Journal of Environmental Science and Engineering B 7 (2018) 103-110
doi:10.17265/2162-5263/2018.03.003

Stepwise Regression: An Application in Earthquakes Localization

Giuseppe Pucciarelli
Dept. of Physics "E. R. Caianiello", University of Salerno, Fisciano 84084, Italy

Abstract: In this paper, an overview of an important feature in statistics field has shown: the stepwise multiple linear regression. Likewise, a link between stepwise multiple linear regression and earthquakes localization has been descripted. Precisely, the aim of this research is showing how stepwise multiple linear regression contributes to solution of earthquakes localization, describing its conditions of use in HYPO71PC, a software devoted to computation of seismic sources' collocation. This aim is reached treating a concrete case, that is computation of earthquakes localization happening on Mount Vesuvius, Italy.

Key words: Stepwise regression, earthquakes localization, Geiger's method, HYPO71PC, Mount Vesuvius.

1. Introduction

Linear regression is probably the simplest and, at same time, the most well-known method of research of a mathematical relation between an explanatory variable and a dependent one of physical quantities. A particular specification is about the case in which explanatory variables are more than ones, that is the case of multiple linear regression. Basic concept is to establish a priori that physical quantities which are at stake are linked with each other in a linear way by means of a mathematical relation, named linear model. The goal is finding the parameters of this model, starting from observed data. A particular type of these statistical models is the so-called SMLR (Stepwise Multiple Linear Regression). In this regression typology, computation of model's parameters happens by means of an algorithm. In this paper, a description of *F*-test as a good criterion for implementation of this linear regression typology has been shown. Otherwise, use of SMLR to solve an important physical problem, earthquakes localization, has been shown. Indeed, to establish in a exact way geographical coordinates of a seismic event is very relevant to obtain all the essential information about it. In particular, an essential point is finding geographical coordinates of seismic events which happen on volcanic areas. If these events happen in a tight period of time and their hypocenters are collocated along volcanic pipe, it could possibly be an imminent magma eruption. This is the reason because an application of computing earthquakes localization in Mount Vesuvius by means of SMLR has been shown. Mount Vesuvius is an active volcano which is placed about 9 km east of Naples, Italy: since a population of 3,000,000 people live in its foothills, it is one of more monitoring volcanoes in the world. The entire process to locate specific earthquakes which occurred along Mount Vesuvius crater by means of SMLR technique (or better yet, by means of a software named HYPO71PC based on SMLR technique) has been described. In this paper, this kind of explanation is preceded by a generic overview about earthquakes localization problem and it is followed by a concise description of the applications which have earthquakes localization as starting point.

Corresponding author: Giuseppe Pucciarelli, Ph.D., research fields: geophysics, volcanology and numerical simulation.

2. Statistical Model: The Stepwise Regression

Mathematically speaking, the regression is a statistical process which allows obtaining an estimate of an assumed value from a dependent variable Y as a function of n independent variables X_1 , ..., X_n. The most common type of regression is the linear one. This is based on the idea that for each sample of observations, there has been a determination of the variable Y and n determinations of the variables X_1 , ..., X_n. The goal is to find a linear relation between these variables. As is well-known, the most simple linear relation is that given by an affine transformation linear that has the Eq. (1):

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i \tag{1}$$

where i varies between the observations, i.e. $i = 1$, ..., n; Y_i are the dependent variables; X_i are the independent variables, called regressors; $\beta_0 + \beta_1 X_i$ is the regression line, called also regression function; ε_i represents the statistical error.

Eq. (1) has the Eq. (2):

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

in which the vectors Y, X and ε are n-dimensional. The values β_0 and β_1 represent the parameters of the model which have to be estimated by the obtained observation samples (so, they are deterministic); instead, the vector ε contains only stochastic parameters to be determined, because for every assumed value from X, it exists in an entire probability distribution that contains values of Y (for this reason, it is impossible to know with certainty the value of Y). Thus, the relation Eq. (2) gives a random variable, where its probability distribution and characteristics are determined from the values of X and the probability distribution of ε. Among all of the methods of linear regression known, the most used is surely the method of least squares.

