Vol. 17, No. 2 (2018) 507-522
Revista Mexicana de Ingeniería Química

A NEW ONLINE COMPUTER PROGRAM (BIDASYS) FOR ORDINARY AND UNCERTAINTY WEIGHTED LEAST-SQUARES LINEAR REGRESSIONS: CASE STUDIES FROM FOOD CHEMISTRY

NUEVO PROGRAMA COMPUTACIONAL EN LÍNEA (BIDASYS) PARA LAS REGRESIONES LINEALES ORDINARIA Y PONDERADA CON INCERTIDUMBRES: ESTUDIO DE CASOS DE QUÍMICA DE ALIMENTOS

M. Rosales-Rivera[1], L. Díaz-González[2*], S.P. Verma[3]

[1]*Doctorado en Ciencias, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma de Estado de Morelos, Av. Universidad 1001, Chamilpa, Cuernavaca, Mor. 62209, México.*

[2]*Departamento de Computación, Centro de Investigación en Ciencias, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma de Estado de Morelos, Av. Universidad 1001, Chamilpa, Cuernavaca, Mor. 62209, México.*

[3]*Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Priv. Xochicalco s/no., Col. Centro, Temixco, Mor. 62580, México.*

Received November 2, 2017; Accepted January 15, 2018

Abstract

A new computer program BiDASys (Bivariate Data Analysis System) is presented for the application of Ordinary and Uncertainty weighted least-squares linear regression models (OLR and UWLR) to experimental data from food chemistry. BiDASys has the following novel aspects: the statistical capability of detecting discordant outliers in bivariate data; new simulated critical values through Monte Carlo for the probability of no-correlation in multivariate samples (n=5-1000); and it is the only available program that can applied the UWLR model. The use of BiDASys is illustrated through three case studies where the relations $^{87}Sr/^{86}Sr_{Must}$ versus $^{87}Sr/^{86}Sr_{soil}$ (from Glera-Prosecco, Italy), $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine,Soil}$ (from Quebec, Canada), and $^{87}Sr/^{86}Sr_{Wines}$ versus $^{87}Sr/^{86}Sr_{Rocks}$ (from Tuscany-Basilicata, Italy) confirms that this isotopic ratio can be used to track the geographical origin of wine and one more case study from Guerrero (Mexico) concerning the influence of breastfeeding time on levels of organochlorine pesticides in human milk.

Keywords: ordinary least-squares linear regression, uncertainty weighted least-squares linear regression, discordancy tests, food chemistry, isotopes, applied statistics.

Resumen

Un nuevo programa BiDASys (Bivariate Data Analysis System) es presentado para la aplicación de los modelos de regresión lineal ordinaria y ponderada con incertidumbres (OLR y UWLR) a datos experimentales de química de alimentos. BiDASys tiene los siguientes aspectos novedosos: capacidad estadística de detectar valores discordantes en datos bivariados; nuevos valores críticos simulados mediante Monte Carlo para la probabilidad de no-correlación en muestras multivariadas (n=5-1000); y es el único programa disponible que aplica el modelo UWLR. El uso de BiDASys es ilustrado a través de tres estudios de caso donde las relaciones $^{87}Sr/^{86}Sr_{Must}$ versus $^{87}Sr/^{86}Sr_{soil}$ (de Glera-Prosecco, Italia), $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine,Soil}$ (de Quebec, Canada) y $^{87}Sr/^{86}Sr_{Wines}$ versus $^{87}Sr/^{86}Sr_{Rocks}$ (de Tuscania-Basilicata, Italia), confirman que esta relación isotópica puede utilizarse para rastrear el origen geográfico del vino y un estudio de caso de Guerrero (México) sobre la influencia del tiempo de lactancia en los niveles de pesticidas organoclorados en leche humana.

Palabras clave: regresión lineal ordinaria, regresión lineal ponderada, pruebas de discordancia, química de alimentos, isótopos, estadística aplicada.

* *Corresponding author. E-mail*: ldg@uaem.mx

doi: 10.24275/10.24275/uam/izt/dcbi/revmexingquim/2018v17n2/Rosales
issn-e: 2395-8472

1 Introduction

In order to explore the relationship between a dependent and an independent variable, linear regression analysis is widely used to achieve valid inferences from experimental data in all scientific and engineering fields (e.g., Draper and Smith, 1998; Chatterjee and Hadi, 2015; Darlington and Hayes, 2016). Many techniques for carrying out regression analysis have been proposed that vary from the conventional ordinary least-squares linear regression (OLR) to weighted least-squares linear regression models (WLR; Barnett and Lewis, 1994; Bevington and Robinson, 2003; Guevara *et al*., 2005; Verma, 2016). Recently, Nunes *et al*. (2015) showed that the use of statistical methods in food science and technology has increased considerably, which implies the importance and need of developing accurate statistical techniques in such applications. Nunes *et al*. (2015) also stated that these statistical methods have been implemented in computational programs usually divided into univariate (graphic analysis and descriptive statistics), bivariate (correlation and linear regression analysis), and multivariate methods (exploratory and classification methods). Strict requirements of international scientific journals allied to the need to correctly interpret the experimental data from the statistical standpoint, have led to a steep increase in the use and development of statistical software. However, many researchers have difficulties in understanding and interpreting crucial statistical concepts and the use of statistical methods for interpreting experimental data (Passari *et al*., 2011; Granato and Calado, 2014; Cozzolino, 2014). Therefore, there is a need of developing better software that would incorporate newer well-tested statistical procedures for handling of experimental data.

Furthermore, in most scientific and engineering fields, the original complete data sets are seldom reported, and only the statistical summary data are generally available. This is a major problem for a correct statistical analysis of experimental data and above all, for ascertaining if the original data were correctly summarized from the statistical point of view, i.e., if the central tendency and dispersion parameters were properly obtained (Barnett and Lewis, 1994). On the other hand, a well-established fact of the scientific research is that when an experiment is performed for the first time, the results often bear too little resemblance to the "truth" being sought (Barnett and Lewis, 1994; Bevington and Robinson, 2003). For all experimental data, random errors and related uncertainties undoubtedly exist, which should be estimated and then reduced by improved experimental techniques and repeated measurements. These errors must always be estimated to establish the validity of the results and interpretation. Thus, error (such as standard deviation) or uncertainty (confidence limits of the mean) estimates on the individual data should actually be reported (e.g., Pope, 1976; Chatterjee and Hadi, 1986, 2015; Barnett and Lewis, 1994; Granato *et al*., 2014; Verma, 2016).

Chemometric techniques have become an important tool in modern analytical chemistry for designing experiments, achieving instrumental calibrations, and ascertaining the quality (precision, accuracy, and uncertainties) of analyses in a wide variety of chemical matrices (Miller and Miller, 2010; Verma, 2016). In food chemistry, many authors (e.g., Raco *et al*., 2015; Fernández *et al*., 2016; Giordani *et al*., 2016; Maslak and Nimmermark, 2017; Foereid, 2017) have used linear regression models. Furthermore, some other publications where the OLR model was applied are as follows: Moreno-Rivas *et al*. (2016) in biosorption of cadmium from aqueous solution by baker's yeast; Borges *et al*. (2017) in analysis of buffalo milk; Gómez-Favela *et al*. (2017) in modelling of water absorption in chickpea seeds; Chaparro *et al*. (2017) in selection of the process parameters and drying protectant to granulated bio-products based on microorganisms; Fuentes-Ortega *et al*. (2017) in study of the process variables of microencapsulation sesame oil by spray drying; Pérez-Grijalba *et al*. (2017) in study of the bio-functionality properties of blackberry juice; and Xu *et al*. (2017) in seasonal and annual variations of atmospheric Hg and Pb isotopes in Xi'an, China. Unfortunately, in such applications most authors do not report or partially report the original data, nor they estimate the individual errors or uncertainties.

The feasibility of reporting errors on individual data will certainly lead in future to a better use of the regression models. Verma (2012) suggested that the use of uncertainty-based weighted least-squares linear regression (UWLR) model should be preferred over other WLR models, because the UWLR carries the connotation of probability or confidence limits. Thus, the use of the UWLR that estimates the total uncertainty for each data point should be considered as the best suitable statistical method for a better analysis

of experimental data (Verma, 2016).

In this work, we present applications of linear regression models in food chemistry. For this purpose, we have developed an online computer program (Bivariate Data Analysis System - BiDASys) for an efficient application of the OLR and UWLR regression models. The discordant outliers may also occur in the linear regressions and should be best handled using proper statistical techniques. BiDASys program allows the application of recursive discordancy tests (Verma *et al*., 2017b), to univariate statistical samples constituted by the studentized residuals (Barnett and Lewis, 1994) for discordant outlier detection and separation. Furthermore, BiDASys can also help us to compare the results from OLR and UWLR models as illustrated in this work.

2 Materials and methods

2.1 Linear regressions

Different types of simple (ordinary) and weighted models are commonly used to explore the relationships between an independent variable (x) and a dependent variable (y). These methods are applied to the data set with the main purpose of obtaining the best linear fit and providing a visual demonstration of the relationship between the data points (Miller and Miller, 2010; Cozzolino, 2014).

