Mak, S., Schorlemmer, D. (2016): What Makes People Respond to “Did You Feel It?”? - *Seismological Research Letters*, *87*, 1, pp. 119—131.

DOI: http://doi.org/10.1785/0220150056

SSA

The Seismological Society of America
400 Evelyn Ave., Suite 201
Albany, CA 94706-1375
(510) 525-5474; FAX (510) 525-7204
www.seismosoc.org

What Makes People Respond to “Did You Feel It?”?

by Sum Mak and Danijel Schorlemmer

ABSTRACT

The data compilation of “Did You Feel It?” (DYFI) and other similar Internet-based macroseismic intensity databases relies on the voluntary responses from Internet users. A region of no responses could mean no perceivable ground shakings or no volunteers submitting responses. We examined the earthquake and socioeconomic conditions that affected the number of DYFI responses received for a region. A resulting statistical model described the expected number of DYFI responses received for an earthquake. We also showed that residents in California and the central and eastern United States followed similar behavior in responding to DYFI, despite the vast difference in seismicity for the two regions. This study allows for a quantitative definition of completeness for DYFI data. The presented modeling technique is applicable to other Internet-based macroseismic intensity databases.

INTRODUCTION

The earthquake ground-motion record is a scarce resource to seismologists. Although the number of seismometer installations solely determines the availability of instrumental records, the availability of macroseismic intensity (Grünthal, 2011) records is essentially anthropogenic. Intensity records are self-selective, meaning that an intensity value itself (i.e., the observable effects of the earthquake) may affect whether the intensity report exists. For historical earthquakes, regions that were sparsely populated and/or suffered relatively minor damage might be less documented in the literature, rendering seismologists with no basis to produce an intensity value. Damage investigations for recent earthquakes have been more systematically performed by government agencies and other parties, improving the completeness of records for many events (see, e.g., Dewey *et al.*, 2002, p. 6, for a description of damage investigation procedures for a modern earthquake).

For modern and future earthquakes, except for those having induced significant damage to qualify case-specific investigations, the major source of intensity reports is likely online questionnaires passively received from Internet users (e.g., the “Did You Feel It?” system, hereafter referred to as DYFI, Wald *et al.*, 1999, 2011; and the “Hai Sentito il Terremoto?” system, Sbarra *et al.*, 2010). The availability of Internet-based macroseismic intensity data (hereafter referred to as “iIntensity”) is likely affected by the self-selection bias because responses are mostly voluntary. A person may be less motivated to submit an online questionnaire if he or she has not been sufficiently “shaken.” The natural result for this bias is that, compared with a stronger shaking, a weaker shaking (assuming it was felt) is less likely to be reported.

Another obvious factor that affects the availability of iIntensity is the population size of the affected region. Wald *et al.* (2011, p. 694) attributed the population density as a key factor to the quantity (and quality) of DYFI data for an earthquake. Naturally, fewer reports are expected for a ground motion felt by only a few individuals, compared with the same ground motion but felt by a crowed.

To the first order, the population size and the ground-shaking level jointly determine the availability of iIntensity. Boatwright and Phillips (2013) estimated the average proportion of the population of a postal ZIP code region that responded to DYFI for various intensity levels. A quantitative estimation of the availability of iIntensity can be used to infer the completeness of intensity data. Data completeness is often influential in how the data should be used. For example, Gasperini (2001), Gómez Capera (2006), and Pasolini *et al.* (2008) discarded low-intensity data when studying the attenuation of intensity with distance because the workers considered them potentially less complete. Gómez Capera *et al.* (2010) did the same when evaluating the seismic hazard of Italy using intensity data. On the other hand, Albarello and D’Amico (2004) conducted a statistical analysis on data of both low-intensity and high-intensity values and concluded that they have similar skewness; they therefore considered that low-intensity data were not incomplete. It is, of course, valid to discard data of questionable quality. It is, however, more desirable if the completeness of quality data can be quantitatively assessed.

The estimation of data availability of iIntensity using only the population size and the ground-shaking level assumes that every potential respondent behaves identically and every earthquake is equally perceived. This could be an oversimplification. The aim of the present study is to quantitatively investigate factors affecting the availability of DYFI data. An expression of the completeness of quality iIntensity is a natural product of the study.

MODELING THE NUMBER OF RESPONSES

The macroseismic intensity of an earthquake is often graphically displayed as a spatial distribution of intensity data points

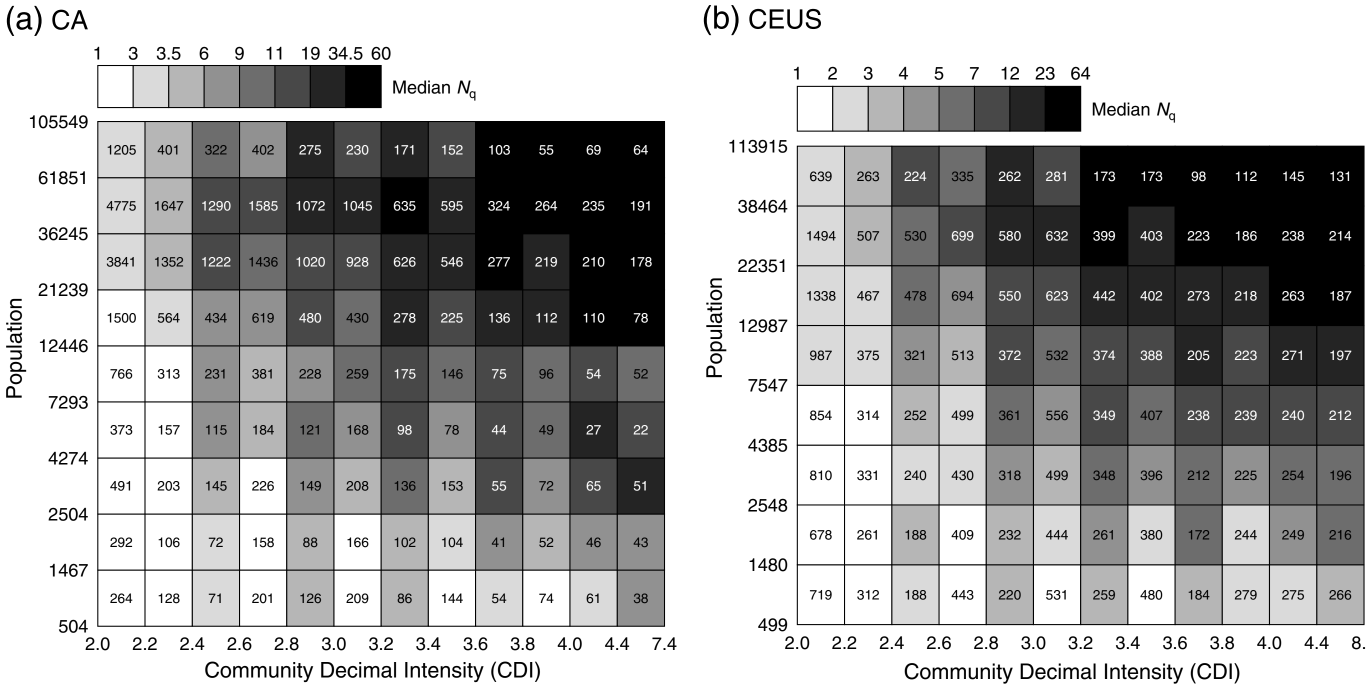
doi: 10.1785/0220150056

(IDPs; Musson and Cecić, 2012, section 12.1). For DYFI in the United States, a community decimal intensity (CDI) is computed for each postal ZIP code region using all responses (i.e., online questionnaires) received from that ZIP region. An IDP for DYFI, therefore, is a ZIP-based intensity value calculated using all responses received from that ZIP region. The availability of DYFI data can be represented by the number of questionnaires (N_q) received for a ZIP region (i.e., for each IDP). Such a number is presumably related to the ground-shaking level (represented by the intensity value) and the population of the ZIP region, as well as other factors describing the perceptibility of the earthquake and the willingness of people to submit an online questionnaire. Figure 1 shows the medians of N_q binned by CDI and population size. They confirm that N_q monotonically increases with both CDI and population size. Such a simple graphical inspection cannot be applied if more factors, in addition to CDI and population size, are considered. A regression analysis is a more suitable tool. In this regression analysis, the response variable (regressand) is N_q, and the factors affecting the availability of DYFI data, including CDI and population size, are the explanatory variables.

