MACRo 2011- International Conference on Recent Achievements in
Mechatronics, Automation, Computer Science and Robotics

Inference in Meteorological Data Taken from January 8, 2011 at Three Different Locations

Lorentz JÄNTSCHI[1,2]

[1]Chair of Chemistry, Faculty of Engineering of Materials and Environment, Technical University of Cluj-Napoca, 400641 Cluj
[2]Chair of Horticulture and Landscape architecture, University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 400372 Cluj
email: lori@academicdirect.org

Manuscript received February 05, 2011; revised March 15, 2011.

Abstract: In thee meteorological observation points were installed identical weather stations and a network infrastructure were created to receive and record the environmental parameters at every minute. A series of six out of forty-four parameters relating air, and solar radiation were taken into the analysis. The inference between observed data in three different locations was analyzed in order to find similarities between locations and/or observables. The analysis revealed important associations between environmental parameters and geographical location.

Keywords: environmental monitoring; environmental parameters; statistical inference; cluster analysis; multiple linear regression

Introduction

The analysis of the inference in meteorological data gives more and more interest due to the development of new forecasting models. Several approaches keep the front line. Thus, neural networks [1] are often used to the cases of incomplete data inference. When series of measurements of different nature are available, the inference in data is often searched using multiple linear regression and principal component analysis [2]. Humidity and temperature analysis provides insight into the potential health impacts of climate change [3]. Dew point analysis has important civil and thermal engineering applications, such as in design of dehumidification equipment [4].

Starting with an home-made equipment for acquisition and measurement of indirect and total solar radiation in 2007 [5] which provided important information concerning the recovering of local solar energy using thermal collectors [6], the environmental parameters acquisition system [7] were later extended with three commercial weather stations, which were able to provide data for analysis of: air movement [8], soil and leaf parameters [9], influence of environmental conditions on fruit growing [10], and apple scab attacks under conditions of excessive rainfalls [11].

In the present study, the inference between observed data in three different locations was analyzed in order to find similarities between locations and/or observables.

Material

Three Vantage Pro2 weather stations were placed and records data (44 observables) at every minute at three different locations.

The location of the observation points are given in *Table 1*.

Table 1: Observation points

Place	Weather station	GPS	Elevation	Distance from ground
Reghin	st1	N 46° 46' 12.41"; E 24° 41' 27.99"	390m	1.5m
USAMV-CN	st2	N 46° 45' 34.00"; E 23° 34' 20.53"	381m	1.5m
UT-CN	st3	N 46° 47' 45.40"; E 23° 37' 34.33"	326m	20m

Six observables were selected (recorded in every observation point in same time) for analysis, being given in *Table 2*.

Table 2: Environmental observables

Observable	Meaning	Measurement unit
t_out	Outside temperature	°C
h_out	Outside relative humidity	%
t_dwp	Outside dew point temperature	°C
w_spe	Wind speed	ms^{-1}
p_bar	Barometric pressure	milibars
s_rad	Solar radiation	Wm^{-2}

The data recorded from 10 to 10 minutes (144 records) were included into the analysis.

Methods

Cluster analysis [12] may be defined as a mathematical way of assignment of a set of observations into subsets (called clusters) so that observations belonging to same cluster are similar in some sense. The clustering problem has been addressed in many contexts and disciplines [13], having applications in meteorological control [14]. Cluster analysis was used in the present study to infer the environmental data from a contingency of six observables and three locations.

Dew point is an estimated variable according to Davis Inc [15] with an approximating formula recommended by WMO [16]. Using same notations as in Table 2, the estimated outside dew point temperature (tedwp) is - according to [15] - given by:

$$\text{vpv} = \text{h\_out} \cdot \frac{6.112}{100} \cdot \exp\left(17.62 \frac{\text{t\_out}}{\text{t\_out} + 243.12}\right), \quad \text{tedwp} = \frac{243.12 \cdot \ln(\text{vpv}) - 440.1}{19.43 - \ln(\text{vpv})}$$

By using our range of observables values (t_out were from -2.1°C to 4.1°C; h_out were from 83% to 93%) a MathCad representation (Figure 1) shown that the tedwp have a monotone approximately plane dependency, which allow us using of multiple linear regression to obtain simpler regression equations estimating t_dwp.