2.1.1 Ordinary least-squares Linear Regression (OLR) model

The most common application of the OLR method aims to create a straight line that minimizes the sum of the squares of the errors (or residuals) generated by the associated equation, as a result of the differences in the observed value and the value anticipated based on the model. Equation (1) describes this model for two variables x and y, where a is the intercept term, b is the slope, and s_a and s_b are their respective standard errors.

$$y = a(\pm s_a) + [b(\pm s_b) * x] \quad (1)$$

However, for the OLR model to be statistically valid, certain assumptions must be fulfilled (Miller and Miller, 2010;): (a) linearity between y and x variables; (b) x is error-free or < 1/10 of the error in y; (c) errors in y are normally distributed; (d) homoscedastic errors in y (constant variance across the entire response range); and (e) errors associated with different observations are independent. Rarely, all assumptions are fulfilled in a given experimental study. Therefore, the OLR models are in general invalid and more sophisticated regression models are required.

2.1.2 Weighted Least-squares Regression (WLR) models

These models are required because not all assumptions for the OLR are fulfilled (e.g. errors from experimental studies do not present a homoscedastic behaviour). These models assign different weights to the data points as an inverse function of the corresponding variances. The York (York, 1968) model, widely used in Earth Sciences, was proposed for isotope data in geochronology, under the assumption of correlated errors in x and y. New York (Mahon, 1996) method was an improvement of older York model. Equation (2) describes the basic WLR model, where symbols are the same as for the OLR and the subscript w refers to the weighted regression. Equation (2) is as follows:

$$y = a_w(\pm s_a w) + [b_w(\pm s_b w) * x] \quad (2)$$

2.1.3 Uncertainty Weighted least-squares Linear Regression (UWLR) model

UWLR is a new weighted linear regression procedure based on total uncertainty estimates, which is considered a better alternative because the use of uncertainty has a probability connotation, here a strict confidence level of 99% (or equivalently, significance level of 1%; Verma, 2016). Before the calculation of the central tendency (e.g. mean $\overline{x}$) and dispersion (e.g. standard deviation s) parameters for univariate samples, it should be mandatory to ascertain that all replicate measurements be free from discordant outliers (Verma *et al*., 2016, 2017a), which can be easily confirmed by computer program UDASys2 (Univariate Data Analysis System 2; Verma *et al*., 2017b). After this statistical procedure, the uncertainty in the *ith* sample x (or y) is calculated as follows:

$$u = (s/\sqrt{n}) * t_{(n-1)} \quad (3)$$

where $t_{(n-1)}$ is the Student t critical value for $(n-1)$ degrees of freedom for the desired confidence level (generally 99% or 95%, two-sided), or significance level of 1% or 5% (α of 0.01 or 0.05). For both samples, the chosen confidence or significance level should be the same. In this way, we have n values of variables x and y with their respective uncertainties

and therefore, we can use the following equation:

$$y = a_w(\pm u_a w) + [b_w(\pm u_b w) * x] \tag{4}$$

where a_w is the weighted intercept term, b_w is the weighted slope and u_{aw} and u_{bw}, are their respective uncertainties (Verma, 2016).

The basic idea from this variation is to handle all the data with uncertainties, which helps to obtain a better interpretation of the relationship from the independent (x) and dependent variable (y). No computer programs seem to be yet available to carry out the UWLR regression. For this reason, we developed a new program BiDASys that applies both OLR and UWLR regression models to datasets. In addition to the regression parameters (intercept, slope, and their respective standard errors or uncertainties) both before and after applying discordancy tests, the linear correlation coefficient r and probability of no-correlation the $P_c(r;n)$ criterion, can also be calculated from our program BiDASys.

2.1.4 *Monte Carlo simulation for the new Pearson correlation coefficient critical values*

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in statistical analysis is the Pearson correlation (r), which is widely used as a measure of the statistical relationship, or association between two continuous variables (Bevington and Robinson, 2003; Cozzolino, 2014; Darlington and Hayes, 2016). We can calculate the Pearson correlation coefficient using the following equation:

$$r = \frac{cov_{x,y}}{\sqrt{s_x^2 * s_y^2}} \tag{5}$$

where $cov_{x,y}$ and $\sqrt{s_x^2 * s_y^2}$ represent the covariance and the square root of the multiplication of the variances of x and y, respectively.

When a linear correlation is statistically significant (e.g. if the linear correlation coefficient, r has a very low probability of no-correlation $Pc(r;n)$ value; Bevington and Robinson, 2003), this relationship can be used to interpret the data and to infer about natural processes. Commercial or freely available software (e.g., SPSS®, JASP, Statistica®, PSPP, SAS, NCSS®) are capable of reporting the r value. However, most users are not familiar with the statistical interpretation of this relationship.

As a part of this work, the required Pearson correlation coefficient critical values for different sample sizes, were newly simulated from samples sizes n_{min}(1)100(5)200(50)500(100)1000, using Monte Carlo procedure (Verma *et al*., 2017b). For the construction of multivariate samples with distribution $N(\mu,\sigma)$, we generated random normal variables Z_i with normal distribution $N(0,1)$ which were used to create the desired bivariate normal vector X as follows (Law and Kelton, 2000):

$$X_i = \mu_i + \sum_{j=1}^{i} c_{ij} Z_j \tag{6}$$

We calculated the Pearson correlation values for several significance levels, using 100,000 repetitions and 190 independent simulation experiments (Verma *et al*., 2017b). The complete table (Pearson correlation (r) values.xlsx) containing the linear correlation coefficient r versus the number of observations n and the corresponding $Pc(r;n)$ value for a large number of significance levels along with the total uncertainty from the Monte Carlo simulations are presented in the supplementary information (Table S1 and S2). Finally, the functional dependence of r corresponding to representative values of $Pc(r;n)$ is plotted on a semi-logarithmic scale as a smooth variation with the number of observations n from 3 to 1000 are presented in Figure 1.

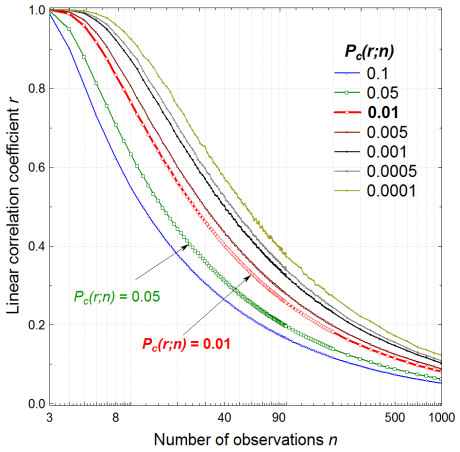

Fig. 1. The linear-correlation coefficient r versus the number of observations n and the corresponding probability $Pc(r;n)$ that the variables are not correlated. The $Pc(r;n)$ values for confidence levels from 90% to 99.99% (equivalent to significance levels from 0.1 to 0.0001) for samples of sizes $n = 3 - 1000$; the green curve corresponds to the $Pc(r;n)$ values for a confidence (or significance) level of 95% (or 0.05; two-sided) whereas the red curve is for the confidence (or significance) level of 99% (or 0.01; two-sided).

2.2 *Statistical procedure for discordant outlier detection and separation*

Under the outlier-based scheme, in univariate samples it should be mandatory as a prior step the identification and separation of discordant outliers before any statistical parameter estimation. It is known that the discordancy tests are used to detect discordant outliers in univariate data to assess the assumption of normality in experimental data, and to enable us to calculate the central tendency (mean) and dispersion (standard deviation) parameters from a set of normally distributed observations (Verma *et al*., 2017b). Discordant outliers may also occur in bivariate samples and should be best handled using proper statistical techniques (Barnett and Lewis, 1994; Bevington and Robinson, 2003). The examination of least-squares residuals for the detection of discordant outliers is one of the most important and effective means for the quality control. There are two types of outliers, those in the response variable (y) which represents model failure, and those with respect to the predictors (x); they can seriously affect the regression model. Graphical methods based on residuals alone will fail to objectively detect these unusual data points (Velleman and Welsch, 1981; Atkinson, 1981; Draper and Smith, 1998; Chatterjee and Hadi, 1986, 2015).

The estimated residuals for the simple regression model do not have constant variance, because the residuals are assumed to come from a common distribution $\sim N(0,\sigma^2)$). One possible approach to apply the univariate discordancy tests is to examine the appropriately weighted estimated residuals (studentized residuals) as suggested and achieved by several workers (e.g., Pope, 1976; Barnett and Lewis, 1994; Bevington and Robinson, 2003; Chatterjee and Hadi, 2015).

For these reasons, in this work we calculated the studentized residuals for both linear regression models (OLR and UWLR) using the following equation (7) proposed by Barnett and Lewis (1994):

$$sr_j = \left| \frac{r_j}{\sqrt{\frac{\sum r_j^2}{n-2}}\sqrt{\left(1-\frac{1}{n}\right)-\left(\frac{(x_j-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}} \right| \qquad (7)$$

where r_j are the residuals calculated by each linear regression, x_j and $\overline{x}$ are each individual value and the mean from the sample x.