N_q, being a nonnegative integer, can be considered as a kind of count data (in this case, the count of DYFI responses). A regression analysis on count data is often treated using a generalized linear model (GLM; e.g., Zuur *et al.*, 2009, chapter 9). A GLM regression is similar to an ordinary least-squares regression (i.e., the simplest form of least-squares curve fitting) in many aspects, except that the response variable does not follow a Gaussian distribution. N_q could potentially be zero, meaning that the ground motion has not been reported to DYFI. The unreported, and therefore unknown, intensity value is a kind of missing data. Omitting the missing data will induce a bias to the regression.

One way to deal with the missing data is to separately estimate the unreported intensity values. One possible estimation is through interpolating an isoseismal map. It is often perceived that the spatial distribution of intensity is smooth so that interpolation should be straightforward. This is not the case, however, in the context of the present study, because regions of missing data often lie at the perimeter or beyond, instead of within, regions where the intensity data are available. Therefore, extrapolation, of which the result is often questionable, is more often needed than interpolation. Another possible estimation is by inference from instrumental records, when available, through a ground motion–to–intensity conversion equation (e.g., Worden *et al.*, 2012). Although inferring an earthquake parameter (e.g., estimating the magnitude of a historical earthquake) or ground motion (e.g., the use of ShakeMap in the central and eastern United States [CEUS]) from macroseismic intensity has been a well-adopted practice and sometimes is the only option, it is a general rule that conversion between intensity measures should be avoided when possible.

The other way to deal with the missing data is to allow them to be missing while correcting the regression to avoid the bias that would otherwise be induced. The zero-truncated model (e.g., Zuur *et al.*, 2009, section 11.2) has been developed for this purpose. The essence for a zero-truncated model is to



▲ **Figure 1.** Median values (grayscale) of the number of responses (N_q) of ZIP-based intensity data points (IDPs) for (a) California and (b) central and eastern United States (CEUS), binned by community decimal intensity (CDI) and population size. The boundaries of the bins are shown by the axis labels. Bins with too few data were combined, so the bin widths are not identical. Population is binned in logarithmic scale. The number in each box indicates the number of IDPs in each bin. The discrete grayscale is selected such that roughly the same number of bins fall into each scale. The data selection process is described in the Influential Factors to the Number of Responses section.

Table 1
Data Winnowing

Location	Filter Steps	Number of IDPs*	Number of Earthquakes	Number of ZIPs
California	Total (2000–2014)	140,840	6578	2164
	$R \le 200$ km	124,048	6293	1747
	CDI > 1	114,300	6221	1747
	$M \ge 4$	43,580	527	1695
	Population ≥ 500	42,871	525	1577
Central and Eastern United States	Total (2000–2014)	86,189	2032	18227
	$R \le 500$ km	74,549	2017	16854
	CDI > 1	70,575	2013	16657
	$M \ge 4$	37,546	89	15700
	Population ≥ 500	36,306	89	14905

*ZIP-based intensity data points (IDPs).

convert the probability distribution of a nonnegative variable into that of a strictly positive one. In a GLM regression on nonnegative count data, the distribution of the response variable N_q is assumed to follow a probability distribution function (PDF), $f(N_q|N_q \ge 0)$. For strictly positive count data, the PDF will become

$$f(N_q|N_q > 0) = \frac{f(N_q|N_q \ge 0)}{1 - f(0)}. \quad (1)$$

This normalization of the PDF is similar to the conversion from a Gutenberg–Richter relation to a truncated Gutenberg–Richter relation (e.g., McGuire, 2004, section 3.3.1).

The remaining works for the modeling are identical to the conventional GLM regression analysis: (1) choose a link function, (2) choose a probability distribution model, (3) choose the explanatory variables, (4) determine the coefficients (usually numerically) by maximizing the likelihood, and (5) justify the choices made in steps 1–3 by model diagnostics.

INFLUENTIAL FACTORS TO THE NUMBER OF RESPONSES

The present study requires three kinds of information, namely the DYFI IDPs, physical parameters of the earthquakes that might affect their perceptibility, and socioeconomic status of the respondents that might affect their willingness to submit an online questionnaire. IDPs and the associated earthquake parameters came from the DYFI system. ZIP-based IDPs from California, which comprise the majority of the DYFI dataset due to the high seismicity in California, were used as the primary dataset in the present study. It is interesting to see if people living in a more seismically silent part of the United States behave differently from Californians. Therefore, IDPs from the CEUS (defined as east to 105° W) were also analyzed. The data were collected from the beginning of the establishment of DYFI (i.e., the year 2000) to the end of 2014.

ZIP-based socioeconomic information came from the summary file 1 of the U.S. decennial census 2010 and from the table DP02 of the American Community Survey 2012 (see Data and Resources). The decennial census is a comprehensive survey that sampled almost the entire United States population, directly surveying basic demographics like the population size, age, and race distribution. The American Community Survey estimates more detailed socioeconomics of the population by random sampling, focusing on populous regions only.

It is practically impossible to completely identify all factors that may affect the DYFI response rate, nor to fully understand the effective mechanism of any particular factor. In this study, we identified and used earthquake and socioeconomic parameters available from the DYFI system and the census database that could conceivably bear some relation to the human reaction to earthquakes as explanatory variables in the regression. The selected explanatory variables and their potential mechanisms are explained below.

Earthquake Parameters

Community Decimal Intensity

CDI is explained in the Modeling the Number of Responses section. The CDI is semicontinuous with an increment of 0.1, but values between 1.0 and 2.0 are not defined. This large gap may violate the linearity assumption of the GLM analysis (see the Modeling Result and Diagnostics section). Therefore, IDPs with CDI of 1.0 were excluded from the analysis.

Magnitude

An earthquake with larger magnitude may be more visible than one with smaller magnitude (e.g., attracting much media coverage) and so may be more likely to induce responses. Earthquakes with magnitude less than 4 were excluded from the analysis, because the human response to microseismicity is uninteresting for most real-world applications. Discarding small earthquakes significantly reduced the amount of data (Table 1) and so relieved much of the computational effort.

Epicentral Distance

People tend to be concerned with things, such as earthquakes, that occur nearby. This is equivalent to the principle of proximity in journalism (e.g., International Press Institute, 1953, pp. 67, 73; Boyd, 2001, p. 19). Far-field IDPs were excluded from the analysis because the linearity assumption of the GLM analysis (see the Modeling Result and Diagnostics section) may not hold for a wide range of distances. Residents in the far field are unlikely to feel the earthquake, and so their motivation to respond to DYFI may not be correctly described by the explanatory variables considered here. IDPs beyond 200 and 500 km from the epicenter, respectively, for California and the CEUS were excluded from the analysis. These thresholds were selected to discard about 10% of IDPs (Table 1).

Focal Depth

A deeper earthquake is often felt in a wider region, increasing the number of people who respond to DYFI. An example is the widely felt M_L 5.1 earthquake that occurred near the city of Parma, Italy, on 23 December 2008, with a depth of 26.7 km (Sbarra *et al.*, 2010, p. 574).

Occurrence Time

The occurrence time of an earthquake may determine people's motion status during the ground shaking, thereby affecting their perception. On the other hand, people who are sleeping (at night) or busy (at day) might be less willing to fill in an online questionnaire. We categorized the earthquake occurrence time as "day" (07:00–15:00), "evening" (15:00–23:00), and "night" (23:00–07:00).