The type of linear regression as soon as view is said simple linear regression, because it is available to have only one independent variable. If instead, there are more independent variables that can be used for the determination of the dependent variable Y, we talk about multiple linear regression. In this case, the equation for the previous problem, becomes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \tag{3}$$

where with respect to Eq. (1), the only change is that the regression line is $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$. This equation has a matrix form:

$$Y = X\beta + \varepsilon \tag{4}$$

in which X is a $k \times n$ matrix, while the vectors Y, ε are n-dimensional and β is k-dimensional. A type of multiple linear regression very used for the applications in the last years is the so-called stepwise regression [1], discovered from Draper and Smith in 1966 [2]. It is especially used when, in the process of regression, problem is solved by means of independent variables which fail to give information on the variance of the dependent variable: stepwise regression is able to solve this sort of problems. In this method, the total variance is partitioned into two parts: the first one which can be explained by the independent variables (explained variance) and the second one which cannot be explained by them (unexplained variance) [3, 4]. In order to do this, given n indipendent variables X_1, ..., X_n, deemed suitable to the explanation of the variance of the dependent variable, it is possible the extraction step by step (hence the name) of the independent variable that has the highest percentage of the explained variance. Now considering the step $k + 1$: the variabile X_{k+1} is related with X_k of the previous step: thus, Y_k is called residual variable, because it is given by Y minus the linear combination of the independent variables already present in the model. In practice:

$$Y_k = Y - \beta_0 - \beta_1 X_1 - \ldots \beta_k X_k \tag{5}$$

The variable X_{k+1} introduced in the model will be the one that will show the highest correlation coefficient, in absolute value, with the residual variable. After establishing the criterion for choosing, an algorithm chooses the independent variables suitable for the explanation of the variance of the dependent variable, after an initial setting of a threshold value with which comparing the diffierent correlation coefficients [1, 2]. However, this procedure is not completely reliable (even considering that the threshold value is subjective), therefore the diffierent correlation coefficients have to satisfy the following null hypothesis:

$$H_0 : b_k + 1 \qquad (6)$$

where b_k+1 is the so-called regression coefficient, i.e. the slope of a line obtained using linear least squares fitting [5, 6]. So, Eq. (6) affirms that the regression coefficient of the independent variable X_{k+1} to input in the model is zero, after the introductions of $X_1, ..., X_k$. This hypothesis is verified through the F value of Fisher, given by

$$F = \frac{MSR}{MSE} = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{m - k - 1}} \qquad (7)$$

in which MSR and MSE are respectively the mean square due to regression and residual; furthermore, the coefficient of determination R^2 is given from:

$$R^2 = 1 - \frac{VAR_{res}}{VAR_{tot}} \qquad (8)$$

where the numerator (respectively denominator) is the sample variance of the estimated residuals (respectively dependent variable). It is clear that $0 \leq R^2 \leq 1$ [7].

3. Limits of the Model and Implementation of the Algorithm

The F value of Fischer is based on a single hypothesis: in the stepwise regression, it is hugely violated [7]. As a result, models are too complex and other parameters could be biased towards 0. Therefore, diffierent alternatives for the stepwise regression are developed. The following methods perform better than stepwise, but their use is not appropriate for statistical knowledge [8]: the most important are LASSO, LAR, CROSS VALIDATION and some experimental procedures (not yet enough reliable). Even though no one methods could substitute for statistical experience, LASSO and LAR (especially) are much better than stepwise: these two methods allow us to consider a wide range of options [8]. Through the F value of Fischer, the model with the introduction of the independent variable X_{k+1} is still statistically significant compared to the previous one, which included the independent variables X_1, ..., X_k: this value is also called F to enter. The introduction of new variables could, however, result in an information loss for some variables introduced previously: to check this phenomenon, authors compare the correlation coefficient with another threshold value, known as F to remove: if they are present, these are immediately deleted. Usually, almost all programs have the possibility to follow step by step the regression procedure, through information about the variables and the signalling of F to enter and F to remove which the program utilizes for this particular step. A further option frequently present in the various programs concerns the possibility of choice, on the part of the user, of variables to be introduced or delete in the model: this choice is made through the control of the values of F to enter and F to remove.

