Modelling the annual precipitation evolution in the region of Dobrudja

Alina Bărbulescu and Judicael Deguenon

***Abstract*—** In this article we describe the regional evolution of precipitation in Dobrudja, a region situated in the South – East of Romania, using the mean annual precipitation collected from 1965 to 2005, at ten hydro-meteorological stations. Firstly, the statistical analyzes of the ten series have been performed. ANOVA followed by the Tukey and the Scheffé tests reveal possible grouping of homogenous data series. Since Sulina series has a particular behaviour, due to its geographical position, models for the precipitation evolution have been built using all the series or only nine (without Sulina) and comparisons have been done. The modelling techniques used were nonparametrics - local linear smoothing and smoothing splines.

***Keywords* —** linear model, nonparametric, precipitation evolution, smoothing.

I. Introduction

Recent studies on weather variations [4], [8] emphasized an increasing precipitation rate of 20 - 40%, in Northern Europe and a decreasing one of 10 - 40% in the South of continent. It was stated that annual zonal average precipitation increased about 7% to 12% for the zones from 30ºN to 85ºN in emerged landmasses in the Northern hemisphere, excepting the Far East [13] [19].

Analyzing patterns, building models and testing their validity is a step in understanding and predicting the weather evolution. The complexity of the problem of modeling meteorological time series derives from the diversity of phenomena that generally affect the climate. Such time series often show non-linear behavior, their analysis constituting a topic of substantial interest in the literature [8] [12] [18]. Changes in the environment may trigger shifts in the process describing the time series. Classical approaches such as the linear model rely on the assumption of a constant data generating process. Often they may fail to obtain adequate models due to the nonlinear dynamic behavior of time series and to the lack of methods adaptation. This makes the problem very well suited for the use other methods, as neural networks, genetic algorithms or hybrid approaches [2] [20] [22].

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 [20]. They are often deployed after the rejection of a parametric model based on specification tests.

To understand the climate variability, a number of studies [4][17] reported the results concerning the analysis of temperature evolution in the Black Sea region, where Romania and particularly Dobrudja is situated. A systematic analysis of precipitation evolution in Dobrudja region has also been done in [1] [3] [5] [6]. Different methods (neural networks, Box – Jenkins techniques, gene expression programming and hybrid algorithms) have been used to determine models for the individual series. We mention that the gene expression programming method gave similar results as ARMA models, for the annual series; for longer series, the adaptive gene expression programming (AdaGEP) or the hybrid model AdaGEP – AR performed better.

Moving from the local to the regional approach, in this article we present the results of modelling the general evolution of precipitation in Dobrudja, obtained by nonparametric methods [11].

The paper is organized as follows. Section II summarizes the theoretical results used in the modeling. Section III gives the methods used for the precipitation modeling and Section IV, the models for the precipitation series. The last section is the concluding remarks.

II. Dobrudja and the Data Base

By its physical and geographical characteristics, Dobrudja represents a special unit of Romanian territory, being a structural and petrographic mosaic, characterized by an accentuated non-uniformity and variety of its active surface. The Black Sea and the Danube have a major influence on the climatic characteristic of this area.

The circulation of atmospheric masses influences the repartition of annual precipitation quantities, which registered small values. The atmospheric masses generally circulate from west to east. The big aquatic unities from the West of Danube and from the East of the Black Sea have the role of thermic barrage. The precipitation decreases from Danube to the Sea,

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).

in Dobrudja plateau, and it increases in the Southern part of region, on the direction from North - East to South - West. The differences in quantities are also due to the relief aspect and to the repartition and structure of vegetation.

The data studied in this article is the annual precipitation series (Fig.1), collected in the period 1965-2005 at ten meteorological stations from Dobrudja region, whose coordinates are given in Table I.

