OYNYL: A new Computer Program for Ordinary, York, and New York Least-Squares Linear Regressions

SURENDRA P. VERMA [1], LORENA DÍAZ-GONZÁLEZ [2], PEDRO SÁNCHEZ-UPTON [2], and E. SANTOYO [1]
[1] Geoenergía, Centro de Investigación en Energía [2] Posgrado en Ingeniería (Energía), sede CIE
Universidad Nacional Autónoma de México
Privada Xochicalco s/n, col. Centro, Temixco, Mor. 62580
MEXICO

Abstract: - We present a synthesis of three least-squares linear regression models (one unweighted – commonly called Ordinary– and two weighted – "York" and "New York"). The corresponding equations and algorithms have been programmed in a new software OYNYL whose essential structure is presented in this paper. The data can be input using either an Excel or a plain text (txt or ASCII) file, and the program can be used for practically any number of pairs, either as "multiple" pairs or "nested" pairs. Further, OYNYL can take into account both correlated and uncorrelated errors. Another novel aspect of our software is that it incorporates an appropriate outlier-detection algorithm. The program saves the regression results in the same format as the initial input data. It also provides a scalable vector graphic output that can be exported to a commercial software such as Corel Draw or a free-access software such as Sodipodi, and, thus, the graph quality can be improved. The use of our new computer program is illustrated through case studies from environment, ecosystems, and development.

Key-Words: - ordinary least-squares linear regression; weighted least-squares linear regression; error propagation; radioactive isotopes; mushroom; biomass; wood; Mexico; Brazil; Java computer language

1 Introduction

Linear regression analysis is widely used in all scientific and engineering fields, including the area of environment, ecosystems, and development. Several models for linear regression have been proposed that vary from the conventional ordinary least-squares regression [1,2] to weighted least-squares "York" and "New York" regression models, proposed by York [3] and Mahon [4], respectively.

At present to the best of our knowledge, no computer program is available to handle all the three types of regressions in an efficient way. We have developed such a versatile, easy to use computer program for this purpose, whose basic characteristics, along with selected case studies from the areas of environment, ecosystems, and development to highlight its use, are presented in this paper.

2 Problem Formulation

Different types of simple unweighted (ordinary) and weighted models are commonly used to explore the relationships between an independent variable (x) and a dependent variable (y). When a linear correlation is statistically significant (e.g., if the linear correlation coefficient, r has a very low probability of no-correlation $P_c(r;n)$ [5] and other linearity tests [6] are valid), this relationship can be used to interpret the data and to infer about natural processes. Discordant outliers may also occur in such situations and should be best handled using proper statistical techniques [7].

We will present only a very brief outline of these different approaches; the reader is referred to the literature references [1-10] for more details on the different types of regressions, including their limitations and applications.

2.1 Ordinary least-squares (OLS) linear regression model

This simple OLS model is the most frequently used method for exploring relationships of experimental data. It assigns equal weights to all data points, irrespective of the experimental errors, and is also called an unweighted model. The simple equation (1) describes this model for two variables x and y, where a and b are the regression parameters intercept and slope, respectively, and s_a and s_b are the corresponding errors.

$$y = a(\pm s_a) + b(\pm s_b)x \qquad (1)$$

However, for the model to be statistically valid certain assumptions must be fulfilled [1,2,8,10]: (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, more sophisticated regression models are required.

2.2 Weighted least-squares (WLS) linear regression models

These more complex WLS models are generally required because all assumptions for the OLS are seldom fulfilled. For example, in most of the experimental studies, the errors are heteroscedastic instead of being homoscedastic. WLS models rely upon assigning different weights to different data points, generally as an inverse function of the corresponding variances. One such model was proposed in 1969 by York [3] for isochron work in geochronology, and was specifically put forth for correlated errors in *x* and *y*. This model is still in wide use. More recently, a refinement and a correction of this model was published by Mahon [4], and was called "New York" model.

Equation (2) describes the basic WLS model, where the symbols are the same as for the OLS and the subscript $_w$ refers to the weighted regression.

$$y = a_w(\pm s_{a_w}) + b_w(\pm s_{b_w})x \qquad (2)$$

2.1.1 York regression model

This weighted linear regression model was developed for isotope data in geochronology, particularly when the *x* and *y* errors are correlated [3]. This model has been a very widely used algorithm in Earth Sciences. It was programmed in some readily available software.

