Studia Universitatis Babes-Bolyai Chemia 61(1): 223-234

Correlation study among boiling temperature and heat of vaporization

Mihaela L. UNGURESAN[a], Lorena L. PRUTEANU[b], Lorentz JÄNTSCHI[a,b,*]

[a] *Technical University of Cluj-Napoca, Faculty of Material Sciences, 103-105 Muncii Blvd., RO-400641, Cluj-Napoca, Romania*
[b] *Babeş-Bolyai University, Faculty of Chemstry and Chemical Engineering, 11 Arany Janos str., RO-400028, Cluj-Napoca, Romania*
* *corresponding author lorentz.jantschi@gmail.com*

ABSTRACT. In this study is communicated a preliminary result from a property-property study on a series of chemical compounds regarding on the obtaining of a quantitative relationship between the properties. The study was conducted on a series of 190 inorganic chemical compounds for which both properties taken into study are known. The correlation analysis revealed that is a strong relationship between the boiling temperature and the heat of vaporization at that temperature, having the variance in the paired series of data explained at over 90%.

Keywords: *Property-property relationships, Distribution analysis, Regression analysis*

INTRODUCTION

Regression analysis and distribution of errors

Even if first studies about binomial expressions were made by Euclid [1], the mathematical basis of the binomial distribution study was put by Jacob Bernoulli [1654-1705]. The Bernoulli's studies, with significance for the theory of probabilities [2], were published 8 years later after his death by his nephew, Nicolaus Bernoulli. In *Doctrinam de Permutationibus & Combinationibus* section of this fundamental work he demonstrates the Newton binomial series expansion. Later, Abraham De Moivre [1667-1754] put the basis of approximated calculus for binomial distribution approximation using the normal distribution [3]. Later, Johann Carl Friedrich Gauss [1777-1855] put the basis of mathematical statistics [4].

The simplest association model is linear association. The model assumes that exist a relationship between two paired characteristics expressed by a straight line. The expression of this association is given by the implicit equation of a straight line: $ax + by + c = 0$. If $a = 0$ then the equation of the line reduces to $by + c = 0$. If further $c \neq 0$ gives a relationship which defines the mean of the Y associated characteristic but no relationship with X. Similarly if $b = 0$ then the equation of the line reduces to $ax + c = 0$ and if further $c \neq 0$ gives a relationship which defines the mean of the X associated characteristic but no relationship with Y. The remained case, if $c = 0$ defines a degenerated linear model in which is no intercept between the characteristics X and Y.

Which expression of the linear equation should be use is a matter of experimental error treatment. Going further, if a linear model defines the relationship between the X and Y characteristics, then if we take samples $(x_i, y_i)_{1 \le i \le n}$ of these two (X and Y) characteristics then the relationship in terms of a experimental error it should still seen.

The information related with the distribution of the error is important. A common assumption is to expect that an error ε_i (or η_i) to occur in equal probability as an error $-\varepsilon_i$ (or $-\eta_i$), and accordingly the distribution of the experimental error is symmetrical.

An experimental design that gives different weight to the errors led to a weighted regression. Usually the weight are function of the observable and/or expectance ($v_i = f(x_i, \hat{x}_i)$, $w_i = g(y_i, \hat{y}_i)$). The reason of giving weight to the errors is to normalize (e.g. the distribution of the errors to become normal or at least to have a known distribution). Another assumption regarding the experimental error (ε_i and η_i) that must be taken in consideration in order to obtain the estimations of the population parameters is 'the experimental error follows a known distribution'.

Structure-activity relationships

Deriving of the first (big) family of molecular descriptors was communicated in [5] and about ten years after the usage potential of this sort of methodology investigating structure-activity relationships were significantly increasing by developing the search with methodology employing genetic algorithms [6].

Relationships that are more common are property-property relationships that often occur due to the intrinsic relationships between the derivatives of thermodynamic functions (see for details [7]).

Non-linear relationships are more difficult to identified due to the unknown dependence form, and are known as being less efficient in prediction than the linear ones, even so in some cases may outperform the linear models (as were obtained in [8]).

