Bindi, D. (2017): The Predictive Power of Ground-Motion Prediction Equations. - *Bulletin of the Seismological Society of America*, *107*, 2, pp. 1005—1011.

DOI: http://doi.org/10.1785/0120160224

THE SEISMOLOGICAL SOCIETY OF AMERICA
400 Evelyn Ave., Suite 201
Albany, CA 94706-1375
(510) 525-5474; FAX (510) 525-7204
www.seismosoc.org

Bulletin of the Seismological Society of America, Vol. 107, No. 2, pp. 1005–1011, April 2017, doi: 10.1785/0120160224

Short Note

The Predictive Power of Ground-Motion Prediction Equations

by D. Bindi

Abstract Although model calibration should be always performed alongside model validation, this is rarely the case when deriving new ground-motion prediction equations (GMPEs). Explanatory modeling (Shmueli, 2010) is often preferred to data-driven predictive approaches, and the analysis of the residual distribution is generally used to support the model qualification. Following previous studies (Kuehn *et al.*, 2009; Scherbaum *et al.*, 2009), this work aims to again stress the importance of validation for assessing the predictive power of GMPEs and for avoiding, or at least limiting, data overfitting. Considering the strong-motion data and models recently analyzed by Roselli *et al.* (2016), I exemplify the application of standard validation approaches based on predictive metrics (Akaike and Bayesian information criteria), resample techniques (cross validation and bootstrap), and validation against new data. The results confirm that GMPEs overfitting the calibration data have a limited predictive power, whereas, regarding validation against new data, the selection of the validation data set should take into account possible regional effects in the ground motion.

Introduction

The selection of alternative ground-motion predictions (GMPEs) is one of the fundamental steps in probabilistic seismic-hazard assessment (Kulkarni *et al.*, 1984; McGuire, 2004). Starting from a set of viable candidates (e.g., Bommer *et al.*, 2010), the selection of models to populate the branches of the logic tree (e.g., Bommer and Scherbaum, 2008; Bommer, 2012) is often supported by measuring the goodness-of-fit with respect to selected data sets (e.g., Delavaud *et al.*, 2012). In particular, Scherbaum *et al.* (2009) introduced an information-theoretic perspective for ranking a set of GMPEs with respect to a set of observations. To perform a relative ranking, Scherbaum *et al.* (2009) proposed using the difference between the base-2 average log likelihood (logLH) computed for the selected models considering a validation data set that is independent of the calibration one. Recently, Roselli *et al.* (2016) proposed a weighting scheme based on the Bayesian information criterion (BIC, Schwarz, 1978) to build an ensemble model from a set of GMPEs. The BIC as well as other criteria like the Akaike information criterion (AIC, Akaike, 1973) introduces a penalty term to correct for the bias in the maximum-likelihood estimators (MLEs) when the model performances are evaluated over the calibration data. This aspect is of particular relevance for GMPEs, which are generally based on complex functional forms with several parameters. As observed by Scherbaum *et al.* (2009), despite the strong overfitting to which the models are exposed, their performance in predicting new data is seldom addressed (e.g., Kuehn *et al.*, 2009; Kaklamanos and Baise, 2011; Mak *et al.*, 2015), and rarely during the calibration phase. Indeed, a GMPE performance is generally evaluated through residual analysis performed with respect to the same calibration data, and the significance of the model coefficients as well as their trade-offs are often not discussed. The goal of the present work is neither to propose new validation tools for assessing the predictive power of GMPEs nor to discuss the best practice for their calibration. Indeed, the goal of this work is to promote the application of validation approaches to assess the predictive power of GMPEs, because such an assessment is not yet a standard practice during the calibration phase. Three standard validation approaches are exemplified using the Italian strong-motion data analyzed by Roselli *et al.* (2016). First, the performance of MLEs for different models constructed from the functional forms of two GMPEs evaluated in Roselli *et al.* (2016) are compared in terms of AIC and BIC values. Second, their predictive power is assessed by applying different resampling techniques based on either cross validation or bootstrap strategies. Finally, two different data sets, one including recent Italian earthquakes and the other extracted from the Next Generation Attenuation-West2 (NGA-West2) flatfile (Ancheta *et al.*, 2014), are used to test the predictive power of the considered models against new data.

Methods

Explanatory and predictive modeling can be defined in different ways (e.g., see Breiman, 2001; Shmueli, 2010). Following Shmueli (2010), an explanatory model is driven by the physical interpretability (theory) of the causal connection (model) between a set of available observations, whereas data generalization is one of the main requirements for predictive modeling, in which the ability of the model to predict new data is crucial. For example, in the seismological context, explanatory and predictive modeling can be exemplified with the retrospective (i.e., testing a set of already existing hypotheses) and prospective (i.e., predicting new observations) testing approaches, respectively (e.g., Jordan, 2014).

