Multivariate Bayesian Regression Analysis Applied to Ground-Motion Prediction Equations, Part 2: Numerical Example with Actual Data

by Danny Arroyo and Mario Ordaz

Abstract  An application of a linear multivariate Bayesian regression model, described in a companion article, to obtain a ground-motion prediction equation (GMPE) using a set of actual ground-motion records and a realistic functional form is presented. Based on seismological grounds and on an adopted functional form, we include a sound discussion about how the prior information required for the model can be defined. For the regression analyses we use two subsets of ground-motion records from the Next Generation of Ground-Motion Attenuation Models (NGA) database. We compare the results obtained with the Bayesian model with those obtained through the one-stage maximum-likelihood and the constrained maximum-likelihood methods. The advantages of the Bayesian approach over traditional regression techniques are discussed.

Introduction

In a companion article (Arroyo and Ordaz, 2010) we have presented a linear multivariate Bayesian technique to compute pseudoacceleration (SA) ground-motion prediction equation (GMPE) models. Because in our companion article we used a set of synthetic data with known statistical parameters, we did not include a sound discussion on how the prior information required by the model can be defined. We find it more useful to present such discussion here, with an application involving real data and a currently favored functional form.

For the example, we have selected two subsets of ground-motion records composed of accelerograms recorded at rock sites, which were included in the Next Generation of Ground-Motion Attenuation Models (NGA) database (Chiou et al., 2008). The first subset is very similar to that used by Idriss (2008) in the development of a GMPE for rock sites in the NGA project, while the second subset was specifically designed to resemble a poorly sampled database. We decided to perform the analysis only for rock sites in order to keep a reasonable length to this article. Therefore, the discussion of the results is focused on the magnitude scaling and on the decay of the ground motion with distance.

Nowadays, the accepted regression methods in the development of GMPEs are the one-stage maximum-likelihood method and the two-stage method. The one-stage maximum-likelihood approach was introduced by Brillinger and Priesler (1984); later Abrahamson and Youngs (1992) and Joyner and Boore (1993, 1994) proposed computational algorithms to implement the method. In addition, Joyner and Boore (1993, 1994) studied how the one-stage maximum-likelihood method and the two-stage method are related; they found that both methods lead essentially to the same results. Hence, we decided to use only the one-stage maximum-likelihood method for the comparisons. Although the Bayesian technique described in the companion article (Arroyo and Ordaz, 2010) is a multivariate model, in order to shorten the presentation, we performed the regression analysis only for peak ground acceleration (PGA). Nevertheless, a similar procedure could be used to set the prior information for other SA ordinates. In some cases the multivariate analysis could become very complex, depending on the correlation between observations and on the adopted functional form. In practice, an analyst can choose between performing multivariate analysis and performing univariate analysis and including the correlation between SA ordinates in another stage of the analysis (Goda and Atkinson, 2009).

It is important to state that we do not intend to develop a new GMPE for the NGA project. Our objectives are (1) to show how prior information can be defined for Bayesian analysis and (2) to compare Bayesian regression analysis and regression analyses performed with commonly used methods. The results show some advantages of the Bayesian technique over the currently accepted regression analysis methods.

Data Set Selection

The first set of ground motions (hereafter referred to as set A) includes 906 accelerograms recorded at rock sites during 44 earthquakes of the NGA database (see the Data and Resources section). We considered 874 accelerograms recorded at sites with average shear-wave velocity, in the upper 30 m of sediments ($V_{s30}$), between 450 and 900 m/sec and 32 records from sites with $V_{s30}$ between 900 and 1428 m/sec. We included these 32 records in view of the fact that they were recorded at sites with National Earthquake Hazards Reduction Program (NEHRP) B classification. We only used records at free field stations and in first floor of buildings with no more
than two stories. The distribution of the records, in terms of the moment magnitude ($M_w$) and the closest horizontal distance to the surface projection of the rupture plane ($R_{JB}$), is shown in Figure 1. For set A, $M_w$ is in the range between 4.5 and 7.9, and $R_{JB}$ is in the range between 1 and 280 km. Set A includes 2 events with unspecified fault type, 16 strike-slip events, 8 normal-slip events, and 18 reverse-slip events. Most of the data included in set A are ground motions recorded during reverse-slip events (700 records) and strike-slip events (154 records). Note that the set becomes sparse for $M_w$ larger than 7 and also for $R_{JB}$ larger than 150 km. Set A is very similar to the one used by Idriss (2008) in the NGA project. However, our set is slightly smaller because we have excluded events that yielded only one record. We believe that set A is one of the most complete databases of rock ground motions that can be found nowadays.