4. Earthquakes: The Geiger's Method

Earthquakes' localization constitutes a particular scientific problem where the goal is computing the coordinates of a point situated inside the Earth from which seismic waves are generated. This point is known as the hypocenter of an earthquake and its outside projection is named epicenter. Hypocentral coordinates which authors want to ascertain are four: three are spatial (latitude, longitude and depth) and the other one is the origin time of seismic waves generation. Historically, earthquakes localization was carried out by means of analysis of outside consequences produced by a seismic event. Now, thanks to seismic network development and a more detailed knowledge of interior structure of the Earth, computing of the four earthquakes localization parameters could be done in a more meticulous way. Earthquakes' localization is based on classical linear motion law, that is:

$$x = vt \qquad (9)$$

where:

- x is the hypocentral distance, that is the distance between hypocenter and seismic station;
- v represents the seismic waves velocity;
- t is the seismic waves traveltime.

Describing the process in a more practical way, earthquakes localization needs three important features, i.e.:

- knowledge of seismic stations' geographical coordinates;
- picking of seismic phases on seismograms and computing of their traveltimes;
- knowledge of velocity structure between hypocenter and seismic stations.

Therefore, earthquakes' localization is a typical inverse problem, that is a problem in which solution consists in an evaluation of parameters of a determinate model (they must be in good agreement with observations). In this particular situation, parameters are the hypocentral coordinates, while the observations are picked seismic phases and model is an Earth hint between hypocenter and seismic stations. Rewriting Eq. (9) for seismic phases P and S, respectively, that is the first waves generated by a hypocenter which arrive to a seismic station, following equations are obtained:

$$\Delta = v_P(t_P - t_0);\ \Delta = v_S(t_S - t_0) \qquad (10)$$

where Δ is the hypocentral distance, v_P and v_S are the velocities of P wave and S wave respectively, t_P and t_S are the P traveltime and the S traveltime and t_0 is the seismic event origin time. Manipulating in an opportune way Eq. (10), Eq. (11) is obtained:

$$t_S - t_P = \frac{\Delta(v_P - v_S)}{v_P v_S} \qquad (11)$$

Replacing Eq. (10) in Eq. (11), Eq. (12) has been obtained:

$$t_S - t_P = [\frac{v_P}{v_S} - 1]t_P - [\frac{v_P}{v_S} - 1]t_0 \qquad (12)$$

Eq. (12) can be understood as an equation of a line in a cartesian plane where t_P is the abscissa and $(t_S - t_P)$ is the ordinate. In this way, through a simple linear interpolation, parameters can be found. By means of slope, it is possible to compute v_P/v_S , while the y-intercept allows the computation of seismic event origin time t_0. From values of these specific parameters, it is easy to obtain Δ through Eq. (11). Earthquakes' localization can be realised in a graphic way by means of so-called circle method [9]. If at least three seismic stations are at our disposal, for each station the epicentral distance has been computed. Then, a circle with the centre arranged on a seismic station and with radius equivalent to epicentral distance is drawn. The point of intersection among all the circles is the localization of an earthquake. A modification of this method is the following. If only P seismic phases and three stations are computed, circle method is applied in the same way as previously described for two of three stations. In this case, earthquakes' localization is the centre of a circle tangent at two drawn circles and passing through the third remaining seismic station. Both linear interpolation and circle method show us the following set of problems:

- Earth between hypocenter and seismic stations is considered as a homogeneus halfspace where the velocity of seismic waves is constant. Obviously, this is not the real case.
- Earth's depth is completely disregarded.
- Consideration of a spherical Earth is completely disregarded.

