Nonparametric methods for fitting the precipitation variability, applied to Dobrudja region

Alina Bărbulescu and Judicael Deguenon

***Abstract*—** Modeling the precipitation evolution at a regional scale is a topic of interest in order to predict the weather evolution and climate change, with their multiple consequences, in a region where the annual drought frequency is about 89%, as in Dobrudja. Therefore, in this article we present the result of modeling the annual and monthly precipitation evolution in the period January 1965 - December 2005, in Dobrudja, a region of situated in the South – East of Romania, between the Danube and The Black Sea. ANOVA, followed by the Tukey LSD and the Scheffe tests led us to the conclusion that between the ten studied series, one has a particular behavior, due to the geographical position of the meteorological station. Thus, the precipitation evolution in the entire region has been modeled using the data provided by the other nine series, with small loss of information. The models proposed by us used nonparametric approach (wavelets and smoothing splines) and the comparison between them has been done.

***Keywords* —** precipitation evolution, principal component analysis, wavelets, smoothing splines.

I. Introduction

CLIMATE change is a topic of worldwide interest for all scientists. Analyzing 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] [18] has been used to predict the precipitation evolution.

If building a good model at local scale is a difficult problem, to describe the behavior of meteorological phenomena at regional scale becomes more complicated. Often, classical methods are not appropriate, the nonparametric approach providing a valuable alternative [6] [16].

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 and monthly 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. The chart of partial mean monthly series is represented in Fig. 2.

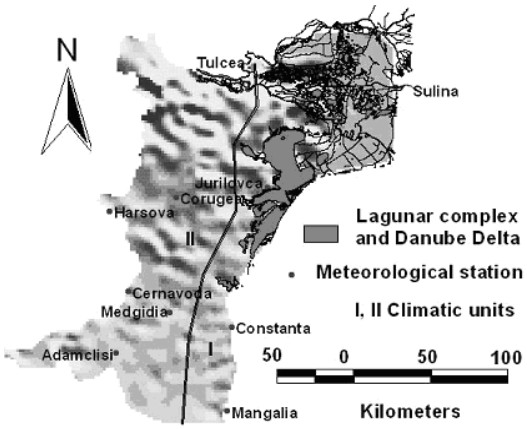

Fig. 1. The region of Dobrudja and the meteorological stations

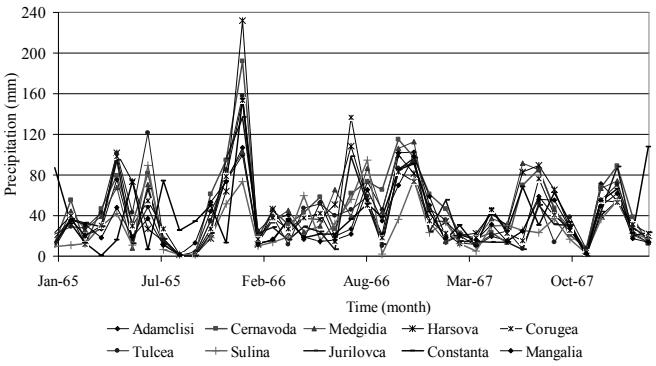

Fig.2. Monthly data series (partial representation)

Manuscript received April 30, 2012: Revised version received -

Alina Bărbulescu is with Ovidius University of Constanța, Faculty of Mathematics and Computers Science, 124, Mamaia Bd., 900527, Constanța, Romania (corresponding author: phone: 0040 744 284444, e-mail: emma.barbulescu@yahoo.com).

Judicael Deguenon is with Université d'Abomey – Calavi, École Polytechnique d'Abomey - Calavi, 01 BP 2009 Cotonou, Benin (e-mail: tjudy73@yahoo.com).

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].

