Talanta 87 (2011) 180–188

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Theil–Sen nonparametric regression technique on univariate calibration, inverse regression and detection limits

Irma Lavagnini, Denis Badocco, Paolo Pastore*, Franco Magno

Department of Chemical Sciences, University of Padova, Via Marzolo 1, 35131 Padova, Italy

ARTICLE INFO

Article history:
Received 6 July 2011
Received in revised form
23 September 2011
Accepted 27 September 2011
Available online 12 October 2011

Keywords:
Theil–Sen regression
Nonparametric confidence region
Tolerance intervals
Detection limit

ABSTRACT

This paper reports the combined use of the nonparametric Theil–Sen (TS) regression technique and of the statistics of Lancaster–Quade (LQ) concerning the linear regression parameters to solve typical analytical problems, like method comparison, calculation of the uncertainty in the inverse regression, determination of the detection limit. The results of this new approach are compared to those obtained with appropriate reference methods, using simulated and real data sets. The nonparametric Theil–Sen regression technique appears a new robust tool for the problems considered because it is free from restrictive statistical constraints, avoids searching for the error nature on x and y, which may require long analysis times, and it is easy to use. The only drawback is that the intrinsic nature of the method may lead to a possible enlargement of the uncertainty interval of the discriminated concentration and to the determination of larger detection limits than those obtainable with the commonly used, less robust, regression techniques.

1. Introduction

Linear regression is a mathematical procedure commonly used to determine the concentration of the identified component of an unknown sample using calibration measurements [1]. Typical problems faced by it in connection with the calibration are: (i) the determination of the detection limit of an instrumental method; (ii) the quantification of the uncertainty of a discriminated (or predicted) variable in the inverse regression; (iii) the method comparison. Usually the regression coefficients are calculated by applying the ordinary least-squares (OLS), the parametric robust iteratively reweighted least-squares (IRLS) in the presence of outliers [2] or, whenever heteroscedasticity is present in the dependent variable, the weighted least-squares (WLS). Under the hypotheses that the independent variable is free of error and that the errors in the dependent variable are normally distributed, the points (i) and (ii) can be satisfactorily solved by the construction of the prediction band about the calibration line [3,4]. This approach also holds when uncertainty is associated to the independent variable (usually the concentration) but pre-assigned (or nominal) constant values of this variable are used in replicate measurements in the calibration procedure [5]. Indeed, the effect of the "making up solutions" factor is taken into account through the overall measurement uncertainty at fixed concentration level which is comprehensive of the contributes of the instrumental and of the concentration variances. If certified reference materials are used as calibration standards and their uncertainty is of the same order of magnitude of that of the response variable, the bivariate least-squares (BLS) regression technique is found suitable to calculate the regression coefficients [6], and to construct both their joint confidence region [7] and the prediction intervals for the response [8].

As about the point (iii) the widely used OLS-based procedure assumes the responses of the reference method as free of error (x-axis) while y-axis refers to the responses obtained by the new method. It applies separate or joint statistical tests on the intercept and slope values obtained by linear regression, to evaluate the significance of the difference from the theoretical values of zero and unity, respectively. When random errors of the same order of magnitude are present in the new method and in the reference one, BLS techniques [8] or the t-paired test procedure [9] may be successfully applied. Anyway, in assessing the accuracy of analytical methods, the hypothesis of normality of the data is assumed or proved to be fulfilled. Lack of Gaussian distribution assumption occurs when the method comparison is performed on the concentrations obtained by calibration from two experimental procedures [10,11]. Nonparametric statistics, like the Wilcoxon T-test [9], offers a remedy since it does not require the knowledge of underlying probabilistic models.

In this paper the successful median-based regression procedure to calculate calibration lines, due to Theil [12] and Sen [13], is the starting point to solve the three problems considered with a new unique technique. The non-parametric Theil–Sen approach (TS) is robust and easy to compute, as widely acknowledged in several

* Corresponding author. Tel.: +39 0498275182; fax: +39 0498275175.
E-mail address: paolo.pastore@unipd.it (P. Pastore).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.09.059

textbooks on nonparametric statistics [14,15]. Very often the Theil's regression method, which collects the slopes of all pairs of points and finds the median of these slopes, is applied for zeroing the effects of outliers [16,17]. Actually its breakdown point is 29%, i.e. nearly one-third of the data can be replaced by contaminants before the parameters are affected, and its stability is improved using repeated medians [18], as proposed in the new regression method developed by Siegel [19,20]. The present paper is aimed to become a unique reference tool for the data treatment in analytical measurements involving regression methods. Its main feature is that it is free from statistical hypotheses which usually affect the traditional regression methods used in analytical chemistry. We develop the methodology firstly describing the Theil and Sen linear regression technique combined with the statistical approach of Lancaster and Quade (LQ) [21]. The approach allows to compare two analytical methods in a nonparametric way. Moreover, two different approaches, both based on the knowledge of the joint confidence region of the regression coefficients, are described to determine the uncertainty of the discriminated variable in the inverse regression. The former approach proposes to construct simultaneous tolerance intervals, i.e. for all concentrations, about the TS calibration line. Tolerance intervals are constructed instead of the more used prediction intervals [3] starting from the confidence band of the calibration line. The calculation of the tolerance intervals reproduces the expression for these intervals in the OLS or WLS theories [22,23]. In those theories the variance of the expected response value, properly bounded above, is put together with the confidence band of the regression line, which in our case accounts for possible errors in both axes. In the latter approach the experimental confidence band of the experimental response matches the joint confidence band of the calibration line [24]. The simultaneous tolerance intervals are also utilized to determine the detection limit of an analytical method [23,25]. The performance of the TS–LQ approach is evaluated at different steps. Firstly, three simulated data sets with random errors added via the Monte Carlo method are used to study the characteristics of the TS–LQ joint confidence regions. Then, real data sets are considered for method comparison, for detection limit estimation of a method and for concentration uncertainty determination in the inverse regression. The results obtained are compared with those based on appropriate, more traditional, regression techniques, chosen among OLS, IRLS, WLS, and BLS, to validate the proposed TS–LQ methodology.

2. Methods

2.1. Theory of the Theil–Sen regression technique

2.1.1. Estimation of the linear regression model parameters

In the simple linear regression model the responses Y_i, $i = 1, 2, \ldots, n$, are dependent on X_i through the relationship

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where errors ε_i are independent identically distributed variables. The unknown distribution function of ε_i is assumed to be continuous with mean zero.