Afterwards, we applied 5 recursive discordancy tests (Verma *et al*., 2017b) with the highest detection power and with the lowest skewness and masking effects, to detect possible discordant outliers in the studentized residuals from bivariate samples, which is another novel aspect of our software.

2.3 *Computer program*

For an efficient application of our statistical methodology proposed for bivariate samples, a computer program BiDASys (Bivariate Data Analysis System) was written in Java Framework ZK (Figure 2). The first part of the program concerns the data validation and activation of the possible regression model that can be applied to an “appropriate" dataset. The input file must have a predetermined format (see supplementary information “Readme" for more details about template).

If the program does not report any errors in the input data file, the user can advance to the next step by selecting one option of the *Regression Analysis* menu, which contains the following options: "Ordinary least-squares Linear Regression - OLR", "Uncertainty Weighted last-squares Linear Regression - UWLR", and "Recommended procedure". After processing all data for OLR, UWLR or both types of regressions ("Recommended procedure" option), optionally, the discordant outlier detection module calculates the studentized residuals (equation 7) for possible discordant outliers. If no discordant outliers were found, the program proceeds to save the results. If any observation is detected as discordant outlier, this is separated, and the regression process is repeated until it does not find any more discordant outliers. Both sets of results (e.g., intercept, slope, and their respective uncertainties, value of r (and R^2), and the $Pc(r;n)$ criterion) and graphics - for input data with and without outliers - are then presented in output files. This information can be easily downloaded by the user. In the report generated by BiDASys, the discordant outlier pairs detected will show an "*" (see "Readme" for more detailed information).

BiDASys has some advantages as compared to the other available software. The user can choose if they want to apply the recursive tests automatically to the studentized residuals calculated by the linear regression models (OLR or UWLR) or if they want only the results for each regression model. In addition, from our "Recommended procedure", the program applies both models to the data and generates the results in one output file, providing the data points to create the graphics with their respective uncertainties and the linear regression line.

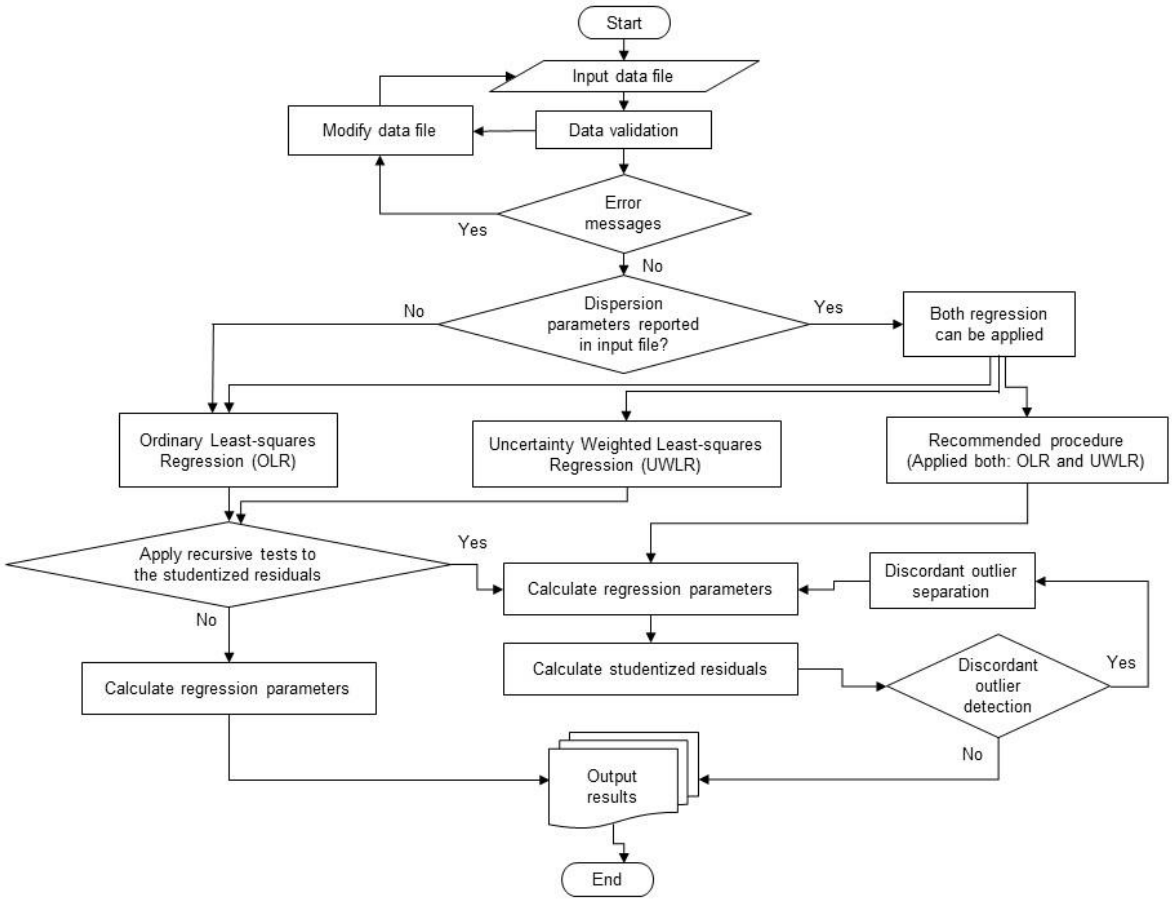

Fig. 2. Schematic flow diagram for the computer program BiDASys.

No other commercial or freely available software (e.g., SPSS®, JASP, Statistica®, PSPP, SAS, NCSS®) is capable of applying the UWLR model and this methodology in such an automatic way.

2.4 *Program availability*

This program, the input data template file, document "Readme", the complete tables of the Pearson correlation (r) values are available from our website http://tlaloc.ier.unam.mx/, after a previous registration onto the server. Once log in, the user can access BiDASys at http://tlaloc.ier.unam.mx/BiDASys, which will be available in the "Online programs" menu (for more details about how to get access to our server, see Readme in the supplementary information).

3 Results

3.1 *Application of BiDASys to food chemistry cases*

The use of this program is illustrated through four examples from different food chemistry case studies. We show the results for the application of both models (OLR and UWLR) and the improved interpretation obtained by using the UWLR instead of the OLR. The results presented in this work were obtained from our statistical procedure applied at a strict confidence level of 99% or significance level of 0.01 (two-sided), or equivalently, at 99.5% or 0.005% (one-sided), respectively. In addition, the original authors only applied the linear regression (OLR for all cases), but did not mention the program used for this application.

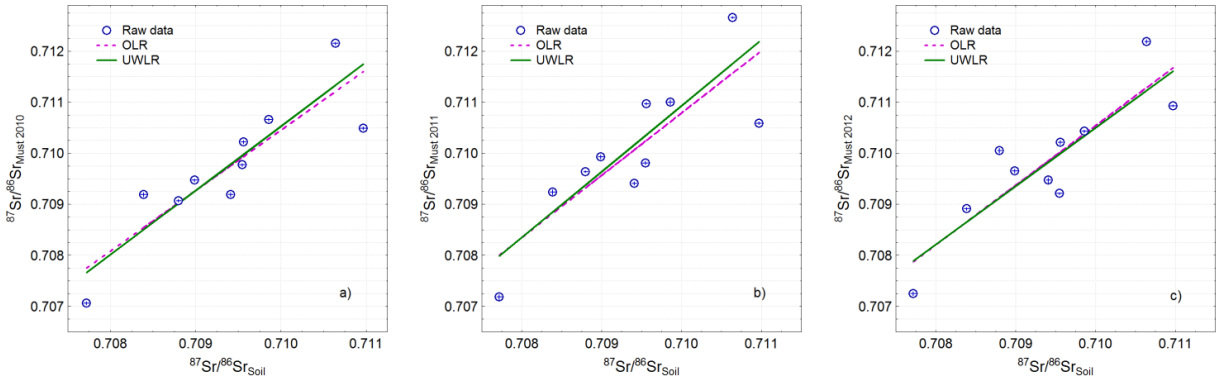

Fig. 3. $^{87}Sr/^{86}Sr_{Must}$s versus $^{87}Sr/^{86}Sr_{soil}$ for Glera vineyards, Veneto Region, Italy (Application Study A1; Petrini *et al*., 2015). a) OLR and UWLR results from the original data for $^{87}Sr/^{86}Sr_{Must}2010$ versus $^{87}Sr/^{86}Sr_{soil}$; and b) OLR and UWLR results from the original data for $^{87}Sr/^{86}Sr_{Must}2011$ versus $^{87}Sr/^{86}Sr_{soil}$; and, c) OLR and UWLR results from the original data for $^{87}Sr/^{86}Sr_{Must}2012$ versus $^{87}Sr/^{86}Sr_{soil}$.

Table 1. Regression parameters and equations from Sr-isotopic compositions of musts versus soil labile fraction (Application Study A1; Petrini *et al*., 2015).