II. 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 [16].

Smoothing splines arise as the solution to the following simple-regression problem: Find the function $\hat{f}(x)$ with two continuous derivatives that minimizes the penalized sum of squares

$$J(\lambda)=\sum_{i=1}^{n}\left|y_i-f(x_i)\right|^2+\lambda\int_{x_{\min}}^{x_{\max}}\left[f''(x)\right]^2dx,$$

where λ is a smoothing parameter and $f''(x)$ is the second derivative of $f(x)$.

The smoothing parameter may be selected by minimizing the mean-squared error of the fit, either by employing a formula approximating the mean-square error, or by some form of cross-validation.

In cross-validation, the data are divided into subsets; the model is successively fit omitting each subset in turn; and then the fit model is used to 'predict' the response for the left-out subset [8].

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:

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

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:

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

- Shrink: $\hat{\beta}_{jk}\leftarrow\text{shrink}(\widetilde{\beta}_{jk})$;

- 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 used 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 [17].

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

III. Results

The normality tests (Kolmogorov – Smirnov, Shapiro – Wilk, Q-Q plot), the correlation study (by using autocorrelation function and the Box-Ljung test), the break tests (the Pettitt, Buishand, Lee and Heginian tests, the Hubert segmentation procedure and the CUSUM procedure) and the homoscedasticity test (Levene) have been performed and the results are presented in [5].

For the series that were not Gaussian noises, different models have been determined, by using Box – Jenkins techniques [4], gene expression programming techniques and hybrid methods, AdaGEP - AR [3].

Since our 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 nonparametric approaches - 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 the Scheffé tests [15] have been performed at the level of significance of 1%. In Table I, we present the results of the Scheffé test on the precipitation mean values, at the level of significance of $\alpha = 0.01$. We remember that in the hypothesis testing, the p-value may be defined as the lowest significance level at which the null hypothesis can be rejected.

Table I. Results of the Scheffé test

Series	Subsets for $\alpha = 0.01$	
	1	2
Sulina	261.6341	
Jurilovca	378.3927	378.3927
Hârşova		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 Principal Component Analysis (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 this 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.

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).

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.

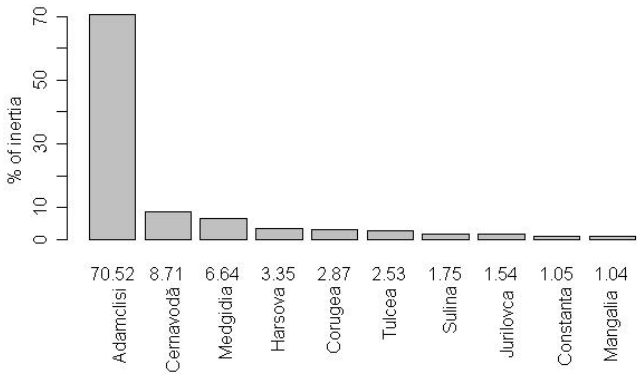

Fig.3. The eigenvalues scree (%) for annual series

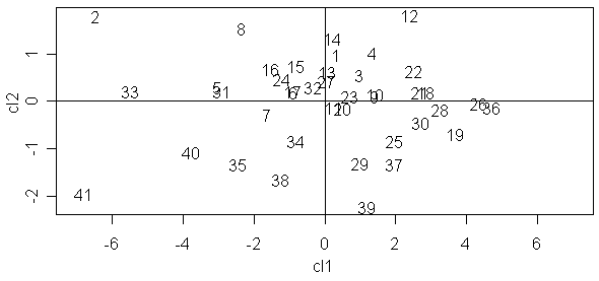

Fig. 4. The distribution of the *i* - th year in the period 1965 – 2005 on the second principal component