2.1.2 New York regression model

This model was an improvement and a correction of the older York model. The main improvement has been with a better estimation of standard errors for slope and intercept values [4]. The model also provides a regression solution that is different from both York and OLS [4]. No computer program seems to be yet available to carry out New York regression.

3 Problem Solution

A new computer program OYNYL (Ordinary and Weighted (York, and New York) Least-Squares linear regressions), which enables us to apply all the three types of linear regressions to an "appropriate" dataset, has been written in Java. Besides, the program also has a built-in function to detect outliers in regressed data – a novel approach as compared to any other regression software. A schematic flow diagram is shown in Fig. 1.

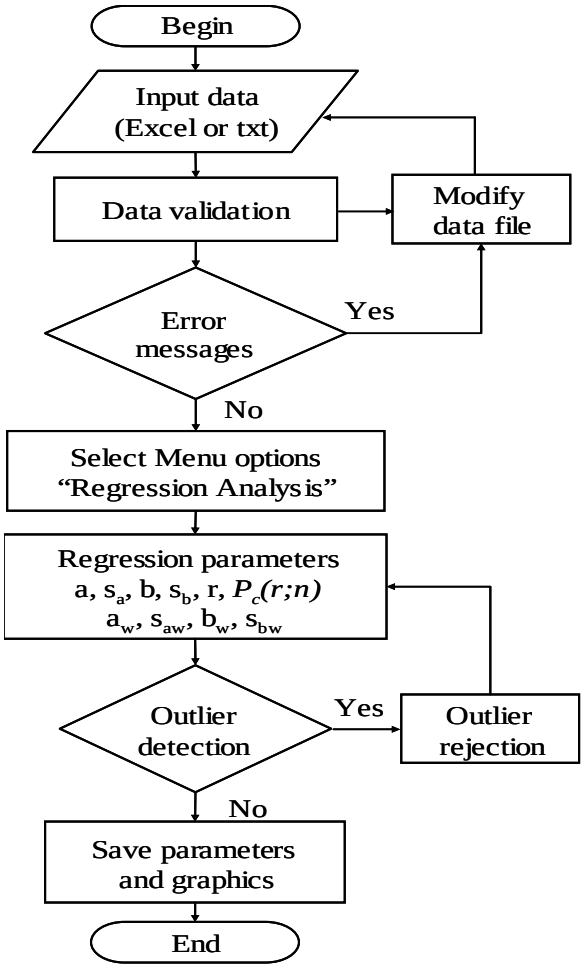

Fig. 1. Schematic flow diagram of OYNYL software written in Java computer language. The regression parameters a, sa, b, sb, r, and $P_c(r;n)$ are respectively intercept, standard error of the intercept, slope, standard error of the slope, linear correlation coefficient, and probability of no-correlation. The subscript $_w$ is for the weighted models.

The data can be input using either an Excel or a plain text format (txt or ASCII) file and the program can be used for any number of pairs of *x* and *y* variables, either as "multiple" (a number of *x*-*y*) pairs or "nested" (all variables as *x* against all as *y*) pairs. The "Data validation" module ensures that

the data are logically correct, e.g., four columns (*n*, *x*, s*x*, Rsd*x*) are used for each variable *x*; data do not contain a letter instead of numbers; and standard deviation (s*x* and s*y*) or relative standard deviation (Rsd*x* and Rsd*y*) values are all positive. Both types of errors –correlated as well as uncorrelated – can be handled. The system asks you to provide a value for the correlation parameter (0 is input for uncorrelated errors, 1 for totally correlated errors, and a value between 0 and 1 for partially correlated errors).

After processing all data for the three types of regressions ("Default process" option), the outlier detection module processes the data for possible discordant outliers using algorithms presented by [7] for the number of data to be regressed up to 100. If no outliers are found, the program proceeds to save the results. If any outlier or outliers are detected as discordant observation(s), these are eliminated, and the regression process is repeated. Both sets of results – for input data with and without outliers – are then presented on the screen, and these can be easily saved in computer files.

This program also provides a graphic output as scalable vector graphics (*.svg) that can either be printed or exported to a commercial software such as Corel Draw or a free-access software such as Sodipodi, and, thus, the graph quality can be improved using conventional software.