Explanatory modeling is generally applied in the development of new GMPEs. Most of the recent GMPEs are based on physical interpretations of source, propagation, and site effects, and numerical simulations often aid both the model calibration (e.g., in constraining the hanging-wall–footwall effects and the nonlinearity of the soil response) and the selection of the functional form. The calibration of the GMPE often involves regression analysis where the residuals between observations and predictions are minimized (see Douglas *et al.*, 2014; Gregor *et al.*, 2014, for a summary of recent developed GMPEs). When complex functional forms are implemented, the obtained models are often overfitting; that is, the models fit details in the data, which are specific features of the analyzed sample, not of the population generating the sample. Because the overfitting limits the predictive power (e.g., Hagerty and Srinivasan, 1991; Forster and Sober, 1994; Shmueli, 2010), some predictive metrics that introduce a penalty term over the model complexities to handle the bias-variance trade-off must be introduced to evaluate the model performance (e.g., Foulser-Pigott and Goda, 2015). Examples are the BIC (Schwartz, 1978) and the AIC (Akaike, 1973).

Given an n-dimensional data set Y_n, AIC and BIC are defined as follows:

$$\text{BIC} = -2\log\text{LH}(\hat{\theta}_n^k|Y_n) + D_k\log(n) \tag{1}$$

$$\text{AIC} = -2\log\text{LH}(\hat{\theta}_n^k|Y_n) + D_k \tag{2}$$

(e.g., Forster and Sober, 1994; Cavanaugh and Neats, 2012), in which $\hat{\theta}_n^k$ denotes the estimator of the parameters θ^k (i.e., the coefficients of the GMPE), obtained by maximizing the likelihood $\text{LH}(\theta^k|Y_n)$ over the k-dimensional parameter space $\Theta^k \subset \Re^k$. In equations (1) and (2), D_k is the number of independent parameters of the model estimated from data. By considering the AIC and BIC selection criteria, models whose complexity is not justified in terms of goodness-of-fit improvements are penalized (and considering large samples, the penalization is stronger considering BIC). A detailed discussion about the application of AIC and BIC for model selection is given by Burnham and Anderson (2004).

Beside the application of the AIC and BIC selection criteria, the predictive power of a group of models can be compared by comparing their accuracy in predicting data not used for the model calibration. When data scarcity does not permit the splitting of the observations into calibration and validation data sets, resampling techniques can be applied. In the following, I consider two different versions of the cross-validation technique (Stone, 1974), namely leave-one-out and k-fold cross validation. In the leave-one-out strategy, the model is calibrated using all data but one, which is then used for validation. The procedure is repeated until all data have been left out once. Several cost functions can be considered to assess the prediction power. In this study, I compute the average squared error given by

$$E_{\text{looCV}} = \frac{1}{N}\sum_{i=1}^{N}[Y_i - \bar{Y}_i(\hat{\theta}_i^k; Y_{-i})]^2, \tag{3}$$

in which the predictions $\bar{Y}_i$ for data i are computed from the MLE $\hat{\theta}_i^k$ evaluated removing sample i from the original data set (i.e., Y_{-i}).

In the k-fold approach, the original data set is first split into k subsets, or folders; then, samples from $k-1$ subsets are jointly used to calibrate the model, whereas samples in the excluded folder are used for validation. The procedure is repeated until all the k subsets have been considered for validation once. The average squared error is computed as follows:

$$E_{k\text{–foldCV}} = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{N_j}\sum_{i=1}^{N_j}[Y_i - \bar{Y}_i(\hat{\theta}_j^k; Y_{-j})]^2, \tag{4}$$

in which N_j, with $j = 1, \ldots, K$, is the number of samples in folder j and the MLE $\hat{\theta}_j^k$ is evaluated excluding the samples belonging to folder j. The different folders are generally selected with the same number of elements. To limit the correlation among recordings in different folders, in this study the folding procedure is applied at the earthquake level (that is, each folder contains almost the same number of earthquakes, but a different number of recordings). In particular, I consider $K = 8$ folders.

The so-called .632+ bootstrap method (Efron and Tibshirani, 1997) is also applied to test the predictive power of the models. As any standard bootstrap technique, the .632+ bootstrap is based on generating a given number of replications of the original data set by randomly selecting samples with repetition. In computing the error for any repetition, only those records not selected in that specific repetition are considered. The error function is computed as follows:

$$E_{\text{training}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}_i[\hat{\theta}_n^k; Y_n])^2 \tag{5}$$

$$E_{\text{boot}} = \frac{1}{\text{NR}}\sum_{j=1}^{\text{NR}}\frac{1}{|C^{-j}|}\sum_{b\in C^{-j}}(Y_b - \bar{Y}_b[\hat{\theta}_j^k; Y_{-j}])^2 \tag{6}$$

$$E_{0.632+} = 0.368E_{\text{training}} + 0.632E_{\text{boot}}, \tag{7}$$

in which NR is the number of bootstrap replications (100 in this study), C^{-j} are the samples of the original data set not