If all x_i are distinct, the point estimator of the slope β_1 proposed by Theil is the median $\hat{\beta}_{1T}$ of the $\binom{n}{2}$ slopes $(\hat{\beta}_1)_{ij} = (y_j - y_i)/(x_j - x_i)$, $1 \le i < j \le n$, where $\binom{n}{2} = (n(n-1))/2$ [12]. In the presence of ties among x_is, namely in the calibration procedure replication of measurements at a fixed concentration level, the Theil estimator of the parameter β_1 was extended by Sen (TS) as the median $\hat{\beta}_{1TS}$ of the $N \le \binom{n}{2}$ slopes $(\hat{\beta}_1)_{ij} = (y_j - y_i)/(x_j - x_i)$ for which $x_i \ne x_j$ [13].

The estimate of the intercept is then calculated as $\hat{\beta}_{0TS} = \text{med}\{y_i - \hat{\beta}_{1TS} x_i\}$, where 'med' stands for the median.

The statistical properties of the estimators of the slope $\hat{\beta}_{1T}$ and $\hat{\beta}_{1TS}$ have been studied in terms of unbiasedness, symmetry of the distribution, and asymptotic distribution in order to obtain the corresponding confidence interval [13,18]. In particular, the TS estimator of the slope $\hat{\beta}_{1TS}$ is unbiased and asymptotically normal for all unknown but continuous distributions of the experimental measurements Y_i. These properties still are valid under the assumption that X_is are random variables [26].

2.1.2. Joint confidence region for the slope and intercept

A joint confidence region for the linear regression parameters estimated by the Theil–Sen method may be obtained based on the sign of the residuals $Y_i - \beta_0 - \beta_1 X_i$, where β_0 and β_1 are guess values. The procedure firstly proposed by Lancaster and Quade [21] was developed as a nonparametric test for hypotheses concerning the β_0 and β_1 parameters, simultaneously. It combines the test which has the power to reject incorrect slopes, based on the statistic T

$$T = \frac{\sum\sum_{i<j} sgn\left\lfloor (Y_i - \beta_1 x_i - Y_j + \beta_1 x_j)(x_i - x_j)\right\rfloor}{\sqrt{N\binom{n}{2}}}, \quad (1)$$

where 'sgn' is the sign function, with the test based on the sign statistic

$$L = \sum_{i=1}^{n} \frac{1 + sgn(Y_i - \beta_0 - \beta_1 x_i)}{2}, \quad (2)$$

which detects incorrect intercepts. The sign function, sgn(a), assumes the value 1, 0, or −1 if a is positive, zero, or negative, respectively.

Since the Kendall statistic T and the sign statistic L are asymptotically normal, the standardized statistics $Z_1 = (T - \mu_T)/\sigma_T$ and $Z_2 = (L - \mu_L)/\sigma_L$, where μ and σ have their usual meaning, are used to compute the statistic $C = Z_1^2 + Z_2^2$, which has an asymptotic chi-squared distribution with 2 degrees of freedom. For small-sample ($n \le 15$) the distribution of the statistic C has been reported in the literature [21].

In the presence of tied observations, if a_n (≥ 2) is the number of distinct sets in the x-sample, $x_1, x_2, \ldots, x_n$, and u_j is the number of elements which are all equal in the jth set, the statistic C is given by [13,21]

$$C = \frac{N\binom{n}{2} T^2}{(1/18)\{n(n-1)(2n+5) - \sum_{j=1}^{a_n} u_j(u_j-1)(2u_j+5)\}} + \frac{(2L-n)^2}{n}, \quad (3)$$

since $\mu_T = 0$,

$$\sigma_T^2 = \frac{(1/18)\{n(n-1)(2n+5) - \sum_{j=1}^{a_n} u_j(u_j-1)(2u_j+5)\}}{N\binom{n}{2}},$$

$$\mu_L = n/2, \quad \text{and} \quad \sigma_L^2 = n/4.$$

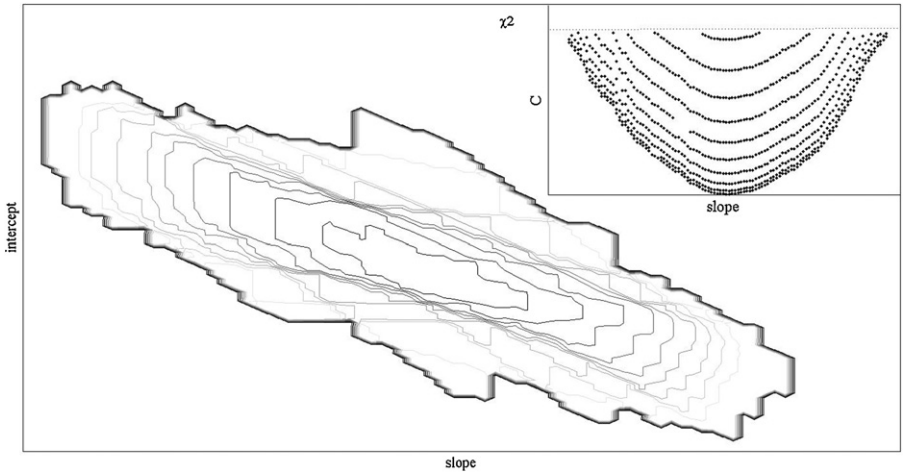

Fig. 1. TS–LQ joint confidence region for the slope and intercept of the regression line. The isolines from the inner to the outer part of the region are relevant to increasing values of the statistic $C \leq {}^{1-\alpha}\chi^2_2$. Inset: statistic C values satisfying the condition $C \leq {}^{1-\alpha}\chi^2_2$ as a function of the slope for various intercept values.

The $(1-\alpha)100\%$ confidence region in the plane slope/intercept, namely the pairs of values of slope and intercept inside the region which are jointly compatible with the data at the α level of significance, may be obtained from the equation $C \leq {}^{1-\alpha}\chi^2_2$, where ${}^{1-\alpha}\chi^2_2$ is the $(1-\alpha)100$ percentage point of the χ^2 distribution with 2 degrees of freedom. Fig. 1 shows the TS–LQ joint confidence region together with the isolines characterized by different levels of significance α. The inset shows the set of C points satisfying the condition $C \leq {}^{1-\alpha}\chi^2_2$ versus the slope for fixed values of the intercept.