Date

The DYFI system was established at the end of the year 1999. The visibility of the system might increase with time due to, for example, the diligence of the U.S. Geological Survey (USGS) public affairs officers. In addition, the Internet access rate likely increased with time. We counted the date as the number of days from 1 January 2000.

Socioeconomic Status

Population Size

Socioeconomic status is explained in the Modeling the Number of Responses section. All ZIP-based socioeconomic information considered in the present study describes the status of residents. They are meaningful to the regression only if respondents are residents of that ZIP region. This assumption might be less correct when there are few residents in the ZIP region. Therefore, ZIP regions with less than 500 residents were excluded from the analysis.

Percentage of Hispanic Population

California has a large Hispanic population. Some ZIP regions are primarily (close to 100%) Hispanic. Ethnic features such as language skills, social circle, interest to public affairs, etc., may affect people's tendency to respond to DYFI. The census database also contains the information of other minor ethnic groups, but, because they often consist of very small proportions of the population, they were not included in the analysis.

Percentage of Educated Population

Those individuals who received higher education (Bachelor's degree or above) might be more connected to the society and so more likely to have heard about the DYFI system.

Percentage of Poor-English-Speaking Population

The DYFI webpage is given in English. People not fluent in English (census definition: speaking English less than "very well") may be reluctant to use the webpage and so are less likely to become respondents.

Percentage of Buildings with Complex Structure

The architecture of a building may affect the occupant's perception to ground shakings. A building with more than 10 units (e.g., apartments) is defined as complex. Complex buildings are likely taller, and occupants at higher stories might feel earthquake ground motions more clearly.

Percentage of Population Living below the Poverty Line; Percentage of Foreign-Born Population

As opposite to the educated population, poor people and immigrants might be less connected to the society and so less likely to have heard about the DYFI system.

Percentage of Veteran Population

Those who have served in the military might be more dutiful in reporting an earthquake to the authority.

Average Household Size

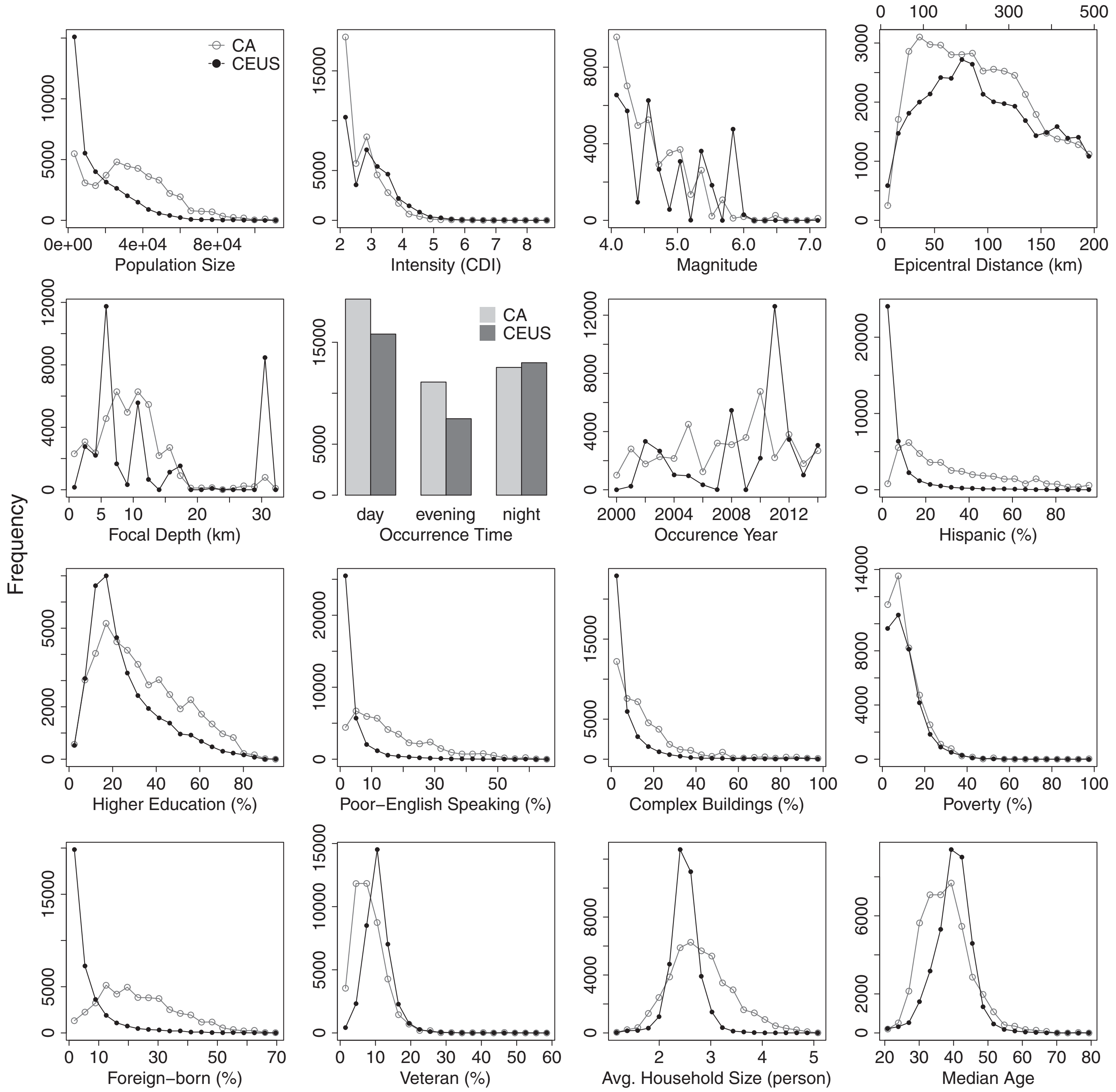
It may be uncommon for multiple people living in the same household to submit separate DYFI questionnaires. For the same population size, a smaller average household size means a larger number of households, and so a larger number of responses may be expected. A very small number of ZIP regions in the dataset have nonzero population but zero households because all residents are living in group quarters (e.g., student dormitory). These ZIP regions were excluded from the analysis.

Median Population Age

DYFI relies on the participation of Internet users. It is well known that older people are less-frequent users of the Internet. Fewer responses may be expected from regions with a high proportion of senior citizens.

The amount of selected data after each step of data winnowing is given in Table 1. The distribution of each explanatory variable is given in Figure 2. Most variables do not distribute uniformly but have a long tail, especially for the CEUS dataset. This is not desirable for regression analysis but is common for real-world data. Some variables are intercorrelated (e.g., intensity, magnitude, and distance; foreign-born and poor-English speaking). The correlations did not cause numerical difficulties in the present study and so were tolerated.

We have no intention to claim that the above-mentioned factors are complete in determining the number of responses to DYFI. For example, Wald *et al.* (2011, p. 694) attributed the prevalence of Internet access as a factor that affects the number of DYFI responses. Although this appears logical, we were not



▲ **Figure 2.** Distribution of explanatory variables. The plot for epicentral distance uses different abscissae for the two lines (bottom, California; top, CEUS). The plots are arranged in the same order as the explanatory variables given in Table 2.

able to locate an authoritative data source of the spatial and temporal distribution of Internet access rate and so did not included this factor in the present study.

MODELING RESULT AND DIAGNOSTICS

We used a logarithmic link function to relate the selected explanatory variables to the number of responses received for a ZIP region:

$$\ln \widehat{N_{\mathrm{q}}} = \beta_0 + \sum_i \beta_i (x_i - \widehat{x_i}), \tag{2}$$

in which $\widehat{N_{\mathrm{q}}}$ is the expected number of DYFI responses received for a ZIP region, x_i are the explanatory variables described in the Influential Factors to the Number of Responses section, $\widehat{x_i}$ is the mean value of the corresponding explanatory variable for the California dataset, and β_i are the coefficients to be determined. The means were subtracted from the covariates in order to bestow on fitted constant (i.e., β_0) the physical meaning as the expected logarithmic number of responses during the day time, given that all other conditions are the California average. The mean-removal process is implemented for continuous explanatory variables (i.e., covariates) only. Categorical explanatory variables (i.e., factors; the only factor used in the current study is occurrence time) have no mean values. There is no physical reason why the expected number of responses has to be related to the explanatory variables linearly. The linearity was assumed to capture the first-order effect of the factors while keeping the model-fitting process simple.