Regression model	*n*	Regression parameters							Equation	
		a	u_a	*b*	u_b	*r*	R^2	*Pc(r;n)*		
$^{87}Sr/^{86}Sr_{Must2010}$ versus $^{87}Sr/^{86}Sr_{Soil}$										
OLR_{lit}	10	-0.133	NR	1.188	NR	NR	0.77	NR	$y = -0.133 + 1.188 * x$	(8)
OLR_{tw}	10	-0.131	0.550	1.186	0.770	0.8776	0.7701	0.0005	$y = -0.131 (\pm 0.550) + 1.186 (\pm 0.770) * x$	(9)
UWLR	10	-0.181	0.163	1.256	0.230	0.8779	0.7706	0.0004	$y = -0.181 (\pm 0.163) + 1.256 (\pm 0.230) * x$	(10)
$^{87}Sr/^{86}Sr_{Must2011}$ versus $^{87}Sr/^{86}Sr_{Soil}$										
OLR_{lit}	10	-0.161	NR	1.227	NR	NR	0.71	NR	$y = -0.161 + 1.227 * x$	(11)
OLR_{tw}	10	-0.158	0.670	1.223	0.945	0.8381	0.7023	0.0013	$y = -0.158 (\pm 0.670) + 1.223 (\pm 0.945) * x$	(12)
UWLR	10	-0.207	0.201	1.292	0.283	0.8408	0.7069	0.0012	$y = -0.207 (\pm 0.201) + 1.292 (\pm 0.283) * x$	(13)
$^{87}Sr/^{86}Sr_{Must2012}$ versus $^{87}Sr/^{86}Sr_{Soil}$										
OLR_{lit}	10	-0.127	NR	1.18	NR	NR	0.78	NR	$y = -0.127 + 1.18 * x$	(14)
OLR_{tw}	10	-0.120	0.541	1.170	0.763	0.8765	0.7682	0.0005	$y = -0.120 (\pm 0.541) + 1.170 (\pm 0.763) * x$	(15)
UWLR	10	-0.102	0.161	1.144	0.226	0.8769	0.7690	0.0005	$y = -0.102 (\pm 0.161) + 1.144 (\pm 0.226) * x$	(16)

The nomenclature used in this table are as follows: OLR_{lit} – ordinary least-squares regression results from Petrini *et al.* (2015); OLR_{tw} – ordinary least-squares regression results obtained in this work; UWLR – uncertainty weighted least-squares regression results obtained in this work; *n* – samples size; *a* – intercept; *b* – slope; *u* – uncertainty at a strict confidence level of 99% two-sided, equivalent to 99.5% one-sided; *r* – correlation coefficient; R^2 – correlation coefficient; *Pc(r;n)* – Probability of no correlation ; NR – Not Reported.

3.1.1 Application Study A1: Glera vineyards, Veneto Region, Italy

The Sr-isotopic systematics was applied to soils and musts from the 2010, 2011 and 2012 vintages in ten distinct Prosecco vineyard farms in the Veneto Region (Italy), which produce 100% Glera Prosecco wine. The aim of the study by Petrini *et al*. (2015) was to test the applicability of the Sr-isotopic method to the Prosecco geographic traceability. In order to further quantify these observations and highlight the possible relationships between soils and the corresponding musts, they applied a statistical approach.

In the cumulative frequency data retrieving, soils and musts belonging to the 2010, 2011 and 2012 harvests follow a normal distribution, which was confirmed by the original authors using statistical tests for normality (Anderson-Darling, Kolmogorov-Smirnoff and similar tests). A linear correlation was modelled by Petrini *et al*. (2015) assuming the $^{87}Sr/^{86}Sr$-isotope ratio of soils as the independent variable and the must (grape) as the dependent variable.

We used the data for Sr-isotopic compositions of

soil labile fractions and musts collected from 2010 to 2012 (Table 1; Petrini *et al.*, 2015) for their processing by BiDASys. The regression results with the corresponding equations reported by Petrini *et al.* (2015) and obtained in this work are listed in Table 1 and plotted in Figure 3.

The regression parameters and results reported by Petrini *et al.* (2015) were as follows: (i) for 2010, an intercept $a = -0.133$, slope $b = 1.188$ and $R^2 = 0.77$ (equation (8); Table 1); (ii) for 2011, an intercept of a = -0.161, slope b = 1.227 and $R^2 = 0.71$ (equation (11); Table 1); and, (iii) for 2012, they obtained an intercept of a = -0.127, slope b = 1.180 and $R^2 = 0.78$ (equation (14); Table 1).

Now, we present the results obtained in this work by using our online computer program BiDASys. In this case study, no discordant outliers were detected. The OLR_{tw} results obtained for 2010 were: $a = -0.131$ with a total uncertainty of $u_a = 0.550$, $b = 1.186$ with a total uncertainty of $u_b = 0.770$, the correlation coefficient of $r = 0.8776$, $R^2 = 0.7701$ and the probability of no correlation $Pc(r;n)$ of 0.0005 (equation (9); Table 1; Figure 3a, purple dashed line). The UWLR results obtained were as follows: $a = -0.181$, $u_a = 0.163$, $b = 1.256$, $u_b = 0.230$, $r = 0.8779$, $R^2 = 0.7706$ and $Pc(r;n) = 0.0004$ (equation (10); Table 1; Figure 3a, green continuous line).

For 2011, the OLR_{tw} results obtained were: $a = -0.158$, $u_a = 0.670$, $b = 1.223$, $u_b = 0.945$, $r = 0.8381$, $R^2 = 0.7023$ and $Pc(r;n) = 0.0013$ (equation (12); Table 1; Figure 3b, purple dashed line). The UWLR results obtained were as follows: $a = -0.207$, $u_a = 0.201$, $b = 1.292$, $u_b = 0.283$, $r = 0.8408$, $R^2 = 0.7069$ and $Pc(r;n)$ value of 0.0012 (equation (13); Table 1; Figure 3b, green continuous line).

Finally, for 2012, the OLR_{tw} results obtained were: $a = -0.120$, $u_a = 0.541$, $b = 1.170$, $u_b = 0.763$, $r = 0.8765$, $R^2 = 0.7682$ and $Pc(r;n) = 0.0005$ (equation (15); Table 1; Figure 3c, purple dashed line). The UWLR results obtained were as follows: $a = -0.102$, $ua = 0.161$, $b = 1.144$, $u_b = 0.226$, $r = 0.8769$, $R^2 = 0.7690$ and $Pc(r;n) = 0.0005$ (equation(16); Table 1; Figure 3c, green continuous line).

We can now mention that the original authors did not report the standard errors and the probability of no-correlation from the regression results. However, the BiDASys program provides the probability of no-correlation for OLR (0.0005, 0.0013, and 0.0005 for 2010, 2011, and 2012, respectively) and UWLR (0.0004, 0.0012, and 0.0005 for 2010, 2011, and 2012, respectively; Figure 4) methods.

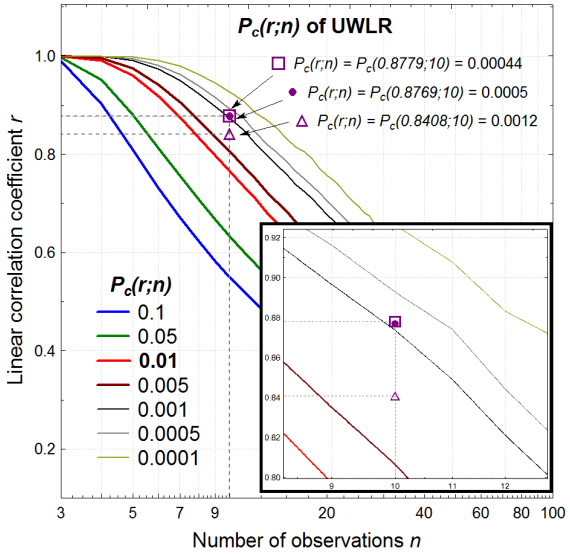

Fig. 4. The linear-correlation coefficient r versus the number of observations n and the corresponding probability Pc(r;n) that the variables ($^{87}Sr/^{86}Sr_{Musts}$ versus $^{87}Sr/^{86}Sr_{soil}$) are correlated. For application study A1 (Petrini *et al.*, 2015), the square, triangle and filled circle symbols correspond to the Pc(r;n) values (0.00044, 0.0012, and 0.0005) obtained for $^{87}Sr/^{86}Sr_{Must}2010$ versus $^{87}Sr/^{86}Sr_{soil}$, $^{87}Sr/^{86}Sr_{Must}2011$ versus $^{87}Sr/^{86}Sr_{soil}$, $^{87}Sr/^{86}Sr_{Must}2012$ versus $^{87}Sr/^{86}Sr_{soil}$, respectively.

Figure 4 shows the linear-correlation coefficient r versus the number of observations n and the corresponding probability $Pc(r;n)$. The Sr-isotope ratio in musts and soil labile fraction are correlated at the 99% confidence level ($Pc(r;n)$ of this relationship is greater than the 99.0% critical value; Figure 4).

