DE GRUYTER OPEN

Mathematical Modelling in Civil Engineering
Vol. 10-No. 4-2014
Doi: 10.2478/mmce-2014-0018

R LANGUAGE: STATISTICAL COMPUTING AND GRAPHICS FOR MODELING HYDROLOGIC TIME SERIES

GABRIELA-ROXANA DOBRE: Assistant Professor, Technical University of Civil Engineering Bucharest, Mathematics and Computer Science Department, e-mail: rx_gabyy@yahoo.com; roxana.dobre2008@gmail.com

Abstract: The analysis and management of Hydrology time series is used for the development of models that allow predictions on future evolutions. After identifying the trends and the seasonal components, a residual analysis can be done to correlate them and make a prediction based on a statistical model. Programming language R contains multiple packages for time series analysis: 'hydroTSM' package is adapted to the time series used in Hydrology, package 'TSA' is used for general interpolation and statistical analysis, while the 'forecast' package includes exponential smoothing, all having outstanding capabilities in the graphical representation of time series. The purpose of this paper is to present some applications in which we use time series of precipitation and temperature from Fagaras in the time period 1966-1982. The data was analyzed and modeled by using the R language.

Keywords: R programming language; Hydrology; Graphics; forecast; Holt-Winters

1. Introduction

The analysis of temperature and precipitation is very important when we study climate changes. Due to the increasing concentrations of greenhouse gases in the atmosphere, the global temperature rise has been accompanied by changes in weather and climate. Climate change studies show an increase of precipitation of 0.5 - 1% per decade in most of the Northern Hemisphere mid and high latitudes [1] and an average global surface temperature increase of about 0.3^0C and 0.6^0C between the late 19th century and 1994.[2]

As for Romania, [3] from data over the period 1961-2007 in 94 meteorological stations, we have an increase of about $2°C$ of the average temperature during summer, winter and spring, and a slight trend of decrease of the average temperature in autumn. The amount of precipitation shows a trend of decrease of the average in summer and winter, and a trend of increase of the amount of precipitation in autumn.

Fagaras Depression is situated in the Southern part of the Transylvanian Basin that is separated from the Romanian Plain in the south by the Southern Carpathians. The climate in the Fagaras area is influenced by the presence of the mountains, which prevent the passage of cold air masses through the South and stop the hot air entering from the South. In Fagaras the annual average temperature is 8.2^0C and the recorded rainfall has annual average values between 600 - 800mm. [4]

The data studied in this article is the daily precipitation and temperature series collected in the period 1966-1982 in the Fagaras area. We consider the problem of modeling the time series of precipitation and temperature using R language.

A seasonal time series consists of a trend component, a seasonal component and an irregular component [11]. The aim of the paper is to analyze time series of precipitations and temperatures from the Fagaras area. After we find the components: the trend component, the seasonal component and the irregular component we make predictions using exponential smoothing with the Holt-Winters method.

2. The R environment

R is a free software environment for statistical computing and graphics, having some advantages in front of other classical statistic programs like SPSS, SAS, STATISTICA, etc. [5]. A few of these advantages are: R is a programming language, an Open Source that can work on Multi-platforms (Windows, Linux, MacOS), and an extendable language that contains over 5300 packages. R Development Core Team updates R and provides support for a large community of R users that share their knowledge.

The R programming environment is freely available through CRAN and can be used for solving practical problems. The R programming environment has many statistical methods and techniques available either built-in or through packages.

In order to use it for the analysis of hydrological time series of the precipitations and the temperatures from the Fagaras area, we use the packages:

- ***HydroTSM*** developed for modeling of hydrological time series. We used version 0.4-2-1 (2014) for the management, analysis and plot of hydrological time series;
- ***TSA*** version 1.01 (2013) is used for time series decomposition and forecast;
- ***Forecast*** version 5.3 (2014) contains the exponential smoothing Holt–Winters method for the analysis and forecast of time series.

Regarding the graphical part, R has good capabilities for representing the time series plots, boxplots and histograms. The boxplot shows the median, first and third quartiles, the data extremes and outliers with the horizontal width of the box proportional to the square root of the size of the group. The histogram is a graphical representation of the data distribution.

