Comparative study on parameter estimation methods for attenuation relationships

Article *in* Journal of Geophysics and Engineering · October 2016

DOI: 10.1088/1742-2132/13/6/912

2 authors:

Farhad Sedaghati
The University of Memphis
10 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Shahram Pezeshk
The University of Memphis
104 PUBLICATIONS **1,228** CITATIONS
SEE PROFILE

Comparative study on parameter estimation methods for attenuation relationships

2016 J. Geophys. Eng. 13 912

(http://iopscience.iop.org/1742-2140/13/6/912)

IOP Publishing Journal of Geophysics and Engineering
J. Geophys. Eng. **13** (2016) 912–927 doi:10.1088/1742-2132/13/6/912

Comparative study on parameter estimation methods for attenuation relationships

Farhad Sedaghati and Shahram Pezeshk

Department of Civil Engineering, The University of Memphis, Memphis, TN 38152, USA

E-mail: spezeshk@memphis.edu

Received 22 September 2015, revised 10 August 2016
Accepted for publication 14 September 2016
Published 14 October 2016

Abstract

In this paper, the performance and advantages and disadvantages of various regression methods to derive coefficients of an attenuation relationship have been investigated. A database containing 350 records out of 85 earthquakes with moment magnitudes of 5–7.6 and Joyner–Boore distances up to 100 km in Europe and the Middle East has been considered. The functional form proposed by Ambraseys *et al* (2005 *Bull. Earthq. Eng.* **3** 1–53) is selected to compare chosen regression methods. Statistical tests reveal that although the estimated parameters are different for each method, the overall results are very similar. In essence, the weighted least squares method and one-stage maximum likelihood perform better than the other considered regression methods. Moreover, using a blind weighting matrix or a weighting matrix related to the number of records would not yield in improving the performance of the results. Further, to obtain the true standard deviation, the pure error analysis is necessary. Assuming that the correlation between different records of a specific earthquake exists, the one-stage maximum likelihood considering the true variance acquired by the pure error analysis is the most preferred method to compute the coefficients of a ground motion predication equation.

Keywords: regression methods, attenuation relationship, Europe and the Middle East, pure error analysis, ground motion predication equation

(Some figures may appear in colour only in the online journal)

Introduction

Ground-motion prediction equations (GMPEs) are widely used to estimate ground motion intensity measures (GMIMs) such as peak ground acceleration (PGA) and spectral accelerations (SA) at different periods. Regression analysis is performed to detect a functional form between variables. The rudimentary formulation of GMPEs is given by

$$Y = f(X_{es}, \theta) + \Delta, \tag{1}$$

where Y is often the natural logarithm or the base 10 logarithm of the ground motion observation, $f(X_{es}, \theta)$ is the ground motion regression model, in which records information are included in X_{es} such as magnitude, distance, fault mechanism, and site condition, and θ is the vector of unknown model coefficients. The total variability of a GMPE is demonstrated by Δ, which has a normal distribution with zero-mean and the standard deviation of σ. This variability is composed of between-events (inter-event) variability, ΔB, and within-event (intra-event) variability, ΔW, which have normal distribution with zero-means and standard deviations of τ and ϕ, respectively. Since the between and within event standard deviations are uncorrelated, the total standard deviation can be obtained form

$$\sigma = \sqrt{\tau^2 + \phi^2}. \tag{2}$$

It should be noted that the notations used in this paper, closely follow the notations used by Al Atik *et al* (2010), since these notations are defined to be used in the development of the next generation attenuation (NGA) models.

The value of the standard deviation of a GMPE has a significant effect on the probabilistic seismic hazard assessment (PSHA) results (Bommer and Abrahamson 2006). In order to decrease the influence of the standard deviation on PSHA results, the amount of the standard deviation should

be reduced. To this end, the GMPE associated with the lower standard deviation is preferred.

Diverse methods have been introduced and developed to compute the unknown parameters of GMPEs. The oldest and most rudimentary method to compute coefficients of a functional form are the least squares estimation, and the weighted least squares estimation. Brillinger and Preisler (1984, 1985) employed the random-effect regression method. Later, McLaughlin (1991) and Joyner and Boore (1993) utilized maximum likelihood estimation. Joyner and Boore (1993) showed that their proposed weighting matrix for the maximum likelihood estimation leads to the same result from the Brillinger and Preisler (1984, 1985) method. Two-stage methods with different weighting matrices were developed by Fukushima and Tanaka (1990), Masuda and Ohtake (1992), and Joyner and Boore (1993). Recently, methods based on genetic algorithm (GA) is getting widespread in which an optimization function is defined to determine the coefficients of GMPEs (Tavakoli and Pezeshk 2005, Sobhaninejad *et al* 2007, Tavakoli and Pezeshk 2007, Cabalar and Cevik 2009, Bagheri *et al* 2011, Yilmaz 2011).

The main goal of this study is to analyze and compare the above-mentioned methods to explore the differences between them as well as to understand advantages and disadvantages of each one. First, these methods are thoroughly described. Then, the considered attenuation relationship and the prepared database are reviewed. Afterward, the LH test introduced by Scherbaum *et al* (2004) in addition to various statistical techniques (Histogram of the standard normal distribution, Q–Q plot, and analysis of variance) and goodness-of-fit measures (t-test, Var-test, Chi-square test, Lillie-test, ks-test, Jarque–Bera test, Anderson–Darling test, and Shapiro–Wilk test) are applied. Finally, the strengths and limitations of each procedure will be compared and discussed.

Parameter estimation methods in linear models

In order to estimate unknown parameters in linear models, the following matrix form is used

$$Y = X\theta + \Delta, \tag{3}$$

where Y is a $n \times 1$ vector of observations (outputs), θ is a $m \times 1$ vector of unknowns, X is a $n \times m$ matrix of variables (inputs), and Δ is a $n \times 1$ vector of deviations. The parameter n represents the number of observations and m represents the number of unknown parameters. The elements of vector Δ have a normal distribution with the average of zero and the variance–covariance matrix which is defined as

$$\text{variance}(\Delta) = V = \sigma^2 W^{-1}, \tag{4}$$

where W is a known weighting matrix of observations, and σ^2 is an unknown positive coefficient which can be acquired from the expected value of the squared residuals. With respect to the type of variance matrix, V, different methods of the regression analysis can be employed to estimate unknown parameters.

Unweighted least squares estimation

The simplest case to solve equation (3), called ordinary least squares (OLS) estimation, occurs once the weighting matrix, W, is a diagonal one with identical elements. In this method, by using properties of matrices, we get

$$\widehat{\theta} = (X^T X)^{-1} X^T Y, \tag{5}$$

in which $\widehat{\theta}$ is the estimation of the unknown coefficients vector, θ. If the matrix X is a square matrix, it could be simply written as

$$\widehat{\theta} = X^{-1} Y. \tag{6}$$

It must be emphasized that this solution is applicable once the inverse of input matrix exists. If the determinant of the matrix, X or $X^T X$ approaches zero, an alternative algorithm called recursive least squares algorithm (Ljung 1999) can be used as follows

$$\begin{cases} \theta_{i+1} = \theta_i + S_{i+1} x_{i+1} (y_{i+1}^T - x_{i+1}^T \theta_i) \\ S_{i-1} = S_i - \dfrac{S_i x_{i+1} x_{i+1}^T S_i}{1 + x_{i+1}^T S_i X_{i+1}}, \quad i = 0, 1, 2, \ldots, n-1, \\ \theta_0 = 0 \qquad S_0 = \lambda I \end{cases} \tag{7}$$

in which I is an $m \times m$ identity matrix, λ is a large arbitrary value, x_i^T is the ith row of the matrix X and y_i^T is the ith element of the vector Y.

Once the unknown parameters are calculated, the estimation of the output vector is

$$\widehat{Y} = X\widehat{\theta}. \tag{8}$$

and; therefore, the difference between the real amounts and predicted values gives the vector of residuals

$$\Delta = \widehat{Y} - Y, \tag{9}$$

where $\widehat{Y}$ is the predicted values of observations, Y is the real amounts of observations and Δ is the vector of residuals.

One of the initial assumptions in the regression analysis is to have a constant variance (homoscedasticity) for the errors in the model to use OLS method. Thus, if this assumption is violated, the OLS method is not appropriate and the effect of non-constant variance (hetersoscedatisticy) should be considered by using a weighting matrix or transformation in the model. When the variance in a model is variable although the estimated coefficients are not biased, the total standard error is inaccurate and the variance of the estimators is not optimum. It is therefore necessary to account for the hetersoscedatisticy effects (e.g. biased standard errors) in the estimations of coefficients to derive best linear unbiased estimators (BLUEs). There are two strategies under the presence of hetersoscedatisticy indicating that the variance is non-constant. In the first strategy, efficient coefficients can be acquired by incorporating an appropriate weighting function in the regression analysis to take into account of hetersoscedatisticy. In the second strategy, the coefficients obtained by OLS are kept, but the variance would not be constant and variable variance should be used. The OLS method is considered as method 1 in this paper.

Weighted least squares (WLS) estimation

Once the weighting matrix is available but either it is off diagonal or the elements of the diagonal matrix are not identical because some observations are less reliable than the others, the OLS method does not result in obtaining BLUEs (Draper and Smith 1981). When the weighting matrix is diagonal, observations are independent but they have different variances, while when the weighting matrix has off diagonal elements, it means that observations are correlated to each other and they are dependent. In this case, the final solution is given by (Searle 1971)

$$\widehat{\theta} = (X^T W X)^{-1} X^T W Y. \tag{10}$$

For the WLS method, different kinds of transformation matrices are proposed. For instance, if the weighting matrix is diagonal, transformation matrix, P, can be estimated by

$$P = \sqrt{W}. \tag{11}$$

If the weighting matrix is off diagonal and it is Hermitian positive-definite, it can be decomposed by using the Cholesky method (Gentle 1998)

$$P^T P = W, \tag{12}$$

in which P is an upper triangular transformation matrix with positive diagonal elements and P^T is the transpose of P. Then, adjusted input and output matrices can be defined as

$$\overline{X} = PX, \tag{13}$$

and

$$\overline{Y} = PY, \tag{14}$$

in which $\overline{X}$ and $\overline{Y}$ are called the transformed input and output matrices. Therefore, using the weighting matrix acts in a way that the input and output matrices are transformed by the matrix P. Thus, the solution of the problem, can be rewritten as

$$\widehat{\theta} = (\overline{X}^T \overline{X})^{-1} \overline{X}^T \overline{Y} = (X^T P^T P X)^{-1} X^T P^T P Y. \tag{15}$$

It should be noted that the $P\Delta$ matrix now has a constant variance covariance matrix. As it is mentioned in the OLS method, once the determinant of matrix is around zero, the recursive least squares method is suggested. Therefore, the recursive algorithm in equation (7) can be applied to estimate coefficients.