As usual, the $(1-\alpha)100\%$ confidence bands for the regression line may be obtained as a by-product of the construction of the joint confidence region, determining at each x the minimum value $y^{-}_{C,1-\alpha}(x)$ and the maximum value $y^{+}_{C,1-\alpha}(x)$ of the y-value set $y = \beta'_0 + \beta'_1 x$, where the parameter couple (β'_1, β'_0) lies on the contour of the confidence region [27]. It is to be outlined that no explicit equation is available for the limits $y^{\pm}_{C,1-\alpha}(x)$ from the TS–LQ theory developed above, being they numerically computed only.

2.2. Method comparison

When two methodologies are compared via the linear regression, the responses in the same units obtained by the two methods from samples at various concentration levels of the analyte should give a straight line of theoretically unity slope and zero intercept. The agreement of the results of the methods is statistically achieved at a given level of significance if the couple (1, 0) falls into the limits of the joint confidence region in the plane slope/intercept.

It is well known that, when the parameters of the regression line are calculated using unweighted or weighted parametric methods, this region is an ellipse which has the center at the estimated regression coefficients $(\hat{\beta}_1, \hat{\beta}_0)$ and semiaxes with lengths depending on the level of confidence chosen and on the residual variance of the regression [28]. If a nonparametric regression technique is used, no equation is available for the joint confidence region and only an approximated shape of the region is obtainable, as indicated above.

2.3. Confidence band based-procedure in inverse regression

Nonparametric confidence bands of the regression line, as previously constructed, are simultaneous, i.e. for all x, in character. This feature prompted the calculation of simultaneous tolerance intervals in order to face two analytical problems linked to the use of the inverse regression: (i) determination of the confidence interval for discriminated concentration value [29], and (ii) determination of the detection limit of an analytical method [23]. The starting point is the equation of the limits $y^{\pm}_{tol,1-\alpha}(x)$ of the two-sided simultaneous tolerance intervals with P% coverage and $(1-\alpha)100\%$ confidence, as developed for OLS regression by Lieberman and Miller [22] in the homoscedastic case:

$$y^{\pm}_{tol,1-\alpha}(x) = \hat{y}(x) \pm \left\{ (2F^{1-\alpha/2} s^2_{y/x})^{1/2} \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2} \right]^{1/2} + N(P) \left(\frac{n-2}{{}^{\alpha/2}\chi^2_{n-2}} s^2_{y/x} \right)^{1/2} \right\}, \quad (4)$$

where $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, $s^2_{y/x} = \sum (y_i - \hat{y}_i)^2/(n-2)$ is the variance of the residuals, $F^{1-\alpha/2}$ is the upper $(1-\alpha/2)$ percentile point of the F-distribution on 2 and $n-2$ degrees of freedom, ${}^{\alpha/2}\chi^2_{n-2}$ is the lower $\alpha/2$ percentile point of the χ^2-distribution with $n-2$ degrees of freedom, and $N(P)$ is the two-sided P percentile point of the unit normal distribution. This interval will cover the proportion P of the measurement population with the specified confidence level $(1-\alpha)$. Since the sum of the first two terms in the right-hand-side of Eq. (4) represents the limits of the $(1-\alpha/2)100\%$ confidence band about the calibration line, the nonparametric $(1-\alpha)100\%$ two-sided simultaneous tolerance intervals with P% coverage may be expressed in the following manner:

$$y^{\pm}_{tol,1-\alpha}(x) = y^{\pm}_{C,1-\alpha/2}(x) \pm N(P) \left(\frac{\nu}{{}^{\alpha/2}\chi^2_{\nu}} s^2 \right)^{1/2}, \quad (5)$$

where $y^{\pm}_{C,1-\alpha/2}(x)$ are the limits of the $(1-\alpha/2)100\%$ confidence band and ν is the number of degrees of freedom of the estimator s^2 of the measurement variance σ^2. It is evident that the latter term in Eq. (5) implies the widely acceptable hypothesis of normality of the experimental responses Y_i apart from any assumption on errors in x-axis.

In the homoscedastic case an estimate of the measurement variance σ^2 is robustly given by [28]:

$$s^2 = \left[\frac{med\{|e_i - \mathrm{med}\{e_i\}|\}}{0.6745} \right]^2 \quad (6)$$

where $e_i = y_i - \hat{\beta}_{0TS} - \hat{\beta}_{1TS} x_i$.

Eq. (6) provides a roughly unbiased estimator of the signal variance σ^2 if n is large and the errors normally distributed [28]. The

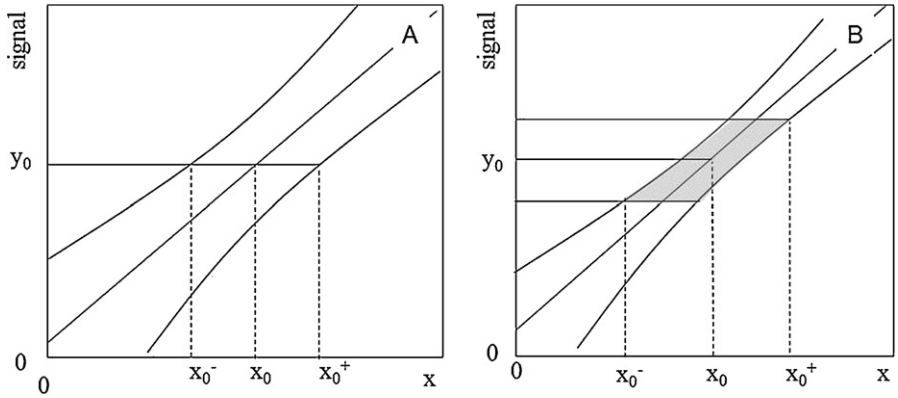

Fig. 2. Graphical determination of the discriminated value x_0 and of its confidence limits x_0^-, x_0^+ in correspondence to a response y_0 using (A) the *Method I*, (B) the *Method II*. The middle line is the TS calibration line; the bounding lines are (A) tolerance functions, (B) regression bands. (B) The regression bands, combined with the confidence interval of y_0, individuate the confidence interval of x_0.

former assumption is satisfied when repeated measurements are made at each x, and the latter one is valid in many experimental situations. The number of degrees of freedom ν may be conservatively taken equal to $n-2$ [30].

