Special Issue Article

Journal of CHEMOMETRICS

Received: 18 February 2010, Revised: 20 May 2010, Accepted: 2 June 2010, Published online 21 October 2010 in Wiley Online Library: 2010

(wileyonlinelibrary.com) DOI: 10.1002/cem.1331

Is your QSAR/QSPR descriptor real or trash?

Rudolf Kiralj[a] and Márcia M. C. Ferreira[a]*

The sign change problem in quantitative structure–activity relationship (QSAR), quantitative structure–property relationship (QSPR) and related studies is the controversy related to the signs of correlation coefficients and regression coefficients of a descriptor in univariate and multivariate regressions, before and after the data split. Among 50 investigated regression models with 227 descriptors extracted from the literature, the sign change problem was shown to have a very high frequency, according to four new criteria proposed in this work for its assessment. The sign change problem can be substantially reduced and even eliminated for a given dataset by statistically based variable selection and by checking for the sign change problem before model validation and interpretation. Knowing the fundamentals of statistics related to the sign change problem, its identification and understanding aid in finding effective means to remedy regression models with this deficiency. Copyright © 2010 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article

Keywords: correlation; univariate regression; multivariate regression; descriptors

1. INTRODUCTION

Multivariate regression models have found fertile ground in chemistry and related sciences, especially within the areas of quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR). As recently noticed by Dearden *et al.* [1], many thousands of works in these areas have been developed, and several publications dealing with guidelines for obtaining statistically correct and satisfactory regression models have been published, as for example the OECD (Organisation for Economic and Cooperation and Development) principles [2]. However, there are still numerous models incorporating various errors, which are classified into 21 types by Dearden *et al.* [1].

Errors in QSAR/QSPR development and use appear and propagate as a consequence of failures to perform certain operations during regression analysis. These failures become rather apparent when basic principles of linear regression analysis are discussed, such as those one can find in the basic statistical literature [3–11] and QSAR/QSPR [1,2,12–20] publications. Twelve concepts related to various types of failures will be discussed in this work, using the following nomenclature.

A single univariate regression for n samples (objects) is a quantitative relationship of the observed dependent variable $\mathbf{y}$ with the independent variable $\mathbf{x}$ (descriptor),

$$\hat{\mathbf{y}} = a + b\mathbf{x} \tag{1}$$

where $\hat{\mathbf{y}}$ is the predicted property, a is the intercept and b is the regression coefficient of $\mathbf{x}$. When there are m independent variables, an analogous expression for each general $\mathbf{x}_j$ variable is

$$\hat{\mathbf{y}} = a_j + b_j\mathbf{x}_j \quad j = 1, 2, \ldots m \tag{2}$$

while the contribution of all variables treated simultaneously is expressed via a general Multiple Linear Regression (MLR) or Partial Least Squares regression (PLS) equation with α as the intercept and β_j as the regression coefficients of $\mathbf{x}_j$:

$$\hat{\mathbf{y}} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_m\mathbf{x}_m \tag{3}$$

Researcher's key role in regression analysis: Typical problems found here are lack of sufficient statistical, chemical or QSAR/QSPR expertise; insufficient human interference into the regression analysis by relying too much on automatic and black-box procedures and consideration of graphical means (variable distribution plots, bivariate distribution plots, residuals plots, etc.) as not necessary.

Randomness: This is an intrinsic characteristic of linear regression, present in various ways. First, the samples (substances) should be randomly distributed in the space of features $\mathbf{x}_j$ and $\mathbf{y}$ for the studied class or classes of compounds, which may be a practical problem when having a small set of similar samples. Second, samples in the data matrix should be ordered so that any systematic variation in variables in the dataset or subsets is avoided, which may be achieved by randomization of the original matrix. Third, data split, especially for external validation, must be based to a certain extent on randomness. Fourth, residuals $\mathbf{y} - \hat{\mathbf{y}}$ must be randomly distributed with respect to any $\mathbf{x}_j$, $\mathbf{y}$ and $\hat{\mathbf{y}}$.

Diversity: Molecular diversity of the dataset accounts for the variation in 2D and 3D structures, and it should provide

* *Correspondence to: M. M. C. Ferreira, Laboratory of Theoretical and Applied Chemometrics, Institute of Chemistry, University of Campinas, Campinas SP 13083-740, Brazil.*
E-mail: marcia@iqm.unicamp.br

a *R. Kiralj, M. M. C. Ferreira*
Laboratory of Theoretical and Applied Chemometrics, Institute of Chemistry, University of Campinas, Campinas SP 13083-740, Brazil

significant variations in all variables, where each variable may be characterized by a certain type of distribution. The number of samples is usually crucial to enable molecular diversity, satisfactory statistical significance achieved in various tests and effective data split. Finally, diversity of descriptors in terms of their nature (steric, electronic, hydrogen bonding, lipophilic, etc.) may aid in describing the complex nature of molecular behavior in QSAR/QSPR.

Normality, independence and linearity conditions: All variables $\mathbf{x}_j$ and $\mathbf{y}$, $\hat{\mathbf{y}}$, and residuals have normal or quasi-normal distribution in the theory of regression analysis. Furthermore, regression coefficients and their errors, correlation coefficients and other statistical parameters also have normal or quasi-normal distribution when several closely related models are compared, as for example in resampling-based model validations. Independence is another important cornerstone of regression analysis. For example, residuals must be independent of variables $\mathbf{x}_j$ and $\hat{\mathbf{y}}$. The so-called independent variables $\mathbf{x}_j$ are considered mutually independent, i.e. one variable does not vary systematically with another and their joint distribution may be other than normal or quasi-normal. Linearity is an additional condition for linear regression, closely related to normality and independence. In this sense, $\mathbf{y}$ linearly depends on the independent variables, $\hat{\mathbf{y}}$ is a linear function of $\mathbf{x}_j$ and regression coefficients are estimated by a linear method. Furthermore, any two quantities (variables $\mathbf{x}_j$, $\mathbf{y}$, $\hat{\mathbf{y}}$, $\mathbf{y}-\hat{\mathbf{y}}$ and statistical parameters) show bivariate normal or quasi-normal distribution in scatterplots, visible as data points situated on concentric ellipses whose major axes lay on the regression line. This line shows the degree and direction of systematic linear variation of one variable with the change in the other, known as correlation. Independence for linear regression means, therefore, that there is no correlation. Normality, independence and linearity concepts, although seeming too banal to be mentioned, are not so rarely misunderstood, misinterpreted and even ignored. Real datasets bring certain deviations from basic assumptions of linear regression. At first, when dealing with a small number of samples, normality conditions for individual and joint $\mathbf{x}_j$–$\mathbf{y}$ distributions are not achieved satisfactorily, while nonlinear relationships between variables and multicollinearity (dependence of a variable on other variables or their linear combinations) are not easily visible. Some types of errors related to normality, independence and linearity still persist in QSAR/QSPR: statistical significance for $\mathbf{x}_j$–$\mathbf{y}$ relations is not checked satisfactorily or not at all by some standard tests (correlation analysis, *t*-test, *F*-test etc.), so statistically irrelevant variables are not dropped; a linear model is not tested for eventual non-linearity; multicollinearity is not inspected satisfactorily to decide which regression method to use (MLR or PLS, for example), so that a redundant dataset may be used in MLR. Such omissions lead to misleading results, unreliable regression coefficients and statistical parameters that underestimate or overestimate the real relationships between $\mathbf{x}_j$ and $\mathbf{y}$. Regression coefficients (Equations (1)–(3)) may be sensitive to inclusion or exclusion of samples, especially when small datasets are involved. They can be unstable with respect to inclusion or exclusion of variables because of relationships between descriptors, due to which each regression coefficient is determined by all $\mathbf{x}_j$ variables. Correlation between variables in QSAR/QSPR is usually quantified by the Pearson correlation coefficient, so that for $\mathbf{x}$ and $\mathbf{y}$ it is defined as the standardized covariance or the scalar product of the respective autoscaled variables (scaled to unit variance):

$$r_{xy} = \frac{(\mathbf{x}-\mathbf{1}\cdot\bar{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{1}\cdot\bar{y})}{(n-1)\sqrt{\sigma_x\sigma_y}},$$
$$\sigma_x^2 = \frac{(\mathbf{x}-\mathbf{1}\cdot\bar{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{1}\cdot\bar{x})}{(n-1)} \text{ and} \quad (4)$$
$$\sigma_y^2 = \frac{(\mathbf{y}-\mathbf{1}\cdot\bar{y})^{\mathrm{T}}(\mathbf{y}-\mathbf{1}\cdot\bar{y})}{(n-1)}$$

where $\mathbf{1}$ is the vector of size $(n\times 1)$ containing ones, T means transposition of a vector and σ_x and σ_y are standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively. When autoscaled variables are used, $\bar{x}$ and $\bar{y}$ are equal to zero and σ_x and σ_y are equal to one, and consequently, r_{xy} is equal to the coefficient b in univariate regression ($a=0$, Equations (1) and (2)). It can be analogously defined for the correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$ as:

$$R_{y\hat{y}} = \frac{(\mathbf{y}-\mathbf{1}\cdot\bar{y})^{\mathrm{T}}(\hat{\mathbf{y}}-\mathbf{1}\cdot\bar{\hat{y}})}{(n-1)\sqrt{\sigma_y\sigma_{\hat{y}}}},$$
$$\sigma_y^2 = \frac{(\mathbf{y}-\mathbf{1}\cdot\bar{y})^{\mathrm{T}}(\mathbf{y}-\mathbf{1}\cdot\bar{y})}{(n-1)} \text{ and} \quad (5)$$
$$\sigma_{\hat{y}}^2 = \frac{(\hat{\mathbf{y}}-\mathbf{1}\cdot\bar{\hat{y}})^{\mathrm{T}}(\hat{\mathbf{y}}-\mathbf{1}\cdot\bar{\hat{y}})}{(n-1)}$$