In 1912, L. Geiger introduced an algorithm to solve the problem of earthquake localization, considering three criticisms mentioned before. This method, justly called Geiger's method, is still used as basis of software for earthquakes localization [10]. Starting point of this method is model of representing Earth as a sequence of parallel layers where in each one of them, seismic waves velocity is constant. Starting from a trial solution for localization, it is named $m_0(x_0, y_0, z_0, t_0)$. After, there is the procedure of computation of the so-called residuals. For the i-th seismic station, Eq. (13) is defined:

$$R_i = T_i - t_i \quad (13)$$

where R_i is the i-th residual that is the simple diffierence between theoretical traveltime T_i and observed (by means of seismograms interpretation) traveltime t_i. Writing T_i in function of m_0, Eq. (13) becomes:

$$R_i = T_0 + \frac{1}{v}[(x_0 - x_i)^2 + (y_0 - y_i)^2 + (z_0 - z_i)^2] - t_i \quad (14)$$

where v is the velocity of considered seismic wave.

Now, rewrite Eq. (14) and apply on it an expansion in Taylor series at m_0. In this way, Eq. (14) becomes:

$$R_i \sim R_i(m_o) + a_i \Delta x + b_i \Delta y + c_i \Delta z + \Delta t \quad (15)$$

where $a_i = \frac{\partial R_i}{\partial x_0}$, $b_i = \frac{\partial R_i}{\partial y_0}$, $c_i = \frac{\partial R_i}{\partial z_0}$ with all partial derivatives calculated in m_0 and $\Delta x = x - x_0$, $\Delta y = y - y_0$, $\Delta z = z - z_0$, $\Delta t = t - t_0$.

Finally, a least squares method to make residual minimum is used. Therefore, hypocentral parameters are computed when $\sum_{i=1}^{n} R_i^2$ is minimum with n seismic stations.

Summarising, to localize an earthquake, this procedure should be:

(1) Acquisition of seismic phases and of observed traveltimes by means of seismograms.

(2) Geographical coordinates of seismic station which have minimum observed traveltime become spatial coordinates of trial solution m_0. While authors assume as t_0, the P seismic wave traveltime registered at that seismic station deprived of a quantity depending on velocity model authors choose.

(3) Computation of residuals.

(4) Appropriate correction of trial solution.

(5) New computation of residuals.

(6) Repetition of (4) and (5) until when $\sum_{i=1}^{n} R_i^2$ is minimum.

It is appropriate to emphasize how solution's goodness depends on an adequate seismic rays coverage, a sufficient knowledge of velocity model and a careful picking of seismic phases.