3. Methods

In the following, we shortly present the procedures to find components of the time series and, by using the exponential method, to make a forecast for a time series.

3.1. Time series analysis

A *stochastic process* is a sequence of random variables $Y_t, t = 0, \pm 1, \pm 2, ...$ that serve as a model for an observed time series. [6]

For a stochastic process Y_t we define:

– the mean function

$$\mu_t = E(Y_t), t = 0, \pm 1, \pm 2, .. \tag{1}$$

– the autocovariance function

$$\gamma_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s,\ t, s = 0, \pm 1, \pm 2.. \tag{2}$$

– the autocorrelation function

$$\rho_{t,s} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t) Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}},\ t, s = 0, \pm 1, \pm 2, .. \tag{3}$$

3.2. Decomposition of the time series

When the series have constant variability relative to their lengths, we can use an additive model for series decomposition:

$$Y_t = \mu_t + X_t \tag{4}$$

where μ_t is a deterministic function and X_t is a white noise process with $E(X_t) = 0$. [6]

Many authors use the classical decomposition model:

$$Y_t = m_t + s_t + X_t \tag{5}$$

where m_t is the trend, s_t is the seasonal effect, and X_t is the residual series. [7]

Many authors use the word *trend* only for a slowly changing mean function, and use the terms *seasonal component* for a mean function that varies cyclically. We did not make such distinctions here. Deterministic models describe the components: linear, seasonal means and cosine trends. [6]

The *trend* is the result of long-term factors and the time trend equations can be fit to the data by using the *method of least squares*.

The linear trend is expressed as

$$\mu_t = \beta_0 + \beta_1 t \tag{6}$$

where the *slope* and *intercept*, β_1, β_0 are unknown parameters.

In order to find a seasonal trend we take the mean function periodic with period 12:

$$\mu_t = \mu_{t-12}, t = 0, \pm 1, \pm 2, ... \tag{7}$$

The model for the cosine trend has the form:

$$\mu_t = \beta_0 + \beta_1 \cos(2\pi f t) + \beta_2 \sin(2\pi f t) \tag{8}$$

where f is the *frequency*, β the amplitude, Φ the phase of curve and $\beta_1 = \beta\cos(\Phi); \beta_2 = \beta\sin(\Phi)$.

3.2. The Holt-Winters exponential smoothing

Another way to find the trend and seasonality is to use the Holt-Winters exponential smoothing. Three parameters control the smoothing: alpha is a smoothing constant for the level, beta for the estimate of the slope of the trend and gamma is a smoothing constant for seasonality estimate. [5], [7]

The level estimate:

$$a_t = \alpha(Y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{9}$$

The trend estimate:

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{10}$$

The seasonality estimate

$$s_t = \gamma(Y_t - a_t) + (1-\gamma)s_{t-p} \tag{11}$$

where $\alpha, \beta, \gamma \in (0,1)$ and p is the length period of time series.

We can use the Holt-Winters exponential smoothing to make short-term forecasts:

$$Y_{t+h} = a_t + b_t h + s_{t-p+1+(h-1) \bmod p} \tag{12}$$

where h is the period to be forecast.

To see if the Holt-Winters model is correctly specified we have to see if the residuals are independently distributed by using the Ljung–Box test. [8]

Shortly, the Ljung–Box test can be described as follows.

The null hypothesis of the Ljung–Box test is:

H_0 : The data are independently distributed.

The test statistic is

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^{\,2}}{n-k} \tag{13}$$

where:

- n is the sample size,
- $\hat{\rho}_k$ is the sample autocorrelation at lag *k*,
- h is the number of lags tested.

The significance level is p and the hypothesis of randomness is rejected if

$$Q > \chi_{1-p,h}^{\,2} \tag{14}$$

where $\chi_{1-p,h}$ is the p-quantile of the chi-squared distribution with h degrees of freedom.