The second set (hereafter referred to as set B) was chosen to simulate a poorly sampled database. We generated this set by excluding some records of set A; the number of records included in set B is 358. As shown in Figure 1, set B includes events with intermediate magnitudes and distances. For set B, $M_w$ is in the range between 5.9 and 7, while $R_{JB}$ is in the range between 25 and 80 km. Set B comprises one event with unspecified fault type, five strike-slip events, two normal-slip events, and eight reverse-slip events. Again, most of the records included in set B are ground motions observed during reverse-slip events (278 records) and strike-slip events (66 records). Of course, it would not be advisable to develop a GMPE with such a poorly sampled set. However, we decided to use this set in the analysis in order to demonstrate the stability of the Bayesian approach.

### Functional Form

For the regression analysis, we used the functional form adopted by Boore and Atkinson (2008; hereafter referred to as BA08) in the NGA project, which is shown in equation (1):

$$y = F_M(M_w) + F_D(R_{JB}, M_w),$$

where $F_M$ and $F_D$ are the magnitude scaling and the distance functions, respectively, and $y$ is the natural logarithm of the GMRotI50 (Boore et al., 2006) of PGA in g units.

$F_D$ and $F_M$ are given by

$$F_D(R_{JB}, M_w) = c_1 \ln \left( \frac{R}{R_{ref}} \right) + c_2 (M_w - M_{ref}) \ln \left( \frac{R}{R_{ref}} \right) + c_3 (R - R_{ref}),$$

$$F_M(M_w) = \begin{cases} e_1 U + e_2 SS + e_3 NS + e_4 RS + e_5 (M_w - M_h) + e_6 (M_w - M_h)^2 & \text{if } M_w \leq M_h, \\ e_1 U + e_2 SS + e_3 NS + e_4 RS + e_5 (M_w - M_h) & \text{otherwise}, \end{cases}$$

where

$$R = \sqrt{R_{JB}^2 + h^2}$$

![Figure 1](image-url)  
**Figure 1.** Sets of ground motions used in the analysis.
and \( c_1, c_2, c_3, e_1, e_2, e_3, e_4, e_5, e_6, e_7, \) and \( h \) are coefficients to be defined by regression analysis. \( U, SS, NS, \) and \( RS \) are dummy variables used to denote unspecified, strike-slip, normal-slip, and reverse-slip fault type, and \( M_{ref}, R_{ref}, \) and \( M_b \) are coefficients to be set in the analysis. We did not group all types of events to compute the unspecified fault type earthquake because we found it more useful, for the objectives of the article, to present results only for strike-slip, normal-slip, and reverse-slip fault type events.

As has been discussed in the companion article (Arroyo and Ordaz, 2010), the proposed multivariate Bayesian scheme is able to include intraevent and interevent variability, the correlation between regression coefficients, and the correlation between SA ordinates at different periods. This level of generality, however, is only achieved if the functional form defined in equation (1) is nonlinear due to the effect of \( h \). Therefore, some considerations must be made in order to apply the Bayesian method.

Although the adopted functional form is nonlinear, it becomes linear once \( h \) is set to a certain value. For this reason, we performed the Bayesian analysis as follows: for several values of \( h \), we computed the regression coefficients through Bayesian analysis (i.e., considering the regression coefficients as random variables with prior and posterior probability density functions). Then, we assigned the value that yields the smaller standard deviation of the residuals to \( h \). This implies that the regression analysis is not fully Bayesian because \( h \) is not considered as a random variable. Nevertheless, the results illustrate the advantages of the Bayesian approach over the traditional methods of regression analysis.

