STUDIA UBB CHEMIA, LXI, 1, 2016 (p. 223-233)
(RECOMMENDED CITATION)

Dedicated to Professor Mircea Diudea on the Occasion of His 65th Anniversary

CORRELATION STUDY AMONG BOILING TEMPERATURE AND HEAT OF VAPORIZATION

MIHAELA L. UNGUREŞAN[a], LORENA L. PRUTEANU[b], LORENTZ JÄNTSCHI[a,b,*], SORANA D. BOLBOACĂ[c]

ABSTRACT. In this paper, a preliminary result from a property-property analysis on a series of chemical compounds in regards of quantitative relationship between two properties is communicated. 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 point and the heat of vaporization at the boiling temperature, having the variance in the paired series of data explained over 90%.

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

INTRODUCTION

Regression analysis and error distribution

Even the 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,

[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 Chemistry and Chemical Engineering, 11 Arany Janos str., RO-400028, Cluj-Napoca, Romania*

[c] *Iuliu Haţieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur str., RO-400349, Cluj-Napoca, Romania*

* *Corresponding author: lorentz.jantschi@gmail.com*

Nicolaus Bernoulli. In *Doctrinam de Permutationibus & Combinationibus* section of this fundamental work he demonstrated 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. The model assumes that there exists 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 $bY + c = 0$. Next, if $c \neq 0$ results in a relationship which defines the mean of Y associated characteristic but no relationship with X. Similarly if $b = 0$ then the equation of line reduces to $aX + c = 0$ and if further $c \neq 0$ leads to a relationship which defines the mean of X associated characteristic but no relationship with Y. The remained case, if $c = 0$ defines a degenerated linear model in which there is no intercept between the characteristics X and Y.

Which expression of the linear equation should be used 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, a relationship in terms of experimental errors would be defined.

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

An experiment design that gives different weights to the errors led to a weighted regression. Usually the weights are function of the observable and/or expectance ($v_i = f(x_i, \hat{x}_i)$, $w_i = g(y_i, \hat{y}_i)$). Weighted errors involve data normalization, e.g. normalization of errors distribution or at least having a known error distribution; knowledge on error distribution is essential in the estimation of population parameters.

Structure-activity relationships

Building of the first (big) family of molecular descriptors was described in [5] and, about ten years after, the usage potential of this sort of methodology investigating structure-activity relationships was significantly increased by joining with genetic algorithms [6].

Relationships commonly called property-property relationships have been developed due to the intrinsic relations between some thermodynamic functions (see for details [7]).

Non-linear relationships are possible but are less desirable, being less efficient in prediction than the linear ones, also more difficult to interpret, even in some cases may over perform the linear models (see [8]).

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

Data on the measured physico-chemical parameters are available for relatively few chemical compounds (a representative source is given in [12]); this is one of the legitimate reasons for developing relationships among properties.

In this paper, a computational study was drawn 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. Our aim was to find if variable transformation leading to normal error distribution would provide significantly simple regression models, able to links the boiling temperature with the heat of vaporization.

RESULTS AND DISCUSSION

The error distribution analysis of the boiling temperature revealed that the normal distribution is rejected at all conventional levels of significance over 20% risk to be in error (Table 1). The analysis of lognormal distribution, has 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 became 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 A-D statistic is 9.31% and the probability associated with the Kolmogorov-Smirnov K-S statistic is 13.34%. Therefore, the data were further transformed by the logarithmic function and analysed again. The probability associated with Anderson-Darling statistic become 9.07%, the probability associated with Kolmogorov-Smirnov statistic become 12.81% (see Figure 1 below) while the estimations of population statistics were μ=6.0873 and σ=0.90038.