When the calibration data are heteroscedastic, the variance s^2 must be substituted for by the signal variance relevant to the concentration x, obtained by fitting the experimental variances at various concentration levels with a variance function. If m measurements are utilized at each concentration x for determining the variance function, the number of degrees of freedom of the modelled variances is $\nu = m-1-p$, where p is the number of the parameters of the variance model.

2.3.1. Uncertainty in the discriminated concentration

The determination of the $(1-\alpha)100\%$ confidence limits x_0^- and x_0^+ for a discriminated concentration x_0 may be performed graphically, intersecting the $(1-\alpha)100\%$ two-sided tolerance intervals, given by Eq. (5), with the line $y = y_0$, where y_0 is the experimental response of the unknown sample, and projecting the intersections onto the concentration axis (Fig. 2A). If y_0 is the average of k sample replicates, the described approach solves graphically the equations:

$$y_0 = y^+_{C,1-\alpha/2}(x_0^-) + N(P)\left(\frac{\nu}{\alpha/2\,\chi^2_\nu}\frac{s^2}{k}\right)^{1/2}, \tag{7a}$$

$$y_0 = y^-_{C,1-\alpha/2}(x_0^+) - N(P)\left(\frac{\nu}{\alpha/2\,\chi^2_\nu}\frac{s^2}{k}\right)^{1/2}. \tag{7b}$$

Since the use of the tolerance interval in the inverse regression can result too conservative, a valid alternative combines a $(1-\alpha/2)100\%$ confidence band of the entire calibration line with a $(1-\alpha/2)100\%$ confidence interval about the sample signal estimated as $y_0 \pm t_{(1-\alpha/4,\upsilon)}(s/\sqrt{k})$, where y_0 is the average of $k \geq 1$ signal replicates and $t_{(1-\alpha/4,\upsilon)}$ is the $(1-\alpha/4)100\%$ value of the t-distribution for ν degrees of freedom [24,29]. When s is the standard deviation of the k replicates the number of degrees of freedom ν is equal to $k-1$, whereas ν is $n-2$ if s is estimated by Eq. (6). The confidence interval limits x_0^- and x_0^+ about the discriminated concentration x_0 are obtained graphically by intersecting the sample confidence interval about y_0 with the confidence band of the calibration curve, and projecting the intersection area onto the concentration axis, as shown in Fig. 2B.

2.3.2. Detection limit

To calculate the detection limit the use of one-sided simultaneous tolerance intervals is more appropriate [25]. The $(1-\alpha)100\%$ one-sided simultaneous tolerance interval is calculated by the relationship

$$y^\pm_{tol,1-\alpha}(x) = y^\pm_{C,1-\alpha}(x) \pm N(P)\left(\frac{\nu}{\alpha\,\chi^2_\nu}s^2\right)^{1/2} \tag{8}$$

where $N(P)$ is the one-sided P percentile point of the unit normal distribution. The procedure to obtain the detection limit x_D with type I-error rate α and type II-error rate β is similar to that followed in the Hubaux–Vos approach [3]. The upper limit $y^+_{tol,1-\alpha}(0)$, that is the limiting value at zero concentration level, represents the "critical level" in the signal domain according to Currie [31] as using $(1-\alpha)100\%$ confidence 99% coverage tolerance, only one measurement per 100 is allowed to be significant with a false positive rate of $\alpha\%$. The projection onto the x-axis of the intersection of the straight line $y = y^+_{tol,1-\alpha}(0)$ with the lower $(1-\beta)100\%$ one-sided tolerance band gives the detection limit in the concentration domain [23].

2.4. Experimental

Three simulated data sets (data set 1–3) were generated using the Monte Carlo method in order to compare the features of the joint confidence regions based on TS–LQ, OLS, BLS and WLS regression techniques [32]. In all cases the theoretical values of the regression linear model coefficients were $\beta_0 = 0$ and $\beta_1 = 1$ and no errors were considered in the x-axis. For the subsequent evaluation of the TS procedure two experimental data sets (data sets 4, 5) were used.

2.4.1. Data set 1

A homoscedastic data set was obtained considering a standard deviation of 0.001 for y values. Nine x values were considered within the range 0–0.8, equally spaced and seven replicate responses were taken on y-axis at each level x.

2.4.2. Data set 2

A heteroscedastic data set was generated using the linear model $\sigma(x) = 0.001 + 0.1x$ for the standard deviation of the measurements.

2.4.3. Data set 3

This data set was obtained from data set 2 with outliers at level $x = 0.7$. The outliers were generated adding to each response a systematic contribute equal to four times the value of the standard deviation at $x = 0.7$.