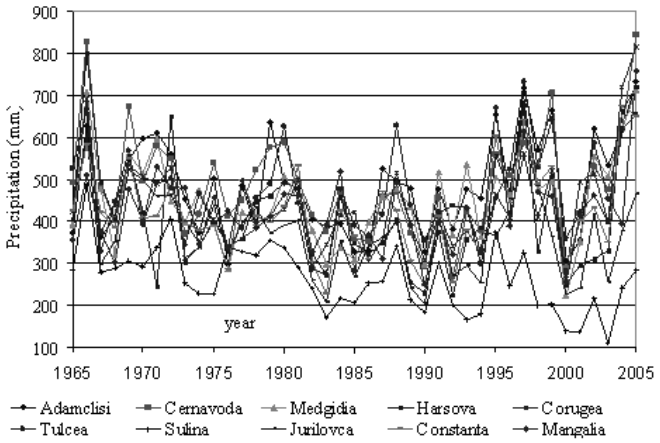

Fig. 1. Annual data series

Table I. The coordinates of meteorological stations

	Station	Lat	Long	Elev. (m)	1965 - 2005 av. precip. (mm)
1	Adamclisi	44:08	28:00	158	484.54
2	Cernavoda	44:21	28:03	87.17	487.60
3	Medgidia	44:15	28:16	69.54	449.92
4	Harsova	44:41	27:57	37.51	408.82
5	Corugea	44:44	28:20	219.2	434.67
6	Tulcea	45:11	28:49	4.36	461.84
7	Sulina	45:09	29:39	2.08	261.63
8	Jurilovca	44:46	28:53	37.65	378.39
9	Constanta	44:13	28:38	12.8	423.04
10	Mangalia	43:49	28:35	6	427.74

III. Methodology

Before the modelling process, data has been checked for accuracy. Different tests (normality, homoscedasticity, independence etc.) have been performed on each series [6].

Since our aim is to describe the regional evolution of precipitation, the model building was preceded by the homogeneity analysis of time series. Therefore, the steps in our modelling procedure were:

I. *Determining the differences* between the precipitation data collected at different hydro-meteorological stations.

For this, Levene test was used to test the null hypothesis that the within – group variances are constant across groups [16].

The null hypothesis is:

$$H_0:\ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2,$$

and its alternative is:

$$H_1:\ \sigma_i^2 \neq \sigma_j^2, \text{ for at least one pair } (i, j).$$

The test' statistic is:

$$W = \frac{n-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (\overline{Z_{i.}} - \overline{Z_{..}})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{ij} - \overline{Z_{i.}})^2},$$

where:

- n is the sample volume (410. in our case),
- k is the number of groups in which the sample is divided ($k = 10$, in this case),
- n_i is the sample size of each group ($n_i = 41$, in this case),
- X_{ij} is the element j in the i - th group,
- $\overline{X_{i.}}$ is the mean of the i-th group,
- $Z_{ij} = \left| X_{ij} - \overline{X_{i.}} \right|$,
- $\overline{Z_{i.}}$ is the mean group of Z_{ij},
- $\overline{Z_{..}}$ is the overall mean of Z_{ij}.

The null hypothesis is rejected at the significance level α, if $W > F_{\alpha, k-1, n-k}$, where $F_{\alpha, k-1, n-k}$ is the upper critical value of the F distribution with k - 1 and $n - k$ degrees of freedom at the significance level of α.

II. Performing *one-way analysis of variance* (ANOVA), to compare the means of the ten data series.

ANOVA tests the null hypothesis that samples in two or more groups are drawn from the same population, using the Fisher distribution, F. To do this, two estimates of the population variance are made, on the assumptions:

- Response variable is normally distributed (or approximately normally distributed),
- Samples are independent,
- Variances of populations are equal,
- Responses for a given group are independent and identically distributed normal random variables.

If the group means are drawn from the same population, the variance between the groups' means should be lower than the variance of the samples. A higher ratio therefore implies that the samples were drawn from different populations [14].

ANOVA is a relatively robust procedure with respect to violation of the normality assumption.

If a significant result has been obtained from an overall *F* – test, investigators often wish to undertake further tests (for example, Tukey HSD and Sheffe or LSD) to determine which particular group means differ [15].

III. Fitting *the precipitation evolution* by different methods:

- Mean multi-annual model,
- Kernel estimation smoothing [10],
- Local linear smoothing [7],
- Smoothing splines.

Nonparametric models can provide accurate methods of data analysis because they make minimal assumptions about the data - generating process. For example, *nonparametric regression* provides a method to estimate the unknown regression curve representing the actual relationship between a covariate and the outcome variable, without making assumptions that the true curve in the population is linear. Therefore, using a nonparametric model to describe the evolution a process of phenomena is a good option since there are fewer restrictions on the data than in the case of using parametric models.