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

Reject	**Boiling temperature**				**Heat of vaporization**			
H_0?	**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 of error (Table 1). Looking for lognormal distribution, the three parameters of lognormal distribution (with the location parameter determined by the maximum likelihood estimation method) were found as being -3.3553. This value was used to transform the observed data. After this transformation ($\Delta H_1 = \Delta H(t_b) + 3.3553$) the data series became lognormal distributed, when the hypothesis of the distribution cannot be rejected at 5% risk of 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 by the logarithmic function and analyzed again. The probability associated with the Anderson-Darling statistic became 23.95%, the probability associated with the Kolmogorov-Smirnov statistic became 16.31% (see Figure 2), and the estimations of the population statistics were $\mu = 3.8313$ and $\sigma = 0.84324$.

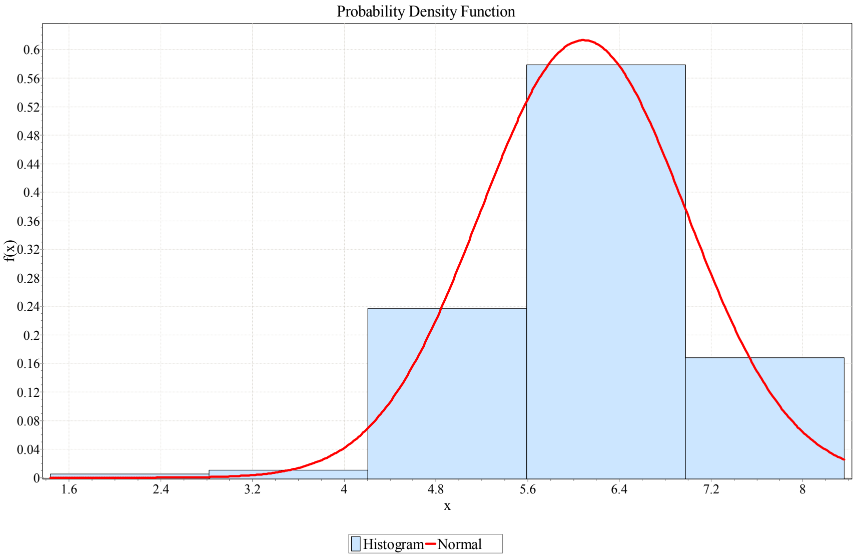

Figure 1. Distribution fit for the transformed boiling temperatures as ln(b.p.(K))

Regression analysis was applied on the original data set and normalized data set and the derived equations were analysed to see if significant differences between models exist. Both the investigated models (created using original and transformed data) proved to be significant (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.

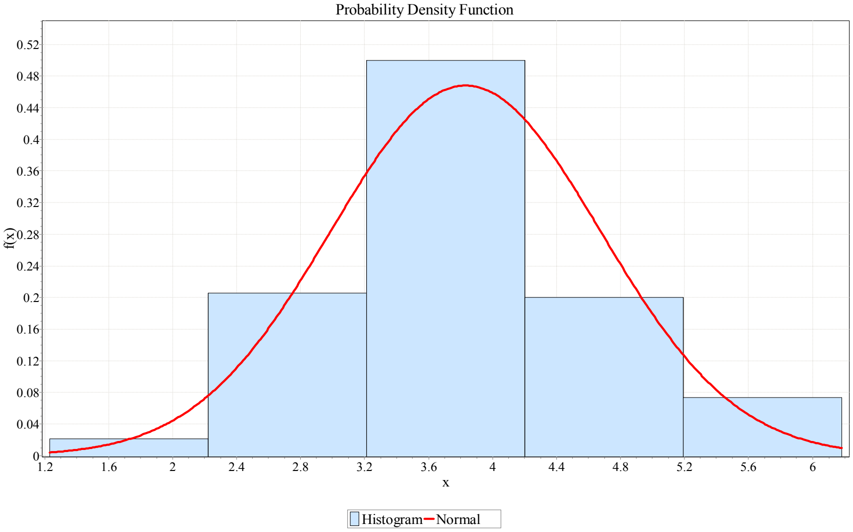

Figure 2. Distribution fit for the transformed heat of vaporization as ln(Δ_{vap}H +3.3553 kJ/mol)

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; the main characteristics of the models are given in Table 3.