In the case of univariate regression, the coefficients (Equations (4) and (5)) are equal in size but can differ in sign. The fitting power of a model can be expressed by the square of $R_{y\hat{y}}$, known as the percentage of explained variance or coefficient of determination, usually calculated in QSAR/QSPR as:

$$R_{y\hat{y}}^2 = 1 - \frac{(\hat{\mathbf{y}}-\mathbf{1}\cdot\bar{\hat{y}})^{\mathrm{T}}(\hat{\mathbf{y}}-\mathbf{1}\cdot\bar{\hat{y}})}{(\mathbf{y}-\mathbf{1}\cdot\bar{y})^{\mathrm{T}}(\mathbf{y}-\mathbf{1}\cdot\bar{y})} \quad (6)$$

The practice of relying exclusively on the coefficients (Equations (4)–(6)) and the residual sum of squares is not sufficient to show how much a model follows the basic assumptions of regression analysis.

Homogeneity: It is not contrary to randomness and diversity, but means that the studied substances must belong to a well defined class or related classes of compounds, with the same chemical behavior, mechanism of biological action, chemical reactivity mechanism or physico-chemical features.

Descriptors transformation: In the case that non-linearity is detected, descriptors can be linearized by applying some transformation or form products. Statistical significance of these transforms with respect to $\mathbf{y}$ should be tested when there is sufficient theoretical background for such transforms.

Dataset refining: This refers to practical achievement of homogeneity, and it consists of outlier detection and separation of essentially distinct groups of samples for separate analyses. Scatterplots are good means to perform these operations.

Variable selection: Frequent problems with respect to this item are as follows: no variable selection is performed; it is done partially or not adequately or the presence of overfitting (excess of descriptors in MLR, or too many latent variables in PLS) was not completely checked during variable selection.

Dataset split: This item means splitting the data samples for validation purposes (external validation and bootstrapping). Problems may happen when a dataset is too small and impossible to split, the splitting is arbitrarily and not based on procedures that would include some random routines and the split does not

preserve statistical and chemical similarity between the new datasets.

Selected descriptors: Descriptors for the final regression model should not be of complex definition, difficult to understand or interpret, but could be decomposed into measurable physical properties. Descriptors with only two distinct values are the other extreme: they do not sufficiently encode molecular diversity and do not prove whether their relation with **y** is linear or nonlinear.

Model validation: Not applying model validation by means of standard validation procedures (leave-one-out crossvalidation, leave-many-out crossvalidation, **y**-randomization, bootstrapping, jackknife, external validation, etc.) increases chances to obtain statistically and chemically unreliable models.

Model interpretation: Statistical significance does not mean practical significance. When talking about causality versus statistical significance, language and logic are important to express causality in the correct way [21]. Mistaking correlation for causality, especially in cases of high correlations, and lack of additional analysis to give an explanation for the cause of variation in **y** [6,22,23] are constant dangers. The presence of correlation between two variables can be observed as one of the following cases [5]: (1) direct correlation with cause-and-effect relationship: **y** is directly caused by **x** (it is possible for the opposite to happen, i.e. **x** is caused by **y** [6,24], or it is difficult to determine the direction of causality [23]); (2) direct correlation has no real cause-and-effect relationship: both variables only affect each other; (3) indirect correlation: both variables are related to a third one or to a set of other variables (confounders); (4) Spurious correlation: accidental association of **x** and **y** is an artifact of various factors originating in the measurement, data analysis and computational procedures; (5) nonsense (chance) correlation: systematic association is obtained by pure coincidence and with no natural cause. One has to distinguish three models or three 'languages' concerning regression analysis, with correct 'translation' between them [25]: the causal model (a hypothetical or theoretical statement about the relation between **x** and **y** that should be proved or confirmed by the analysis), the observational model (noticed relations between **x** and **y** from collected data) and the statistical model (regression equation). Mechanistic interpretation of a QSAR/QSPR model is its chemical validation, used to establish the practical significance of the model. The model is verified if it follows chemical knowledge and is based on causality and not on chance correlations. The ultimate goal of QSAR/QSPR is to identify the fundamental properties (descriptors) for certain **y** [26]. Interpretation of regression coefficients should be done taking into account their statistical significance and the way they were obtained (i.e. considering procedural differences between MLR and PLS [27]).

After having worked for several years with QSAR and QSPR applications and model validations, the authors of this work were encouraged to further consider validation procedures for regression models (as presented recently [20]), especially one that has not been taken sufficiently into account (also noticed during discussions at the CMTPI 2009 conference): the regression model's self-consistency or the sign change problem. This phenomenon consists of two facts:

(1) The signs of the Pearson correlation coefficient r_{xy} (Equation (4); r in further text) for one or more descriptors in univariate regression are not preserved during data splitting, i.e. coefficients for the complete (r_c), training (r_t) and external validation (r_e) sets do not satisfy the condition:

$$\mathrm{sgn}(r_c) = \mathrm{sgn}(r_t) = \mathrm{sgn}\ (r_e) \tag{7}$$

(2) The signs of r (Equation (4)) from univariate regression and the respective regression coefficient sign from multivariate regression (Equation (3)) for a given descriptor do not satisfy the equality:

$$\mathrm{sgn}(r_c) = \mathrm{sgn}(\beta_c) = \mathrm{sgn}(\beta_t) \tag{8}$$

where β_c and β_t are regression coefficients for a particular descriptor for the complete and training sets, respectively. An illustrative example is a MLR-QSPR model for predicting boiling points of alkanes [28] (in further text: dataset 36), in which the first two of five descriptors show the signs of β_c values opposite to those of the respective r_c values (Figure 1). A regression model is self-consistent, i.e. the sign change problem is absent when both conditions (Equations (7) and (8)) are valid for all descriptors. In this work, the sign change problem is discussed in detail for a significant number of QSAR and QSPR models from the literature. The purpose of this discussion is to alert researchers to this serious problem, which puts into question the significance and chemical validity of QSAR and QSPR models, and to offer methods for understanding and solving the problem that occurs because of multiple omissions, discussed in the earlier

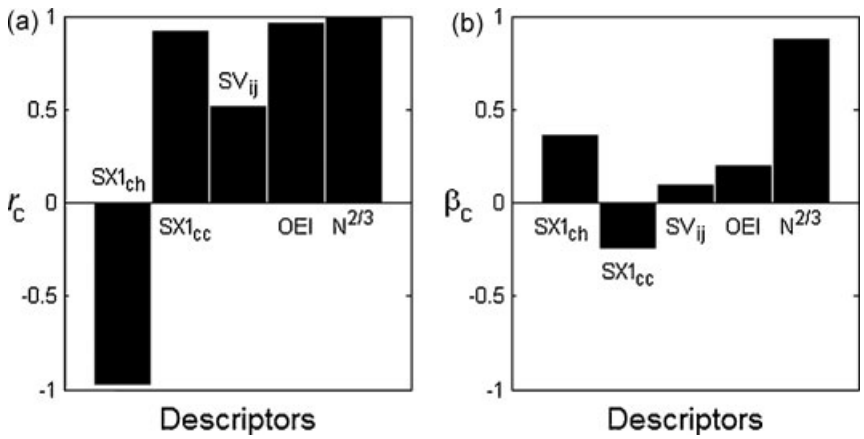

Figure 1. Bar graphs for a MLR-QSPR model for prediction of boiling points of alkanes [28]. The first two descriptors show the opposite signs for β_c and the respective r_c values.

paragraphs, mainly misuse of statistical tools and misinterpretation of models. Even using large datasets and having satisfactory model validation is not a guarantee for the absence of the sign change problem.

2. METHODOLOGY

Datasets for which QSAR and QSPR regression models had been built using MLR, PLS and Principal Component Regression (PCR) were searched in the literature according to the following rules: (1) QSAR datasets were preferred because QSAR is usually more complex to study than QSPR; (2) original datasets were published either with the paper or as supplementary information; (3) papers published after the year 2000 were considered, with preference for recent works; (4) extremely small datasets (less than 15 samples) were not used; (5) Among datasets in the same paper, the one resulting in the model with the best compromise between its statistics and the number of used descriptors was selected; (6) Dataset split was used as in the original publications whenever sufficient information for splitting was provided. Otherwise, a new split was performed in this work based on hierarchical cluster analysis (HCA) [12,17,20].