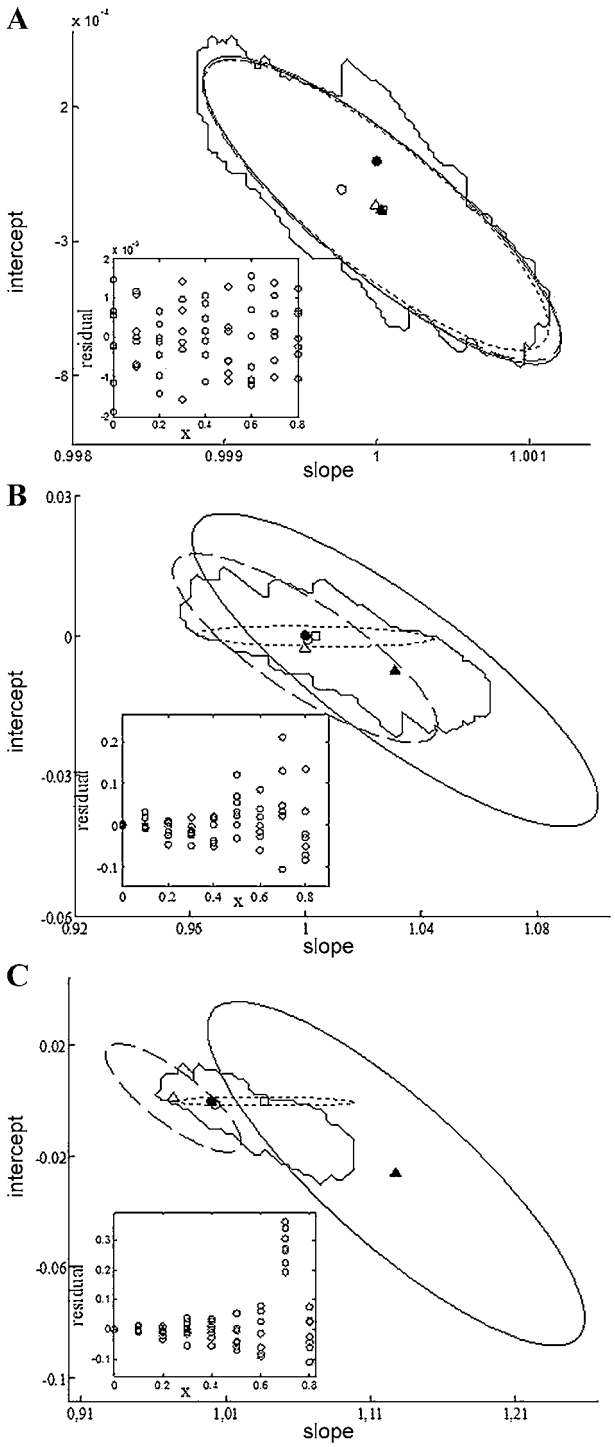

Fig. 3. 95% joint confidence region based on TS–LQ, OLS, IRLS, and WLS methods for the simulated data set 1 (A), 2 (B), 3 (C). TS–LQ, wiggly continuous line; OLS, continuous line; IRLS, broken line; WLS, dotted line. (●) Theoretical couple (1, 0) of the parameters of the regression line; estimated couples by TS (○), OLS (▲), IRLS (△), and WLS (□). Insets: residuals obtained from TS regression technique for the homoscedastic data set 1 (A), the heteroscedastic data set 2 (B), and the heteroscedastic with outliers data set 3 (C).

2.4.4. Data set 4

Experimental data regarding the analysis of volatile thiols at trace level in wines were used to construct the joint confidence regions in the application of the TS–LQ and the BLS approaches to compare the analytical procedures [11,33]. The thiols analyzed were: 3-mercaptohexan-1-ol (3-MH); 3-mercaptohexyl acetate (3-MHA) and the methods compared for both analytes were: purge and trap (P&T), headspace solid phase micro-extraction (HS-SPME), and solid phase micro-extraction (SPE) all followed by gas chromatographic–mass-spectrometry analysis (GC–MS).

2.4.5. Data set 5

The determination of the confidence interval of an unknown concentration and of the detection limit was done using GC–MS measurements of chloromethane in water in terms of peak area ratios of chloromethane and of the internal standard (fluorobenzene) [4]. The linear calibration model was used from 0 to 0.4 $\mu g\,L^{-1}$ and the calibration design provided five concentration levels replicated ten times. The experimental data appeared heteroscedastic in nature.

2.4.6. Weighting functions

The IRLS regression approach adopted uses iteratively re-weighted least-squares, with the weight functions of the *i*th residual at the *j*th iteration calculated as [2,20,34]:

$$w_i^{(j)} = \begin{cases} (1 - r_i^{2(j)})^2 & \text{for } |r_i^{(j)}| \le 1 \quad i = 1, 2, \ldots, n \\ 0 & \text{else} \end{cases},$$

where $r_i^{(j)} = (y_i - \hat{\beta}_0^{(j)} - \hat{\beta}_1^{(j)} x_i)/(4.695 \cdot s \cdot [1 - (1/n) - ((x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2)]^{1/2})$ and s is an estimate of the standard deviation of the errors in y-axis given by Eq. (6).

When heteroscedasticity is present, the weight factor w_i in the WLS regression is assumed equal to the inverse of the variance $s^2(x_i)$ at the concentration level x_i.

2.5. Statistical analysis

The Cochran test was used to check the scedasticity of the data sets considered [28]. The 5% significance level was used for all tests.

2.6. Software availability

The computations were done using home-made Matlab subroutines (Matlab for Microsoft Windows ver. 7.10, The MathWorks, Inc., Natick, MA).

3. Results and discussion

3.1. Qualitative analysis of the joint confidence regions

Fig. 3 shows the joint confidence regions obtained by applying the TS–LQ, OLS, IRLS, and WLS, regression techniques to three simulated data sets. The insets show the residuals obtained with the TS regression method as a function of the independent variable x, for illustrating the scedasticity and the presence of outliers in the considered data set. The TS–LQ confidence region is not exactly an ellipse but its irregular contour can be fitted by an ellipse. For each data set the comparison of the confidence regions is done in terms of shape, area, and position of the ellipses in the plane slope/intercept. The ellipse shape may be described by two factors, the ellipticity E, i.e. the ratio major axis/minor axis, and of the tilt measured by the angle γ° between the major axis and the slope axis. Table 1 shows the E and γ° values calculated with the following equations:

$$E = \left(\frac{\sum w_i (1 - \bar{x}_w \mathrm{tag}\,\gamma^\circ)^2 + D_w \mathrm{tag}^2\,\gamma^\circ}{\sum w_i (\bar{x}_w + \mathrm{tag}\,\gamma^\circ)^2 + D_w} \right)^{1/2},$$

$$\gamma = \frac{1}{2} \mathrm{arctag} \frac{2\bar{x}_w}{1 - \bar{x}_w^2 - D_w / \sum w_i},$$