These methods are nonparametric, since they trace the dependence of a response variable (y_i) on one or several predictors without specifying in advance the function that relates the response to the predictors:

$$E(y_i) = f(x_i)$$

where $E(y_i)$ is the mean for the i^{th} of *n* observations.

Nonparametric regression is therefore distinguished from linear regression, in which the function relating the mean of y_i to the x_i is linear in the parameters,

$$E(y_i) = \alpha + \beta x_i ,$$

and from traditional nonlinear regression, in which the function relating the mean of y_i to the x_i, though nonlinear in its parameters, is explicitly specified:

$$E(y_i) = f(x_i, \gamma) .$$

There is a large literature on nonparametric regression analysis. For more extensive introduction to the subject see [9][10].

In the mean *multi – annual model*, the precipitation variation is a described by a polygonal line, with the vertices $(t_i, \overline{y}_i)$, where $\overline{y}_i$ is the average of annual mean precipitations, registered at all the stations in the year t_i. The $\overline{y}_i$ can be calculated as simple or weighted means, by *Thiessen's polygons method* [17].

A kernel smoother uses a set of weights, defined by a kernel, to produce the estimate at each target value. The weight given to the *j* - th point in producing the estimate at y_0 is defined by $c_0 / \lambda K\left[\left|\frac{x-x_0}{\lambda}\right|\right]$, where $K(t)$ is the kernel, λ - the window – width (bandwidth), and c_0 is a constant, usually chosen so that the weights sum to unity [10].

Local polynomial smoothing [9] is a method in which at each point in the data set, a low-degree polynomial is fit to a subset of the data, with explanatory variable values near the point whose response is being estimated. The polynomial is fit using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The value of the regression function for the point is then obtained by evaluating the local polynomial using the explanatory variable values for that data point.

If the order of the local polynomial is 1, we discuss about a local linear fit.

A cubic smoothing spline is a function $s(t)$ that minimizes

$$J(\lambda) = \sum_{i=1}^{n} |y_i - s(t_i)|^2 + \lambda \int [s''(t)]^2 dt ,$$

where λ is a roughness penalty and $s''(t)$ is the second derivative of $s(t)$.

IV. Results

I. The precipitation' box plot is presented in Fig. 3. A number was assigned (Subject_ID) to each station as follows: 1– Adamclisi, 2 – Cernavodă, 3 – Medgidia, 4 – Hârșova, 5 – Corugea, 6 – Tulcea, 7 – Sulina, 8 – Jurilovca, 9 – Constanta, 10 – Mangalia.

Analysing the chart we remark that that eight series present two or three outliers.

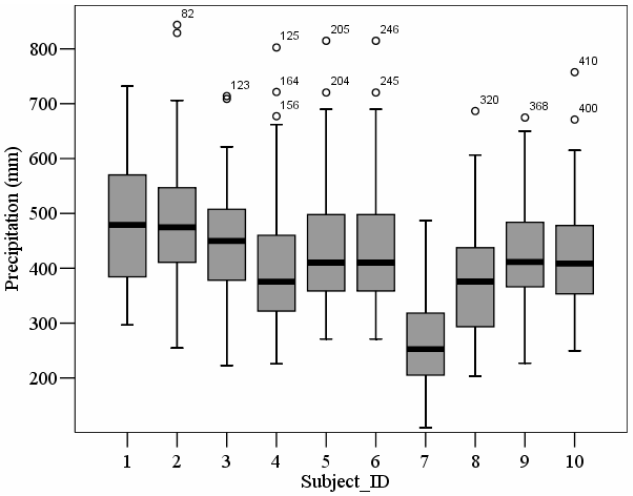

Fig. 3. Box plot of annual amount of precipitation in mm of 10 stations in Dobrudja for the period 1965 - 2005

The result of Levene's test (Table II) is consistent with our examination of the box plots. We didn't find enough evidence for a departure from the homogeneity assumption since the *p* - value, Sig.= 0.49 > 0.05 = α .