Each dataset was autoscaled, and then for each descriptor, univariate linear regression **x**–**y** (Equations (1) and (2)) was carried out using the complete dataset and the training set after data splitting. Three Pearson correlation coefficients (r_c and r_t, and r_e when external validation set contained more than six samples) and two regression coefficients (β_c and β_t) were obtained. The regression vectors were normalized by their Euclidean norm, to be comparable with the regression coefficients. Additional parameters, *F*-functions, were defined to characterize descriptors with the sign change:

$$F_1 = \mathrm{sign}(r_c r_t)\sqrt{|r_c r_t|} \qquad F_2 = \mathrm{sign}(r_c r_e)\sqrt{|r_c r_e|} \tag{9}$$

$$F_3 = \mathrm{sign}(r_c \beta_c)\sqrt{|r_c \beta_c|} \qquad F_4 = \mathrm{sign}(r_c \beta_t)\sqrt{|r_c \beta_t|} \tag{10}$$

Data matrices, (r_c, r_t, r_e, β_c and β_t) and (F_1, F_2, F_3 and F_4) were established when sufficiently large external sets were available, otherwise matrices (r_c, r_t, β_c and β_t) and (F_1, F_3 and F_4) were constructed. All the matrices were analyzed by means of HCA (with complete linkage) and principal component analysis (PCA) [12,17] to find features by which acceptable descriptors could be distinguished from those with the sign change. Using statistical parameters to form matrices for exploratory analysis can provide interesting results, as has been shown previously [18].

3. RESULTS AND DISCUSSION

3.1. Data matrices

Extensive search in the QSAR, QSPR and related literature resulted in 50 datasets (meaning 50 regression models), ready or almost ready for use, error-free and satisfying the search criteria. Among the models that comprised 227 descriptors (Table S1 in Supporting Information containing variables r_c, r_t, r_e, β_c, β_t, F_1, F_2, F_3 and F_4), 34 were typical QSAR model and eight were QSPR models, while the rest could be associated either to QSAR or QSPR. Most (42) were MLR models, the others were PLS, but no PCR models were found. A rational HCA-based data split into training and external validation sets had to be performed for 25 models. The entire literature search reflects several difficulties in carrying out this and similar studies, the purpose of which is to contribute to statistical and chemometrical advances in the areas of regression analysis applied to QSAR and QSPR. The main reason for this to occur is the default habit that there is no need to publish the datasets used (more often in QSAR than in QSPR), and this may put in question even the reproducibility of several publications, which was also pointed out recently by Dearden *et al.* [1]. As a rule, datasets from more complicated studies, such as 3D- and 4D-QSAR, are never published.

The sign change problem, as will be shown in the later text by defining four criteria for its assessment, turns out to be a very serious and frequent problem in QSAR and QSPR studies, since results in this work (Table I) are good representatives of the literature in the areas in question. The main reasons for this to occur are as follows: (1) procedural (use of automatic and fast black-box procedures with little or no human intervention); (2) educational (insufficient statistical and chemometrical knowl-

Table I. Statistical summary for QSAR, QSPR and other regression models from the literature

Parameter[a]	Models[b]	Descriptors[c]
Total no. cases	50 (100%)	227 (100%)
No. cases with no sign change problem (Criterion I)	18 (36%)	169 (74%)
No. cases with acceptable descriptor type (Criterion II)	10 (20%)	131 (58%)
No. cases with acceptable **x**–**y** scatterplot (Criterion III)	13 (26%)	142 (63%)
No. cases satisfying all criteria (Criterion IV)	9 (18%)	114 (50%)

[a] Counts and relative counts of models and descriptors satisfying the following criteria: criterion I—the absence of the sign change problem, i.e. the sign consistency of all Pearson correlation coefficients and regression coefficients for a descriptor in the complete and split datasets, is confirmed; criterion II—the presence of a real descriptor, i.e. descriptor with no sign change problem and no extreme change in its contribution to regression models during the data split, is necessary, while other descriptor types are considered unsuitable; criterion III—the presence of a descriptor which has acceptable scatterplot relative to the modeled variable **y** is necessary, i.e. the scatterplot has a reasonable bivariate **x**–**y** distribution or it may be easily remedied by excluding some outliers or distinct samples, or performing modest descriptor's linearization; criterion IV—intersection of criteria I, II and III.
[b] Regression models based exclusively on descriptors satisfying criteria I–IV.
[c] Descriptors satisfying criteria I–IV.

edge) and (3) conceptual, i.e. a misconception that no interpretation and rigorous model validation of regression model is necessary when the paper's goal is not a concrete application (testing new descriptors, theoretical approaches, computational procedures and software). The last item may be rooted in confusing statistics with mathematics [22], a viewpoint that was inherited from these areas some decades ago. To our knowledge, although there are several papers dealing with variable selection and model validations in QSAR and QSPR, there is no recent textbook or monograph devoted exclusively to these aspects of QSAR and QSPR models.

3.2. Criterion I: the simplest check for the sign change problem

An independent variable is free of the sign change problem if all its five parameters (r_c, r_t, r_e, β_c and β_t, or only four when there is no r_e for small external validation sets) satisfy Equations (7) and (8). A regression model is free of sign change if all of its independent variables are free of it. The simplest way to detect the sign change problem is to calculate the values of the five parameters for all descriptors and compare their signs (as shown in Table S1 in Supporting Information). If any descriptor fails in this test, it has to be rejected or replaced by a new one or eventually modified, a new model has to be constructed and inspected for the sign change problem and so on, until a new variable selection results in a reliable model, which should be then validated by other procedures.

The four *F*-functions (Equations (9) and (10)) were defined to show graphically various effects of the sign change problem on the statistics of several QSAR/QSPR models. As can be seen in Figure 2, the *F*-functions are symmetric when plotted against r_c, as expected. The data points placed in the negative regions of the *F*-functions mean that the corresponding independent variables incorporate the sign change problem and, consequently, regression models containing these descriptors are not reliable. F_1 and F_2 account for data splits, where the sign change problem is more problematic for the latter (external validation sets) than for the former (training sets after splitting). Failure in F_1 and F_2 indicates inadequate data splitting or problematic descriptors that do not allow satisfactory splitting. While F_1 has a very few and F_2 a modest number of data points in the negative regions, F_3 and F_4, which account for regression models before and after splitting, respectively, have very well defined and similar X-shaped scatterplot profiles. The profiles show clearer than the scatterplot for F_2 that the sign change problem can occur even for high correlations between **x** and **y** (r_c). F_3 and F_4 mean that an independent variable **x** shows controversial direction of correlation with **y**, for example, positive in univariate and negative in multivariate regression models. Models based on such descriptors are, of course, unreliable in many senses, as will be discussed later.

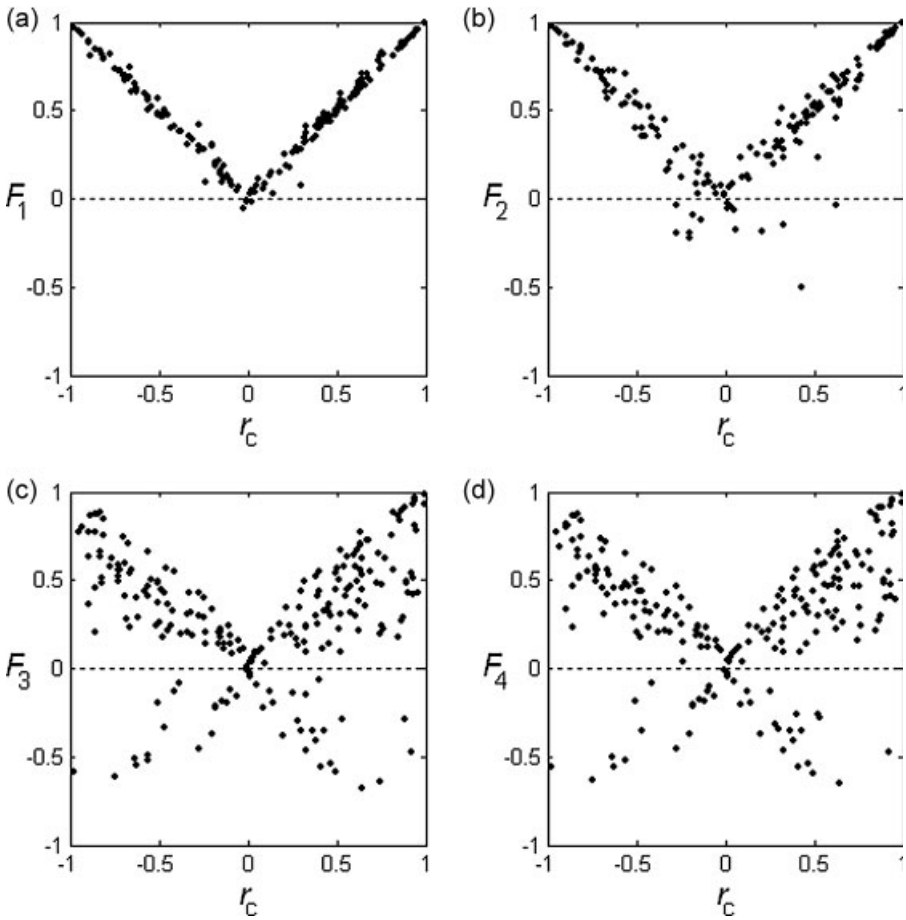

Figure 2. Scatterplots of the four *F*-functions versus the correlation coefficient r_c, by which the presence of the sign change problem can be easily detected in negative regions of the *F*-functions.