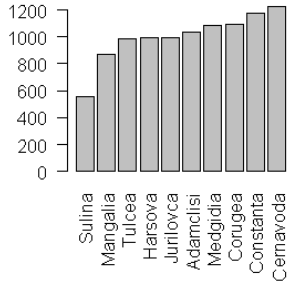

Fig. 5. Profile of stations' contribution on the first axis

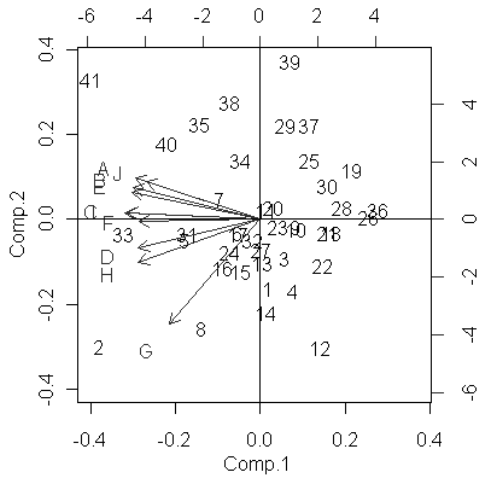

Fig. 6. Simultaneous representation of years and stations axis - the biplot
(A – Adamclisi, B – Medgidia, C – Cernavoda, D – Harsova, E – Corugea, F – Tulcea, G – Sulina, H – Jurilovca, I – Constanta, J – Mangalia)

3.1. Models for the annual precipitation evolution in Dobrudja region

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

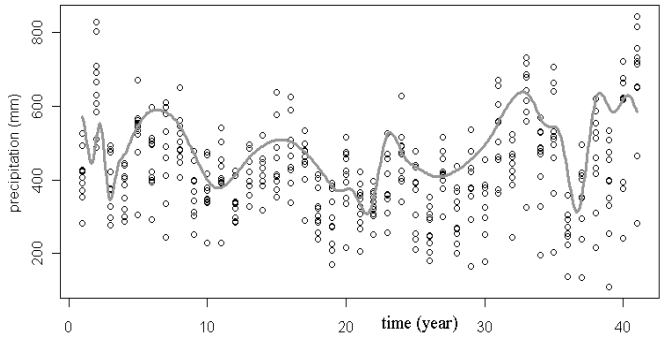

Fig. 7. The model for the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using ten series

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

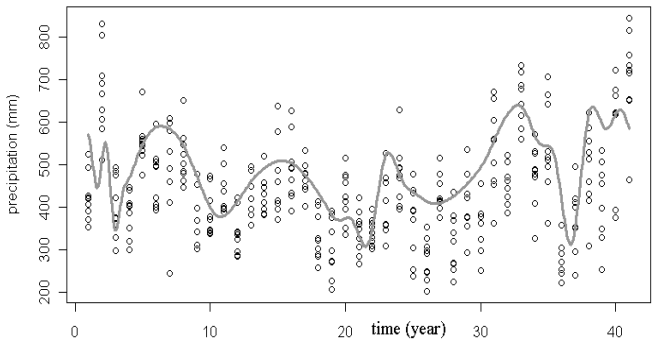

Fig. 8. The wavelets model for the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using Group 1

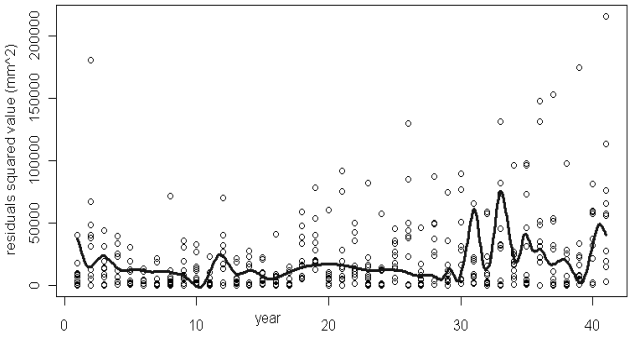

Fig. 9. Residual in the model for the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using ten series