Table 1

Main Features of the Considered Data Sets

Data Set	$[N_{\mathrm{rec}}, N_{\mathrm{eq}}, N_{\mathrm{st}}]$*	**M**†	R_{JB} (km)‡	SOF [N,R,S,U]§	EC8 [A,B,C,D,E]‖
ITACA	[957,152,311]	4.1–6.9	0–300	[684,151,87,35]	[401,271,219,22,44]
ITACA-SEL	[479,41,172]	5–6.9	0–150	[399,62,18,0]	[215,145,111,8,0]
IT-VAL	[1335,67,367]	4.1–6.1	0–150	[564,0,259,0]	[373,671,260,0,31]
NGA-SEL	[1362,58,825]	4.5–7	0–150	[41,773,548,0]	[79,707,529,47,0]

ITACA, ITalian ACcelerometric Archive; ITACA-SEL, a subset of ITACA; IT-VAL, Italian data independent from ITACA.

*[Number of records, number of earthquakes, number of stations].

†Magnitude range.

‡Joyner–Boore distance range.

§SOF, style of faulting [normal, reverse, strike slip, unknown].

‖Eurocode 8 (2004) site classes from A to E.

included in the bootstrap replication j, and $|C^{-j}|$ is the number of such samples.

Models and Data

The computation of BIC (and AIC) for models which are not MLEs has little statistical interest. Therefore, I consider the ITalian ACcelerometric Archive (ITACA) data (Luzi *et al.*, 2008) also analyzed by Roselli *et al.* (2016) to derive a set of MLEs based on the functional forms of two GMPEs analyzed by Roselli *et al.* (2016). The functional forms are similar to those of Cauzzi and Faccioli (2008) and Bindi *et al.* (2011), in the following referred to as CF and IT, respectively. The functional form of model CF is

$$\ln(Y_{\mathrm{CF}}) = e_1 + e_2(M - M_{\mathrm{ref}}) + e_3 \ln(R_{\mathrm{hypo}}) + \delta\mathrm{EC8}_s + \delta\mathrm{SOF}_e + \epsilon, \tag{8}$$

whereas the IT functional form is

$$\ln(Y_{\mathrm{IT}}) = b_1 + [b_2 + b_3(M - M_{\mathrm{ref}})] \ln\left(\sqrt{R_{\mathrm{JB}}^2 + b_4^2}/R_{\mathrm{ref}}\right) + b_5\left(\sqrt{R_{\mathrm{JB}}^2 + b_4^2} - R_{\mathrm{ref}}\right) + F(M) + \delta\mathrm{EC8}_s + \delta\mathrm{SOF}_e + \epsilon, \tag{9}$$

in which the magnitude term $F(M)$ is defined as

$$F(M) = \begin{cases} b_6(M - M_h) + b_7(M - M_h)^2 & \text{for } M \le M_h \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

In equations (8) and (9), $\delta\mathrm{EC8}_s$ is a random effects term for each site class of Eurocode 8 (2004); that is, $\mathrm{EC8} = [A, B, C, D, E]$; $\delta\mathrm{SOF}_e$ is a random effects term depending on the style of faulting (SOF), considering the four classes: normal (N), reverse (R), strike slip (S), and unknown (U); and ϵ is the residual distribution. The site effects and the SOF are introduced in equations (8) and (9) as random effects on the offset coefficients e_1 or b_1, respectively. For the prediction of new values, $\delta\mathrm{EC8}_s$ and $\delta\mathrm{SOF}_e$ are used as scalar adjustment to the median; that is, applied along with the fixed-effects components of the model, their variances removed from the aleatory variability. Because the main goal of this work is to discuss the model validation, the decomposition of the residuals into the between- and within-event components is not required. Model IT considers the Joyner–Boore distance R_{JB}, whereas CF implements the hypocentral distance R_{hypo}. Following Bindi *et al.* (2011), the reference distance R_{ref}, the reference magnitude M_{ref}, and the hinge magnitude M_h are set equal to 1, 5, and 6.75 km, respectively. The regressions are performed for the peak ground acceleration (PGA).

To test the predictive power of the GMPEs as a function of the model complexity, I consider three further versions of the IT functional form, where either the apparent anelastic attenuation (b_5), the quadratic magnitude term (b_7), or both, are dropped:

$$\text{Model A}: b_7 = b_5 = 0 \quad \text{in equation(9)} \tag{11}$$

$$\text{Model B}: b_5 = 0 \quad \text{in equation(9)} \tag{12}$$

$$\text{Model C}: b_7 = 0 \quad \text{in equation(9)}. \tag{13}$$

The MLEs for the models described by equations (8)–(13) are computed using ITACA as the calibration data set. The models are also calibrated considering a selection of data corresponding to the magnitude and distance ranges analyzed in Cauzzi and Faccioli (2008; see Table 1), referred to as ITACA-SEL.