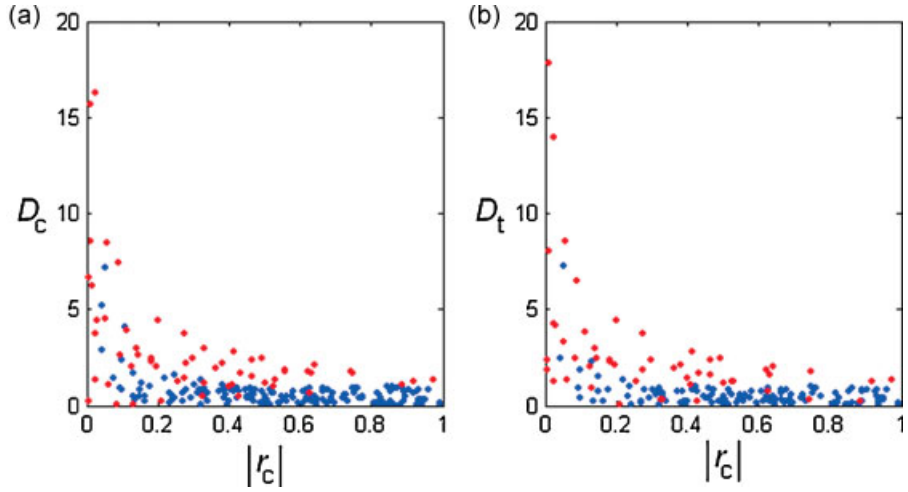

Figure 3. Scatterplots of the change D in regression coefficients versus the absolute value of the correlation coefficient r_c, which occurs when the independent variables are first used in univariate and then in multivariate regression. (a) Complete datasets and (b) training sets, where blue and red data points denote variables without and with the sign change problem, respectively.

Other effects of the sign change problem can be noticed when absolute or relative differences between the parameters (r_c, r_t, r_e, β_c and β_t) are plotted against the absolute value of r_c (Figures 3 and 4):

$$D_c = \left|\frac{r_c - \beta_c}{r_c}\right| \qquad D_t = \left|\frac{r_t - \beta_c}{r_c}\right| \tag{11}$$

$$|D_{te}| = |r_t - r_e| \qquad \left|\frac{D_{te}}{r_c}\right| = \left|\frac{r_t - r_e}{r_c}\right| \tag{12}$$

where D_c and D_t account for the difference between regression coefficients from univariate (Equations (1) and (2)) and multivariate (Equation (3)) models for complete and training sets, respectively, while $|D_{te}|$ and $|D_{te}/r_c|$ present disagreements between training and external validation sets in terms of correlation coefficients. In general, the weaker the correlation of **x** with **y**, i.e. the lower the $|r_c|$, the greater are the values of the D parameters, meaning increasing instability of **x** due to data splitting or multivariate modeling. This trend seems to be exponential in Figure 3 and super-exponential in Figure 4b. When variables with and without (red and blue dots, respectively in colored version of Figures 3 and 4) the sign change problem are distinguished, it is clear that the blue dots are predominant at high correlations and the red dots prevail at low correlations. The red dots show, on average, worse performance of the corresponding variables over the whole range of $|r_c|$. Figures 3 and 4b indicate that the blue dots are almost uniformly distributed in the range of $|r_c|$ from 0.3 to 1, meaning that the corresponding variables are characterized by at most 100% change in regression coefficients (D_c and D_t) and also in

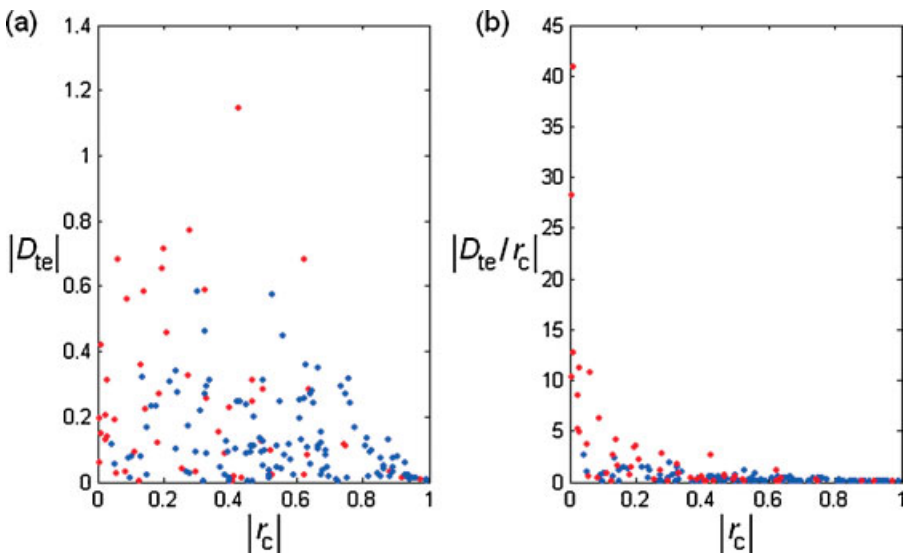

Figure 4. The difference D_{te} between the correlation coefficients for training (r_t) and external validation (r_e) sets, plotted versus the absolute value of the correlation coefficient r_c. (a) The absolute value of D_{te}, and (b) the absolute value of D_{te} relative to r_c, where blue and red data points denote variables without and with the sign change problem, respectively.

correlation coefficients due to splitting ($|D_{te}/r_c|$). The value of $|r_c|$ equal 0.3 has been described as a reasonable limit for molecular descriptors in QSAR/QSPR [20], around which the values of the lowest $|r_c|$ for pre-selected descriptors by means of *t*- or *F*-tests, are concentrated. This agrees with the general recommendation, launched in statistics by Cohen [11], that the value 0.3 distinguishes low from moderate correlations. Therefore, one of the origins of the sign change problem, as noticed in Figures 3 and 4, is the use of statistically non-significant descriptors, i.e. ones with low values of $|r_c|$ in univariate regression. Such descriptors were neither rejected by the *t*- or *F*-tests nor inspected in distribution histograms and scatterplots, they were not tested for non-linearity, and maybe, the data split was not adequate. As a consequence, these variables were not identified as problematic during variable selection and later in model validations, so that the models turned out to be falsely good [20]. Distribution histograms for regression coefficients $|r_c|$, $|r_t|$ and $|r_e|$ (refer to Figures S1–S3 given in Supporting Information) show that a significant fraction of descriptors is placed below the value of 0.3, even with a local peak at 0.1.

According to criterion I, three quarters of descriptors are free of the sign change problem, which accounts for hardly over one third of satisfactory QSAR/QSPR models (Table I). This is an extremely worrying situation that needs urgent action of QSAR/QSPR researchers.

3.3. Criterion II: the six types of independent variables—three noise and three descriptor types

The sign change problem seems to be a complex phenomenon, because of which an empirical extension of criterion I, called criterion II, is being established. Criterion II explores the signs and absolute values of the parameters (r_c, r_t, r_e, β_c and β_t), and takes into account four items: the sign change, training correlations, external correlation and contributions to the model. Detailed rules for this criterion are presented in Table II, where it is seen that one must carefully inspect combinations of the four items to correctly assign classes to variables. One must have in mind that these rules are fast and easy to apply but are of qualitative nature. More quantitative treatments of the parameters should include *t*- or *F*-test for statistical significance of $\mathbf{x}_j$–$\mathbf{y}$ correlations and calculation of errors or confidence intervals for regression coefficients (as done in the usual way for MLR or by applying jackknifing for PLS [27]). The nomenclature of the six classes is based on the following rules:

(1) Independent variables are divided into descriptors (statistically significant correlations with **y**) and noise variables (i.e. noise with respect to **y**). Descriptors have correlation coefficients satisfying criteria $|r_c| > 0.3$ and $|r_t| > 0.3$, which is not valid for noise variables.
(2) The noise variables are divided into: (a) unstable noise: *unstable* because the sign change problem occurs in the