Table 3. Characteristics of models in the 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;
MAPE = mean absolute percentage error

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

A training and test analysis was conducted to assess the validity of the identified model, with 126 compounds in the training set and 64 in the test set. The equation for model with original data is given in Eq(3)

$$\hat{Y} = 25.218+0.107*X \qquad \text{Eq(3)}$$
$$R^2_{Tr} = 0.9741;\ n = 126$$
$$R^2_{Ts} = 0.9334;\ n = 64$$

where $\hat{Y}$ approximates the heat of formation at boiling point temperature ($\hat{Y} \sim \Delta H(t_b)$) and X is the boiling point temperature ($X = t_b$, in Celsius degrees).

The equation for model with transformed data is given in Eq(4):

$$\hat{Y} = -2.145 +0.982*X \qquad \text{Eq(4)}$$
$$R^2_{Tr} = 0.9633;\ n = 126$$
$$R^2_{Ts} = 0.8888;\ n = 64$$

where $\hat{Y}$ approximates the logarithm of the heat of formation at boiling point temperature ($\hat{Y} \sim \ln(\Delta H(T_b)+3.3553)$) and X is the logarithm of the boiling point temperature ($X = \ln(T_b)$, in Kelvin).

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

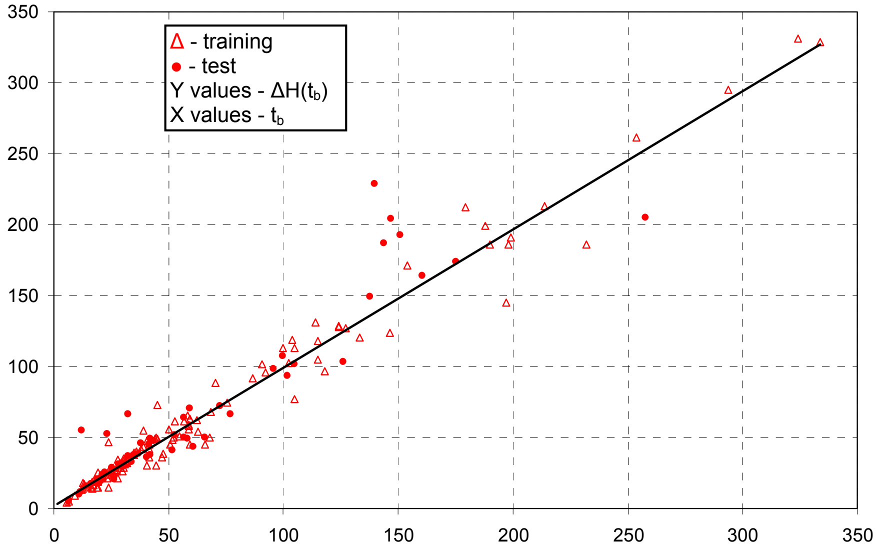

Figure 3. Training vs. test analysis: original data

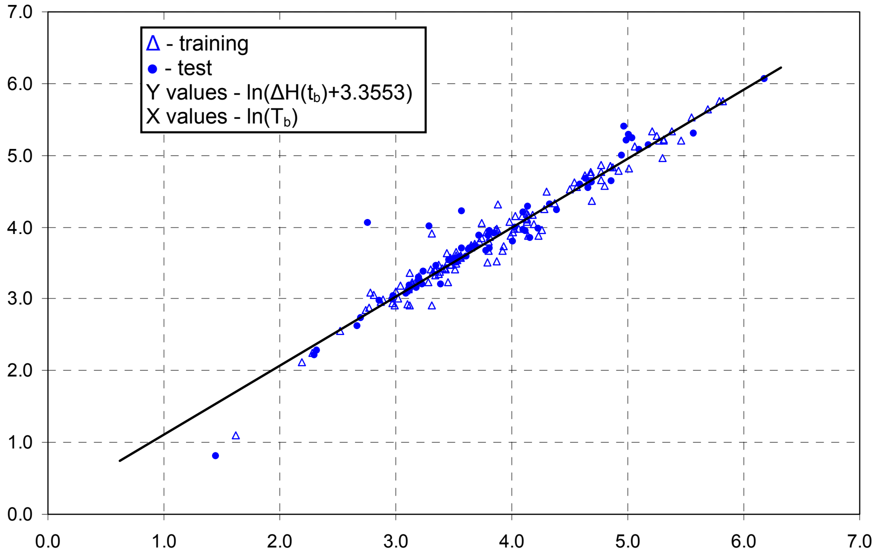

Figure 4. Training vs test analysis: transformed data

The training vs test analysis shows an apparently better agreement in both training and test sets (see Eq3 and Eq4) when the untransformed data are used. This should not lead to the conclusion that is better to be used the untransformed data, because for the untransformed data the assumptions of the linear regression are not accomplished. Also, the measurement units for the transformed data are not the same with the measurement units for original data. For instance, turning back to the original measurement units (by raising to the exponent of Y and $\hat{Y}$ values) the determination in the test set (between Y and $\hat{Y}$) for transformed data becomes 0.9321, a much closer value to that given in Eq3. This case - of having lower agreement when the data are properly transformed to accomplish the requirements of the regression analysis - is much more important than it seems – because, usually, the agreements are reported without checking the accomplishment of the requirements. The explanation of this fact relies on the intrinsic procedure of obtaining the coefficients, namely on the minimization of the sum of squares between observed and estimated values. If there exists some points at ends of the interval of values with large departures (differences between observed and estimated values, see for instance [12]) then the minimization of the sum of squares has the tendency to follow it, ignoring and penalizing other departures, and the same idea applies for the correlation coefficient.