Physico-chemical properties such as the heat of vaporization are of technical interest for designing of devices working at transition phase between gaseous and liquid state [9, 10,11].

For a small part of chemical compounds are available measurements of physico-chemical parameters and at least for a part of them the values are regularly updated (a representative source is given in [12]); this is one of the reasons for which seeking for relationships among properties is legitimate.

In this work, a computational study was conducted for a series of 190 inorganic chemical compounds to relate the molar enthalpy (heat) of vaporization ($\Delta_{vap}H$) at the normal boiling point (t_b) referred to a pressure of 101.325 kPa (760 mmHg) with their boiling point (t_b). Our aim was to investigate if transformation of variables to lead to normal distribution gives a significant simple regression model linking the boiling temperature with heat of vaporization.

MATERIALS AND METHODS

The data were taken from a recent edition of the serial containing reference physical and chemical data [13] and refers both the boiling point and the heat of vaporization, the primary study reporting these values being [14].

The chemical compounds included in the study, listed in the ascending order of their boiling point, are: helium (He), hydrogen (H_2), Neon (Ne), Nitrogen (N_2), Fluorine (F_2), Argon (Ar), Oxygen (O_2), Krypton (Kr), Fluorine monoxide (F_2O), Nitrogen trifluoride (NF_3), Silane (SiH_4), Xenon (Xe), Phosphorus(III) fluoride (PF_3), Chlorine fluoride (ClF), Boron trifluoride (BF_3), Fluorosilane ($SiFH_3$), Trifluorosilane (SiF_3H), Diborane (B_2H_6), Germane (GeH_4), Phosphine (PH_3), Hydrogen chloride (HCl), Phosphorus(V) fluoride (PF_5), Difluorosilane (SiF_2H_2), Tetrafluorohydrazine (N_2F_4), Chlorotrifluorosilane ($SiClF_3$), Hydrogen bromide (HBr), Arsine (AsH_3), Nitrosyl fluoride (NFO), Hydrogen sulfide (H_2S), Difluorine dioxide (F_2O_2), Arsenic(V) fluoride (AsF_5), Phosphorothioc trifluoride (PSF_3), Stannane (SnH_4), Phosphorus(III) chloride difluoride ($PClF_2$), Perchloryl fluoride ($ClFO_3$), Thionyl fluoride (SOF_2), Hydrogen selenide (H_2Se), Sulfur tetrafluoride (SF_4), Hydrogen iodide (HI), Chlorine (Cl_2), Tetrafluorodiborane (B_2F_4), Ammonia (NH_3), Dichlorodifluorosilane ($SiCl_2F_2$), Chlorosilane ($SiClH_3$), Stibine (SbH_3), Disilane (Si_2H_6), Sulfur dioxide (SO_2), Nitrosyl chloride (NClO), Hydrogen telluride (H_2Te), Bromosilane ($SiBrH_3$), Chlorine monoxide (Cl_2O), Thionitrosyl fluoride (FNS), Dichlorosilane (Cl_2H_2Si), Chlorine dioxide (ClO_2), Chlorine trifluoride (ClF_3), Boron trichloride (BCl_3), Phosphorus(III) dichloride fluoride (PCl_2F), Tungsten(VI) fluoride (WF_6), Tetraborane(10) (B_4H_{10}), Bromine fluoride (BrF), Digermane (Ge_2H_6), Trichlorosilane ($SiHCl_3$), Rhenium(VI) fluoride (ReF_6), Molybdenum(VI) fluoride (MoF_6), Hydrazoic acid (HN_3), Bromine pentafluoride (BrF_5), Aluminum borohydride (AlB_3H_{12}), Sulfur trioxide (SO_3), Osmium(VI) fluoride (OsF_6), Vanadium(V) fluoride (VF_5), Trisilane (Si_3H_8), Iridium(VI) fluoride (IrF_6), Arsenic(III) fluoride (AsF_3), Tetrachlorosilane ($SiCl_4$), Bromine (Br_2), Diphosphine (P_2H_4), Pentaborane(11) (B_5H_{11}), Dibromosilane ($SiBr_2H_2$), Sulfuryl chloride (SO_2Cl_2), Hydrogen disulfide (H_2S_2), Thionyl chloride ($SOCl_2$), Phosphorus(III) chloride (PCl_3), Germanium(IV) chloride ($GeCl_4$), Boron tribromide (BBr_3), Water (H_2O), Iodine pentafluoride (IF_5), Selenium tetrafluoride (SeF_4), Phosphoryl chloride (PCl_3O), Tribromosilane ($SiHBr_3$), Trigermane (Ge_3H_8), Hydrazine (N_2H_4), Tin(IV) chloride ($SnCl_4$), Chromium(VI) dichloride dioxide ($CrCl_2O_2$), Bromine trifluoride (BrF_3), Vanadyl trichloride ($VOCl_3$), Arsenic(III) chloride ($AsCl_3$), Titanium(IV) chloride ($TiCl_4$), Hydrogen peroxide (H_2O_2), Vanadium(IV) chloride (VCl_4), Tetrabromosilane ($SiBr_4$), Rhenium(VI) oxytetrafluoride (ReF_4O), Phosphorus(III) bromide (PBr_3), Iodine (I_2), Rhenium(VII) dioxytrifluoride (ReF_3O_2), Tungsten(VI) oxytetrafluoride (WOF_4), Molybdenum(VI) oxytetrafluoride (MoF_4O), Germanium(IV) bromide ($GeBr_4$), Phosphoryl bromide (PBr_3O), Gallium(III) chloride ($GaCl_3$), Tin(IV) bromide ($SnBr_4$), Boron triiodide (BI_3), Molybdenum(V) fluoride (MoF_5), Antimony(III) chloride ($SbCl_3$), Arsenic(III) bromide ($AsBr_3$), Rhenium(V) fluoride (ReF_5), Phosphorus(III) iodide (PI_3), Tantalum(V) fluoride (TaF_5), Tungsten(VI) oxytetrachloride ($WOCl_4$), Osmium(V) fluoride (OsF_5), Titanium(IV) bromide ($TiBr_4$), Niobium(V) fluoride (NbF_5), Tantalum(V) chloride ($TaCl_5$), Niobium(V) chloride ($NbCl_5$), Aluminum bromide ($AlBr_3$), Molybdenum(V) chloride ($MoCl_5$), Gallium(III) bromide ($GaBr_3$), Phosphorus (P), Tetraiodosilane (SiI_4), Antimony(III) bromide ($SbBr_3$), Mercury(II) chloride ($HgCl_2$), Mercury(II) bromide ($HgBr_2$), Tungsten(VI) chloride (WCl_6), Gallium(III) iodide (GaI_3), Tantalum(V) bromide ($TaBr_5$), Mercury(II) iodide (HgI_2), Mercury (Hg), Tin(IV) iodide (SnI_4), Titanium(IV) iodide (TiI_4), Aluminum iodide (AlI_3), Tellurium tetrachloride ($TeCl_4$), Antimony(III) iodide (SbI_3), Arsenic(III) iodide (AsI_3), Bismuth trichloride ($BiCl_3$), Sulfur (S), Bismuth tribromide ($BiBr_3$), Beryllium chloride ($BeCl_2$), Beryllium iodide (BeI_2), Tin(II) chloride ($SnCl_2$), Tin(II) bromide ($SnBr_2$), Indium(I) bromide (BrIn), Zinc bromide ($ZnBr_2$), Selenium (Se), Indium(I) iodide (InI), Tin(II) iodide (SnI_2), Thallium(I) chloride (ClTl), Zinc chloride ($ZnCl_2$), Cadmium iodide (CdI_2), Cadmium (Cd), Thallium(I) bromide (BrTl), Thallium(I) iodide (ITl), Cadmium bromide ($CdBr_2$), Lead(II) iodide (PbI_2), Lead(II) bromide ($PbBr_2$), Thorium(IV) chloride ($ThCl_4$), Titanium(III) chloride ($PbCl_2$), Titanium(III) chloride ($TiCl_3$), Cadmium chloride ($CdCl_2$), Tellurium (Te), Chromium(II) chloride ($CrCl_2$), Molybdenum(VI) oxide (MoO_3), Lead(II) fluoride (PbF_2), Thallium(I) sulfide (STl_2), Sodium hydroxide (NaOH), Titanium(II) chloride ($TiCl_2$), Zinc fluoride (ZnF_2), Silver(I) bromide (AgBr), Silver(I) iodide (AgI), Silver(I) chloride (AgCl), Bismuth (Bi), Lithium hydroxide (LiOH), Lithium fluoride (LiF), Thorium(IV) fluoride (ThF_4), Lead (Pb), Cadmium fluoride (CdF_2), Barium (Ba), Gallium (Ga), Aluminum (Al), Germanium (Ge), Gold (Au), and Boron (B).