Table II. Decision rules for variable classification according to criterion II

Variable type	Sign change	Basic correlations[a]	External correlation[b]	Contribution to regression model[c]
Real descriptor	Absent	$\|r_c\| \geq 0.3$ and $\|r_t\| \geq 0.3$	$\|r_e\| \geq 0.3$	$\|\beta_c\| \geq 0.001$ and $\|\beta_t\| \geq 0.001$
Quasi descriptor	Absent	$\|r_c\| \geq 0.3$ and $\|r_t\| \geq 0.3$	(1) $\|r_e\| \geq 0.3$ and [(a) $\|\beta_c\| \geq 0.001$ and $\|\beta_t\| < 0.001$, or (b) $\|\beta_c\| < 0.001$ and $\|\beta_t\| \geq 0.001$, or (c) $\|\beta_c\| < 0.001$ and $\|\beta_t\| < 0.001$], or (2) $\|r_e\| < 0.3$ and [$\|\beta_c\| \geq 0.001$ and $\|\beta_t\| \geq 0.001$], or (3) $\|r_e\| < 0.3$ and [(a) $\|\beta_c\| \geq 0.001$ and $\|\beta_t\| < 0.001$, or (b) $\|\beta_c\| < 0.001$ and $\|\beta_t\| \geq 0.001$, or (c) $\|\beta_c\| < 0.001$ and $\|\beta_t\| < 0.001$]	
Anti descriptor	Present	$\|r_c\| \geq 0.3$ and $\|r_t\| \geq 0.3$	Any	Any
Real noise	Absent	(1) $\|r_c\| \geq 0.3$ and $\|r_t\| < 0.3$, or (2) $\|r_c\| < 0.3$ and $\|r_t\| \geq 0.3$, or (3) $\|r_c\| < 0.3$ and $\|r_t\| < 0.3$	Any	$\|\beta_c\| < 0.3$ and $\|\beta_t\| < 0.3$
Unstable noise	Present	(1) $\|r_c\| \geq 0.3$ and $\|r_t\| < 0.3$, or (2) $\|r_c\| < 0.3$ and $\|r_t\| \geq 0.3$, or (3) $\|r_c\| < 0.3$ and $\|r_t\| < 0.3$	Any	Any
Hidden noise	Absent	(1) $\|r_c\| \geq 0.3$ and $\|r_t\| < 0.3$, or (2) $\|r_c\| < 0.3$ and $\|r_t\| \geq 0.3$, or (3) $\|r_c\| < 0.3$ and $\|r_t\| < 0.3$	Any	(1) $\|\beta_c\| \geq 0.3$ and $\|\beta_t\| < 0.3$, or (2) $\|\beta_c\| < 0.3$ and $\|\beta_t\| \geq 0.3$, or (3) $\|\beta_c\| < 0.3$ and $\|\beta_t\| < 0.3$

[a] The coefficients $|r_c|$ and $|r_t|$ are the absolute values of the Pearson correlation coefficients between a descriptor and the dependent variable **y** for the complete dataset and for the training set after data split, respectively.
[b] The coefficient $|r_e|$ is the absolute value of the Pearson correlation coefficient between a descriptor and the dependent variable **y** for the external validation set after data split.
[c] The coefficients $|\beta_c|$ and $|\beta_t|$ are the normalized regression vectors for the complete dataset and for the training set after data split, respectively.

data split or when building multivariate models, or in both; (b) hidden noise: *hidden* in a multivariate model, with no sign change problem but with very significant contribution to the model ($|\beta_c| > 0.3$ or $|\beta_t| > 0.3$, or both) and (c) real noise: *real* because it behaves as expected for a noise variable, with no sign change problem and with low or modest contribution to the model ($|\beta_c| < 0.3$ or $|\beta_t| < 0.3$, or both).

(3) The descriptors are divided into: (a) anti descriptor: *anti* because its behavior in the multivariate model is *against* its univariate nature, i.e. it undergoes the sign change problem; (b) quasi descriptor: *quasi* because the descriptor does not suffer from the sign change problem, but it has very bad performance in splitting ($|r_e| < 0.3$) or in its contribution to the model ($|\beta_c| < 0.001$ or $|\beta_t| < 0.001$, which has not been observed in this work), or both and (c) real descriptor: *real* because the descriptor is probably a true descriptor, satisfying all the conditions of criterion II.

Practical application of criterion II consists of careful inspection of the parameters (r_c, r_t, r_e, β_c and β_t), by which each descriptor is classified into one of the six classes, as shown in Table S2 in Supporting Information, where the descriptors from the only acceptable class (real descriptors) are marked in pink. That is why the noise descriptors can be called 'trash descriptors', while anti and quasi descriptors perhaps can be remedied by some action such as simple linearization, removal of a very few outliers and distinct samples and sample separation into distinct classes. The reader should have in mind that, due to possible instability of regression coefficients, a noise variable can be real in one model, unstable in another and even hidden in some other. Similarly, a descriptor may seem real in one model and anti or quasi in another. However, there is a high probability that a good descriptor used in various models will stay in the same class, as will be shown in subsection 3.7 for dataset 48. In fact, whether some descriptor is real, i.e. it belongs to the desired class of descriptors, can be noticed from various regression models during variable selection, resampling-based model validations, or by exclusion or addition of some descriptors.

Criterion II shows that only less than 60% of the investigated descriptors are acceptable, which corresponds to one-fifth of the studied QSAR/QSPR models (Table I). This is an extremely pre-occupying statistics, and also discouraging if no change can be expected in near future.

3.4. Criterion III: a simple inspection and characterization of *x*–y scatterplots

This criterion consists of graphical inspection of distribution histograms of **x** and **y**, and of **x**–**y** bivariate distribution in the form of scatterplots, by which normality of joint and individual distributions is tested. A descriptor is characterized as: (1) good (statistically very significant), and certainly it should be used for multivariate modeling because of its well-defined linear relation with **y**; (2) acceptable (statistically significant), when there are some problems in the bivariate distribution, but the descriptor can be used in multivariate modeling the way it is, or via modest data modification (descriptor linearization or exclusion of a very few outliers or distinct samples) and (3) not acceptable (statistically not significant), because it cannot be used for multivariate modeling. Conditional acceptance under (2) accounts for well-defined nonlinearity, and the presence of one or a very few outliers or distinct samples that weakly or moderately affect the means and standard deviations of the bivariate distribution. However, when serious problems with a descriptor under (3) are noticed in the scatterplot, then no reasonably modest modification can change this situation. These are as follows: large dispersion of data points without defined linear or nonlinear relations between **x** and **y**; too many outliers or distinct samples or even groups, a few outliers or distinct groups but far away from the majority of data points (i.e. with strong effect on the statistics of the bivariate plot) and the existence of only two distinct values of indicator variable **x**.

The 227 descriptors are analyzed and characterized with additional diagnostic detail in Table S2 which can be found in the Supporting Information, where the suitable descriptor types (good and acceptable descriptors) are marked pink. Similar to criterion II, criterion III shows that only 63% of descriptors can be used for multivariate models, the large majority of which only after modest modification. In other words, only one-quarter of regression models could be considered as reliable (Table I). Criteria II and III agree in terms of descriptors' acceptability and unacceptability for 184 descriptors (81%), meaning that the possibility for the sign change problem to occur may be greatly reduced by graphical means.

3.5. Criterion IV: intersection of criteria I, II and III

This criterion is simply the intersection of the three previous criteria, classifying descriptors as reliable and not reliable. A descriptor that satisfies this criterion (reliable) does not have the sign change problem, it is a real descriptor, and its **x**–**y** scatterplot is good or at least acceptable. Only one half of the investigated descriptors pass this criterion, which corresponds to less than one-fifth of the regression models studied (Table I), raising the question of the sense and usefulness of the great majority of QSAR and QSPR studies. What may be a positive observation is the fact that the models with at least one descriptor satisfying criterion IV make up a large majority (46 or 92%), meaning that it is possible to obtain reliable descriptors in QSAR and QSPR.

3.6. Chemometrical approach to criteria I–IV by means of exploratory analysis

Exploratory analysis was performed on four data matrices: (1) *F*-functions F_1, F_2, F_3 and F_4 for 174 descriptors; (2) *F*-functions F_1, F_3 and F_4 for 227 descriptors; (3) statistical parameters r_c, r_t, r_e, β_c and β_t for 174 descriptors and (4) statistical parameters r_c, r_t, β_c and β_t for 227 descriptors. The results are shown in Figure 5 and Figures S4–S14 with comments in Supporting Information. The scores plots and sample dendrograms are colored to be consistent with Figures 1–3, where statistically most significant and least significant descriptors are in blue and red, respectively, while other transitional classes are in other corresponding colors. The purpose of this analysis is to enable a graphical aid in understanding and treatment of the sign change problem in terms of the four proposed criteria.