The weighting matrix for the WLS method can be determined by using either grouped data to estimate the variance for each group, absolute residuals, or squared residuals versus a specific variable. The WLS method is called method 2 in this paper.

Unweighted one-stage maximum likelihood estimation

For the one-stage maximum likelihood estimation, it is assumed that observations have a normal distribution, and accordingly, the probability density function of the observations L (West *et al* 2007) is expressed as

$$L = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y - X\theta)^T V^{-1} (Y - X\theta)\right], \tag{16}$$

where $||$ is the determinant and n is the number of observations. For an assumed V, maximizing equation (16) with respect to the vector θ leads to the same results obtained from OLS method.

Weighted one-stage maximum likelihood estimation

When the variance–covariance matrix, V, is either off diagonal or diagonal with unequal elements, and this matrix is unknown, the maximum likelihood method can be used.

By substituting $\sigma^2 W^{-1}$ instead of V in equation (16) and then taking a natural logarithm, it can be expressed as

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\left|W^{-1}\right| - \frac{1}{2\sigma^2}\left[(Y - X\theta)^T W (Y - X\theta)\right]. \tag{17}$$

For a constant W, the vector θ can be estimated by using equation (10). As a result, $\ln(L)$ is a function of σ^2, and in order to maximize $\ln(L)$, its differentiation with respect to σ^2 has to be equal to zero.

To derive an overall unbiased estimation of σ^2, the following equation is proposed (Chatterjee and Hadi 2006)

$$\sigma^2 = \frac{1}{n - m}\left[(Y - X\theta)^T W (Y - X\theta)\right], \tag{18}$$

where m is the rank of matrix X and $n - m$ displays the degrees of freedom.

For a specific W, values of $\widehat{\theta}$, σ^2 and $\ln(L)$ values can be calculated. These values are calculated for different assumed W's, and then the final answer corresponds to the W that generates the maximum value of $\ln(L)$.

Joyner and Boore (1993) proposed a block diagonal variance covariance matrix

$$V = \sigma^2 W^{-1} = (\tau^2 + \phi^2)\begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & v_{N_e} \end{bmatrix}, \tag{19}$$

where N_e is the number of events in the database, and v_i is defined as

$$v_i = \begin{bmatrix} 1 & \gamma & \cdots & \gamma \\ \gamma & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \gamma \\ \gamma & \gamma & \cdots & 1 \end{bmatrix}, \tag{20}$$

where γ is defined as the following equation

$$\gamma = \frac{\tau^2}{\tau^2 + \phi^2}, \tag{21}$$

in which τ^2 and ϕ^2 are the variances related to the between-events and within-event variabilities, respectively. The rank of the matrix v_i is equal to R_i, the number of records for the event i (Joyner and Boore 1993). The weighted maximum likelihood method is considered as method 3 in this paper.

Two-stage methods

The very first step in this method is to separate variables that are identical for a specific event such as magnitude and fault mechanism from the other variables which are different for various records of a specific event such as distance and soil condition. The linear equation can be written as

$$Y = B_0 + \sum_{i=1}^{m} B_i f_i(x_i) + \sum_{j=1}^{n} A_j g_j(x_j), \tag{22}$$

where $f_i(x_i)$ and $g_j(x_j)$ consist of identical variables and unequal variables for the same event, respectively. B and A are coefficients in this equation, B_0 is a constant in the equation, n is the number of unequal variables, and m is the number of identical variables. Also, the total number of records for all earthquakes and stations is N, and the total number of different events is N_e.

Now, function H can be defined as

$$H = B_0 + \sum_{i=1}^{m} B_i f_i(x_i). \tag{23}$$

In the first stage, unknown parameters related to the unequal variables for the same event, A_j, are calculated by using the following equation

$$Y1 = X1\theta 1 + \Delta 1, \tag{24}$$

where

$$Y1 = [Y_k], \tag{25}$$

$$\theta 1 = [A_1, A_2, \ldots, A_n, H_1, H_2, \ldots, H_{N_e}]^T, \tag{26}$$

and

$$X1 = [g_1(x_{k1}), g_2(x_{k2}), \ldots, g_n(x_{kn}), E_{k1}, E_{k2}, \ldots, E_{kN_e}], \tag{27}$$

for $k = 1, 2, \ldots, N$ and $\Delta 1$ is a vector containing intra-event residuals, ΔW, with a normal distribution $N(0, \phi^2)$. E_{kl}'s are dummy variables and $E_{kl} = 1$ if the record k comes from the event l; otherwise, $E_{kl} = 0$.Then, equation (23) can be rearranged as

$$\widehat{H} = B_0 + \sum_{i=1}^{m} B_i f_i(x_i) + (\widehat{H} - H) + \Delta B, \tag{28}$$

where ΔB is the intrinsic variability for inter-event residuals in the aforementioned equation. At this point, unequal variables, A_j, are calculated. Now the identical variables, B_i, can be estimated by the following equation

$$Y2 = X2\theta 2 + \Delta 2, \tag{29}$$

where

$$Y2 = [\widehat{H}_1, \widehat{H}_2, \ldots, \widehat{H}_{N_e}]^T, \tag{30}$$

$$\theta 2 = [B_0, B_1, B_2, \ldots, B_m]^T, \tag{31}$$

and

$$X2 = [1, f_1(x_{l1}), f_2(x_{l2}), \ldots, f_m(x_{lm})], \tag{32}$$

for $l = 1, 2, \ldots, N_e$ and $\Delta 2$ is a vector containing errors containing $\widehat{H} - H$ and ΔB. Since the $\widehat{H} - H$ and ΔB vectors are uncorrelated, the variance–covariance matrix of $\Delta 2$ can be written as

$$V_2 = \mathrm{Var}(\widehat{H} - H) + \tau^2 I, \tag{33}$$

where τ^2 is the variance of the vector ΔB and I is the identity matrix. $\mathrm{Var}(\widehat{H} - H)$ is defined as

$$[\mathrm{Var}(\widehat{H} - H)]_{i,k} = [\mathrm{Var}(\widehat{\theta} 1)]_{i+n,k+n}, \tag{34}$$

and the right hand side of the above mentioned equation can be estimated by

$$\mathrm{Var}(\widehat{\theta} 1) = (X1^T X1)^{-1}\phi^2. \tag{35}$$

in which ϕ^2 is the variance of the first stage. Finally, the likelihood of samples can be obtained from the following equation

$$L2 = (2\pi)^{-\frac{N_e}{2}} |V_2|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}(Y2 - X2\theta 2)^T V_2^{-1}(Y2 - X2\theta 2)\right]. \tag{36}$$

Now the problem can be easily solved as a WLS problem if the weighting matrix ${V_2}^{-1}$ is known. Since, τ^2 is unknown, therefore ${V_2}^{-1}$ is unknown. Different assumptions are suggested by Fukushima and Tanaka (1990), Masuda and Ohtake (1992), and Joyner and Boore (1993), for this weighting matrix. In this paper the Fukushima and Tanaka (1990), and Joyner and Boore (1993) weighting matrices are used to calculate the unknown parameters of the ground motion equation.

Joyner and Boore (1993) proposed a diagonal weighting matrix as

$$w_i = \left(\frac{\phi^2}{R_i} + \tau^2\right)^{-1}, \tag{37}$$

where R_i is the number of records for the event i, and ϕ^2 is the variance of the first stage. τ^2 can be calculated by iterating starting from zero to obtain the maximum likelihood.

Fukushima and Tanaka (1990) suggested using a diagonal weighting matrix in which each earthquake's weight is equal to the number of records for that event. Two-stage maximum likelihood method with no weighting matrix, Boore and Joyner weighting matrix, and Fukushima and Tanaka weighting matrix are called method 4, method 5, and method 6 in this paper.

Genetic algorithm

A GA is used to solve optimization problems based on methods that originate from the natural and biological evolution such as selection, mutation, and cross over (Holland 1975, Goldberg *et al* 1989). In a GA, the optimization problem is formulated by defining an objective function. Then, a set of initial solutions called an initial population to the objective function is generated. Each solution (called a chromosome) is a combination of all unknown parameters that need to be estimated to satisfy the optimization problem (Gen and Cheng 2000).

The next step is to find the values of the objective function for the population. The initial population can now be improved by using genetic operators including selection, cross over, and mutation. The fitter chromosomes from the initial population are selected to create the next population. Each step of generating a new population is called a generation. To generate the next population, the selected chromosomes can be mixed two by two (parent) to create a new chromosome (child) which has the properties of its parent. Another way to create a new chromosome for the next population is using chromosomes from the previous generation (Gen and Cheng 2000). In this situation, the parameters inside of the chromosome are modified in order to keep the genetic diversity and avoid being stuck in local minima.

This process repeats until either the maximum number of generations is reached or the value of the objective function is satisfactory, and then this iterative algorithm is terminated. Finally, the answer to the optimization problem is the solution from the last generation which matches better with the objective function. The GA method is considered as method 7 in this paper.

Database

The database for this paper is built based on the CD ROM of European and Middle Eastern strong motion (Ambraseys *et al* 2000). The database includes 462 triaxial strong motion records from 110 earthquakes within 261 stations in Europe and the Middle East for corrected acceleration, velocity, and displacement (Ambraseys *et al* 2004).

All published empirical GMPEs from 1964 are provided by Douglas (2003) and Douglas (2014a). Most of these GMPEs until 2000s were developed for PGA and 5% damped linear elastic pseudo-absolute response spectral acceleration (PSA) ordinates at different periods (Douglas 2014b); however, the estimation of other GMIMs such as PGV, PGD, Arias intensity, and duration in addition to PGA and PSA are nowadays getting popular (Douglas 2012). Of course, particular GMIMs such as PGA and PSA at 0.2 and 1 s are very important due to the usage of them in performing PSHA (Tavakoli and Pezeshk 2005, Rezaeian *et al* 2015). In this study, we use values of PGA and 5% damped linear elastic PSA at 0.2 and 1 s as representative GMIMs to assess the performance of described regression methods. In this study, we use the geometric mean of two horizontal components because it is the most widely used GMIM in GMPEs (Beyer and Bommer 2006, Douglas 2014a). Various schemes have been defined on how to treat two horizontal components in order to be employed in the GMPEs such as arithmetic mean, geometric mean, larger component, random component for each accelerogram. Boore *et al* (2006) proposed two new definitions for the GMIM that can be used in GMPEs called GMRotD50 (GM indicates geometric mean, Rot implies rotations, D stands for period-dependent rotations, and 50 means that the 50 percentile value is used for the measure) or GMRotI50 (GM indicates geometric mean, Rot implies rotations, I stands for period-independent rotations, and 50 means that the 50 percentile value is used for the measure). The advantage of these measures is sensor orientation independency that yields in removing the uncertainty caused by the sensor orientation in GMPEs. Later, Boore (2010) proposed another definition to be considered as a GMIM called RotD50 in which no geometric mean is used to derive the new measure. NGA-west attenuation relationships (Bozorgnia *et al* 2014) are based on RotD50 definition for the GMIMs. Beyer and Bommer (2006) demonstrated that decreasing the aleatory uncertainties of GMPEs due to adopting GMRotD50 or GMRotI50 as the reference GMIM is less than 2%. Here, we employ the same database for all regression methods; and thus, the usage of geometric mean will equivalently affect results obtained from all regression methods.

