A New Approach to Estimate a Mixed Model–Based Ground Motion Prediction Equation

Behrooz Tavakoli,[a)] M.EERI, and Shahram Pezeshk,[a)] M.EERI

A derivative-free approach based on a hybrid genetic algorithm (HGA) is proposed to estimate a mixed model–based ground motion prediction equation (attenuation relationship) with several variance components. First, a simplex search algorithm (SSA) is used to reduce the search domain to improve the convergence speed. Then, a genetic algorithm (GA) is employed to obtain the regression coefficients and the uncertainties of a predictive equation in a unified framework using one-stage maximum-likelihood estimation. The proposed HGA results in a predictive equation that best fits a given ground motion data set. The proposed HGA is able to handle changes in the functional form of the equation. To demonstrate the solution quality of the proposed HGA, the regression coefficients and the uncertainties of a test function based on a simulated ground motion data set are obtained. Then, the proposed HGA is applied to fit two functional attenuation forms to an actual data set of ground motion. For illustration, the results of the HGA are compared with those used by previous conventional methods. The results indicate that the HGA is an appropriate algorithm to overcome the shortcomings of the previous methods and to provide reliable and stable solutions. [DOI: 10.1193/1.2755934]

INTRODUCTION

Derivation of ground motion prediction equation and estimation of its coefficients and uncertainties are significant components of seismic hazard analysis for seismically active regions. The predictive equations in such regions may be reliably estimated from statistical calculations based on extensive ground motion data recorded in a region. There are several nonlinear mathematical functions that relate a given ground motion parameter (e.g., peak ground acceleration [PGA]) to seismological parameters of a seismic event in a data set, such as earthquake magnitude, source-to-site distance, style of faulting, and local site conditions.

A statistical regression procedure is performed to estimate the residual error (the difference between an observation and an estimated value) and the regression coefficients in a given predictive equation. An extensive verification is required to investigate whether a proposed predictive equation provides a good description of the ground motion data. It is common in probabilistic seismic hazard studies to distinguish between various uncertainties to better understand ground motions at a given site. The discussion of the partitioning of uncertainty is ambiguous. The probability seismic hazard analysis

[a)] Department of Civil Engineering, The University of Memphis, Memphis, TN 38152

Earthquake Spectra, Volume 23, No. 3, pages 665–684, August 2007; © 2007, Earthquake Engineering Research Institute

(PSHA) generally distinguishes between epistemic uncertainty (due to lack of data and knowledge) and aleatory uncertainty (random or apparently random variability)—see Toro et al. (1997) for more details. The residual error or the sigma (σ) term in predictive equations, which may be made up of several variance components, is treated as aleatory variability. The decomposition of residual error into two variance components dates back to Brillinger and Preisler (1984). They partitioned the residual error into two parts, namely intra- (within) event and inter- (between) event terms. Joyner and Boore (1993) incorporated three variance components into a ground motion prediction equation, namely earthquake-to-earthquake component, site-to-site component, and record-to-record component. The site-to-site component and the record-to-record component are generally lumped into the intra-event variability term due to the limited number of recordings available from different earthquakes for a given site. The partitioning between inter-event and intra-event variability is key to the understanding of the nature of the scatter, but has not been formally incorporated into PSHA yet. In general, a ground motion prediction equation is defined as a nonlinear mixed model incorporating both regression coefficients (fixed effects) and uncertainties with several variance components (random effects).

There are four conventional methods to perform a statistical regression analysis to develop a ground motion prediction equation (attenuation relationship) and to estimate the associated uncertainties. These four methods are (1) one-stage weighted least-squares regression (Campbell 1989); (2) two-stage weighted least-squares regression (Joyner and Boore 1993); (3) one-stage maximum-likelihood regression, which was first introduced by Brillinger and Preisler (1984, 1985), then improved by Abrahamson and Youngs (1992), and later re-examined by Joyner and Boore (1993); and (4) Bayesian expectation-maximization (EM) regression (Chen and Tsai 2002). The aim of all these regression methods is to provide the most accurate estimates of the regression coefficients and variance components. These methods provide explicit statistical regression procedures for estimating the variance components. The one-stage maximum-likelihood methods (all parameters are determined simultaneously) are used to give a more accurate partitioning of the variance components than the least-squares methods. In the one-stage methods, an EM algorithm (Brillinger and Preisler 1985, Chen and Tsai 2002) or a certain search algorithm (Abrahamson and Youngs 1992, Joyner and Boore 1993) is applied to obtain the maximum likelihood estimates of the variance components. Some of these methods (e.g., Brillinger and Preisler 1985, Chen and Tsai 2002), which are based on the EM algorithm, do not necessarily work in the absence of good initial estimates that appear to guarantee convergence of the algorithm employed. Unreasonable initial estimates might lead to biased estimates of the variance components. Abrahamson and Youngs (1992) suggested an alternative algorithm to maximize the likelihood of the set of observations without EM, which is considered to give more stability. Although the 1992 Abrahamson and Youngs algorithm provides an explicit formula for the variance component estimates, an additional regression procedure is required to estimate the regression coefficients.

To estimate a general ground motion prediction equation, there is a need for a flexible search algorithm to obtain statistically the best regression coefficients and variance

of components under a unified framework. The search algorithm must be capable of handling changes in the functional form in the attenuation curves with no additional regression analyses. This would imply that all model parameters and the uncertainties are estimated simultaneously (one-stage method) and there is no need to construct derivatives of the predictive equation. Genetic algorithm (GA) is a directed stochastic search method (Holland 1975, Goldberg 1989) based on the principles of natural selection. A hybrid genetic algorithm (HGA) is a combination of the GA with a simulated stochastic method to reduce the search domain and find the suitable sequence of initial guesses for learning and estimating the best regression coefficients and uncertainties in a ground motion prediction equation.

In this study, we develop an alternative approach based on a hybrid genetic algorithm (HGA) for one-stage maximum-likelihood estimation of mixed models. To demonstrate the solution quality of the proposed HGA, the regression coefficients and the uncertainties of a test function, which is based on a simulated ground motion data set, are obtained. Then, the proposed HGA is applied to fit two functional attenuation forms to an actual data set of ground motion recordings in Taiwan using data provided in Chen and Tsai (2002). To illustrate the strengths and limitations of the proposed algorithm, the model parameters and the residual error in the predictive equation are estimated using the previous conventional algorithms, and then compared with those determined by the proposed algorithm. Finally, as another example of the HGA application, we define a complex functional attenuation form and determine the best estimate of the regression coefficients and the variance components from the actual ground motion data.