Table II. Test of homogeneity of variances

Levene statistic	df1	df2	Sig.
0.940	9	400	.490

(df1 = k -1, df2 = n – k are the degrees of freedom and Sig. is the p-value.)

II. We can now assess the effect of geographical position of stations on precipitation. The one-way ANOVA table, including the *F*-test is shown in Table III.

We deduce a significant effect of station position on precipitation, since $F(9,400) = 13.004,\ p < 0.05$.

Table III. ANOVA table for precipitation data

	Sum of Squares	df	Mean Square	F	p –val
Between Groups	1519346.3	9	168816.2	13	0.00
Within Groups	5192774.1	400	12981.9		
Total	6712120.4	409			

Table IV. Results of the Tukey HSD and the Sheffé tests

Test	ID	Subsets for α = 0.05		
		1	2	3
Tukey	7	261.63		
	8		378.39	
	4		408.82	408.82
	9		423.04	423.04
	10		427.74	427.74
	5		434.66	434.66
	6		434.66	434.66
	3		449.92	449.92
	1			484.54
	2			487.59
	Sig.	1.000	0.126	0.058
Scheffé	7	261.63		
	8		378.39	
	4		408.82	408.82
	9		423.04	423.04
	10		427.74	427.74
	5		434.66	434.66
	6		434.66	434.66
	3		449.92	449.92
	1			484.54
	2			487.59
	Sig.	1.000	0.527	0.370

Performing the Tukey HSD and the Sheffé tests it results that there are the following possibilities to group the meteorological stations, taking into account the mean annual precipitation:

Group 0: Sulina;

Group I: Jurilovca, Hârşova, Constanta, Mangalia, Corugea, Tulcea, Medgidia;

Group II: Hârşova, Constanta, Mangalia, Corugea, Tulcea, Medgidia, Adamclisi, Cernavodă.

Both tests give the same grouping possibilities as it can be seen in Table IV.

The results of these tests are consistent to the observation in Fig. 3. Also, the fact that Sulina form a distinct group is in concordance to its particular position. It is situated 13 km offshore, the amount of precipitation is influenced by the presence of Danube and the Black Sea.

Since the aim is to describe the general evolution of precipitation in Dobrudja, we can follow the ways: (*A*) considering all the ten series, (*B*) eliminating at least Sulina series and using nine series, (*C*) considering Group I or (*D*) considering the Group II, to determine the global model.

In this article we shall present only the first two approaches.

4.1. Results for integral data series (10 *stations*)

In this subsection we discuss the results obtained if we work with all ten data series, i.e. in the case (*A*).

The curves fit by the annual average, the annual weighted average (by Thiessen polygons methods) and the kernel smoothing (using the Gaussian kernel, with a bandwidth of 2 and 3) methods are represented in Fig. 4.

Analysing the residuals in the first two models (Figs. 5 and 6) we reject the normality hypothesis (Fig. 7) and we accept that of correlation (Fig. 8).

The corresponding standard deviation was respectively: 87.35 and 99.18.

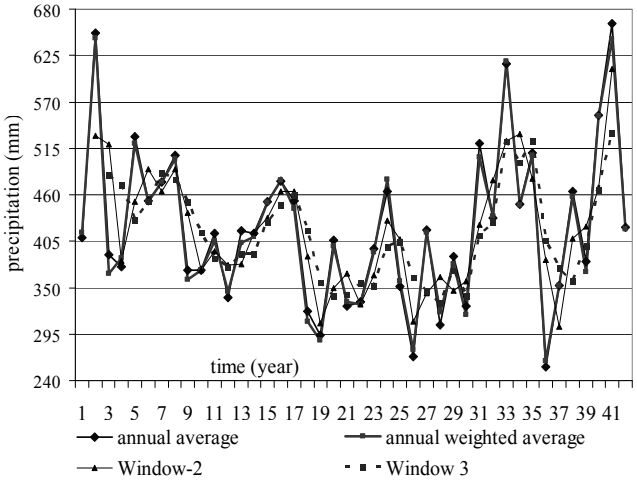

Fig. 4. Annual average and smoothing kernel models for the precipitation evolution in Dobrudja region, using ten data series

