Maximum Direction to Geometric Mean Spectral Response Ratios using the Relevance Vector Machine

Y. Dak Hazirbaba, J. Tezcan , Q. Cheng
Southern Illinois University Carbondale, IL, USA

SUMMARY:
The 2009 edition of the NEHRP Recommended Provisions for Seismic Regulations for New Buildings and Other Structures redefined the design levels of ground motions as the spectral accelerations in the direction of maximum horizontal response. The 2009 Provision maps have been converted from the geometric mean maps assuming that the maximum direction to geometric mean ratio (MD/GMRotI50) depends only on the vibration period, and the standard deviation of maximum direction motions is the same as that of geometric mean motions. A detailed investigation of the MD/GMRotI50 ratio is essential to understanding the ground motion levels intended for seismic design. In this paper, we use records from strike-slip earthquakes to investigate the dependence of the logarithm of the ratio on magnitude, distance and rupture directivity and estimate its standard deviation using the Relevance Vector Machine regression. Our results suggest that different sets of variables govern the ratio for different vibration periods.

Keywords: Maximum direction to geometric mean ratio, Relevance Vector Machine, Bayesian inference

1. INTRODUCTION

The 2009 edition of the NEHRP Recommended Provisions for Seismic Regulations for New Buildings and Other Structures redefined the design levels of ground motions as the spectral accelerations in the direction of maximum horizontal response. The Ground Motion Prediction Equations (GMPEs) developed in the Next Generation of Ground Motions (NGA) project used in the updated seismic design maps predict a rotated geometric mean called GMRotI50 (Boore et al., 2006), not the spectral acceleration in the direction of maximum response (MD). As opposed to GMRotI50, spectral maximum is not a geometric-mean measure. The 2009 Provision maps have been converted from the geometric mean maps by setting the maximum direction to geometric mean ratio (MD/GMRotI50) equal to 1.1 for short periods and 1.3 for mid-periods, based on the work of Huang et al. (2008). Implicit in this conversion are the following assumptions: (1) The maximum to geometric mean ratio depends only on the vibration period; (2) The standard deviation of maximum direction motions is the same as that of geometric mean motions. A detailed investigation of the MD/GMRotI50 ratio is essential to a better understanding of the ground motion levels intended for seismic design.

In this study, we use 599 pairs of strike-slip records from the PEER-NGA database (http://peer.berkeley.edu/nga) to investigate the dependence of the maximum to geometric mean ratio on magnitude, distance and rupture directivity. Using different subsets of these three predictors as input, we perform Relevance Vector Machine (RVM) regression for the natural logarithm of the ratio for vibration periods between 0.05 second to 4 seconds.

The remainder of the paper is organized as follows. Section 2 describes the maximum direction ground motions, and summarizes recent research findings regarding the variables affecting the MD/GMRotI50 ratio. Section 3 describes the RVM regression model. After a brief introduction to the RVM, the input and output variables, and the ground motion data are described, and the RVM regression algorithm is

outlined in this section. Section 4 presents the findings on the influence of different sets of predictors on the logarithmic ratio, and compares the predictions of the RVM model to those from two existing parametric models. Sections 5 concludes the paper by summarizing the findings.

2. MAXIMUM DIRECTION MOTIONS

The maximum direction motion at period T is defined as the maximum pseudo-acceleration of a single degree of freedom system with natural vibration period of T, taking into account all possible rotations in the horizontal plane. For a given acceleration record, the maximum direction motions can be computed by successively rotating the two orthogonal components through 90^o in 1^o increments, and extracting the maximum pseudo-acceleration for each period. The two components can also be combined into a single time series which will then be rotated through 180^o (Boore, 2010). Alternatively, for each period, the pseudo-accelerations for the two components can be plotted, and the maximum spectral ordinate can be determined from the point furthest from the origin, as illustrated in Fig. 2.1. The figure shows the 5%-damped pseudo-accelerations for T=1.0 second, for the Pacoima Dam recording of the 1971 San Fernando Earthquake. For this record, the 1-second spectral ordinate in the direction of maximum response is 1.42g, which corresponds to a rotation angle of 31^o.