To validate the models against new data, two different data sets are considered. The first (referred to as IT-VAL) includes recordings relevant to earthquakes that occurred in Italy after 2009 (i.e., earthquakes not contained in ITACA); the second one (referred to as NGA-SEL) includes a selection of records extracted from the NGA-West2 flatfile (Ancheta *et al.*, 2014). In particular, I removed the Italian records from the Campbell and Bozorgnia (2014) selection and, to validate the models over the calibration ranges, I considered only magnitudes in the 4.5–7 range and distances less than 150 km. Discussions about the possibility to extrapolate the

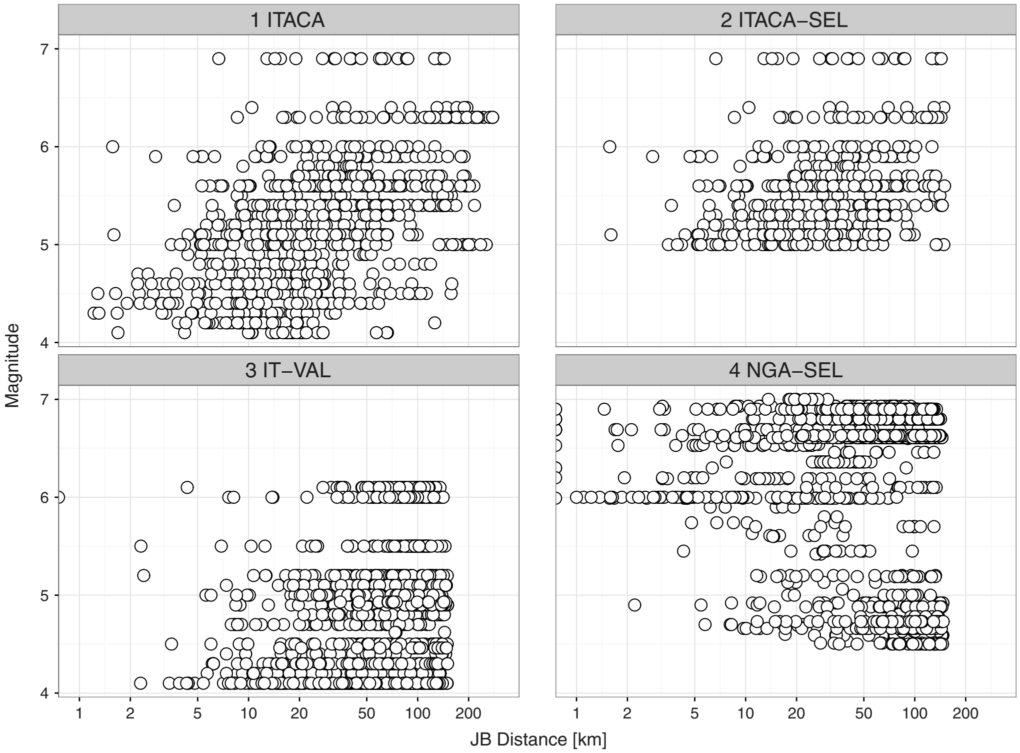

Figure 1. Magnitude–distance scatter plot for four considered data sets: (1) ITACA, ITalian ACcelerometric Archive data set used for calibration; (2) ITACA-SEL, a subset of ITACA, also used for calibration; (3) IT-VAL, Italian data independent from ITACA, used for validation against new data; (4) NGA-SEL, records extracted from the Next Generation Attenuation-West2 (NGA-West2) data set, used for validation against new data. JB, Joyner–Boore.

models outside the calibration ranges are beyond the goal of this work. The magnitude–distance scatter plots for the four data sets are shown in Figure 1, whereas Table 1 lists their main characteristics.

Results

Predictive Metric

Table 2 summarizes the performance of the considered models in terms of logLH, AIC, and BIC; the standard deviation σ of the residuals is also listed. Although, as expected, the most complex model (i.e., IT) shows the lowest logLH, model B shows the best performance in terms of both AIC and BIC. Therefore, the introduction of the apparent anelastic term (coefficient b_5 in equation 9) is not justified in terms of the balance between goodness-of-fit and model complexity. Indeed, a significance test performed over b_5 confirms that it is not significantly different from zero at any confidence level (because its t-value is less than 1). To avoid overfitting, the AIC values suggest that one should remove this parameter. It is worth noting that the contribution of the anelastic term to the seismic attenuation increases with distance. When larger distances are analyzed, the contribution of the anelastic term could become significant, as expected from theory. Considering that the average attenuation along a ray path depends on the traveled distance, the application of the anelastic attenuation term over distances larger than the calibration one should be avoided unless constrained by further information. Model B shows the lowest AIC (and BIC) value also when the ITACA-SEL data set is considered. Regarding the CF model, the larger BIC, AIC, and σ values suggest that this model is underfitting the data.

