American Journal of Applied Sciences

Original Research Paper

Least Absolute Deviation Regression and Least Squares for Modeling Retention Indices of Set Compounds Food and Pollutants of the Environment

Fatiha Mebarki, Khadija Amirat, Salima Ali Mokhnach and Djellol Messadi

Department of Chemistry, Laboratory of Environmental security and Food,
Badji Mokhtar Annaba University, Annaba, Algeria

Article history
Received: 22-12-2016
Revised: 04-06-2017
Accepted: 05-06-2017

Correspondins Author:
Khadija Amirat
Department of Chemistry,
Laboratory of Environmental
security and Food, Badji
Mokhtar Annaba University,
Annaba, Algeria
kadijatoumi@yahoo.com

Abstract: Considering the importance of the statistical analysis of regression in modeling based separately on study for Quantitative structure retention indices on Carbowax 20 M (I^{Cw20M}) and OV-101 columns ($I^{OV\text{-}101}$) relationships (QSRR) are determined for 114 pyrazines. The detection of influential observations for the standard least squares regression model is a problem which has been extensively studied. Least Absolute Deviation regression diagnostics offers alternative dicapproaches whose main feature is the robustness. Here a nonparametric method for detecting influential observations is presented and compared with other classical diagnostics methods. With have been applied for modeling separately retention indices of the same set of (89 pyrazines of Training and 25 of Test) eluted on Columns OV-101 and Carbowax-20M, using theoretical molecular descriptors derived from DRAGON Software and validating the results in the state approached graphically by Probability plot of the error and approached tests statistics of Anderson-Darling, in finished by the confidence interval thanks to robustness concept to check if errors distribution is really approximate.

Keywords: Least Absolute Deviation Regression, Robustness, Outliers, Leverage Points, Tests Statistics, Environmental

Introduction

Since the 1970 the environment term is used to indicate the global Ecologic context, i.e., the whole of the conditions physical, chemical, biological climatic and geographic conditions, in which are developed living conditions and humans being in particular. Air, earth, water, natural resources, flora, fauna, people and their social interactions are included.

The volatile heterocyclic constitute a significant family of odorous molecules, particularly interesting in the field of chemistry of the flavours and the odor can be regarded as a local pollution and a limited harmful effect to the bordering population of the potential sources. They represent more than one quarter of the 5 000 volatile compounds characterized up to now in our food

Pyrazines are heterocycles very present in our food. More than 80 derived from pyrazines are identified in a great number of cooked food, as bread, meat, torrefied coffee, the cocoa or hazel nuts; they are aromatizing compounds (Li *et al.*, 2014; Buchbauer, 2000).

Stanton and Jurs (1989), have used QSRR methodology to develop Models to link structural features of 107 pyrazines differently substituted, to their retention indices obtained up on two different polarities columns (OV-101 and Carbowax-20M). The equations have been calculated with the help of multilinear regression, the choice of the explanatory variables (topological, electronic and physical properties) being achieved by progressive elimination (Small and Jurs, 1983), among the 85 individual Molecular descriptors obtained for each whole molecule. The retention Indices (IR) obtained on each column are treated separately, while by drawing from the same sets of descriptors. The calculated models with 6 explanatory variables provide high standards errors (S = 23 units of index - u.i. - on OV-101 and S = 36.33 u.i .up on Carbowax- (20 M) which do not predict good predictive capacities for these models, which let to suppose nonlinear relations between descriptors and property (IR) studied (Mebarki *et al.*, 2016).

Science Publications

A large number of other estimation methods aimed at achieving robustness have been suggested and a considerable body of literature has developed. See for example, Gonin and Money (1989; Dodge, 1987) and the references therein. Generally the robust estimators in the literature can be classified as M-estimators, L-estimators, or R-estimators. Probably most attention has been paid to the *L*estimators, for other type estimators, Judge *et al.* (1985).

The robustness of Least Absolute Deviation method in relation with influential observations and its susceptibility to leverage point which are largely studied in literature (Dodge, 1987; 1997). We propose non parametric method Least Absolute Deviation (LAD) to detect the influential observations (aberrant and affect leverage) in comparison with least squares method.

The tests of normality as whereas theory-driven methods include the normality test such Anderson Darling test. However, seier classified the test of normality into major categories test, empirical and normality distribution of *the observed data.*

The Durbin-Watson statistic is conditioned on the order of the observations (rows). Minitab assumes that the observations are in a meaningful order, such as time order. The Durbin-Watson statistic determines whether or not the correlation between adjacent error terms is zero. To reach a conclusion from the test, you will need to compare the displayed statistic with lower and upper bounds in a table. If D > upper bound, no correlation exists; if D < lower bound, positive correlation exists; if D is in between the two bounds, the test is inconclusive.

The objective of this work aims at using QSRR methodology, in the approach Method Least Absolute Deviation/Least Square (LAD/OLS), to model retention indices of (114) pyrazines (113 taken from Stanton and Jurs (1979) (1) and one compound (2-VinylPyrazine) taken from Mihara and Enomoto (1985), the molecular descriptors are only calculated starting from the chemical structure of the compounds.

The linear statistical model for fixed effects will be examined relationships between retention index and different descriptors for two columns [(between retention indices of non polar column (OV- 101) and descriptor of Connectivity indices (are among the most popular topological indices (it is a descriptor of Structure-Activity Analysis), descriptor of Geometrical descriptors (representation of a molecule involves the knowledge of the relative positions of the atoms in 3D space) and descriptor of 3D-Molecule Representation of Structures based on Electron diffraction (3D-MoRSE); for relationships between retention index of polar column (CRW-20M) and descriptor of Connectivity indices (are among the most popular topological indices), descriptor of 2D autocorrelations (are molecular descriptors which describe how a considered property is distributed along a topological molecular structure) and descriptor of 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction)] by two robust methods for the evaluation of regression parameters starting from robust coefficients of regression most popular by the appendices. We have based ourselves on comparison between the two methods, application field (DA) will be discussed using Williams diagram which presents residues of standardized prediction according to the levers values (hi) (Eriksson *et al*., 2003; Tropsha *et al*., 2003). We present the state approached graphically by Probability plot of the error and approached statistics tests (Anderson-Darling), in finished by the confidence interval of compatibility at normal law to validated results of approached state between two methods for a risk α = 5% (Nornadiah and Yah, 2011; Damodar *et al*., 2009).

Methodology

The Data Set

Molecular software Hyperchem 6.03 (AL-Noor and Asmaa, 2013) is used to represent the molecules, by employing semi-empirical method AM1 (Dewar *et al*., 1985; Holder, 1998) to obtain final geometries. The implied compounds in this study have the general structure 1.

The retention data for the114 compounds chromate graphed on stationary phases OV-101 and CRW-20M have been taken from (113 taken from Stanton and Jurs (1979) (1) and 1 compound (2-VinylPyrazine) taken from (Mihara and Enomoto, 1985) and are enumerated in Table 1.

Descriptor Generation

The optimized geometries are transferred in software dragon from data-processing software version 5.4, for calculation of 1320 descriptors while operating on 89 pyrazines of test; subsets of descriptors are chosen by genetic algorithm, these descriptors can be separate in four categories: Topological, geometrical, physical and electronic descriptors have accounts of way and molecular indices of connectivity included. The geometrical descriptors included sectors of shade, the length with the reports/ratios of width, volumes of van der Waals, the surface and principal moments of inertia. The calculated descriptors of physical property included the molecular refringency of polariz ability and molar. The electronic descriptors included most positive and most negative described by Kaliszan.

By employing the software Mobydigs (Todeschini *et al*., 2009) and by maximizing the coefficient of prédiction Q^2 and minimal R^2 of S (the error).