Another advantage of BiDASys program is that it provides the uncertainty values at the strict confidence level of 99% obtained for each regression method. From the results obtained in this case study, we can also infer that the UWLR model (shaded rows in Table 1) presents systematically lower uncertainties for the calculated values of the intercept and slope (e.g., for 2012, $u_a = 0.161$ and $u_b = 0.226$) compared to the OLR results (for 2012, $u_a = 0.541$ and $u_b = 0.763$). Finally, because of the lower uncertainty and $Pc(r;n)$ values for the UWLR (Table 1; Figure 4), we may also infer that the UWLR results better support the conclusions reported by the original authors (Petrini *et al.*, 2015).

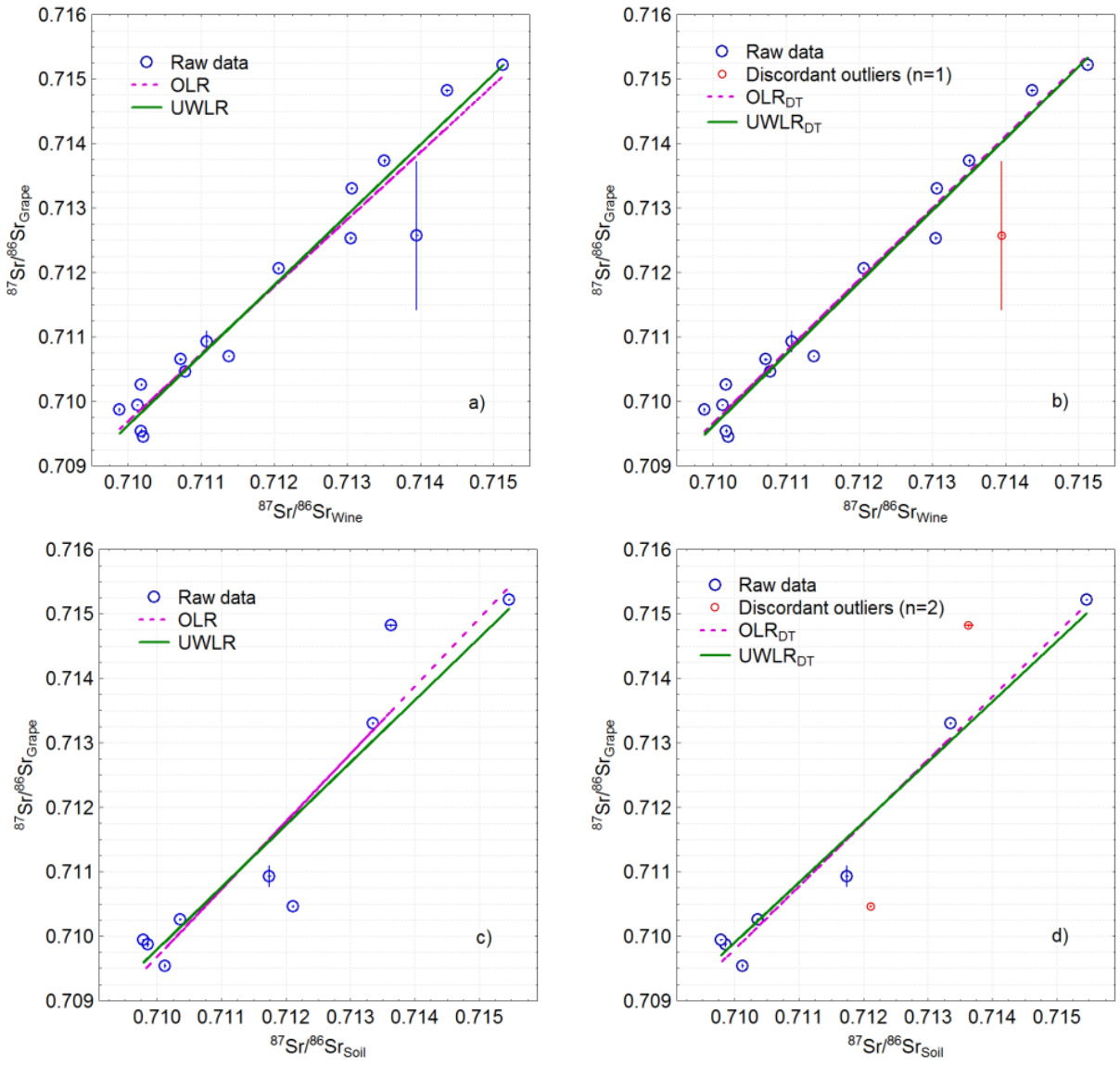

Fig. 5. $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$ and $^{87}Sr/^{86}Sr_{soil}$ for vineyards from different wine producing areas of Quebec, Canada (Application Study A2; Vinciguerra *et al*., 2016). a) OLR and UWLR results from the original data for $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$; b) OLR_{DT} and $UWLR_{DT}$ results after the application of recursive discordancy tests for $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$; c) OLR and UWLR results from the original data for $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{soil}$; and, d) OLR_{DT} and $UWLR_{DT}$ results after the application of recursive discordancy tests for $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{soil}$.

3.1.2 Application Study A2: Vineyards from different wine producing areas of Quebec, Canada

Vinciguerra *et al*. (2016) reported new Sr-isotope data for soils, grapes, and wine derived from 13 vineyards from different regions of Quebec (Canada), suggested traceability of the isotope composition from the soil, through the grapes to the wine, and provided constraints for determining the geographic origin of locally produced wines. The data for $^{87}Sr/^{86}Sr$ ratios of wine, grape and labile soil fractions from Table 1 of Vinciguerra *et al*. (2016) were used for their processing by BiDASys. The regression results with the corresponding equations obtained by Vinciguerra *et al*. (2016) and those obtained in this study are included in Table 2 and plotted in Figure 5.

Table 2. Regression parameters and equations from Sr-isotopic of grape versus wine and grape versus soil, using OLR and UWLR models (Application Study A2; Vinciguerra *et al*., 2016).

Regression model	*n*	Regression parameters							Equation	
		a	u_a	*b*	u_b	*r*	R^2	*Pc(r;n)*		
					$^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$					
OLR_{lit}	16	-0.03	NR	1.04	NR	NR	0.94	NR	y = -0.03 + 1.04 * x	(17)
OLR_{tw}	16	-0.031	0.150	1.043	0.211	0.969	0.9394	<0.00005	y = -0.031 (± 0.150) + 1.043 (± 0.211) * x	(18)
UWLR	16	-0.063	0.049	1.088	0.070	0.971	0.9428	<0.00005	y = -0.063 (± 0.049) + 1.088 (± 0.070) * x	(19)
OLR_{DT}	15	-0.079	0.109	1.111	0.153	0.987	0.9737	<0.00005	y = -0.079 (± 0.109) + 1.111 (± 0.153) * x	(20)
$UWLR_{DT}$	15	-0.0803	0.0351	1.113	0.049	0.988	0.9753	<0.00005	y = -0.0803 (± 0.0351) + 1.113 (± 0.049) * x	(21)
					$^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Soil}$					
OLR_{lit}	9	-0.04	NR	1.06	NR	NR	0.88	NR	y = -0.04 + 1.06 * x	(22)
OLR_{tw}	9	-0.036	0.360	1.05	0.51	0.9398	0.8832	0.00008	y = -0.036 (± 0.360) + 1.05 (± 0.51) * x	(23)
UWLR	9	0.023	0.105	0.968	0.148	0.9395	0.8826	0.00008	y = 0.023 (± 0.105) + 0.968 (± 0.148) * x	(24)
OLR_{DT}	7	0.014	0.203	0.98	0.285	0.9872	0.9746	<0.00005	y = 0.014 (± 0.203) + 0.98 (± 0.285) * x	(25)
$UWLR_{DT}$	7	0.046	0.052	0.935	0.072	0.9873	0.9747	<0.00005	y = 0.046 (± 0.052) + 0.935 (± 0.072) * x	(26)

See Table 1 footnote for more information. OLR_{DT} – ordinary least-squares regression results obtained in this work after the application of recursive tests to studentized residuals; $UWLR_{DT}$ – uncertainty weighted least-squares regression results obtained in this work after the application of recursive tests to studentized residuals.

Vinciguerra *et al*. (2016) report the corresponding linear relationship $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$ as y = -0.03 + 1.04 * x (equation (17)) and R2 = 0.94 and for $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{soil}$ as y = -0.04 + 1.06 * x (equation (22)) and R^2 = 0.88 (Table 2).

Now, using our statistical procedure through BiDASys, we present the regression results from OLR_{tw} and UWLR models before and after the detection of discordant outliers. These values were detected as follows: one discordant outlier from the relationship $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{Wine}$ (Figure 5a-b) and two discordant outliers from the relationship $^{87}Sr/^{86}Sr_{Grape}$ versus $^{87}Sr/^{86}Sr_{soil}$ (Figure 5c-d).