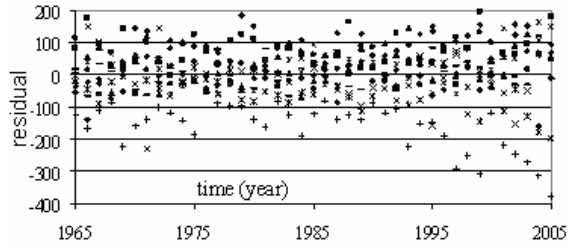

Fig. 5. Residual in the annual average model for the precipitation evolution in Dobrudja region, obtained by using ten data series

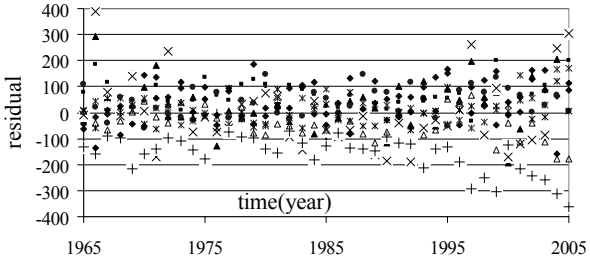

Fig. 6. Residual in the annual weighted average model for the precipitation evolution in Dobrudja region, obtained by using ten data series

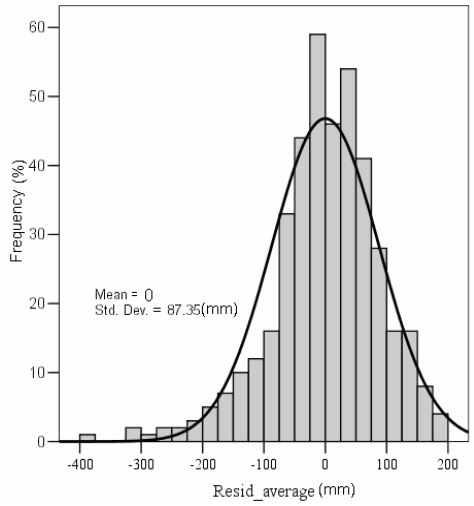

Fig. 7. Histogram of residual in annual average model

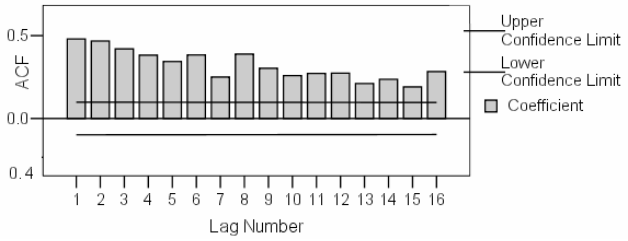

Fig. 8. Autocorrelation function of residual in weighted annual average model

Kernel smoothing (with a bandwidth of 2 or 3) produces Gaussian (Table 5) and correlated residual.

Table 5. Kolmogorov – Smirnov and Shapiro – Wilk normality tests [21]

	Kolmogorov - Smirov			Shapiro - Wilk		
	stat	df	sig	stat	df	sig
window 2	0.041	400	0.104	0.996	400	0.454
window 3	0.028	390	0.200	0.998	390	0.957

In what follows we discuss the results of local linear fit (Fig. 9), obtained by locfit package from R software.

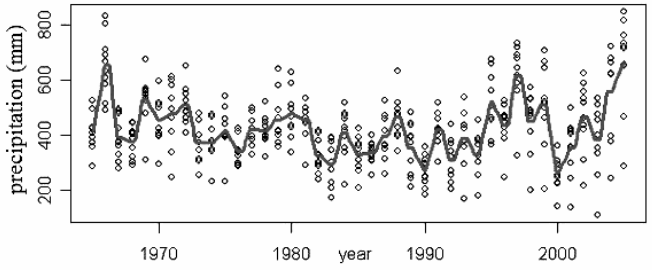

Fig. 9. Model obtained by local linear fit with optimal bandwidth h = 0.6988397

The optimal bandwidth (h = 0.6988397) was chosen by the general cross validation principle (GCV) and the residual variance estimation was done for the same optimal bandwidth (Fig. 10).