GENERAL MIXED MODEL–BASED PREDICTIVE EQUATION

The following nonlinear regression model is used to denote a mixed-based ground motion prediction equation

$$Y_{ij} = f(\boldsymbol{x}_{ij}, \boldsymbol{\theta}) + \sum_{\varphi=1}^{c} \boldsymbol{X}_{\varphi}\boldsymbol{b}_{\varphi} + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the jth ground motion parameter (e.g., PGA) from the ith event ($i = 1, \ldots, n$, and $j = 1, \ldots, n_i$), n is the number of events, and n_i is the number of recording for the ith event. An event represents a group of recordings (cluster) that are stochastically dependent, such as the set of strong-motion recordings collected during a single earthquake. The predictive equation consists of three terms to reflect the clustered nature of the ground motion data and the involvement of random effects. The first term of Equation 1, $f(\boldsymbol{x}_{ij}, \boldsymbol{\theta})$, is a known nonlinear functional form. The vector $\boldsymbol{x}_{ij}$ is a vector of independent variables including the earthquake magnitude (M_i) and the source-to-site distance (R_{ij}), and $\boldsymbol{\theta}$ is a vector of fixed effects for regression coefficients. The second term of Equation 1 is used to model the inter-event variations among clusters. The vector $\boldsymbol{b}_{\varphi}$ denotes a specific vector of random effects for the φ-factor, and c is the number of factors to be included in the analysis, such as earthquake-specific, site-specific, and path-specific factors. The matrix $\boldsymbol{X}_{\varphi}$ is the incidence matrix for random effects. The last term, ε_{ij}, represents intra-event variations within the clusters, which is the residual error

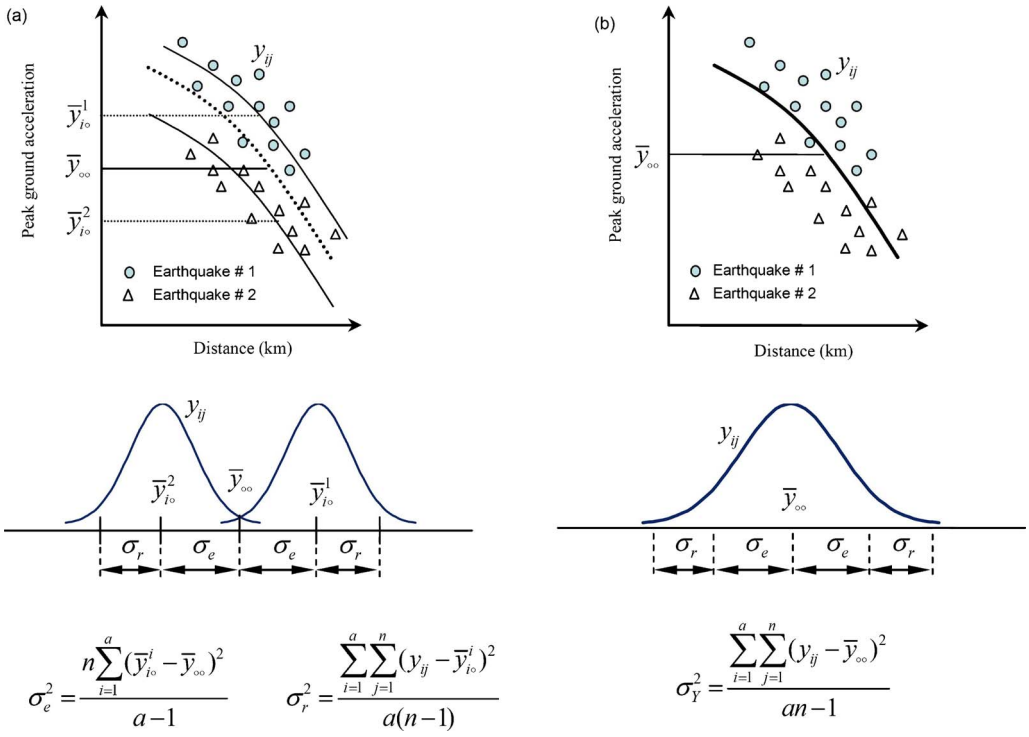

Figure 1. Total deviation is partitioned into two type of errors, namely, inter- and intra-event terms. The inter-event term is a vector of specific random effect such as earthquake-specific components, and the intra-event term is the residual error. The inter-event (random effect) and the intra-event (random error) terms are normally distributed with means zero and variances σ_e^2 and σ_r^2, respectively.

for the *j*th recording from the *i*th event. The random effects and the random error (ε_{ij}) are normally distributed with zero means and variances σ_φ^2 and σ_r^2, respectively. The variance σ_φ^2 is independent of the variance σ_r^2; therefore, the variance of a ground motion parameter is estimated to be $\sigma_Y^2 = \sigma_\varphi^2 + \sigma_r^2$. The variance σ_φ^2 may be partitioned into three types of variance components: σ_e^2, σ_s^2, and σ_p^2. The variance σ_e^2 represents the earthquake-specific deviation obtained for each earthquake magnitude, the variance σ_s^2 represents the site-specific deviation obtained for each site at the different magnitudes, and the variance σ_p^2 represents the path-specific deviation obtained for each record at the different sites.

Figures 1a–b show how the vector of deviations about overall ground motion mean may be partitioned into two components of variance. Suppose that the ground motion data consist of two earthquakes with the same magnitude, as shown in Figure 1a–b, each of which represents a group of recordings from a random location in a given seismic region. When two groups are being compared, the ground motion recordings from all the

groups are involved in computing a total ground motion mean (Y_∞). The total deviation (σ_Y) is based on how far each recording in each group differs from this total mean. The total deviation is decomposed into two error terms. The inter-event term (σ_e) represents the deviation of each group mean from the overall mean, while the intra-event term (σ_r) represents the deviation of each individual recording from the relevant group mean.

The ground motion data (Y_{ij}) is represented as a Gaussian random variable with the overall mean, $\boldsymbol{E}(Y_{ij})$, and variance-covariance matrices ($\boldsymbol{V}$) obtained by the following equations (Searle 1971):

$$\boldsymbol{E}(Y_{ij}) = f(\boldsymbol{x}_{ij}, \boldsymbol{\theta}) \tag{2}$$

and

$$\boldsymbol{V} = \mathrm{var}(Y_{ij}) = \sum_{\varphi=1}^{c} \sigma_\varphi^2 \boldsymbol{X}_\varphi \boldsymbol{X}_\varphi^t + \sigma_r^2 \boldsymbol{I}_N \tag{3}$$

where N is the total number of ground motion data points, $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix, $\boldsymbol{X}_\varphi$ is the incidence matrix for random effects, and the superscript t denotes matrix transposition. Assuming the ground motion data has a multivariate normal distribution with mean $\boldsymbol{E}(Y_{ij})$ and the variance-covariance matrix $\boldsymbol{V}$, a multidimensional normal probability density function can be used as a likelihood-based estimate of the parameter values. The log-likelihood of $(\boldsymbol{\theta}, \boldsymbol{V} | Y_{ij})$ under the Gaussian model is:

$$\log l(\boldsymbol{\theta}, \boldsymbol{V} | Y_{ij}) = -\frac{1}{2}\{N \log(2\pi) + \log|\boldsymbol{V}| + [Y_{ij} - f(\boldsymbol{x}_{ij}, \boldsymbol{\theta})]^t \boldsymbol{V}^{-1} [Y_{ij} - f(\boldsymbol{x}_{ij}, \boldsymbol{\theta})]\} \tag{4}$$

where $|\boldsymbol{V}|$ is the determinant of the variance-covariance matrix and the superscript t denotes matrix transposition. Equation 4 is considered as the objective function of HGA and has to be solved for the elements of $\boldsymbol{\theta}$ and the variance components inherent in $\boldsymbol{V}$. The maximum-likelihood estimates of $(\boldsymbol{\theta}, \boldsymbol{V})$ are obtained by maximizing the right-hand side of the equation. Maximizing $\log l(\boldsymbol{\theta}, \boldsymbol{V} | Y_{ij})$ in Equation 4 or equivalently minimizing $-2 \log l(\boldsymbol{\theta}, \boldsymbol{V} | Y_{ij})$ is often an expensive and ill-conditioned problem. For instance, it is more difficult to construct the partial derivatives of Equation 4 to find the maximum-likelihood estimators for a given nonlinear predictive equation with more random factors (φ). The methods based on derivation require calculating the inverse of a matrix with a size equal to the number of the random effects in each iteration, and the linearization of a given nonlinear regression function with a Taylor's-series expansion about the regression coefficients. Finding the best search defined by the conventional optimization procedure often involves solving an inflexible large-scale maximization problem, in particular for a complex nonlinear predictive equation. Thus, the derivative-free methods provide a flexible alternative to the algorithms used currently for the regression analysis of strong-motion data.

In this study, we propose an alternative search method based on a hybrid genetic algorithm to find the best value of parameters $(\boldsymbol{\theta}, \boldsymbol{V})$ that minimize the objective function. The objective function directly sets up the basis for selection of parameters, each of

which represents a candidate solution to do the best curve fitting. When the total number of ground motion data points and random effects is large, finding the best estimate according to a HGA is more appropriate. The variance estimates of the variance components can be obtained directly after finding the variance components (e.g., σ_e^2, σ_s^2, and σ_r^2) based on the HGA. Details of the variance estimates following Searle (1970) are given in the Appendix.

OPTIMIZATION USING HGA

The HGA is a directed stochastic search technique (a derivative-free approach) that is able to provide an optimal solution to compute the vector of the model parameter values ($\boldsymbol{\theta}$) and the variance components ($\boldsymbol{V}$) defined in Equation 4. The basic idea is to maintain a population of possible solutions that evolves and improves over time through a process of competition and controlled variation. The HGA is different from conventional random algorithms since it combines the elements of directed and stochastic search by using the process of natural selection. The HGA uses first a simplex search algorithm (Lagarias et al. 1998) to reduce the search domains for each parameter, and then a genetic algorithm (GA) to randomly generate an initial population within the reduced search space. The search domains are estimated for the model parameter values based on a fixed effects regression (no assumption is made about the random effects). This assumption involves choosing the model parameters that minimize the sum of squares of deviations of the observations from their expected values defined in Equation 4. The reduced search domains are only considered to improve the convergence speed of the HGA. Therefore, unreasonable initial values of variance components do not cause a problem.

The HGA is used to estimate simultaneously the new model parameters and the variance components in Equation 4. A HGA consists of initialization, evaluation, reproduction, crossover, and mutation. An initial population of possible solutions to Equation 4 is first constructed in a random way and represented in a vector form. These vectors are of the same length and are called strings ($\boldsymbol{S}$) or chromosomes. The length of each string (L) is determined by the number of model parameters (regression coefficients) and variance components (uncertainties) used in the ground motion prediction equation. A string vector form may be expressed as

$$\boldsymbol{S}_{i,j} = (\boldsymbol{\theta}, \sigma_\varphi^2, \sigma_r^2) \quad i = 1, \ldots, M \quad j = 1, \ldots, L \tag{5}$$

where M is a population size, and is usually chosen to be more than twice the string length. Each value of this population array is encoded into a binary string with a known number of bits (N_b) assigned for the representation of the level of accuracy or range of each variable. Each row of the population array is a string represented by a binary string of all encoded solutions.

To examine the practical performance of various aspects of the proposed HGA, we considered the following test function:

Table 1. The population array for three regression coefficients of the test function as a sample to illustrate the process of encoding, decoding, and crossover (bold numbers) in an iteration of the HGA

	String (j)				Binary String Length		
Population Size (i)	1 θ_1	2 θ_2	3 θ_3		1 θ_1	2 θ_2	3 θ_3
1	20.274	0.1333	19.843	Decode	**00101111**	00100010	00101110
2	100.51	0.6667	40.549		**11101001**	10101010	01011110
3	40.549	0.5294	13.372	→	01011110	10000111	00011111
4	59.961	0.7804	65.137		10001011	11000111	10010111
5	80.235	0.4275	28.039	Encode	10111010	01101101	01000001
6	37.529	0.2314	18.981		01010111	00111011	00101100
				←			
				—			
				,			
	…	…	…		…	…	…

$$Y = \theta_1 \exp(-\theta_2 X) + \theta_3[(X-\theta_4)^{-2} + \theta_5]^{-1} + \varepsilon \tag{6}$$

where θ_1 through θ_5 are the regression coefficients and $\boldsymbol{\varepsilon}$ is the uncertainty. To simulate a data set, we used $\boldsymbol{\theta} = [107, 0.629, 20, 1.9, 0.75]$. Then, a random number of the form $8 \times [rand(n,1) - 0.5]$ is utilized to the n data points to simulate $\boldsymbol{\varepsilon}$. The HGA goal is to maximize Equation 4 to estimate the vector $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ using the simulated data set. The error term is assumed to be sampled from a normal distribution with mean zero and unknown variance, σ^2. Following the HGA, we obtained the search domain for the regression coefficients $\theta_1 \in [0, 110]$, $\theta_2 \in [0, 1]$, $\theta_3 \in [0, 110]$, $\theta_4 \in [0, 10]$, and $\theta_5 \in [0, 1]$. As a sample for the illustration of string vector, a population array for the first three regression coefficients θ_1, θ_2, and θ_3 are listed in Table 1. The decoding from a binary string into a decimal number is calculated by the following relationship:

$$D_k = \left(\sum_{j=0}^{N_b - 1} \alpha_j \times 2^j \right) \frac{u_k - l_k}{2^{N_b} - 1} + l_k \quad k = 1, \ldots, N \tag{7}$$

where D_k is a certain population assigned for the decimal representation of the kth parameter bounded by $[l_k, u_k]$, and α_j is a binary representation of D_k with N_b bits. For example, if the search domain for the parameter θ_1 is [0, 110], then the binary string (01011110) with length of $N_b = 8$ is decoded into a corresponding decimal number (40.549) as shown in Table 1.