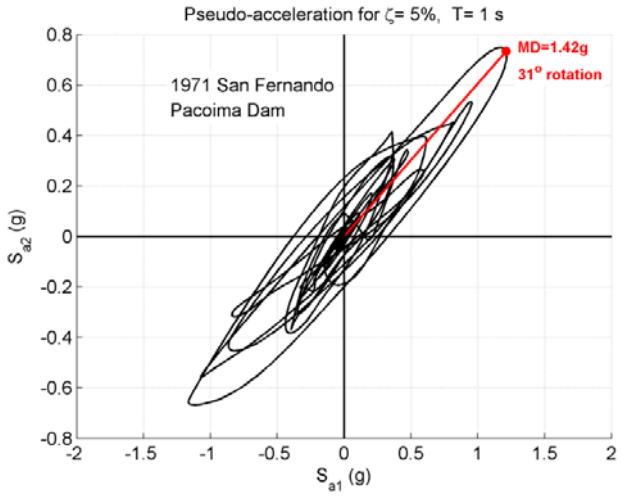

Figure 2.1. 5%-damped pseudo-accelerations and the spectral maximum for T=1.0 s, for the 1971 San Fernando earthquake recorded at Pacoima Dam station

Several recent studies have focused on the relationship between the spectral maximum and the geometric mean. e.g., Beyer and Bommer (2006), Watson-Lamprey and Boore (2007), Campbell and Bozorgnia (2007), Huang et al. (2008). Beyer and Bommer (2006) used 949 records representing a wide range of magnitude, distance, fault mechanism and site class to investigate the relationship between different definitions of horizontal ground motions. They developed piecewise linear functions for the mean and the standard deviation of the natural logarithm of the maximum to geometric mean ratio as a function of period. Although they used as-recorded geometric mean, not the rotated geometric mean $GMRotI50$, the difference between the two is typically less than 3% (Boore et al., 2006). In a similar study, Watson-Lamprey and Boore (2007) used 3397 records to develop conversion factors for the average and the standard deviation of the logarithm of spectral ordinates, for different definitions of horizontal ground motions. Using least squares regression, they developed an equation for $ln\ (MD/GMRotI50)$ as a function of magnitude, rupture distance and radiation pattern. They use the parameter $\cos(2\theta_{MidFault})$ to represent the radiation pattern for strike-slip faults, where $\theta_{MidFault}$ is the angle between the station and the fault strike, measured from the midpoint of the fault. Their results are in close agreement with those reported in Beyer and Bommer (2006). Huang et al. (2008) analyzed 147 pairs of near-fault records with rupture distance smaller than 15 km, and moment

magnitude at least 6.5 to study the relationship between different definitions of horizontal spectra. They found that the median of the ratio of maximum direction to the average NGA-predicted GMRotI50 is dependent on period and the Somerville directivity parameters.

3. THE RVM MODEL

Introduced by Tipping (2000), the Relevance Vector Machine (RVM) is a supervised machine learning tool that overcomes several limitations of traditional approaches. By combining the computational advantage of the use of kernel functions with an efficient algorithm for maximizing the marginal likelihood function, the RVM constructs a nonparametric model capable of probabilistic predictions. The RVM treats the model coefficients and the standard deviation as individual random variables with different variances, and searches for the most likely model based on the observations While the RVM algorithm can be used in both regression and classification problems with some adjustments (Tipping and Faul, 2003), in this paper, we restrict our attention to the regression algorithm.

The RVM adopts a very flexible model structure where the coefficients and the number of terms in the model are determined based on the observations. Using a fixed functional form can introduce bias into the predictions, making the model susceptible to over-fitting. In ground motion modeling, the lack of full understanding of earthquake processes makes it difficult to determine the proper form of the functional input-output relationship. While the lack of a sufficient number of records has so far prevented full exploitation of nonparametric modeling, this approach is expected to gain increasing attention with the continual growth in seismic databases.