In this database, the moment magnitude, **M**, is considered for the magnitude scale. Records that do not have moment magnitudes have been removed. In addition, records that have moment magnitudes less than 5.0 have been omitted. Joyner and Boore distance, R_{JB}, is used for the distance scale. Records with Joyner and Boore distances less than or equal 100 km are considered in this database.

Records included in the database all have known fault mechanisms. The Frohlich and Apperson (1992) approach is applied to classify fault mechanisms. Therefore, four classifications, thrust, normal, strike-slip, and odd, are defined based on the *B*, *P*, and *T* axes plunges for each record. Once the plunge of *P* axis is more than 60°, the earthquake is considered as normal. When the plunge of *B* axis is greater than 60°, the earthquake is classified as strike slip. Once the plunge of *T* axis is more than 50°, the event is considered as thrust. Otherwise, the earthquake is considered as odd.

Records with unknown site conditions have been removed from the database. Boore *et al* (1993) method has been used to categorize site conditions based on the average shear-wave velocity in the upper 30 m of the site profile. Accordingly, site conditions are divided into 4 groups: very soft, soft, stiff, and rock. According to Boore *et al* (1993), once $V_{S30} < 180$ m s^{-1}, the soil is considered as very soft; once $180 \leqslant V_{S30} < 360$ m s^{-1}, the soil is considered as soft, once $360 \leqslant V_{S30} < 750$ m s^{-1}, the soil is classified as stiff; otherwise, the soil is categorized as rock'.

A total of 350 records created by 85 earthquakes from the data of the CD ROM have been collected as the final database. This database has 27 singly recorded earthquakes out of 85. Table 1 and figure 1 present the distribution of the data with respect to the site condition and fault mechanism. Since there are only 7 records with the very soft soil site condition, the soft and very soft soil classifications are combined and considered as one group. Figure 2 displays the distribution of the database with respect to the magnitude and distance.

Functional form

To evaluate the aforementioned methods for the estimation of empirical attenuation relationships coefficients, Ambraseys *et al* (2005) equation is considered. The proposed functional form is

Table 1. The distribution of the database with respect to the site condition and the fault mechanism.

	Very soft	Soft	Stiff	Rock	Total	Percent (%)
Normal	5	23	58	52	138	40
Strike-slip	1	15	24	49	89	25
Thrust	1	11	33	14	59	17
Odd	0	18	25	21	64	18
Total	7	67	140	136	350	
Percent (%)	2	19	40	39		

$$\log(y) = a_1 + a_2\,\mathrm{M} + (a_3 + a_4\mathrm{M})\log\left(\sqrt{d^2 + a_5^2}\right) + a_6 S_\mathrm{S} + a_7 S_\mathrm{A} + a_8 F_\mathrm{N} + a_9 F_\mathrm{T} + a_{10} F_\mathrm{O}, \tag{38}$$

where a_1 through a_{10} are the unknown parameters in the equation. $S_\mathrm{S} = 1$ or $S_\mathrm{A} = 1$ if the soil conditions are soft or stiff, respectively; otherwise, they are equal to zero. $F_\mathrm{N} = 1$, $F_\mathrm{T} = 1$, or $F_\mathrm{O} = 1$ if the fault mechanisms are normal, thrust, or odd, respectively; otherwise, they are equal to zero. y is the GMIM of interest such as PGA and SA at different periods, and d is the Joyner and Boore distance.

Application of different methods

Equation (38) is nonlinear in coefficients due to the multiplication of a_3 and a_4 with a_5. The first step to apply regression methods is to linearize the equation using Taylor series. Therefore, matrices Y, θ, and X can be defined as follows

$$Y = [\log y_1, \log y_2, \ldots, \log y_\mathrm{N}]^T, \tag{39}$$

where N is the number of records in the database which is 350

$$\theta = [a_1, a_2, a_3, a_4, \Delta a_5, a_6, a_7, a_8, a_9, a_{10}]^T, \tag{40}$$

and

$$X = \begin{bmatrix} 1 & M_i & \log\left(\sqrt{d_i^2 + a_5'^2}\right) & M_i \log\left(\sqrt{d_i^2 + a_5'^2}\right) & & & & & \\ & & \dfrac{\partial}{\partial a_5}\left[(a_3' + a_4' M_i)\log\left(\sqrt{d_i^2 + a_5^2}\right)\right]_{a_5 = a_5'} & & S_{\mathrm{S}i} & S_{\mathrm{A}i} & F_{\mathrm{N}i} & F_{\mathrm{T}i} & F_{\mathrm{O}i} \end{bmatrix}$$

$$= \left[1 \quad M_i \quad \log\left(\sqrt{d_i^2 + a_5'^2}\right) \quad M_i \log\left(\sqrt{d_i^2 + a_5'^2}\right) \quad \frac{a_3' a_5' + a_4' a_5' M_i}{d_i^2 + a_5'^2} \quad S_{\mathrm{S}i} \quad S_{\mathrm{A}i} \quad F_{\mathrm{N}i} \quad F_{\mathrm{T}i} \quad F_{\mathrm{O}i}\right], \tag{41}$$

in which i is the record number, and a_3', a_4', and a_5' are trial amounts for a_3, a_4, and a_5. $\Delta a_5'$ is a perturbation of the starting point of a_5 coefficient that aim in improving the solution. Calculation of the coefficients is therefore an iterative process in which the parameter a_5' is updated by $a_5' + \Delta a_5'$. Any positive number except zero can be used as an starting point for a_5' (Joyner and Boore 1993). This iteration is terminated once the amount of $\Delta a_5'/a_5'$ is less than a desired limit which is considered 10^{-6} in this paper because the coefficients are estimated with 5 decimal places. The initial assumption for a_3, a_4, and a_5 is 1.

Since there are many similar records, the determinant of the X^TX matrix approaches to zero for OLS method (method 1). Accordingly, the recursive algorithm defined by equation (7) is used to calculate coefficients of the equation. The result of this method is mentioned in tables 2–4 for PGA, PSA at 0.2, and PSA at 1 s, respectively. The coefficients obtained by Ambraseys *et al* (2005) are also reported for comparison. It is worth mentioning that the Ambraseys *et al* (2005) coefficients are derived using the larger horizontal component.

Douglas and Smit (2001) suggested that if their database is divided into 2 km by 0.2M unit intervals, records gathered in each bin can be assumed as repeated records. Douglas (2004a, 2004b) showed that using the analysis of variance could be a useful tool to investigate the regional dependency of strong motion data. In order to perform analysis of variance, intervals of 5 km by 0.25M_s units were used by Douglas (2004a). Since the number of records in this study is limited, using those intervals leads to loss of many records in this investigation. Therefore, after exploring different unit intervals for the distance and magnitude, the database is divided into 10 km by 0.2M unit intervals to have sufficient records in each bin. The equations from figures 3 and 4 display the dependency of the standard deviations estimated from unit intervals on magnitude and distance as well as distribution of them, respectively.

Gradients of the fitted lines depict that there is a decrease in the standard deviation, σ, with increasing the magnitude and distance as Youngs *et al* (1995), Campbell (1997), and Ambraseys *et al* (2005) reported. Although the slope for the distance dependency is near to zero, the dependency of the error on the magnitude is considerable. In this regard, the dependency of the standard deviation on the magnitude cannot be rejected at the significance level of 5%. In fact, it can be said that earthquakes with higher magnitude are more informative and therefore have more weight. Hence, there are two approaches to treat hetersoscedatisticy. In the first approach, the coefficient derived from OLS method are used but the variance should be updated based on either the prior knowledge or performing pure error analysis. In the second approach, a WLS regression analysis (method 2) can be employed in which the weighting matrix can be derived

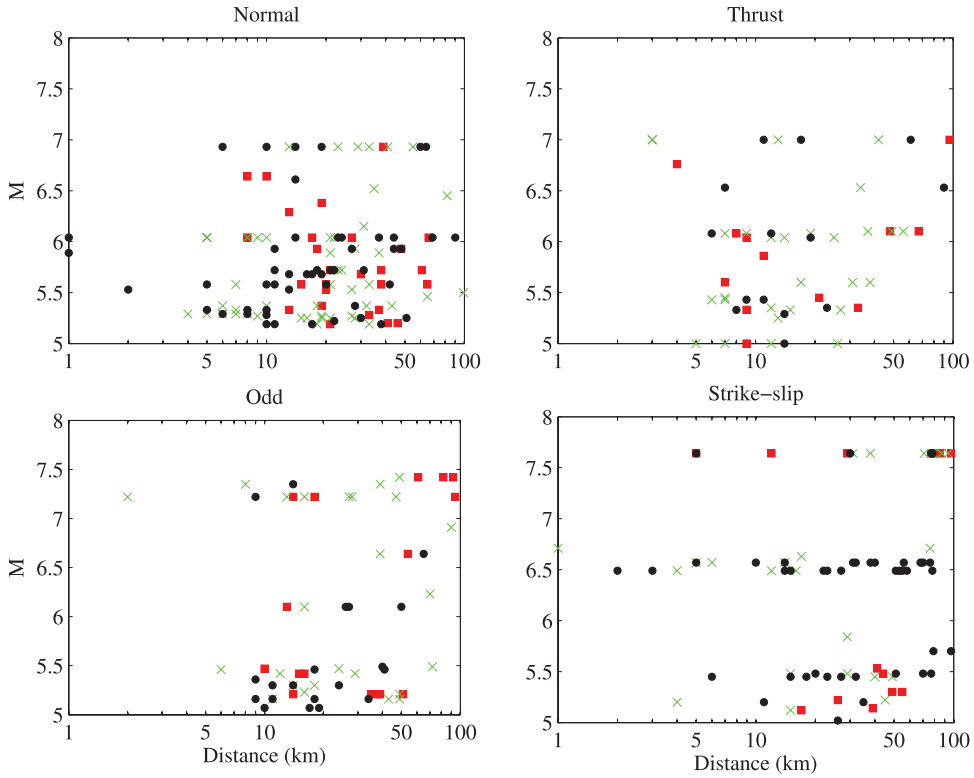

Figure 1. The distribution of the database used in this study in the magnitude-distance database for different fault mechanisms. Rectangular signs represent soft soil, cross signs are for stiff soil, and circle signs represent rock.