Through three operation rules based on Darwin's natural selection, the HGA performs a directed search for the best solution by maximizing Equation 4. The first rule is *reproduction/selection*. During the reproduction phase, each string is assigned a fitness value derived from its raw performance measure given by the objective function. This

value is used in the selection to bias toward more fit strings. The strings are descended according to their fitness values. Highly fit strings, relative to the whole population, have a high probability of being selected for the next population whereas less fit strings have a correspondingly low probability. Once the strings have been assigned a fitness value, they can be chosen from the population, with a probability according to their relative fitness, and recombined to produce the next generation.

The second rule is *crossover*. Crossover or mating allows pairs of strings from the population to combine their better features to create improved strings for the next population. All strings are paired at random in such a way that each string belongs to only one pair. Each of the pairs in the population undergoes crossover rule with a probability p_c. Any pair not selected for crossover is placed directly into a new population array. As shown in Table 1, consider $\boldsymbol{S}_{1,1}=(\alpha_{1,1},\ldots,\alpha_{1,N_b})=(00101111)$ and $\boldsymbol{S}_{2,1}=(\alpha_{2,1},\ldots,\alpha_{2,N_b})=(11101001)$ to be two binary strings of θ_1, with the size of $N_b=8$ from the current population, that have been selected for crossover. A position $k \in \{1,2,\ldots,N_b-1\}$ as a crossover point is randomly chosen and two new strings are produced. If the crossover point is 5, for example, then the new solutions are:

$$\boldsymbol{S}'_{1,1}=(\alpha_{1,1},\ldots,\alpha_{1,k},\alpha_{2,k+1},\ldots,\alpha_{2,N_b})=(00101\mathbf{001}) \tag{8}$$

$$\boldsymbol{S}'_{2,1}=(\alpha_{2,1},\ldots,\alpha_{2,k},\alpha_{1,k+1},\ldots,\alpha_{1,N_b})=(11101\mathbf{111}), \tag{9}$$

$\boldsymbol{S}'_{1,1}$ and $\boldsymbol{S}'_{2,1}$ are placed in a new binary string and $\boldsymbol{S}_{1,1}$ and $\boldsymbol{S}_{2,1}$ would be removed from the current population.

The last rule is *mutation*. Mutation gives the algorithm an opportunity to branch into previously unexplored regions of the domain space by arbitrarily altering one or more bits of a selected string. Each bit of every string undergoes mutation with the probability p_m. In the simple case, for each bit in a new population a random number is generated between [0, 1]. If the random number is greater than the probability p_m, the bit is unchanged. Otherwise, the bit is placed by a reverse random bit of each number represented by strings to make a new population array.

The population is now relabeled as a new population array and the cycle of operations is repeated. This process of natural selection continues until some termination criterion (e.g., number of generations) is met, at which time the best string achieved is generally taken as the optimized solution.

Comparison of the best parameter estimation of test function together with the true function and the corresponding simulated data are plotted in Figure 2. The bias of each regression coefficient is listed in Table 2, where the bias is the difference between the estimated and the true values. The maximum error is 2.92% and is associated with regression coefficient θ_4. The error can be reduced by increasing the population size or the number of generations. This would imply that the HGA has a small bias overall, and is an appropriate method to fit complex nonlinear functions to a given data set.

Figure 3 gives an overview of the proposed HGA to determine the best estimate of

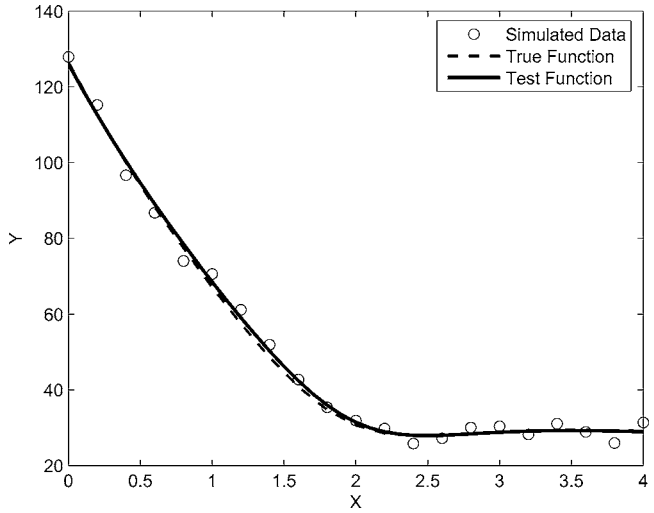

Figure 2. Comparison of the true value of the parameters with those values estimated by the HGA for the test function discussed in this study.

the model parameters and variance components in a certain ground motion prediction equation. The proposed HGA used to compute the vectors $\boldsymbol{\theta}$ and $\boldsymbol{V}$ in a given predictive equation is summarized as follows:

1. Construct the search domains for the vectors $\boldsymbol{\theta}$ and $\boldsymbol{V}$ using the SSA to improve the convergence speed.
2. Generate a random population of M strings within the search domains (candidate solutions for the problem).
3. Evaluate the fitness of each string in the population and find an optimum solution.
4. Generate a new population by repeating the following steps until the new population reaches population size M:

Table 2. Simulation results for the population size of 40 based on 100 generations

	Parameter Values				
HGA	θ_1	θ_2	θ_3	θ_4	θ_5
True	107.0	0.629	20.0	1.90	0.75
Estimated	106.7417	0.612	19.7522	1.9554	0.7646
Bias	−0.2583	−0.017	−0.2478	0.0554	0.0146
Error	−0.24%	−2.72%	−1.24%	2.92%	1.95%

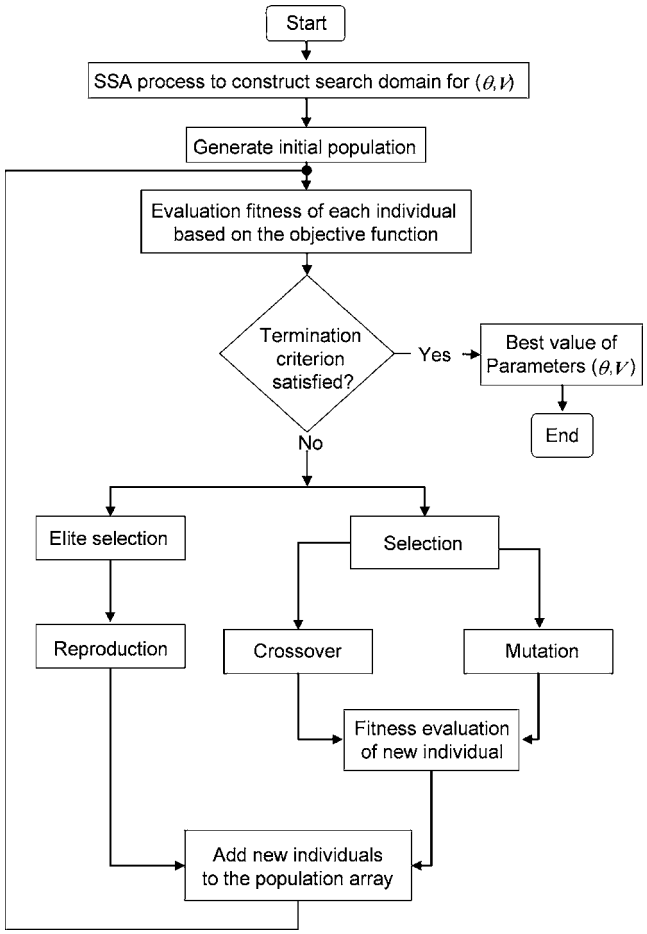

Figure 3. A flowchart for an alternative search fitting method based on a hybrid genetic algorithm to find the best value of parameters $(\boldsymbol{\theta}, \boldsymbol{V})$.