The proper functioning of the program was ascertained by processing of the data for isochron analysis (Table 3; [4]), for which results obtained from OYNYL were identical to those presented by the author of the New York algorithm (Table 4; [4]).

The use of this program is illustrated through examples from the three research areas covered in this Conference.

3.1 Application Example of Environment:

Data for two radioactive isotopes: one relatively short-lived ^{137}Cs (half-life ~ 30.17 y [11]) and the other very long-lived ^{40}K (half-life ~ 1.28×10^9 y [11]), together with numerous chemical elements, in soil samples and edible wild mushroom species from the central part of the Mexican Volcanic Belt (Salazar, State of Mexico and surroundings of ~ 20 km radius, in semi-natural forest ecosystems, at altitudes between 3000 and 3700 m asl) were reported by [12].

We used the data for 15 mushroom species collected between 1993 to 1999 (Table 4; [12]) for their processing by OYNYL. The results for Rb-^{137}Cs pair are shown in Fig. 2. The upper part of Fig. 2 lists a part of the input data (n=15) used for running OYNYL (note only a part of these data are shown in Fig. 2 although all of them are saved in the computer file); the middle part gives the statistical results for the three regressions, in which "No Outlier" means no outlying observations were detected in this dataset. The word "BEFORE" in the first row (see the middle part of Fig. 2) means that these results correspond to the statistical results *before* the application of "Outlier detection". The rows 2 to 4 present the results of the three types of linear regressions (Ordinary, York, and New York). The linear correlation coefficient, r estimated for the Rb-^{137}Cs pair (0.91163) is practically the same as that calculated by the original authors [12].

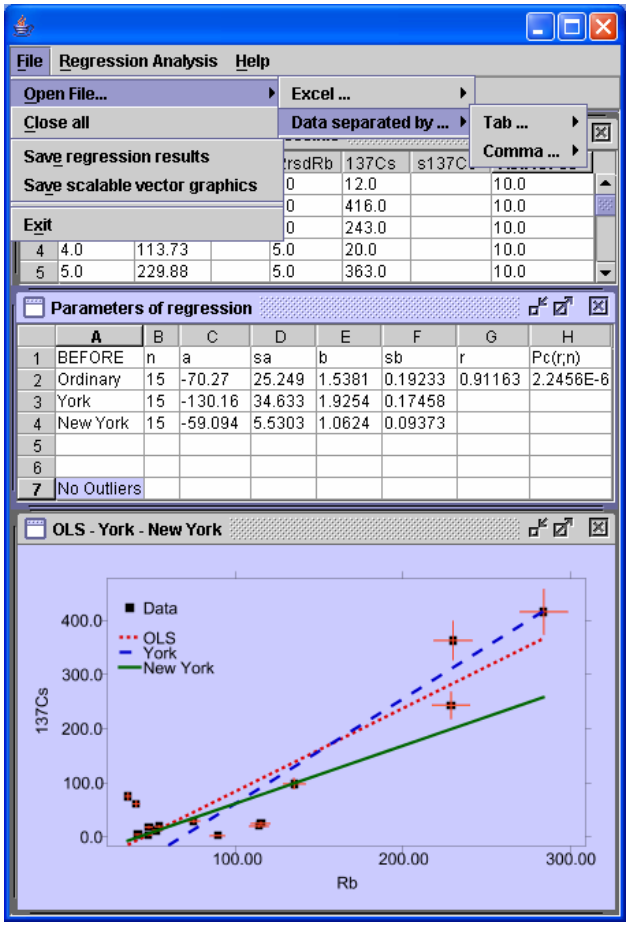

Fig. 2. Output screen from OYNYL for the Rb-^{137}Cs pair for 15 mushroom species from Mexico [12]. The first part of the "Menu" options for "File" are also clearly seen in the upper part of this three-part diagram.

The final column in this second part of Fig. 2 gives the "probability of no-correlation", $P_c(r;n)$. Note a small value (2.2456E-6) of this parameter is equivalent to a high probability that a linear correlation exists between these two parameters (Rb and ^{137}Cs). Because errors were not reported for individual data (only Rsd values were indicated by

[12]), we processed the Rb and ^{137}Cs data assuming 5% and 10% errors, respectively. Generally, such equal %Rsd are not realistic for chemical data, and error estimates on individual data should be obtained and reported [2,13]. The feasibility of reporting individual errors has been shown recently for chromatographic data [9]. This practice will certainly lead in future to a better use of the regression models, such as those programmed in OYNYL, in different fields of science and engineering.