The test case analyzed in this work has been inspired by Roselli *et al.* (2016). In their work, the authors proposed the usage of BIC for ranking the predictive performances of a set of candidate GMPEs similar to those calibrated in the present study. To compare the results of this study with Roselli *et al.* (2016), the logLH is also computed for the B-2011 (Bindi *et al.*, 2011) and CF-2008 (Cauzzi and Faccioli, 2008) models considered by the authors. Because in this case θ^k is not an MLE, the values of AIC and BIC have no particular statistical meaning. Therefore, Table 3 shows the logLH values, the mean residual (bias) and the standard deviation of the residuals. Because B-2011 was calibrated over a data set very similar to ITACA (only small differences due to data selection could exist), the bias is almost zero and the logLH is smaller than for CF-2008, which was calibrated considering mainly Japanese data. Indeed, CF-2008 shows a positive bias, suggesting possible regional differences between the high-frequency ground-motion scaling in Japan and Italy. Even when the penalty term applied by Roselli *et al.* (2016); (i.e., 14log[957], in which 14 is the number of degrees of

Table 2

Goodness-of-Fit and Predictive Metric Computed for Different Models and Two Different Data Sets

Data Set	Model	D^k	logLH	AIC	BIC	σ*
ITACA	CF	6	−1168	2348	2377	0.8035
($N = 957$)	A	8	−1123	2261	2300	0.7732
	B	9	−1117	2251	2295	0.7686
	C	9	−1122	2263	2306	0.7705
	IT	10	−1116	2253	2301	0.7685
ITACA-SEL	CF	6	−534	1080	1105	0.7259
($N = 479$)	A	8	−513	1042	1075	0.6967
	B	9	−493	1003	1041	0.6673
	C	9	−513	1044	1081	0.6967
	IT	10	−493	1005	1047	0.6672

logLH, log likelihood; AIC, Akaike information criterion; BIC, Bayesian information criterion; CF, Cauzzi and Faccioli (2008); IT, Bindi *et al.* (2011).

*The standard deviation of the residual distribution.

Table 3

Log Likelihood (logLH), Mean (Bias), and Standard Deviation (σ) of the Residual Distribution for Two Models

Data Set	Model	logLH	Bias	σ
ITACA ($n = 957$)	CF-2008	−1251	0.15	0.87
	B-2011	−1108	0.01	0.77
ITACA-SEL ($n = 479$)	CF-2008	−560	0.23	0.75
	B-2011	−505	0.01	0.68

CF-2008, Cauzzi and Faccioli (2008); B-2011, Bindi *et al.* (2011).

Table 4

Mean Squared Error for Different Resampling Techniques

Model	CV-LOO*	CV-k-fold†	Bootstrap‡
CF	0.66	0.69	0.64
A	0.61	0.62	0.61
B	0.60	0.62	0.61
C	0.62	0.64	0.65
IT	0.62	0.64	0.66

*Leave-one-out cross validation (equation 3).

†k-fold cross validation with $k = 8$ (equation 4).

‡.632+ bootstrap (equation 7).

Table 5

Results of Validation against New Data

Data Set	Model	logLH	Bias	σ
IT-VAL ($n = 1335$)	CF	−1907	−0.32	0.84
	A	−1778	−0.11	0.84
	B	−1820	0	0.86
	C	−1809	−0.22	0.83
	IT	−1820	−0.13	0.85
NGA-SEL ($n = 1362$)	CF	−1690	0.34	0.77
	A	−1709	0.33	0.77
	B	−1844	0.52	0.74
	C	−1605	0.22	0.76
	IT	−1676	0.39	0.73

freedom considered by the authors for B-2011) is added to the deviance (defined as −2logLH) of the ITA10 model, the obtained value is still less than the deviance computed for CF-2008 without the penalty term. The same conclusion is reached when ITACA-SEL is considered. Therefore, different from what was observed by Roselli *et al.* (2016), B-2011 describes the data set to which the model was fitted better than CF-2008 in terms of both goodness-of-fit and penalized deviance.

Validation through Resampling Techniques

The results of the predictive power evaluation through resampling techniques are given in Table 4. Although the different models produce similar mean squared errors, models B (equation 12) and A (equation 11) are preferred by all techniques. In particular, model B is selected by the leave-one-out cross validation, confirming the asymptotic behavior of AIC to converge to the leave-one-out solution (Stone, 1977).