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.

Conducting of the analysis without checking the assumptions of the analysis may lead to incorrect results, usually tending to produce more explanatory power than it is.

MATERIALS AND METHODS

Data were taken from a recent edition of the serial containing reference physical and chemical data [13] and refers to both the boiling point and the heat of vaporization, the primary study reporting these values being [14,15].

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:

- Analyses the distribution of the boiling temperature values; if the values are not normally distributed, then find the transformation which normalizes it;
- Analyses 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, draw the regression analysis;
- After identification of the regression model, use the inverse of the transformations, which normalizes the data to analyses the model.

The analysis of the distribution was conducted by EasyFit [16] and the analysis of regression was conducted by Excel [17]. The distribution parameters were estimated using the Maximum Likelihood Method (MLE, [18]), and the agreement between the observations and the model were measured using Anderson-Darling statistic ([19]) and Kolmogorov-Smirnov statistic ([20, 21]).

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 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. S.D. Bolboacă, L. Jäntschi, *Combinatorial Chemistry & High Throughput Screening*, **2013**, *16(4)*, 288.
13. W.M. Haynes (Ed.), CRC Handbook of Chemistry and Physics. 95th Edition, Chapman and Hall/CRCnetBASE, Boca Raton, FL, Internet Version, **2015**.
14. D.R. Lide, CRC Handbook of Chemistry and Physics, 90th Edition Internet Version. Boca Raton, FL: Chapman and Hall/CRCnetBASE, **2010**.
15. 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.
16. MathWave Technologies, 2009. EasyFit Proffesional v.5.2 (software). Web site: *http://mathwave.com*
17. Microsoft® Excel® 2002 SP3, Copyright© Microsoft Corporation 1987-2001.
18. R.A. Fisher, *Messenger of Mathematics*, **1912**, *41*, 155.
19. T.W. Anderson, D.A. Darling, *Annals of Mathematical Statistics*, **1952**, *23(2)*, 193.
20. A. Kolmogorov, *Annals of Mathematical Statistics*, **1941**, *12(4)*, 461.
21. N.V. Smirnov, *Annals of Mathematical Statistics*, **1948**, *19(2)*, 279.