In order to relate the properties, the following methodology of analysis was applied:

- Analysis the distribution of the boiling temperature values; if the values are not normally distributed, then find the transformation which normalizes it;
- Analysis the distribution of the heat of vaporization values; if the values are not normally distributed, then find the transformation which normalizes it;
- On the normalized data, by keeping the association given by the chemical compound on which these properties were measured, conduct the regression analysis;
- After identification of the regression model, use the inverse of the transformations, which normalizes the data to analyze the model.

The analysis of the distribution was conducted with EasyFit [15] and the analysis of regression was conducted with Excel [16]. The distribution parameters were estimated using the Maximum Likelihood Method (MLE, [17]), and the agreement between the observations and the model were measured using Anderson-Darling statistic ([18]) and Kolmogorov-Smirnov statistic ([19, 20]).

RESULTS AND DISCUSSION

The distribution analysis of the boiling temperature revealed that the normal distribution is rejected at all conventional levels of significance over 20% risks to be in error (Table 1). The analysis of lognormal distribution, were found that the location parameter determined by the maximum likelihood estimation is -309.79. This value is near to -273.15 and suggests that a transformation of the scale from Celsius degrees to Kelvin degrees will lead to normalization of data. Indeed, after this transformation (T=t°C+273.15) the data series become lognormal distributed, and the hypothesis of the distribution cannot be rejected at a significance level of 5%. Thus, the probability associated with the Anderson-Darling statistic is 9.31% and the probability associated with the Kolmogorov-Smirnov statistic is 13.34%. Therefore, the data were further transformed with logarithm function and analyzed again. The probability associated with Anderson-Darling statistic become 9.07% and the probability associated with Kolmogorov-Smirnov statistic become 12.81% (see Figure 1 below), while the estimations of the population statistics were μ=6.0873 and σ=0.90038.

Table 1. H_0 (Data follow normal distribution): Results for different significance levels

Reject H_0?	**Boiling temperature**				**Heat of vaporization**			
	α=0.2	**α=0.1**	**α=0.05**	**α=0.01**	**α=0.2**	**α=0.1**	**α=0.05**	**α=0.01**
K-S	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
A-D	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
CS	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

K-S = Kolmogorov-Smirnov; A-D = Anderson-Darling; CS = Chi-square

The distribution analysis of the heat of vaporization also revealed that the normal distribution is rejected at all conventional levels of significance over 20% risk being in error (Table 1). Looking for lognormal distribution, were found the three parameters lognormal distribution with the location parameter determined by the maximum likelihood estimation method as being -3.3553. This value was used to transform the observed data. After this transformation (ΔH_1 = $\Delta H(t_b)$+3.3553) the data series become lognormal distributed, when the hypothesis of the distribution cannot be rejected at 5% risk being in error. Thus, the probability associated with the Anderson-Darling statistic is 23.53% and the probability associated with the Kolmogorov-Smirnov statistic is 16.34%. Therefore, the data were further transformed with logarithm function and analyzed again. The probability associated with the Anderson-Darling statistic become 23.95% and the probability associated with the Kolmogorov-Smirnov statistic become 16.31% (see Figure 2), and when estimations of the population statistics were μ = 3.8313 and σ = 0.84324.