In order to facilitate the comparisons, we set \( R_{ref} = 1 \) km, \( M_{ref} = 4.5, \) and \( M_b = 6.75, \) which are the values used in the BA08 model. The adopted functional form is equivalent to the original BA08 model with \( V_{S30} = 760 \) m/sec.

**Prior Information**

For the Bayesian approach, in order to stabilize the regression analysis, prior information about unknown parameters needs to be defined. For a given value of \( h \), prior mean values of \( c_1, c_2, c_3, e_1, e_2, e_3, e_4, e_5, e_6, \) and \( e_7 \) were set as follows.

We assumed that the amplitude Fourier spectrum at the source, regardless of fault type, can be described by the two-corner point-source model proposed by Atkinson and Silva (2000):

\[
A(f) = C(2\pi f)^2 M_0 \left( \frac{1 - e}{1 + (f/f_a)^2} + \frac{e}{1 + (f/f_b)^2} \right),  \quad (5)
\]

where

\[
C = \frac{R_p V F S}{4 \pi \rho \beta^3}, \quad (6)
\]

\[
\log f_a = 2.181 - 0.496 M_w, \quad (7)
\]

\[
\log f_b = 2.41 - 0.408 M_w, \quad (8)
\]

\[
\log \varepsilon = 0.605 - 0.255 M_w. \quad (9)
\]

\( M_0 \) is the seismic moment, \( R_p = 0.55 \) is the radiation pattern, \( V = 0.70711 \) is a factor that takes into account the equal partitioning of energy in the two horizontal components, \( F_S = 2 \) is the free surface amplification, \( \rho = 2.8 \) g/cm\(^3\) is the mass density in the vicinity of the source, and \( \beta = 3.5 \) km/sec is the shear-wave velocity. Implicit in the two-corner point-source model of equation (5) is a Brune point-source model with a stress drop of 80 bars; further details can be found in Atkinson and Silva (2000).

The source spectrum defined in equation (5) is modified to take into account the propagation through the crust by the following processes:

1. Geometric spreading. Following Atkinson and Silva (2000), we assumed that the geometric spreading can be modeled by \( R^{-1} \) to a distance of 40 km and \( R^{-0.5} \) for \( R \) greater than 40 km.
2. Anelastic attenuation. We modeled the anelastic attenuation with a factor \( e^{-(\pi f/\beta)^2} \), with \( Q = 180 \) for generic rock sites.
3. Crustal amplification. We modeled the amplification of \( S \) waves as they propagate upward through materials of progressively lower velocity with the factors proposed by Boore and Joyner (1997) for generic rock sites.
4. The near-surface attenuation is modeled with factor \( e^{-\pi \kappa f} \) (Anderson and Hough, 1984). Following Atkinson and Silva (2000) we used \( \kappa = 0.03 \) for California ground motions.
5. Near-source saturation. According to Atkinson and Silva (2000), the two-corner point-source model is able to simulate the near-source saturation observed in finite-fault simulations through the use of an equivalent point-source depth computed as \( h = -0.05 + 0.15M_w \). Thus, for the point-source model, \( R = \sqrt{D_{cl}^2 + h^2} \) with \( D_{cl} \) being the closest distance to the fault plane.

Through random vibration theory and the amplitude spectrum described previously, we generated a synthetic set of PGAs for different combinations of \( M_w \) and \( R_{BH} \). We assumed that the geometry of the fault plane is such that \( R_{BH} = D_{cl} \). Using this synthetic database, we computed the prior mean values of the regression coefficients, for each value of \( h \), using the least-squares method. Because we assumed that the point-source model is the same for all fault types, the prior mean values of \( e_2, e_3, \) and \( e_4 \) are equal.