For GLM regression on count data, the two most popular probability distributions (i.e., $f(\cdot)$ in equation 1) that the response variable is assumed to follow are the Poisson and the negative binomial distributions; the former is a special case of the latter with a shape parameter of unity. We used the negative binomial model. The result justified this choice, because the shape parameter obtained is significantly different from unity. With the zero-truncation adjustment explained in equation (1), the likelihood function for a negative binomial model is (Zuur *et al.*, 2009, their equation 11.7):

$$\ell = \prod_{i=1}^{N} \left\{ \frac{\Gamma(N_{qi} + \sigma)}{\Gamma(\sigma)\Gamma(N_{qi} + 1)} \left(\frac{\sigma}{\widehat{N_{qi}} = +\sigma} \right)^{\sigma} \times \left(1 - \frac{\sigma}{\widehat{N_{qi}} + \sigma} \right)^{\widehat{N_{q_i}}} \left[1 - \left(\frac{\sigma}{\widehat{N_{qi}} + \sigma} \right)^{\sigma} \right]^{-1} \right\}, \quad (3)$$

in which $\Gamma(\cdot)$ is the gamma function, σ is the shape parameter of the model, N is the total number of IDPs, and N_{qi} and $\widehat{N_{q_i}}$ are the observed and expected number of responses, respectively, for the ith IDP; the latter is described by equation (2). A set of coefficients was determined by numerically maximizing the log likelihood (Table 2). Population size and distance were logarithmically transformed to enhance numerical stability. A likelihood-ratio test (e.g., DeGroot and Schervish, 2012, pp. 543–555) shows that all the explanatory variables are significant (Table 3), and so the full model (i.e., the model including all explanatory variables) is preferred.

Model diagnostics using residual plots are given in Figures 3–4. In general, a sufficient regression model should produce pattern-free residual plots. For residual diagnostics of GLM regression, there are additional cautions compared with that of ordinary least-squares (OLS) regression. First, the Pearson residual (residual divided by standard deviation), instead of the residual, was used because of the heteroskedastic nature of GLM regression. Second, the definition of a "pattern-free" residual plot is sometimes unclear due to the nonsymmetry of the negative binomial distribution, especially under unbalanced data. A set of synthetic residual plots is therefore given in Figure 5 as an example of residual plots from a good-fit model. The synthetic residuals were generated using synthetic N_q values that were randomly generated following a negative binomial model with coefficients given in Table 2. Patterns in Figures 3–4 that are not seen in Figure 5 are signals of potential model insufficiency.

The residuals show no unusual pattern for most variables except for intensity and distance, in which the residuals for large intensity and small distance values are concentrated (compared with those in Fig. 5). This implies that a linear relation

Table 2
Fitted Coefficients to be Used in Equation (2)

Fitted Coefficient	Explanatory Variable (x_i)	Acronym	California	CEUS	$\widehat{x_i}$
β_0	(Constant)		2.046	2.075	
β_1	Population size	$\log_e(pop)$	0.6891	0.7830	10.01
β_2	Macroseismic intensity	CDI	0.8051	0.5381	2.628
β_3	Magnitude	*mag*	1.490	1.267	4.579
β_4	Distance	$\log_e R$	−1.229	−1.171	4.337
β_5	Focal depth	*Depth*	0.03960	0.04183	9.506
β_6	Occurrence time (evening)	*Time*	0.2647	0.7850	
β_6	Occurrence time (night)	*Time*	−0.1943	−0.1905	
β_7	Date		0.0002307	0.0003055	2972
β_8	Percentage of Hispanic population	*PctHisp*	−0.01281	−0.01195	31.69
β_9	Percentage of educated population	*PctHighEd*	0.01312	0.01880	32.86
β_{10}	Percentage of poor-English-speaking population	*PctPoorEng*	0.0113	0.02048	16.00
β_{11}	Percentage of complex building	*PctHU10unit*	0.002198	0.004993	15.78
β_{12}	Percentage of poverty population	*PctPoverty*	−0.01438	−0.004600	10.51
β_{13}	Percentage of foreign-born population	*PctForeignBorn*	−0.01047	−0.02102	23.63
β_{14}	Percentage of veteran population	*PctVeteran*	−0.007800	0.01935	8.196
β_{15}	Average household size	*AvgHHsize*	−0.5107	−0.7196	2.808
β_{16}	Median age	*MedianAge*	−0.02622	−0.02564	37.76
σ	(Shape parameter)		0.5296	0.5252	

Table 3
Likelihood Ratio Test

Explanatory Variable	California d.d.	California p	CEUS d.d.	CEUS p
$\log_e$(pop)	5930	$<1 \times 10^{-4}$	8437	$<1 \times 10^{-4}$
CDI	2334	$<1 \times 10^{-4}$	1276	$<1 \times 10^{-4}$
Mag	5799	$<1 \times 10^{-4}$	4207	$<1 \times 10^{-4}$
$\log_e R$	6295	$<1 \times 10^{-4}$	5375	$<1 \times 10^{-4}$
Depth	838.8	$<1 \times 10^{-4}$	430.0	$<1 \times 10^{-4}$
Time	646.9	$<1 \times 10^{-4}$	1695	$<1 \times 10^{-4}$
Date	1820	$<1 \times 10^{-4}$	385.9	$<1 \times 10^{-4}$
PctHisp	315.6	$<1 \times 10^{-4}$	62.11	$<1 \times 10^{-4}$
PctHighEd	357.5	$<1 \times 10^{-4}$	727.4	$<1 \times 10^{-4}$
PctPoorEng	25.58	$<1 \times 10^{-4}$	21.69	$<1 \times 10^{-4}$
PctHU10unit	10.56	1.154×10^{-3}	27.95	$<1 \times 10^{-4}$
PctPoverty	131.1	$<1 \times 10^{-4}$	13.62	2.24×10^{-4}
PctForeignBorn	34.71	2.41×10^{-3}	57.05	$<1 \times 10^{-4}$
PctVeteran	10.57	1.149×10^{-3}	69.95	$<1 \times 10^{-4}$
AvgHHsize	332.0	$<1 \times 10^{-4}$	295.4	$<1 \times 10^{-4}$
MedianAge	263.7	$<1 \times 10^{-4}$	187.7	$<1 \times 10^{-4}$