I. Select two strings from the current population, giving preference to highly fit strings (high fitness values). Automatically copy the fittest string to the next generation.

II. With a given crossover probability p_c, crossover the strings to form two new strings. If no crossover was performed, a new string is an exact copy of a string in the current population.

III. With a given mutation probability p_m, randomly swap two bits of each number represented by strings to make a new string.

IV. Copy the new string into a new population.

5. Copy the newly generated population over the existing population.

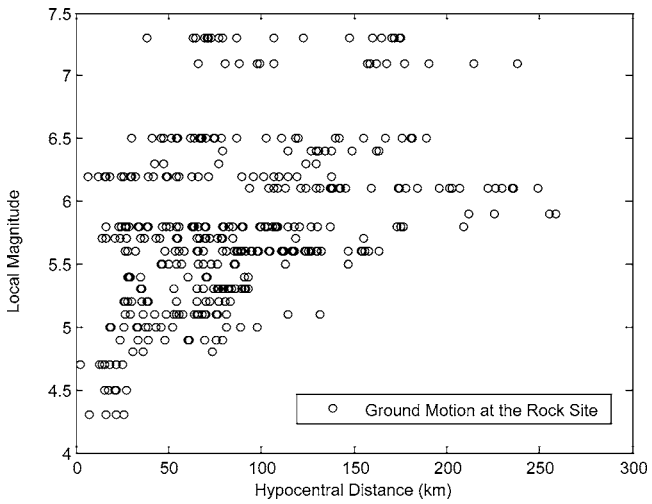

Figure 4. The distribution of 424 recordings in the PGA data set of Chen and Tsai (2002) plotted as a function of magnitude and hypocentral distance.

6. If the loop termination condition is satisfied, then stop and return the best solution in current population.
7. Otherwise, go to Step 3.

The choice of p_c and p_m depends on the nature of the objective function. Despite this fact, a value of p_c between 0.6 and 0.9 (Herrera et al. 1998) and a value of p_m between $1/N_b$ and 0.5 (Back 1993) are often recommended to promote exploration and population diversity.

TWO EXAMPLES OF THE HGA APPLICATION

As the first example, we employed the proposed HGA to fit a typical strong ground motion data set to the following general predictive equation:

$$\log_{10} y_{ij} = \theta_1 + \theta_2 M_i + \theta_3 M_i^2 + \theta_4 R_{ij} + \theta_5 \log_{10}(R_{ij} + \theta_6 10^{\theta_7 M_i}) + \varepsilon_{ij} \qquad (10)$$

where the y_{ij} value is the geometric mean of two horizontal peak ground accelerations for the jth recording from the ith event in cm/sec^2, M_i is the local magnitude, R_{ij} is the hypocentral distance (km), ε_{ij} is a total residual (random effects and random error), and θ_1 through θ_7 are the regression coefficients to be determined. In this study, we used the same ground motion data set and the general ground motion perdition equation considered by Chen and Tsai (2002).

Figure 4 shows the distributions of 48 earthquakes used in this study in terms of magnitude and distance. There are 424 recordings from 48 earthquakes with magnitudes greater than 4.0 in the ground motion data set. As shown in Figure 4, large earthquakes are recorded at greater distances than small earthquakes.

Table 3. The HGA parameters used to estimate the best-fitting attenuation relationship

HGA Parameters	Values Used
Population size (M)	**200**
Maximum number of generations	**100**
Probability of crossover (p_c)	**0.6**
Probability of mutation (p_m)	**0.04-0.1**
Length of strings	**25 bits for each parameter**
Search domain for the vector of $\boldsymbol{\theta}$	**[−5 5]**
Search domain for the vector of $\boldsymbol{V}$	**[0.1 0.5]**
Termination criterion	**100**

The HGA parameters used in this study are listed in Table 3. The proposed HGA is performed to obtain the optimal values of vectors $\boldsymbol{\theta}$ and $\boldsymbol{V}$ given in Equation 10. The evolution of the objective function is plotted in Figure 5. There are some ups and downs in the convergence of the objective function since the best solution is not retained at each generation and the algorithm is allowed to explore the entire search domain.

The optimum solution with minimizing the objective function is obtained in 25 generations. However, the proposed algorithm continued to iterate pending the termination criterion in order to search for a better solution.

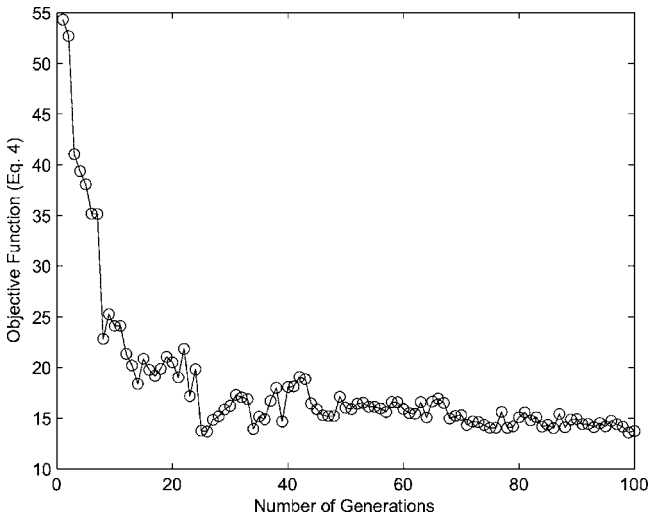

Figure 5. Maximum number of generations used to estimate the best result of parameters and uncertainties in the ground motion prediction equation for a data set of ground motion recordings in Taiwan.