These prior mean values cannot be considered as the true values because they were obtained from a very simple theoretical model, which is independent from fault type and fault geometry. Moreover, many parameters required to define the Fourier amplitude spectrum are valid only for ground motions recorded in California. Therefore, a measure of the degree of certainty in the prior mean values should be
placed when defining their prior covariance matrix (see Arroyo and Ordaz, 2010, for further details). Through this prior covariance matrix, the Bayesian model accounts for the uncertainty in the prior mean values. We set the prior covariance matrix as a diagonal matrix, which implies that, a priori, we believe that the regression parameters are uncorrelated. The regression parameters with the largest uncertainties are $e_2$, $e_3$, and $e_4$ because they depend on effects such as fault type, fault geometry, and site effects. Hence, we assigned a very large value, equal to 10,000, to their variances. This implies that $e_2$, $e_3$, and $e_4$ will not be controlled by their prior mean values, so they will be free to attain the values that yield the best fit to data. For parameters $c_1$, $c_2$, $c_3$, $e_5$, $e_6$, and $e_7$ we used a variance that implies a coefficient of variation equal to 0.59, as it was done in a previous study (Ordaz et al., 1994). Note that regression coefficients could be greatly constrained during the regression analysis by assigning a very small variance to their prior mean values.

Because we perform the analysis just for PGA only, the prior mean value of the variance of the residuals is needed for the analysis. As prior mean value we used $\Sigma_p = 0.49$, which means that we a priori believe that the standard deviation of the residuals is equal to 0.7, a value frequently found in available GMPEs. Also, the degree of certainty in $\Sigma_p$ should be defined. This degree of certainty is controlled by parameter $\nu$: the larger the value of $\nu$, the larger the degree of certainty on $\Sigma_p$. We set $\nu = 7$, which is the minimum value that could be used in order to give a finite value to the variance of $\Sigma_p$; further details can be found in Arroyo and Ordaz (2010).

Finally, the parameters of the prior probability density of $\gamma_c$ (the parameter that accounts for the correlation between observations for the same earthquake) need to be specified. Because the prior density of $\gamma_c$ is a scalar beta density, we set $a = b = 1.5$, where $a$ and $b$ are the parameters of the beta density (further details can be found in the companion article). This implies that our prior knowledge about $\gamma_c$ is vague because the prior density of $\gamma_c$ is very flat.

### Results

We developed the following four GMPEs in order to assess the performance of the Bayesian technique with respect to the one-stage maximum-likelihood method:

1. Bayesian analysis with set A (hereafter referred to as Bayes-A), $\Sigma$
2. Bayesian analysis with set B (hereafter referred to as Bayes-B), $\Sigma$
3. Maximum-likelihood analysis with set A (hereafter referred to as ML-A), $\Sigma$
4. Maximum-likelihood analysis with set B (hereafter referred to as ML-B), $\Sigma$

Also, we compared the GMPEs listed previously with the BA08 model, which is defined in equation (1). The BA08 model was obtained with the two-stage method, although some coefficients were constrained during the analysis; hence, the BA08 model can be considered as a case of the constrained maximum-likelihood method. It is worth noting that the data set used to develop the BA08 model is different from the data set used in this article, because the authors modified the SA values at different soil conditions to obtain equivalent observations for a reference velocity of 760 m/sec, based on the work of Choi and Stewart (2005).

As has been discussed in the companion article (Arroyo and Ordaz, 2010), the posterior mean values of the regression parameters must be computed using the Gibbs sampling method. Hence, the number of terms ($K$) for the Gibbs sampling needs to be defined. The value of $K$ required to attain convergence depends on the correlation between observations and should be defined by inspection of the numerical results. For the considered set of ground motions we found that convergence was attained with small values of $K$; normally, no more than 100 terms were required. One of the disadvantages of the Bayesian technique is the great numerical effort required in the Gibbs sampling method; the results presented in this article were computed on three standard personal computers. All the regression analyses required for the Bayes-A