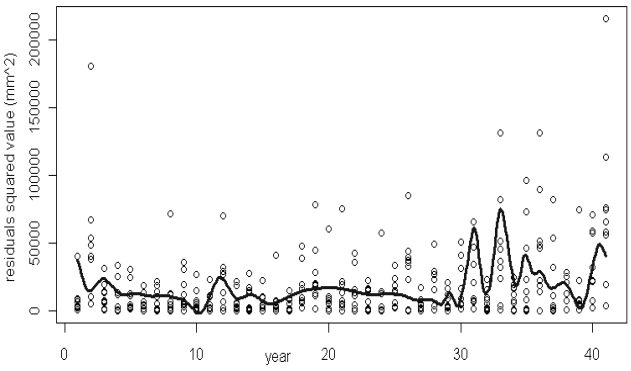

Fig. 10. Residual in the model for the precipitation variability in Dobrudja region, obtained by wavelets method (hard threshold) using Group 1

The residual standard deviations don't significantly differ. They are respectively 11.81 and 11.51, 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.

Analyzing the models, we remark some periodic oscillations of precipitation evolution in the studied period, suggesting a possible parametrical model with harmonic components. The comparison between the results is presented in Table II.

Table II. 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

3.2. Models for the monthly precipitation evolution in Dobrudja region

The evolution of monthly series was also described by different models, using gene expression programming (GEP) and a combined algorithm ARIMA – GEP [2], [3].

To find the homogenous groups of monthly series, Scheffe's test was applied (Table III).

Table III. Sheffe's test for monthly data

Series	Subsets for α = 0.01	
	1	2
Sulina	21.803	
Jurilovca		31.533
Harsova		34.069
Constanta		35.253
Mangalia		35.645
Corugea		36.222
Medgidia		37.494
Tulcea		38.487
Adamclisi		40.379
Cernavoda		40.633
Sig.	1.000	0.131

It results that Sulina series is the single member of one group, so to model the evolution of the monthly mean precipitation over Dobrudja, at least this series can be removed.

The results of principal components analysis of the monthly data are presented in Figs. 11 – 15.

The conclusion of the principal component analysis is that Sulina series has the smallest contribution to the description of the general precipitation variability in Dobrudja region. Therefore, for comparison reason, the models were built for the entire data and for the series without Sulina.

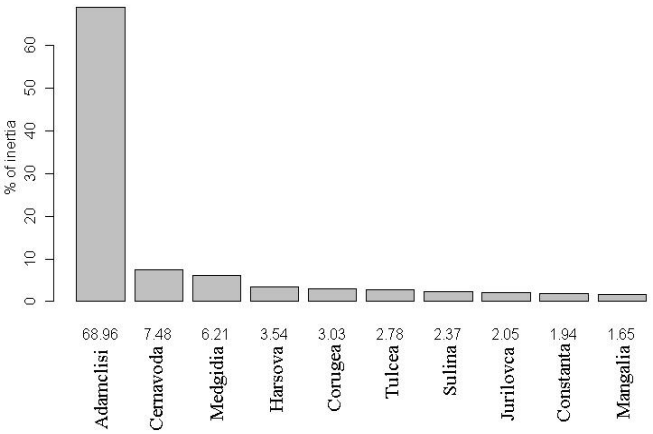

Fig.11. The eigenvalues scree (%) for monthly data

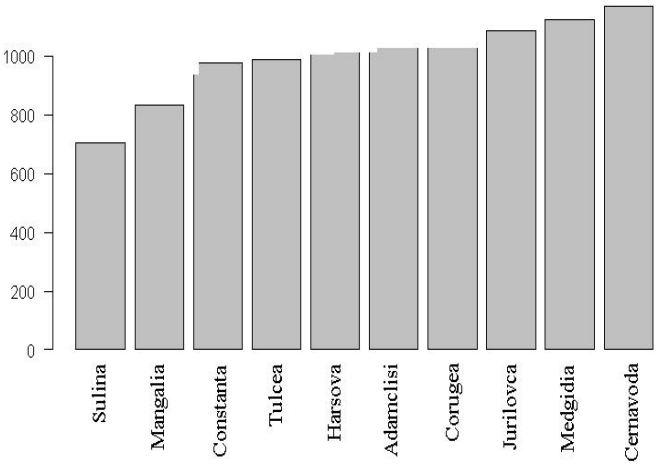

Fig.12. Profile of stations contribution on first axis (monthly data)