4. Results and discussion

4.1. Main capabilities of the hydroTSM package

In order to make a basic analysis of the monthly values of the precipitation and temperature at the Fagaras station, from 01/Jan/1966 up to 31/Dec/1982, we convert the daily data to monthly data and obtain the data from table 1. [5]

Table 1

The summary statistics of the time series of monthly precipitation and temperature from Fagaras with data from 01/Jan/1966 up to 31/Dec/1982

Index	Precipitation	Temperature	Observation
Min	0.10	-9.53	Minimum
1st Qu.	26.98	0.60	First quartile
Median	45.95	7.81	Median
Mean	61.61	7.64	Mean value
3rd Qu.	88.75	15.32	Third quartile
Max.	230.60	19.24	Maximum
IQR	61.77	14.72	Interquartile Range= 3rdQu-1stQu
sd	47.22	7.99	Standard deviation
cv	0.77	1.04	Coefficient of variation
Skewness	1.29	-0.26	Skewness
Kurtosis	1.71	-1.20	Kurtosis
n	204	204	Total number of elements

To highlight the characteristics of the monthly time series of precipitation and temperature from Fagaras, we made different graphs: time series plots, boxplots and histograms.

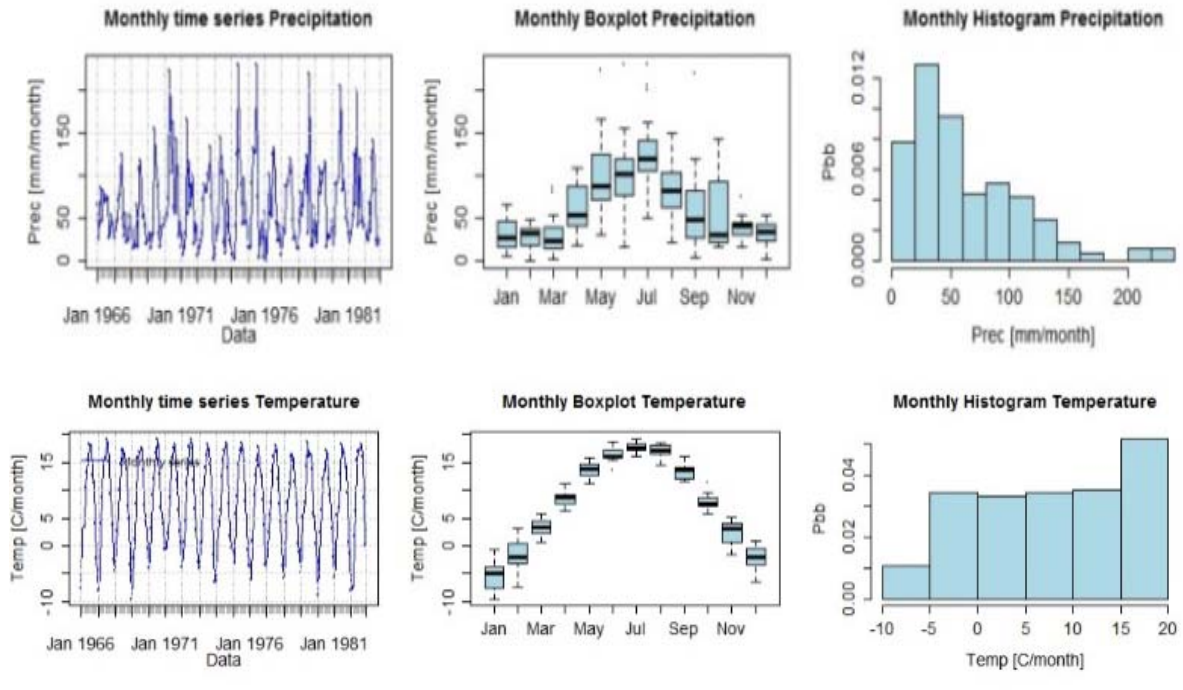
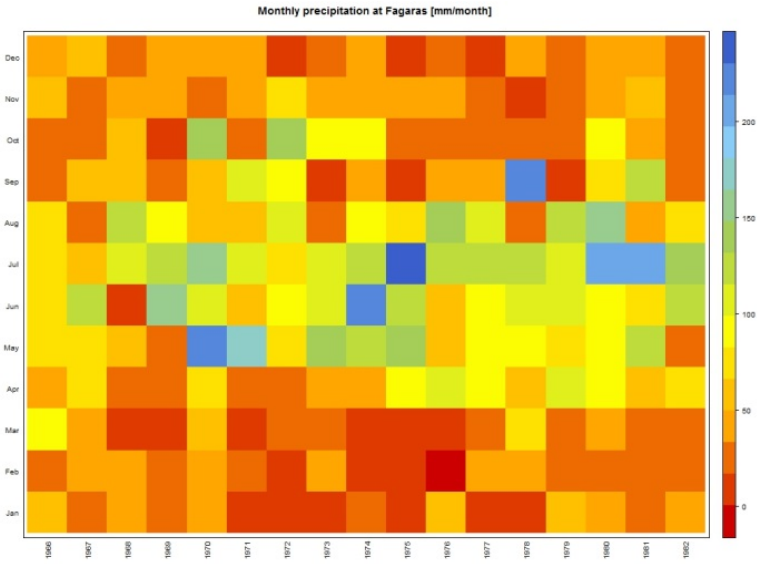