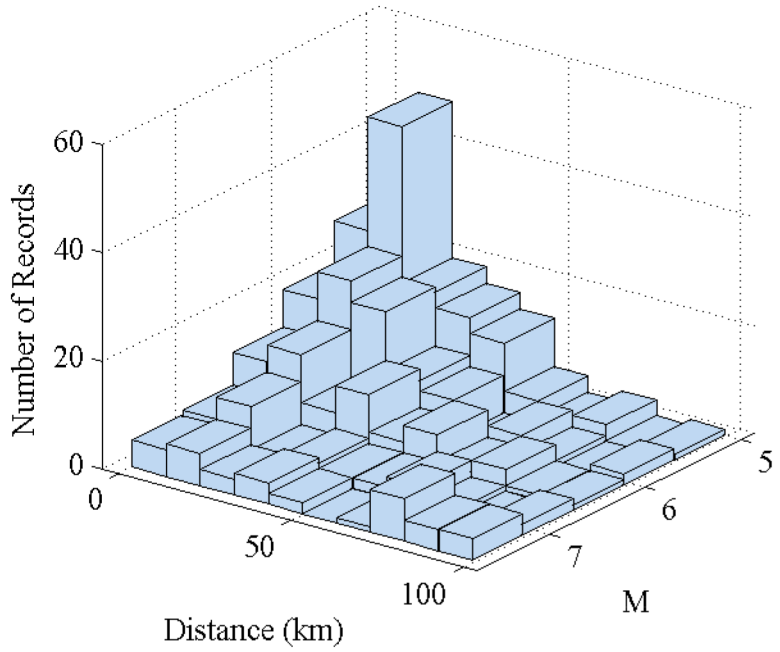

Figure 2. The distribution of the database with respect to the magnitude and distance.

from the dependency of the error on magnitude. Therefore, the weighting matrix which is determined by using grouped data is defined as

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & w_N \end{bmatrix}, \tag{42}$$

where

$$w_i = \frac{1}{\sigma_i^2} = \frac{1}{(0.59163 - 0.06026M_i)^2}, \tag{43}$$

in which σ_i^2's are variances of the bins. To estimate the coefficients of the equation the recursive algorithm is used and these coefficients are tabulated in tables 2–4 for PGA, PSA at

Table 2. Coefficients of the GMPEs obtained for PGA.

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	ADSS05
a_1	1.305 48	2.001 06	1.423 97	1.198 29	1.202 43	1.225 84	1.028 69	2.522
a_2	0.013 57	−0.105 12	0.007 61	0.034 63	0.030 66	0.018 64	0.055 09	−0.142
a_3	−2.406 43	−2.859 04	−2.524 33	−2.377 50	−2.377 50	−2.377 50	−2.272 62	−3.184
a_4	0.207 08	0.281 36	0.220 46	0.206 84	0.206 84	0.206 84	0.186 87	0.314
a_5	7.640 47	7.279 80	7.435 78	5.792 73	5.792 73	5.792 73	7.459 75	7.600
a_6	0.121 60	0.154 00	0.102 68	0.055 83	0.055 83	0.055 83	0.117 70	0.137
a_7	0.008 34	0.019 79	0.014 68	0.007 44	0.007 44	0.007 44	0.002 24	0.050
a_8	0.004 46	0.003 46	−0.050 73	−0.123 84	−0.080 97	0.001 09	0.019 05	−0.084
a_9	0.084 11	0.089 77	0.064 43	0.075 11	0.065 69	0.078 29	0.105 45	0.062
a_{10}	−0.050 97	−0.022 27	−0.056 34	−0.066 39	−0.056 00	−0.040 17	−0.034 17	−0.044
τ	—	—	0.274 10	0.275 84	0.275 84	0.275 84	—	
φ	—	—	0.115 14	0.130 37	0.120 00	0.113 73	—	
σ	0.295 90	0.297 13	0.297 60	0.305 10	0.300 81	0.298 37	0.296 11	

Note: σ is the overall unbiased standard deviation for each method. Also, note that the ADSS05 coefficients are derived using the larger horizontal component instead of the geometric mean of two horizontal components.
Method 1: OLS; method 2: WLS; method 3: one-stage maximum likelihood; method 4: two-stage maximum likelihood with no weighting matrix; method 5: two-stage maximum likelihood with Boore and Joyner weighting matrix; method 6: two-stage maximum likelihood with Fukushima and Tanaka weighting matrix; method 7: GA; ADSS05: Ambraseys *et al* (2005).

Table 3. Coefficients of the GMPEs obtained for PSA at 0.2 s.

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	ADSS05
a_1	1.009 60	1.295 80	1.027 10	0.641 80	0.607 30	0.610 30	1.422 60	2.632
a_2	0.095 80	0.049 60	0.103 70	0.156 80	0.154 60	0.146 70	0.036 30	−0.109
a_3	−2.037 50	−2.232 70	−2.067 50	−1.773 20	−1.773 20	−1.773 20	−2.342 90	−2.990
a_4	0.161 90	0.192 20	0.162 60	0.127 80	0.127 80	0.127 80	0.206 40	0.289
a_5	7.481 50	7.626 20	7.432 30	5.240 70	5.240 70	5.240 70	7.994 80	8.100
a_6	0.093 80	0.097 00	0.081 50	0.035 10	0.035 10	0.035 10	0.095 90	0.124
a_7	0.029 60	0.033 90	0.031 00	0.020 60	0.020 60	0.020 60	0.027 80	0.070
a_8	0.066 80	0.070 80	0.019 10	−0.113 60	−0.010 40	0.067 10	0.069 40	−0.033
a_9	0.114 10	0.118 00	0.093 40	0.069 90	0.089 80	0.115 70	0.115 30	0.090
a_{10}	−0.056 40	−0.029 60	−0.068 80	−0.108 70	−0.069 10	−0.044 10	−0.055 00	−0.039
τ	—	—	0.305 80	0.308 40	0.308 40	0.308 40	—	
φ	—	—	0.090 20	0.115 80	0.091 40	0.084 30	—	
σ	0.318 00	0.318 20	0.318 80	0.329 40	0.321 60	0.319 70	0.318 20	

Note: σ is the overall unbiased standard deviation for each method. Also, note that the ADSS05 coefficients are derived using the larger horizontal component instead of the geometric mean of two horizontal components.
ADSS05: Ambraseys *et al* (2005).

0.2, and PSA at 1 s, respectively. It should be noted that if the sample size (number of records) is small, the determined coefficients might not be accurate because the weight is estimated from the grouped data.

The weighted one-stage maximum likelihood method (method 3) is an iterative process. In this method, the natural logarithm of likelihoods (equation (17)) of different values of γ are calculated to find the one that is associated with the maximum likelihood of the pdf of observations. The value of the γ can vary from 0 to 1. After exploring this interval, it is found out that the γ related to the maximum likelihood is located between 0.1 and 0.2. Figure 5 shows that the maximum likelihood occurs when $\gamma = 0.15$. The coefficients corresponding to this method with $\gamma = 0.15$ are listed in tables 2–4 for PGA, PSA at 0.2, and PSA at 1 s, respectively.

In the first stage of the two-stage method, coefficients according to the terms that are not constant for records of a specific earthquake are determined. Then, the remaining coefficients, a_1, a_2, a_8, a_9, and a_{10} are calculated in the second stage. Equation (23) can be defined as

$$H = a_1 + a_2 M + a_8 F_N + a_9 F_T + a_{10} F_O, \tag{44}$$

and the unknown coefficients vector for the stage 1 is

$$\theta 1 = \left[a_3, a_4, \Delta a_5, a_6, a_7, H_1, H_2, \ldots, H_{N_e}\right]^T. \tag{45}$$

After finding vector $\theta 1$, vector $Y2$ and matrix $X2$, then vector $\theta 2$, can be determined using the equation (29). These coefficients with considering different weighting matrices, no weighting matrix (method 4), Boore and Joyner (1993) weighting matrix (method 5), and Fukushima and Tanaka (1990) weighting matrix (method 6) are provided in tables 2–4 for PGA, PSA at 0.2, and PSA at 1 s, respectively. It should be noted that for these three methods, coefficients a_3, a_4, a_5, a_6, and a_7 are identical. The final

Table 4. Coefficients of the GMPEs obtained for PSA at 1.0 s.

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	ADSS05
a_1	−2.129 90	−1.737 90	−1.477 70	−0.898 10	−0.801 10	−0.715 90	0.994 50	−1.359
a_2	0.521 90	0.444 70	0.420 40	0.321 30	0.307 30	0.291 40	0.010 40	0.403
a_3	−1.394 60	−1.607 20	−1.960 30	−2.427 60	−2.427 60	−2.427 60	−2.962 80	−1.848
a_4	0.038 20	0.079 50	0.126 10	0.201 10	0.201 10	0.201 10	0.279 50	0.124
a_5	7.328 00	6.665 70	6.413 50	5.222 00	5.222 00	5.222 00	7.896 40	6.000
a_6	0.378 30	0.386 70	0.392 90	0.419 00	0.419 00	0.419 00	0.281 30	0.357
a_7	0.158 10	0.157 50	0.184 70	0.186 20	0.186 20	0.186 20	0.103 80	0.211
a_8	0.004 30	0.017 90	−0.047 00	−0.077 70	−0.042 00	−0.017 60	0.117 70	−0.013
a_9	−0.045 20	−0.023 80	−0.042 70	0.041 10	−0.073 50	−0.082 10	0.348 50	0.024
a_{10}	−0.053 70	−0.047 60	−0.036 20	0.037 40	−0.050 70	−0.058 60	0.331 20	−0.101
τ	—	—	0.370 40	0.365 40	0.365 40	0.365 40	—	
φ	—	—	0.167 70	0.199 20	0.183 60	0.182 50	—	
σ	0.403 80	0.404 30	0.406 60	0.416 10	0.408 90	0.408 40	0.435 60	

Note: σ is the overall unbiased standard deviation for each method. Also, note that the ADSS05 coefficients are derived using the larger horizontal component instead of the geometric mean of two horizontal components.
ADSS05: Ambraseys *et al* (2005).