Table 1. Experimentally determined Retention Indices for pyrazines on OV-101 and Carbowax-20 M

n°	Compounds	ov-101	Compounds	IR(cw)
1	Pyrazine	710	Pyrazine	1179
2	Methylpyrazine	801	Methylpyrazine	1235
3	2,3-dimethylpyrazine	897	2,3-dimethylpyrazine	1309
4	2,5-dimethylpyrazine	889	2,5-dimethylpyrazine	1290
5	2,6-dimethylpyrazine	889	2,6-dimethylpyrazine	1300
6	Trimethylpyrazine	981	Trimethylpyrazine	1365
7	Trimethylpyrazine	1067	Trimethylpyrazine	1439
8	Ethylpyrazine	894	Ethylpyrazine	1300
9	2-ethyl-5-methylpyrazine	980	2-ethyl-5-methylpyrazine	1357
10	2-ethyl-6-methylpyrazine	977	2-ethyl-6-methylpyrazine	1353
11	2,5-dimethyl-3-ethylpyrazine	1059	2,5-dimethyl-3-ethylpyrazine	1400
12	2,6-dimethyl-6-ethylpyrazine	1064	2,6-dimethyl-6-ethylpyrazine	1415
13	2,3-dimethyl-5-ethylpyrazine	1066	2,3-dimethyl-5-ethylpyrazine	1421
14	2,3-diethylpyrazine	1065	2,3-diethylpyrazine	1417
15	2,3-diethyl-5-methylpyrazine	1137	2,3-diethyl-5-methylpyrazine	1459
16	Propylpyrazine	986	Propylpyrazine	1374
17	2-methyl-3-propylpyrazine	1072	2-methyl-3-propylpyrazine	1438
18	2,3-dimethyl-5-propylpyrazine	1154	2,3-dimethyl-5-propylpyrazine	1500
19	2,5-dimethyl-3-propylpyrazine	1142	2,5-dimethyl-3-propylpyrazine	1474
20	2,6-methyl-3-propylpyrazine	1151	2,6-methyl-3-propylpyrazine	1493
21	Isopropyl pyrazine	949	Isopropylpyrazine	1316
22	2,3-dimethyl-5-isopropylpyrazine	1112	2,3-dimethyl-5-isopropylpyrazine	1431
23	Butylpyrazine	1088	Butylpyrazine	1474
24	2-butyl-3-methylpyrazine	1121	2-butyl-3-methylpyrazine	1459
25	3-butyl-3,5-dimethylpyrazine	1184	3-butyl-3,5-dimethylpyrazine	1487
26	3-butyl-3,6-dimethylpyrazine	1196	3-butyl-3,6-dimethylpyrazine	1514
27	5-butyl-2,3-dimethylpyrazine	1254	5-butyl-2,3-dimethylpyrazine	1600
28	Isobutyl pyrazine	1043	Isobutylpyrazine	1406
29	2,3-dimethyl-5-isobutylpyrazine	1200	2,3-dimethyl-5-isobutylpyrazine	1525
30	2-isobutyl-3,5,6-trimethylpyrazine	1263	2-isobutyl-3,5,6-trimethylpyrazine	1556
31	sec-butylpyrazine	1040	sec-butylpyrazine	1394
32	5-sec-butyl-2,3-dimethylpyrazine	1194	5-sec-butyl-2,3-dimethylpyrazine	1500
33	Pentylpyrazine	1192	Pentylpyrazine	1575
34	2,3-dimetyl-5-pentylpyrazine	1352	2,3-dimetyl-5-pentylpyrazine	1700
35	Isopentylpyrazine	1157	Isopentylpyrazine	1530
36	2,3-dimetyl-5-isopentylpyrazine	1317	2,3-dimetyl-5-isopentylpyrazine	1655
37	(2-methylbutyl) pyrazine	1151	(2-methylbutyl) pyrazine	1527
38	2,3-dimethyl-5-(2-methylbutyl) pyrazine	1306	2,3-dimethyl-5-(2-methylbutyl) pyrazine	1636
39	2-(2-methylbutyl)-2,5,6-trimethylpyrazine	1363	2-(2-methylbutyl)-2,5,6-trimethylpyrazine	1661
40	(2-methyl-3-pentyl) pyrazine	1240	(2-methyl-3-pentyl) pyrazine	1606
41	(2-ethylpropyl) pyrazine	1121	(2-ethylpropyl) pyrazine	1449
42	(1-methylbutyl) pyrazine	1133	(1-methylbutyl) pyrazine	1471
43	2,3-demethyl-5-(2-methylpentyl) pyrazine	1377	2,3-demethyl-5- (2-methylpentyl) pyrazine	1710
44	Hexylpyrazine	1293	Hexylpyrazine	1668
45	Octylpyrazine	1495	Octylpyrazine	1845
46	2-methyl-3-octylpyrazine	1546	2-methyl-3-octylpyrazine	1956
47	2-methyl-5-(2-methylbutyl)-3-octylpyrazine	1923	2-methyl-5-(2-methylbutyl)-3-octylpyrazine	2200
48	2-methyl-6-(2-methylbutyl)-3-octylpyrazine	1962	2-methyl-6-(2-methylbutyl)-3-octylpyrazine	2264
49	Methoxypyrazine	877	Methoxypyrazine	1306
50	2-methoxy-3-methylpyrazine	954	2-methoxy-3-methylpyrazine	1339
51	2-methoxy-5-methylpyrazine	969	2-methoxy-5-methylpyrazine	1358
52	3-ethyl-2-methoxypyrazine	1037	3-ethyl-2-methoxypyrazine	1400
53	3-isopropyl-2-methoxypyrazine	1078	3-isopropyl-2-methoxypyrazine	1400
54	5-isopropyl-3-methyl-2-methoxypyrazine	1170	5-isopropyl-3-methyl-2-methoxypyrazine	1467
55	5-sec-butyl-3-methyl-2-methoxypyrazine	1250	5-sec-butyl-3-methyl-2-methoxypyrazine	1536
56	5-isobutyl-3-methyl-2-methoxypyrazine	1257	5-isobutyl-3-methyl-2-methoxypyrazine	1556
57	3-methyl-2-methoxy-5-(2-methylbutyl) pyrazine	1362	3-methyl-2-methoxy-5-(2-methylbutyl)pyrazine	1664
58	3-methyl-2-methoxy-5-(2-methylpentyl) pyrazine	1444	3-methyl-2-methoxy-5-(2-methylpentyl)pyrazine	1737
59	Ethoxypyrazine	959	Ethoxypyrazine	1348
60	2-ethoxy-3-methylpyrazine	1029	2-ethoxy-3-methylpyrazine	1385
61	2-ethoxy-5-methylpyrazine	1047	2-ethoxy-5-methylpyrazine	1418
62	2-ethoxy-3-ethylpyrazine	1101	2-ethoxy-3-ethylpyrazine	1439
63	2-ethoxy-3-isopropylpyrazine	1143	2-ethoxy-3-isopropylpyrazine	1431