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>BA08</th>
<th>Prior Value</th>
<th>Bayes-A</th>
<th>Bayes-B</th>
<th>ML-A</th>
<th>ML-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>$-0.66050$</td>
<td>$-1.02357$</td>
<td>$-1.09301$</td>
<td>$-1.63000$</td>
<td>$-1.72878$</td>
<td>$-6.10325$</td>
</tr>
<tr>
<td>$c_2$</td>
<td>$0.11970$</td>
<td>$0.20363$</td>
<td>$0.12960$</td>
<td>$0.18900$</td>
<td>$0.18122$</td>
<td>$1.30826$</td>
</tr>
<tr>
<td>$c_3$</td>
<td>$-0.01151$</td>
<td>$-0.00250$</td>
<td>$-0.00463$</td>
<td>$-0.00199$</td>
<td>$0.00074$</td>
<td>$0.04262$</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$-0.50350$</td>
<td>$0.29240$</td>
<td>$0.47797$</td>
<td>$1.74000$</td>
<td>$2.27514$</td>
<td>$7.78966$</td>
</tr>
<tr>
<td>$e_3$</td>
<td>$-0.75472$</td>
<td>$0.29240$</td>
<td>$0.38781$</td>
<td>$1.68000$</td>
<td>$2.17682$</td>
<td>$7.47312$</td>
</tr>
<tr>
<td>$e_4$</td>
<td>$-0.50970$</td>
<td>$0.29240$</td>
<td>$0.69095$</td>
<td>$2.17000$</td>
<td>$2.42496$</td>
<td>$7.91966$</td>
</tr>
<tr>
<td>$e_5$</td>
<td>$0.28805$</td>
<td>$0.09655$</td>
<td>$0.02416$</td>
<td>$0.02800$</td>
<td>$-0.30972$</td>
<td>$-3.57822$</td>
</tr>
<tr>
<td>$e_6$</td>
<td>$-0.10164$</td>
<td>$-0.13966$</td>
<td>$-0.13903$</td>
<td>$-0.16700$</td>
<td>$-0.18023$</td>
<td>$0.21229$</td>
</tr>
<tr>
<td>$e_7$</td>
<td>$0.00000$</td>
<td>$0.09655$</td>
<td>$0.02565$</td>
<td>$0.02460$</td>
<td>$-0.44444$</td>
<td>$-6.78643$</td>
</tr>
<tr>
<td>$\gamma_c$</td>
<td>$0.26820$</td>
<td>$0.50000$</td>
<td>$0.40129$</td>
<td>$0.45900$</td>
<td>$0.30847$</td>
<td>$0.27288$</td>
</tr>
<tr>
<td>$h$</td>
<td>$1.35$</td>
<td>$2.0$</td>
<td>$2.5$</td>
<td>$2.5$</td>
<td>$12.1$</td>
<td>$17.5$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>$0.69297$</td>
<td>$0.7$</td>
<td>$0.66797$</td>
<td>$0.78152$</td>
<td>$0.65449$</td>
<td>$1.14914$</td>
</tr>
<tr>
<td>$b$</td>
<td>$-0.140$</td>
<td>$-0.171$</td>
<td>$-0.158$</td>
<td>$-0.164$</td>
<td>$-0.158$</td>
<td>$-0.158$</td>
</tr>
</tbody>
</table>
model (one analysis for each value of \( h \) used in the computations) were obtained in three weeks, while the computations for the Bayes-B model were obtained in one week. However, the time of analysis can be reduced if a more powerful computer (i.e., workstation or supercomputer) is used.

In Table 1, the computed parameters for the GMPEs are included. For reference, we have also included the prior mean value of each parameter used in the Bayesian analysis. Note that \( \sigma \) values reported in Table 1 are the standard deviations of the residuals related to set A, except for the case of the prior value of \( \sigma \), where the reported value is that used in the Bayesian analysis. In Table 1, the value of \( \sigma \) related to the BA08 model is larger than the value reported in the NGA project (\( \sigma = 0.564 \)) because we used a different set of ground-motion records. We also include the mean value of the residual \( (b) \) for each GMPE with respect to set A; results suggest that the level of bias is similar for all the GMPEs.

For models ML-A and ML-B, the coefficients \( c_1 \) and \( c_3 \) have physically debatable values. Note that even with set A, positive values of \( c_3 \) and quite large values of \( c_1 \) are obtained with the maximum-likelihood method. This means that the information contained in the data set is not enough to

![Figure 2](image-url)  
Figure 2. Comparison of GMPEs for strike-slip events; open symbols are observations contained in set A.
properly solve the regression analysis for $c_1$ and $c_3$ simultaneously, as Boore and Atkinson (2008) have stated. On the other hand, both Bayesian models lead to $c_1$ and $c_3$ values that seem physically consistent. This is an advantage of the Bayesian technique with respect to the maximum-likelihood method because the Bayesian method was able to simultaneously solve for $c_1$ and $c_3$ parameters. Also, $e_5$ and $e_7$ attain unrealistic values with the maximum-likelihood method, while coefficients obtained with the Bayesian method are physically correct.