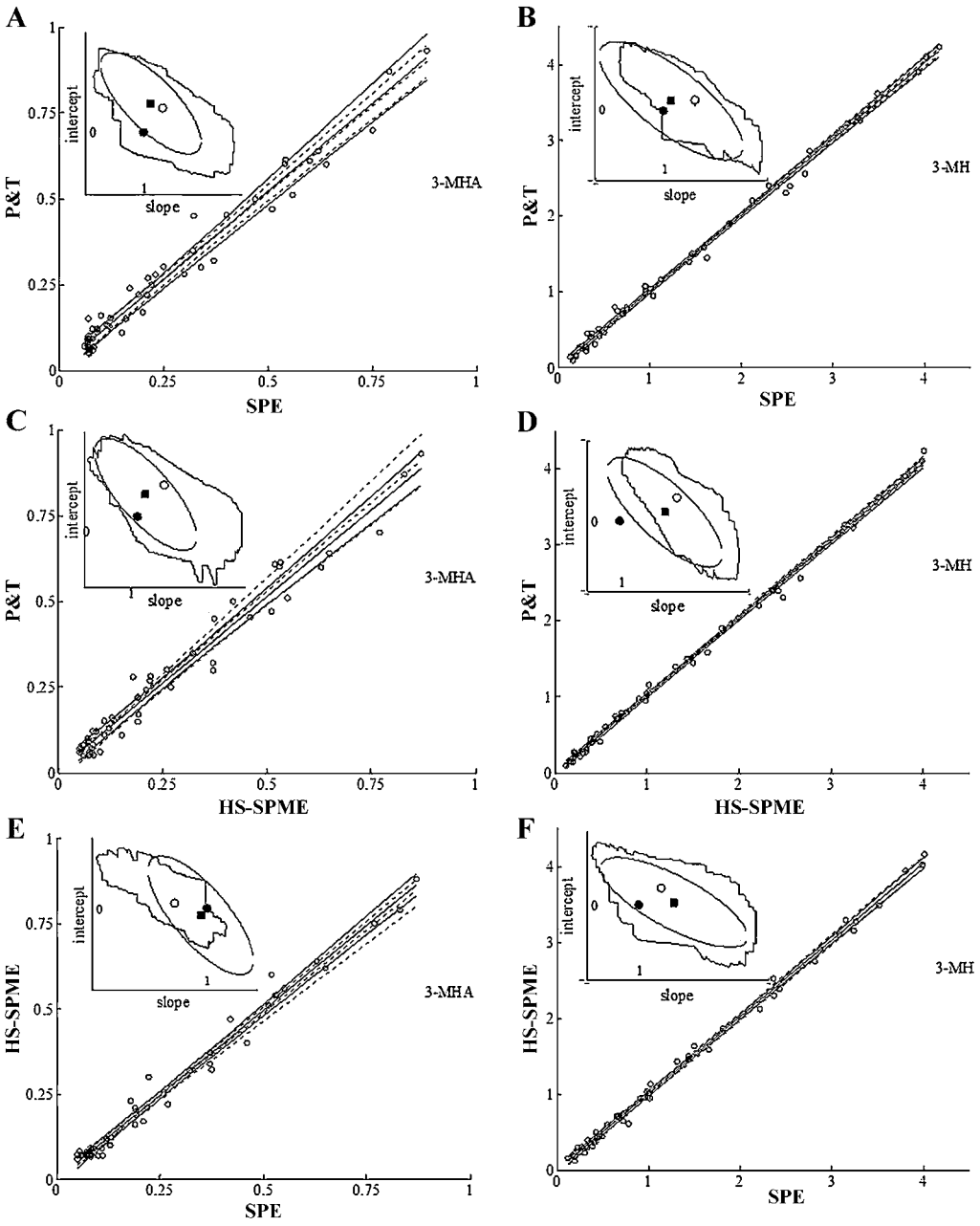

Fig. 4. Regression design comparison of P&T, HS-SPME, and SPE methods for sampling two volatile thiols (3-MHA and 3-MH) at trace level in wines for real data set 4 applying TS–LQ and BLS regression techniques. Continuous lines are calibration function and the 95% confidence band for the TS–LQ regression method; dotted lines are calibration function and the 95% confidence band for the BLS technique. Insets: TS–LQ and BLS 95% joint confidence regions, and slope and intercept obtained using TS (○), and BLS (■) regression approach. (●) Theoretical point (1, 0).

where w_i is the ith weight, $\bar{x}_w = \sum w_i x_i / \sum w_i$, and $D_w = \sum w_i (x_i - \bar{x}_w)^2$ for the OLS, WLS, and IRLS techniques. In the OLS case w_i is obviously equal to 1, in the WLS procedure w_i comes from the adopted linear model for the standard deviation of the measurements, while in the IRLS approach the weights of the final iteration were inserted in the formulas for the ellipticity and the tilt. The E and γ° values reported for the TS procedure were determined from the ellipses obtained fitting the contour of the joint confidence regions. For all the three data sets the shape of the TS–LQ region is quite similar to that of IRLS ellipse, as a consequence of the robustness of both methods. The WLS regression method, apart from the homoscedastic data set 1, which makes coincident OLS and WLS confidence ellipses, shows small γ° values, because small $\bar{x}_w$ values are produced by larger weights at low concentration levels. The increase of ellipticity is consequence of a better estimate of the intercept produced by the WLS regression method, which gives more importance to the less uncertain data near to zero concentration. The area of the joint confidence region depends on the uncertainty of the calibration line [27]. The confidence regions

Table 1
Ellipticity, E, and angle γ° (expressed in degree unit) between the major axis of the ellipse and the slope axis, relevant to TS–LQ, OLS, IRLS, and WLS 95% joint confidence regions for the data sets 1–3.

Data set	TS–LQ		OLS		IRLS		WLS	
	E	γ°	E	γ°	E	γ°	E	γ°
1	4.5	23	4.5	23	4.5	23	4.6	24
2	4.5	19	4.5	23	4.5	23	19	0.61
3	4.7	19	4.6	23	4.5	23	36	0.14