Table 1. Continuo

64	2-ethoxy-5-isopropyl-3-methylpyrazine	1230	2-ethoxy-5-isopropyl-3-methylpyrazine	1500
65	2-ethoxy-5-isobutyl-3-methylpyrazine	1314	2-ethoxy-5-isobutyl-3-methylpyrazine	1584
66	5-sec-butyl-2-ethoxy-3-methylpyrazine	1306	5-sec-butyl-2-ethoxy-3-methylpyrazine	1566
67	2-ethoxy-3-methy-5-(2-methylbutyl) pyrazine	1415	2-ethoxy-3-methy-5-(2-methylbutyl) pyrazine	1693
68	(methylthio) pyrazine	1076	2-ethoxy-3-methy-5-(2-methypentyl) pyrazine	1771
69	3-methyl-2-(methylthio) pyrazine	1151	(methylthio) pyrazine	1600
70	5-methyl-2-(methylthio) pyrazine	1163	3-methyl-2-(methylthio) pyrazine	1616
71	3-ethyl-2-(methylthio) pyrazine	1237	3-ethyl-2-(methylthio) pyrazine	1695
72	3-isopropyl-2-(methylthio) pyrazine	1273	3-isopropyl-2-(methylthio) pyrazine	1692
73	3-isopropyl-3-(methylthio) pyrazine	1362	3-isopropyl-3-(methylthio) pyrazine	1737
4	5-sec-butyl-3-methyl-2-(methylthio) pyrazine	1441	5-sec-butyl-3-methyl-2-(methylthio) pyrazine	1800
75	5-isobutyl-3-methyl-2-(methylthio) pyrazine	1446	5-isobutyl-3-methyl-2-(methylthio) pyrazine	1816
76	3-methyl-5-(2-methylbutyl)-2-(methylthio) pyrazine	1552	3-methyl-5-(2-methylbutyl)-2-(methylthio) pyrazine	1941
77	3-methyl-5-(2-methylpentyl)-2-(methylthio) pyrazine	1638	3-methyl-5-(2-methylpentyl)-2-(methylthio) pyrazine	2008
78	(ethylthio) pyrazine	1148	(ethylthio) pyrazine	1635
79	2-ethylthio-3-methylpyrazine	1215	2-ethylthio-3-methylpyrazine	1655
80	2-ethylthio-5-isopropyl-3-methylpyrazine	1418		2-
hylthio-5-isopropyl-3-methylpyrazine		1769		
81	5-sec-butyl-2-ethylthio-3-methylpyrazine	1494	5-sec-butyl-2-ethylthio-3-methylpyrazine	1832
82	2-ethylthio-5-isobutyl-3-methylpyrazine	1496	2-ethylthio-5-isobutyl-3-methylpyrazine	1843
83	2-ethylthio-3-methyl-5-(2-methylbutyl) pyrazine	1602	2-ethylthio-3-methyl-5-(2-methylbutyl) pyrazine	1951
84	2-ethylthio-3-methylyl-5-(2-methylpentyl) pyrazine	1686	2-ethylthio-3-methylyl-5-(2-methylpentyl) pyrazine	2026
85	Phenoxypyrazine	1415	Phenoxypyrazine	2104
86	2-methyl-3-phenoxypyrazine	1465	2-methyl-3-phenoxypyrazine	2103
87	5-isopropyl-3-methyl-2-phenoxypyrazine	1620	5-isopropyl-3-methyl-2-phenoxypyrazine	2114
88	5-sec-butyl-3-methyl-2-phenoxypyrazine	1694	5-sec-butyl-3-methyl-2-phenoxypyrazine	2173
89	5-isobutyl-3-methyl-2-phenoxypyrazine	1706	5-isobutyl-3-methyl-2-phenoxypyrazine	2209
90	3-methyl-5-(2-methylpentyl)-2-phenoxypyrazine	1807	3-methyl-5-(2-methylpentyl)-2-phenoxypyrazine	2301
91	(phenylthio) pyrazine	1606	(phenylthio) pyrazine	2400
92	3-methyl-2-(phenylthio) pyrazine	1658	3-methyl-2-(phenylthio) pyrazine	2399
93	5-isopropyl-3-methyl-2-(phenylthio) pyrazine	1806	5-isopropyl-3-methyl-2-(phenylthio) pyrazine	2375
94	5-sec-butyl-3-methyl-2-(phenylthio) pyrazine	1874	5-sec-butyl-3-methyl-2-(phenylthio) pyrazine	2430
95	5-isobutyl-3-methyl-2-(phenylthio) pyrazine	1882	5-isobutyl-3-methyl-2-(phenylthio) pyrazine	2452
96	3-methyl-5-(2-methylbutyl)-2-(phenylthio) pyrazine	1985	3-methyl-5-(2-methylbutyl)-2-(phenylthio) pyrazine	2569
97	3-methyl-5-(2-methylpentyl)-2-(phenylthio) pyrazine	2064	3-methyl-5-(2-methylpentyl)-2-phenylthio) pyrazine	2669
98	Acetylpyrazine	993	Acetylpyrazine	1571
99	2-acetyl-3-methylpyrazine	1061	2-acetyl-3-methylpyrazine	1567
100	2-acetyl-5-methylpyrazine	1093	2-acetyl-5-methylpyrazine	1625
101	2-acetyl-6-methylpyrazine	1089	2-acetyl-6-methylpyrazine	1618
102	2-acetyl-3-ethylpyrazine	1138	2-acetyl-3-ethylpyrazine	1617
103	2-acetyl-3,5-dimethylpyrazine	1153	2-acetyl-3,5-dimethylpyrazine	1629
104	Chloropyrazine	861	Chloropyrazine	1351
105	2,3-dichloropyrazine	1032	2,3-dichloropyrazine	1581
106	2-chloro-3-methylpyrazine	951	2-chloro-3-methylpyrazine	1399
107	2-chloro-3-ethylpyrazine	1044	2-chloro-3-ethylpyrazine	1467
108	2-chloro-3-isobutylpyrazine	1187	2-chloro-3-isobutylpyrazine	1575
109	2-chloro-5-isipropyl-3-methylpyrazine	1173	2-chloro-5-isipropyl-3-methylpyrazine	1505
110	5-sec-butyl-2-chloro-3-methylpyrazine	1256	5-sec-butyl-2-chloro-3-methylpyrazine	1577
111	2-chloro-5-isobutyl-3-methylpyrazine	1264	2-chloro-5-isobutyl-3-methylpyrazine	1600
112	2-chloro-3-methyl-5-(2-methylbutyl) pyrazine	1371	2-chloro-3-methyl-5-(2-methylbutyl) pyrazine	1710
113	2-chloro-3-methyl-5-(2-methylpentyl) pyrazine	1456	2-chloro-3-methyl-5-(2-methylpentyl) pyrazine	1789
114	2-VinylPyrazine	907	2-VinylPyrazine	1392

Regression Analysis

The analysis of the multiple linear regressions was carried out with two methods by software Matlab (2009) for (Least Absolute Deviation) and Minitab (16) for (OLS).

We considers the multiple model of regression wich is given by (Berlin, 1982):

$$y_i = \beta_0 + \sum_{j=2}^{p-1} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

Detection of meaningless statements and with action leverage according to the method of least squares is a problem which is largely studied. Diagnosis by the Least Absolute Deviation regression offers alternative approaches whose principal characteristic is robustness. In our study a non-parametric method to detect the meaningless statements and point's lever is applied and compared with the traditional method of diagnosis (least squares).

Least Squares OLS Method

This is carried out with software Minitab 16, method OLS with is applied to multiple regression which consists in defining the β estimate which minimizes:

$$\sum ei^2 = \sum \left(y_i - \beta_0 - \sum x_{ij} \right)^2 \qquad (2)$$

Least Absolute Deviations (LAD) Method

The analysis of linear regression multiple is carried out with software Matlab (2009), by using the Least Absolute Deviations (LAD) method, which is one of the principal alternatives to the method of least squares when it is a question of estimating parameters of regression on, which minimizes the absolute values but not the values with square of the term of error. Least Absolute Deviation Method applied to the multiple regression consists in defining the β estimates which minimize (Dodge and Jureckova, 2000, Dodge, 2004):

$$\sum |ei| = \sum \left| y_i - \beta_0 - \sum \beta x_{ij} \right| \qquad (3)$$

Results and Discussion

An ideal model is one that has a high R value, a smallest value of standard error, starting from independent variables. The best models found has 3 descriptors for each stationary phase by using the software Moby Digs are given below.

The criterion for identifying a compound as an outlier is that compound is diministed by three or more of six standard statistical tests used to detect outliers in regression analysis. These tests were (1) residual, (2) standardized residual, (3) Studentized residual, (4) leverage, (5) DFFITS, (6) Cook's distance. The residual is the difference between real value and the value predicted by the regression equation. The standardized residual is the residual divided by difference models of regression equation. The Studentized residual is the residual of forecast divided by proper model difference.

Leverage allows for the determination of a point the influence.

DFFITS describes difference in the fits of the equation caused by displacement of a given observation and Cook's distance describes the change of a model coefficient by the displacement of indicated point.

The definition of each descriptor is given Table 2.

The coefficient of multiple determinations (R^2) indicates the amount of variance in data is a explained by the model. The standard error of regression coefficient is given in each case and n indicates of molecules involved in regression analysis procedure.

The Best Models

IR (OV-101) : (XMOD, FDI, Mor 06 v); S = 18.379, R^2 = 99.4, n = 89 compounds
IR (CRW20M) : (RDCHI, GATS1p, Mor 02 m); S = 34.933, R^2 = 98.08, n = 89 compounds

The best tree parametric model was constructed using:

[OV-101: Modified Randi connectivity index (XMOD) (is a molecular descriptor proposed as the sum of atomic properties, accounting for valence electrons and extended connectivities in the H-depleted molecular graph using a Randic connectivity index-type formula), Folding Degree Index (FDI) (is the largest eigenvalue of the distance/distance matrix, normalised dividing it by the number of atoms nAT. This index tends to one for linear molecules (of infinite length) and decreases in correspondence with the folding of the molecule. Thus, it can be thought of as a measure of the folding degree of the molecule because it indicates the degree of departure of a molecule from strict linearity) and (Mor06v) (3D-MORSE-signal 06/weighted by atomic Vander Waals volumes (Mor06v) (3D-MoRSE) (3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves.3D-MoRSE the descriptors are calculated for five different atomic properties *w*: the unweighted case (u), atomic mass (m), the van der Waals volume (v), the Sanderson atomic electro negativity (e) and, the atomic polarizability (p). (CRW-20M: Reciprocal Distance Randi-type Index (RDCHI) (is defined on the analogy of the Randic connectivity index X1, where the vertex degrees are substituted by the row sums of the reciprocal distance matrix. Moreover, the reciprocal distance squared Randictype-index RDSQ is obtained from the RDCHI index substituting the exponent-1/2 with 1/2.), Geary Autocorrelation -log 1/weighted by atomic polariz abilities (GATS1p) (2D autocorrelations calculated by DRAGON are spatial autocorrelations calculated on a H-depleted molecular graph weighted by atom physico-chemical properties (i.e., the atom weightings *w*) and include: Autocorrelations GATS calculated by the Geary coefficient) and 3D-MORSE-signal 02/weighted by atomic masses (Mor02m)].