3.1. Input and Output Variables

The output variable in the RVM model is the logarithmic ratio $ln\ (MD/GMRotI50)$ for the following vibration periods: T=0.05 second, T= 0.1 second, T= 0.5 second, T=1.0 second, T=2 second, and T= 4 second.

We consider three predictors (moment magnitude, rupture distance and directivity), and define eight models each using a different variable subset. To represent rupture directivity, we use the Somerville parameter $Xcos\theta$ where X is the fraction of fault along strike that ruptures toward site, and θ is the azimuth angle between the fault plane and the ray path to the site (Somerville et al., 1997). The input variables used in the RVM models are listed in Table 3.1. A tick mark in the table indicates that the variable is included in the model. Model 1 does not use any predictors; it simply outputs the average logarithmic ratio for each period. This model serves as a baseline to measure the improvement achieved by using a predictive variable set, instead of simply using the average logarithmic ratio.

Table 3.1. Input Variables used in the Models

	Input Variables		
Model	Magnitude (M)	Distance (R_{rup})	Directivity ($Xcos\theta$)
RVM_1 (Baseline)			
RVM_2	✓		
RVM_3		✓	
RVM_4			✓
RVM_5	✓	✓	
RVM_6		✓	✓
RVM_7	✓		✓
RVM_8	✓	✓	✓

3.2. Ground Motion Records

The ground motion records used in training the RVM models come from the PEER-NGA database (http://peer.berkeley.edu/nga). We use strike-slip records with rupture distance less than 200 km. The records with missing rupture distance or Somerville parameter were excluded. The magnitude and distance distribution of the 599 records is shown in Fig. 3.1.

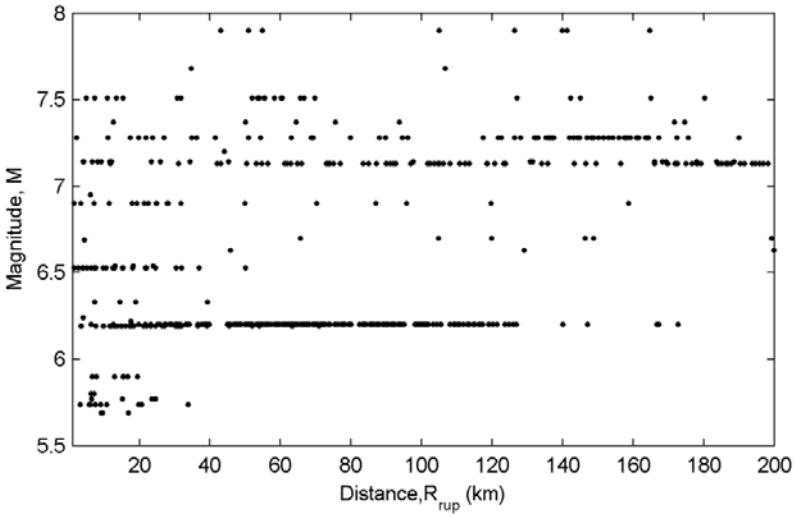

Figure 3.1. Magnitude and distance distribution of the selected records

3.3. The RVM Regression Algorithm

Given a set of N input-target pairs, where the i[th] pair is represented as $(x_i,\ t_i)$, the RVM adopts the kernel representation (Smola and Schölkopf, 2004) to express the unknown functional relationship. The regression function is written as a linear combination of basis (kernel) functions:

$$f(x) = \sum_{i=1}^{N} w_i\ K(x, x_i) + w_0 \tag{3.1}$$

where $w_i,\ i = 1 \ldots N$ are the model weights, w_0 is a constant term, and the K is a kernel function. In this study, we use the radial basis function (RBF) as the kernel. The RBF is defined as