Fig.1-a) Precipitation time series with day, month and year frequency;
b) Temperature time series with day, month and year frequency

The monthly boxplots from figures 1a and 1b describe the seasonal distributions of precipitation and temperature. We can see an increasing trend from December to July and after, a progressive decrease until December.

According to Table 1 and Figure 1a, the precipitation has a positive skewness, so we have an asymmetrical distribution with a long tail to the right and the positive kurtosis shows a distribution more peaked than the Gaussian distribution while in Figure 1b the temperature series has left a longer tail and a flatter distribution.

For identifying the dry/wet months from the monthly precipitation and the hot/cold months from the time series of monthly temperature, we plotted a matrix of the monthly values for each year.



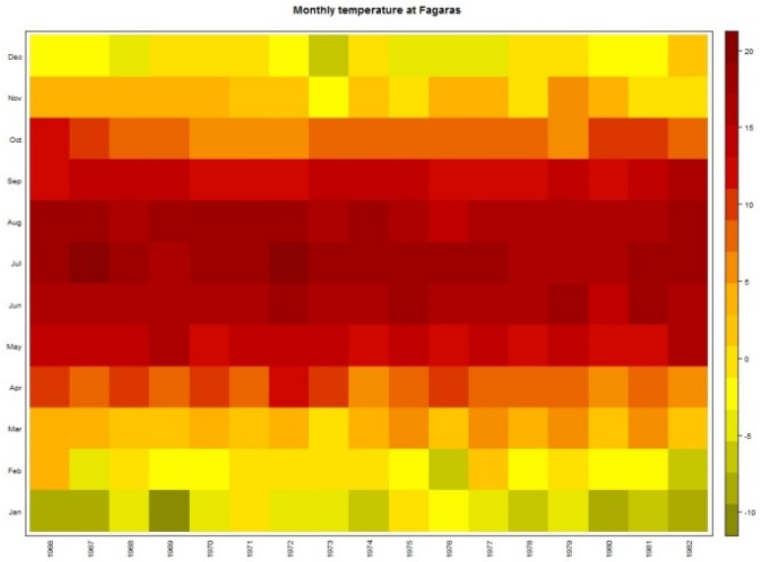


Fig. 2-a) Matrix plot of the monthly precipitations at Fagaras;
b) Matrix plot of monthly temperatures at Fagaras

The package *hydroTSM* contains many functions that can make the conversion between daily, monthly, annual and seasonal data.
Regarding the seasonal data, we obtained the following values:

Table 2

Seasonal analysis for precipitation and temperature

Season	Average seasonal values of precipitation	Average seasonal values of temperature
DJF	87.8	-2.97
MAM	190.4	8.48
JJA	309.5	17.10
SON	151.6	7.94

Analyzing table1, table 2 and figures 1a and 2a, the monthly precipitation repartition reveals that: for the smaller quantities registered in the winter period, from December to February, an average seasonal value of precipitation of 87.8 mm resulted with a minimum of 0.1 mm on February 1976, while in the summer period the average seasonal value of precipitation was 309mm, with a maximum of 230 mm in July 1975.