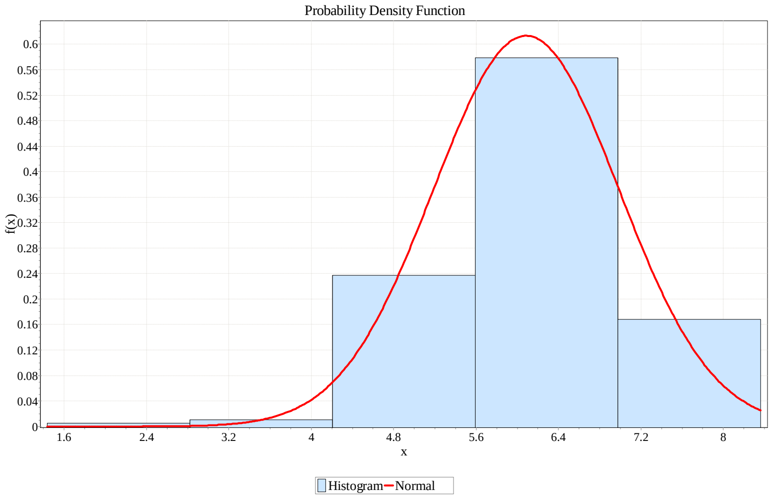

Figure 1. Distribution fit for the transformed boiling temperatures as ln(b.p.°C+273.15)

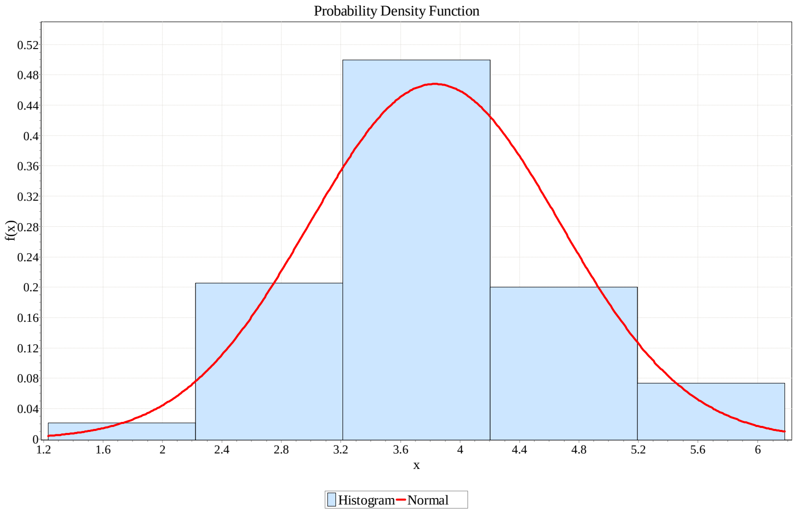

Figure 2. Distribution fit for the transformed heat of vaporization as ln(HVAP kJ/mol+3.3553)

Regression analysis was applied on original data set and normalized data set and the obtained equations were analyzed to see if significant differences between models exist. Both investigated models (created using original and transformed data) proved significantly (Table 2), with a higher contribution to the intercept for the model obtained on original data and of the heat of vaporization on the model with transformed data.