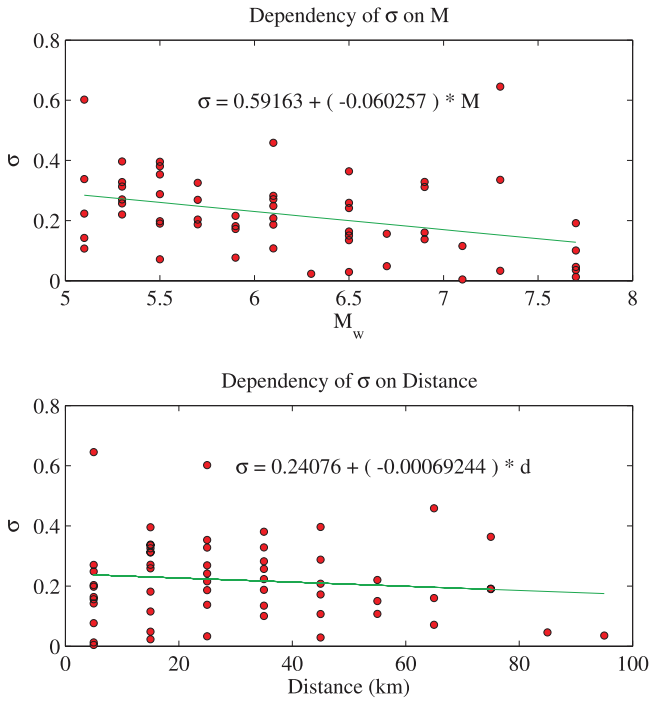

Figure 3. The dependency of the standard deviation on the magnitude and distance. Dots demonstrate the standard deviation obtained from bins and solid lines represent the fitted lines.

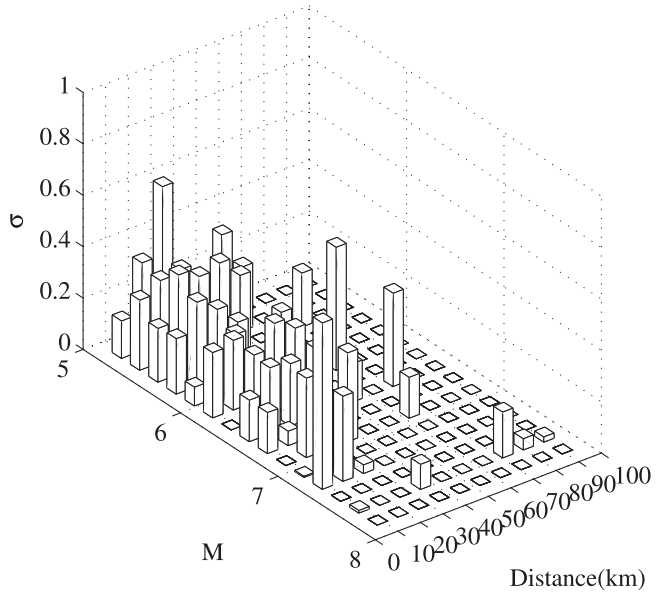

Figure 4. The distribution of the standard deviations for unit intervals. Distance intervals are 10 km and magnitude intervals have a length of 0.2.

coefficients for the Boore and Joyner method are determined when $\tau = 0.12$.

In order to apply the GA (method 7), the objective function which is an unbiased estimator of the variance is defined as

$$\text{Objective Function} = \frac{1}{N-10}\sum_{i=1}^{N}[\log(y_i) - \log(\hat{y}_i)]^2 , \quad (46)$$

where N is the number of records in the database, $N-10$ is the number of the degrees of freedom (10 is the number of unknown parameters), y_i is the observed value, and $\hat{y}_i$ is the estimated value of the ground motion parameter for the record i.

For the selection of chromosomes, the roulette wheel definition is used. The cross over probability p_c and mutation probability p_m are defined 0.8 and 0.05, respectively. The probability of the mutation is a small number to avoid losing suitable chromosomes. The number of generation and the size of population are 1000 and 200, respectively (Tavakoli and Pezeshk 2005, Sobhaninejad *et al* 2007). It should be pointed out that repetition of the GA estimation does not generate identical coefficient. In this regard, we run this algorithm twenty times and the result for each coefficient is then computed by taking the mean of estimated values for that coefficient. The estimated coefficients for the geometric mean of PGA and PSA at 0.2 and 1 s values of horizontal components are provided in tables 2–4, respectively.

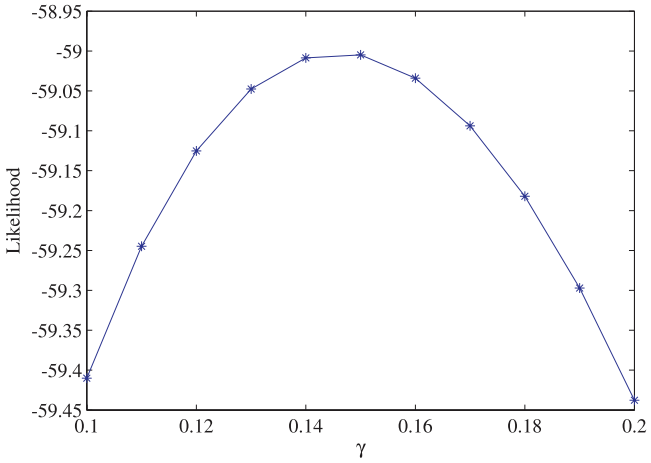

Figure 5. γ versus the likelihood for the one-stage maximum likelihood method. The peak is located at $\gamma = 0.15$.

Comparison of different methods

As it is seen from tables 2–4, the estimated coefficients from different methods are not the same or even in similar ranges. The overall unbiased standard deviation for each method is also given in tables 2–4. The minimum standard deviation is for the method 1 (OLS) and the maximum standard deviation is derived for the method 4. Statistical comparison can be an appropriate way to compare regression methods with each other to see which method gives the best answer. In order to assess the accuracy of different methods the residual analysis is utilized to check whether the basic assumptions of errors are valid. Therefore, various statistical tests are chosen and applied to compare these methods. These tests can be applied on residuals defined as the difference between the observed and predicted values. The residual, R, in models is obtained by

$$R = \log(y_{\text{observed}}) - \log(y_{\text{predicted}}). \tag{47}$$

These residuals tend to have a correlation with the variables (West *et al* 2007), since they are normalized by dividing them to the corresponding standard deviation. Therefore, normalized residuals, Z, are defined as

$$Z = \frac{P(\log(y_{\text{observed}}) - \log(y_{\text{predicted}}))}{\sigma}. \tag{48}$$

in which P is the transformation matrix and σ is the overall unbiased standard deviation given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\log(y_{\text{observed}}) - \log(y_{\text{predicted}}))^2}{N-10}}, \tag{49}$$

where N is the number of data and $N-10$ is the number of the degrees of freedom. 10 is the number of unknown parameters. Normalized residuals should follow the standard normal distribution in which the mean is zero and the standard deviation is one. Following tests are used to compare the described regression methods.

Graphical

The first graphical method is the Histogram which graphically summarizes characteristics of the distribution of a dataset (Montgomery and Runger 2003, Mendenhall and Sincich 2007). Histograms or bars in figure 6 illustrate the distribution of the residuals obtained by using the geometric mean of PGA values of horizontal components, so the spread of data and skewness can be graphically observed. The red lines display the probability distribution function (pdf) of the standard normal distribution. The number of bins equal to the square root of the number of elements in data. As it can be seen the distributions of the residuals are not perfect, but they approximately follow the standard normal distribution.

A Quantiles–Quantiles (Q–Q) plot shows the quantiles of a normal distribution versus the quantiles of the sample which is standard residuals here in order to compare the sample distribution with the standard normal distribution (Montgomery and Runger 2003, Mendenhall and Sincich 2007). Quantiles are the inverse of the cumulative distribution function (cdf) of a population at defined intervals. The '+' signs will be linear, if the sample is drawn from a normal distribution. Figure 7 indicates that the quantiles of the normalized residuals obtained by using the geometric mean of PGA values of horizontal components from methods 2 and 3 are matched on the quantiles of the standard normal distribution. For the remaining methods, there is a slight difference especially at the upper bound of the normalized residuals.

Goodness-of-fit tests

The basic assumption for the normalized residuals is to follow the standard normal distribution. Hence several statistical tests have been proposed to understand the goodness of a fit for a model (Montgomery and Runger 2003, Mendenhall and Sincich 2007). We employ some of these statistical measurements to investigate the quality of fits by different regression methods for the residuals obtained by using the geometric mean of PGA values of horizontal components.

The first test is the t-test in which the normalized residuals are assessed to see if the mean is zero (null hypothesis). If the null hypothesis cannot be rejected, the test statistic has a Student's t distribution. P-values of the t-test are tabulated in table 5. When a p-value which describes the probability of the null hypothesis is equal to or less than the considered significance level which is 0.05 in this paper, then the null hypothesis would be rejected at this significance level. Results state that all methods fail to reject the null hypothesis.

The second test is the Var-test which is a Chi-square measure and it evaluates whether the normalized residuals have a unit variance (null hypothesis). The p-values corresponding to this test from table 5 show that null hypothesis cannot be rejected in all described methods.

The next test is the Chi-square goodness-of-fit test that recognizes if the normalized residuals come out of the standard

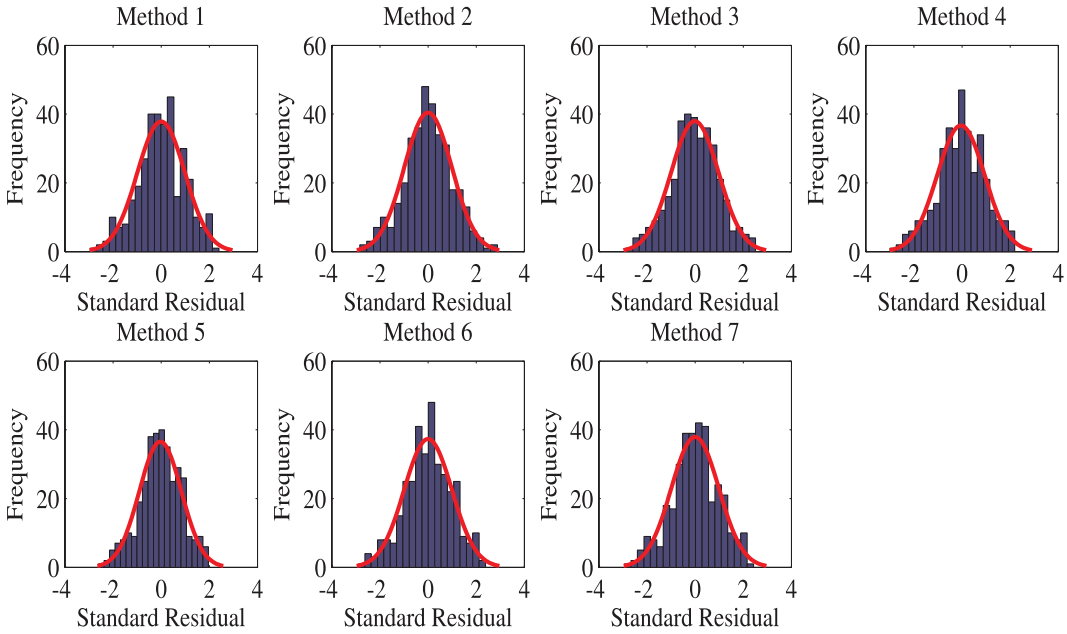

Figure 6. Histograms of the normalized residuals for different regression methods. Solid line shows the normal distribution with a mean of 0 and standard deviation of 1,$N(0, 1)$ and bars represent the distribution of the standard residuals.

