Risk Analysis DOI: 10.1111/risa.12455

Statistical Modeling of Fire Occurrence Using Data from the Tōhoku, Japan Earthquake and Tsunami

Dana Anderson,[1] Rachel A. Davidson,[1,*] Keisuke Himoto,[2] and Charles Scawthorn[3]

In this article, we develop statistical models to predict the number and geographic distribution of fires caused by earthquake ground motion and tsunami inundation in Japan. Using new, uniquely large, and consistent data sets from the 2011 Tōhoku earthquake and tsunami, we fitted three types of models—generalized linear models (GLMs), generalized additive models (GAMs), and boosted regression trees (BRTs). This is the first time the latter two have been used in this application. A simple conceptual framework guided identification of candidate covariates. Models were then compared based on their out-of-sample predictive power, goodness of fit to the data, ease of implementation, and relative importance of the framework concepts. For the ground motion data set, we recommend a Poisson GAM; for the tsunami data set, a negative binomial (NB) GLM or NB GAM. The best models generate out-of-sample predictions of the total number of ignitions in the region within one or two. Prefecture-level prediction errors average approximately three. All models demonstrate predictive power far superior to four from the literature that were also tested. A nonlinear relationship is apparent between ignitions and ground motion, so for GLMs, which assume a linear response-covariate relationship, instrumental intensity was the preferred ground motion covariate because it captures part of that nonlinearity. Measures of commercial exposure were preferred over measures of residential exposure for both ground motion and tsunami ignition models. This may vary in other regions, but nevertheless highlights the value of testing alternative measures for each concept. Models with the best predictive power included two or three covariates.

KEY WORDS: Boosted regression tree; earthquake; fire; generalized additive model; generalized linear model

1. INTRODUCTION

The 2011 Tōhoku, Japan earthquake and tsunami caused at least 348 reported fires—more than any other earthquake in history. By comparison, approximately 285 were documented in Kobe, Japan (1995), 12 in Niigata, Japan (1964),[1] 110 in Northridge, California (1994), and 26 in Loma Prieta, California (1989).[2] The Tōhoku fires occurred in a variety of land area types from urban to rural, and were caused by two distinct hazards—ground motion and tsunami inundation. As they were all part of a single event, the data set describing the fires could be collected at one time, ensuring a consistency not possible when compiling data from multiple events over many years. Because of the size and features of the fire data set it generated, this event offers a unique opportunity to improve the statistical

[1]Department of Civil and Environmental Engineering, University of Delaware, Newark, DE, USA.
[2]Building Research Institute, Tachihara 1, Tsukuba, Ibaraki 305-0802, Japan.
[3]SPA Risk LLC, San Francisco, CA, USA and Kyoto University (retired), Kyoto, Japan.
*Address correspondence to Rachel A. Davidson, Department of Civil and Environmental Engineering, University of Delaware, Newark, DE 19716, USA; tel: +1-302-831-4952; rdavidso@udel.edu.

models of postearthquake ignitions that rely on such data and that are critical for planning for the emergency response needs and total losses that can result from such fires.

Previous efforts to model postearthquake ignitions fall into two main categories.[3] In the first,[4–7] the probabilities of different mechanisms of ignition (e.g., utility damage, overturning of objects) are estimated separately and combined using fault or event trees. In the second category, which dates back to Kawasumi,[8] statistical models are developed using data from past earthquakes. Although they differ in the data and the specific response variables and covariates they use, almost all previous statistical models have regressed some measure of ignition rate on a single measure of earthquake intensity (e.g., ignitions per sq ft of building area vs. peak ground acceleration, PGA), apparently using least squares regression.[9–11] Ren and Xie[12] estimate the number of ignitions in each geographic area unit as the product of ignition rate from the regression and total building area of the unit. Others[2,13,14] then simulate ignitions for each area unit assuming they follow a Poisson process with that product as the Poisson parameter. Cousins and Smith[15] assume ignitions are normally distributed with mean ignition rate from the regression and a standard deviation of 1. Davidson[16] introduced the use of generalized linear models (GLMs) and generalized linear mixed models (GLMMs) for this application for the first time. Unlike previous models, the approach uses discrete, nonnegative ignition counts as the response variable, examines many possible covariates, and uses a small unit of study to ensure homogeneity in variable values for each area unit. Nevertheless, the data in that analysis were only available for selected jurisdictions for each of six earthquakes, which made it impossible to fully capture the zero counts, i.e., the places where ground shaking was strong enough to cause ignitions but did not. The analysis also did not fully characterize the model's predictive ability with out-of-sample validation.