Analyzing table1, table 2 and figures 1b and 2b the monthly temperature repartition reveals that: smaller temperatures are registered in the winter period, from December to February when we obtain an average seasonal value of temperature -2.97^0C with a minimum of -9.53^0C in January 1969, while in the summer period the average seasonal temperature was 17.1^0C with a maximum of 19.24^0C in July 1967. The seasonal values of precipitation and temperature show four generally recognized calendar-based seasons.

4.2 Estimate trends using TSA

In order to find the pattern of the time series we make graphical decomposition to see the component for the precipitation time series: trend, seasonal and irregular part. [5]
Regarding the *linear model* for the monthly precipitation and temperature time series given by equation (6), the coefficient of determination R^2 is almost 0, so that a fitting line to these data is not appropriate.
In order to see the *seasonal variation* given by equation (7) we create a vector which contains the twelve parameters $\beta_1, \beta_2, ..., \beta_{12}$ of the expected average monthly temperature and precipitation.

Table 3

Seasonal Trends

Month	Estimate temperature	Estimate precipitation
January	-5.2869	29.912
February	-1.6348	27.571
March	3.3480	31.259
April	8.3927	60.676
May	13.6636	98.494
June	16.3559	100.541
July	17.7641	128.318
August	17.1630	80.641
September	13.3702	59.800
October	7.9176	52.053
November	2.5594	39.729
December	-1.8854	30.376

The calculated R-squared for the average monthly temperature is 0.9765 so we have a very good fitting of data points with this statistical model and one can explain this by the inclination change toward the sun of the Northern Hemisphere. The calculated R-squared for the average monthly precipitation is 0.8019, approximately 80 % of the variation of the precipitation series explained by the seasonal component.

The seasonal component [6] does not take into account the differences or similarities between two close periods so we try to model the cyclical component with cosine curves given by equation (8) that incorporates the smooth change expected from one period to the next, while the seasonality is still preserved.

Table 4

Cosine Trend Model

Coefficients	Temperature	Precipitation
β_0	7.644	61.614
β_1	-10.9485	-41.464
β_2	0.1866	7.224

Multiple R-squared for the average monthly temperature is 0.9448 and for the average monthly precipitation is 0.3993. This can be seen in the graphics of the fitting of data.

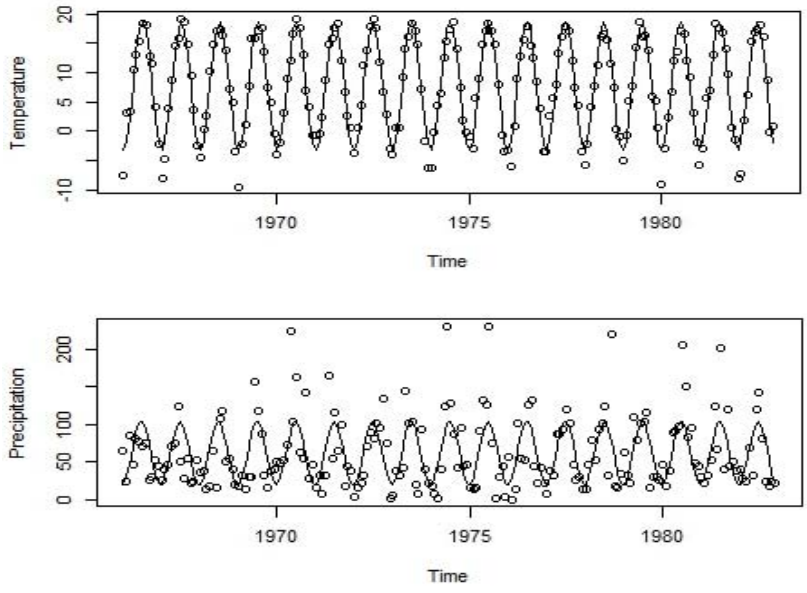


Fig.3- Cosine trends for the temperature and precipitation

The best model for precipitation and temperature is the model with the seasonal component.