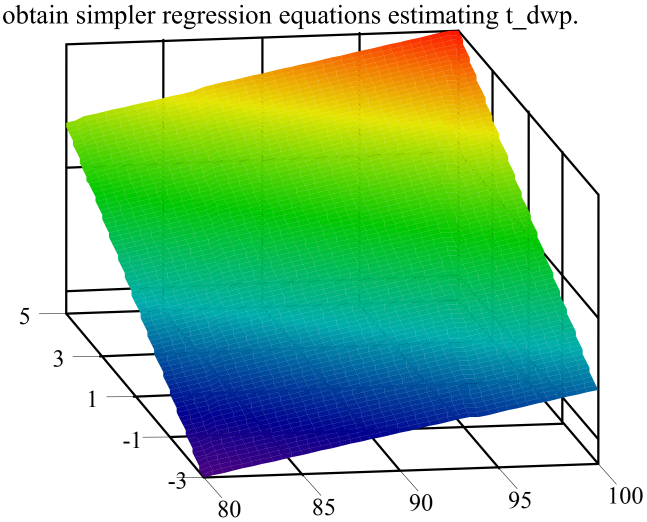

Figure 1: tedwp(h_out,t_out) variable plot (h_out: 80% to 100%; t_out: -3°C to 5°C)

Correlation analysis, initially introduced to measure the strength of linear dependence [17], was later extended to a more general case to infer monotone dependences [18]. Correlation analysis was used in the present study to compare the Dew point estimation method reported in [15] with recorded from weather stations values of the dew point.

Multiple linear regression uses various strategies [19] to minimize the disagreement between a set of observables [20] under assumption of linear dependence [17], having many environmental analysis applications [21]. Multiple linear regression were used in the present study to construct multiple linear relationship (MLR) models for dew point.

Results

The analysis of correlation between observed dew points and calculated ones according to [15] are given in Table 3. Table 3 contains the correlation analysis of all data (from all three weather stations, the data being taken pair by pair (432 pairs).

Table 3: Correlation analysis between calculated (tedwp) and recorded (t_dwp) dew point

Correlation [18]	Pearson	Spearman	K-Tau-a	K-Tau-b	K-Tau-c	Gamma
Coefficient	0.9997	0.9990	0.9666	0.9861	0.9644	0.9982
Wrong model probability*	0.0e-1	0.0e-1	9.6e-1088	1.2e-1077	1.8e-1077	1.3e-1111

*from 'Student t' test

Correlation coefficients from Spearman to Gamma express different measures of monotone association (their definitions uses different manners of ties treatment).

A tree diagram for 6 (observables) times 3 (locations) using 144 observations were obtained when single linkage of Euclidian distances were calculated (Figures 2 to 4).