5. Application: Examples of Earthquakes Localization on Mount Vesuvius

Mount Vesuvius is an active stratovolcano which is situated about 9 km east of Naples, Italy, and a short distance from the shore. Since a population of 3,000,000 people live in its vicinity, it is one of more monitoring volcanoes in the world. And one of monitoring strategies is earthquakes localization. Indeed, earthquakes localization is very important in volcanic areas because a contingent high concentration of seismic events could be due to magma rising along the volcanic pipe. Magma rising causes a gradual deformation of volcanic pipe rocks which can trigger seismic events. And often magma rising could forerun an eruption. Therefore, accuracy of earthquakes localization in volcanic areas could prove to be fundamental. In this paper, a computed earthquakes localization of 2,315 seismic events which occurred on Mount Vesuvius in a period from 06/12/1986 to 31/12/1999 and was registered by 41 seismic stations belonging to OV-INGV Mount Vesuvius monitoring network is shown. It is important to underline that only P and S seismic phases have been chosen for localization. Starting point is choice of an opportune velocity model to schematise Earth. This last one is a standard 1-D velocity model for Mount Vesuvius area (see Fig. 1) which has already been in other previous works with satisfactory results. This model schematises Earth in a depth interval between 0 and 30 kilometers. Model is composed of six layers and in each of them P wave velocity is assumed to be constant. Then, earthquakes localization is computed by means of HYPO71PC software. This software is a FORTRAN code and it represents one of more sofisticated versions of original code HYPO71PC. Other similar versions are HYPOELLIPSE and HYPOINVERSE. This software is nowadays widely used for earthquakes localization. The reason for this is the following one. This software makes use of stepwise multiple linear regression to carry out six points of earthquakes localization procedure which have been previously described in Chapter 5. In particular, it carries out the point (6) and it shows lots of benefits in accuracy and convergence of solution. In this particular case, an input file named VESUVIO.PHA is supplied to Hypo71PC. This file is organised in the following way: First of all, it presents the command HEAD followed by the number of characters used to write a heading above each earthquake in the output. This command is put equal to 20. Then, the command RESET is written. This command allows setting values of the software test variables in a way which is straightly connected to input data. In this particular case, seven test variables are test. Among these, the critical *F* value for stepwise multiple regression is set. In Hypo71PC, this value should be included between 0.5 and 2 according to number and quality of available P and S traveltimes. Indeed, if this value is lesser than 0.5, then coefficients matrix of Eq. (15) could be ill-conditioned. Instead, if this value is greater than 2, then Geiger's iteration may finish prematurely and therefore a good localization could be not obtained. For avoiding it, critical *F* value is set equal to 0.5 because there are not lots of available S traveltimes. Then, geographical coordinates (latitude and longitude) of seismic stations which have recorded earthquakes to localize must be written. After the realization of this procedure, velocity model must be choosen. Hypo71PC procures two options: station delay model and variable first layer model. In first case, there is the simple addition or subtraction to traveltimes of a number which is stated for each seismic station. In second one, this stated quantity—previously described—is interpreted as a thickness to add to first layer. Therefore, each seismic station has the same P-wave velocity but thickness of first two layers under each seismic station varies. This typology of model is often chosen when there is the case of different travel paths, that is to say, paths do not come from an uniform reason. Fortunately, this is not the case and therefore station delay model is chosen. Following step is writing a record in which parameters of trial solution and control indicators for the software are indicated. In these parameters of the trial focal depth, the distance from epicenter and others concerning the earthquakes localizations are set (see Ref. [11]). It is important to underline how there are other possible options that can be activated, especially, options concerning geographical coordinates of a trial solution, but this is not the case. Therefore, for each earthquake, Hypo71PC chooses as geographical coordinates of trial solution those of seismic station which is nearest to earthquake localization. Finally, in the last part of input file traveltimes are found. They are organised in the following way. For each earthquake, this is the structure of the script: the acronym of seismic station which has recorded that earthquake, the symbol of analysed seismic phase (that is, P), P-traveltime at that station with accuracy down to centiseconds and, when they have been recorded, S-traveltime at that station with accuracy down to centiseconds. When all traveltimes records regarding to a specific earthquake have been written, then the character "10" in the record immediately following is written. This particular record is considered as a pause between traveltimes of a seismic event and traveltimes of another one. The output file is named VESUVIO.PRT: it is divided into two parts, where in the first one the most important thing is given from the root mean square error of time residuals, that is:

$$RMS = \sqrt{\frac{R_i}{NO}} \tag{16}$$

where R_i is the quantity described in Eq. (14), while *NO* is the number of station readings used in locating the earthquake. The second part of output file is about the characteristics of seismic station: for further information (see Ref. [11]). The result of earthquakes