Table 2. Definitions of descriptors used in the retention index prediction models

Descriptors	The definition
XMOD	Modified Randi connectivity index
FDI	folding degree index
Mor06v	(3D-MORSE-signal 06/weighted by atomic Vander Waals volumes
RDCHI	reciprocal distance Randi-type index
GATS1p	Geary autocorrelation -log 1/weighted by atomic polarizabilities
Mor02m	3D-MORSE-signal 02/weighted by atomic masses

Using a significance level of 0.05, the Anderson-Darling normality test (Fig. 1) (A-Squared = 0,134; OV-101, A-Squared = 0,270; Crbowax- 20 $M < v_{cri} = 0.752$) indicates that the resting pulse data follow a normal distribution But it disturbance that if outliers may be present in the measurements.

Auto Correlation of the Residus

Values of the statistics of Durbin-Watson (Durbin and Watson, 1951), [d = 1,47910; OV-101/D = 1,29968; Carbowax-20M] are the greater than higher values given by the tables, respectively for 3 regresses and for reasonable risk α = 0.05, which expresses positive auto correlation of residues which establishes each time the independence of the residues include the absence of autocorrelation that if outliers may be present in the measurements.

Column RCW -20 M

Column OV -101

The diagnostic statistics joined together in Table 3 make it possible to make comparisons and to draw several conclusions.

All relevant statistical parameters are reported in Table 3.

Values of $R^2$and R^2_{adj} attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter F).

The model is robust, the difference between R^2 and Q^2 is small (0.05% of Colum OV-101 and 0.22% of Colum CRW-20M). The model demonstrates a very good stability in internal validation while bootstrapping confirms the internal (Q^2bOO) predictivity and stability of the model. SDE Pext is a little bit different from SDEP. The model works slightly worse in external prediction than in internal prediction.

Correlation Matrix between Retention Indices and the Selected Descriptors

Column OV-101

ov-101	XMOD	FDI	
XMOD	0,986		
	0,000		
FDI	-0,039	-0,152	
	0,715	0,154	
Mor06v	0,181	0,059	0,274
	0,089	0,582	0,009

Column CRW-20M:

	IR (cw)	RDCHI	GATS1p
RDCHI	0,893		
	0,000		
GATS1p	-0,375	0,044	
	0,000	0,681	
Mor02m	0,896	0,930	-0,024
	0,000	0,000	0,821

The matrix of correlation Table 4, obtained using the order Correlation of software MINITAB, shows that the descriptors are more or less correlated between them (r≥0,39 for a p = 0,045<α = 0.05).

All the descriptors respectively are correlated with the retention index of the CRW-20M phase except the GATS1p descriptor is correlated less and with the retention index of phase OV -101 descriptor (XMOD) is correlated and the Descriptors (FDI, Mor06v) less correlated.

The Least Squares method of estimation of parameters of linear (regression) models performs well provided that the residuals are well not behaved. However, models with the disturbances that are prominently non-normally distributed or follow a normal distribution But it disturbance and contain sizeable outliers fail estimation by the Least Squares method. An intensive research has established that in such cases estimation by the Least Absolute Deviation (LAD) method performs well.

Multiple linear Regression Comparison Robust Regression of OLS and Least Absolute Deviation

We will try More particularly 2 estimate methods for the vector $((\beta_0^*,\beta_1^*,...,\beta_k^*)$ of Parameters:

- Method of ordinary least squares, the most known and the most used.
- The method Least Absolute Deviation (LAD) (Sum of the absolute values of the errors) (Machabert, 2014).

Table 3. Statistics diagnostic for the selected models

Colum	Models	R^2	Q^2	Q^2boot	Q^2ext	R^2adj	Kx
OV-101	X1sol Mor06v AMR	99,44	99,39	99,35	97,5	99,42	51,36
			Kxy	SDEP	SDEC	F	s
		65,5	18,736	17,961	4987,6	18,38	
		R2	Q2	Q2boot	Q2ext	R2adj	Kx
CRW-20M	RDCHI GATS1p Mor02m	98,08	97,86	97,72	77,02	98,01	46,61
		Kxy	SDEP	SDEC	F	s	
		63,91	36,044	34,139	1444,5	34,93	

Table 4. Least absolute deviation estimates for model

Predictor	Coef	SE Coef	T	P
Constant	-946	100,237	-9,44	0,000
XMOD	29,1	5,216	5,58	0,000
FDI	1174.4	65,36	17,97	0,000
Mor06v	70,4	10,909	6,453	0,000

Table 5. Least squaresestimates for model

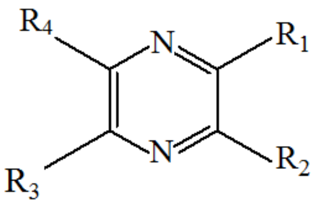
Predictor	Coef	SE Coef	T	P
Constant	-809,4	107,2	-7,55	0,000
XMOD	292,454	0,2535	115,35	0,000
FDI	1028,3	108,5	9,48	0,000
Mor06v	70,453	6,266	11,24	0,000

Table 6. Least absolute deviation estimates for model

Predictor	Coef	SE Coef	T	P
Constant	859,72	94,47	9,10	0,000
RDCHI	527,46	44,679	11,805	0,000
GATS1p	-630,74	20,68	-30,5	0,000
Mor02m	28.36	19,582	1,45	0,000

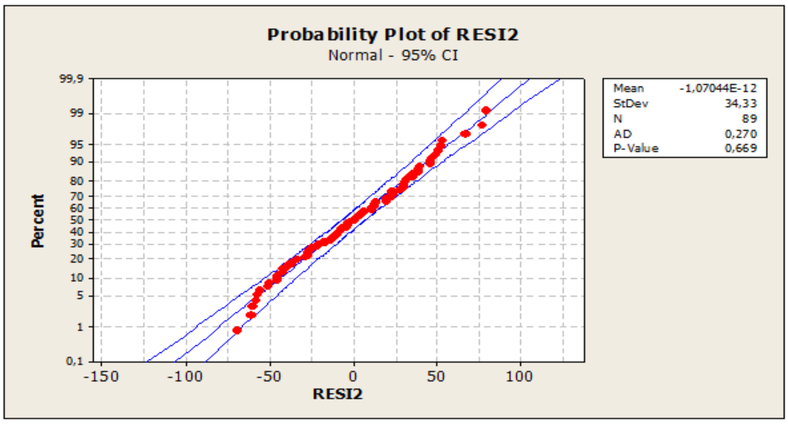
Table 7. Least squares estimates for mode

predictor Coef SE Coef T P				
Constant	852,37	44,50	19,15	0,000
RDCHI	512,52	33,40	15,34	0,000
GATS1p	-636,05	24,61	-25,85	0,000
Mor02m	32,671	4,612	7,08	0,000

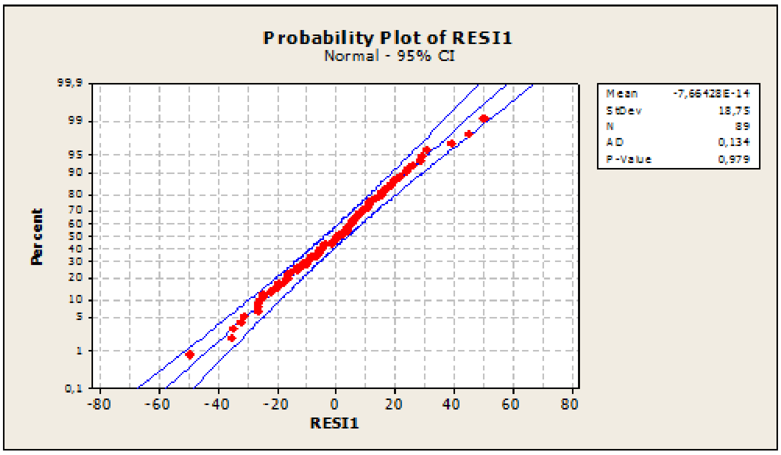


R1: H, alkyl, alkoxy, alkylthio, aryloxy, arylthio, acetyl, chloro.
R2: H, alkyl, chloro, vinyl.
R3: H, alkyl.
R4: H, alkyl.