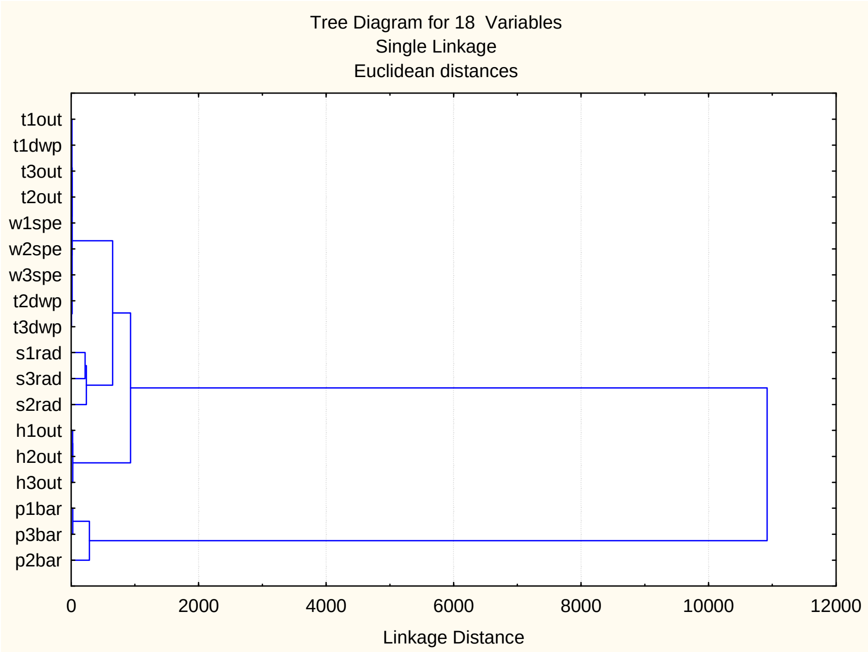

Figure 2: Linkage distance between observables × locations

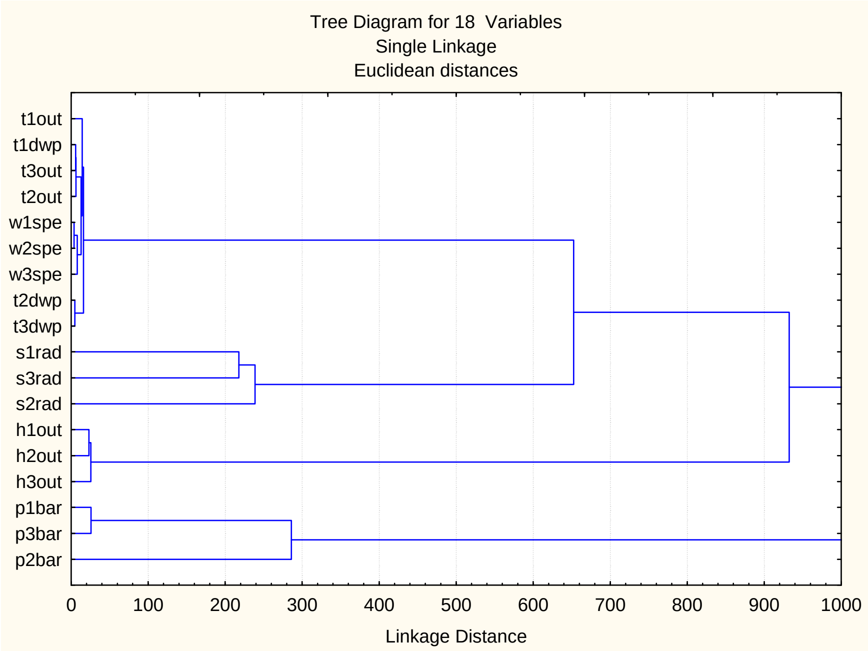

Figure 3: Zoom in 0 to 1000 range of Figure 1

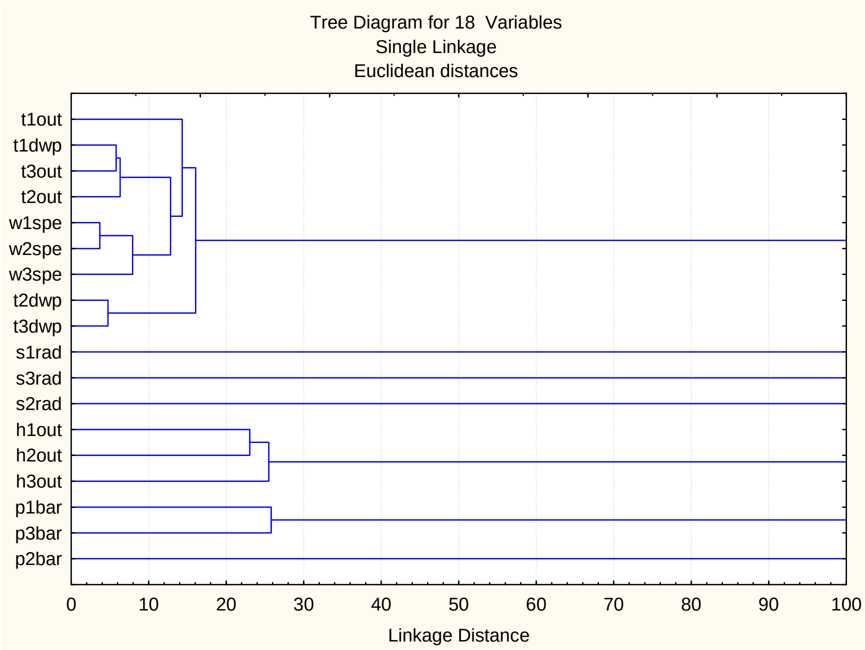

Figure 4: Zoom in 0 to 100 range of Figure 1

Table 4 gives four multiple linear regression analyses between t_dwp (as dependent variable) and t_out and h_out (as independent variables).