For the first relationship, the OLR_{tw} results are: $y = -0.031\ (\pm\ 0.150) + 1.043\ (\pm\ 0.211) * x, r = 0.969$, R^2 = 0.9394 and $Pc(r;n) < 0.00005$, consistent with the results from the original authors; and for UWLR, the results are: y = -0.063 (± 0.049) + 1.088 (± 0.070) * x, r = 0.971, R^2 = 0.9428 and $Pc(r;n) < 0.00005$ (Table 2; Figure 5a). After the detection and separation of one discordant outlier the results for OLR_{DT} are: $y = -0.079\ (\pm\ 0.109) + 1.111\ (\pm\ 0.153) * x, r = 0.987$, R^2 = 0.9737 and $Pc(r;n) < 0.00005$, and $UWLR_{DT}$ results are: $y = -0.0803\ (\pm\ 0.0351) + 1.113\ (\pm\ 0.049)$ * x, r = 0.988, R^2 = 0.9753 and $Pc(r;n) < 0.00005$ (Table 2; Figure 5b).

For the second relationship, the OLR_{tw} results are: y = -0.036 (± 0.360) + 1.05 (± 0.51) * x, r = 0.9398, R^2 = 0.8832 and $Pc(r;n)$ = 0.00008, and the results for UWLR are: y= 0.023 (± 0.105) + 0.968 (± 0.148) * x, r = 0.9395, R^2 = 0.8826 and $Pc(r;n)$ = 0.00008 (Table 2; Figure 5c). After the detection of two discordant outliers, the results for OLR_{DT} are: $y = 0.014\ (\pm\ 0.203) + 0.98\ (\pm\ 0.285) * x, r = 0.9872$, R^2 = 0.9746 and $Pc(r;n) < 0.00005$, and $UWLR_{DT}$ results are: $y = 0.046\ (\pm\ 0.052) + 0.935\ (\pm\ 0.072)$ * x, r = 0.9873, R^2 = 0.9747 and $Pc(r;n) < 0.00005$ (Table 2; Figure 5d).

We can clearly appreciate the difference between before and after the application of both regression models, where the results provided by the $UWLR_{DT}$ regression present higher values of correlation coefficients (0.9753, 0.9747) compared with the values reported by the original authors (0.94, 0.88).

In addition, we can appreciate that the $UWLR_{DT}$ (shaded rows in Table 2) presents systematically lower uncertainties (u_a = 0.0351, 0.052; and u_b = 0.049, 0.072) for both relationships compared to the OLR_{DT} (u_a = 0.109, 0.203; and u_b = 0.153, 0.285).

The values detected as discordant outliers are clearly visible from the plots in the original paper but the corresponding probability cannot be ascertained from these plots. Our approach (use of the BiDaSys program) can improve the conclusions presented by Vinciguerra *et al*. (2016). This confirms that the $^{87}Sr/^{86}Sr$-isotope ratios can be used to track the geographical origin of wine.

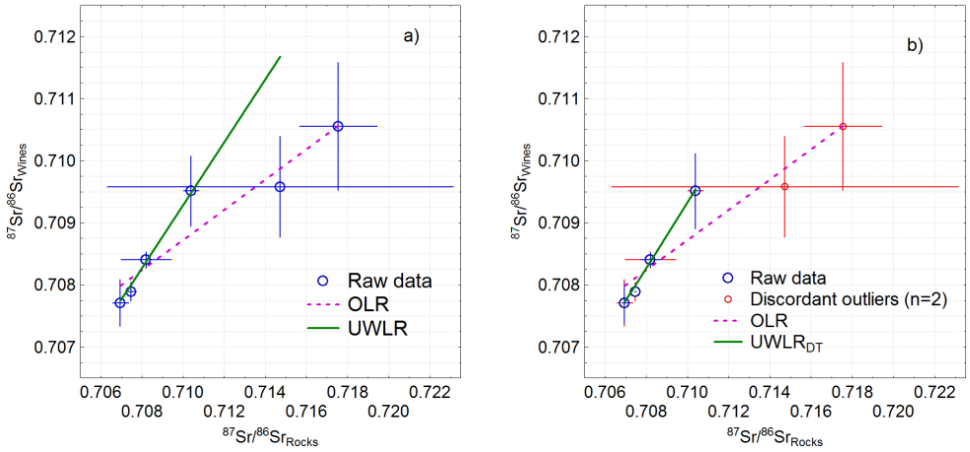

Fig. 6. $^{87}Sr/^{86}Sr_{Wines}$ versus $^{87}Sr/^{86}Sr_{Rocks}$ for vineyards from Tuscany to Basilicata, Italy (Application Study A3; Marchionni *et al*., 2013). a) OLR and UWLR results from the original data; and b) OLR and UWLR$_{DT}$ results after the application of recursive discordancy tests to the studentized residuals.

Table 3. Regression parameters and equations from Sr-isotopic of wines versus rocks, using OLR and UWLR models (Application Study A3; Marchionni *et al*., 2013).

Regression model	*n*	Regression parameters							Equation	
		a	u_a	*b*	u_b	*r*	R^2	*Pc(r;n)*		
		$^{87}Sr/^{86}Sr_{Wines}$ versus $^{87}Sr/^{86}Sr_{Rocks}$								
OLR$_{lit}$	6				NR				NR	
OLR$_{tw}$	6	0.536	0.141	0.243	0.199	0.9422	0.8877	0.00249	$y = 0.536\ (\pm 0.141) + 0.243\ (\pm 0.199) * x$	(27)
UWLR	6	0.35	0.098	0.507	0.138	0.9564	0.9148	0.00144	$y = 0.350\ (\pm 0.098) + 0.507\ (\pm 0.138) * x$	(28)
UWLR$_{DT}$	4	0.3342	0.0203	0.528	0.0287	0.9971	0.9942	0.00145	$y = 0.3342\ (\pm 0.0203) + 0.528\ (\pm 0.0287) * x$	(29)

See Table 1 footnote for nomenclature.

3.1.3 *Application Study A3: Vineyards from Tuscany to Basilicata, Italy*

The wine regions selected for Marchionni *et al*. (2013) are distributed along the Italian Peninsula, from Tuscany to Basilicata. The selected wine areas are Tuscany, Latium, Campania, Basilicata, Chianti Classico and Giglio Island. The wine Sr-isotope compositions are then cross-checked with geological and isotopic data of the rocks of the production areas to verify the relationship, if any, between wines and their geological isotopic characteristics. We used the values for $^{87}Sr/^{86}Sr$ of the wines from the same production area and those rocks of the substratum (Table 3; Marchionni *et al*. 2013) for their processing by BiDASys. The regression results obtained in this work are presented in Table 3 and Figure 6.

The results for OLR$_{tw}$ are shown in equation (27), with r = 0.9422, R^2 = 0.8877 and a $Pc(r;n)$ value of 0.00249 (Figure 6a). For UWLR, results from equation (28) and the values for r, R^2, and $Pc(r;n)$ are 0.9564, 0.9148, and 0.00144, respectively. We can appreciate that the uncertainty values and the correlation coefficients obtained by the UWLR model are lower in comparison with those obtained by the ORL model. After the detection of two discordant outliers (Figure 6b), the UWLR$_{DT}$ results (equation (29)) for the linear relationship between those wines and rocks are: r = 0.9971, R^2 = 0.9942, and $Pc(r;n)$ = 0.00145. In several cases, the $^{87}Sr/^{86}Sr$ value of wines overlaps with the values of the rock of the substratum of the vineyards from which the wines are produced, which can be sustainable with the results obtained by the UWLR$_{DT}$ model (Table 3; Figure 6a-b). Because four of the six wine areas under consideration in this study are characterized by vineyards cultivated mainly or partially on volcanic terrains. The two discordant outliers detected correspond to these wine areas, Chianti Classico (sedimentary substrata) and Giglio Island (granitic rocks), which agrees with the conclusions reported by Marchionni *et al*. (2013). However, this was not clearly obtained by the OLR model, as shown in Table 3 and Figure 6a-b.

Table 4. Regression parameters and equations of pp'DDE versus Days of lactation, Σ-DDT versus Days of lactation, pp'DDE versus Weeks, and Σ-DDT versus Weeks (Application Study A4; Chávez-Almazán *et al*., 2015).