4.3. Forecasts Using Holt-Winters Exponential Smoothing

For our forecasting experiments we used the first 15 years (1966-1980) of data values as history by using monthly values of precipitation. [11]
The estimated values of the parameters with the Holt-Winters method are:

Table 5

Holt-Winters exponential smoothing constants

Smoothing parameters	
alpha	0.0124
beta	0.0456
gamma	0.1700

Coefficients	
a	61.5734
b	0.1589
s1	-25.0646
s2	-28.4599
s3	-22.8383
s4	18.7979
s5	39.4857
s6	48.9741
s7	77.9200
s8	38.2382
s9	9.0780
s10	-5.0105
s11	-17.6685
s12	-26.7912

Estimated parameters alpha and beta are very close to zero, so the forecast for the level and the trend are based on further observations in the past. Gamma is relatively low, so this indicates that the estimate of the seasonality at the current point in time is based upon both recent observations and some observations in a more distant past. If the values of parameters are high (close to 1) the estimate of the components are based upon very recent observations.

The forecast were made for the two remaining years (1981-1982) and then the resulting values are compared with those observed for the same period [12]. We obtain 0.77 as a correlation coefficient for the generated and measured data. Since we obtain a strong uphill linear relationship, we can make a prediction for the next 2 years (1983-1984), which is not included in the original time series.

Table 6

The forecast values and prediction interval for the precipitation time series, which corresponds to the January 1981-December 1982 period

Month	Observed	Forecast	Lo80	Hi80	Lo95	Hi95
Jan-81	31.0	36.7	-13.8	87.2	-40.5	113.9
Feb-81	20.9	33.4	-17.1	83.9	-43.8	110.6
Mar-81	31.8	39.2	-11.3	89.7	-38.0	116.4
Apr-81	52.7	81.0	30.5	131.5	3.8	158.2
May-81	124.0	101.9	51.3	152.4	24.6	179.1
Jun-81	66.3	111.5	61.0	162.0	34.3	188.7
Jul-81	200.4	140.6	90.1	191.1	63.3	217.9
Aug-81	40.2	101.1	50.6	151.6	23.8	178.4
Sep-81	120.1	72.1	21.6	122.6	-5.2	149.4
Oct-81	44.0	58.2	7.6	108.7	-19.1	135.4
Nov-81	51.3	45.7	-4.9	96.2	-31.7	123.0
Dec-81	38.0	36.7	-13.9	87.2	-40.6	114.0
Jan-82	39.9	38.6	-12.9	90.0	-40.1	117.2
Feb-82	27.1	35.3	-16.1	86.8	-43.3	114.0
Mar-82	23.4	41.1	-10.3	92.6	-37.6	119.8
Apr-82	68.3	82.9	31.5	134.4	4.2	161.6
May-82	32.9	103.8	52.3	155.2	25.0	182.5
Jun-82	119.8	113.4	61.9	164.9	34.7	192.1
Jul-82	141.1	142.5	91.0	194.0	63.8	221.3
Aug-82	81.3	103.0	51.5	154.5	24.2	181.8
Sep-82	23.0	74.0	22.5	125.5	-4.8	152.8
Oct-82	16.9	60.1	8.5	111.6	-18.8	138.9
Nov-82	24.4	47.6	-4.0	99.1	-31.3	126.4
Dec-82	22.7	38.6	-13.0	90.2	-40.3	117.5