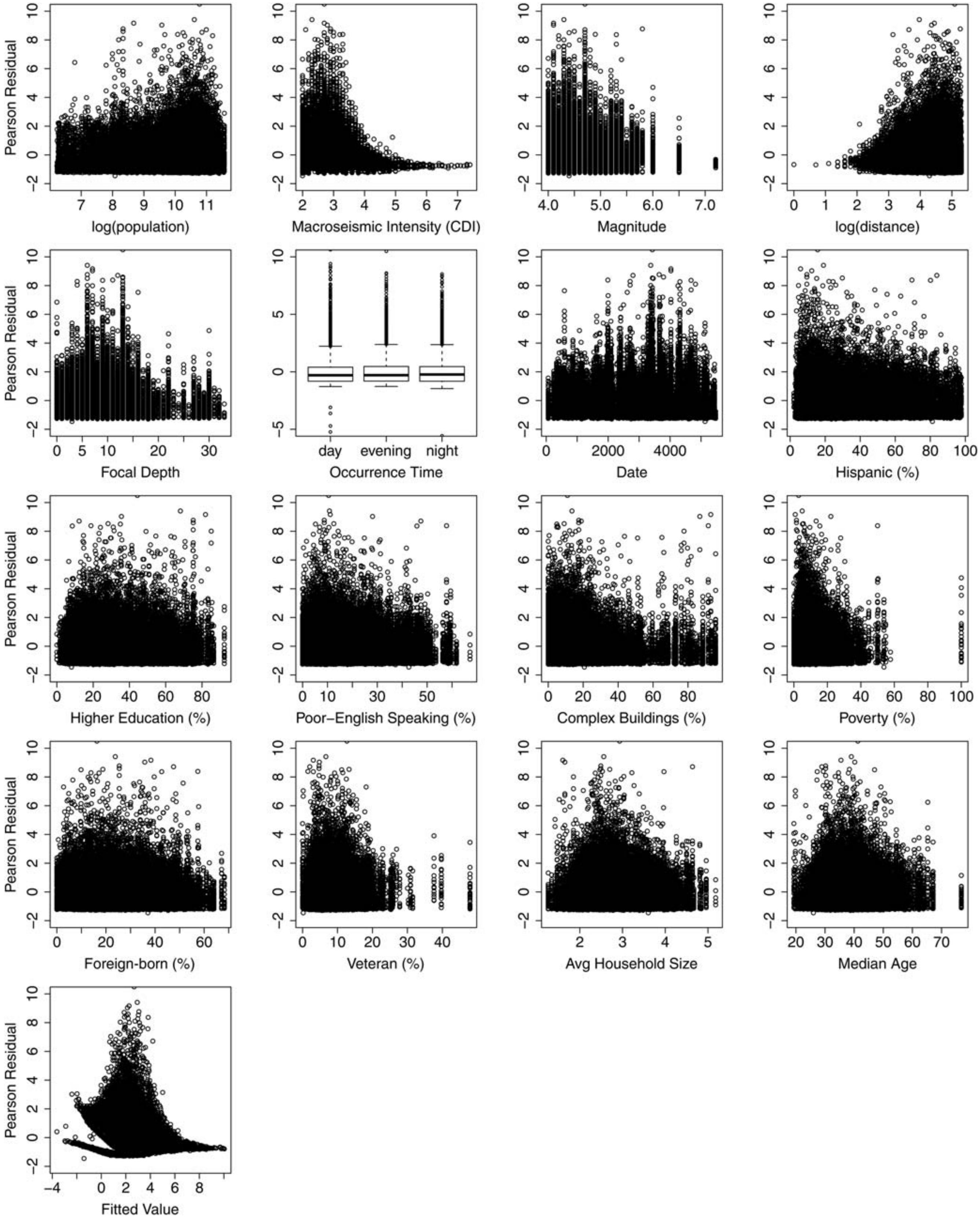
A small p value means the full model is significantly better. See the first paragraph of the Discussion section for the meaning of deviance difference (d.d.).

may not be sufficient for the two variables; there may be a saturation effect at large intensity and small distance. The relation between intensity and number of responses could be complicated, including at least two effects: (1) the proportion of people feeling the ground shaking increases with intensity, and (2) the motivation of people who felt the ground shaking to report it may increase if the shaking is more severe. The first effect, and likely also the second, will saturate at high intensity levels. On the other hand, Wald *et al.* (2011, p. 700) anticipated that when the ground motion is strong enough to induce widespread damage, the number of DYFI responses will be reduced due to the interruption of Internet service and power. This in fact happened during the 2009 L'Aquila (Italy) earthquake. Sbarra *et al.* (2010, p. 576) reported that no online questionnaires were received from the epicenter region in the first few days after the earthquake. However, they also reported that, after a few days, a large number of responses were received from the epicenter region, primarily submitted by people who suffered relatively minor damage. Factors specifically affecting large-intensity values may make the linearity assumption of the regression analysis invalid. In the present study, the number of IDPs with high intensity and short distance is very small (less than 5% of IDPs have CDI ≥ 4.4), and so the insufficient linear functional form might not seriously affect the regression. Because the purpose of the present study is to capture the first-order effect, further and more complicated modeling (e.g., using the generalized addictive model) was not conducted to look for a more accurate functional form for intensity and distance. In addition, the small amount of data at those ranges may not warrant a more complicated model.

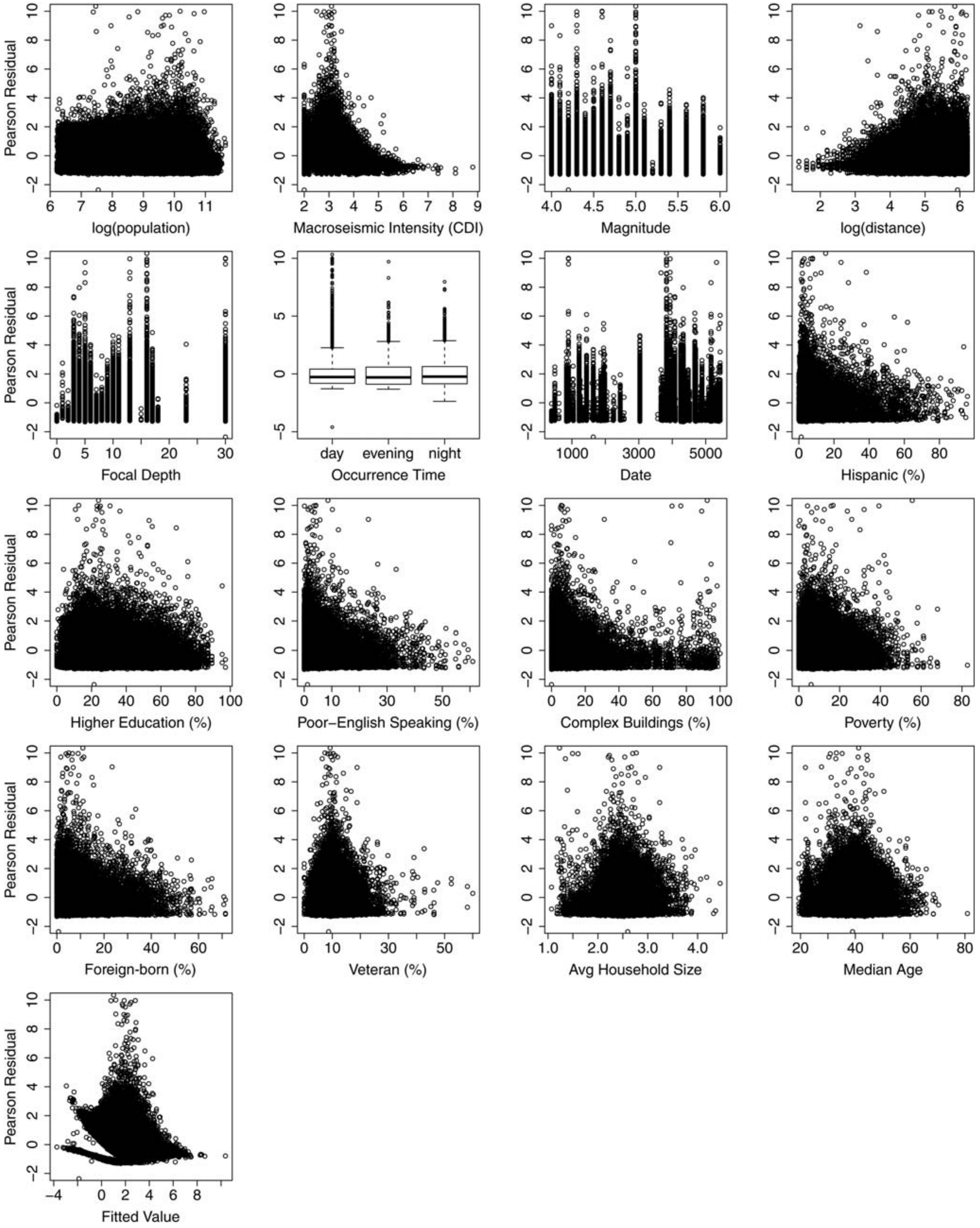
The goodness of fit of the model can be visualized through comparing the model-predicted distribution of N_q with the observed one (Fig. 6). This is similar to the concept of a QQ plot for OLS regression or the concept of "calibration" used by meteorologists to assess predictive models (Jolliffe and Stephenson, 2012, chapter 2.10). The comparison shows that, for $N_q > 3$, the model matched the observations quite well. There were more IDPs with $N_q = 1$ to 3 than that predicted by the model. Rare factors not considered in the Influential Factors to the Number of Responses section might produce particularly notable effects on IDPs of small N_q. For example, the existence of a single dedicated seismologist or Earth sciences student who is diligent in reporting to DYFI whenever there is an earthquake would efficiently turn N_q to one when it otherwise should be zero by the prediction using factors considered in the present study. Similarly, an observer's awareness of earthquakes may be substantially increased by subscriptions to Internet services such as the Earthquake Notification Service (ENS), ShakeCast, and Prompt Assessment of Global Earthquakes for Response (PAGER) of the USGS. There is currently no way to include such factors into the analysis.