Fig. 1. Structure of pyrazine



(a)



(b)

Fig. 2. Diagram of percentage of normality's of the residues

The advantage large of the Least Absolute Deviation (LAD) method is robustness, i.e., that the estimators are not impact by the extreme values, (they are known as "robust"). It is thus particularly interesting to use the method Least Absolute Deviation LAD if one is in the presence of aberrant values in comparison with Least Squares (OLS) method.

Comparison of Hyperplanes of Regression

The model has been estimated by first by Least Squares (OLS,) and then by Least Absolute Deviation, Running the least squares and Least Absolute Deviation regression yields the estimates given in Table.

Column OV-101

Column CRW -20M

All the variables for the two models is strongly statistically significant in the two columns with method least squares and the method Least Absolute Deviation (Table 4-7).

We noticed that calculated of β least squares are not very different for the regression with β the Least Absolute Deviation on the two columns, except, calculated.

β_1 and β_3 least squares is almost the same ones as for the regression with β_1 and β_3 Least Absolute Deviation on column OV-101 (Table 4-7).

Thus it is relevant to remake a verification in presences of aberrant values using the following phases (Fig. 3):

Hyper plane of regression can radically vary with the change of hyper plane coefficients.

Graphical Comparisons of Alternative Regression Models

The application field has been discussed with the help of Williams diagram.

Column CRW-20M

Column OV-101

The analysis of the residues shows that the observations (82 68 14 1) raised residues in the two estimates and the observations (72, 2) raised residue with the Least Absolute Deviation estimate and lever by least square also observation (2, 4) raised residue and influential observations in the two estimates in the whole of validation on column OV -101 and column CRW -20 M the observations (1, 7, 85) raised residues in the two estimates, the observation (86) raised residues with the Least Absolute Deviation estimate and lever by least square also observation (2,3) raised residues and influential observations with Least Absolute Deviation but it with the least squares estimate the observation (2) influential observation butthe observation (3) lever whole of validation.

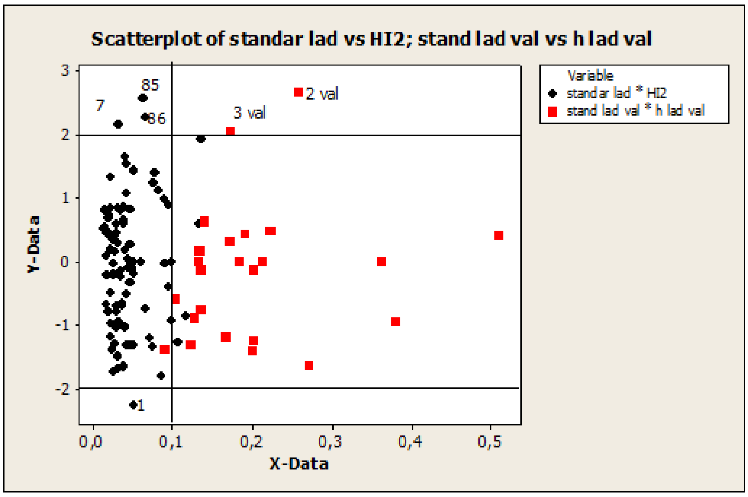
After elimination of the collective aberrant points between the two methods and after the secondary treatment one has the observation (83) raised residues in the two estimates also the observation 2influential observation in the whole of validation in the two estimates on column CRW -20 M and on column OV -101 the observations (1,69) raised residues in the two estimates and the observation 81 the observations raised residues in the least squares estimate also observation (2) influential observation in the least squares estimate.

Thus finally the models in which the meaningless statements were removed become:

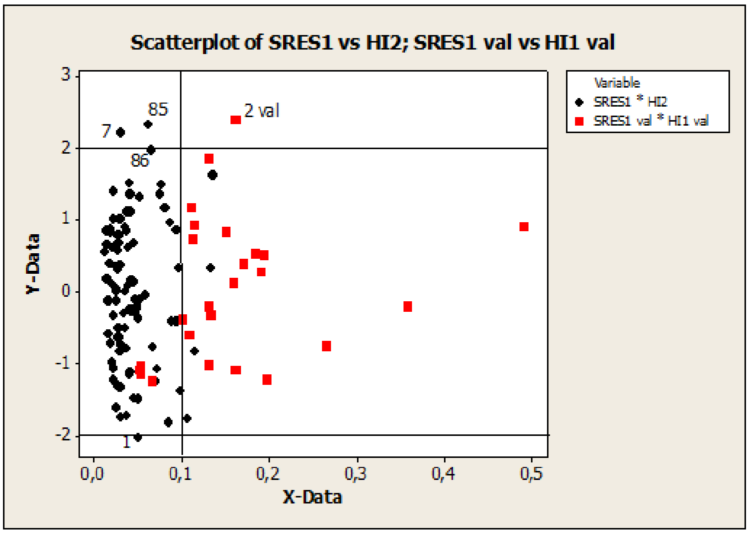
Column OV-101

Least Absolute Deviation:

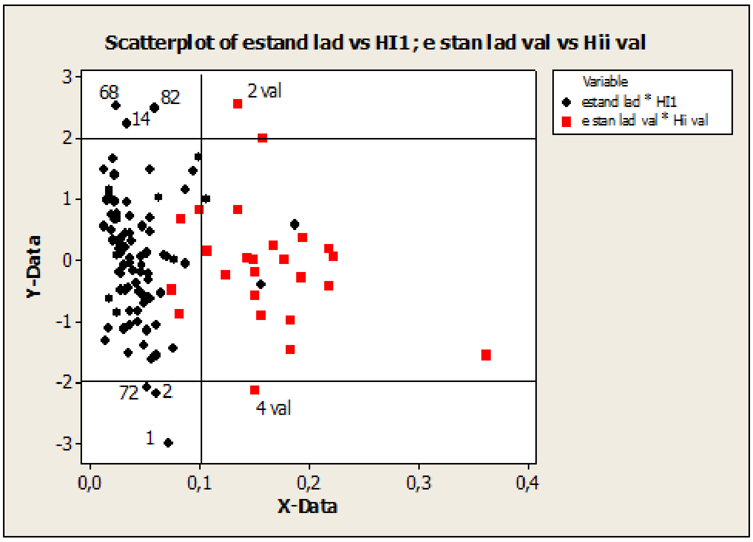
$$y = -946 + 29.1\,XMOD + 1174.4\,FDI + 70.4\,Mor06v \qquad (4)$$


Scatterplot of standar lad vs HI2; stand lad val vs h lad val
Variable
◆ standar lad * HI2
■ stand lad val * h lad val
85
7
86
2 val
3 val
1
Y-Data
X-Data
3
2
1
0
-1
-2
0,0
0,1
0,2
0,3
0,4
0,5

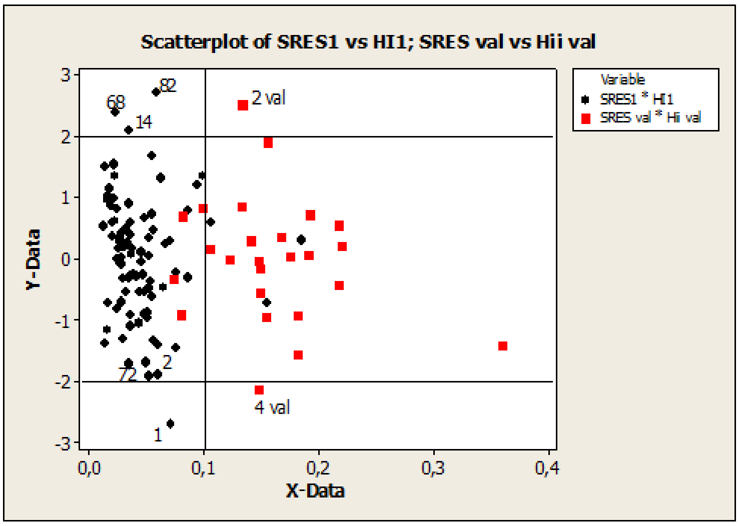
(a)


Scatterplot of SRES1 vs HI2; SRES1 val vs HI1 val
Variable
◆ SRES1 * HI2
■ SRES1 val * HI1 val
85
7
86
2 val
1
Y-Data
X-Data
3
2
1
0
-1
-2
0,0
0,1
0,2
0,3
0,4
0,5

(b)