The four *F*-functions for 174 descriptors, due to certain mutual similarity (Figure 2), undergo some data contraction, so that two principal components (PCs) describe 95% of the total variance (PC1: 71% and PC2: 24%). A rather clear separation in the scores plots (Figure 5) is visible between descriptors with and without the sign change problem (criterion I), descriptors and noise variables (criterion II), good/acceptable and not acceptable descriptors (criterion III) and between reliable and not

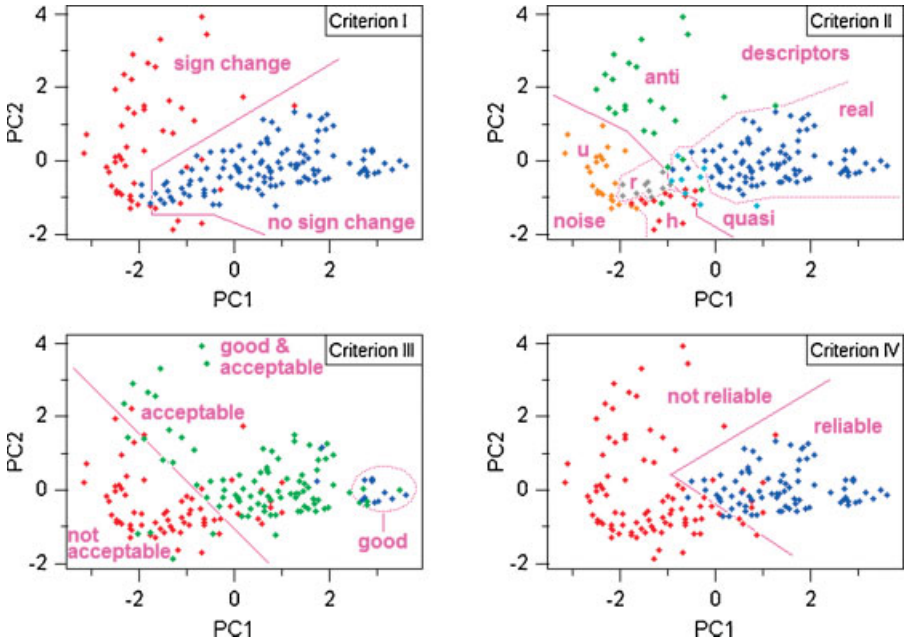

Figure 5. The PC1-PC2 scores plots from the PCA analysis of the four *F*-functions (F_1, F_2, F_3 and F_4), with classes of descriptors marked in different colors and separated by arbitrarily drawn magenta lines. Criterion I: blue—no sign change, red—sign change present; Criterion II: blue, cyan, green—real, quasi, anti descriptors, respectively, and gray, orange, red—real, unstable, hidden noise, respectively; Criterion III: blue—good, green—acceptable, red—not acceptable **x**–**y** scatterplots; Criterion IV: blue—reliable, red—not reliable descriptors. Abbreviations **u**, **h** and **r** in pink stand for unstable, hidden and real noise, respectively, according to criterion II.

reliable descriptors (criterion IV). Six types of descriptor and noise variables can be also distinguished (criterion II). The reader can notice that descriptors with the best performance in all criteria always lie in the region of blue data points (in colored version of Figure 5), defined by the highest values of PC1 and values of PC2 close to zero. The most problematic descriptors are placed at the lowest values of PC1 and around zero of PC2. Distinction of the descriptor classes can be also noticed in the HCA dendrogram (Figure S4), which consists of the greater (A) and smaller (B) cluster, with 111 and 63 descriptors, respectively. What all the four criteria have in common is the general distribution of the classes, so that cluster A contains the majority of statistically significant descriptors, while most of the problematic descriptors are in cluster B. It is interesting to notice that the two clusters show some differentiation with respect to two items (see Figure S5 with comments): descriptors from QSAR/QSAR-like models versus those from QSPR/QSPR-like models and descriptors from PLS versus those from MLR models. When PCA and HCA are applied to three *F*-functions for 227 descriptors, similar results to these just described are obtained (see Figures S6–S8 with comments).

The five parameters (r_c, r_t, r_e, β_c and β_t) for 174 descriptors also undergo some data contraction, so two PCs describe 97% of the original variance (PC1: 80% and PC2: 17%). Data points in the scores plots (Figure S9), when placed farther from the origin along PC1, mean higher correlation of descriptors with **y**, and, when placed distant from the origin along PC2, mean the sign change problem. PC1 can be identified as a degree of the correlation, while PC2 accounts for its correct direction, so the best descriptors are always at extremes of PC1 and around zero PC2, while the worst are spread around the origin and extremes of PC2. The HCA dendrogram shows equilibrated distribution of all descriptors classes between two clusters C and D (Figure S10), including PLS-MLR and QSAR-QSPR distinctions (Figure S11), while distinctions between classes can be observed at the subcluster level. When applying PCA and HCA to the four parameters (r_c, r_t, β_c and β_t) for 227 descriptors, similar results to these just described are obtained (see Figures S12–S14 with comments).

3.7. Rationalizing the origins of the sign change problem with some examples

Both graphical and numerical approaches to inspect statistical significance of correlations between all descriptors and the dependent variable **y** are the crucial steps for regression analysis. During these procedures linear or nonlinear trends, outliers and distinct groups are observed, and it can be determined which samples and descriptors will be used in further calculations. Guyon *et al*. [29] have pointed out that, when roles of each descriptor in univariate and multivariate regression models are compared, falsely relevant and falsely irrelevant variables may appear. Therefore, four types of descriptors can be identified: (1) truly irrelevant (not significant in univariate and multivariate regression); (2) falsely irrelevant (not significant in univariate but significant in multivariate regression); (3) falsely relevant (significant in univariate but insignificant in multivariate regression) and (4) truly relevant (significant in univariate and multivariate regression). Comparing this classification with I–II, relations indeed can be found, and are obvious in case of criterion II: all noise variables are irrelevant (truly or falsely), real descriptors are truly relevant and other descriptors are falsely relevant.

Falsely irrelevant descriptors, as nicely demonstrated by Guyon *et al*. [29], are not descriptors that, after inspecting their relations with **y**, have low correlation and act as relevant in multivariate models (hidden or unstable noise variables according to criterion II). A falsely irrelevant descriptor is weakly correlated to **y**, but is

also a so-called confounder, a causal variable of another variable that is well-correlated to **y**. Therefore, the correlation between the confounder and **y** is indirect, but is of theoretical importance. The interaction between the confounder and its effect variable can be modeled as their product or another more complicated term. Such a new variable is an independent variable that must be well correlated to **y** and treated in the same way as other descriptors.

Falsely relevant descriptors are certainly the most frequent type of problematic descriptors, in many cases accompanied with the sign change problem. There are several causes for the existence of such descriptors, to which MLR and PLS models are not immune: (a) failure in eliminating noise variables; (b) presence of outliers and distinct samples; (c) other significant deviations of the **x**–**y** distribution from the normal bivariate distribution; (d) untreated nonlinearity; (e) Simpson's paradox [11,29,30]; (f) multicollinearity (this includes high descriptors' intercorrelations and redundancy in the covariance matrix in MLR, and some other correlation effects in PLS) and (g) overfitting (use of too many descriptors in MLR and too many latent variables in PLS). Simpson's paradox accounts for distinct classes of samples, which, when put together, show trends in the **x**–**y** scatterplot that are significantly different from those for the individual classes. This can be seen for topological descriptor SX_{1CC} (Figure 6), which was used in a QSPR model to predict boiling points of alkanes (refer to dataset 36 from Table S1 in Supporting Information, with the sign change problem as shown in Figure 1) [28]. SX_{1CC} is an anti descriptor, with acceptable **x**–**y** scatterplot after linearization, and not reliable. There are seven classes of alkanes, isomer groups for which correlations of SX_{1CC} with the boiling point are mutually similar and negative, and not positive as for all 150 alkanes. Since some important information about the classes was ignored when putting them together, the common correlation is a false, artificial effect that does not reflect the true relation between SX_{1CC} and boiling point.

Figures 3 and 4 have already been discussed in detail with respect to the increasing instability of descriptors in terms of their regression coefficients as a function of decreasing correlation to **y**. Such an instability can be noticed during variable selection in a concrete QSAR or QSPR study. The stability of three descriptors from the final MLR model for prediction of human toxicity HAP [31] (dataset 48 in Table S1), LUMO, Human Liver (human liver toxicity as $-\log IC_{50}$) and N_O (number of oxygen atoms), can illustrate this trend well (Figure 7). The complete descriptors pool in this QSAR-like study [31] contained 22 variables, so each of the three descriptors is used in 21 bivariate linear regressions to predict HAP (Table S3 in Supporting Information), and all β_{1c} values are plotted against $|r_{1c}|$ for the three descriptors. The values in Table S3 are for the complete dataset, and very similar results were obtained for the training set based on the split from Table S1 (not shown). According to criterion II, Human Liver and LUMO were previously characterized as real descriptors (see Table S2), and the same is confirmed by 21 bivariate regressions (Figure 7), where LUMO is somewhat more stable than Human

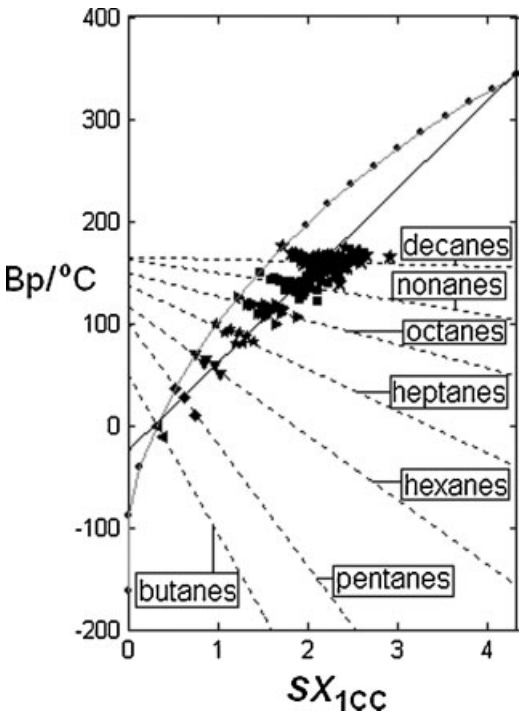

Figure 6. Illustration of Simpson's paradox in the scatterplot of boiling point (Bp) against a topological descriptor SX_{1CC} from QSPR dataset 36 on alkanes [28]. The regression line for all compounds is the solid line, the seven regression lines for isomer groups are dashed and the interpolating curve for normal alkanes is drawn in gray. Data points for distinct isomer groups are drawn using different symbols.