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \qquad \gamma > 0 \tag{3.2}$$

where γ is a parameter controlling the width of the kernel. Assuming zero-mean Gaussian noise i.e., $n_i \sim \mathcal{N}(0, \sigma_n^2)$, the target value of the i[th] observation can be written as

$$t_i = f(x_i) + n_i \qquad i = 1, \ldots, N. \tag{3.3}$$

Equation (3.3) can be rewritten in matrix form as

$$t = \Phi \mathrm{w} + n, \tag{3.4}$$

where Φ is the basis matrix of size $N \times (N+1)$ where $\Phi_{ij} = K(x_i, x_{j-1})$ and $\Phi_{i1} = 1$, $w = (w_0, \ldots, w_N)^T$, and $t = (t_1, \ldots, t_N)^T$. Assuming independence of observations, the likelihood function (the probability of observing the data given the model) can be written as:

$$p(t|w,\sigma_n^2) = (2\pi\sigma_n^2)^{-\frac{N}{2}}\, e^{-\frac{1}{2\sigma_n^2}\|t-\Phi\mu\|^2} \tag{3.5}$$

where $\mu = (\mu_0, \ldots, \mu_N)^T$ is the vector containing the mean values of the weights. Assigning a Gaussian prior with mean 0 and variance $1/\alpha_i$ for the i[th] component of w, and assuming independence of the weights, the conditional probability of the weights given the inverse variances $\alpha = (\alpha_0, \ldots, \alpha_N)$ can be written as (MacKay, 1992b)

$$p(w|\alpha) = \prod_{i=0}^{N} \mathcal{N}(0,\, 1/\alpha_i). \tag{3.6}$$

The posterior distribution of weight vector can be obtained as:

$$p(w|t,\alpha,\sigma_n^2) = (2\pi)^{-\frac{N+1}{2}}\ |C|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(w-\mu)^T C^{-1}(w-\mu)} \tag{3.7}$$

where μ is the mean vector given by

$$\mu = \sigma_n^{-2} C\, \Phi^T t, \tag{3.8}$$

and C is the covariance matrix given by

$$C = [\sigma_n^{-2}\Phi^T\Phi + A\,]^{-1} \quad \text{with} \quad A = \begin{bmatrix} \alpha_0 & \ldots & \ldots & 0 \\ \vdots & \alpha_1 & & \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & \cdots & \alpha_N \end{bmatrix}. \tag{3.9}$$

The marginal likelihood is determined by integrating out the weights as follows:

$$p(t|\alpha,\sigma_n^2) = (2\pi)^{-\frac{N}{2}}\ |H|^{-\frac{1}{2}}\, e^{-\frac{1}{2}t^T H^{-1} t}, \tag{3.10}$$

where $H = \sigma_n^2 I_N + \Phi A^{-1}\Phi^T$ and I_N is the identity matrix of size N. The values of α_i and σ_n^2 maximizing Eqn. (3.10) can be found iteratively as follows (MacKay, 1992a):

$$(\alpha_i)_{new} = \frac{1-\alpha_i C_{ii}}{\mu_i^2} \tag{3.11}$$

$$(\sigma_n^2)_{new} = \frac{\|t-\Phi\mu\|^2}{N-\sum(1-\alpha_i C_{ii})}. \tag{3.12}$$

Because the nominator in Eqn. (3.11) is a positive number with a maximum value of 1, α_i tending to infinity implies that $w_i = 0$. As a consequence, the corresponding basis function (or, equivalently, the column in matrix Φ) can be removed from the model. The procedure for determining the weights and the noise variance can be summarized as follows:

1) Form the basis matrix Φ.
2) Initialize $\alpha = (\alpha_0, \ldots, \alpha_N)$ and σ_n^2.
3) Compute the covariance matrix C using Eqn. (3.9).
4) Compute the mean vector μ using Eqn. (3.8).
5) Update α and σ_n^2 using Eqn. (3.11) and Eqn. (3.12).
6) If $\alpha_i \to \infty$, set $w_i = 0$ and remove the corresponding column in Φ.
7) Go to Step 3 until convergence.
8) Set the remaining weights equal to μ.