Using a new data set compiled for the Tōhoku earthquake and tsunami, this article offers two main contributions. First, we introduce for this application two additional model types—generalized additive models (GAMs) and boosted regression trees (BRTs)—and compare their performance with the GLMs found to be best in Davidson.[16] In the process, we also improve estimation of each model's predictive power, i.e., how well it will predict the number and locations of ignitions in a future earthquake and the magnitude and type of errors to expect. Second, we introduce and compare new models for ground-motion- and tsunami-generated postearthquake ignitions in Japan. The best ground motion models demonstrate predictive power far superior to those available in the literature. To our knowledge, no such model for tsunami-generated ignitions previously existed for any region. In Sections 2 and 3, we describe the data and the three types of statistical models, respectively. The model section process is discussed in Section 4, followed by the analysis results for the ground motion and tsunami ignition models, respectively, in Sections 5 and 6.

2. DATA DESCRIPTION

2.1. Data Compilation

The study area was defined to be the region in which ignitions were physically possible, which we assumed to be composed of the 17 prefectures that experienced at least one damaged building or fatality (not including Hokkaido) (Fig. 1). This area, which includes 126,768 km^2 and 54.7 million people (34% of the total area and 43% of the total population of Japan, respectively), corresponds approximately to the region that experienced PGA $\geq$ 0.035 g, based on data used in the analysis (Section 2.3). Municipalities were taken to be the area units for the analysis, but for the nine largest cities in the area that are administratively divided into ku's (Chiba, Hamamatsu, Kawasaki Sagamihara, Saitama, Sendai, Shizuoka, Tokyo, and Yokohama), we used the ku's instead. These area units were chosen because many variables are available for municipality/ku, and they are of similar size for analysis and small enough so that covariates are approximately homogeneous within them. Together, 701 municipalities and 56 ku's cover the entire study region. Because 72 municipalities and 6 ku's were missing ignition data (Section 2.2), the final data set included 629 municipalities and 50 ku's (Fig. 1).

Overlaying data from several sources in a geographic information system (GIS), we compiled a data set that includes a value for each variable in Table I for each area unit. For ground motion and inundation covariates, which varied spatially, we used the average over each area unit and over the inundated portion of each area unit, respectively. For analysis, because of the differences in ignition mechanisms, the data set was divided into the *ground*

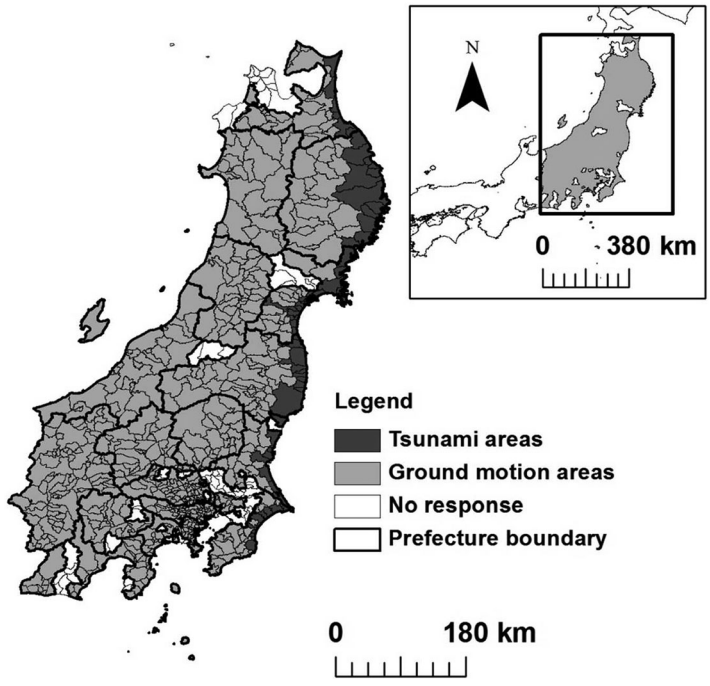

Fig. 1. Map of study area and its location in Japan (inset).

motion data set, which included the 615 area units with no tsunami inundation ($x_{area} = 0$) and only ignitions identified as ground motion generated; and the *tsunami data set* composed of the 64 area units with some inundation ($x_{area} > 0$) and only ignitions identified as tsunami generated. The two were analyzed separately. Anderson[17] includes the complete data sets.