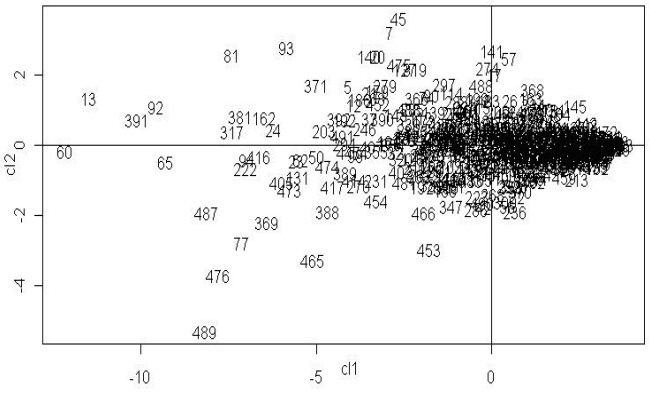

Fig.13. The distribution of the i - th month in the period January 1965 – December 2005 on the second principal component

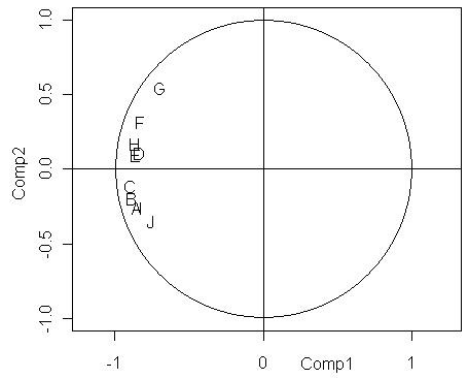

Fig.14. The stations contributions – correlation circle for monthly data
(A - Adamclisi, B – Medgidia, C – Cernavoda, D – Harsova, E – Corugea, F – Tulcea, G – Sulina, H – Jurilovca, I – Constanta, J – Mangalia)

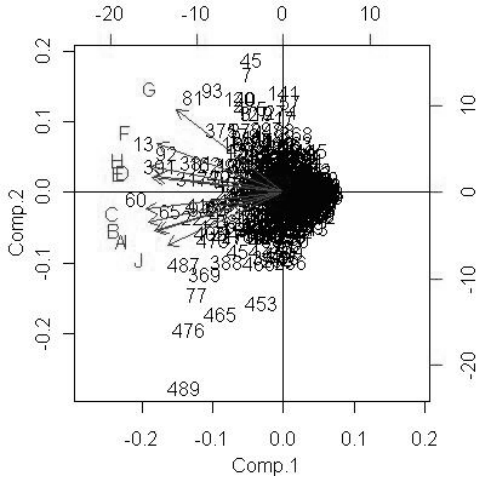

Fig. 15. The biplot of monthly data

In the models and the residual obtained by wavelets methods are presented in Figs. 16 - 19, and those obtained by smoothing splines, in Figs. 20 - 23. The months are numbered from 1 to 492, where 1 corresponds to January 1965 and 492 to December 2005.