Regression model	*n*	Regression parameters							Equation	
		a	u_a	*b*	u_b	*r*	R^2	*Pc(r;n)*		
mg/kg lipid of pp'DDE versus Days of lactation										
OLR_{lit}	161	NR	NR	NR	NR	-0.216	NR	0.006	NR	
OLR_{tw}	161	1.241	0.354	-0.0139	0.0129	-0.2164	0.0468	0.00609	$y = 1.241\ (\pm 0.354) - 0.0139\ (\pm 0.0129) * x$	(30)
OLR_{DT}	109	0.965	0.179	-0.0144	0.0063	-0.4996	0.2496	<0.00005	$y = 0.965\ (\pm 0.179) - 0.0144\ (\pm 0.0063) * x$	(31)
mg/kg lipid of ∑-DDT versus Days of lactation										
OLR_{lit}	161	NR	NR	NR	NR	-0.222	NR	0.005	NR	
OLR_{tw}	161	1.346	0.373	-0.0149	0.0136	-0.2215	0.0491	0.00444	$y = 1.346\ (\pm 0.373) - 0.0149\ (\pm 0.0136) * x$	(32)
OLR_{DT}	102	1.163	0.174	-0.0163	0.0061	-0.5736	0.3291	<0.00005	$y = 1.163\ (\pm 0.174) - 0.0163\ (\pm 0.0061) * x$	(33)

See Table 1 footnote for nomenclature

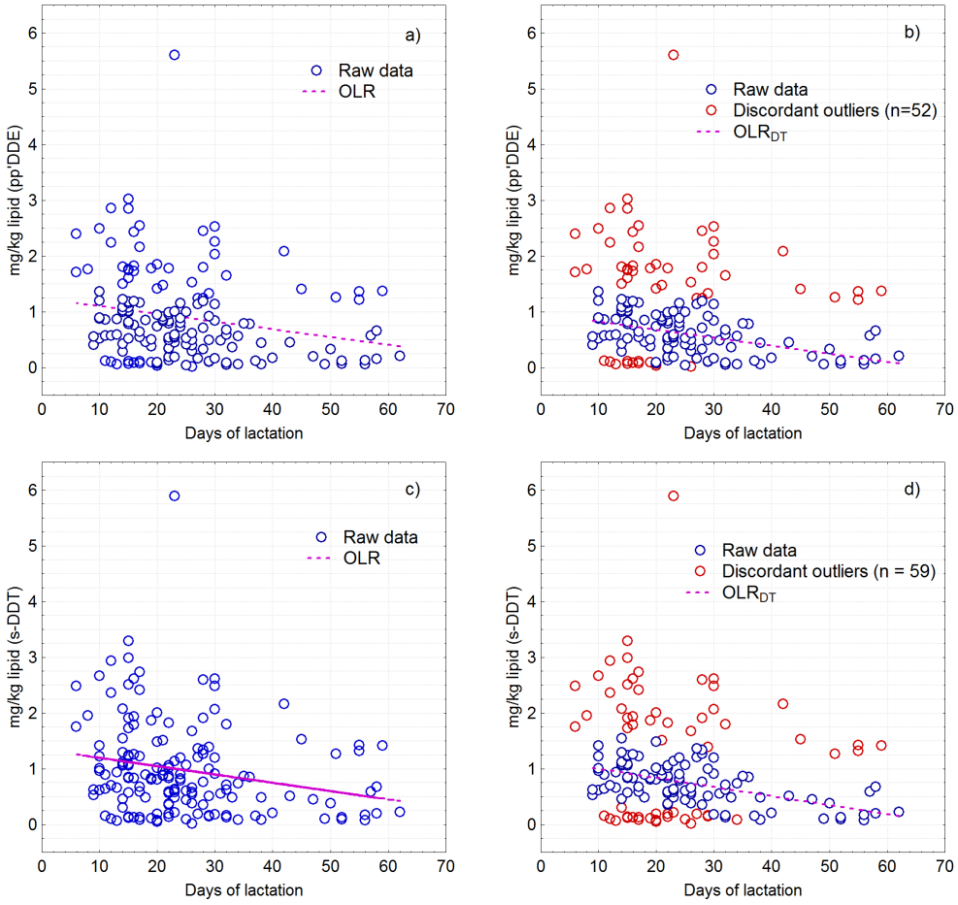

Fig. 7. pp′DDE levels and Σ-DDT levels versus Days of lactation in Breast-Milk from mothers that lived in Guerrero, Mexico (Application Study A4; Chávez-Almazán *et al*., 2015). a) OLR regression results for pp'DDE levels versus days of lactation; b) OLR_{DT} regression results after the application of recursive discordancy tests for pp'DDE levels versus days of lactation; c) OLR regression for Σ-DDT levels versus days of lactation; d) OLR_{DT} regression results after the application of recursive discordancy tests for Σ-DDT levels versus days of lactation.

3.1.4 *Application Study A4: Human milk of a population from Guerrero, Mexico*

The final example is from Chávez-Almazán *et al*. (2015). The aim of their study was to determine variations in organochlorine pesticide levels in breast milk through the grouping of donor according to the stage of lactation in which they were in. The quantities of organochlorine pesticides in breast milk were expressed as milligrams per kilogram of lipid base (mg/kg lipid). The regression results obtained by Chávez-Almazán *et al*. (2015) and those obtained in this study are summarized in Table 4.

Chavez-Almazán *et al*. (2015) report values for the relationship pp'DDE versus Days of lactation of $r = -0.216$ and a $Pc(r;n)$ value of 0.006 and for the relationship Σ-DDT versus Days of lactation of $r = -0.222$ and a $Pc(r;n)$ value of 0.005 (Table 4). Because errors were not reported for the individual data, we processed the original data with the OLR model (Table 4; Figure 7).

For the relationship of pp'DDE versus Days of lactation, the OLR_{tw} results are: $y = 1.241$ (± 0.354) - 0.0139 (± 0.0129) * $x, r = -0.2164$, $R^2 = 0.0468$ and $Pc(r;n) = 0.00609$, consistent with the results from the original authors (Table 4; Figure 7a). After the detection and separation of 52 discordant outliers the results for OLR_{DT} were as follows: $y = 0.965$ (±0.179) - 0.0144 (±0.0063) $*x, r = -0.4996$, $R^2 = 0.2496$ and $Pc(r;n) < 0.00005$ (Table 4; Figure 7b).

For the relationship of Σ-DDT versus Days of lactation, the OLR_{tw} results were as follows: $y = 1.346$(± 0.373) - 0.0149 (± 0.0136) $*x, r = -0.2215$, $R^2 = 0.0491$ and $Pc(r;n) = 0.00444$, also consistent with the results from the original authors (Table 4; Figure 7c). After the detection and separation of 59 discordant outliers, the results for OLR_{DT} are: $y = 1.163$(±0.174) - 0.0163 (±0.0061) $*x, r = -0.5736$, $R^2 = 0.3291$ and $Pc(r;n) < 0.00005$ (Table 4; Figure 7d).

Chávez-Almazán *et al*. (2015) applied the Kolmogorov-Smirnov normality test to their data, but this test does not allow the detection of outlying observations. It is clear from the present methodology that the original statistical samples were not drawn from a single normal population. Furthermore, for this case study we encourage to the original authors to estimate and report the individual errors of their experimental data to apply the UWLR model, to improve their interpretation.

4 Discussion

The four case studies selected for the illustration of BiDASys clearly show the high potential of this software in food chemistry. Highlighting the comparison of the results from OLR and UWLR regression models, the UWLR scheme is more robust, because it systematically presents small uncertainty values of regression parameters (intercept and slope) and small probability of the no-correlation value, which is equivalent to a high probability that a linear correlation exists. This was most notable in situations where at least one discordant outlier was identified and separated. Through this software, we encourage the use of $UWLR_{DT}$ regression method that estimates the total uncertainty of each data point, which is better than the conventional practice (use of OLR), because UWLR has a probability connotation, here a strict confidence level of 99% was used. Our program (BiDASys) clearly showed that the $UWLR_{DT}$ approach gave more consistent results than the original authors. With regard to experimental data compilation, we highly recommend that the researchers start estimating and reporting errors of individual experimental data as is customary in isotopic studies.

Finally, BiDASys is the only computer program capable of applying the UWLR regression model. For the studentized residuals, the discordant outlier detection module can be applied to sample sizes $n = 5 - 30,000$. For $n \leq 1000$, BiDASys applies 5 recursive tests, due to the current available critical values (Verma *et al*., 2017b) and for $1000 < n \leq 30,000$, applies 13 single-outlier discordancy tests. As a future work, the critical values simulation for the recursive tests and the Pearson correlation coefficient (r) values will be carried out to sample sizes up to 30,000.

Conclusions

The following conclusions can be drawn from this study:

1. Application of the proposed statistical methodology was facilitated by the new online computer program BiDASys. This program should be very useful for the handling of bivariate data in all science and engineering fields,

including food chemistry.

2. The UWLR regression method outlined in the present work seems to perform better than OLR unweighted regression model. UWLR clearly and consistently presented smaller uncertainty values of regression parameters (intercept and slope) than OLR model.

3. Simulation of the new precise and accurate critical values for the probability of no-correlation $Pc(r : n)$ criterion for sample sizes up to 1000 is required for a better interpretation of statistical significant linear correlation. These critical values can be used for all applications.

4. Identification and separation of discordant outliers coupled to the regression methods clearly showed more consistent results than the original authors.

5. The importance of estimating and reporting errors of individual experimental data could lead to use the best weighted regression model, such as UWLR practiced in this work.

Acknowledgements

This work was partly supported by the DGAPA-PAPIIT grant IN100816. Mauricio Rosales-Rivera thanks CONACYT for the doctoral fellowship. We are much grateful to Luis A. Chávez-Almazán for sending us the original complete data related to the paper by Chávez-Almazán *et al*. (2016). We are grateful to the editor Jaime Vernon for efficient handling of our manuscript. We also sincerely thank to the three anonymous reviewers who provided useful comments to improve our presentation.

References

Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika 68*, 13-20.

Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data*. 3rd ed. Chichester, U.K: John Wiley and Sons, 256pp.

Bevington, P.R., Robinson, D.K. (2003). *Data Reduction and Error Analysis for the Physical Sciences*. McGraw Hill, Boston, 336pp.

Borges, M.V., Alves, M.F., Chaves, M.A., Egito, A.S., Gross, E., Ferrão, S.P.B. (2017). Chemical, structural and proteomic profile of buffalo milk powder produced in mini spray dryer. *Revista Mexicana de Ingeniería Química 16*, 67-76.

Chaparro, M. L., Céspedes, E., Cruz, M., Castillo-Saldarriaga, C. R., Gómez-Álvarez, M. I. (2017). Fluidized bed drying of a granulated prototype based on a potential probiotic yeast Meyerozyma Guilliermondii: selection of process parameters and drying protectant. *Revista Mexicana de Ingeniería Química 16*, 347-357.

Chatterjee, S., Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science 1*, 379-393.

Chatterjee, S., Hadi, A.S. (2015). *Regression Analysis by Example*. John Wiley & Sons, 424pp.

Chávez-Almazán, L.S., Díaz-Ortíz, J., Alarcón-Romero, M., Davila-Vazquez, G., Saldarriaga-Noreña, H., Sampedro-Rosas, L., López-Silva, S., Santiago-Moreno, A., Rosas-Acevedo, J.L., Waliszewski, S.M. (2015). Influence of breastfeeding time on levels of organochlorine pesticides in human milk of a Mexican population. *Bulletin of Environmental Contamination and Toxicology 96*, 168-172.

Cozzolino, D. (2014). The use of correlation, association and regression to analyse processes and products. Mathematical and statistical methods in food science and technology. Wiley Blackwell, *Chichester*, 19-30.

Darlington, R.B., Hayes, A.F. (2016). *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*. Guilford Publications, 661pp.

Draper, N. R., Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons, 736pp.

Fernández, J.A., Niñirola, D., Ochoa, J., Orsini, F., Pennisi, G., Gianquinto, G., Egea-Gilabert, C. (2016). Root adaptation and ion selectivity affects the nutritional value of salt-stressed hydroponically grown baby-leaf *Nasturtium officiale* and *Lactuca sativa*. *Agricultural and Food Science 25*, 230-239.

Foereid, B. (2017). Phosphorus availability in residues as fertilizers in organic agriculture. *Agricultural and Food Science 26*, 25-33.

Fuentes-Ortega, T., Martínez-Vargas, S. L., Cortés-Camargo, S., Guadarrama-Lezama, A. Y., Gallardo-Rivera, R., Baeza-Jimenéz, R., Pérez-Alonso, C. (2017). Effects of the process variables of microencapsulation sesame oil (*Sesamum indica* L.) by spray drying. *Revista Mexicana de Ingeniería Química 16*, 477-490.

Giordani, E., Ancillotti, C., Petrucci, W.A., Ciofi, L., Morelli, D., Marinelli, C., Checchini, L., Furlanetto, S., Del Bubba, M. (2016). Morphological, nutraceutical and sensorial properties of cultivated *Fragaria vesca* L. berries: influence of genotype, plant age, fertilization treatment on the overall fruit quality. *Agricultural and Food Science 25*, 187-201.

Gómez-Favela, M. A., García-Armenta, E., Reyes-Moreno, C., Garzón-Tiznado, J. A., Perales-Sánchez, J. X. K., Caro-Corrales, J. J., Gutiérrez-Dorado, R. (2017). Modelling of water absorption in chickpea (*Cicer arietinum* L) seeds grown in Mexico's northwest. *Revista Mexicana de Ingeniería Química 16*, 179-191.

Granato, D., Calado, V.M.A. (2014). The use and importance of design of experiments (DOE) in process modelling in food science and technology. *Mathematical and Statistical Methods in Food Science and Technology*, 1-18.

Granato, D., Calado, V.M.A., Jarvis, B. (2014). Observations on the use of statistical methods in food science and technology. *Food Research International 55*, 137-149.

Guevara, M., Verma, S.P., Velasco-Tapia, F., Lozano-Santa Cruz, R., Girón, P. (2005). Comparison of linear regression models for quantitative geochemical analysis: An example using x-ray flourescence spectrometry. *Geostandards and Geoanalytical Research 29*, 271-284.

Law, A.M., Kelton, W.D. (2000). *Simulation Modeling and Analysis*. McGraw Hill, Boston, 760 pp.

Mahon, K.L. (1996). The New "York" Regression: Application of an improved statistical method to geochemistry. *International Geology Review 38*, 293-303.

Marchionni, S., Braschi, E., Tommasini, S., Bollati, A., Cifelli, F., Mulinacci, N., Mattei, M., Conticelli, S. (2013). High-precision $^{87}Sr/^{86}Sr$ analyses in wines and their use as a geological fingerprint for tracing geographic provenance. *Journal of Agricultural and Food Chemistry 61*, 6822-6831.

Maslak, K., Nimmermark, S. (2017). Thermal energy use for dehumidification of a tomato greenhouse by natural ventilation and a system with an air-to-air heat exchanger. *Agricultural and Food Science 26*, 56-66.

Miller, J.N., Miller, J.C. (2010). *Statistics and Chemometrics for Analytical Chemistry*. Prentice Hall, Essex. 271pp.

Moreno-Rivas, S. C., Armenta-Corral, R. I., Frasquillo-Félix, M. C., Lagarda-Díaz, I., Vázquez-Moreno, L., Ramos-Clamont Montfort, G. (2016). Biosorción de cadmio en solución acuosa utilizando levadura de panadería (*Saccharomyces cerevisiae*). *Revista Mexicana de Ingeniería Química 15*, 843-857.

Nunes, C.A., Alvarenga, V.R., Sant'Ana, A.S., Santos, J.S., Granato, D. (2015). The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Research International 75*, 270-280.

Passari, L.M.Z.G., Soares, P.K., Bruns, R.E. (2011). Estatística aplicada à química: dez dúvidas comuns. *Quimica Nova 34*, 888-892.

Pérez-Grijalba, B., García-Zebadúa, J. C., Ruíz-Perez, V. M., Téllez-Medina, D. I., Guzmán-Gerónimo, R. I., Mora-Escobedo, R. (2017). Biofunctionality, colorimetric coefficients and microbiological stability of blackberry (*Rubus fructicosus* var. Himalaya) juice under microwave/ultrasound processing. *Revista Mexicana de Ingeniería Química 17*, 13-28.

Petrini, R., Sansone, L., Slejko, F.F., Buccianti, A., Marcuzzo, P., Tomasi, D. (2015). The $^{87}Sr/^{86}Sr$ strontium isotopic systematics applied to Glera vineyards: A tracer for the geographical origin of the Prosecco. *Food chemistry 170*, 138-144.

Pope, J. (1976). The statistics of residuals and detection of outliers. NOAA Technical Report Nos 66, NGS 1.

Raco, B., Dotsika, E., Poutoukis, D., Battaglini, R., Chantzi, P. (2015). O-H-C isotope ratio determination in wine in order to be used as a fingerprint of its regional origin. *Food Chemistry 168*, 588-594.

Velleman, P. F., Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician 35*, 234-242.

Verma, S.P. (2012). Geochemometrics. *Revista Mexicana de Ciencias Geológicas 29*, 276-298.

Verma, S.P. (2016). *Statistical Analysis of Compositional Data*. CDMX Mexico City, Mexico: Universidad Nacional Autónoma de México (book in Spanish), 746pp.

Verma, S.P., Díaz-González, L., Pérez-Garza, J.A., Rosales-Rivera, M. (2016). Quality control in geochemistry from a comparison of four central tendency and five dispersion estimators and example of a geochemical reference material. Arabian Journal of Geosciences 9(20), 740.

Verma, S.P., Díaz-González, L., Pérez-Garza, J.A., Rosales-Rivera, M. (2017a). Erratum to: Quality control in geochemistry from a comparison of four central tendency and five dispersion estimators and example of a geochemical reference material. *Arabian Journal of Geosciences 10*, 24.

Verma, S.P., Rosales-Rivera, M., Díaz-González, L., Quiroz-Ruiz, A. (2017b). Improved composition of Hawaiian basalt BHVO-1 from the application of two new and three conventional recursive discordancy tests. *Turkish Journal of Earth Sciences 26*, 331-353.

Vinciguerra, V., Stevenson, R., Pedneault, K., Poirier, A., Hélie, J.F., Widory, D. (2016). Strontium isotope characterization of wines from Quebec, Canada. *Food chemistry 210*, 121-128.

Xu, H., Sonke, J.E., Guinot, B., Fu, X., Sun, R., Lanzanova, A., Candaucap, F., Shen, Z., Cao, J. (2017). Seasonal and annual variations in atmospheric Hg and Pb isotopes in Xi'an, China. *Environmental Sciences & Technology 51*, 3759-3766.

York, D. (1968). Least-squares fitting of a straight line with correlated errors. *Earth and Planetary Science Letters 5*, 320-324.