Table 4: Multiple linear regression analysis results for t_dwp=MLR(t_out,h_out)

Observations*	t1dwp	t2dwp	t3dwp	t_dwp
Raw data: (Y=t_dwp, X_1=t_out, X_2=h_out)				
Number	144	144	144	432
Model: $\hat{Y}=a_0+a_1\cdot X_1+a_2\cdot X_2$; Assumption: $\hat{Y}$ is a normal estimates for Y (see [22] for details)				
Intercept	-15.24 $\notin$ CI_{t_dwp}	-15.48 $\in$ CI_{t_dwp}	-15.72 $\notin$ CI_{t_dwp}	-15.49
t_out coefficient	.9823 $\notin$	.9902 $\notin$	.9769 $\notin$	.9854
h_out coefficient	.1534 $\notin$	.1560 $\in$	.1589 $\notin$	.1561
Linear association measure: $r=r(Y,\hat{Y})$ (see [17] for details)				
Pearson r	0.9994	0.9997	0.9996	0.9997
Common sense statistics: confidence intervals for coefficients (see [22] for details)				
$CI_{Intercept}$	[-15.60, -14.88]	[-15.69, -15.28]	[-15.93, -15.52]	[-15.61,-15.37]
$CI_{t\_out\ coefficient}$	[.9760, .9885]	[.9860, .9944]	[.9707, .9830]	[.9829,.9878]
$CI_{h\_out\ coefficient}$	[.1495, .1572]	[.1537, .1583]	[.1565, .1613]	[.1548,.1575]

* t1dwp: from st1; t2dwp: from st2; t3dwp: from st3; t_dwp: all together; CI: at 95% probability coverage

Followings can be observed in *Table 4*:

÷ $Intercept_{t1dwp}$=-15.24 $\notin$ [-15.61,-15.37]=$CI(Intercept)_{t\_dwp}$; idem for t3dwp;

÷ $Coef(t\_out)_{t1dwp}$=.9823 $\notin$ [.9829,.9878]=$CI(t\_out)_{t\_dwp}$; idem for t2dwp and t3dwp;

÷ $Coef(t\_out)_{h1dwp}$=.1534 $\notin$ [.1548,.1575]=$CI(t\_out)_{t\_dwp}$; idem for h3dwp;

Discussions

Correlation analysis from *Table 3* reveals that the reported (in [15]) formula of calculation doesn't "fit exactly" on the weather station given data. The main result of the analysis given in *Table 3* is that a 'monotone association' is less likely than a particular case of it, 'linear association'. A better agreement was obtained when the data are correlated (0.9997) than their ranks are correlated (all others below 0.9990); this is not the expected result when a formula of calculation are applied, but on the contrary, when experimental error occurs [23]; thus, the only conclusion which can be drawn from here is that the reported formula (in [15]) and their implementation in the weather station device (or software) has some (minor) leaks.

The tree diagram from *Figures 2-4* reveals the degree of the similarity between observables. The observed groups of data had shown that:

÷ Wind speed (due to it's stationery almost all the time behavior - no wind activity in over 75% of the cases) are one of the best group of relatives (clearly shown on *Figure 4*); inside this group the association between observations from Reghin (st1) and USAMV-CN (st2) is attributed to the relative altitude from ground of the observation points: both weather stations are at near to the ground level while UT-CN (st3) observation point is at about 20m from the ground, and thus is expected that wind activity to be somehow different - is almost twice much more wind activity (62 vs. 36 snapshots out of 144 with wind activity) and is over twice more intense (54.7 ms^{-1} vs. 20.2 ms^{-1} sums of wind speeds from 144 moments) at observation point st3 (Cluj-Napoca, 20m from ground) vs. observation point st2 (Cluj-Napoca, near to the ground); due to this fact we can draw the conclusion that wind speed observations array is a better location indicator in terms of the height from the ground position;

÷ Dew points from st2 and st3 observation points are relatives due to similar environmental conditions (temperature, relative humidity, barometric pressure) but are much relatives than any of them (*Figures 3 and 4*); this similarity should be assigned to the neighborhood of the observation points (placed at about 5.8 Km one to the other); due to this fact, we can draw the conclusion that dew point observations array is a better location indicator than others in terms of horizontal geographical position;

÷ The next cluster of interest groups wind speeds, outside temperatures and dew points (*Figure 4*); this is a surprising association, being known [16] that dew point is in relationship with temperature (t_out) and relative humidity (h_out);