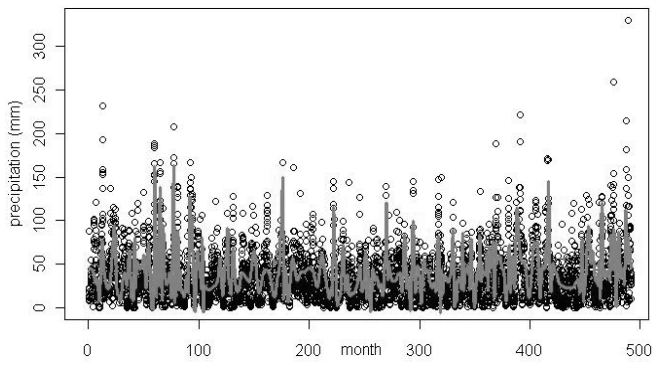

Fig. 16. Wavelets method (hard threshold) – monthly series

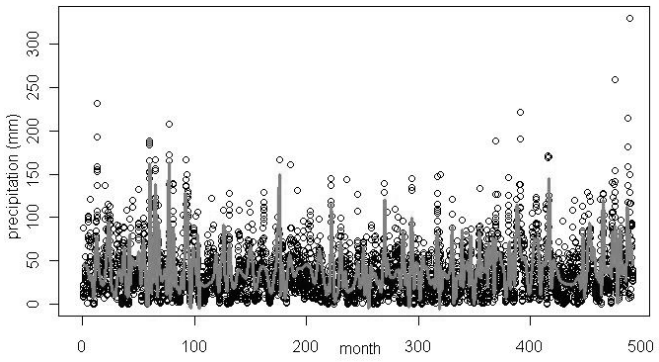

Fig. 17. Wavelets method (hard threshold) – monthly series, without Sulina

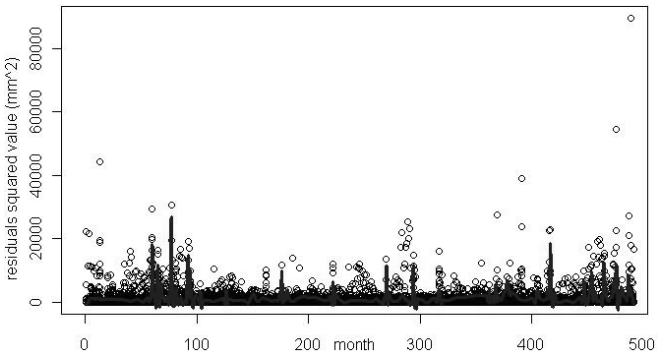

Fig. 18. Residual in wavelets model (hard threshold) – monthly series

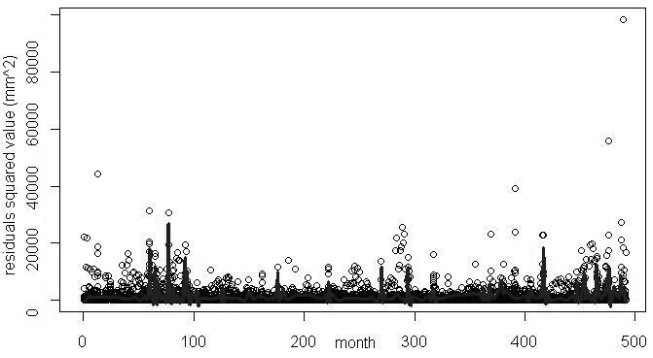

Fig.19. Residual in wavelets model (hard threshold) – monthly series, without Sulina