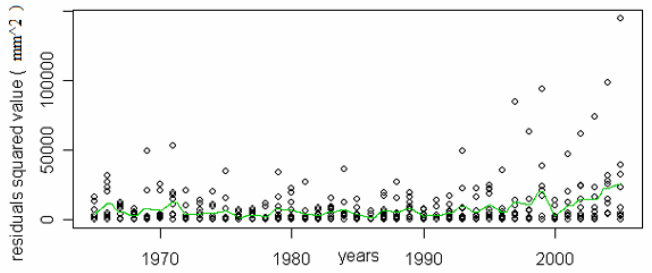

Fig. 10. Residual variance estimation in the model obtained by the local linear fit with optimal bandwidth h = 0.6988397

The question which arises is why we didn't perform a parameterized linear regression:

$$y_i = a + bt_i + \varepsilon_i ,$$

(with y_i - precipitation, t_i – time, ε_i – residual) instead of the local linear fit.

In order to perform a parameterized regression, some general conditions, must be satisfied, as:

i. The model is linear in y_i ;

ii. The values y_i are observed without errors;

iii. The model' mean residual is zero;
iv. The residuals' variance is constant;
v. The residuals are not correlated;
vi. The residuals don't depend on the explicative variable (as is presumed in time series analysis)

Let us discuss only *iv*. and *i.*, in this order.

After a calculus, it results that:

$$a = 0.1918 \text{ and } b = 41.1244.$$

Firstly, the residuals are not homoscedastic, as it can be seen in Fig.11, where they are plotted against the calculated values.

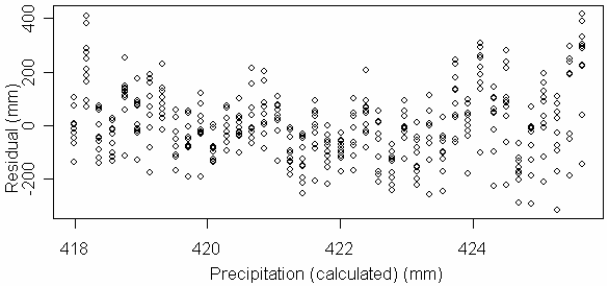

Fig. 11. Residual representation in the parameterized linear model

Secondly, the *t*-tests on the regression coefficients prove that they didn't significantly differ from zero, since the probability to accept the hypothesis that the coefficients are zero is high: 0.97, respectively 0.72 (Table 6).

Table 6. The results of *t* - tests on the linear model coefficients

	Estimate	Std. Err.	t value	Prob.
slope	411.244	1063.236	0.039	0.97
intercept	0.1918	0.5356	0.358	0.72

Also, the residual variance is very high (6718850) and the model's adequation (0.00031) is very small, proving that only 0.031% of variable time acts for explaining the variable precipitation.

Now, we return to the models presented till now. One of the comparison criteria for the fit quality is the residual standard deviation. In our case, the smallest value, 87.06, was obtained for the local linear model.

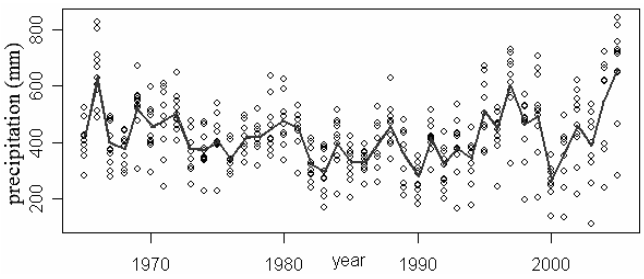

Fig. 12. Annual rainfall fit using smoothing splines

The values calculated for the smoothing parameter and the number of knots were: $\lambda = 0.117317$, $\mathbf{K} = 34$. The residual variance estimation is presented in Fig. 13.

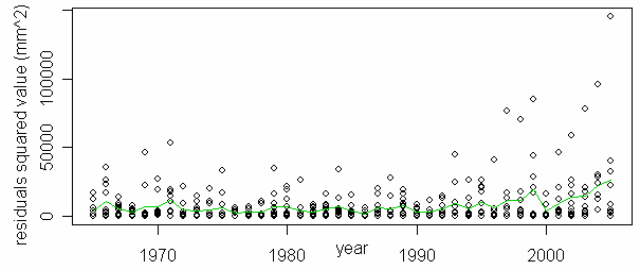

Fig. 13. Residual variance estimation in the model obtained by smoothing splines method

The residuals are normally distributed and their variance is comparable to that obtained by the previous method.