Scatterplot of estand lad vs HI1; e stan lad val vs Hii val
Variable
◆ estand lad * HI1
■ e stan lad val * Hi val
68
82
14
2 val
72
2
4 val
1
Y-Data
X-Data
3
2
1
0
-1
-2
-3
0,0
0,1
0,2
0,3
0,4

(c)



(d)

Fig. 3. Diagram of Williams of the residues of prediction standardized according to the lever (a, c) Least absolute deviation method (Training, Test); (b, d) Least squares method (Training, Test)

Least Squares:

$$y = -886 + 29{,}1\ XMOD + 1115\ FDI + 70.9\ Mor06v \qquad (5)$$

Column CW -20M

Least Absolute Deviation:

$$y = -859{,}72 + 527.46\ RDCHI \\ -630.74\ GATS1p + 28.37\ Mor02m \qquad (6)$$

Least Square:

$$y = 842{,}527\ RDCHI - 625\ GATS1p + 29{,}2\ Mor02m \qquad (7)$$

We noticed besides that calculated β can approach that regression with β Least Absolute Deviation on the two columns into precise calculated (β_1 and β_3) least squares are almost the same ones as for regression with (β_1 and β_3) Least Absolute Deviation and on the order same with (β_0 and β_2) on OV 101 and calculated β_1 least squares are almost the same ones as for regression with β_1 Least Absolute Deviation on CRW -20 M and on the order same with (β_1, β_3 and β_4).

The analysis of the residues shows that in this case All the observation of Least Absolute Deviation method between (-2, 2), but it the analysis of the residues of least squares method shows that the observations [OV-101: Training - test (2), CRW-20 M: Training- (46)] the Least Absolute Deviation estimate given good result On the other hand estimate least squares Fig. 4:

Graphical Comparisons of Alternative Regression Models

Column CRW-20M

Column OV-101

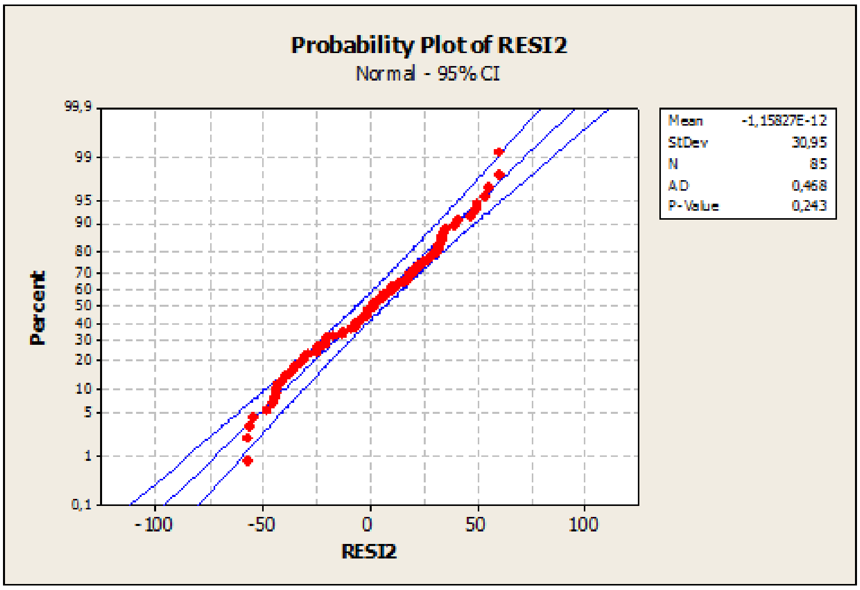
We notice no change of the coefficients of the right-hand side after feeding of the aberrant point what translates the line is stable which expresses that the Least Absolute Deviation method born not sensitive to the presences of the aberrant values thus we report that the Least Absolute Deviation method is a stable method and more robust.

To conform the approach between the two methods and to deduce the robust method between them, There is a set of tests of normality (of standard errors or residues…) indeed, thanks to robustness concept, we can used simple techniques (descriptive e.g. Statistics, technical graphs) to check if the distribution of data is really approximate.

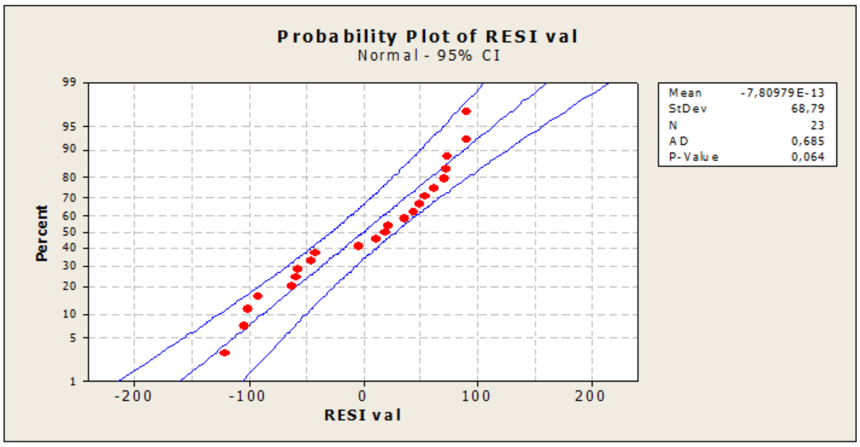
Any test is associated *a risk* known as of first species years works us, we will adopt it risk $\alpha = 5\%$.

Comparisons of the Tests of Normality of the Errors between Method Least Absolute Deviation and Least Squares in Approached State

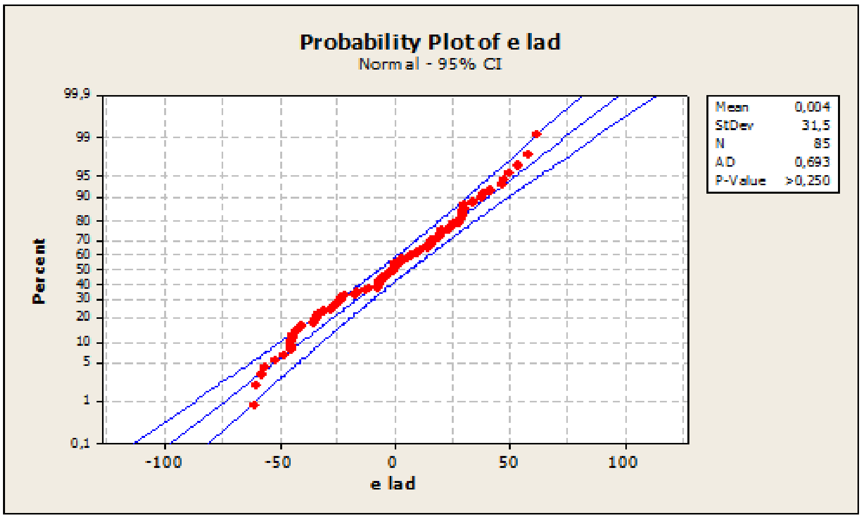
Software Minitab 16 proceeds automatically in estimating two principal parameters of the normal law (μ the Mean (OV-101:0, CRW-20M: 0), σ the variation-type (OV-101:10.35, CRW-20M:14.84) for least squares one applying the same principle with the Least Absolute Deviation method but one used (the median (OV-101: -1.57, CRW-20M:0.01) σ variation-type (OV-101:10.26, CRW-20M:15.08) and with the principal number in the state approached to the two columns (OV-101: n = 83, CRW-20 M: n = 85).