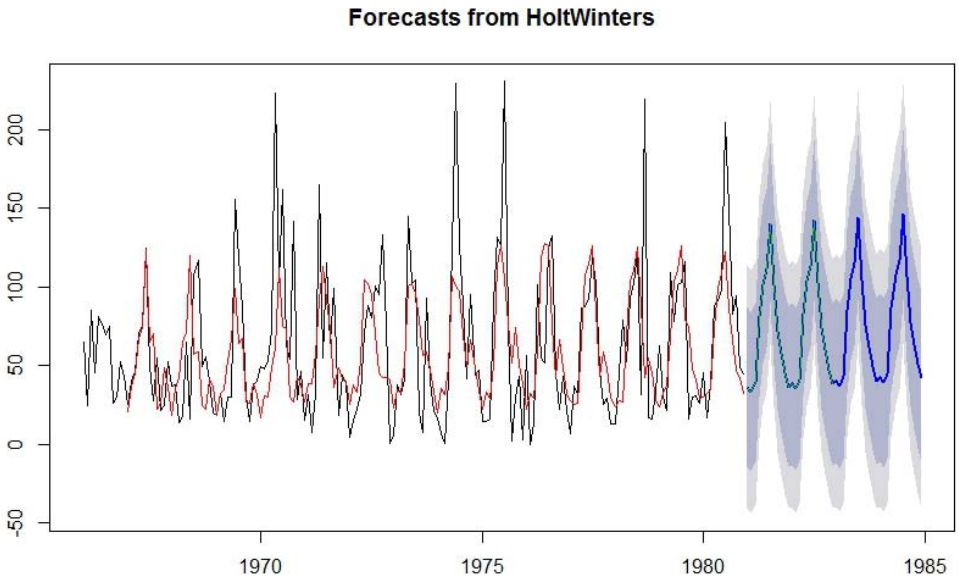


Fig.4 - The graphics of the monthly precipitation time series from January 1966 up to December 1980 (black); Holt-Winters exponential smoothing (red) for the precipitation time series from January 1966 up to December 1980; Holt-Winters exponential smoothing (green) for the precipitation time series from January 1981 up to December 1982; Holt-Winters exponential smoothing (blue) for the precipitation time series from January 1983 up to December 1984; Shaded areas show 80% (dark grey) and 95% (light grey) prediction intervals.

In order to see if the Holt-Winters exponential smoothing provides an adequate predictive model for the precipitation time series, we make a correlogram in order to check if the residuals show non-zero autocorrelations at lags 1-20. Autocorrelation function indicates how a time series is related to itself over time. [11]

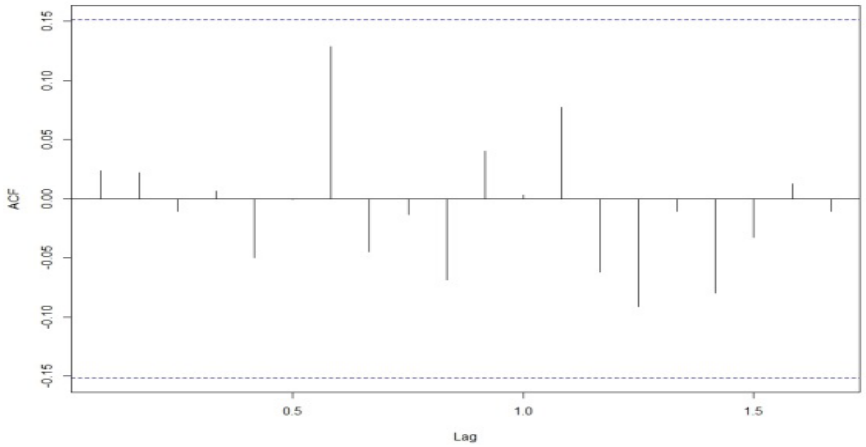


Fig. 4 - The autocorrelation function for the residuals of precipitation time series

The correlogram shows that the autocorrelations for the in-sample forecast errors do not exceed the significance bounds for lags 1-20. From the Ljung-Box test we obtain a p-value of 0.9686 for the $\text{lag} = 20$ so it is above 0.05, which indicates non-significance **of autocorrelation values.**

To see if our forecast is good enough for our purposes and that it cannot be improved, we check if the forecast errors are normally distributed using a histogram. [11]