Validation against New Data

Table 5 shows the performance of the models against new data, considering either Italian data (IT-VAL data set) or recordings from regions outside of Italy (NGA-SEL data set). Regarding the Italian data, model A shows the lowest logLH, whereas model B is the only one with a null bias. The large bias and logLH obtained for model CF (equation 8) confirm that this model is probably underfitting. The recordings in NGA-Sel are relevant to earthquakes that occurred outside of Italy. The results in Table 5 show a different picture about the performance of the various models: in this case, model B shows the worst performance in terms of logLH and bias, whereas model C (equation 13) shows the lowest logLH and bias. These results suggest that the samples in NGA-SEL cannot be considered as being generated by the same (unknown) physical process that generated the Italian data used for calibration. In other words, these results suggest that regional effects can have an impact on ground motion variability, as has been recognized in previous studies, where regional attributes have been included in the GMPE (e.g., Campbell and Bozorgnia, 2014; Kotha *et al.*, 2016).

Conclusions

The calibration of GMPEs usually follows an explanatory modeling approach where the causal connection between observables reflects theoretical expectations. Such GMPEs are based on complex combinations of explanatory variables such as magnitude, fault-to-site distance, proxies for site amplification, and for finite-fault effects, etc. Moreover, results from numerical simulations are often exploited for constraining nonlinear site effects, for describing hanging-wall–footwall source effects, and for including basin-related site amplifications. The overall quality of the GMPE is gen-

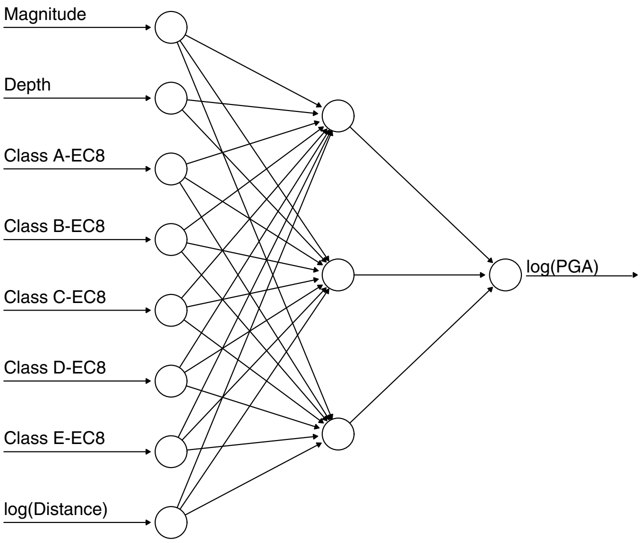

Figure 2. An example of a neural network calibrated using the ITACA data set. EC8, Eurocode 8 (2004).

erally assessed by evaluating its goodness of fit with respect to the calibration data set and analyzing the statistical features of the residual distribution (e.g., bias, variance, dependences on the explanatory variables, etc.). Because the measurable data are only providing a limited representation of the underlying (unknown) physical process (nature), the GMPEs obtained following the explanatory approach are exposed to overfitting which, in turn, limits their predictive power. Indeed, it has been often observed in the literature that the increase of strong-motion data availability over the years has corresponded to an increase in GMPE complexity, without a significant reduction in the overall aleatory variability σ (e.g., Bozorgnia *et al.*, 2015; Malhotra, 2015).

To control the overfitting, the model calibration should be performed along with model validation. Considering as an example the Italian strong-motion data, in this study I compared the predictive power of models with different levels of complexity, starting from two GMPEs analyzed by Roselli *et al.* (2016). I applied different validation approaches such as the usage of predictive metrics introducing a penalty term for the logLH depending on the model complexity (AIC and BIC), resampling techniques that split the available data into learning and testing subsets (cross validation and bootstrap), and validation against new data. In terms of the balance between goodness of fit and parsimony of the model, the predictive metrics and the resampling techniques suggested the use of a model with an intermediate level of complexity. The introduction of the apparent anelastic attenuation term does not improve the goodness of fit enough to justify the increase in model complexity. Moreover, the trade-off affecting the parameters controlling the attenuation with distance allows one to remove a parameter with little impact on the overall predictive power. On the other hand, whereas the validation against new Italian data confirms the outcomes of the resampling analysis, the validation against new data generated in regions different from Italy indicated the need to consider regional effects.

As a final remark, it is worth remembering that other modeling approaches can be applied to derive GMPEs. In particular, in many fields of science, data-driven approaches have shown their suitability for generating models with high predictive powers (e.g., Breiman, 2001). In the context of GMPE calibration, examples of data-driven approaches include the application of artificial neural networks (e.g., Derras *et al.*, 2012, 2014). To exemplify their application to the case study at hand, Figure 2 shows a simple neural network that includes one hidden layer with three neurons. The input neurons are the magnitude, the logarithm of the distance, the hypocentral depth, and the EC8 site classes. Despite its relatively simple structure, the calibrated network shows good performance both in terms of goodness of fit and predictive power. For example, the logLH and σ values for the calibration data set (ITACA) are -1085 and 0.75, respectively, whereas the logLH, bias, and σ of the predictions relevant to IT-VAL are -1650, 0.09, and 0.82, respectively. When compared with results shown in Tables 2 and 5, these values confirm the potentiality of such an approach.