It can be argued that the fact that the Bayesian technique leads to coefficients that are physically correct has limited merit because the really important issue is the predicted value related to the GMPE. However, the GMPEs obtained through Bayesian analysis yield more stable predictions than their maximum-likelihood counterparts, as can be observed in Figure 2.

In Figure 2, we present a comparison of the predicted PGA decay with distance for all GMPEs together with the prior model, for strike-slip events. Unsurprisingly, the ML-B model is very unstable and yields satisfactory predictions only for combinations of $M_w$ and $R_{JB}$ that were included in set B. On the other hand, the predictions computed with the ML-A model are more stable, although a debatable change in the slope of the curves is observed for $R_{JB} > 300$ km. The predictions computed with the Bayes-B model are clearly more stable than those obtained with the ML-B model and are close to the predictions of the Bayes-A model, except for $R_{JB} < 20$ km, where the Bayes-B model predicts larger values of PGA. Note that according to $\sigma$ values in Table 1, the Bayes-B model is considerably more accurate than its maximum-likelihood counterpart when it is applied to predict values of set A. This is an advantage of the Bayesian technique with respect to the maximum-likelihood approach because both models were developed with the same poorly sampled set.

We observed that predictions of the Bayesian models are close to predictions of the prior model for small values of $R_{JB}$; conversely, for large values of $R_{JB}$ the predictions of the Bayesian models are far from the prior model. In order to clarify the reasons for this observation, in Table 2 we have included the weighting factors for the Bayesian models. Normally, in Bayesian analysis, posterior mean values of regression coefficients can be expressed as the weighted average between their prior mean values and their maximum-likelihood estimators. In Table 2, $W_d$ and $W_p$ are the weighting factors for the maximum-likelihood estimator and for the prior mean value, respectively. According to Table 2, coefficients $e_2$, $e_3$, and $e_4$ are controlled only by the information contained in the database (i.e., $W_p$ is much smaller than $W_d$) because a very large prior variance was assigned to them in the Bayesian analysis. In the case of the distance term coefficients ($c_1$, $c_2$, and $c_3$), the influence of the prior information decreases as the quality of the set increases. For the Bayes-A model, the prior information has effect only in $c_3$, while in Bayes-B model the prior information has considerable effect on both $c_2$ and $c_3$. On the other hand, the coefficients of the magnitude scaling term remain controlled by its prior mean value even with set A; we observe a significant effect of data only in $e_6$. Interestingly, we note that the same level of prior uncertainty was assigned to $c_1$, $c_2$, $c_3$, $e_5$, $e_6$, and $e_7$; hence, the results suggest that the information contained in the set was able to identify that the prior distance term is not supported by the data, especially at large distances. Conversely, the information contained in the data set was not able to disregard the prior information in the magnitude scaling term because as $M_w$ increases set A becomes sparse (note that in set A there are only four events with $M_w < 5$ and two events with $M_w > 7.5$). As can be observed in Figure 2, the largest differences between ML-A and Bayesian models are observed for $R_{JB} < 20$ km, where there is a lack of data. In order to solve this lack of data, the Bayesian analysis merged the information contained in the data set with the prior model and yielded a GMPE that is very close to observed values in the regions of well-sampled data and that is close to the prior model in those regions were the set is poorly sampled. Clearly, this is another advantage of the Bayesian analysis over the maximum-likelihood approach.

We acknowledge that ML-A and ML-B models are extreme examples of regression analysis because, in practice, most analysts would try to stabilize the regression analysis constraining the range of some coefficients, such that unnecessary influence due to one parameter does not affect the estimation of other parameters. This idea improves the stability of the maximum-likelihood approach; for instance, in the development of the BA08 model, $c_3$ and $h$ were constrained to $-0.01151$ and 1.35, respectively (Boore and Atkinson, 2008). The fixed value of $c_3$ was set using additional data from three small California events that are not included in the NGA project and solving the first-stage analysis with $c_2 = 0$ and $c_1 = -0.8$. According to the weighting factors, the decision of using a constrained value of $c_3$ seems reasonable because the information contained in the data set is scarce to completely define its value. However, in the maximum-likelihood approach, it is not clear which parameters need to be constrained in the analysis. For instance, the results suggest that the use of constrained values of $e_5$ and $e_7$ would have also been reasonable. Although the weighting