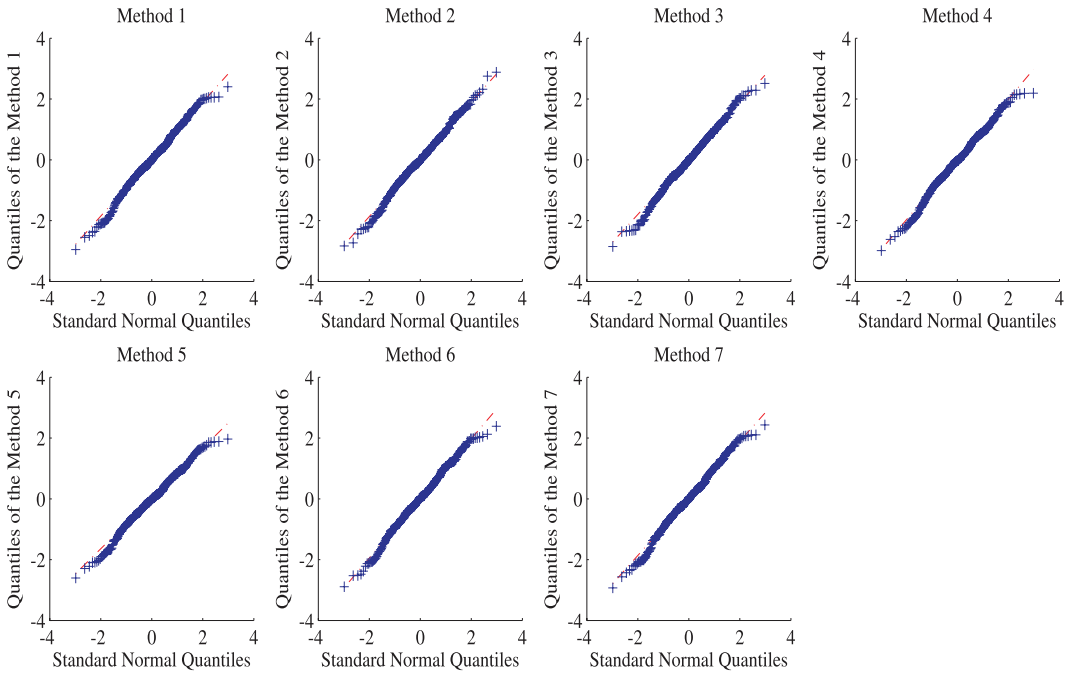

Figure 7. *Q*–*Q* plots of the normalized residuals for explained regression methods. Solid line shows the normal distribution with a mean of 0 and standard deviation of 1, and '+' signs represent the quantiles of each regression method.

normal distribution. Calculated *p*-values from table 5 display the null hypothesis cannot be rejected in all methods.

The next test is the Lilliefors test (Lillie-test) in which residuals are compared to observe if they follow a normal distribution. In this test the mean and variance are unknown. Lillie-test *p*-values are gathered in table 5. As it can be seen all methods fail to reject the null hypothesis at the desired significance level.

The fifth test is the ks-test (Kolmogorov–Smirnov test) that determines whether the cdf of the normalized residuals is matched with the cdf of the standard normal residual (null hypothesis). The *p*-values are listed in table 5. According to the estimated values null hypothesis cannot be rejected in all methods.

The next test is the Jarque–Bera test which evaluates if the skewness and kurtosis of the normalized residuals correspond to the standard normal distribution. The *p*-values are provided in table 5. These amounts show that all models fail to reject the null hypothesis.

The seventh test is the Anderson–Darling test which checks if the pdf of the normalized residuals matches the pdf of the

Table 5. The *p*-values of different goodness-of-fit tests.

Method	*t*-test	Var-test	Chi-square	Lillie-test	ks-test	Jarque–Bera test	Anderson–Darling test	Shapiro–Wilk test
1	0.999 87	0.750 23	0.139 11	0.268 45	0.628 33	0.352 49	0.193 89	0.104 16
2	0.995 85	0.750 23	0.193 97	0.392 33	0.697 54	0.500 00	0.293 71	0.532 73
3	0.812 09	0.748 62	0.297 85	0.215 13	0.682 80	0.500 00	0.543 39	0.375 34
4	0.395 66	0.729 77	0.107 88	0.340 01	0.743 24	0.275 47	0.152 02	0.088 41
5	0.479 15	0.680 60	0.079 91	0.487 23	0.173 64	0.281 09	0.191 14	0.089 52
6	0.999 90	0.750 23	0.258 12	0.389 29	0.719 91	0.322 72	0.312 59	0.139 37
7	0.998 26	0.750 23	0.140 83	0.184 10	0.553 32	0.387 26	0.178 59	0.122 52

Note: Method 1: OLS; method 2: WLS; method 3: one-stage maximum likelihood; method 4: two-stage maximum likelihood with no weighting matrix; method 5: two-stage maximum likelihood with Boore and Joyner weighting matrix; method 6: two-stage maximum likelihood with Fukushima and Tanaka weighting matrix; method 7: GA.

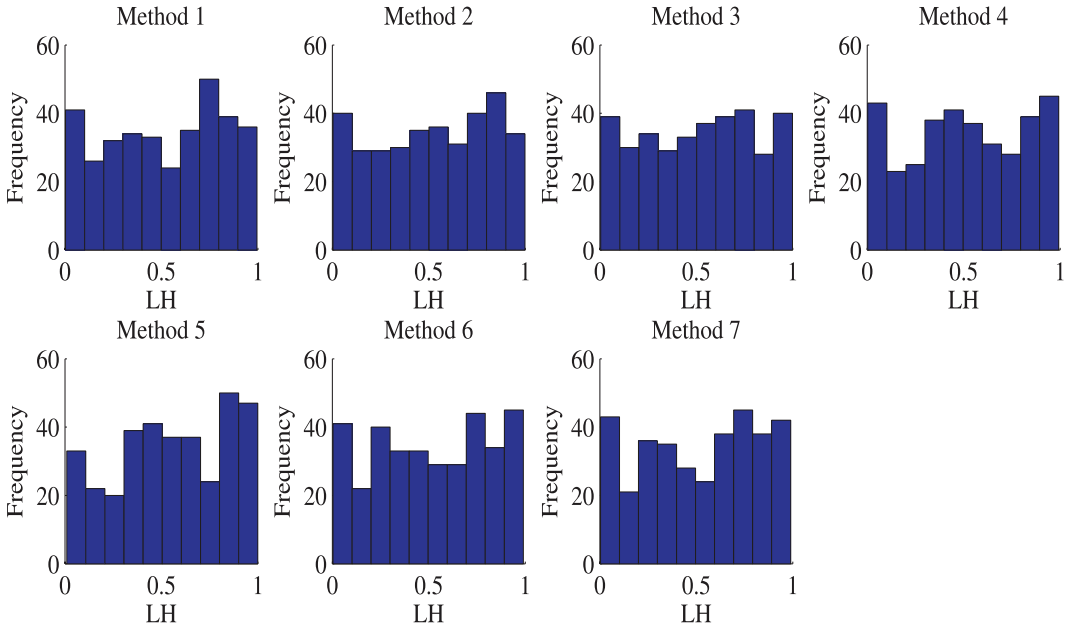

Figure 8. The distributions of the LH values for described regression methods. Each histogram has 10 bins with length of 0.1.

standard normal distribution. The *p*-values provided in table 5 indicate that the null hypothesis cannot be rejected in all methods.

The last one, Shapiro–Wilk test is a test to determine whether the residuals are drawn from a normal distribution. The *p*-values corresponding to this test are given in table 5 and they show that all methods fail to reject the null hypothesis.

A Larger *p*-value implies that the model is more confident (Scherbaum *et al* 2004). Regarding to these results, method 2 has the largest *p*-value for the Var-test, Jarque–Bera test, and Shapiro–Wilk test. Method 3 has the largest *p*-value for the Chi-square test, Jarque–Bera test, and Anderson–Darling test. Also, it can be inferred that method 4 and method 5 have the weakest performance compared to the remaining regression methods since they have 4 and 3 smallest *p*-values respectively.

LH test

Most of the goodness-of-fit measures can only check one of the assumptions related to the standard normal distribution. Scherbaum *et al* (2004) have introduced a likelihood-based method (LH test) in which all the assumptions would be assessed. In the LH test the likelihood of an observation that is a normalized residual for being equal to or greater than Z_0 with considering both tails of the distribution is given by

$$\mathrm{LH}(|Z_0|) = \frac{2}{\sqrt{2\pi}} \int_{Z_0}^{\infty} \exp\left(\frac{-Z^2}{2}\right) \mathrm{d}Z, \qquad (50)$$

where Z_0 is a normalized residual. If the normalized residuals are completely matched with the standard normal distribution, the amounts of the LH test would be uniformly distributed between 0 and 1 and the median of the LH values should be 0.5. The distributions of the LH values for the residuals obtained by using the geometric mean of PGA values of horizontal components are demonstrated in figure 8. The values of the LH are uniformly distributed for methods 2 and 3. The median of LH (MEDLH) is given in table 6. In addition, the mean, median, and standard deviation of the normalized residuals are tabulated in table 6 (MEANNR, MEDNR, and STDNR, respectively). Moreover, standard deviations corresponding to the median, mean, and standard deviation and estimated using the 'delete-1' jackknife resampling (Shao and Tu 1995) procedure (Scherbaum *et al* 2004) are provided in table 6.

Table 6. The results of the LH test.

Method	MEDLH	σ	MEANNR	σ	MEDNR	σ	STDNR	σ
1	0.521 84	0.000 36	0.000 01	0.002 83	0.029 31	0.002 17	0.987 02	0.001 98
2	0.519 91	0.000 37	0.000 27	0.002 83	−0.011 88	0.001 28	0.987 02	0.002 07
3	0.538 96	0.000 39	−0.012 55	0.002 83	−0.017 41	0.000 16	0.986 94	0.001 98
4	0.508 33	0.000 39	−0.044 82	0.002 83	−0.027 29	0.000 51	0.986 00	0.001 98
5	0.539 31	0.000 38	−0.033 03	0.002 50	−0.012 13	0.003 61	0.982 16	0.001 75
6	0.509 26	0.000 36	−0.000 01	0.002 83	0.036 89	0.001 00	0.987 02	0.001 95
7	0.525 08	0.000 36	−0.000 12	0.002 83	0.020 89	0.000 47	0.987 02	0.001 98

Note: MEDLH: median of LH; MEANNR: mean of normalized residual; MEDNR: median of normalized residuals; STDNR: standard deviation of normalized residuals; σ: corresponding standard deviation.