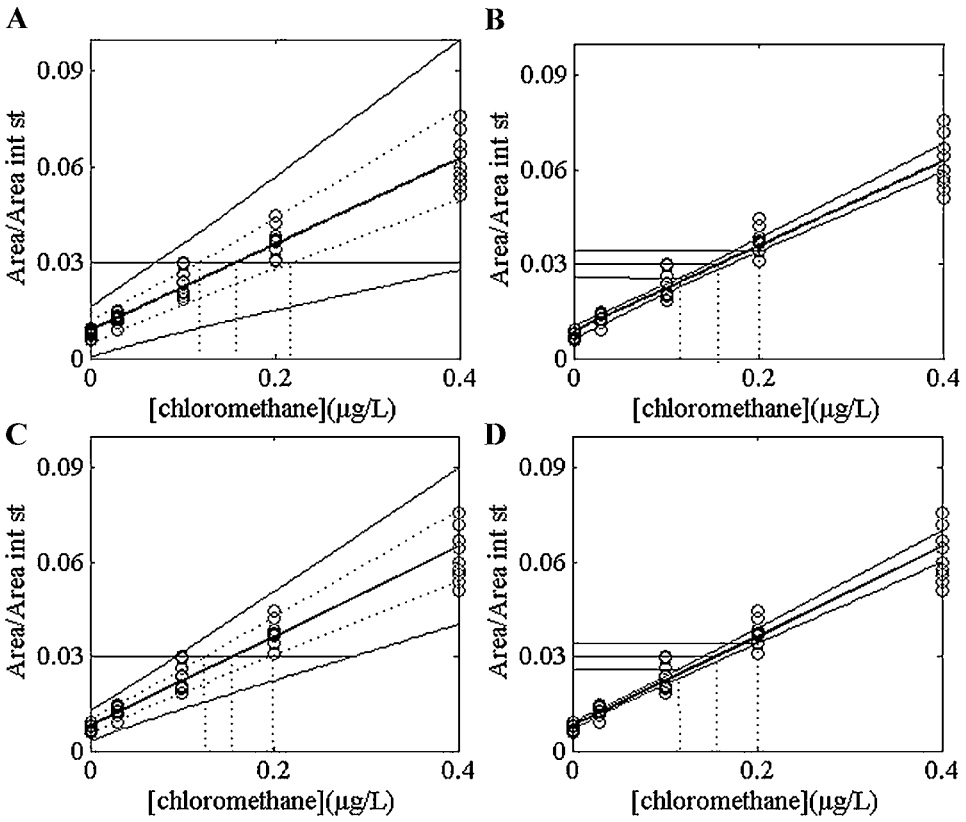

Fig. 5. Uncertainty graphical determination of the discriminated concentration value relevant to the signal value $y_0 = 0.03$ using the real data set 5. Figures A and B refer to the TS regression method ($y = 0.008 + 0.142X$, $R^2 = 0.998$); figures C and D refer to WLS regression ($y = 0.009 + 0.135X$, $R^2 = 0.997$). Figures A and C: *Method I*, 95% confidence limits x_0^-, x_0^+ coming from 95% confidence 99% coverage two-sided simultaneous tolerance intervals, $k = 1$ (continuous line) and $k = 10$ (dotted line) signal replicates. Figures B and D: *Mehod II*, combination of the 97.5% confidence band of the calibration curve with the 97.5% confidence interval for the sample signal ($k = 1$).

relevant to the homoscedastic data set 1 with no outliers are quite similar (Fig. 3A), whereas data sets 2 and 3 furnish WLS-ellipses with small area (see Fig. 3B and C), that indicates a major precision of the WLS method for the data sets considered. Further, for data sets 2 and 3, the TS–LQ and IRLS methods give quite similar areas of the respective confidence regions and smaller than those obtained with the OLS method. Finally, the position of the ellipses, given by the couple of the estimated parameters $(\hat{\beta}_1, \hat{\beta}_0)$ with respect to the theoretical couple (1, 0), indicates the bias of the regression method. In particular, the regression methods used furnish parameter values very close to each other and close to the theoretical one for the data sets 1 and 2, except the OLS method in the case of heteroscedasticity. For data set 3 the couple (1, 0) falls out of the OLS joint confidence region, as expected in the presence of heteroscedasticity and of outliers (Fig. 3C). Moreover, the TS–LQ method appears more robust than the IRLS procedure, whereas the WLS method is not affected by the presence of anomalous data with large uncertainty. This favorable characteristics of the WLS technique fails when the outlier responses have low variances.

3.2. Method comparison

Fig. 4 shows the results for the comparisons between couples of three different sampling methods of two thiols in fifty-two wine samples. On each graph the coordinates of the points are the concentrations obtained from the calibration curves relevant to the two sampling techniques compared. Besides to the TS and BLS regression lines, each figure shows the 95% confidence bands, and, in the insets, the two 95% joint confidence regions in the plane slope/intercept, the theoretical couple (1, 0), and the estimated couples of the regression parameters. The TS procedure, suitable to non normal x, y values, is compared to the BLS approach, which assumes normality of the errors in both axes. The validity of the use of BLS regression technique in method comparison dealing with discriminated concentrations lies on the assumption that these variables are approximately normal [31]. The variances of the discriminated concentrations used in the BLS regression were calculated applying the error propagation [1,4]. Table 2 summarizes the results of applying the TS–LQ and BLS regression approaches. Both regression methods give the same results with the exception of the analyte 3-MHA when the sampling methods P&T and SPE are compared (see Fig. 4A and first column of Table 2). In particular, the TS technique does not detect differences among the discriminated concentrations, whereas BLS does. These results indicate that the former technique is more conservative, as previously shown considering simulated data, where the areas of the confidence regions obtained for weighted regression approaches were generally smaller. This

Table 2
Regression-based comparison of three experimental sampling techniques (P&T, HS-SPME, SPE) of two analytes (3-MHA, 3-MH) in wines (data set 4). TS and BLS regression techniques were used to define the 95% joint confidence regions and to verify whether the theoretical point (1, 0) falls (yes) or does not fall (no) within these regions.

	P&T/SPE	P&T/HS-SPME	HS-SPME/SPE
Regression Method		3-MHA	
TS	Yes	Yes	Yes
BLS	No	Yes	Yes
		3-MH	
TS	Yes	No	Yes
BLS	Yes	No	Yes

Table 3
Discriminated x_0 ($\mu g\,L^{-1}$) value, with 95% confidence limits x_0^- ($\mu g\,L^{-1}$) and x_0^+ ($\mu g\,L^{-1}$), obtained with TS and WLS linear regression approaches using *Method I* and *Method II* for data set 5. The x_0 value comes from the single ($k=1$) or average ($k=10$) value $y_0=0.03$. Measurements refer to nominal chloromethane concentration of 0.15 $\mu g\,L^{-1}$.

		Method I		*Method II*	
		$k=1$	$k=10$	$k=1$	$k=10$
Regression method	x_0	(x_0^-, x_0^+)	(x_0^-, x_0^+)	(x_0^-, x_0^+)	(x_0^-, x_0^+)
TS	0.158	(0.070, 0.551)	(0.116, 0.220)	(0.117, 0.201)	(0.137, 0.180)
WLS	0.155	(0.093, 0.286)	(0.125, 0.197)	(0.116, 0.200)	(0.134, 0.179)

Table 4
Critical limit x_C ($\mu g\,L^{-1}$) and detection limit x_D ($\mu g\,L^{-1}$)for data set 5 obtained with TS and WLS linear regression approaches at 95% confidence, 99% coverage, one-sided tolerance interval.

TS		WLS	
x_C	x_D	x_C	x_D
0.041	0.130	0.028	0.073

feature gives to the TS technique a smaller resolving power in method comparison.