Table 4. Results of the example application to parameter estimates and standards errors

Parameters	HGA[a]	B&P[b]	J&B[c]	C&T[d]
θ_1	**−3.4712**	−3.507	−3.767	−4.366
θ_2	**2.2639**	2.221	2.507	2.540
θ_3	**−0.1546**	−0.144	−0.177	−0.172
θ_4	**0.0021**	0.0017	0.0019	0.0017
θ_5	**−1.8011**	−1.833	−2.025	−1.845
θ_6	**0.0490**	0.0875	0.016	0.0746
θ_7	**0.2295**	0.203	0.386	0.221
σ_r	**0.2203**	0.2349	0.2358	0.2358
σ_e	**0.2028**	0.2057	0.2075	0.2128

[a] HGA = The algorithm used in this study.
[b] B&P = Billinger and Preisler algorithm (1985).
[c] J&B = Joyner and Boore algorithm (1993).
[d] C&T = Chen and Tsai algorithm (2002).

The final HGA results obtained in this study for the best-fit shape of the predictive equation are provided in Table 4. The attenuation shape for local magnitude 5.5 is illustrated in Figure 6, which plots the observed PGA for a subset of the data with magnitudes 5.0–6.0, in comparison to the predictive equation. The regression analysis for the

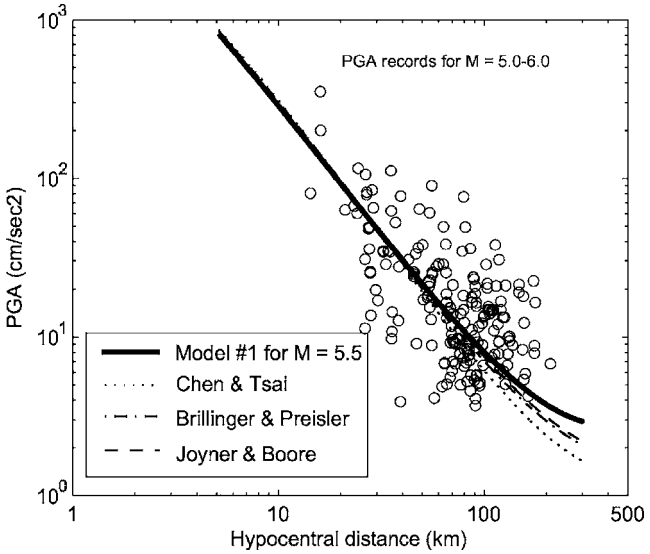

Figure 6. The geometric mean of two horizontal peak ground accelerations in cm/sec^2 for events of M5.0–6.0 (circles) compared to the predictive model (Model #1) developed from the HGA (line) and the previous search algorithms.

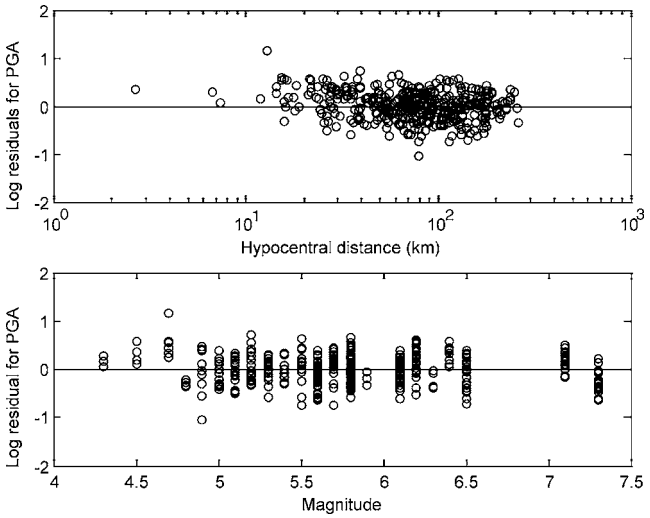

Figure 7. Log residuals (=log observed-log predicted PGA) for the regression of the ground motion data versus hypocentral distance and magnitude.

fit of the data to the predictive equation is also performed by using the algorithms of Brillinger and Preisler (1985), Joyner and Boore (1993), and Chen and Tsai (2002). The results are listed in Table 4 for comparison. The proposed HGA produces a solution that is in good agreement with the previous studies, with a slightly better fit (smaller error term).

The log residual is defined as the difference between the log of the observed ground motion amplitude and the log of the predicted ground motion amplitude according to Equation 10. Figure 7 illustrates the total residuals (random effects and random error) as a function of hypocentral distance and magnitude. There are no apparent trends in the residuals. The total residuals have been partitioned into two variance components (σ_e^2, σ_r^2). When we mix the vector of fixed effects $(\boldsymbol{\theta})$ with the vector of random effect, for the earthquake-specific component (σ_e^2), the residual error can be plotted against the hypocentral distance as shown in Figure 8. Comparison of the total residuals with the residual error shows that the prediction errors can be reduced when the random effects are corrected.

The total standard deviation of $\log_{10} y_{ij}$ in the regression is estimated to be 0.299. In Table 4, the variances of the two variance components represented by (σ_e^2, σ_r^2) can be estimated by the variance-covariance matrix defined in the Appendix. In this case, we ignore the effect of site-specific deviation (σ_s); hence the variance-covariance matrix reduces to the following inverse matrix:

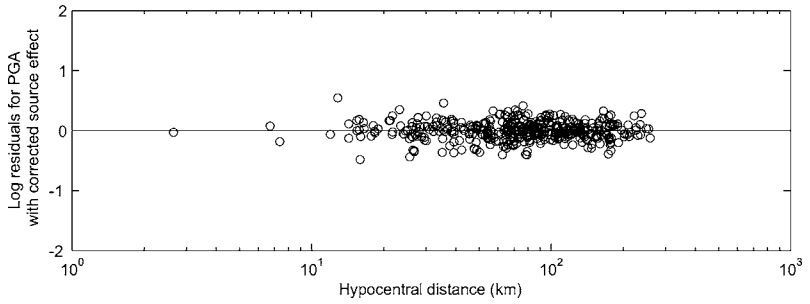

Figure 8. Corrected ground motion residuals for the regression of the ground motion data versus hypocentral distance.

$$\begin{pmatrix} v(\sigma_e^2) & \mathrm{cov}(\sigma_e^2,\sigma_r^2) \\ \mathrm{cov}(\sigma_e^2,\sigma_r^2) & v(\sigma_r^2) \end{pmatrix} = 2\begin{pmatrix} \tau_{ee} & \tau_{er} \\ \tau_{er} & \tau_{rr} \end{pmatrix}^{-1}. \tag{11}$$

In this study, the variance estimates are evaluated based on the estimated variance components and the number of recordings from an earthquake. Following the equations provided in the Appendix and the ground motion catalog used in this study, the Equation 11 is determined to be:

$$\begin{pmatrix} v(\sigma_e^2) & \mathrm{cov}(\sigma_e^2,\sigma_r^2) \\ \mathrm{cov}(\sigma_e^2,\sigma_r^2) & v(\sigma_r^2) \end{pmatrix} = \begin{pmatrix} 9.77E-05 & -1.82E-06 \\ -1.82E-06 & 1.25E-05 \end{pmatrix}. \tag{12}$$

The log residuals demonstrate that the attenuation model of Equation 10 provides a satisfactory description of the ground motion data at distances of up to 100 km. The geometric spreading of body waves does not have a spherical shape beyond 100 km, since the direct shear waves are superimposed by waves reflected from the Moho discontinuity at distances of the order of 100 km (typically of the order of twice the Moho depth). As shown in Figure 6, the trend of PGA versus distance changes beyond 100 km because of the effect of geometric spreading.