The "File" menu is also purposely shown in the upper part of this diagram (Fig. 2). The "Open file" option allows a data file to be opened and processed. The next submenu linked to this option enables the user to select an "Excel" or a txt file ("Data separated by …" either "Tab" or "Comma").

The "Close all" option helps to close all windows, including the input data file, for a new data file to be opened and processed. The "Save regression results" serves the purpose of saving all output data in a new file with the name "_out" automatically added to the initial data file; this output file is saved in the same file folder as the input data file.

The "Save scalable vector graphics" option allows the user to save the graph or graphs generated in a given application. The file name is the same as the input data file, with the "_graphic1" added to this name. This is the graphics file for the output generated before (see "BEFORE" in the second part of the diagram; Fig. 2) the "Outlier detection" (see Fig. 1 for more details). If no outliers are detected, this is the only graph file that can be generated in a given application. However, if one or more outliers are present, these are automatically deleted before a second round of processing, which is also automatically carried out. In this case, a second graphic output is also generated (with the name "_graphic2" added to the initial file name). Finally, the "Exit" option is for closing the entire application.

The graph in the lower part of Fig. 2 shows the three regression lines obtained for the Rb-^{137}Cs pair for mushroom data. The inferred error-bars on the experimental data are also shown, which actually explain why the three types of regressions result in different best-fit lines.

3.2 Application Example of Ecosystems:

Biomass estimation equations were derived by [14] in a study of thornscrub from Tamaulipa, north-eastern Mexico. Because individual data were not available, average and standard deviation data for 10 shrub species based on 15 measurements from each species (Table 1; [14]) were processed using OYNYL. We fitted linear regressions to log-transformed data (natural logarithm ln) from Table 1 of [14]. To illustrate the use of OYNYL, we present in Fig. 3 the results for the basal diameter-top height (DB-H) pair for the ten shrub species.

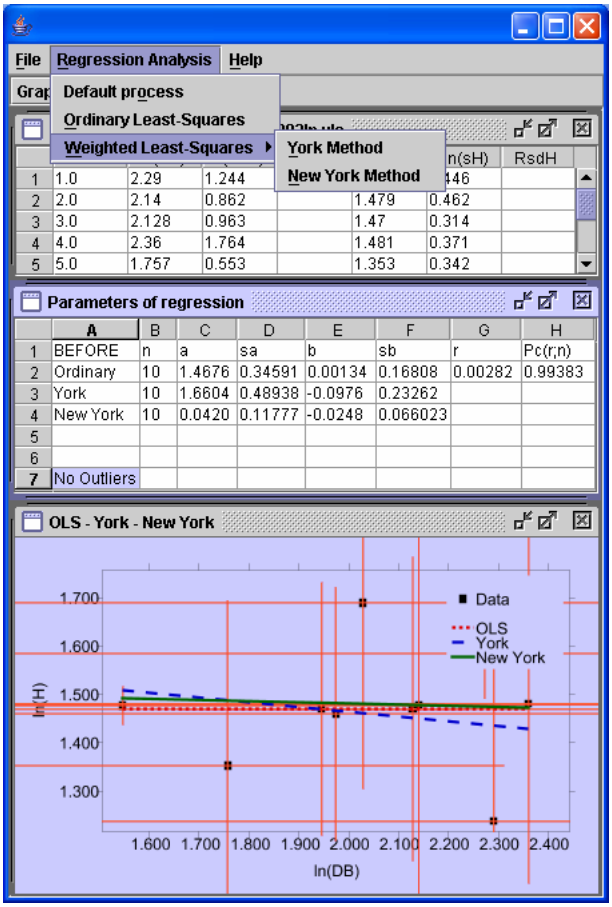

Fig. 3. Output from OYNYL for DB-H pair (DB – basal diameter; H – top height) for 10 shrub species from Mexico [14].