The input vectors corresponding to the remaining nonzero weights are called the "relevance vectors". After the weights and the noise variance are determined, the predictive mean for a new input x_* can be found as follows:

1) Using the relevance vectors $(\mathrm{r}_1, \mathrm{r}_2 \ldots, \mathrm{r}_{\mathrm{Nr}})$, where form the basis matrix Φ_* using

$$\Phi_* = [1 \;\; K(\mathrm{x}^*, \mathrm{r}_1) \;\; K(\mathrm{x}^*, \mathrm{r}_2) \ldots \;\; K(\mathrm{x}^*, \mathrm{r}_{\mathrm{Nr}})]^{\mathrm{T}} \tag{3.13}$$

2) Compute the expected value of the output using

$$f(x_*) = \mu^T \Phi_* \tag{3.14}$$

Total predictive variance $(\sigma_*{}^2)$ can be determined by adding the variance due to uncertainty in the weights $\left(\Phi_*^{\mathrm{T}} \mathrm{C} \Phi_*\right)$ to the variance of the noise (σ_n^2).

4. RESULTS

4.1. Variable Selection

The eight models listed in Table 3.1. have been trained, and five-fold cross validation errors have been computed. More details on training and cross validation can be found in e.g. Tezcan and Piolatto (2012) and Tezcan and Cheng (2012). Table 4.1 lists the cross-validation root-mean-square errors (RMSE) corresponding to the six vibration periods. For each period, the lowest error has been marked using superscript **.

Table 4.1. Cross validation root-mean-square errors (RMSE)

		Cross validation RMSE					
Model	Variables	T=0.05 s.	T=0.1 s.	T=0.5 s.	T=1 s.	T=2 s.	T=4 s.
RVM_1 (Baseline)	-	0.0757	0.0797	0.1010	0.1024	0.1104	0.1219
RVM_2	M	0.0725	0.0763	0.0961	0.0996	0.1071	0.1174
RVM_3	lnR	0.0700**	0.0745**	0.0946	0.0995	0.1097	0.1200
RVM_4	$Xcos\theta$	0.0726	0.0772	0.0940	0.0989	0.1059	0.1139**
RVM_5	M, lnR	0.0723	0.0752	0.0939**	0.0989	0.1058	0.1191
RVM_6	$lnR, Xcos\theta$	0.0718	0.0763	0.0951	0.0993	0.1078	0.1176
RVM_7	$M, Xcos\theta$	0.0722	0.0756	0.0958	0.0970**	0.1046**	0.1166
RVM_8	$M, lnR, Xcos\theta$	0.0722	0.0764	0.0946	0.0980	0.1056	0.1167

The RMSE values in the table show that while the baseline model (RVM_1), which does not use any predictors, has the highest cross validation error for each period, the improvement achieved by introducing the predictors is not great. The listed RMSE values suggest that different sets of variables may govern the behavior of the maximum-to-geometric mean ratio for different periods. However, there is no indication of a strong relationship between $ln\,(MD/GMRotI50)$ and any of the variable sets listed.

4.2. Comparisons with existing models

In this section, we evaluate RVM_5 model for two magnitudes (M=6 and M=7), and two distances (R=10 km and R=100 km), and compare its predictions to two existing models: the model by Beyer and Bommer (2006), and the model by Watson-Lamprey and Boore (2007). These models will be referred to as BB06 and WLB07, respectively. The reason for selecting the RVM_5 out of the eight RVM models is its similarity to the WLB07 model in terms of the predictors used. The WLB07 model uses M, lnR, and an optional radiation pattern term, while the BB06 model depends only on period. We evaluated the predictions of the WLB07 model without the radiation pattern term, making its predictor set identical to that of the RVM_5 model.