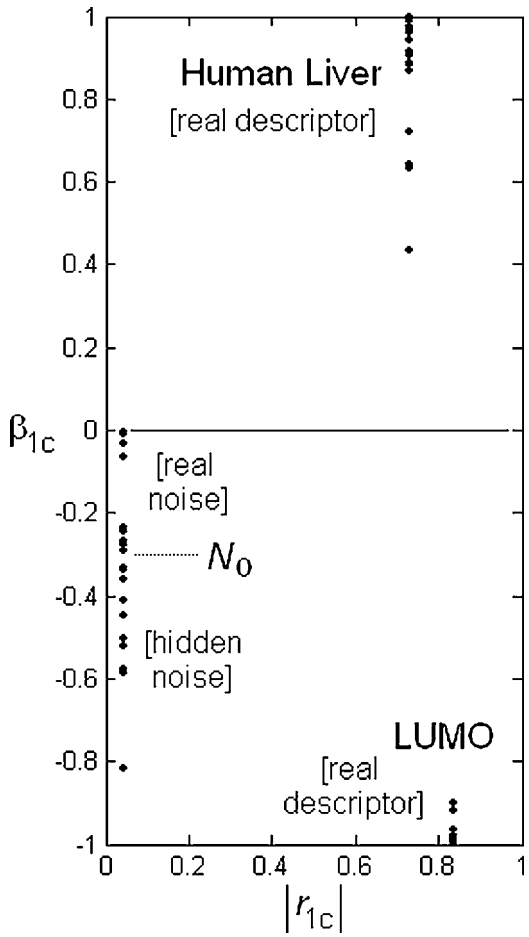

Figure 7. Illustration of descriptors' instability in terms of β_{1c} plotted against $|r_{1c}|$ for all bivariate regressions generated from the complete descriptors pool for dataset 48. The three descriptors are those from the final model for prediction of human toxicity HAP [32].

Liver due to its higher correlation with HAP. N_O, which was characterized as hidden noise and with serious problems in its **x**–**y** scatterplot (Table S2), now shows larger variations in β_{1c}, because of which it behaves as real noise and hidden noise. One may notice some very high absolute values of β_{1c}, which is a good illustration of falsely relevant variables.

The paper 'Multicollinearity' in Salkind's statistical encyclopedia [11] summarizes the effect of multicollinearity on MLR: (1) large fluctuation (instability) of the estimated regression coefficients caused by small perturbations to the model (inclusion or exclusion of some variables or samples); (2) the sign change problem and (3) the possibility that the full regression model is statistically significant and, at the same time, some or all regression coefficients are not significant. This way, the inherent nature of the regression coefficient [5,6,8] is lost (see Table S3 with comments in Supporting Information). Geometrically speaking, the MLR equation (Equation (3)) is not represented well by a hyperplane in the $m+1$-dimensional space [3], because the projections of the hyperplane onto the coordinate planes, the regression lines, do not have the same slopes as the regression lines from the corresponding univariate regressions (Equation (2)), and even may change directions (the sign change problem).

Regression coefficients in MLR and PLS are complicated rational functions of all correlation coefficients that quantify linear relations between descriptors and **y** and also between descriptors, as can be seen when comparing bivariate regressions MLR and PLS (Table III). (More details about these expressions and more complicated cases of MLR are given in Table S4 and Appendices 1 and 2 of Supporting Information.) Computationally speaking, MLR becomes useless in cases of pronounced multicollinearity because of inversion of the covariance matrix which is not a problem for PLS. A simpler form of multicollinearity, i.e. descriptor intercorrelations, is included in regression coefficients in MLR and also in PLS (Table III). Besides that, all descriptors participate in defining regression coefficients, literally they 'compete' or 'concur' [5], and it is difficult to predict which will be the value and sign of a particular regression coefficient. Descriptors that share basically the same information will have their contributions similar in the multivariate model but smaller than in the respective univariate regressions [32]. It seems from expressions for regression coefficients that multicollinearity enhances this competition, which may result in distinguishing descriptors as 'winners' (important to the model) and 'losers' (not really important, even with changed signs). The PLS example presented has a didactical purpose to show that intercorrelations are indeed important for PLS, that descriptors 'compete' also in PLS and that intercorrelations are taken into account differently in PLS than in MLR. Peterangelo and Seybold warn [33] that, even when multicollinearity is present to a significant extent, it may happen that descriptors, highly correlated with others, contain some important information and should stay in the model.

Coming back to the bivariate MLR models associated to dataset 48 (Figure 7), it is possible to show how correlations between descriptors affect regression coefficients. When absolute values of regression coefficients β_{1c} from 21 bivariate models are plotted against the absolute value of correlation coefficients between **y** and other variables (r_{2c}) or of coefficients of intercorrelations (r_{12}) (data are in Table S3), systematic variations can be noticed for descriptors Human Liver (Figure S15), LUMO (Figure S16) and N_O (Figure S17). LUMO, which is the most stable among the three descriptors (Figure 7), is not affected by the presence of other descriptors and multicollinearity (Figure S16). The contribution of Human Liver to the bivariate models is modestly affected by multicollinearity and descriptor's competition (Figure S15). However, the noise variable N_O illustrates rather clearly that the contribution of such variables to models decreases with the increase of multicollinearity and the correlation of other variables with **y**.

A more complex form of multicollinearity in MLR is due to significant correlations between some descriptors and the linear combination of two or more descriptors, in which the former is not important for the model, known as redundancy [34], and is related to overfitting in MLR (excess of descriptors that is not so obvious when observing the correlation matrix). This is very difficult to observe graphically, and perhaps dispersion of the data points in Figures S15 and S17 is caused by it.

It is interesting to investigate how overfitting in PLS affects the sign change problem. For this purpose, some parameters were calculated for all eight datasets for PLS models (defined in Table S1) for all latent variables, as is shown in Figure 8, and Table S5 and Figure S18 with comments in Supporting Information. The average F_3-function ($\overline{F_3}$) was calculated as the sum of F_3 values for all descriptors in a model and then divided by the number m of descriptors. When plotted as a function of the number of latent variables (N_{LV}), it is noticeable that the general tendency of $\overline{F_3}$ is decreasing, and in cases of the sign change problem this is very pronounced, up to values around 0.1 or even below. It is well-known in the literature that PLS overfitting occurs when the number of latent variables is greater than the optimum number

Table III. Multicollinearity effects on regression coefficients in bivariate MLR and PLS[a]

Univariate regression(s)[b]	Multivariate regression	MLR without multicollinearity	MLR with multicollinearity	PLS with one LV
$\hat{\mathbf{y}} = b_1\mathbf{x}_1$	$\hat{\mathbf{y}} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2$	$\beta_1 = b_1 = r_1$	$\beta_1 = \frac{r_1 - r_2 r_{12}}{1 - r_{12}^2} \neq b_1$	$\beta_1 = \frac{r_1\left(r_1^2 + r_2^2\right)}{r_1 + r_2 + r_1 r_2 r_{12}^2} \neq b_1$
$\hat{\mathbf{y}} = b_2\mathbf{x}_2$		$\beta_2 = b_2 = r_2$	$\beta_2 = \frac{r_2 - r_1 r_{12}}{1 - r_{12}^2} \neq b_2$	$\beta_2 = \frac{r_2\left(r_1^2 + r_2^2\right)}{r_1 + r_2 + r_1 r_2 r_{12}^2} \neq b_2$

[a] All variables are autoscaled, i.e. scaled to unit variance.
[b] Univariate regressions for descriptors $\mathbf{x}_1$ and $\mathbf{x}_2$. The predicted values of the dependent variable **y** are marked with $\hat{\mathbf{y}}$, and b_1 and b_2 are regression coefficients for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The Pearson correlation coefficients for correlations between descriptors $\mathbf{x}_1$ and $\mathbf{x}_2$ and **y** are r_1 and r_2, respectively, while r_{12} is the correlation coefficient for descriptors' intercorrelation.