<table>
<thead>
<tr>
<th>$e_i$</th>
<th>$W_p$</th>
<th>$W_d$</th>
<th>$e_i$</th>
<th>$W_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>0.0104</td>
<td>0.9896</td>
<td>0.0684</td>
<td>0.9316</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.0367</td>
<td>0.9633</td>
<td>0.5607</td>
<td>0.6393</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.4034</td>
<td>0.5966</td>
<td>0.9776</td>
<td>0.0224</td>
</tr>
<tr>
<td>$e_4$</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0002</td>
<td>0.9998</td>
</tr>
<tr>
<td>$e_5$</td>
<td>0.0001</td>
<td>0.9999</td>
<td>0.0003</td>
<td>0.9997</td>
</tr>
<tr>
<td>$e_6$</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0002</td>
<td>0.9998</td>
</tr>
<tr>
<td>$e_7$</td>
<td>0.9998</td>
<td>0.0001</td>
<td>0.9999</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**Table 2**

Weighting Factors for Bayesian Models
factors are available only after the Bayesian analysis has been performed, we believe that the possibility to quantitatively assess which parameters are not well constrained by the data is another advantage of the Bayesian technique over the maximum-likelihood approach. Of course, an experienced analyst would know which parameters to constrain and which parameters to leave free. But, in our view, the Bayesian formalism allows for a more ordered and systematic inclusion of all the knowledge that is not contained in the data.

In Figures 3 and 4 we compare predictions for reverse-slip events and normal-slip events; in the comparisons we do not include the ML-B model because similar trends as those shown in Figure 2 were observed. In addition, in Figure 5 we compare the mean residuals for set A considering several \( M_w \) and \( R_{JB} \) bins for reverse-slip events. As can be observed, all models yield very similar trends in the residuals, with the exception of Bayes-B model for \( R_{JB} = 0 \) and 100 km, where large residuals are observed as a consequence of the lack of data in set B (although not shown, similar results were observed for other fault type events). Also, the predicted value of the Bayesian models are very close to the predicted values of the BA08 model in the range \( 5 < M_w < 7 \) and \( 10 < R_{JB} < 150 \), that is, where most data are concentrated.

Figure 3. Comparison of GMPEs for reverse-slip events; open symbols are observations contained in set A.
From Figure 5 and Table 1 we conclude that the BA08, Bayes-A, and ML-A models have the same accuracy for set A. However, in those magnitude–distance regions in which the set is sparse, significant differences between the Bayesian models and BA08 model are observed. For $R_{JB} < 10$ km the ratios between predicted values are roughly 1.5, which are similar to the differences observed for other NGA models (Abrahamson et al., 2008). For $R_{JB} = 200$ km the ratios between predicted values are about 1.5, but for $R_{JB} > 200$ km the differences rapidly increase, and it is impossible to define which model yields the correct predicted value. This problem has been recognized in the development of the BA08 model because its authors emphasize that their model should be used only for $R_{JB} < 200$ km. For long distances, the prior model overpredicts the PGA values, especially for large $M_w$ (see Figs. 2–4) and the Bayesian models yield predicted values that are closer to the ML-A model. Considering that the $c_3$ value used in the constrained maximum-likelihood method may not be valid for ground motions outside the California region, we consider that the Bayesian models are more robust; nevertheless, this conclusion is very speculative and certainly more data are needed to validate it. Fortunately,
for such large distances the estimation of ground-motions parameters is not critical from an engineering point of view.