3.2. Results for the data series without Sulina

The calculus was conducted in the same mode as in the previous section. The charts of fit curves are presented in Fig.14, those of residuals squared values in the case of local polynomial smoothing and smoothing splines in Figs. 15 and 16 and comparisons between the residual standard deviations, in Table 7.

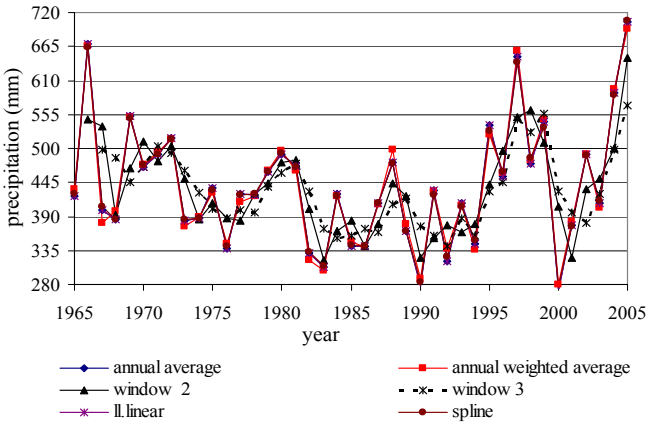

Fig. 14. Models of precipitation variation

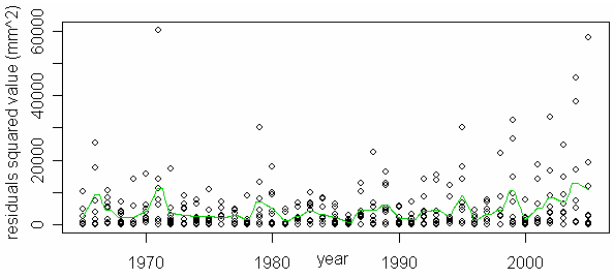

Fig. 15. Residual variance estimation (without Sulina series) in the model obtained by local linear smoothing

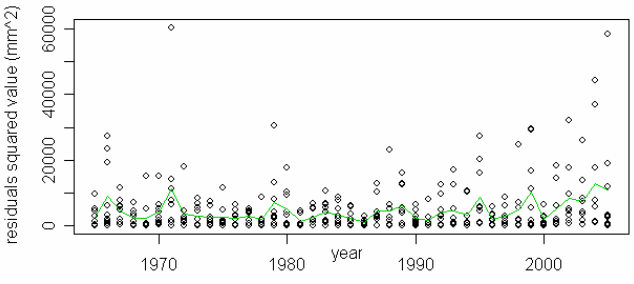

Fig. 16. Residual variance estimation (without Sulina series) in the model obtained by smoothing splines

Table 7. Comparison of residual standard deviations

	average	weighted average	Window_2 (Window_3)	local linear smoothing	splines smoothing
Data	87.35	99.18	105.51 (109.01)	87.06	87.64
Data without Sulina	67.78	68.39	96.91 (110.09)	67.76	67.93

It can be seen that in 5 of 6 cases there is an improvement in the trend estimation, since there is a decreasing of about 20% in standard deviations.

We have to mention that in the second case (data without Sulina) the residuals are normally distributed (Fig.17) and the best result was obtained by the local linear smoothing method.

Also, the p-values corresponding to the models designed by using the data without Sulina are higher (and higher than 0.46) than those in case when Sulina has been considered, confirming that the models obtained are better.