DISCUSSION

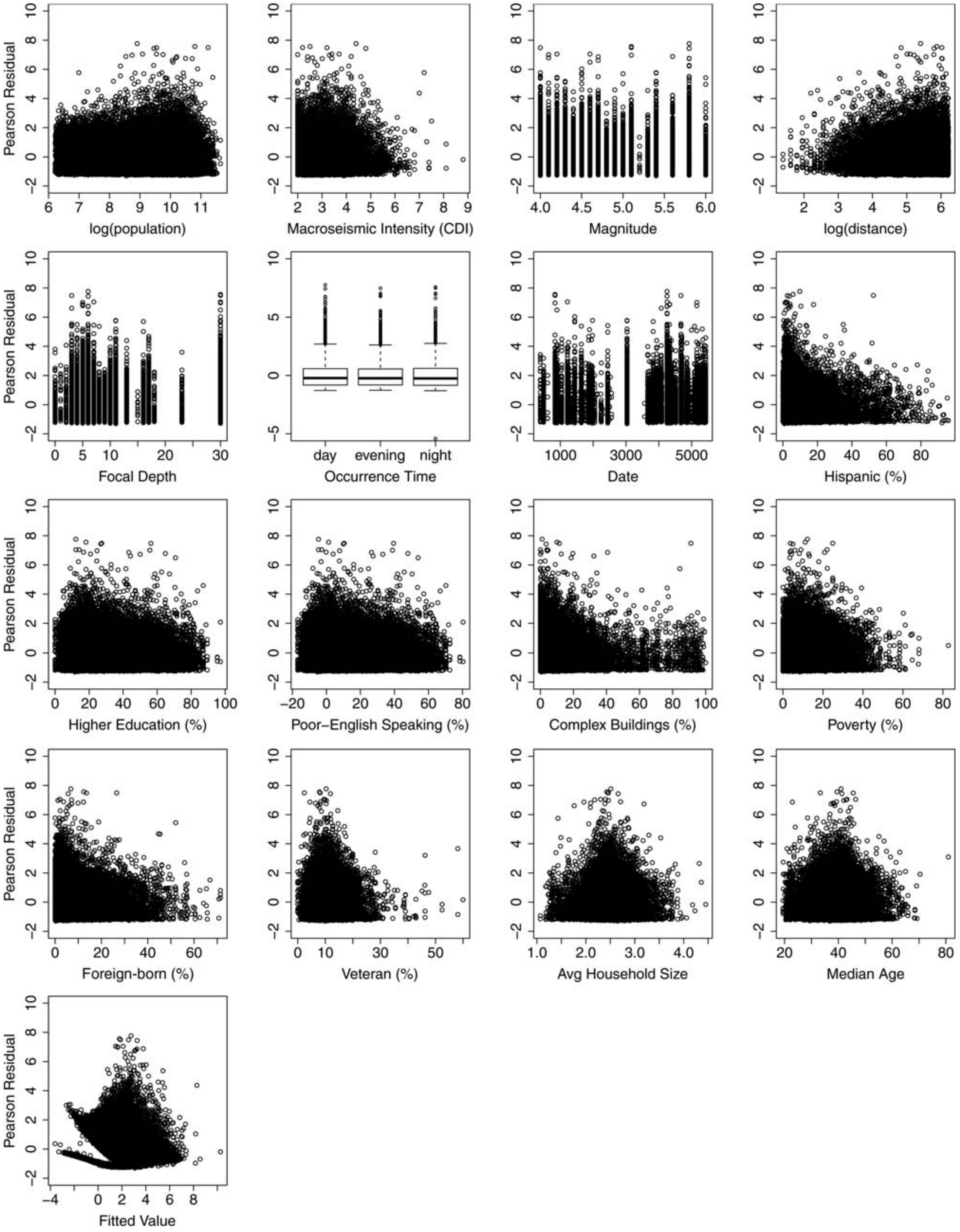
The signs of the fitted coefficients (Table 2) largely confirmed the conjectured effects of each explanatory variable described in the Influential Factors to the Number of Responses section. The expected number of DYFI responses increased with intensity, magnitude, focal depth, time since the establishment of DYFI, population, educated population size, and proportion



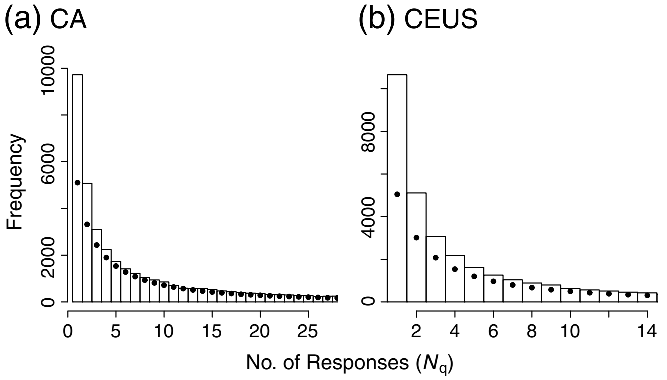
▲ **Figure 3.** Pearson residual plots for the California model.



▲ **Figure 4.** Same as Figure 3 but for the CEUS.



▲ **Figure 5.** Example of pattern-free residuals for a generalized linear model (GLM) regression. This figure is identical to Figure 4, except that the N_q values are not observed, but synthesized using a negative binomial model with coefficients given in Table 2. Note its similarity to Figures 3 and 4, except for CDI and perhaps $\log_e R$.

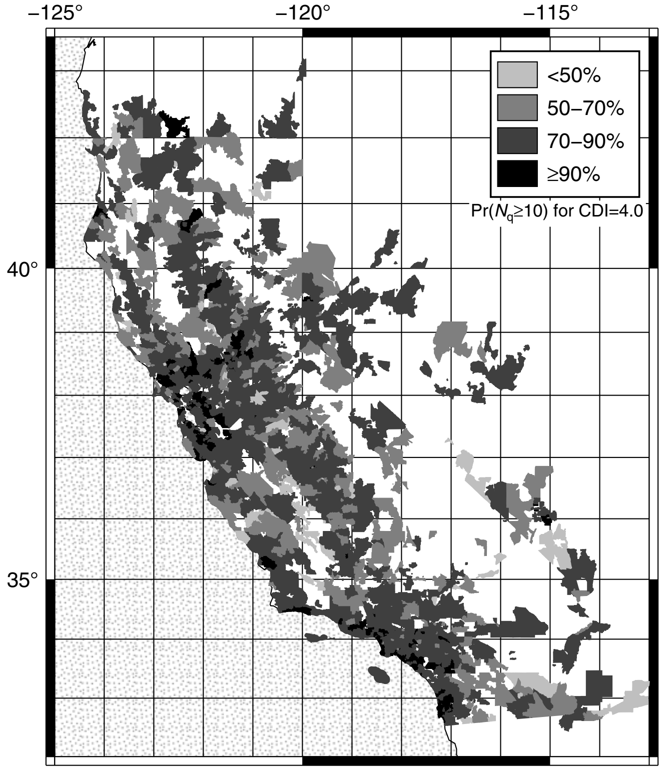


▲ **Figure 6.** Observed distribution of N_q (bars) versus model-predicted distribution (solid dots) for (a) California and (b) CEUS. N_q has a long tail, and only the lower 80% of data are shown.

of complex buildings. It decreased with distance, Hispanic population, poor-English speaking population, high-poverty population, foreign-born population, average household size, and median population age. The deviance difference (i.e., change in deviance) after dropping each explanatory variable from the full model (Table 3) provides a sense of how each variable contributed to the fit of the model; the larger the difference is, the more the dropped variable contributes to the fit. The deviance for GLM regression (Zuur *et al.*, 2009, section 9.5.3) is the analog of R^2 (coefficient of determination) for OLS regression. The deviance difference for GLM, therefore, can be understood as the difference in R^2 for OLS between two models. Judging from the deviance difference, magnitude and distance were the two most important parameters after intensity and population size. Generally, the physical earthquake parameters contributed more to the fit than did the socioeconomic parameters.