Table 2. Characteristics of obtained models

Model	Original data	Normalized data
R^2	0.9574	0.9259
R^2_{adj}	0.9572	0.9255
RMSE	14.16	0.23
MAE	38.86	0.63
MAPE	7.77	0.18
F (p)	4224 (<0.0001)	2349 (<0.0001)
Int [95%CI]	24.80 [22.46; 27.14]	-1.65 [-1.88; -1.43]
Coeff [95%CI]	0.11 [0.10; 0.11]	0.90 [0.86; 0.94]

R^2 = determination coefficient;
R^2_{adj} = adjusted determination coefficient;
RMSE = root mean square error;
MAE = mean absolute error;
F = Fisher’s statistic; p = probability to be in error;
Int = intercept; 95%CI = 95% confidence interval;
Coeff = the value of coefficient associated to heat of vaporization

Our findings showed that the model created on original data (R=0.9785) had a significantly (p=0.0059) higher correlation coefficient compared with the model obtained on transformed data (R=0.9622). However, the values of the root mean square error (RMSE) and the mean absolute error (MAE) showed that the model obtained on transformed data is more reliable (small values of both RMSE and MAE).

The leave-one-out analysis was carried out to assess the internal validity of the models and the main characteristics of the models are given in Table 3.

Table 3. Characteristics of models in leave-one-out analysis

Model	Q^2	RMSE	MAE	MAPE	F_{loo} (p_{loo})
Original data	0.9551	14.47	7.69	0.40	3995 ($p<0.0001$)
Transformed data	0.9182	0.24	0.15	0.05	2104 ($p<0.0001$)

Q^2 = determination coefficient in leave-one-out (loo) analysis
RMSE = root mean square error;
MAE = mean absolute error;

The root mean square error, mean absolute error and mean absolute percent error are smaller in the model with transformed data (Table 3) and thus sustain the validity and reliability of this model even that the determination coefficient is smaller compared to that obtained on original data.

A training and test analysis was conducted to assess the validity of identified model, with 126 compounds in training set and the 64 in test set. The equation for model with original data is given in Eq(1)

$Y = 25.218+0.107*X$ Eq(1)

$R^2_{Tr} = 0.9741$; $n = 126$

$R^2_{Ts} = 0.9334$; $n = 64$

while the equation for model with transformed data is given in Eq(2):

$Y = -2.145 +0.982*X$ Eq(2)

$R^2_{Tr} = 0.9633$; $n = 126$

$R^2_{Ts} = 0.8888$; $n = 64$

Graphical representation of performances in training and test analysis is given in Figure 3 for original data and in Figure 4 for transformed data.