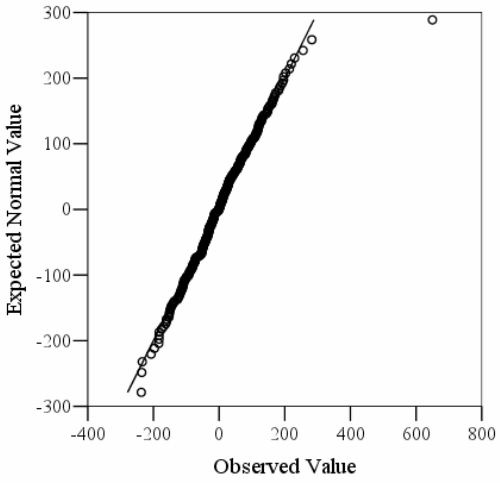

Fig. 17. Q - Q plot of residual in Window_2 model

V. Conclusion

In this article different nonparametric methods were used to determine the trend of precipitation evolution in Dobrudja, since the conditions necessary to build a linear parametric regression model were not satisfied. Taking into account the results of ANOVA and eliminating Sulina series, a better model (in terms of residual variance and p-values) for the trend was determined. It was expected, since Sulina is the single station situated offshore (13 km), in the Danube Delta, so the climate is different from those of other meteorological stations. The same analysis redone after removing Sulina and Jurilovca series leads to a small improvement in residual variances.

In all the cases, the best result was obtained by local linear smoothing method.

In another article we shall present the results of principal components analysis on the ensemble of precipitations series, proving that the influence of Sulina and Jurilovca series in explaining the precipitation variability in Dobrudja region is very small.

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, "Meteorological Time Series Modelling Based on Gene Expression Programming", *Recent Advances in Evolutionary Computing*, 2009, pp. 17 – 23.

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

[4] A. Bărbulescu, E. Băutu, "Mathematical models of climate evolution in Dobrudja", *Theoretical and Applied Climatology*, Vol.100, pp. 29 – 44, DOI 10.1007/s00704 – 009 – 0160 – 7.

[5] A. Bărbulescu, E. Pelican, "On the Sulina Precipitation Data Analysis Using the ARMA models and a Neural Network Technique", *Recent Advances in Mathematical and Computational Methods in Science and Engineering*, 2008, pp. 508 – 511.

[6] A. Bărbulescu, E. Pelican, "ARIMA models for the analysis of the precipitation evolution", *Recent Avances in Computers*, 2009, pp. 221 – 226.

[7] W. S. Cleveland, J. S. Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting", *Journal of the American Statistical Association*, vol. 83, 1986, pp. 569 – 610.

[8] C. T. Dhanya, D. N. Kumar, "Nonlinear ensemble prediction of chaotic daily rainfall", *Advances in Water Resources*, 33(3), 2010, 327 – 347.

[9] J. Fox, *Multiple and Generalized Nonparametric Regression*, Sage, Thousand Oaks CA, 2000

[10] T. J. Hastie, R.J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, London, 1990

[11] D. Hedeker., R.D. Gibbons, *Longitudinal data analysis*, John Wiley and Sons, 2006

[12] Y.-S. T. Hong, P.A. White P. A., "Hydrological modeling using a dynamic neuro - fuzzy system with on-line and local learning algorithm", *Advances in Water Resources*, **32**(1), 2009, pp. 110 – 119.

[13] J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. Van der Linden, X. Dai, K. Maskell., C.A. Johnson, *Climate Change: The Scientific Basis*, Cambridge University Press, 2001

[14] D. Howell, *Statistical Methods for Psychology*, Duxbury, 2002

[15] S. Landau, B.S. Everitt, *A Handbook of Statistical Analyses using SPSS*, Chapmann and Hall/CRC, Boca Raton, 2004

[16] H. Levene, Robust tests for equality of variances. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. In I. Olkin et al. (eds.), Stanford University Press, 1960, pp. 278-292.

[17] C. Maftei, A. Barbulescu, "Statistical analysis of climate evolution in Dobrudja region", *Lecture Notes in Engineering and Computer sciences*, WCE 2008, vol.II, IAENG, 2008, pp. 1082 – 1087.
[18] W.W Ng, U.S. Panu, "Comparisons of traditional and novel stochastic models for the generation of daily precipitation occurrences", *Journal of Hydrology***,** 380(1-2), 2010, pp. 222 – 236.
[19] PESETA project, http://peseta.jrc.es
[20] Z. Sen, A. Ôztopal, "Genetic algorithms for the classification and prediction of precipitation occurrence", *Hydrological Sciences-Journal* 46(2), 2001, pp. 255 – 267.
[21] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples)", *Biometrika*, Vol. 52, No. ¾, 1965, pp. 591-611.
[22] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", *Neurocomputing*, 50, 2003, pp. 159 – 175.