Although the interpretations of the effect of many explanatory variables, as given in the Influential Factors to the Number of Responses section, are straightforward based on common sense, it is often not difficult to conceive opposite interpretations that appear to be equally plausible; it is important to verify any interpretative statements with data. The interpretation is further complicated by the intercorrelation among multiple variables. For example, a portion of the Hispanic population could be foreign-born and probably does not speak English fluently. The significance of all three variables (*PctHisp*, *PctForeignBorn*, and *PctPoorEng*) implies that one cannot simply attribute the negative effect of the number of DYFI responses of the Hispanic population to their language proficiency and foreign-born nature.

An exposure variable (i.e., coefficient fixed to be unity) is sometimes used in a GLM regression. In the present study, the population size could be considered as an exposure variable, implying that, after all corrections, the number of responses simply scales with the population size. This is the assumption made by Boatwright and Phillips (2013). The regression result shows that the coefficient of $\log_e(pop)$ is significantly different from unity, meaning that the population size cannot be treated as an exposure variable. This either means that there are important explanatory variables missing or the relation between population size and number of responses is not that simple. Because, in practical terms, the list of explanatory variables is never completely clear, it is better to treat the population size as a combination of both the count of residents and a feature of the living environment instead of simply as an exposure variable. There could be alternative representations of resident count. We tried to replace population size with number of households, and with the labor force size. The resulting models were not better in fit. Therefore, the population size is still the best representation of the count of residents in a ZIP region.



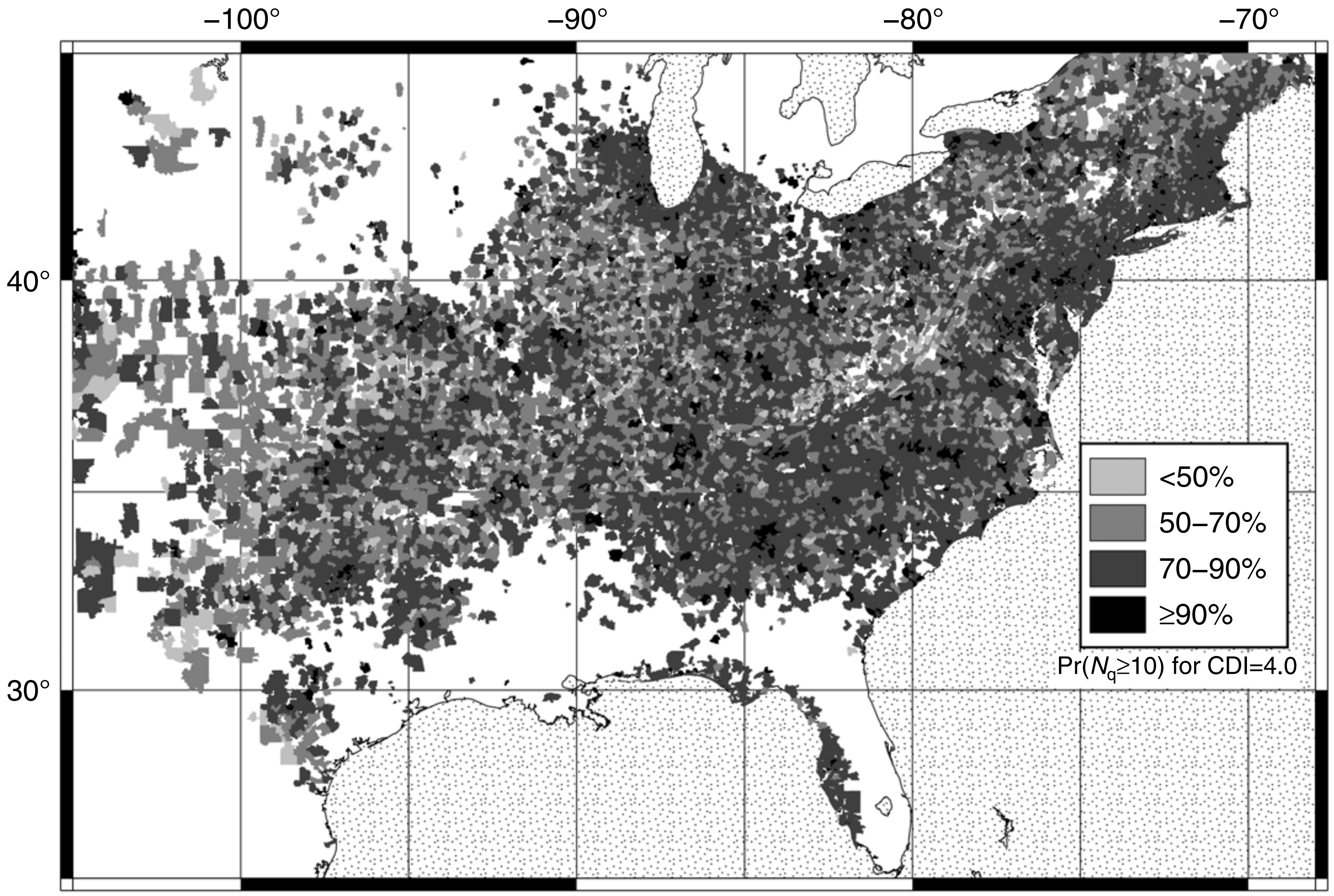
▲ **Figure 7.** ZIP-based probability for $N_q \geq 10$ when CDI = 4.0, for magnitude 5, distance of 30 km, depth of 10 km, and during day time at the end of the year 2014. The white area is not modeled because of data nonexistence.

The fitted coefficients for the datasets of California and of the CEUS are surprisingly similar. The fitted constant (β_0) represents the logarithmic number of responses when all conditions are the California average. Comparing the fitted constants between the two datasets is a simple way to inspect the similarity of the DYFI response patterns of the two regions. The consistent results on two sets of independent data provide a degree of confidence to the accuracy of the analysis. It also implies that, in spite of the very different seismicity rates and thereby the residents's experience on earthquakes, the response patterns for the two regions follow some similar mechanism. Wald *et al.* (2011, *p.* 694) reported that a surprisingly large number of responses were received from earthquakes in the eastern United States, where earthquakes are infrequent and residents have presumably lower awareness of earthquakes. They considered it as an evidence of uncorrelation between earthquake awareness and DYFI response rate. The present study confirms this claim by data analysis. The major dissimilarity between the two sets of fitted coefficients is on the variable *PctVeteran*; the coefficient is negative for California but positive for the CEUS. The underlying reason is unclear. Nevertheless, *PctVeteran*'s contribution to the model fit was relatively minor (Table 3).