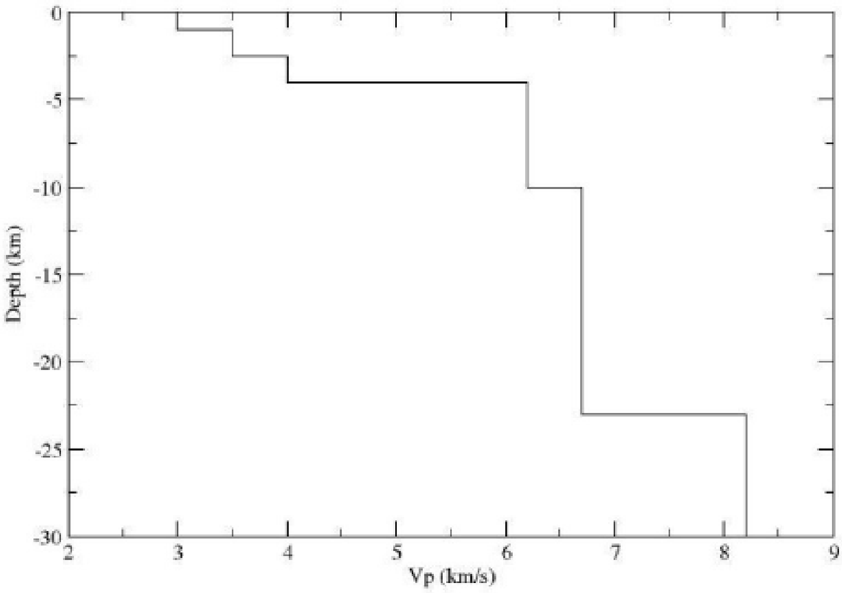

Fig. 1 1D velocity model for Mount Vesuvius area used as input for computing earthquakes localizations.

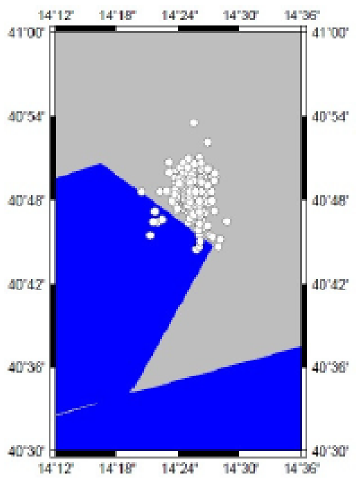

Fig. 2 Earthquake localization in Mount Vesuvius area: the hypocenters are represented by these white circles in the figure.

localization in Mount Vesuvius area is shown in Fig. 2.

6. Conclusions

In this paper, an interesting characteristic of an important topic in statistical field is described: the SMLR. Especially, the aim is to underline a particular link between statistics and physics, explaining how SMLR can be used in the solution of a very important physical problem: the earthquakes localization. Indeed, SMLR is used to implement Geiger's method [10] by means of a subroutine of software HYPO71PC [11, 12]. Its mode operation has been described through a concrete example as computation of some Mount Vesuvius earthquakes localization. Understanding how SMLR could be connected with this relevant physical problem is very important because a good earthquakes localization represents a starting point for other geophysical research techniques as seismic tomography [13, 14], focal mechanism implementation [15], good resolution of epicentral distance [16] and obtaining of graphical images by means of GMT (Generic Mapping Tools) [17]. Of course, other methods for locating earthquakes exist, for example, Metropolis-Gibbs Method [18]. It is based on a 3D velocity model. This kind of method has already been used to locate earthquakes on Mount Vesuvius [19]. Method which has been described in this paper where SMLR is implemented is based on a 1D velocity model. But 3D velocity model tests have already been carried out [20]. For future studies, development of these 3D velocity model tests using data treated in this paper could be appropriate.

This map is given by the geographical coordinates, i.e. latitude and longitude.

References

[1] Burnham, K. P., and Anderson, D. R. 2002. *Model Selection and Multimodel Inference.* New York: Springer.