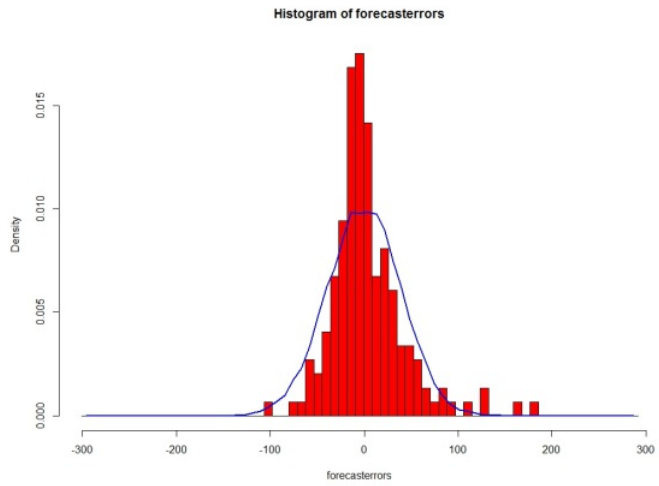


Fig. 5- The normal distribution of the forecast errors of the precipitation time series

Holt-Winters exponential smoothing method provides an adequate predictive model for the precipitation time series from Fagaras, because the distribution of the forecast errors seems to be normally distributed with mean and zero constant variance over time and the Ljung-Box test showed that there is little evidence of non-zero autocorrelations.

5. Conclusion

The purpose of this article is to present some results related to the modeling of the hydrologic time series of precipitation and temperature from Fagaras meteorological station in the period 1966-1982 by using the R language.

The paper presents the main capabilities of the three R packages: *hydroTSM*, *TSA* and *forecas*t with applications. We used the *hydroTSM* package for the management, analysis and the plots that capture the information about the central tendency, distribution and frequency. In order to describe some deterministic components: linear, seasonal, means and the cosine trends components, we estimate parameters and investigate the efficiency of these regression methods by using the *TSA* package. By using the *forecast* package, we make predictions of future events related to precipitation in the Fagaras area based on the Holt-Winters method. We conclude that this method provides an adequate predictive model, which probably cannot be improved more. In a future study, we will try to make a better predictive model by trying other different methods: ARIMA, Markov, or by using artificial intelligence methods.

References

[1] Sayemuzzaman, M., Jha, MK. (2014). *Seasonal and annual precipitation time series trend analysis in North Carolina*, Atmospheric Research, Volume 137, pp 183–194. http://dx.doi.org/10.1016/j.atmosres.2013.10.012
[2] IPCC Fifth Assessment Report: Climate Change. (2013) , http://www.ipcc.ch/report/ar5/
[3] Climate change Romania, http://www.climateadaptation.eu/romania/climate-change/
[4] Raport de mediu–Plan Urbanistic General Municipiul Fagaras, http://www.primaria-fagaras.ro/urbanism/PUG-2013/raport%20mediu%20revizuit%20mai%202013.pdf
[5] The Comprehensive R Archive Network. http://cran.r-project.org/
[6] Cryer, J. D., Chan, K-S. (2008).*Time Series Analysis with Applications in R*, Springer
[7] Brockwell, P. J., Davis R. A. (2002). *Introduction to Time Series and Forecasting*. Springer-Verlag New York, Inc
[8] Ljung, G. M.; Box, G. E. P. (1978). *On a Measure of a Lack of Fit in Time Series Models.* Biometrika 65 (2), pp 297–303.
[9] Barbulescu A., Deguenon, J. (2011). *Mathematical models for extreme monthly precipitation*, Ovidius University Annals, Series: Civil Engineering, issue 13, pp. 93 – 104, http://revista-constructii.univ-ovidius.ro/doc/anale/2011.pdf
[10] Cowpertwait, P. S.P. (2006). *Introductory Time Series with R*, Springer Science+Business media
[11] Coghlan, A. (2014). *Using R for Time Series Analysis*, https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf
[12] Miroiu, M., Petrehus, V., Zbaganu G. (2008-2011): *Initiere in R pentru persoane cu pregatire matematica*, POSDRU/56/1.2/S/32768