Probability Plot of RESI2
Normal - 95% CI
Percent
99,9
99
95
90
80
70
60
50
40
30
20
10
5
1
0,1
-100
-50
0
50
100
RESI2
Mean -1,15827E-12
StDev 30,95
N 85
AD 0,468
P-Value 0,243

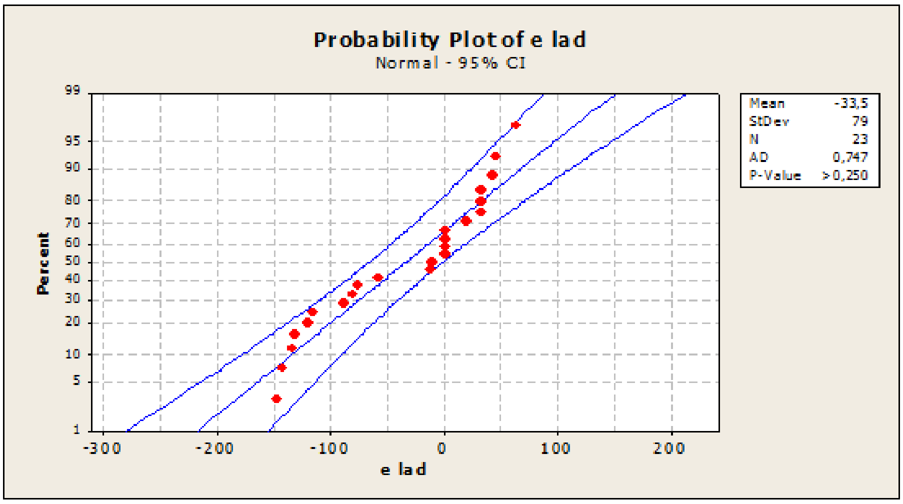
(a)


Probability Plot of RESI val
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-200
-100
0
100
200
RESI val
Mean -7,80979E-13
StDev 68,79
N 23
AD 0,685
P-Value 0,064

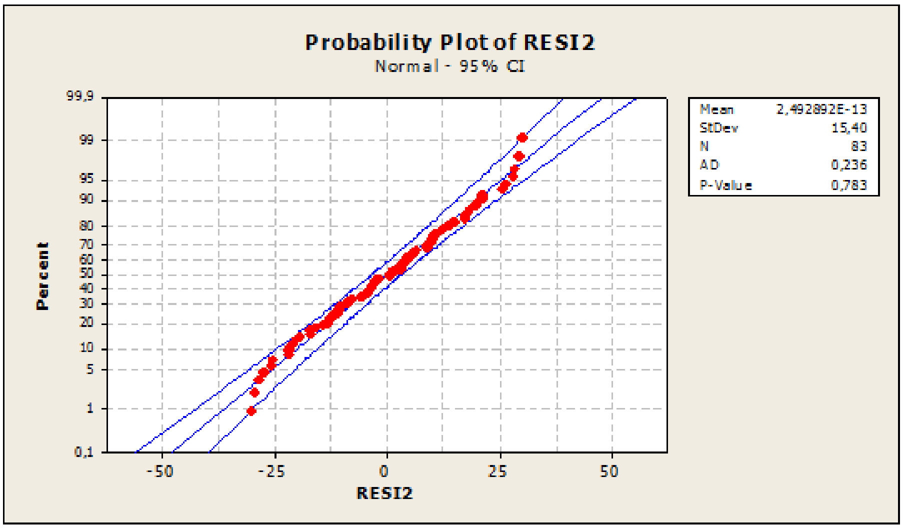
(b)


Probability Plot of e lad
Normal - 95% CI
Percent
99,9
99
95
90
80
70
60
50
40
30
20
10
5
1
0,1
-100
-50
0
50
100
e lad
Mean 0,004
StDev 31,5
N 85
AD 0,693
P-Value >0,250

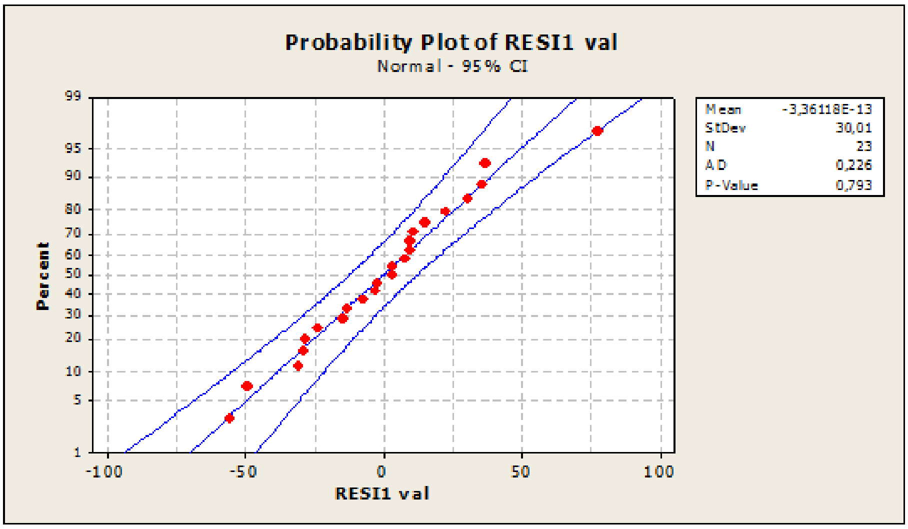
(c)


Probability Plot of e lad
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-300
-200
-100
0
100
200
e lad
Mean -33,5
StDev 79
N 23
AD 0,747
P-Value >0,250

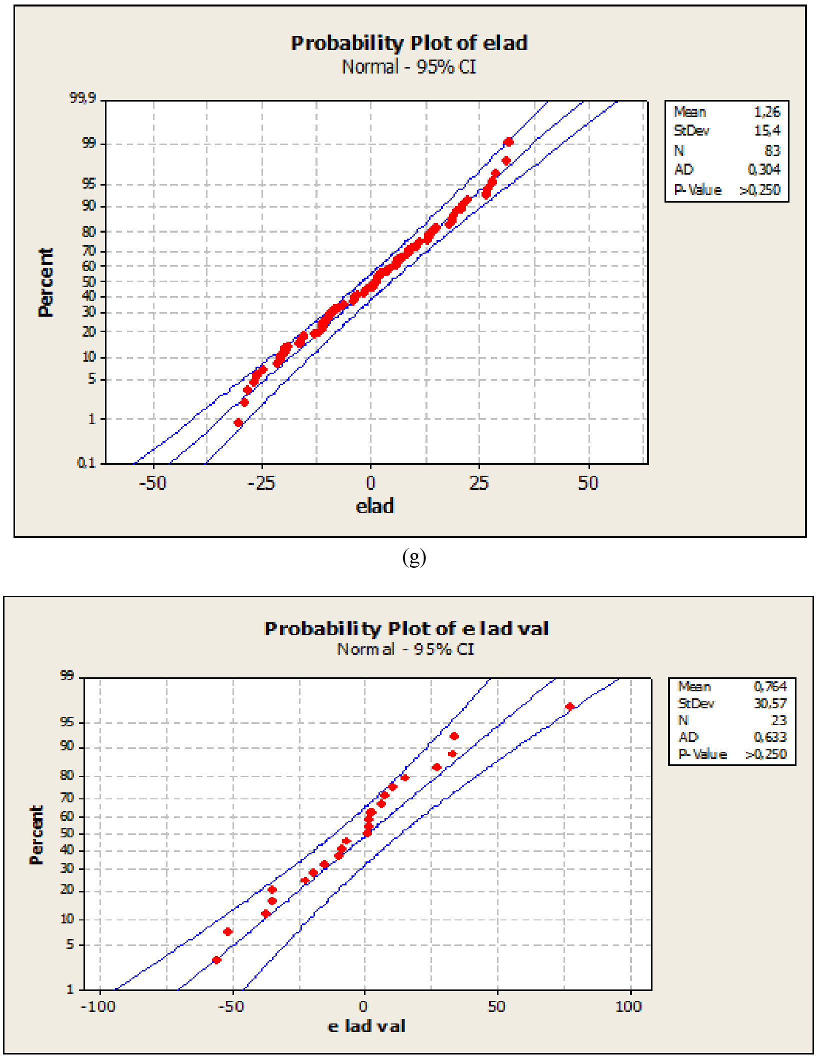
(d)


Probability Plot of RESI2
Normal - 95% CI
Percent
99,9
99
95
90
80
70
60
50
40
30
20
10
5
1
0,1
-50
-25
0
25
50
RESI2
Mean 2,492892E-13
StDev 15,40
N 83
AD 0,236
P-Value 0,783

(e)


Probability Plot of RESI1 val
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-100
-50
0
50
100
RESI1 val
Mean -3,36118E-13
StDev 30,01
N 23
AD 0,226
P-Value 0,793

(f)



(g)

(h)

Fig. 4. Diagram of normality percentage of residues (Training, Test) (a, c, e, g) Training; (b, d, f ,h) Test

Graphic Tests

Probability Plot of Error

To check normality of errors of a model of regression is to carry out Probability stud of residues.

Column CW -20M

Least Squares Method

Least Absolute Deviation Method

Column OV-101
Least Squares Method

Least Absolute Deviation Method

A normal distribution with the two columns appears to fit your data sample fairly well.

The plotted points form a reasonably straight line.

Test of Anderson-Darling

In our work, one finds us that Anderson-Darling (AD) [OV- 101: (Least Absolute Deviation) = 0.364 with value of p>0.250, (least squares) = 0.236 with value of p = 0.783, n = 83], [CRW-20M: (Least Absolute Deviation) Anderson-Darling (AD) = 0,693 with value of p>0.250, (least squares) = 0,468 with value of p = 0.243 n = 85] < AD critique = 0.752 with p>0.1 to 5%, the assumption of normality is compatible with our data with Least Absolute Deviation method and least squares.

