

# Weather Forecast using SPSS Statistical Methods

Daniela Șchiopu , Elia Georgiana Petre, Cătălina Negoită

Petroleum – Gas University of Ploiești, Bd. București, No. 39, 100680, România, Catedra de Informatică  
e-mail: daniela\_schiopu@yahoo.com ; elia\_petre@yahoo.com; catalinanegoita@yahoo.com

## Abstract

*This paper presents a case study of using SPSS 13.0 in weather prediction. The data were collected from 2001 till 2005 and it was made a prediction of future temperatures encountered in this region. For this, we used two methods provided by SPSS 13.0, such as factor analysis and linear regression.*

**Key words:** *weather prediction, factor analysis, correlation, regression*

## Introduction

Weather prediction deals with the application of science and technology to predict the state of the atmosphere for a future time and a given location. There is a variety of methods to weather forecast. Weather warnings are an important forecast because they are used to protect life and environment.

Starting the training data, we use SPSS software for weather forecast. SPSS [4] (which stands for Statistical Package for the Social Sciences) encapsulates advanced mathematical and statistical expertise to extract predictive knowledge that when deployed into existing processes makes them adaptive to improve outcomes. Thus, together with these advantages offered by this software, we can predict the outcomes before they occur.

## Case Study

For this application, we used the public available data provided by Hong Kong Observatory site [3].

The structure of the database is:

- *an* – the year when the data were registered;
- *luna* – the month when the data were registered;
- *pres\_medie* – average pressure;
- *max\_zilnica* – temperature daily maximum;
- *medie\_temp* – temperatures average;
- *min\_zilnica* – temperature daily minimum;
- *umidit\_relat* – average relative humidity;
- *cant\_nori* – average amount of cloud;
- *precipitatii* – total rainfall.

The temperatures from the database were registered in Celsius degrees.

Then, the database is completed with the data about weather corresponding from 2001 to 2005 (Figure 1).

	an	luna	pres_medie	max_zilnica	medie_temp	min_zilnica	umidit_relativa	cant_nori	precipitatii	var
22	2003	October	1008,7	29,6	27,6	26,0	81,0	66,0	394,0	
23	2003	November	1014,8	27,1	25,3	23,6	71,0	44,0	48,6	
24	2003	December	1021,9	25,7	23,0	20,9	77,0	65,0	2214,3	
25	2004	January	1020,4	17,5	15,8	14,2	73,0	67,0	51,0	
26	2004	February	1018,4	19,4	17,2	15,4	79,0	62,0	51,8	
27	2004	March	1015,8	21,1	19,1	17,7	80,0	76,0	104,3	
28	2004	April	1012,6	25,3	22,9	21,2	83,0	76,0	147,2	
29	2004	May	1008,8	28,6	26,3	24,5	82,0	71,0	194,4	
30	2004	June	1006,8	30,8	28,6	26,8	78,0	66,0	144,7	
31	2004	July	1006,7	30,5	28,4	26,6	82,0	76,0	386,7	
32	2004	August	1003,5	30,7	28,4	26,5	82,0	64,0	488,5	
33	2004	September	1009,9	29,9	27,6	25,9	77,0	62,0	167,3	
34	2004	October	1016,1	27,5	24,9	22,9	64,0	36,0	2,3	
35	2004	November	1018,6	24,5	22,7	21,1	73,0	50,0	,4	
36	2004	December	1019,3	21,1	19,1	17,2	70,0	40,0	,0	
37	2005	January	1019,6	17,9	15,9	14,2	75,0	63,0	5,9	
38	2005	February	1017,9	18,3	16,5	14,8	86,0	91,0	19,2	
39	2005	March	1018,7	19,7	17,4	15,2	80,0	76,0	52,6	
40	2005	April	1014,1	25,4	23,1	21,3	83,0	79,0	32,3	
41	2005	May	1007,9	28,8	27,0	25,3	85,0	82,0	508,6	
42	2005	June	1004,6	29,7	27,8	25,9	85,0	85,0	893,9	
43	2005	July	1006,4	31,8	29,1	26,9	80,0	59,0	360,4	
44	2005	August	1004,8	30,2	28,0	26,3	85,0	74,0	971,3	
45	2005	September	1009,2	30,4	28,2	26,5	81,0	69,0	352,6	
46	2005	October	1014,9	28,3	26,2	24,4	72,0	57,0	6,6	
47	2005	November	1016,6	25,0	23,0	21,4	77,0	55,0	1,6	
48	2005	December	1021,2	19,1	17,0	15,0	64,0	59,0	9,5	
49	.	.	.	.	.	.	.	.	.	.
50	.	.	.	.	.	.	.	.	.	.
51	.	.	.	.	.	.	.	.	.	.

Fig. 1. The SPSS database

Outside of the primary variables, we transform average temperatures from Celsius degrees in Fahrenheit degrees. The transformation formula is (1), where *C* is Celsius and *F* is Fahrenheit.

$$F = \frac{9}{5} * C + 32 \tag{1}$$

### Linear Regression

Regression analysis is a statistical tool that can produce data predictions [2]. The basic principle behind regression is to use one or more variables to predict another variable of interest. The predictor variables are known as independent variables and the variable that is being predicted is known as the dependent variable.

In our example, we consider the regression between the atmospheric pressure and the average temperature. There is only a predictor variable (*pres\_medie*) and a dependent variable (*medie\_temp*). Using SPSS regression linear method, we obtain regression coefficients, shown in Figure 2. Unstandardized (B) and standardized (beta) coefficients can be used in prediction equation. B is the origin and Beta is the slope of the regression right line. For instance, if an atmospheric pressure is registered being 1017.5, we can estimate the average temperature using the relation  $medie\_temp = B + Beta * pres\_medie = 736.274 - 0.704 * 1017.5 = 19.95$ .

Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	736.274	55.110		13.360	.000
	pres_medie	-.704	.054	-.886	-12.930	.000

a. Dependent Variable: medie\_temp

Fig. 2. The coefficients for linear regression

The values predicted in this way are estimations, while the correlation between variables is not perfect. The error from estimations is not directly proportional with the correlation between variables (the correlation graph points being more far away from the regression right line).

## Factor Analysis

It is a reductive method in which new factors are built based on existing relationships between variables. A period of time, factor analysis was used only in psychology. After noticing the good results of this method, it was applied in the economical analysis and it has become an established statistical method [1].

The main challenge of factor analysis is to find variables that have as much variation (information) "common" as possible, so that after concentration little useful information is lost.

Using SPSS, we can apply this method by accessing the menu: *Statistics – Data Reduction – Factor*, and the *Factor Analysis* window will appear (Figure 3). We can select the variables needed, as well as various characteristics of the variables that may be calculated like mean, standard deviation, correlation matrix, number of factors to be determined, factor score coefficient matrix. The exclusion of the case list wise and the suppression of the absolute values less than 0.10 can be done as well.

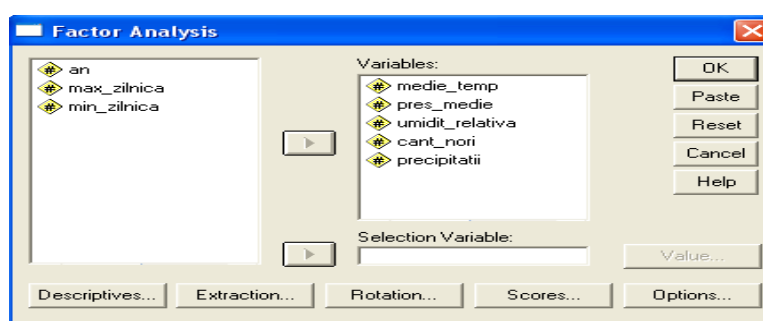


Fig. 3. The Factor Analysis window

After the proper selection of the needed characteristics, we can press the OK button.

Beside the results described above, the method output presents the Total Variance Explained, the Component Matrix and the new variables found.

In the *Total Variance Explained* section (Figure 4) it is noticed that 5 new variables are generated, but only 2 of them explain our system more properly.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.830	56.606	56.606	2.830	56.606	56.606	2.166	43.324	43.324
2	1.268	25.359	81.964	1.268	25.359	81.964	1.932	38.640	81.964
3	.744	14.888	96.853						
4	.100	2.004	98.857						
5	.057	1.143	100.000						

Extraction Method: Principal Component Analysis.

Fig. 4. Total Variance Explained section

On *Extraction Sums of Squared Loadings* columns (Figure 4), there are explained variance and cumulative variance for two factors in the context of initial factorial solution (non-rotating). Variant explained by each factor is distributed between factor 1 - 56,606% and factor 2 - 25.359%. Together, they explain 81.964% of the variation of analyzed values. The rest up to 100% remain unexplained by this model factorial.

On *Rotation Sums of Squared Loadings* columns (Figure 4), we have the same values, but after the procedure of rotation. A redistribution of variation explained by each of the items: factor 1 – 43.324% and factor 2 – 38.640% has been made. So, using a rotation method of the redistribution between the two factors has been made.

Another section of the output is *Component Matrix* in which the weights of the each variable in the correlation factors selected are been presented. Having in mind the selected settings made, values less than 0.10 are not included in the table.

Using the last section, the *Rotated Component Matrix*, which presents data after the factors rotation in a form more easily to analyze, we will formulate the method conclusions (Figure 5).

*Component 1* is formed only by variable *medie\_temp* with a rate of 0.973%. Other variables were not taken into consideration because they were in a negative proportion or they have a ratio under 0.6. Since this factor is represented by only one variable, the new name should be: *temperature*.

The second component consists of two variables: *cant\_nori* (0.972%) and *umidit\_relativa* (0.915%) and it has been renamed *specific\_ploii* (rain).

	Component	
	1	2
medie_temp	.973	
pres_medie	-.892	-.303
precipitatii	.560	.241
cant_nori		.972
umidit_relativa	.321	.915

Fig. 5. The Rotated Component Matrix

## Conclusions

The present application can forecast an unknown value, on the basis of some real, known values, using techniques that don't use too many scientific details. The factor analysis is important because it reduces a large number of variables into less factors.

## References

1. Field, A.P. - *Discovering statistics using SPSS (second edition)*, London: Sage publications, 2005, Chapter 15, <http://www.sussex.ac.uk/Users/andyf/factor.pdf>, accessed on September 2009
2. SPSS Inc. - *Time Series Analysis and Forecasting with SPSS Trends 9.0 (course guide)*, 1999, <https://www.vcu.edu/s/ts/manuals/spss/guides/Time%20Series%20and%20Forecasting.pdf>, accessed on September 2009
3. \*\*\* - *Hong Kong Observatory*, <http://www.weather.gov.hk/wxinfo/pastwx/ywx.htm>, accessed on September 2009
4. \*\*\* - *SPSS*, <http://www.spss.com/>, accessed on September 2009

## Prognoza vremii utilizând metode statistice din SPSS

### Rezumat

Lucrarea prezintă un studiu de caz al aplicării metodelor statistice furnizate de SPSS 13.0 în realizarea de prognoze meteo. Datele despre starea vremii au fost înregistrate între anii 2001 – 2005 într-o anumită regiune și s-a făcut o predicție referitoare la temperaturile viitoare care pot fi întâlnite în această zonă. Metodele furnizate de programul SPSS 13.0 sunt analiza factorială și regresia liniară.