[2] Wessel, P., and Smith, W. H. F. 1991. "Free Software Helps Map and Display Data EOS." *Advancing Earth and Space Science* 72 (41): 441-6.

[3] Cohen, J., and Cohen, P. 2010. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edition, Taylor & Francis e-Library.

[4] Chatterjee, S., and Hadi, A. S. 2012. *Regression Analysis by Example*, 5th edition, John Wiley & Sons.

[5] Hansen, P. C., Pereyra, V., and Scherer, G. 2012. *Least Squares Data Fitting with Applications.* Baltimore: Johns Hopkins University Press.

[6] Lawson, C., and Hanson, R. 1987. *Solving Least Squares Problems, Series: Classics in Applied Mathematics*. Book 15, Society for Industrial and Applied Mathematics.

[7] Harrell, F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag.

[8] Flom, P. L., and Cassell, D. L. 2007. "Stopping Stepwise: Why Stepwise and Similar Selection Methods Are Bad, and What You Should Use." Presented at Northeast SAS Users Group (NESUG) Annual Conference.

[9] Piana Agostinetti, N., and Chiarabba, C. 2008. "Seismic Structure beneath Mt. Vesuvius from Receiver Function Analysis and Local Earthquakes Tomography: Evidences for Location and Geometry of the Magma Chamber." *Geophys. J. Int.* 175 (3): 1298-308.

[10] Geiger, L. 1912. "Probability Method for the Determination of Earthquake Epicenters from Arrival Time Only." *Bulletin St. Louis University* 8: 60-71.

[11] Lee, W. H. K., and Valdes, C. M. 1985. "HYPO71PC: A Personal Computer Version of the HYPO71 Earthquake Location Program." U.S Geological Survey, Open File Report 85749.

[12] Lee, W. H. K., and Lahr, J. C. 1975. *"HYPO71: A Computer Program for Determining Hypocenter, Magnitude and first Motion Pattern of Local Earthquakes*." U. S. Geological Survey, Open File report, pp. 75-311.

[13] Lahr, J. C. 1979. "HYPOELLIPSE: A Computer Program for Determining Hypocenter, Magnitude and first Motion Pattern of Local Earthquakes." U. S. Geological Survey, Open File report, pp. 79-431.

[14] Seber, G. A. F., and Lee, A. J. 2003. *Linear Regression Analysis*, 2nd edition. Hoboken, NJ: John Wiley & Sons.

[15] Labay, A., and Haeussler, P. J. 2007. "3D Visualization of Earthquake Focal Mechanism Using Arcscene." U. S. Geological Survey.

[16] Wallace, T. C., and Lay, T. 1995. *Modern Global Seismology*. San Diego: Academic Press.

[17] Wessel, P., Smith, W. H. F., Scharroo, R., Luis, J. F., and Wobbe, F. 2013. "Generic Mapping Tools: Improved Version Released." *EOS, Transaction AGU* 94: 409-10.

[18] Lomax, A., Virieux, J., Volant, P., and Berge, C. 2000. *Probabilistic Earthquake Location in 3D and Layered Models: Introduction of a Metropolis-Gibbs Method and Comparison with Linear Localizations,* edited by C. H Thurber, N. Rabinowitz, Kluwer, Amsterdam: Advances in Seismic Event Location, pp. 101-34.

[19] Lomax, A., Zollo, A., Capuano, P., and Virieux, J. 2001. "Precise, Absolute Earthquake Location under Somma-Vesuvius Volcano Using a New 3D Velocity Models." *Geophys. J. Int.* 146: 313-31.

[20] Chen, H., Chiu, J. M., Pujol, J., Kwanghee, K., Chen, K. G., Huang, B. S., et al. 2005. "A Simple Algorithm for Local Earthquake Location Using 3D Vp and Vs Models: Test Examples in the Central United States and in Central Eastern Taiwan." *Bull. Seism. Soc. Am.* 96 (1): 288-305.