÷ Three other clusters groups the observables by their locations; by degree of association these groups are (*Figure 3*): Outside relative humidity (at linkage distance of about 26), Solar radiation (at linkage distance of about ten times higher), and Barometric pressure (at linkage distance of about eleven times higher);

The multiple linear regression analysis from *Table 4* reveals that even if the models are high statistically significant, this is not a enough condition to be accepted as to be true - under assumption that the equation t_dwp=MLR(t_out,h_out) exists, then must be true when the model are feed with all possible data and should converge to it's true value when number of observations increases; more than that, outside of the 95% probability confidence interval should be only 5% of the possible cases, while analysis given in *Table 4* reveals that in 7 out of 9 cases the coefficients are outside of the 95% confidence interval. This result should be correlated with the fact that the data are split according to their location of observation. Due to this fact we can draw the conclusion that dew point temperatures depends (in small amount) by at least one parameter more than temperature t_out and relative humidity h_out (such as wind speed, as *Figure 4* strongly suggest).

Somebody may say: Why the linear regression was considered? What was the reason (only the simplicity)? - and we should raise an answer to these questions too. Indeed, the simplicity characterizes the linear regression. But more, as we depicted in Figure 3, for small ranges (as we have) of our observables (tedwp, h_out, t_out) values at least to a monotonic tedwp(h_out,t_out) function we should expect. More than that, correlation analysis from *Table 3* reveals that 'monotone association' is less likely than a particular case of it, 'linear association'. Note that this it not means by necessity that the model is linear, it means only that the linear model is a 'approximating enough' model for the association in small ranges of the values of the observables and is according to our common manner of $\sin(\Theta) \sim \Theta$ approximations when Θ are small enough. Going further with this reasoning, a polynomial $\hat{Y}=f(X,Z)$ of a given range is expected to increase the accuracy of the estimates $\hat{Y}$, but not of the confidence of the coefficients of the model too.

Another point of view expecting an answer may be: May down-sampling from one minute to ten minutes to produce loosing of information? - Always down-sampling produces the loose of the information. Thus, the question should be reformulated: the loose of the information may affect the interpretation? - And here the answer is - definitely no. When a model cannot be rejected (as the proof given in *Table 4*) the increasing of the number of observations (let's say ten times more observations) the only expectation which may have is to obtain even smaller confidence intervals for same probability coverage (in an exact ratio of square root of ten smaller, given by the expression of standard error of the sampling as function of population standard deviation and the sample size). Thus, the increasing of the sample size is expected not to reject the observations drawn from *Table 4* ($\text{Intercept}_{\text{t1dwp}} \notin \text{CI}_{95\%}(\text{Intercept})_{\text{t\_dwp}}$ and the following ones), but on contrary, to even strongly proof it.

The present study focused on the inference of a small group of six out of forty-four weather stations provided parameters (more exactly six out of twenty-two weather related parameters). The selection of the five out of six parameters (outside dew point temperature entering by default into the analysis) was based on physical reasoning - the known relationship giving an approximation ('approximating enough model') of dew point temperature as function of temperature and relative humidity as well as other three parameters characterizing three physical phenomena which may infer the occurrence of the dew point - solar radiation, air pressure and air movement. The study revealed (see *Figure 4*) that the most dew point estimates affecting parameter from all three is wind speed. Thus, the study reveals that a better estimating accuracy for dew point should be obtained when the model of the estimate it include together with temperature and relative humidity the wind speed too.

Conclusions

Cluster analysis of the data recorded on January 8, 2011 in three different locations revealed that:

÷ Wind speed observations array is a very good location indicator in terms of terms of the height from the ground position.
÷ Dew point observations array is a very good location indicator in terms of horizontal geographical position.
÷ Dew point temperature model depending on air temperature and relative humidity proved to have not enough accuracy when location are changed, and at least one more parameter (such as wind speed, as present study shown) are necessary in the equation to correct it's accuracy.