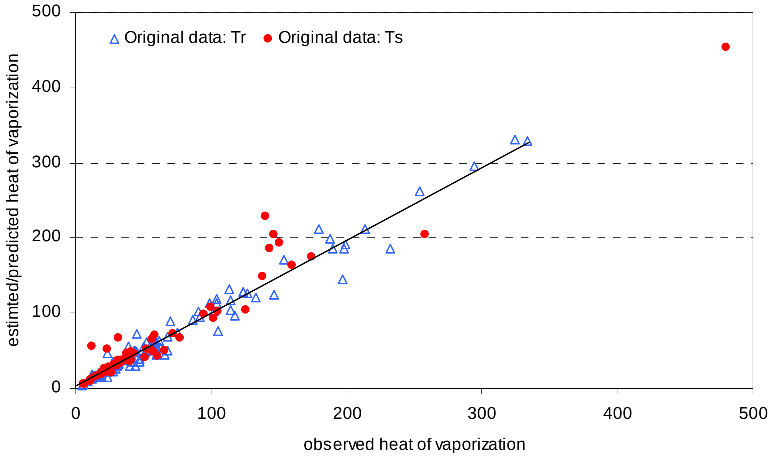

Figure 3. Training vs test analysis: original data

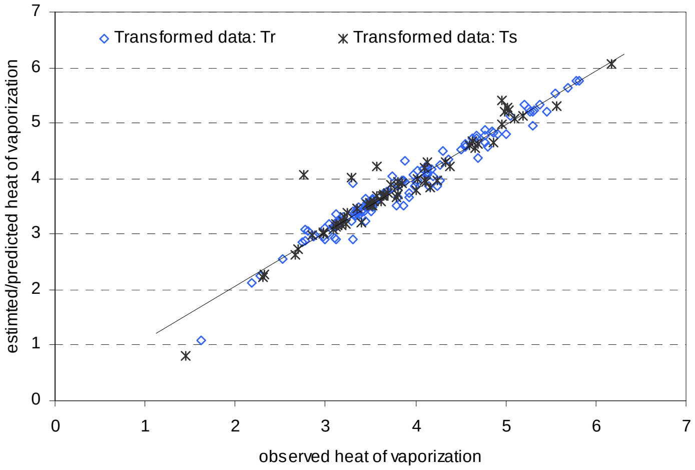

Figure 4. Training vs test analysis: transformed data

CONCLUSIONS

As can be concluded from this analysis, it seems that these two properties (boiling point and heat of vaporization at the boiling point) have a large part of their variance explained by one to each other and are suitable for a more detailed study meant to increase the explanatory power.

ACKNOWLEDGMENTS

This paper is dedicated to Prof. Mircea V. Diudea with the occasion of his 65th birthday.

REFERENCES

1. J. L. Coolidge, *The American Mathematical Monthly*, **1949**, *56(3)*, 147.
2. J. Bernoulli, "Ars Conjectandi", Thurnisius, Basel, **1713**.
3. A. De Moivre, "Approximatio ad Summam Terminorum Binomii (a+b)n in Seriem expansi" (presented privately to some friends in 1733), In: The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play (2nd ed.), London, W. Pearson, **1738**, p. 235-243.
4. J. C. F. Gauss, *Comment. Societ. R. Sci. Gottingensis Recentiores*, **1823**, *5*, 33.
5. L. Jäntschi, Prediction of Physical, Chemical and Biological Properties using Mathematical Descriptors (in Romanian). PhD Thesis in Chemistry (PhD Advisor: Prof. Dr. Mircea V. DIUDEA). Cluj-Napoca: Babeş-Bolyai University, **2000**.
6. L. Jäntschi, Genetic Algorithms and their Applications (in Romanian). PhD Thesis in Horticulture (PhD Advisor: Prof. Dr. Radu E. SESTRAŞ). Cluj-Napoca: University of Agricultural Sciences and Veterinary Medicine, **2010**.
7. L. Jäntschi, S. D. Bolboacă, *Journal of Computational Science*, **2014**, *5(4)*, 597.
8. L. Jäntschi, S. D. Bolboacă, *Studia Universitatis Babeş-Bolyai. Series Chemia*, **2010**, *LV(4)*, 61.
9. C. N. Markides, R. B. Solanki, A. Galindo, *Applied Energy*, **2014**, *124*, 167.
10. L. R. Erickson, E. K. Ungar, *AIAA SPACE 2013 Conference and Exposition*, **2013**, 99748.
11. M. Son, J. Koo, W. Cho, E. Lee, *Journal of Thermal Science*, **2012**, *21(5)*, 428.
12. W. M. Haynes (Ed.), CRC Handbook of Chemistry and Physics. 95th Edition, Chapman and Hall/CRCnetBASE, Boca Raton, FL, Internet Version, **2015**.
13. D. R. Lide, CRC Handbook of Chemistry and Physics, 90th Edition Internet Version. Boca Raton, FL: Chapman and Hall/CRCnetBASE, **2010**.
14. M. W. Chase, Jr., C. A. Davies, J. R. Downey, Jr. D. J. Frurip, R. A. McDonald, A. N. Syverud, *Journal of Physical and Chemical Reference Data*, **1985**, *14(S1)*, 1856p.
15. MathWave Technologies, 2009. EasyFit Proffesional v.5.2 (software). Web site: http://mathwave.com
16. Microsoft® Excel® 2002 SP3, Copyright© Microsoft Corporation 1987-2001.
17. R. A. Fisher, *Messenger of Mathematics*, **1912**, *41*, 155.
18. T. W. Anderson, D. A. Darling, *Ann Math Stat*, **1952**, *23(2)*, 193.
19. A. Kolmogorov, *Ann Math Stat*, **1941**, *12(4)*, 461.
20. N. V. Smirnov, *Ann Math Stat*, **1948**, *19(2)*, 279.