Furthermore, the continuous increase of available strong-motion data, and the possibility of easily merging them with weak motion data recorded by seismological networks and retrieved through standardized web applications (e.g., Incorporated Research Institutions for Seismology and European Integrated Data Archive, see Data and Resources), suggests that data-driven and parametric modeling approaches could be integrated in an attempt to minimize the combination of bias and estimation variance, accepting the occasional sacrifice of theoretical accuracy for improved empirical precision (Shmueli, 2010).

Data and Resources

The R software (R Development Core Team, 2008; http://www.R-project.org, last accessed May 2016) has been used in this study to perform the regressions. In particular, the packages lme4 (Bates *et al.*, 2015; https://cran.r-project.org/web/packages/lme4/news.html, last accessed May 2016) and neuralnet (https://cran.r-project.org/web/packages/neuralnet/index.html, last accessed May 2016) have been used for mixed-effect regressions and calibration of the neural network, respectively. The Italian strong-motion data are available at http://itaca.mi.ingv.it (last accessed June 2016). The Next Generation Attenuation-West2 (NGA-West2) flatfile is available at http://peer.berkeley.edu/ngawest2/databases/ (last accessed June 2016). Incorporated Research Institutions for Seismology (IRIS) portal can be accessed at http://www.iris.edu; European Integrated Data Archive (EIDA) portal can be accessed at http://www.orfeus-eu.org/eida/ (both last accessed June 2016).

Acknowledgments

The author thanks F. Cotton, S.-R. Kotha, A. Oth, S. Parolai, M. Pittore, M. Haas, A. Strader, and E. Rivalta for triggering fruitful discussions. The author also thanks K. Fleming who proofread the manuscript. The support of the editorial board and an anonymous reviewer is also acknowledged.

References

Akaike, H (1973). Information theory and an extension of the maximum likelihood principle, in *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Editors), Akademiai Kiado, Budapest, Hungary, 267–281.

Ancheta, T. D., R. B. Darragh, J. P. Stewart, E. Seyhan, W. J. Silva, B. S.-J. Chiou, K. E. Wooddell, R. W. Graves, A. R. Kottke, D. M. Boore, *et al.* (2014). NGA-West2 database, *Earthq. Spectra* **30,** 989–1005.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4, *J. Stat. Software* **67,** no. 1, 1–48.

Bindi, D., F. Pacor, L. Luzi, R. Puglia, M. Massa, G. Ameri, and R. Paolucci (2011). Ground motion prediction equations derived from Italian strong motion database, *Bull. Earthq. Eng.* **9,** 1899–1920.

Bommer, J. J. (2012). Challenges of building logic trees for probabilistic seismic hazard analysis, *Earthq. Spectra* **28,** 1723–1735.

Bommer, J. J., and F. Scherbaum (2008). The use and misuse of logic-trees in PSHA, *Earthq. Spectra* **24,** 997–1009.

Bommer, J. J., J. Douglas, F. Scherbaum, F. Cotton, H. Bungum, and D. Faeh (2010). On the selection of ground-motion prediction equations for seismic hazard analysis, *Seismol. Res. Lett.* **81,** 794–801.

Bozorgnia, Y., J. P. Stewart, T. Kishida, D. M. Boore, K. W. Campbell, G. M. Atkinson, B. S.-J. Chiou, I. M. Idriss, W. J. Silva, and R. R. Young (2015). Response to discussion by P. Malhotra on NGA-West2 research project, *Earthq. Spectra* **31,** 1879–1884.

Breiman, L. (2001). Statistical modeling: The two cultures, *Stat. Sci.* **26,** 199–231.

Burnham, K. P., and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection, *Socio. Meth. Res.* **33,** no. 2, 261–304.

Campbell, K. W., and Y. Bozorgnia (2014). NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5%-damped linear acceleration response spectra, *Earthq. Spectra* **30,** 1087–1115.

Cauzzi, C., and E. Faccioli (2008). Broadband (0.05 to 20 s) prediction of displacement response spectra based on worldwide digital records, *J. Seismol.* **12,** 453–475.

Cavanaugh, J. E., and A. A. Neats (2012). The Bayesian information criterion: Background, derivation, and applications, *WIREs Comput. Stat.* **4,** 199–203, doi: 10.1002/wics.199.

Delavaud, E., F. Scherbaum, N. Kuehn, and T. Allen (2012). Testing the global applicability of ground-motion prediction equations for active shallow crustal regions, *Bull. Seismol. Soc. Am.* **102,** no. 2, 707–721.

Derras, B., P.-Y. Bard, and F. Cotton (2014). Towards fully data-driven ground-motion prediction models for Europe, *Bull. Earthq. Eng.* **12,** 495–516.