References

[1] Barber, C., Bockhorst, J., Roebber, P., "Auto-Regressive HMM Inference with Incomplete Data for Short-HorizonWind Forecasting", *Advances in Neural Information Processing Systems*, 23:136-144, 2010.
[2] Hausmann S., Pienitz R., "Seasonal climate inferences from high-resolution modern diatom data along a climate gradient: a case study", *Journal of Paleolimnology*, 38(1):73-96, 2007, DOI: 10.1007/s10933-006-9061-2.
[3] Barreca, A., "Climate Change, Humidity, and Mortality in the United States", *Repec EconPapers* 906(R1):1-55, 2009.
[4] Li, Z., Liu, X.-H., Lun, Z., Jiang, Y., "Analysis on the ideal energy efficiency of dehumidification process from buildings", *Energy and Buildings*, 42(11):2014-2020, 2010.
[5] Bălan, M.C., Damian, M., Jäntschi, L., "Preliminary Results on Design and Implementation of a Solar Radiation Monitoring System", *Sensors*, 8(2):963-978, 2008.
[6] Bălan, M.C., Jäntschi, L., Bolboacă S.D., Damian, M., "Assessment of Thermal Solar Collectors Behaviour in Transitory Regime", *Polish Journal of Environmental Studies*, 19(1):231-241, 2010.
[7] Bălan, M.C., Damian, M., Jäntschi, L., "Solar Radiation Monitoring System", *Actual Tasks on Agricultural Engineering*, 36:507-517, 2008.
[8] Bălan, M.C., Bolboacă, S.D., Jäntschi, L., Weather Monitoring: Wind Analysis (May, 2009; GPS: Lat. N46°45'35"; Long. E23°34'19"), *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, 66(1):7-9, 2009.
[9] Bălan, M.C., Jäntschi, L., Bolboacă, S.D., Ştefu, M., Sestraş, R.E., "Weather monitoring setup and analysis of air, soil and leaf parameters", *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, 66(2):672-679, 2009.

[10] Bălan, M.C., Bolboacă, S.D., Sestraş, R.E., Jäntschi, L., "Experimental Setup to Study the Local Renewable Energy Potential and the Environment Influence on Fruits Growing", *Actual Tasks on Agricultural Engineering*, 37:265-271, 2009.

[11] Petrişor, C., Mitre, V., Mitre, I., Bălan, M.C., Jäntschi, L., Sestraş, A.F., Bolboacă, S.D., Sestraş, R.E., "Response of Some New Apple Varieties to Natural Infection with Apple Scab, under Conditions of Excessive Rainfalls", *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, 67(1):483, 2010.

[12] Tryon, R.C. "Cluster analysis; correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality", Edwards Brothers, Ann Arbor, MI, 1939 (122p), LOC: 39025023.

[13] Frades, I., Matthiesen, R., "Overview on techniques in cluster analysis", Methods in molecular biology (Clifton, N.J.) 593:81-107, 2010.

[14] Dorling, S.R., Davies, T.D., Pierce, C.E., "Cluster analysis: A technique for estimating the synoptic meteorological controls on air and precipitation chemistry - method and applications", *Atmospheric Environment - Part A General Topics* 26A(14): 2575-2581, 1992.

[15] Davies Inc., "Derived Variables in Davis Weather Products", *Application Note* (Rev A) 28:1-23, May 11, 2006.

[16] World Meteorological Organization, "Guide to Meteorological Instruments and Methods of Observation", Geneva, Switzerland, 6th Ed., 1996.

[17] Pearson, K., "Regression, heredity, and panmixia", *Philosophical Transactions of the Royal Society of London, Ser. A*, 187:253-318, 1896.

[18] Bolboacă S.D., Jäntschi, L., "Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds", *Leonardo Journal of Sciences*, 5(9):179-200, 2006.

[19] Jäntschi, L., Bolboacă, S.D., "The Jungle of Linear Regression Revisited", *Leonardo Electronic Journal of Practices and Technologies*, 6(10):169-187, 2007.

[20] Jäntschi, L., Bolboacă, S.D., "Observation vs. Observable: Maximum Likelihood Estimations according to the Assumption of Generalized Gauss and Laplace Distributions", *Leonardo Electronic Journal of Practices and Technologies*, 8(15):81-104, 2009.

[21] Haan, C.T., "Statistical methods in hydrology", Iowa State University Press, Ames, IA, 1977 (378p), LOC: 77011734.

[22] Jäntschi, L., "Annex 3. Descriptive and inferential statistics" In: "Genetic Algorithms and Their Applications", PhD Thesis in Horticulture (Supervisor: Prof. Sestraş R. E.), UASVM Cluj-Napoca, 2010.

[23] Bolboacă S.D., Jäntschi, L., "Comparison of QSAR Performances on Carboquinone Derivatives", *TheScientificWorldJOURNAL*, 9(10):1148-1166, 2009, DOI: 10.1100/tsw.2009.131.