2.2. Ignition Data

Ignition data were collected by the Committee for Postearthquake Fire Research of the Japan Association for Fire Science and Engineering (JAFSE), which included Himoto, co-author of this article. They sent hardcopy mail surveys to all 297 fire services in the 17 affected prefectures.[11] Surveys were mailed in April–May 2012, and reminders were sent in the end of 2012 to the beginning of 2013. In all, 258 surveys (87%) were returned complete. The survey included 12 open-ended questions asking the number of ignitions, and for each ignition, its location (street address), whether it was within the tsunami-inundated area or not, estimated occurrence time, reported time, apparent cause, fire type, consequent losses, and how firefighting activity was conducted.[18] Many of the questions were extracted from the official "Kasai Hokoku" fire report that fire services are required to submit to the larger regional agencies.

The 78 (10%) area units for which ignition data are missing appear to be randomly distributed geographically (Fig. 1), and have similar distributions for the covariates as the area units for which data are available.[17] With no reason to believe otherwise, we thus assume that the missing data are what is known as missing completely at random (MCAR), i.e., unrelated to the number of ignitions or to the covariate values.[19] The observed data can then be thought of as a random subsample of the hypothetically complete data, and omitting those area units from the analysis should not introduce bias in the results.

For consistency, we included only ignitions that (1) were identified as ground motion or tsunami generated (as determined by the JAFSE team based on the specified cause), and (2) occurred within 10 days of the earthquake (i.e., by 11:59 pm March 21).[11] The 10-day cutoff includes most reported earthquake- or tsunami-generated ignitions (84%), and could be considered the ignitions that would be of most interest to emergency responders and risk managers. The data set includes ignitions that self-extinguished or were extinguished by citizens, but

Table I. Definition, Mean, and Coefficient of Variation (COV) of Variables Used in Each Data Set

Concept		Variable Definition	Ground Motion Data Set		Tsunami Data Set	
			Mean	COV	Mean	COV
Ignitions	y_g	Ground motion ignitions in 10 days	0.2	3.3	–	–
	y_t	Tsunami ignitions in 10 days	–	–	1.9	2.0
Ground motion	x_{psa03}	Average PSA[a], 0.3 s (g)	0.4	0.9	0.8	0.5
	x_{psa10}	Average PSA, 1 s (g)	0.2	0.6	0.3	0.5
	x_{psa30}	Average PSA, 3 s (g)	0.1	0.5	0.1	0.4
	x_{pgv}	Peak ground velocity, PGV (cm/s)	18.8	0.5	31.5	0.3
	x_{pga}	Peak ground acceleration, PGA (g)	0.2	0.8	0.4	0.6
	x_{ii}	Instrumental intensity	6.1	0.2	7.4	0.1
Inundation	x_{area}	Area that experienced inundation[b] (m^2)	–	–	7,935,734	1.4
	x_{depth}	Average inundation depth[c] (m)	–	–	2.8	0.7
Exposure	x_{pop}	Population[b]	83,033	1.5	57,284	1.2
	x_{res}	Area of residential zoning[b] (1,000s m^2)	7,054	1.5	7,884	1.4
	x_{estab}	Number of business establishments[b]	3,846	1.5	2,698	1.3
	x_{com}	Area of commercial zoning[b] (1,000s m^2)	805	1.6	819	1.5
	x_{indus}	Area of industrial zoning[b] (1,000s m^2)	1,948	1.7	3,638	1.8
Vulnerability	x_{pwood}	% houses that are wooden	20.4%	1.0	28.6%	0.9
	x_{pdam3}	% houses collapsed	0.1%	6.9	4.2%	2.3
	x_{pdam2}	% houses with moderate damage	0.4%	6.4	2.0%	1.9
	x_{pdam1}	% houses with minor damage	2.0%	3.5	4.6%	1.8
	$x_{pdam123}$	% houses with at least minor damage	2.4%	3.5	10.8%	1.5
Damaged buildings	x_{dam3}	Num. collapsed houses[b]	10.2	11.1	572	2.5
	x_{dam2}	Num. houses with moderate damage[b]	82.1	11.2	883	4.6
	x_{dam1}	Num. houses with minor damage[b]	349	5.0	2,149	3.4
	x_{dam123}	Num. houses with at least minor damage[b]	441	6.1	3,605	3.3
	x_h	Num. houses[d]	35,534	1.5	24,847	1.3