The variation of focal depth would affect the shape of the attenuation curves at near-source distances. Thus, the effect of amplitude saturation (a constant amplitude value as distance is decreased) should be considered in the plot of a given ground motion prediction equation. In this case, a complex predictive equation is required to explain both the geometric attenuation of seismic waves at distances of beyond 100 km, and the saturation effect at near-source distances. One approach involves fitting several attenuation curves to the data and restricting the use of each curve to specified intervals of distance. This approach is particularly appropriate when no well-defined simple curve can be found to summarize the ground motion data. Piecewise fitted curves are rarely used in ground motion prediction equations based on empirical data, where near-source saturation effects and the change in attenuation rate at large distances are usually handled by the inclusion of a pseudo-depth coefficient and a combination of logarithmic and linear distance terms.

As the second example of the HGA application, the following complex functional form (Tavakoli and Pezeshk 2005) is utilized to fit the ground motion equation to the data set:

$$\ln(Y_{ij}) = \theta_1 + \theta_2 M_i + \theta_3(8.5 - M_i)^{2.5} + \theta_9 \ln(r_{rup} + 4.5) + (\theta_4 + \theta_{13} M_i)\ln R_{ij} + (\theta_8 + \theta_{12} M_i)R_{ij} + \varepsilon_{ij} \quad r_{rup} \leqslant 70km \tag{13a}$$

$$\ln(Y_{ij}) = \theta_1 + \theta_2 M_i + \theta_3(8.5 - M_i)^{2.5} + \theta_9 \ln(r_{rup} + 4.5) + \theta_{10} \ln\left(\frac{r_{rup}}{70}\right) + (\theta_4 + \theta_{13} M_i)\ln R_{ij} + (\theta_8 + \theta_{12} M_i)R_{ij} + \varepsilon_{ij} \quad 70 < r_{rup} \leqslant 130km \tag{13b}$$

$$\ln(Y_{ij}) = \theta_1 + \theta_2 M_i + \theta_3(8.5 - M_i)^{2.5} + \theta_9 \ln(r_{rup} + 4.5) + \theta_{10} \ln\left(\frac{r_{rup}}{70}\right) + \theta_{11} \ln\left(\frac{r_{rup}}{130}\right) + (\theta_4 + \theta_{13} M_i)\ln R_{ij} + (\theta_8 + \theta_{12} M_i)R_{ij} + \varepsilon_{ij} \quad r_{rup} \geqslant 130km. \tag{13c}$$

In these terms, r_{rup} (km) is a rupture distance and defined as the closest distance to the fault rupture, and M_i is moment magnitude for ith event. The finite-fault geometry causes the average distance from the observation point to the fault to introduce extended-source effects, since at any point we cannot be close to the entire fault plane. This implies that there is a pseudo-depth (effective focal depth), which will appear to be the source of radiation if it is treated as a point-source model. Thus, the distance measure R_{ij} includes a magnitude dependence to illustrate the effect of the extended source on the shape of the attenuation curve based on a pseudo-depth, which is given by Campbell and Bozorgnia (2003):

$$R_{ij} = \sqrt{r_{rup}^2 + (\theta_5 \exp[\theta_6 M_i + \theta_7(8.5 - M_i)^{2.5}])^2}. \tag{14}$$

Seismogenic depth of 3 km is also used to measure rupture distance from hypocentral distance. Finding the best-fitting curve with changing regression functional forms by using the previous algorithms (e.g., Joyner and Boore 1993, Chen and Tsai 2002) often involves solving a complicated and large-scale minimization problem. Constructing derivatives of Equation 13 with respect to regression coefficients and using more random factors are the main problems in these algorithms. The HGA is used directly to determine the unknown regression coefficients in Equation 13. In this way transition points are incorporated in a fitted curve simply as the boundary points between adjacent predictive equations. The HGA parameters are the same as those used in the first example (Table 3). The final result is the best estimate of the coefficients that fits the predicted model to the ground motion data set. The regression coefficients are estimated as 4.2298, 0.7699, −0.0252, −1.501, −0.5571, 0.3495, −0.0069, −0.0032, 0.0094, 1.2125, −0.7826, 0.0003, and 0.0001 for θ_1 through θ_{13}, respectively. The standard deviations are estimated to be σ_e=0.1995 and σ_r=0.1875. The result of the ground motion values for an earthquake of magnitude 5.5 is compared with the previously mentioned methods as shown in Figure 9. The average focal depth of 10 km is used to convert various distances to the horizontal distance. The discrepancy in ground motions between Equations

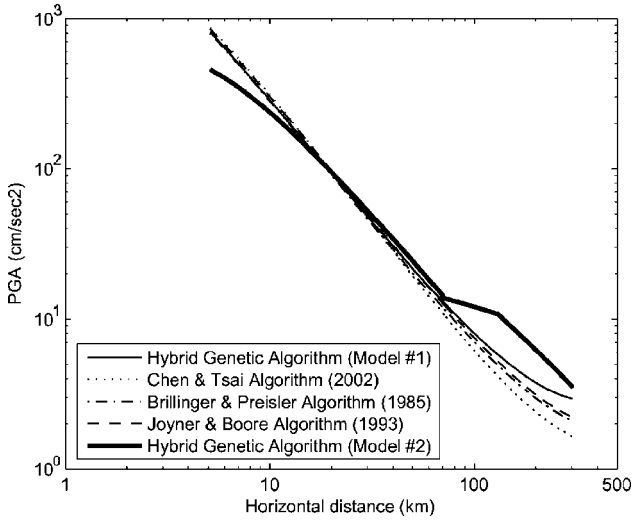

Figure 9. Comparison of peak ground accelerations (Model #2) for an event of M5.5 developed in this study (thick solid lines) with the predictive equation (Model #1) developed from the HGA (line) and the previous search algorithms.

13 and 10 is significant at short distances and at distances above 100 km. Therefore, the selection of predictive equation is a crucial factor in the magnitude-distance ranges that are significant to seismic hazard analysis.

DISCUSSION AND CONCLUSIONS

We have proposed a search algorithm for the estimation of ground motion prediction equations and the associated components of variance. The algorithm combines the elements of directed and stochastic search to reduce the search domain of parameters, and in turn the time of process. The HGA can be applied to complex predictive equations with several variance components. The proposed algorithm can easily cope with a larger number of variance components compared with existing algorithms, although it takes more time to reach an optimized result.