Interval of Confidence

The interval confidence and the risqe *a* constitute a complementary approach thus (an estimate approach) the most used interval confidence is interval confidence has 100(1-*a*) = 95%.

The Column OV-101:

Training : Least Absolute Deviation: (-31.52, 29), least squares (-30.18, 30.18)
Test : Least Absolute Deviation (-59.15, 60.68), least squares (-58.82, 58.82)

The Column CRW-20M:

Training : Least Absolute Deviation: (-61.73, 61.74), least squares (-60.66, 60.66)
Test : Least Absolute Deviation (-135.9, 135.8), least squares (-136.6, 136.6)

The data may be compatible with the hypothesis also that the limited values of the interval are center which expresses the mean and the median which verifies position 95% that the 50th percentile for the population the center of the acceptance zone the null hypothesis.

Completely all the graphic and statistical tests is accepted data of the approached state between the two methods especially the test of Anderson-Darling the value of the Least Absolute Deviation method closer to least squares method and Interval of The value of confidence these result is formed L approximate of two method.

Conclusion

PYRAZINes are compounds naturally presents in food and taking part in their odour, contray to their biodegradation, pyrazine formation has been intensively studied.

Modeling of retention indices of 114 pyrazines (89 Training and 25 Test) eluted out of two columns various OV -101, the best tree parametric model was constructed using.

[OV-101 with Modified Randi connectivity index (XMOD), Folding Degree Index (FDI) and (3D-MORSE-signal 06/weighted by atomic Vander Waals volumes (Mor06v); CRW-20M with Reciprocal distance Randi-type Index (RDCHI), Geary autocorrelation -log 1/weighted by atomic polariz abilities (GATS1p) and 3D-MORSE-signal 02/weighted by atomic masses (Mor 02 m)].

The Column of OV-101 and CRW-20M by two methods Least Absolute Deviation and least squares are based on the following comparisons.

The comparison of the equations of the hyper planes:

L'equations of least squares is closer to Least Absolute Deviation after elimination of the aberrant points for the β_2 (Least Absolute Deviation) $\cong \beta_2$ (least squares) and the other coefficient remaining with the same order for column OV-101 for the column CRW-20 M the β_1 (Least Absolute Deviation) $\cong \beta_1$ (least squares) and the other coefficient remaining with the same order after the secondary treatments for the checking of presence of aberrant values (training: 1, 2, 14, 68, 72, 82 test: 2, 4) (training: 1, 7, 85, 86, test: 2, 3) on column (OV -101) and (training: 1, 7, 85, 86, test: 2, 3) for the CRW-20M- column) and to be able to compare them By using the following stage.

Graphic comparison: The applicability is discussed using the diagram of Williams in dependence.

Lastly, it is noted that Least Absolute Deviation is a robust estimator not sensitive to the presences of the aberrant values thus we report that the Least Absolute Deviation method is a stable and robust method.

Used test of normality's of the errors by graphic and statistical test. One applied compatibility with the normal law, but using the degree $\alpha = 0.05$. Too one confirmed approached graphically by Probability plot of the error One notes that the test to accept the assumption of normality is that of Anderson-Darling, in finished by the confidence interval with one p-been worth sup 0.1 on the columns.

It general this study is shown that results by the two estimates theoretical (equation) and graph give good results expressed by the models.

Acknowledgement

We would like to thank Salima Khanouch, PhD (informatique) for writing the algorithm of least absolut deviation Method and Mr hessene (Teacher of English Language) for reviewing the language in the manuscript.

Author's Contributions

Fatiha Mebarki: Good Developed methods of least absolut deviation and least squares, Developed deference's Softwares (Matlab, Minitab, Tanagra, genetic Algorithm)and participated in all experiments, coordinated the data-analysis.

Khadija Amirat: Developed deference's Softwares (Matlab, Minitab, Tropsha, SVM, genetic Algorithm) and participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Salima Ali Mokhnach: Correction of the work and chef laboratories.

Djellol Messadi: Is the main researcher of project and chef of tree laboratories, designed the research plan and organized the study.

Ethics

This article is original and to the best knowledge of the authors has not been published before. The authors confirm that there are no ethical issues involved.

References

Berlin, G.B., 1982. The Pyrazine. 1st Edn., J. Wiley, New York, ISBN-10: 0471381195, pp: 687.

Buchbauer, G., 2000. Threshold-based structure-activity relationships of pyrazines with bell-pepper Flavor. J. Agric. Food Chem., 48: 4273-4278. PMID: 10995349

Damodar, N.G. and C.D. Porter, 2009. Basic Econometrics. 5st Edn., McGraw-Hill Irwin,, Boston, ISBN-10: 0071276254, pp: 922.

Dewar, M.J.S., E.G. Zoebisch, E.F. Ealy and J.J.P. Stewart, 1985. AM1: A new general purpose quantum mechanical model. J. Am. Chem. Soc., 107: 3902-3909.

Dodge, Y. and J. Jureckova, 2000. Adaptive Regression. 1st Edn., Springer Science and Business Media, New York, ISBN-10: 1441987665, pp: 177.

Dodge, Y., 1987. Statistical Data Analysis Based on the Li-Norm and Related Methods. 1st Edn., North-Holland, Amsterdam, ISBN-10: 0444702733, pp: 464.

Dodge, Y., 1997. L1-Statistical Procedures and Related Topics. 1st Edn., Institute of Mathematical Statistics, Hayward, ISBN-10: 0940600439, pp: 498

Dodge, Y., 2004. Statistique: Dictionnaire Encyclopédique. 1st Edn., Springer Science and Business Media, Paris, ISBN-10: 2287720944, pp: 662.

Dragon 5.4, http://www.disat.unimib.it

Eriksson, L., J. Jaworska, A. Worth, M. Cronin and R.M. Mc Dowell *et al*., 2003. Methods for reliability, uncertainty assessment and applicability evaluations of regression based and classification QSARs. Environ. Health Perspect., 111: 1361-1375.

Gonin, R. and A.H. Money, 1989. Linear L_P-norm Extimation. 1st Edn., Marcel Dekker, New York.

Holder, A.J., 1998. AM1, Encyclopedia of Computational Chemistry. Scheleyer, P.V.R., N.L. Allinger, T. Clarck, J. Gasteiger and P.A. Kollman *et al*. (Eds.), Wiley, Chichester, pp: 1-8.

Hyperchem 6.03, (Hypercube), http://www.hyper.com.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee, 1985. The Theory and Practice of Econometrics. 2nd Edn, Wiley, New York, ISBN-10: 047189530X, pp: 1019

Li, W., C.L. Heth and S.C. Rasmussen, 2014. Thieno[3,4-b] pyrazine-based oligothiophenes: Simple models of donor-acceptor polymeric materials. Phy. Chem. Chemicale Phys. J., 28 : 7231-40.

Machabert, T., 2014. Modèles en très grande dimension avec des outliers. Théorie, Simulations, Applications Paris.

Matlab, R., 2009. Minitab, release 16.1, statistical software, 2003.

Mebarki, F., K. Amirat, S.A. Mokhnache and D. Messadi, 2016. Treatment by alternative methods of regression gas chromathographic retention indices of 35 pyrazines. Int. J. Instrument. Control Syst., 6: 1-14.

Mihara, S. and N. Enomoto, 1985. Calculation of retention indices of pyrazines on the basis of molecular structure. J. Chromatogr., 324: 428-430.

Moby Digs 1.1, http://www.disat.unimib.it

Nornadiah, M.R. and Y.B. Yah, 2011. Power Comparaisons of shapiro-wilk, Kolmogorov-smornov, lillieffors and Anderson-Darling tests. J. Statistique Modell. Analyt., 2: 21-33.

Small, G.W. and P.C. Jurs, 1983. Interactive computer system for the simulation of carbon-13 nuclear magnetic resonance spectra. Anal. Chem., 55: 1121-1127. DOI: 10.1021/ac00258a033

Stanton, D.T. and P.C. Jurs, 1989. Computer-assisted predict of gaschromatographicretention indexes of pyrazines. Anal. Chem., 61: 1328-1332.

Todeschini, R., D. Ballabio, V. Consonni, A. Mauri and V. Pavan, 2009. MobyDigs 1.1, Copyright TALETE srl.

Tropsha, A., P. Gramatica and V.K. Grombar, 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Combi. Sci., 22: 69-76.