The median (50^{th} percentile) and the 84^{th} percentile values of the $MD/GMRotI50$ ratios from the three models for M=6 and R=10 km are shown in Fig 4.1. The M=6 and R=100 km case is shown in Fig.4.2. The predictions for M=7 are shown in Fig.4.3. and Fig.4.4. for R= 10 km and R=100 km,

respectively. The RVM_5 predictions have not been smoothed.

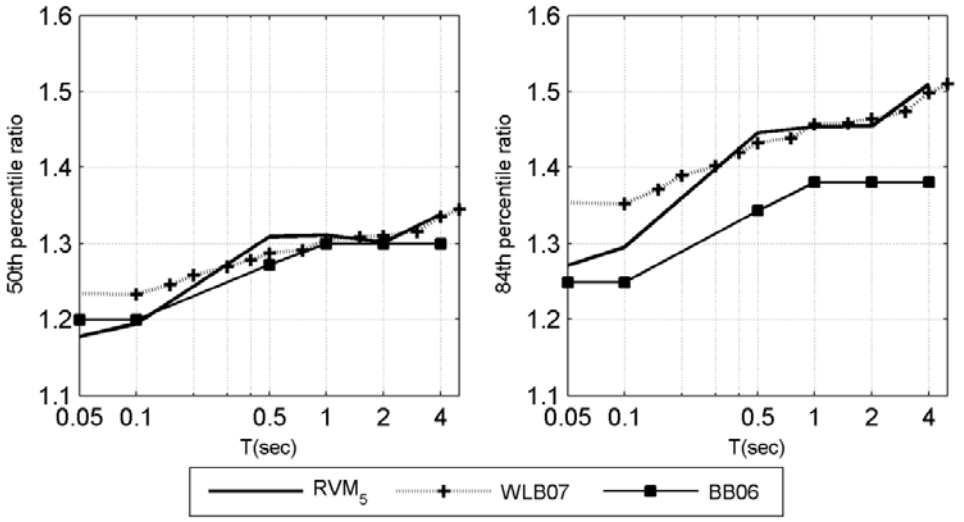

Figure 4.1. Predictions for the median and 84th percentile values of $MD/GMRotI50$ for M=6 and R=10 km

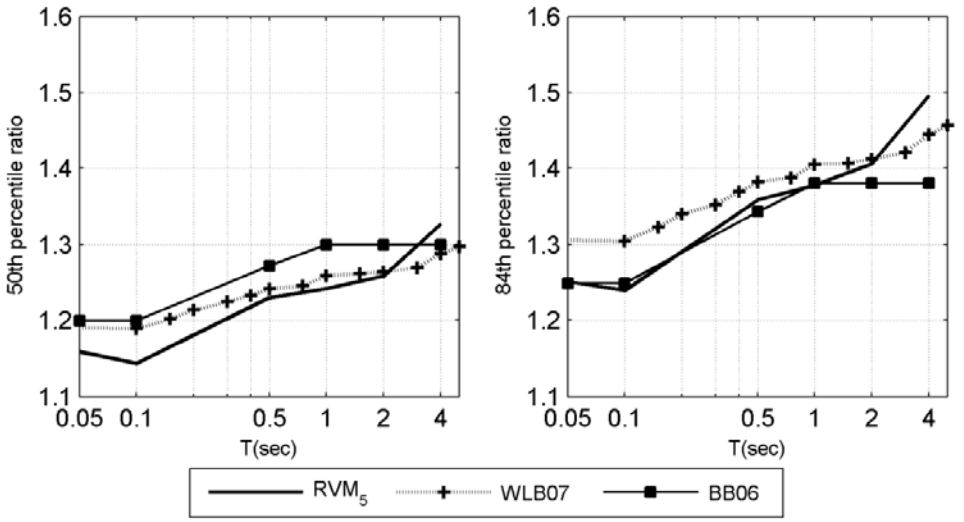

Figure 4.2. Predictions for the median and 84th percentile values of $MD/GMRotI50$ for M=6 and R=100 km