Because individual data were not available, the results could not be directly compared with those obtained by [14] using the log-transformed linear regressions. This application shows that the linear regression of log-transformed base diameter and height data is not statistically significant (see the very low value of r = 0.00282 and the very high value of $P_c(r;n)$ = 0.99393 in the first row of the middle part of the diagram, Fig. 3). The sub-horizontal regression lines are the further proof of a statistically invalid linear correlation between these two variables. We must, however, mention that the rather large standard deviation values reported by [14] for the shrub data had to be taken as the assumed experimental errors of these regressions

(Fig. 3). In fact, actual experimental errors on individual data should be used whenever available.

In Fig. 3 we have purposely opened the details of the "Regression analysis" menu. It has three options: (i) Default process – indicates that OYNYL will commonly carry out all the three types of regressions if no specific action is chosen by the user; (ii) Ordinary Least-Squares – this is the standard OLS model; and (iii) Weighted Least-Squares – when the WLS model is selected, it has the further sub-option of choosing between the York method [3] and the New York method [4].

3.1 Application Example of Development:

Our third application was for aboveground biomass and wood volume data for samples of trees from the Rio Negro farm in Nhecolandia Pantanal, Mato Grosso do Sul, Brazil [15]. Once again, because the original raw data were not available, we used the processed data for biomass percentage (%) of tree components for two sets of trunk-branch pairs (combined data for 5-15 and 15-25 cm diameter) reported in Table 2 of [15]. The results are shown in Fig. 4.

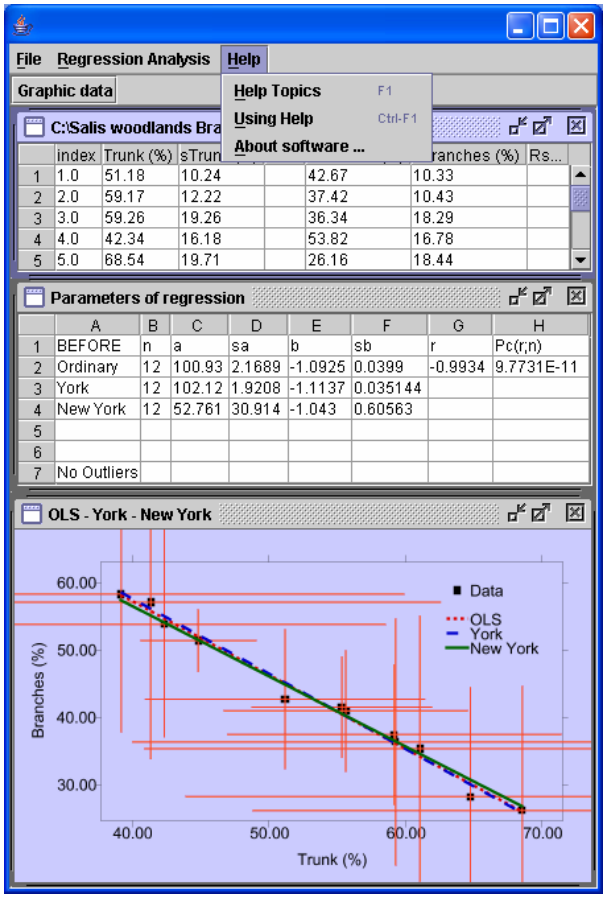

Fig. 4. Output from OYNYL for trunk-branch pair for five species and one group of 11 species (n=12) from Brazil [15].

The authors [15] used a commercial software, but because of the unavailability of the raw data to us, our results could not be compared to those obtained from this commercial software. In Fig. 4, the inverse correlation is fully expected given the complementary nature of the percentage data. This kind of data belongs to a "closed" system and the "closure" problem should be dealt in a different way by log-ratio transformations [2,16] before the regression analysis.

Also in Fig. 4, the "Help" menu is purposely activated. It provides three help options: (i) Help Topics – allows a quick search on a specific topic; (ii) Using Help – gives some details on how to use the Help option; (iii) About software - includes some clarifications about the OYNYL software, such as the literature references for further reading.

4 Discussion

The different applications selected for the illustration of OYNYL clearly show the high potential of this software. Through this software, it has been possible to bring at the same place three different types of linear regressions used by researchers in different fields of science and engineering.

The next step in computer programming should be to incorporate into the software more statistical tests for the linearity of data, in addition to the $P_c(r;n)$ criterion [2,6]. Concerning the experimental data collection, on the other hand, we highly recommend that researchers start estimating and reporting reliability (errors) of individual data [2,9,13]. This will certainly enable the user of OYNYL to take advantage of the full potential of this software.