The HGA process starts with a population of solutions to find a theoretical attenuation curve, then continues by optimizing and fitting the theoretical curve to the ground motion data. The HGA focuses on a population of attenuation coefficients and variance components, each of which is generated randomly within a certain search domain obtained using SSA. Coefficients and variance components are grouped in variable sets, each of which is composed of a series of strings to define a possible solution for the problem. The performance of the variables, as described by the objective function and the constraints, is represented by the fitness of each variable. A mathematical expression calculates a fitness value for each solution of the objective function.

Comparison of the numerical results with those obtained in the previous studies cited

herein shows that the HGA performs successfully in estimating the parameters in mixed model–based ground motion prediction equation with several variance components. The algorithm maintains a population of potential solutions, whereas all other methods process a single point of the search space. The HGA, unlike most existing models, is independent of some supplementary information, such as derivatives, to solve a complex problem. The algorithm only uses an objective function and several quite simple genetic operations for the potential solution to the problem. The flexibility of the algorithm allows solving problems with a series of changing regression functional forms that partition the attenuation function's domain. The transition points can be incorporated in a fitted attenuation curve simply as the boundary points between adjacent curves.

ACKNOWLEDGMENTS

We wish to thank the many people who contributed information and criticism. We thank Chu-Chuan Peter Tsai for providing the data set used in this study and the computer program to calculate Bayesian expectation-maximization regression. We also thank three anonymous reviewers for their comments and suggestions. This work was partially funded by the Tennessee Department of Transportation.

APPENDIX: VARIANCES OF LARGE SAMPLE MAXIMUM-LIKELIHOOD ESTIMATORS

Suppose that the ground motion data consist of N records coming from M earthquakes and S sites. Using the results of Searle (1970), the estimated variance-covariance matrix of (σ_e^2, σ_s^2 and σ_r^2) is given by the following relationship:

$$\begin{pmatrix} v(\sigma_e^2) & \operatorname{cov}(\sigma_e^2,\sigma_s^2) & \operatorname{cov}(\sigma_e^2,\sigma_r^2) \\ \operatorname{cov}(\sigma_e^2,\sigma_s^2) & v(\sigma_s^2) & \operatorname{cov}(\sigma_s^2,\sigma_r^2) \\ \operatorname{cov}(\sigma_e^2,\sigma_r^2) & \operatorname{cov}(\sigma_s^2,\sigma_r^2) & v(\sigma_r^2) \end{pmatrix} = 2\begin{pmatrix} \tau_{ee} & \tau_{es} & \tau_{er} \\ \tau_{es} & \tau_{ss} & \tau_{sr} \\ \tau_{er} & \tau_{sr} & \tau_{rr} \end{pmatrix}^{-1}$$

with

$$m_{ij} = n_{ij}\sigma_s^2 + \sigma_r^2$$

$$A_{ipq} = \sum_{j=1}^{S_i} (n_{ij}^p / m_{ij}^q), \quad \text{for integers } p \text{ and } q$$

$$q_i = 1 + \sigma_e^2 A_{i11}$$

and

$$\tau_{ee} = \sum_{i=1}^{M} A_{i11}^2 / q_i^2$$

$$\tau_{es} = \sum_{i=1}^{M} A_{i22} / q_i^2$$

$$\tau_{er} = \sum_{i=1}^{M} A_{i12}/q_i^2$$

$$\tau_{ss} = \sum_{i=1}^{M} (A_{i22} - 2\sigma_e^2 A_{i33}/q_i + \sigma_e^4 A_{i22}^2/q_i^2)$$

$$\tau_{sr} = \sum_{i=1}^{M} (A_{i12} - 2\sigma_e^2 A_{i23}/q_i + \sigma_e^4 A_{i12}A_{i22}/q_i^2)$$

$$\tau_{rr} = \sum_{i=1}^{M} (A_{i02} - 2\sigma_e^2 A_{i13}/q_i + \sigma_e^4 A_{i12}^2/q_i^2) + \left(N - \sum_{i=1}^{M} S_i\right) \Big/ \sigma_r^4$$

in which M is the number of events, and n_{ij} is the number of recording for the ith event and the jth site. Thus, the total number of records can be obtained by $N=\Sigma_{i=1}^{M}\Sigma_{j=1}^{S_i} n_{ij}$.

REFERENCES

Abrahamson, N. A., and Youngs, R. R., 1992. A stable algorithm for regression analysis using the random effects model, *Bull. Seismol. Soc. Am.* **82**, 505–510.

Back, T., 1993. Optimal mutation rates in genetic search, *Proceedings of the Fifth International Conference on Genetic Algorithms, San Mateo, Calif.*, 2–8.

Brillinger, D. R., and Preisler, H. K., 1984. An exploratory analysis of the Joyner-Boore attenuation data, *Bull. Seismol. Soc. Am.* **74**, 1441–1450.

——, 1985. Further analysis of the Joyner-Boore attenuation data, *Bull. Seismol. Soc. Am.* **75**, 611–614.

Campbell, K. W., 1989. *Empirical prediction of near-source ground motion for the Diablo Canyon power plant site*, San Luis Obispo County, California, *USGS Open-File Report 89–484*, U.S. Geological Survey, Washington, D.C.

Campbell, K. W., and Bozorgnia, Y., 2003. Updated near-source ground motion (attenuation) relations for the horizontal and vertical components of peak ground acceleration and acceleration response spectra, *Bull. Seismol. Soc. Am.* **93**, 314–331.

Chen, Y. H., and Tsai, C. C. P., 2002. A new method for estimation of the attenuation relationship with variance components, *Bull. Seismol. Soc. Am.* **92**, 1984–1991.

Goldberg, D. E., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, Reading, Mass.

Herrera, F., Lozano, M., and Verdegay, J., 1998. Tackling real-coded genetic algorithms: operators and tools for behavioral analysis, *Artif. Intell. Rev.* **12**, 265–319.

Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor.

Joyner, W. B., and Boore, D. M., 1993. Methods for regression analysis of strong-motion data, *Bull. Seismol. Soc. Am.* **83**, 469–487.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E., 1998. Convergence properties of

the Nelder-Mead simplex method in low dimensions, *SIAM J. Optim.* **9**, 112–147.

Searle, S. R., 1970. Large sample variances of maximum likelihood estimators of variance components, *Biometrics* **26**, 505–524.

——, 1971. *Linear Models*, Wiley, New York, NY.

Tavakoli, B., and Pezeshk, S., 2005. Empirical-stochastic ground-motion prediction for eastern North America, *Bull. Seismol. Soc. Am.* **95**, 2283–2296.

Toro, G. R., Abrahamson, N. A., and Schneider, J. F., 1997. Model of strong ground motions from earthquakes in central and eastern North America: best estimated and uncertainties, *Seismol. Res. Lett.* **68**, 41–57.

(Received 28 March 2006; accepted 13 February 2007)