Derras, B., P.-Y. Bard, F. Cotton, and A. Bekkouche (2012). Adapting the neural network approach to PGA prediction: an example based on the KiK-net data, *Bull. Seismol. Soc. Am.* **102,** 1446–1461.

Douglas, J., S. Akkar, G. Ameri, P.-Y Bard, D. Bindi, J. J. Bommer, S. S. Bora, F. Cotton, B. Derras, M. Hermkes, *et al.* (2014). Comparisons among the five ground-motion models developed using RESORCE for the prediction of response spectral accelerations due to earthquakes in Europe and the Middle East, *Bull. Earthq. Eng.* **12,** 341–358.

Efron, B., and R. Tibshirani (1997). Improvements on cross-validation: The .632+ bootstrap method, *J. Am. Stat. Assoc.* **92,** no. 438, 548–560.

Eurocode 8 (2004). Design of structures for earthquakes resistance, part 1: General rules, seismic actions and rules for buildings, EN 1998-1, European Committee for Standardization (CEN), Brussels, Belgium.

Forster, M., and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions, *Br. J. Philos. Sci.* **45,** 1–35.

Foulser-Pigott, R., and K. Goda (2015). Ground-motion prediction models for Arias intensity and cumulative absolute velocity for Japanese earthquakes considering single-station sigma and within-event spatial correlation, *Bull. Seismol. Soc. Am.* **105,** 1903–1918.

Gregor, N., N. A. Abrahamson, G. M. Atkinson, D. M. Boore, Y. Bozorgnia, K. W. Campbell, B. S.-J. Chiou, I. M. Idriss, R. Kamai, E. Seyhan, *et al.* (2014) Comparison of NGA-West2 GMPEs, *Earthq. Spectra* **30,** 1179–1197.

Hagerty, M. R., and V. Srinivasan (1991). Comparing the predictive powers of alternative multiple regression models, *Psychometrika* **56,** no. 1, 77–85.

Jordan, T. H. (2014). The prediction problems of earthquake system science, *Seismol. Res. Lett.* **85,** no. 4, 767–769.

Kaklamanos, J., and L. G. Baise (2011). Model validation and comparisons of the Next Generation Attenuation of ground motions (NGA-West) project, *Bull. Seismol. Soc. Am.* **101,** 160–175.

Kotha, S.-R., D. Bindi, and F. Cotton (2016). Partially non-ergodic region specific GMPE for Europe and Middle-East, *Bull. Earthq. Eng.* **14,** no. 4, 1245–1263.

Kuehn, N., F. Scherbaum, and C. Riggelsen (2009). Deriving empirical ground-motion models: Balancing data constraints and physical assumptions to optimize prediction capability, *Bull. Seismol. Soc. Am.* **99,** 2335–2347.

Kulkarni, R. B., R. R. Youngs, and K. J. Coppersmith (1984). Assessment of confidence intervals for results of seismic hazard analysis, *Proc. of the Eighth World Conf. on Earthquake Engineering*, San Francisco, California, July 1984, Vol. 1, 263–270.

Luzi, L., S. Hailemikael, D. Bindi, F. Pacor, F. Mele, and F. Sabetta (2008). ITACA (ITalian ACcelerometric Archive): A web portal for the dissemination of Italian strong-motion data, *Seismol. Res. Lett.* **79,** no. 5, 716–722, doi: 10.1785/gssrl.79.5.716.

Mak, S., R. A. Clements, and D. Schorlemmer (2015). Validating intensity prediction equations for Italy by observations, *Bull. Seismol. Soc. Am.* **105,** no. 6, 2942–2954.

Malhotra, P. (2015). Discussion of NGA-West2 research project, *Earthq. Spectra* **31,** no. 3, 1875–1878.

McGuire, R. K. (2004). Seismic hazard and risk analysis, *EERI Monograph MNO-10*, Earthquake Engineering Research Institute, Oakland, California, 187 pp.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0.

Roselli, P., W. Marzocchi, and L. Faenza (2016). Toward a new probabilistic framework to score and merge ground-motion prediction equations: The case of the Italian region, *Bull. Seismol. Soc. Am.* **106,** 720–733.

Scherbaum, F., E. Delavaud, and C. Riggelsen (2009). Model selection in seismic hazard analysis: An information-theoretic perspective, *Bull. Seismol. Soc. Am.* **99,** no. 6, 3234–3247.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Stat.* **6,** no. 2, 461–464.

Shmueli, G. (2010). To explain or to predict?, *Stat. Sci.* **25,** no. 3, 289–310, doi: 10.1214/10-STS330.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Stat. Soc. B* **36,** 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Stat. Soc. B* **39,** 44–47.

GFZ German Research Centre for Geosciences
Telegrafenberg
14473 Potsdam, Germany
bindi@gfz-potsdam.de

Manuscript received 13 July 2016;
Published Online 31 January 2017