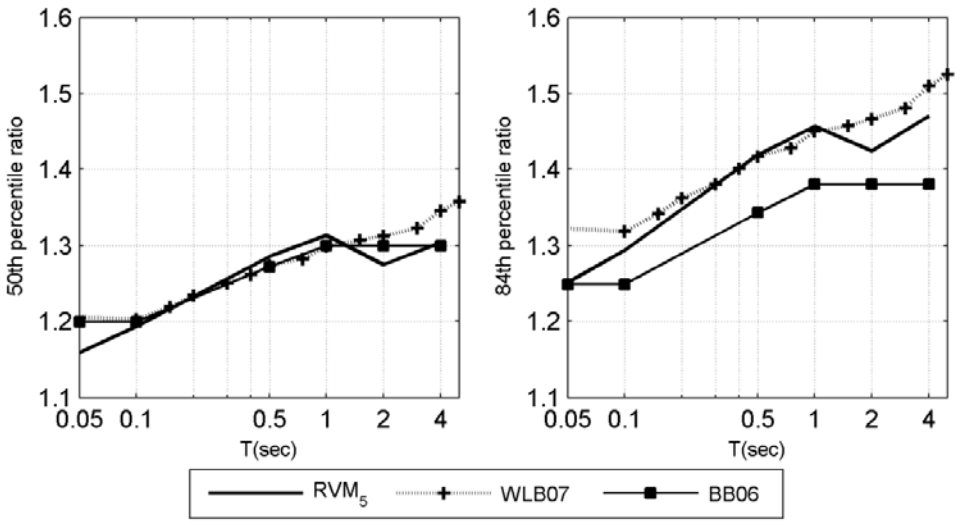

Figure 4.3. Predictions for the median and 84th percentile values of $MD/GMRotI50$ for M=7 and R=10 km

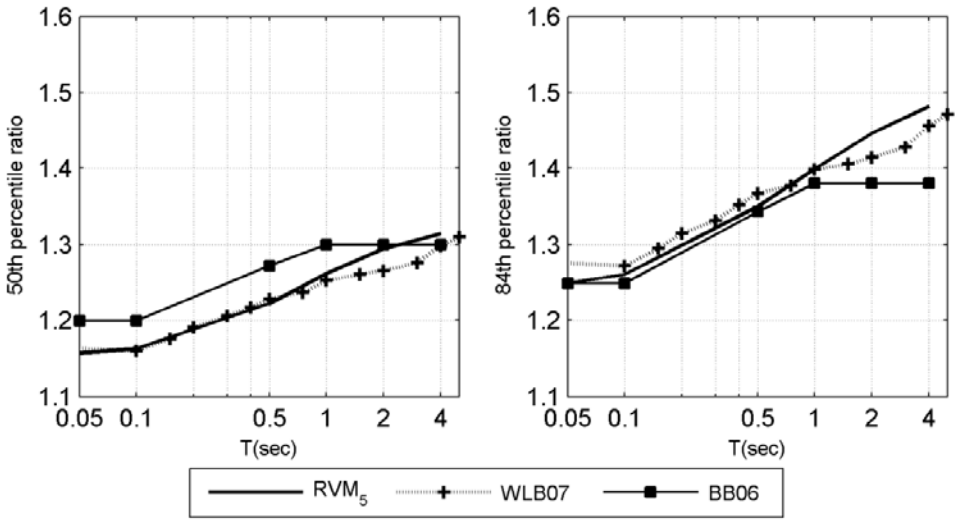

Figure 4.4. Predictions for the median and 84th percentile values of $MD/GMRotI50$ for M=7 and R=100 km

The logarithmic standard deviations from the three models are listed in Table 4.2. The values from the RVM_5 model are approximately double the BB06 values, and very close to the WLB07 values.