5 Conclusions

A new computer program OYNYL was successfully developed and applied to three different case studies in the areas of environment, ecosystems, and development. It is envisioned that OYNYL should be very useful for bi-variate data handling in all science and engineering fields, including the environment, ecosystems, and development.

6 Acknowledgements

We are grateful to A. Quiroz-Ruiz for an efficient computer maintenance of the Geoenergy group. E. Santoyo thanks the project DGAPA-PAPIIT: IN104703-3 for partial support. P. Sánchez-Upton is grateful to the Comisión Federal de Electricidad (CFE) for granting permission to participate in this work at CIE-UNAM. We are also thankful to two

anonymous reviewers for evaluating an earlier version of our paper.

References:

[1] M. Guevara, S.P. Verma, F. Velasco-Tapia, R. Lozano-Santa Cruz, and P. Girón, Comparison of Linear Regression Models for Quantitative Geochemical Analysis: An Example using X-Ray Flourescence Spectrometry, *Geostandards and Geoanalytical Research*, Vol. 29, No. 3, 2005, pp. 271-284.

[2] S.P. Verma, *Estadística Básica para el Manejo de Datos Experimentales: Aplicación en la Geoquímica (Geoquimiometría)*, UNAM, Mexico, 2005.

[3] D. York, Least-squares Fitting of a Straight Line with Correlated Errors, Earth and Planetary Science Letters, Vol. 44, No. 10, 1969, pp. 1079-1086.

[4] K.L. Mahon, The New "York" Regression: Application of an Improved Statistical Method to Geochemistry, International Geology Review, Vol. 38, No. 4, 1996, pp. 293-303.

[5] P.R. Bevington and D.K.Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, Boston, 2003.

[6] J. Andaverde, S.P. Verma, and Santoyo E., Uncertainty estimates of static formation temperatures in boreholes and evaluation of regression models. *Geophysical Journal International*, Vol. 160, No. 3, 2005, pp. 1112-1122..

[7] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, Chichester, 1994.

[8] S.P. Verma, J. Andaverde, and E. Santoyo, Statistical Evaluation of Methods for the Calculation of Static Formation Temperatures in Geothermal and Oil Wells using an Extension of the Error Propagation Theory. *Journal of Geochemical Exploration*, Vol. 89, No. 1-3, 2006, pp. 398-404.

[9] E. Santoyo, M. Guevara, and S.P. Verma, Determination of lanthanides in international geochemical reference materials by reversed-phase high performance liquid chromatography: An application of error propagation theory to estimate total analysis uncertainties, *Journal of Chromatography A*, Vol. 1118, No. 1, 2006, pp. 73-81.

[10] J.N. Miller and J.C. Miller, *Statistics and chemometrics for analytical chemistry*, Prentice Hall, Essex, 2000.

[11] F.W. Walker, G.J. Kirouac, and F.M. Knolls, *Chart of the Nuclides*, General Electric, San Jose, 1977.

[12] M.I. Gaso, N. Segovia, O. Morton, M.L. Cervantes, L. Godinez, P. Peña, and E. Acosta, ^{137}Cs and relationships with major and trace elements in edible mushrooms from Mexico, *Science of the Total Environment*, Vol. 262, No. 1-2, 2000, pp. 73-89.

[13] P. De Bièvre, Measurement results without statements of reliability (uncertainty) should not be taken seriously, Accreditation and Quality Assurance, Vol. 2, No. 6, 1997, p. 269.

[14] J. Návar, J. Nájera, and E. Jurado, Biomass Estimation Equations in the Tamaulipan Thornscrub of North-Easterm Mexico, *Journal of Arid Environments*, Vol. 52, No. 2, 2002, pp. 167-179.

[15] S.M. Salis, M.A. Assis, P.P. Mattos, and A.C.S. Pião, Estimating the Aboveground Biomass and Wood Volume of Savanna Woodlands in Brazil's Pantanal Wetlands based on Allometric Correlations, *Forest Ecology and Management*, Vol. 228, No. 1-3, 2006, pp. 61-68.

[16] S.P. Verma, M. Guevara, and S. Agrawal, Discriminating four tectonic settings: Five new geochemical diagrams for basic and ultrabasic volcanic rocks based on log-ratio transformation of major-element data, *Journal of Earth System Science*, in press, 2006.