[a]PSA = pseudo-spectral acceleration.
[b]We took the natural log of all exposure variables, damage variables, and x_{area} before using them in models.
[c]Averaged only over the portion of the area unit with nonzero inundation depth.
[d]Number of houses was used to compute x_{pwood} and to apply the models from the literature (Section 4.2).

not those that did not require firefighting from the beginning. The final ground motion data set included 528 (86%) area units with zero ignitions, 54 with one ignition, 23 with two, four with three, three with four, and one each with five, six, and 11 ignitions. The tsunami data set included 33 (52%) zero counts, 11 area units with one ignition, six with two, five with three, and nine with four or more (with a maximum of 22 ignitions in one area unit).

2.3. Covariate Data

The choice of candidate covariates was guided by a simple conceptual framework describing the causes of ignitions (Fig. 2), together with data availability. The main reported mechanisms of postearthquake fires in the absence of a tsunami include damage to electric power and gas lines, and overturning or falling of appliances, heating equipment, candles, or other contents.[2,7] Ignitions can be caused directly by building damage (e.g., damage to a wall may cause a water heater to overturn and ignite), but they can also occur in the absence of building damage if, for example, an appliance overturns due to acceleration in an otherwise undamaged building. In the case of tsunami inundation, additional ignition mechanisms include damage to liquefied propane gas cylinders for home heating, automobiles, and oil tanks.[20,21] All these types of ignitions ultimately are expected to be directly related to hazard, exposure (buildings, their contents, and the utilities and cars associated with them), and building vulnerability. We identified multiple possible covariates to represent each of these concepts (Fig. 2; Table I). Within a concept (e.g., ground motion), the candidate covariates listed are alternative measures of the same idea, and as such are highly correlated (i.e., $\rho \geq 0.75$).[17] The variance inflation factors (VIFs) are also quite high. For a negative binomial (NB) GLM with all

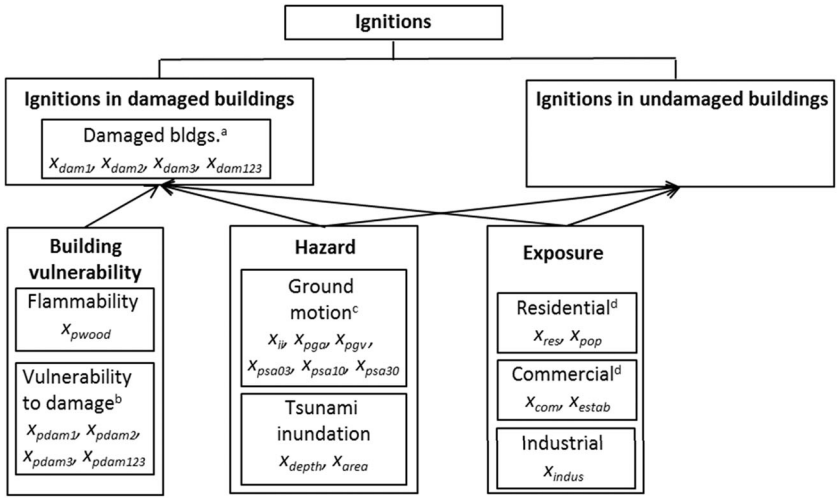

Fig. 2. Conceptual framework of ignitions and candidate covariates (variable definitions in Table I).

covariates, the average VIF across covariates is 474, and for all but two covariates, they are over 2.5. By contrast, in the models considered in the cross-validation study, in which no more than one covariate is chosen for each concept, the VIFs are all approximately 1.0, indicating no multicollinearity problems. Given that the covariates within a concept are alternative indicators of the same idea and are highly correlated, the intention is not to include all of them in a single model, but for each concept to choose the most promising covariate of the alternatives. Thus, we defined the following rules for inclusion of the covariates in a model: include at most one of the six ground motion covariates; at most one of the residential or commercial exposure covariates; at most one of the building damage covariates; and at most one of the percentage of building damage covariates. The framework also suggested investigation of interactions between hazard, exposure, and building vulnerability. This approach was developed to avoid blind data mining and multicolliearity problems, and to end up with a model that could be easily interpreted and that is not more complex than necessary to facilitate future application.