Table 7. ANOVA table.

Source	SS	df	MS	F	Prob > F
Columns	0.73	6	0.122 16	0.129 48	0.992 67
Error	2304.71	2443	0.943 39		
Total	2305.45	2449			

Note: SS: The sum of squares for each source; df: The degrees of freedom for each source; MS: The mean squares for each source; F: F-statistic.

Scherbaum *et al* (2004) have used the values of the LH median, plus the mean, median and standard deviation of the normalized residuals to rank different ground motion models for seismic hazard analysis. Based on this scheme models fall into 4 groups in this ranking scheme. Models are ranked class C, as the lowest acceptable group once the LH median is greater than 0.2 and the absolute values of the median and mean of the normalized residuals with corresponding standard deviations are less than 0.75 and the standard deviation for the normalized residuals smaller than 1.5. If the LH median is more than 0.3 and the absolute values of the median and mean of the normalized residuals with corresponding standard deviations are less than 0.5 and the standard deviation for the normalized residuals smaller than 1.25, models is ranked class B. Models are ranked class A, as the most satisfactory group when the LH median is larger than 0.4 and the absolute values of the median and mean of the normalized residuals with corresponding standard deviations are less than 0.25 and the standard deviation for the normalized residuals smaller than 1.125. Eventually, models that could not satisfy these criteria fall into the unacceptable group, class D. As it can be seen from table 6, all methods are classified as class A.

Analysis of variance

Since table 6 clearly shows that the estimated means of normalized residuals obtained by using the geometric mean of PGA values of horizontal components from different regression methods are not equal, the technique called analysis of variance (ANOVA) is used to compare the means of normalized residuals obtained from these methods (Montgomery and Runger 2003, Mendenhall and Sincich 2007). In one-way ANOVA, the null hypothesis here is defined as if samples (normalized residuals) come from populations with equal mean which is zero. The results from one-way ANOVA are tabulated in table 7 and the box plot of these methods is demonstrated in figure 9. A box plot graphically shows the centrality and skewness of the distribution of a dataset. This plot also indicates the 25th, 50th (median), and 75th percentiles, in addition to minimum, maximum and outlier values (Montgomery and Runger 2003, Mendenhall and Sincich 2007). In accordance with the *p*-values from table 7, the null hypothesis cannot be rejected at the 5% significance level. Figure 9 also depicts that there is no evidence of significant discrepancy between the normalized residuals derived from various regression methods.

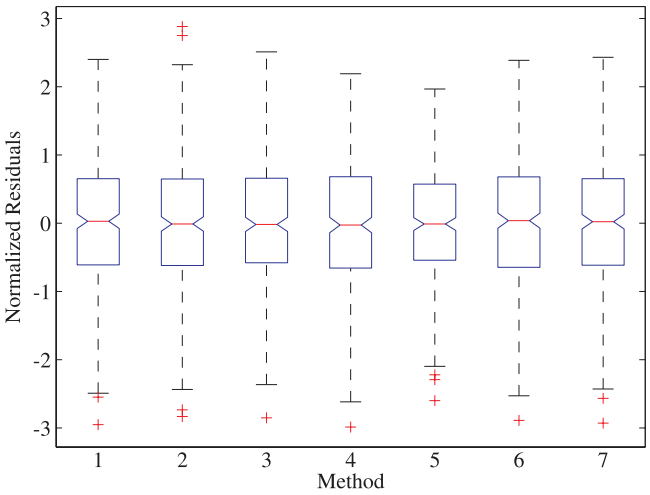

Figure 9. Box plot of the normalized residuals. On each box, the line in the middle of the box is the median. The down and up sides of the box show the 25th and 75th percentiles of the sample. The horizontal lines at the end of each column called whiskers present the minimum and maximum of the sample. Finally the '+' sign shows the outliers in each sample.

Table 8. The dependency of the standard deviation on magnitude and distance.

Method	Magnitude	Distance
1	0.591 63–0.060 26*M*	0.240 76–0.000 69*d*
2	0.355 40–0.004 50*M*	0.288 40–0.000 51*d*
3	0.580 34–0.058 35*M*	0.239 68–0.000 64*d*
4	0.562 50–0.055 01*M*	0.242 09–0.000 63*d*
5	0.553 01–0.053 74*M*	0.239 78–0.000 61*d*
6	0.545 35–0.052 78*M*	0.238 21–0.000 61*d*
7	0.595 02–0.060 73*M*	0.241 04–0.000 68*d*

Table 9. Standard deviations of unknown parameters in each regression method.

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
a_1	0.722 53	0.705 04	0.746 55	0.654 29	0.730 68	0.639 86	0.748 83
a_2	0.105 49	0.096 57	0.109 94	0.098 50	0.110 00	0.096 32	0.108 59
a_3	0.490 21	0.475 16	0.506 51	0.450 60	0.503 21	0.440 66	0.506 06
a_4	0.071 75	0.065 42	0.074 74	0.067 58	0.075 47	0.066 09	0.073 68
a_5	0.879 00	0.841 80	0.816 04	0.720 72	0.804 86	0.704 82	0.895 89
a_6	0.043 93	0.041 87	0.044 36	0.045 28	0.050 57	0.044 29	0.044 14
a_7	0.036 83	0.036 09	0.036 46	0.038 00	0.042 44	0.037 16	0.037 01
a_8	0.043 70	0.042 74	0.060 04	0.045 05	0.050 31	0.044 06	0.043 92
a_9	0.053 81	0.051 91	0.070 03	0.055 58	0.062 07	0.054 35	0.054 05
a_{10}	0.050 46	0.045 60	0.066 95	0.052 04	0.058 11	0.050 89	0.050 70

Constancy of the variance

Constancy of the variance of residuals is one of the basic assumptions about the residuals and the violation of this assumption is considered as a deficiency in the model. To overcome this problem, it is suggested to use a transformation on the variables in the functional form or to consider an appropriate weighting matrix in the regression analysis (Draper and Smith 1981). Another treatment is to keep those estimated coefficients but to use the true variable variance for the prediction of various percentile of ground motion (Ambraseys *et al* 2005).

One way to check constancy of the variance assumption is to use the pattern of residuals against variables. When this pattern is like a funnel it illustrates that the variance is variable and therefore using a transformation or WLS method is necessary to satisfy the basic assumptions of the regression analysis. In this paper, the pure error analysis (Draper and Smith 1981) is performed to estimate the dependency of the standard error on the magnitude and distance. These equations are tabulated in table 8.

No weighting matrix was used for methods 1, 4, and 7, but a weighting matrix is considered in the remaining methods. As it can be observed from the gradients of the reported equations in table 8, the dependency of the standard deviation on the distance is insignificant, while the dependency on the magnitude still exists and it is considerable even for methods in which the weighting matrix is applied except method 2. In this regard, the dependency of the standard deviation on the magnitude cannot be rejected at the 5% significance level and it should be accounted for in the regression analysis. Therefore, all methods, except the WLS that accounts for the weight which is obtained from the pure error analysis of residuals, do not yield in obtaining BLUEs. It means that methods in which a blind weighting matrix is supposed are not helpful, so there is a need to use the pure error analysis of residuals to estimate the accurate weight and then consider it in the regression analysis.

Variance of the estimators

The variance covariance matrices of the unknown parameters using the geometric mean of PGA values of the horizontal components for different methods are estimated and

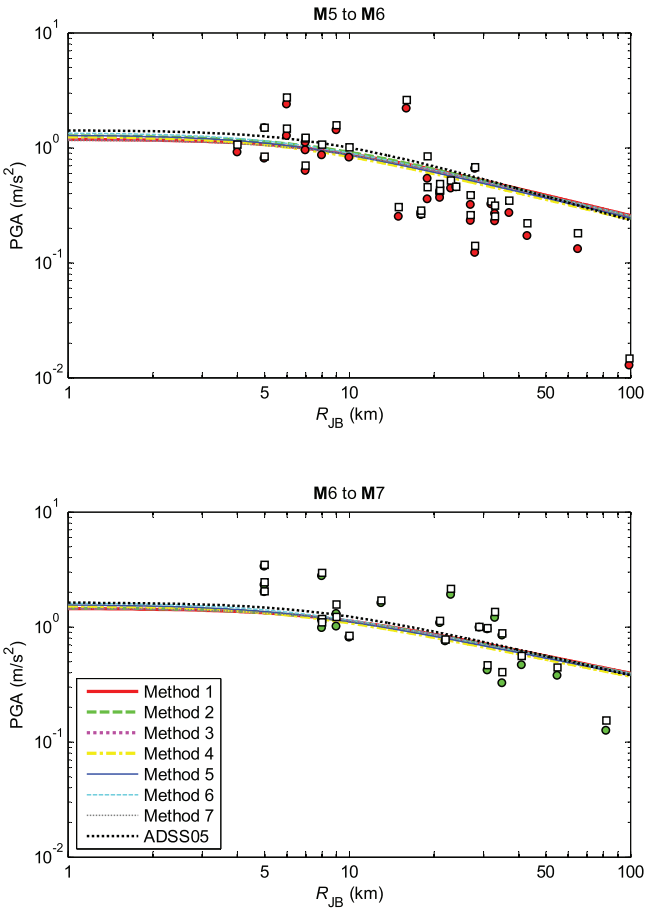

Figure 10. The comparison between estimated accelerations from different regression methods and observations for a normal fault at a stiff site. Circle and rectangular represent the geometric mean and larger horizontal component, respectively.

the diagonal elements are considered to calculate the standard deviation of each coefficient. All standard deviations for these methods are given in table 9. As it was expected the method 2 (WLS) has the minimum standard deviations for the most coefficients.

Therefore, using a weighting matrix derived from the pure error analysis leads to having the optimal standard deviations for the coefficients. Furthermore, table 9 implies that the coefficients a_1, a_3, and a_5 possess the highest standard deviations

and it can be confirmed by table 2 since these coefficients have the larger discrepancy for different applied methods.

Comparison with data

To visually explore the differences between the regression methods, we plot the decay rate (attenuation) of estimated accelerations with distance as well as observations (see figure 10). To this end, stiff site condition and normal fault mechanism are considered in obtaining accelerations using the functional form since the most data available are in this category. As can be seen, the difference between the estimated accelerations obtained from different regression method is insignificant particularly at distances more than 5 km. The small difference between decay rates at short distances less than 5 km can be attributed to the lack of data for this range of distance. On the other hand, all methods overestimate the acceleration at long distances. This can be related to the effect of ignoring the anelastic or intrinsic attenuation (Sedaghati and Pezeshk 2016) term in the functional form. In fact, this plot demonstrates that choosing an appropriate functional form is very important, whereas the influence of selecting a specific regression method to acquire the coefficients of that chosen functional form is insignificant. In figure 10, the GMPE proposed by Ambraseys *et al* (2005) for the larger component of PGAs is also plotted for comparison. As can be seen, this equation is above all other GMPEs derived in this study, since we used the geometric mean of horizontal components.