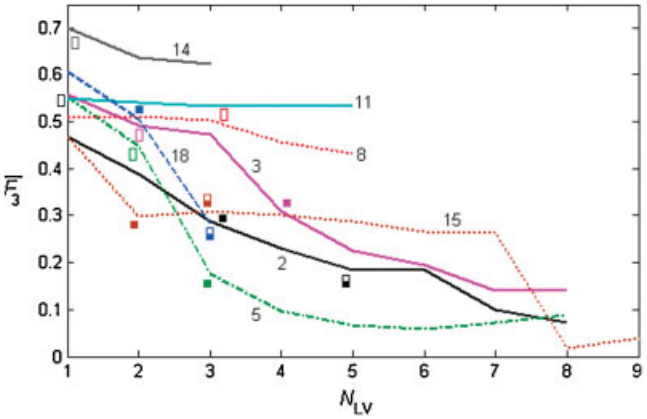

Figure 8. The average F_3-functions ($\overline{F_3}$) for the complete dataset as a function of the number of latent variables (N_{LV}) for eight PLS datasets, are marked with numbers and different colors (2—black, 3—magenta, 5—green, 8—red, 11—cyan, 14—gray, 15—brown and 18—blue). The values of N_{LV} for the PLS models from literature are marked with small rectangular boxes (which are semi-solid if the sign change has been detected). Small solid squares mean the first latent variable at which the sign change occurs.

[15,35–37]. Leave-one-out crossvalidation has been frequently reported as not being always adequate for determination of this parameter.

A more general vision about the sign change problem of all 50 regression models can be obtained when parameters $\overline{F_3}$, its analogue $\overline{F_4}$ (average F_4-function) and other parameters based on descriptors with the sign change problem are calculated, as shown in Table S6 and Figures S19–S21 with comments in Supporting Information. In these plots, $\overline{F_3}$ and $\overline{F_4}$ always decrease with m, indicating the presence of the sign change problem in some critical interval of values, especially when several descriptors are used. The extent of the sign change problem may be disastrous, when most descriptors in a model show sign changes. All these trends are common to MLR and PLS models, to QSAR/QSAR-like as well to QSPR/QSPR-like models.

4. CONCLUSIONS

The sign change problem of descriptors in QSAR, QSPR and related studies is defined in this paper as the controversy in the signs of contributions of a particular descriptor when its correlation coefficients and regression coefficients are compared for univariate and multivariate regression models, using both complete and split datasets. This is a very serious and frequent problem that puts into question the majority of the investigated regression models, which is the most probable situation in the literature. Four criteria, created in this work, were applied to 50 QSAR and QSPR models with 227 descriptors by means of a set of rules and by exploratory analysis, in order to characterize descriptors, identify problematic ones and improve regression models so that the sign change problem could be avoided. The origins of the sign change problem are discussed in terms of violations of the fundamental statistical and QSAR/QSPR rules, especially in terms of multicollinearity. For practical purposes, it is recommended that the proposed or some other equivalent criteria be used in descriptors' pre-selection and variable selection, and that the sign change problem is checked prior to model validation. A final and very important step to eliminate the sign change problem is model interpretation, still omitted in several QSAR, QSPR and related studies. If the described procedures are not sufficient to eliminate the sign change problem, then the dataset must be changed by introducing new descriptors or samples (substances), or as the ultimate step, the problem must be redefined.

Acknowledgements

The authors acknowledge The State of São Paulo Funding Agency (FAPESP) for financial support and Dr Carol H. Collins for English revision.

REFERENCES

1. Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* 2009; **20**: 241–266.
2. OECD *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69. OECD: Paris, 2007. Available at: http://www.oecd.org/dataoecd/55/35/38130292.pdf [8 May 2008].
3. Weisberg S. *Applied Linear Regression* (3rd edn). Wiley series in probability and statistics. Wiley: Hoboken, NJ, 2005.
4. Segupta D, Jammalamadaka SR. *Linear Models: An Integrated Approach*, Series on multivariate analysis, Vol. 6. World Scientific: River Edge, NJ, 2003.
5. Schroeder LD, Sjoquist DL, Stepha PE. *Understanding Regression Analysis: An Introductory Guide*, Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07–057. Sage: Newbury Park, CA, 1986.
6. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models* (5th edn). McGraw-Hill: Boston, 2005.
7. Rencher AC. *Methods of Multivariate Analysis* (2nd edn). Wiley series in probability and mathematical statistics. Wiley: New York, 2002.
8. Chatterjee S, Hadi AS. *Regression Analysis by Example* (4th edn). Wiley series in probability and mathematical statistics. Wiley: Hoboken, NJ, 2006.
9. Timm NH. *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer-Verlag: New York, 2002.
10. *Regression analysis and other linked articles*. Wikipedia—the free encyclopedia. Wikimedia Foundations, San Francisco, CA. Available at: http://en.wikipedia.org/wiki/Regression_analysis [13 September 2009].
11. Salkind NJ (ed). *Encyclopedia of Measurement and Statistics*, vols. 1–3. Sage: Thousand Oaks, CA, 2007.
12. Livingstone D. *Data Analysis for Chemists*. Oxford University Press: Oxford, 2002.
13. Kubinyi H. Quantitiative Structure-activity relationships in drug design. In *Encyclopedia of Computational Chemistry*, vol. 4, von Schleyer PR (ed). Wiley: Chichester, UK, 1998; 2309–2320.
14. Jurs PC. Quantitative structure-property relationships QSPR. In *Encyclopedia of Computational Chemistry*, vol. 4, von Schleyer PR (ed). Wiley: Chichester, UK, 1998; 2320–2330.
15. Vandeginste BGM, Massart L, Buydens LMC, de Jong S, Lewi PJ, Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics, Part B*, Data Handling in Science and Technology, vol. 20B. Elsevier: Amsterdam, 1998; 366–371, 383–420.
16. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry. vol. 11. Wiley-VCH: Weinheim, 2000.
17. Ferreira MMC. Multivariate QSAR. *J. Braz. Chem. Soc.* 2003; **13**: 742–753.
18. Kiralj R, Ferreira MMC. On heteroaromaticity of nucleobases. Bond lengths as multidimensional phenomena. *J. Chem. Inf. Comput. Sci.* 2003; **43**: 787–788.
19. Leach AR, Gillet VJ. *An Introduction to Cheminformatics*. Springer: Dordrecht, 2007; 75–97.

20. Kiralj R, Ferreira MMC. Basic Validation procedures for regression models in QSAR and QSPR studies: Theory and application. *J. Braz. Chem. Soc.* 2009; **20**: 770–787.
21. Bennett J. *Events and Their Names*. Clarendon: Oxford, 1988.
22. Ben-Zvi D, Garfield J (eds). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Kluwer: Dordrecht, 2004.
23. Graham A. *Developing Thinking In Statistics*. Open University: London, 2006.
24. Adler I, Adler R. *Probability and Statistics for Everyman: How to Understand and Use the Laws of Chance*. Signet Science Library Books. Signet: New York, 1966; 176–177.
25. Shipley B. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press: Cambridge, UK, 2004; 21–25.
26. Charton M. The nature of topological parameters. I. Are topological parameters 'fundamental properties'? *J. Comput.-Aid. Mol. Des.* 2003; **17**: 197–209.
27. Burnham AJ, MacGregor JF, Viveros R. Interpretation of regression coefficients under a latent variable regression model. *J. Chemom.* 2001; **15**: 265–284.
28. Cao CZ, Jiang LH, Yuan H. Eigenvalues of the bond adjacency matrix extended to application in physicochemical properties of alkanes. *Internet Electron. J. Mol. Des.* 2003; **2**: 621–641.
29. Guyon I, Aliferis C, Elisseeff A. Causal feature selection. In *Computational Methods of Feature Selection*, Liu H, Motoda H (eds). Taylor & Francis: Boca Raton, FL, 2008; 63–85.
30. Rücker G, Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Med. Res. Methodol.* 2008; **8**: article 34.
31. Lessigiarska I, Worth AP, Netzeva TI, Dearden JC, Cronin MTD. Quantitative structure-activity-activity and quantitative structure-activity investigations of human and rodent toxicity. *Chemosphere* 2006; **65**: 1878–1887.
32. Graham MH. Confronting multicollinearity in ecological multiple regression. *Ecology* 2003; **84**: 2809–2815.
33. Peterangelo SC, Seybold PG. Synergistic interactions among QSAR descriptors. *Int. J. Quant. Chem.* 2004; **96**: 1–9.
34. Flury BW. Understanding partial statistics and redundancy of variables in regression and discriminant analysis. *Am. Stat.* 1989; **43**: 27–31.
35. Mark H, Workman J. *Chemometrics in Spectroscopy*. Academic Press: New York, 2007; 165–166.
36. Gemperline P (ed). *Practical Guide to Chemometrics* (2nd edn). CRC Press: Boca Raton, FL, 2006.
37. Varmuza K, Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics* (chapter 4). CRC Press: Boca Raton, FL, 2009;