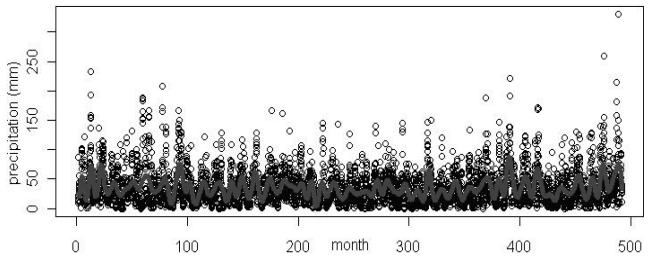

Fig. 20. Splines smoothing – monthly series

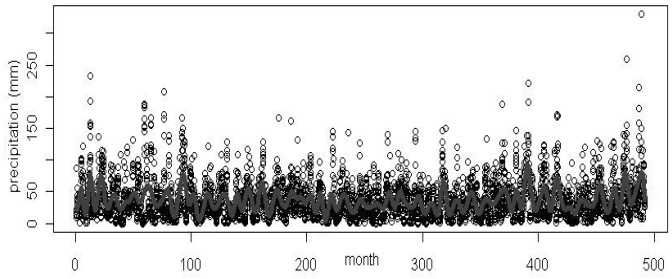

Fig. 21. Splines smoothing – monthly series, without Sulina

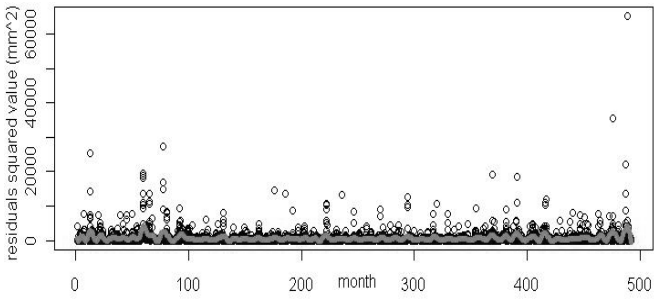

Fig. 22. Residual in splines smoothing – monthly series

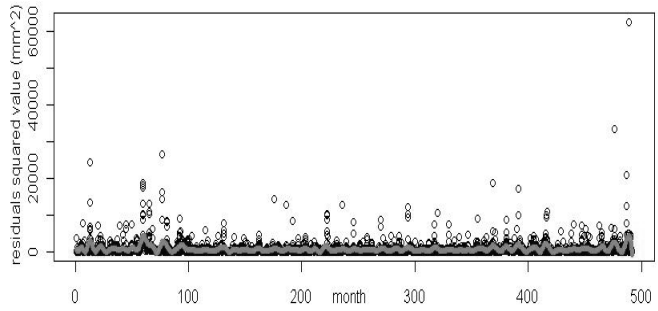

Fig. 23. Residual in splines smoothing – monthly series, without Sulina

The trends (respectively the residual), are similar when we consider all the ten series and when we remove Sulina, in the case of wavelets model. The same happens when the smoothing splines regression was used.

Analyzing the trends from the Figs. 16 and 20 (respectively Fig. 17 and 21), we remark that the seasonality is revealed in both cases, but the curve is smoother when the smoothing splines technique has been used. Also, the last procedure produces stationary residuals, as it can be seen in Figs. 22 and 23.

The overall residual standard deviations have also been calculated and are presented in Table IV.

They are sensibly higher in the case of wavelets method, but the removal of Sulina series from the set of studied series doesn't change the results, so for the modeling purposes, it can be neglected without loss of information.

This remark is also confirmed by the p-values calculated, that are very close one to the other.

Table IV. Comparison of residual standard deviation from wavelets method on monthly data

Method	Wavelets		Smoothing splines	
Series	All	without Sulina	All	without Sulina
Std. dev.	37.27	37.78	26.13	26.34

IV. Conclusion

In this article we described the global evolution of annual and monthly precipitation in Dobrudja for the period 1965 - 2005, by wavelets and smoothing splines methods.

Since Sulina series has particular characteristics, due to its geographical situation (13 km offshore, in the Danube Delta), it was removed, from the data set taken into account in the modeling process, with small loss of information. This result is also confirmed when another nonparametric method – smoothing splines – was used. We mention that in the case of the global models obtained by splines regression for annual data, a decreasing of residual variance of 9% can be reported, after the removal of Sulina series from the data set.

For the monthly data, a decreasing of 30% of the residual variance has been registered, when the smoothing splines method has been used for modeling.

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*, vol. 3, no. 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. Bărbulescu, E. Băutu, "Alternative models for time series", *An. Șt. 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. Șerban (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] H. Scheffé, *The Analysis of Variance*, Wiley, New York, 1959
[16] 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.
[17] L. Smith, A tutorial on Principal Components Analysis, 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
[18] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", *Neurocomputing*, 50, 2003, pp. 159 – 175.