In the case of $R_{JB} < 10$ km the Bayesian models are closer to the prior model because, as has been discussed in the former paragraphs, the information contained in the data set is scarce and the prior information controls the results. With the available data, it is impossible to define which GMPE yields the correct predicted value. However, we note that predictions related to the prior model match some observations in set A (see Figs. 2–4) at this range of $R_{JB}$. This observation is not surprising. It must be recalled that, in order to obtain our prior distributions, we used the stochastic approach, whose ability to predict ground motions for distances shorter than 100 km has been recognized for many years (Hanks and McGuire, 1981; Boore, 1983; Boore and Atkinson, 1987; Atkinson and Silva, 1997, 2000). With the Bayesian technique, all previous knowledge contained in the stochastic approach (or in finite-fault simulations) can be formally merged with the information contained in available sets of recorded ground motions, in order to construct a model that, in our view, is more robust than ground-motion models based on the maximum-likelihood approach, especially in zones where data sets become sparse. Perhaps this is the main advantage of the Bayesian technique over the maximum-likelihood approaches.

In addition, some authors have correctly expressed the need for stochastic models in the study of ground-motion parameters (Atkinson and Silva, 2000). Firstly, the stochastic models provide a conceptual framework for understanding some of the basic physical parameters that control observed ground motions (Atkinson and Silva, 2000). Secondly, comparison of a GMPE obtained with the stochastic approach with purely empirical models allows us to identify possible errors in the extrapolation of empirical GMPE (Atkinson and Silva, 2000). Based on the results presented in this study we can add another reason: the stochastic models are also useful for generating prior information to be incorporated in Bayesian analysis.

Conclusions

We have presented an application of a Bayesian method to compute PGA GMPEs using sets of real ground-motion records and a currently favored functional form. We included a discussion about how, based on seismological grounds, the prior information required for the model can be defined.

The numerical computations required by the Bayesian model are more extensive than the numerical effort required by the maximum-likelihood approach. However, we believe that the Bayesian method has the following advantages that justify its application:

1. With scarce data, the Bayesian approach is more accurate than the one-stage maximum-likelihood method.
2. The Bayesian technique is more stable than the maximum-likelihood approach, especially when a large number of free parameters must be determined by the regression analysis.
3. With the Bayesian analysis it is possible to quantitatively assess which parameters are not well constrained by the data set used.
4. With the Bayesian technique, all the knowledge that is not contained in the data (e.g., the analyst’s experience, results from the stochastic approach, and results from
finite-fault simulations) can be formally merged with the information contained in available sets of ground motions in order to construct a model that works well even in zones where data sets are sparse.

Although it could be argued that advantages (3) and (4) can also be included in traditional analysis by fixing some regression coefficients (as it has been done in most available GMPEs), the Bayesian technique provides a formal framework to properly include these aspects in the regression analysis. For practical applications involving thousands of ground-motion records and current standard computers, perhaps the best solution in the development of GMPEs is to use the Bayesian approach interactively with the maximum-likelihood approach.

It is worth noting that, although we have applied the Bayesian technique to the case of PGA, where most of the records included in the NGA database are usable, there are other periods for which the number of usable records decreases considerably; from the results presented in this article, we believe that in such cases the Bayesian approach could lead to greater improvements in the development of GMPEs.

Based on the presented results, we noticed that there are several future research issues related to the Bayesian approach. For example, the use of other functional forms, the use of finite-fault simulations as prior information, and the development of GMPEs that include prior site amplification factors. For the purpose of further research, our computer code to perform Bayesian analysis is freely available on request.

Finally, we do not think that our PGA ground-motion model is better than the BA08 model, nor better than any other GMPE. We simply have tried to show the application of the Bayesian method presented in a companion article (Arroyo and Ordaz, 2010) with a set of real data and a realistic functional form.

Data and Resources

The Next Generation Attenuation of Ground Motions project is found at http://peer.berkeley.edu/products/nga_project.html (last accessed April 2010).

Acknowledgments

The authors appreciate two anonymous reviewers who greatly contributed to improve the original manuscript. Constructive comments by Gail Atkinson and the assistance of Juan Casillas in editing are also appreciated.

References


Departamento de Materiales
Universidad Autónoma Metropolitana
Av. San Pablo # 180. Colonia Reynosa Tamaulipas
Azcapotzalco CP 02200
México D.F., México
aresda@correo.azc.uam.mx

Instituto de Ingeniería
UNAM Ciudad Universitaria
Coyocán CP 04510
México D.F. México
MOrdazs@pumas.iingen.unam.mx
(D.A.)

Manuscript received 2 October 2009