Table 4.2. Comparison of standard deviation of $ln\ (MD/GMRotI50)$

Period (sec)	Standard deviation		
	RVM_5	WLB07	BB06
0.05	0.076	0.093	0.040
0.1	0.081	0.092	0.040
0.5	0.099	0.107	0.054
1.0	0.104	0.110	0.060
2.0	0.111	0.111	0.060
4.0	0.120	0.115	0.060

5. CONCLUSION

This paper uses Relevance Vector Machine (RVM) regression, a probabilistic sparse kernel model, to investigate the dependence of the maximum direction to geometric mean (MD/GMRotI50) ratio on magnitude, rupture distance, and directivity. Not requiring a fixed functional form makes RVM suitable for determining the influence of different subsets of predictors on the dependent variable being modeled. In addition, because RVM is a probabilistic model, the standard deviations are automatically computed using Bayesian inference, instead of being estimated from the fitting errors.

Eight RVM models, each using a different subset of the three predictors (magnitude, rupture distance, and directivity) as the input and the logarithm of the (MD/GMRotI50) ratio as the output, have been trained using 599 pairs of horizontal acceleration records from strike-slip earthquakes, taken from the PEER database. Six vibration periods ranging from 0.05 second to 4 seconds were considered. The performance of each model was measured using five-fold cross validation errors.

The results suggest that different subsets of the three predictors considered (magnitude, distance and directivity) may govern the maximum-to-geometric mean ratio for different periods. However, there is no indication of a strong relationship between $ln\,(MD/GMRotI50)$ and any of the variable subsets.

ACKNOWLEDGEMENT

This study was supported in part by a grant (CMMI-1100735) from the National Science Foundation.

REFERENCES

Beyer, K. & Bommer, J. J. (2006). Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion. *Bulletin of the Seismological Society of America.* **96:4A,** 1512-1522.

Boore, D. M. (2010). Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of motion. *Bulletin of the Seismological Society of America.* **100:4,** 1830-1835.

Boore, D. M., Watson-Lamprey, J. & Abrahamson, N. A. (2006). Orientation-independent measures of ground motion. *Bulletin of the Seismological Society of America.* **96:4A,** 1502-1511.

Campbell, K. W. & Bozorgnia, Y. (2007). Campbell-Bozorgnia NGA ground motion relations for the geometric mean horizontal component of peak and spectral ground motion parameters. *PEER Report 2007/02.* 238.

Huang, Y. N., Whittaker, A. S. & Luco, N. (2008). Maximum spectral demands in the near-fault region. *Earthquake Spectra.* **24:1,** 319-341.

Mackay, D. J. C. (1992a). Bayesian interpolation. *Neural computation.* **4:3,** 415-447.

Mackay, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural computation.* **4:5,** 720-736.

Smola, A. J. & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing.* **14:3,** 199-222.

Somerville, P. G., Smith, N. F., Graves, R. W. & Abrahamson, N. A. (1997). Modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity. *Seismological Research Letters.* **68:1,** 199-222.

Tezcan, J. & Cheng, Q. (2012). A nonparametric characterization of vertical ground motion effects. *Earthquake Engineering & Structural Dynamics.* **41:3,** 515-530.

Tezcan, J. & Piolatto, A. (2012). A probabilistic nonparametric model for the vertical-to-horizontal spectral ratio. *Journal of Earthquake Engineering.* **16:1,** 142-157.

Tipping, M. E. (2000). The Relevance Vector Machine. *Proceedings of the First 12 Conferences in Advances in Neural Information Processing Systems*. 652–658

Tipping, M. E. & Faul, A. (2003). Fast marginal likelihood maximisation for sparse Bayesian models. *Ninth International Workshop on Artificial Intelligence and Statistics*.

Watson-Lamprey, J. A. & Boore, D. M. (2007). Beyond SaGMRotI: Conversion to SaArb, SaSN, and SaMaxRot. *Bulletin of the Seismological Society of America.* **97:5,** 1511-1524.