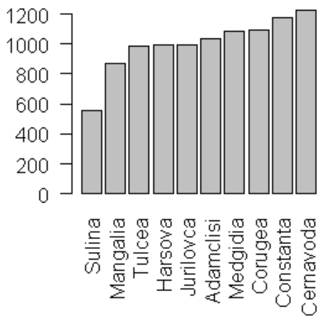

Fig. 5. Profile of stations' contribution on first axis

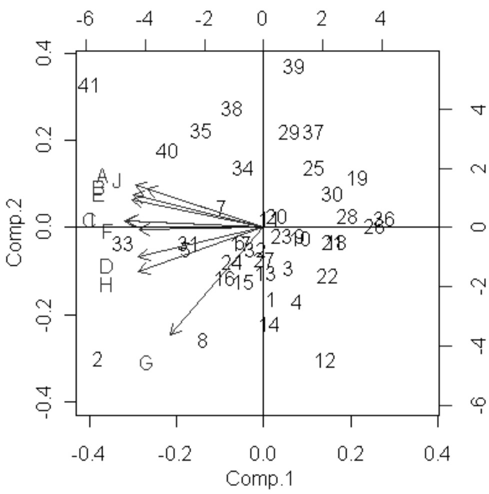

Fig. 6. Simultaneous representation of years and stations axis - the biplot (A – Adamclisi, B – Medgidia, C – Cernavoda, D – Harsova, E – Corugea, F – Tulcea, G – Sulina, H – Jurilovca, I – Constanta, J – Mangalia)

For comparison reasons models have been built using the ten series, as well as with the nine series remain after the removal of Sulina, group called in the following Group 1.

The models are presented in Figs. 7 and 8 and the residuals, in Figs. 9 and 10. They were obtained using the R software.

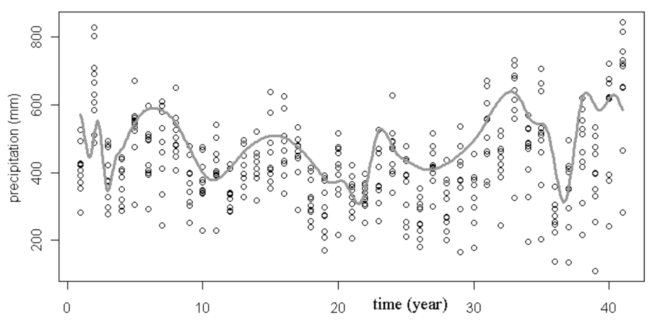

Fig. 7. The model for the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using ten series

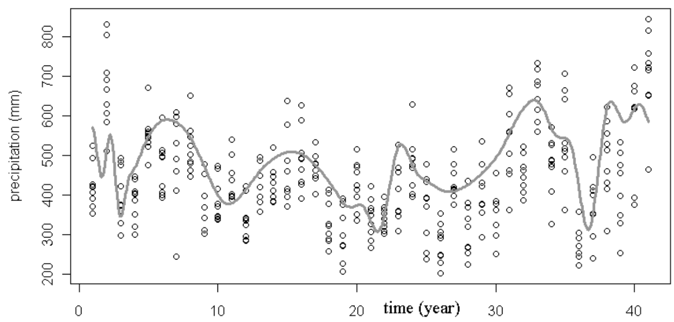

Fig. 8. The model of the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using Group 1

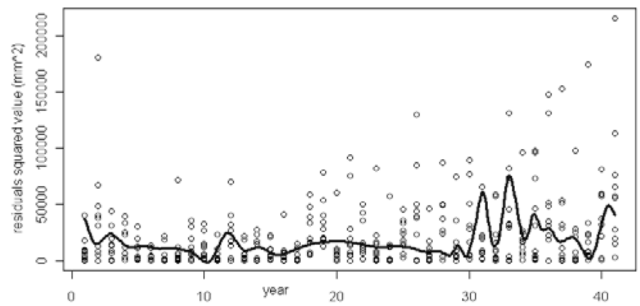

Fig. 9. Residual in the model of the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using ten series

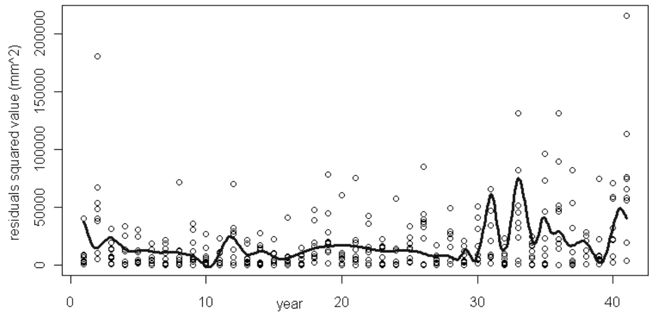

Fig. 10. Residual in the model of the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using Group 1

The residual standard deviations don't differ significantly (Table 2), coming to confirm once again the preliminary results of PCA.