We also used geometric mean of SA values of two horizontal components at periods of 0.2 and 1 s to investigate the performance of different regression methods. All aforementioned tests are done for SAs at 0.2 and 1 s and results are similar to the results inferred using PGA.

Summary and conclusions

OLS estimation supposes that all events have an identical weight and reliability and it ignores the correlation between different records from a specific event. This estimation has a simple algorithm and the solution can be quickly found. WLS estimation assumes that some earthquakes are more informative and consequently have higher weight than the other earthquakes, but it neglects the correlation between records measured for an earthquake at different sites. One-stage maximum likelihood estimation by Joyner and Boore (1993) accounts for the correlation between different records and different earthquakes. Two-stage methods with various proposed weighting matrices separate the coefficients which are constant for records of a specific earthquake such as the magnitude and fault mechanism from the other coefficients like the distance and site condition. These methods are computationally inefficient. The GA has a simple concept but it is computationally very slow. In addition the solution is not unique and each time this method runs various results can be obtained. We can summarize the results as follows:

- Pure error analysis should be performed as a part of developing GMPEs to gain insight about the true variance of the derived equations which can be applied to estimate different percentiles of ground motions to be used in PSHA.
- Various goodness-of-fit tests, graphical methods, and the LH test demonstrate that the performance of the WLS estimation and one-stage maximum likelihood method is better in comparison to the other considered regression methods. Of course, the WLS methods is not perfect because it neglects the correlation between records. Thus, the one-stage maximum likelihood method in which the true variance obtained from the pure error analysis is considered can be selected as the best regression method to derive coefficients of GMPEs.
- The ordinary GA is very slow and does not result in improving the standard deviation. Also, the statistical tests illustrate that the distribution of residuals of GA is very similar to the distribution of residuals of OLS. Therefore, the OLS estimation is preferred compared to the ordinary GA.
- Two-stage methods are considered as worst regression methods because not only are slow and complicated but also yield in higher standard deviations for the functional form. Further, the pure error analysis shows that the true variance is not constant even after considering a weighting matrix in the regression procedure.
- Choosing an appropriate functional form is very important to develop GMPEs, whereas the influence of selecting a specific regression method to acquire the coefficients of that chosen functional form is insignificant.

Acknowledgments

We are indebted to Professor Nicholas N Ambraseys who provided us the CD ROM of European and Middle Eastern strong motion data released by Ambraseys *et al* (2000). May his soul rest in peace. The authors would also like to thank anonymous reviewers for their thoughtful comments to improve this paper. This project was partially supported by the Tennessee Department of Transportation.

References

Ambraseys N N, Douglas J, Sarma S K and Smit P M 2005 Equations for the estimation of strong ground motions from shallow crustal earthquakes using data from Europe and the Middle East: horizontal peak ground acceleration and spectral acceleration *Bull. Earthq. Eng.* **3** 1–53

Ambraseys N N, Douglas J, Sigbjornsson R, Bergethierry C, Suhadolc P, Costa G and Smit P M 2004 Dissemination of European strong-motion data *13th World Conf. on Earthquake Engineering (Vancouver, Canada)* vol 2

Ambraseys N N, Smit P M, Beradi R, Rinaldis D, Cotton F and Berge C 2000 *Dissemination of European Strong-motion Data (CD ROM Collection, Directorate-General XII Environmental and Climate Programme ENV4-CT97-0397)* (Brussels: European Commission)

Al Atik L, Abrahamson N, Bommer J J, Scherbaum F, Cotton F and Kuehn N 2010 The variability of ground-motion prediction models and its components *Seismol. Res. Lett.* **81** 794–801

Bagheri A, Ghodrati Amiri G, Khorasani M and Haghdoust J 2011 Determination of attenuation relationships using an optimization problem *Int. J. Optim. Civil Eng.* **4** 597–607

Beyer K and Bommer J J 2006 Relationships between median values and between aleatory variablitlies for different definitions of the horizontal component of motion *Bull. Seism. Soc. Am.* **96** 1512–22

Bommer J J and Abrahamson N A 2006 Why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates? *Bull. Seism. Soc. Am.* **96** 1967–77

Boore D M 2010 Orientation-independent, non geometric-mean measures of seismic intensity from two horizontal components of motion *Bull. Seism. Soc. Am.* **100** 1830–5

Boore D M, Joyner W B and Fumal T E 1993 Estimation of response spectra and peak accelerations from western North American earthquakes: an interim report *US Geol. Surv.* Open-File Report 93-509

Boore D M, Watson-Lamprey J and Abrahamson N A 2006 Orientation-independent measures of ground motion *Bull. Seism. Soc. Am.* **96** 1502–11

Bozorgnia Y *et al* 2014 NGA-West2 research project *Earthq. Spectra* **30** 973–87

Brillinger D R and Preisler H K 1984 An exploratory analysis of the Joyner–Boore attenuation data *Bull. Seism. Soc. Am.* **74** 1441–50

Brillinger D R and Preisler H K 1985 Further analysis of the Joyner–Boore attenuation data *Bull. Seism. Soc. Am.* **75** 611–4

Cabalar A F and Cevik A 2009 Genetic programming-based attenuation relationship: an application of recent earthquakes in turkey *Comput. Geosci.* **35** 1884–96

Campbell K W 1997 Empirical near-source attenuation relationships for horizontal and vertical components of peak ground acceleration, peak ground velocity, and pseudo-absolute acceleration response spectra *Seism. Res. Lett.* **68** 154–79

Chatterjee S and Hadi S A 2006 *Regression Analysis by Example* 4th edn (New York: Wiley)

Douglas J 2003 Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates *Earth-Sci. Rev.* **61** 43–104

Douglas J 2004a An investigation of analysis of variance as a tool for exploring regional differences in strong ground motions *J. Seismol.* **8** 485–96

Douglas J 2004b Use of analysis of variance for the investigation of regional dependence of strong ground motions *13th World Conf. on Earthquake Engineering (Vancouver, Canada)* vol 375

Douglas J 2012 Consistency of ground-motion predictions from the past four decades: peak ground velocity and displacement, Arias intensity and relative significant duration *Bull. Earthq. Eng.* **10** 1339–56

Douglas J 2014a Ground motion prediction equations 1964–2014 www.gmpe.org.uk/

Douglas J 2014b Fifty years of ground-motion models *Proc. 2nd European Conf. on Earthquake Engineering and Seismology (2ECEES) (a Joint Event of the 15th ECEE and 34th General Assembly of the ESC) (August 2014)*

Douglas J and Smit P M 2001 How accurate can strong ground motion attenuation relations be? *Bull. Seism. Soc. Am.* **91** 1917–23

Draper N R and Smith H 1981 *Applied Regression Analysis* 2nd edn (New York: Wiley)

Frohlich C and Apperson K D 1992 Earthquake focal mechanisms, moment tensors, and the consistency of seismic activity near plate boundaries *Tectonics* **11** 279–96

Fukushima Y and Tanaka T 1990 A new attenuation relation for peak horizontal acceleration of strong earthquake ground motion in Japan *Bull. Seism. Soc. Am.* **80** 757–83

Gen M and Cheng R 2000 *Genetic Algorithms and Engineering Optimization* (New York: Wiley)

Gentle J E 1998 *Numerical Linear Algebra for Applications in Statistics* (Berlin: Springer)

Goldberg D E, Korb B and Deb K 1989 Messy genetic algorithms: Motivation, analysis, and first results *Complex Syst.* **3** 493–530

Holland J 1975 *Adaptation in Natural and Artificial Systems* 1st edn (Cambridge, MA: MIT Press)

Joyner W B and Boore D M 1993 Methods for regression analysis of strong-motion data *Bull. Seism. Soc. Am.* **83** 469–87

Ljung L 1999 *System Identification: Theory for the User* 2nd edn (Englewood Cliffs, NJ: Prentice Hall)

Masuda T and Ohtake M 1992 Comment on 'A new attenuation relation for peak horizontal acceleration of strong earthquake ground motion in Japan' by Y Fukushima and T Tanaka *Bull. Seism. Soc. Am.* **82** 521–2

McLaughlin K L 1991 Maximum likelihood estimation of strong-motion attenuation relationships *Earthq. Spectra* **7** 267–279

Mendenhall W and Sincich T 2007 *Statistics for Engineering and the Sciences* 5th edn (Upper Saddle River, NJ: Pearson Prentice Hall)

Montgomery C D and Runger C G 2003 *Applied Statistics and Probability for Engineers* (New York: Wiley)

Rezaeian S, Petersen M D and Moschetti M P 2015 Ground motion models used in the 2014 US National seismic hazard maps *Earthq. Spectra* **31** S59–84

Shao J and Tu D 1995 *The Jackknife and Bootstrap* (New York: Springer) (doi 10.1007/978-1-4612-0795-5)

Scherbaum F, Cotton F and Smit P 2004 On the use of response spectral-reference data for the selection and ranking of ground-motion models for seismic-hazard analysis in regions of moderate seismicity: the case of rock motion *Bull. Seism. Soc. Am.* **94** 2164–85

Searle S R 1971 *Linear Models* (New York: Wiley)

Sedaghati F and Pezeshk S 2016 Estimation of the coda-wave attenuation and geometrical spreading in the New Madrid seismic zone *Bull. Seismol. Soc. Am.* **106** 1482–98

Sobhaninejad G, Noorzad A and Ansari A 2007 Genetic algorithm (GA): a new approach in estimation strong ground motion attenuation relationships *4th Int. Conf. on Earthquake Geotechnical Engineering (Thessaloniki, Greece)*

Tavakoli B and Pezeshk S 2005 Empirical-stochastic ground-motion prediction for Eastern North America *Bull. Seism. Soc. Am.* **95** 2283–96

Tavakoli B and Pezeshk S 2007 A new approach to estimate a mixed model–based ground motion prediction equation *Earthq. Spectra* **23** 665–84

West T B, Welch B K and Galecki T A 2007 *Linear Mixed Models: a Practical Guide Using Statistical Software* (London: Chapman and Hall)

Yilmaz S 2011 Ground motion predictive modelling based on genetic algorithms *Nat. Hazard Earth Syst.* **11** 2781–9

Youngs R R, Abrahamson N, Makdisi F I and Sadigh K 1995 Magnitude-dependent variance of peak ground acceleration *Bull. Seism. Soc. Am.* **85** 1161–76