The reliability of a CDI value depends on the number of responses used to compile an IDP: the larger the number of responses from which an IDP is compiled, the more stable the CDI value is (see fig. 3 of Worden *et al.*, 2012). Hough (2013) found a related phenomenon in historical macroseismic intensity data: the intensity for a whole city could be disproportionally influenced by a few dramatic effects emphasized by archival accounts. The present study provides an objective way to measure the completeness of quality DYFI data. If a quality IDP is defined as one compiled from at least $N_{\min}$ responses, the completeness is represented by the probability for a ZIP region to produce such a quality IDP. Take $N_{\min} = 10$, the probability to have quality IDPs of CDI = 4.0 induced by a hypothetical earthquake of magnitude 5 at 10 km depth and 30 km from the site during day time at the end of the year 2014 is given in Figures 7–8. Because the population size is a critical factor in determining the number of responses, the maps are largely another form of population map. The two figures show that the probability for most ZIP regions to produce at least 10 responses under the above-mentioned conditions is mostly ≥ 70%. The darkest regions, corresponding to the probability of ≥ 90%, overlap with the most populated regions. The high probabilities indicate that IDPs with CDI ≥ 4 are generally very likely of good quality. Other options of $N_{\min}$ may fit other applications; the model presented here will yield the data completeness accordingly.

CONCLUSION

We statistically explained and predicted the number of DYFI responses by population size, intensity, and various earthquake and socioeconomic parameters. The effects of these parameters on the number of responses were found to largely agree with common sense. To correctly predict the number of responses, the most important factors were found to be population size, intensity, magnitude, and distance. The response behavior to DYFI for residents in California and the CEUS were comparable despite the very different seismicity rates for the two regions. Voluntary responses to DYFI appeared to follow a common mechanism with small regional dependence. DYFI data of intensity values 4 or above were likely of good quality. The present study



▲ **Figure 8.** Same as Figure 7 but for the CEUS.

provides a quantitative measure of data completeness and so assists the data selection of potential analyses based on DYFI data. We consider the presented statistical modeling technique applicable to other similar databases of Internet-based macroseismic intensity.

DATA AND RESOURCES

All “Did You Feel It?” (DYFI) data used in the analysis were downloaded from earthquake.usgs.gov/earthquakes/dyfi. Census data were downloaded from factfinder.census.gov. Statistical analyses were conducted using R (R Development Core Team, 2008), particularly with the package VGAM (cran.r-project.org/web/packages/VGAM) for the GLM fitting. Figures 1, 7, and 8 were prepared using Generic Mapping Tools (Wessel *et al.*, 2013). All websites were last accessed on December 2015.

ACKNOWLEDGMENTS

A discussion with Robert Clements initiated this study. A discussion with David Harte on modeling techniques further shaped the study. Vince Quitoriano provided crucial help in data retrieval. David Wald pointed the first author to a few important references. Jim Dewey and David Wald provided useful comments on an early version of the manuscript. Comments from C. Bruce Worden, Rémy Bossu, two additional anonymous reviewers, and Associate Editor Susan Hough substantially improved the manuscript. The study was developed under the financial support by the Global Earthquake Model Foundation.

REFERENCES

Albarello, D., and V. D’Amico (2004). Attenuation relationship of macroseismic intensity in Italy for probabilistic seismic hazard assessment, *Bollettino di Geofisica Teorica ed Applicata* **45,** no. 4, 271–284.

Boatwright, J., and E. Phillips (2013). Exploiting the demographics of “Did You Feel It?” responses to estimate the felt areas of moderate earthquakes, *Seismol. Res. Lett.* **84,** no. 1, 147, poster T1.

Boyd, A. (2001). *Broadcast Journalism: Techniques of Radio and Television News*, 5th Ed., Focal Press, Oxford, England. ISBN: 0-240-51571-4.

DeGroot, M. H., and M. J. Schervish (2012). *Probability and Statistics*, 4th Ed., Pearson, Boston, Massachusetts, ISBN: 978-0-321-50046-5.

Dewey, J. W., M. G. Hopper, D. J. Wald, V. Quitoriano, and E. R. Adams (2002). Intensity distribution and isoseismal maps for the Nisqually, Washington, earthquake of 28 February 2001, *U.S. Depart. Interior U. S. Geol. Surv. Open-File Rept. 02-346*, 60 pp.

Gasperini, P. (2001). The attenuation of seismic intensity in Italy: a bilinear shape indicates the dominance of deep phases at epicentral distances longer than 45 km, *Bull. Seismol. Soc. Am.* **91,** no. 4, 826–841, doi: 10.1785/0120000066.

Gómez Capera, A. A. (2006). Seismic hazard map for the Italian territory using macroseismic data, *Earth Sci. Res. J.* **10,** no. 2, 6790.

Gómez Capera, A. A., V. D’Amico, C. Meletti, A. Rovida, and D. Albarello (2010). Seismic hazard assessment in terms of macroseismic intensity in Italy: A critical analysis from the comparison of different computational procedures, *Bull. Seismol. Soc. Am.* **100,** no. 4, 1614–1631, doi: 10.1785/0120090212.

Grünthal, G. (2011). Earthquake, intensity, in *Encyclopedia of Solid Earth Geophysics, Encyclopedia of Earth Sciences*, H. K. Gupta (Editor), Springer, Dordrecht, The Netherlands, 237–242, ISBN: 978-90-481-8701-0.

Hough, S. E. (2013). Spatial variability of “Did You Feel it?” intensity data: Insights into sampling biases in historical earthquake intensity distributions, *Bull. Seismol. Soc. Am.* **103,** no. 5, 2767–2781, doi: 10.1785/0120120285.

International Press Institute (1953). *The Flow of the News*, International Press Institute, Zurich, Switzerland, ISBN: 0405047517.

Jolliffe, I. T., and D. B. Stephenson (Editors) (2012). *Forecast Verification—A Practitioner’s Guide in Atmospheric Science*, 2nd Ed., Wiley-Blackwell, Chichester, England, ISBN: 978-0-470-66071-3.

McGuire, R. K. (2004). *Seismic Hazard and Risk Analysis*, Earthquake engineering research institute, Oakland, California, ISBN: 0943198011.

Musson, R. M., and I. Cecić (2012). New Manual of Seismological Observatory Practice (NMSOP-2), IASPEI, in *Intensity and Intensity Scales*, chapter 12, GFZ German Research Centre for Geosciences, Potsdam, Germany.

Pasolini, C., P. Gasperini, D. Albarello, B. Lolli, and V. D'Amico (2008). The attenuation of seismic intensity in Italy, part I: Theoretical and empirical backgrounds, *Bull. Seismol. Soc. Am.* **98,** no. 2, 682–691, doi: 10.1785/0120070020.

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

Sbarra, P., P. Tosi, and V. De Rubeis (2010). Web-based macroseismic survey in Italy: Method validation and results, *Nat. Hazards* **54,** 563–581, doi: 10.1007/s11069-009-9488-7.

Wald, D., V. Quitoriano, L. Dengler, and J. Dewey (1999). Utilization of the Internet for rapid community intensity maps, *Seismol. Res. Lett.* **70,** 680–697, doi: 10.1785/gssrl.70.6.680.

Wald, D. J., V. Quitoriano, B. Worden, M. Hopper, and J. W. Dewey (2011). USGS “Did You Feel It?” Internet-based macroseismic intensity maps, *Ann. Geophys.* **54,** no. 6, 688–707, doi: 10.4401/ag-5354.

Wessel, P., W. H. F. Smith, R. Scharroo, J. F. Luis, and F. Wobbe (2013). Generic Mapping Tools: Improved version released, *Eos Trans. AGU* **94,** 409–410, doi: 10.1002/2013EO450001.

Worden, C. B., M. C. Gerstenberger, D. A. Rhoades, and D. J. Wald (2012). Probabilistic relationships between ground-motion parameters and modified Mercalli intensity in California, *Bull. Seismol. Soc. Am.* **102,** no. 1, 204–221, doi: 10.1785/0120110156.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith (2009). Mixed effects models and extensions in ecology with R, in *Statistics for Biology and Health*, Springer, New York, New York, ISBN: 978-0-387-87457-9.

Sum Mak
Danijel Schorlemmer
Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences
Helmholtzstraße 6
14467 Potsdam, Germany
smak@gfz-potsdam.de