The p – value calculated for the second model is slightly bigger then for the first one, proving that the second model is better than the first one.

For the same groups of series, the smoothing splines method was also applied, to obtain a model for the global evolution of precipitation in Dobrudja region. The comparison between the results is presented in Table2.

Table 2. Comparison of residual standard deviation from wavelets method and smoothing splines applied to the annual data

	Wavelets method		Smoothing splines	
Series	All	Group 1	All	Group 1
std. dev.	139.57	132.58	87.64	67.93

Analyzing the model, we remark some periodic oscillations of precipitation evolution in the studied period, suggesting a possible parametrical model with harmonic components.

4 Conclusion

In this article we described the global evolution of annual precipitation in Dobrudja for the period 1965 - 2005, by wavelets method.

Since Sulina series has particular characteristics, due to its geographical situation (13 km offshore, in the Danube Delta), it was removed, without loss of information, from the data set taken into account in the modeling process. This result is also confirmed when another nonparametric method – smoothing splines – is used. We didn't present in details the last result, but we mention that in the case of the global models obtained by splines regression, a decreasing of residual variance of 9% can be reported, after the removal of Sulina series from the data set.

The main advantage of wavelet shrinkage is that it is highly adaptive to irregular signals as well as smooth ones, so it can be used, without restriction concerning the distribution of the time series. The nonparametric models obtained suggested a new parametrical model, with a harmonic trend, that will be presented in another article.

Finally, we mention that the extremes are better captured by this type of models, contributing to the increasing of their prediction accuracy.

References:

[1] A. Bărbulescu, E. Băutu, Time Series Modeling Using an Adaptive Gene Expression Programming Algorithm, *International Journal of Mathematical Models and Methods in Applied Sciences*, 3(2), 2009, pp. 85 – 93.

[2] A. Bărbulescu, E. Băutu, ARIMA Models versus Gene Expression Programming in Precipitation Modeling, *Recent Advances in Evolutionary Computing*, 2009, pp. 112 – 117.

[3] A. Bărbulesc, E. Băutu, Alternative models for time series, *An. Şt. Univ. "Ovidius" Constanța*, 17(3), 2009, pp. 45 – 68.

[4] A. Bărbulescu, E. Pelican, ARIMA models for the analysis of the Precipitation Evolution, *Recent Advances in Computers*, 2009, pp. 221 – 226.

[5] A. Bărbulescu, C. Şerban (Gherghina), C. Maftei, Statistical Analysis and Evaluation of Hurst Coefficient for Annual and Monthly Precipitation Time Series, *WSEAS Transactions on Mathematics*, Issue 10, vol. 9, Oct. 2010, pp. 791 – 800.

[6] I. Daubechies, *Ten lectures in wavelets*, SIAM, 2002

[7] I. De Falco, A. Della Cioppa, E. Tarantino, A Genetic Programming System for Time Series Prediction and Its Application to El Niño Forecast, *Advances in Soft Computing*, 32, 2005, pp. 151 – 162.

[8] J. Fox, S. Weisberg, *An R Companion to Applied Regression*, Second Edition, Sage Publications, 2011

[9] R. J. Kuligowski, A. P. Barros, Experiments in short term precipitation forecasting using artificial neural networks, *Monthly Weather Review*, 126, 1998, pp. 470 – 482.

[10] C. Maftei, A. Bărbulescu, Statistical analysis of climate evolution in Dobrudja region, *Lecture Notes in Engineering and Computer Sciences*, WCE 2008, vol. II, 2008, pp.1082 – 1087.

[11] V. Masson - Delmotte et al., Changes in European precipitation seasonality and in drought frequencies revealed by a four-century-long tree-ring isotopic record from Brittany, western France, *Climate Dynamic*, 24, 2005, pp. 57–69.

[12] T. D. Mitchell, P. D. Jones, An improved method of constructing a database of monthly climate observations and associated high-resolution grids, *International Journal of Climatology*, 25, 2005, pp. 693–712.

[13] B. Qian, H. Xu, J. Corte-Real, Spatial - temporal structures of quasi-periodic oscillations in precipitation over Europe, *International Journal of Climatology*, 20, 2000, pp. 1583–1598.

[14] J. S. Racine, Nonparametric Econometrics: A Primer, *Foundations and Trends in Econometrics*, 3 (1), 2008, pp. 1– 88.

[15] B. W. Silverman, Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting, *J. R. Statist. Soc. B*, 47(l), 1985, pp. 1- 52.

[16] L. Smith, A tutorial on Principal Components Analysis, 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[17] G. P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing*, 50, 2003, pp. 159 – 175.