3.3. Uncertainty in the inverse regression

Fig. 5 shows the graphical determination of the 95% confidence interval of a discriminated value x_0 for a single or an average response $y_0=0.03$ when TS–LQ or WLS tolerance intervals (*Method I*) or the alternative method (*Method II*) are applied to the data set 5. Table 3 reports the relevant limits x_0^- and x_0^+ of the 95% confidence interval. The WLS technique was used to validate the TS–LQ approach in establishing the uncertainty in inverse regression since the experimental data set 5 fulfilled the hypotheses required by the weighted least squares procedure. As expected, the width of the confidence intervals of the discriminated concentration decreases increasing the number k of the unknown sample replicates. Asymmetry is present in the confidence intervals and *Method I* yields confidence intervals larger than *Method II*. Moreover, the nonparametric regression technique TS–LQ leads to large confidence intervals when the *Method I* is used, but quite similar to those obtained with WLS regression when the alternative *Method II* is applied. The conservative character of the TS–LQ methodology combined with the tolerance intervals and low number of replicate samples k may produce very large confidence intervals on the discriminated concentration, as shown in Table 3, where an upper limit appears out of the calibration range on the x-axis.

3.4. Detection limit

Table 4 shows the values of the critical and detection limits ($\alpha=0.05$, $\beta=0.05$) based on the tolerance intervals obtained with TS–LQ and WLS regression approaches using the data set 5. The differences between the values found using TS–LQ and WLS approaches may be easily explained considering the robustness of the nonparametric method together with the role of the decreasing variance at low concentration in the WLS technique.

4. Conclusions

The nonparametric Theil–Sen regression approach, characterized by no assumption regarding data statistics, was applied to solve three analytical problems, namely, method comparison, estimation of the confidence interval for a concentration obtained in inverse regression, and determination of the detection limit. Method comparison was performed via construction of the joint confidence region for the regression coefficients, which may contain or not the theoretical point defined by unity slope and zero intercept. The joint confidence region obtained from the statistical Lancaster–Quade methodology appears ellipse-like and therefore is easily comparable to the ellipses typical of the well known OLS/WLS/BLS regression techniques. Determination of the concentration uncertainty in the inverse regression was achieved either using the simultaneous tolerance intervals (*Method I*) or an alternative method (*Method II*), where the confidence band of the calibration line matches the confidence interval of the sample signal. In both methods the presence of errors in both axes x and y was taken into account to construct the confidence band for the calibration curve, whereas the assumption of normality in the signal domain is inserted as added term (*Method I*) or treated in a separate way (*Method II*). The limits of the confidence intervals using TS–LQ indicate that the Theil–Sen approach, even being of more general use, furnishes analogous results of the WLS technique when heteroscedastic real data sets are considered and the alternative procedure (*Method II*) is applied. A larger detection limit was obtained for the nonparametric approach owing to the presence of two opposite effects in the two compared methods, namely, the robustness of the TS procedure, producing an enlargement of the tolerance interval, and a decreased width of the WLS tolerance interval, produced by lower variances at low concentration levels. Anyway, this drawback does not affect the advantage achieved by the use of a unique, robust regression technique in quite different analytical problems.

Acknowledgement

The financial support of the Italian Ministry for Universities and Research (MIUR) is gratefully acknowledged.

References

[1] J.N. Miller, Analyst 116 (1991) 3.
[2] R.W. Holland, R.E. Welsch, Commun. Stat. A: Theory Methods 6 (1977) 813.
[3] A. Hubaux, G. Vos, Anal. Chem. 42 (1970) 849.
[4] I. Lavagnini, F. Magno, Mass Spectrom. Rev. 26 (2007) 1.
[5] J. Mandel, The Statistical Analysis of Experimental Data, Wiley, New York, 1967.
[6] J.M. Lisy, A. Cholvadova, J. Kutej, Comput. Chem. 14 (1990) 189.
[7] J. Riu, F.X. Rius, Anal. Chem. 68 (1996) 1851.
[8] F.J. del Río, J. Riu, F.X. Rius, J. Chemometr. 15 (2001) 773.
[9] D.L. Massart, B.G.M. Vandegiste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.
[10] L.M. Schwarz, Anal. Chem. 48 (1976) 2287.
[11] B. Fedrizzi, G. Versini, I. Lavagnini, D. Badocco, G. Nicolini, F. Magno, Anal. Chim. Acta 621 (2008) 38.
[12] H. Theil, Proc. K. Ned. Akad. Van Wetens. 53 (1950) 386.
[13] P.K. Sen, J. Am. Stat. Assoc. 63 (1968) 1379.
[14] P. Sprent, Applied Nonparametric Statistical Methods, CRC Press, New York, 1993.
[15] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, Wiley, New York, 1999.
[16] P.J. Rousseuw, A.M. Leroy, Robust regression and Outlier Detection, Wiley, New York, 1987.
[17] L. Glasser, J. Chem. Educ. 84 (2007) 533.
[18] E.J. Dietz, Am. Stat. 43 (1989) 35.
[19] A.F. Siegel, Biometrika 69 (1982) 242.
[20] R. Rajkó, Anal. Lett. 27 (1994) 215.
[21] J.F. Lancaster, D. Quade, J. Am. Stat. Assoc. 80 (1985) 393.
[22] G.J. Lieberman, R.G. Miller, Biometrika 50 (1963) 155.
[23] R.D. Gibbons, Statistical Methods for Groundwater Monitoring, McGrawHill, New York, 1994.
[24] J.S. Garden, D.G. Mitchell, W.N. Mills, Anal. Chem. 52 (1980) 2310.

[25] M.E. Zorn, R.D. Gibbons, W.C. Sonzogni, Anal. Chem. 69 (1997) 3069.
[26] X. Wang, J. Nonparametr. Stat. 17 (2005) 107.
[27] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, Wiley, New York, 1960.
[28] N.R. Draper, H. Smith, Applied Regression Analysis, Wiley, New York, 1998.
[29] R.G. Miller, Simultaneous Statistical Inference, McGrawHill, New York, 1966.
[30] A.M. Gross, J. Am. Stat. Assoc. 72 (1977) 341.
[31] L.A. Currie, Chemometr. Intell. Lab. Syst. 37 (1997) 151.
[32] P.C. Meier, R.E. Zünd, Statistical Methods in Analytical Chemistry, Wiley, New York, 1993.
[33] B. Fedrizzi, G. Versini, I. Lavagnini, G. Nicolini, F. Magno, Anal. Chim. Acta 596 (2007) 291.
[34] A.E. Beaton, J.W. Tukey, Technometrics 16 (1974) 147.