Data for the ground motion and tsunami inundation hazard covariates were found from the U.S. Geological Survey Shakemap archives[22] and Geospatial Information Authority of Japan,[23] respectively. Population and establishment data were collected from the Japanese Statistics Bureau.[24] Data on number of wooden buildings, number of houses (to normalize wooden buildings), and zoning were all collected from the Ministry of Land and Infrastructure.[25,26] The vulnerability to damage and damaged buildings information was found through the Fire and Disaster Management Agency.[27] To avoid numerical problems associated with highly skewed distributions, before fitting any models, we took the log of all exposure and damage covariates, as well as area of tsunami inundation x_{area} (e.g., using $lx_{res} = \ln(x_{res})$ instead of x_{res}).

3. STATISTICAL MODEL TYPES

The goal of this analysis was to develop a model to predict the expected number of ignitions in each area unit i as a function of the attributes of the area unit captured in the covariate vector $\vec{x}_i$. Three model types were considered—GLMs, GAMs, and BRTs. These were selected to represent examples of parametric, nonlinear, and nonparametric model types, respectively. The GLMs allow comparison with a previous similar analysis for California.[16] The GAMs extend those to allow the possibility of nonlinearity in the covariates, which was considered a potential issue. The BRTs provide an example of a different, nonparametric approach. Although other types of methods are certainly possible (e.g., random forest, multivariate adaptive regression splines, or support vector machines), these three provide a reasonable diversity of approaches given the nature of the problem. All models were fitted using *R* software version 3.0.2 using default settings except where noted.[28] Poisson GLMs, NB GLMs, GAMs, and BRTs were fitted using the *glm* {stats}, *glm.nb* {MASS

v7.3-29}, *gam* {mgcv v1.7-27}, and *gbm* {gbm v2.1-0.3} functions (in the noted package), respectively.[29–32]

3.1. Generalized Linear Models

GLMs are a generalization of ordinary linear regression that allow for response variables that are not normally distributed.[33] Poisson and NB GLMs are types of GLMs that are useful when the response variable represents nonnegative, integer count data, as in this analysis. In a Poisson GLM, the observations of the response variable y_i are assumed to be generated from a Poisson distribution, and the mean of the distribution, $\mu_i \equiv E[y_i|\vec{x}_i]$, is related to covariates $\vec{x}_i$ through Equation (1), where $\vec{\beta}$ is a vector of unknown parameters.[34] Thus, the parameter μ_i of the distribution (the estimated mean) varies by area unit i depending on the values of the covariates $\vec{x}_i$ for that area unit.

$$\ln(\mu_i) = \vec{x}_i^T \vec{\beta} = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ji} \qquad (1)$$

An NB GLM is similar except that the counts y_i are assumed to be generated from an NB distribution. A Poisson GLM includes the implicit assumption that, conditional on the covariates, the mean and variance are equal. In reality, data are often overdispersed (i.e., the conditional variance is greater than the conditional mean), and thus the Poisson assumption is not appropriate. In that case, an NB GLM may be preferred because in the NB model, the conditional mean is still μ_i, but the conditional variance is as defined in Equation (2), where $\alpha \geq 0$ is the overdispersion parameter. The NB also has a larger expected proportion of zero counts and a thicker right tail than the Poisson.[34]

$$\text{Var}[y_i|\vec{x}_i] = \mu_i + \alpha\mu_i^2 \qquad (2)$$

When $\alpha = 0$, the NB model reduces to the Poisson, so a larger estimated α value indicates NB is a more appropriate model than a Poisson. In this analysis, preliminary modeling suggested that the Poisson GLM was not appropriate, and thus we focused on NB GLM models.

3.2. Generalized Additive Models

GAMs are an extension of GLMs in which the only change is that the linear terms in Equation (1), $\beta_j x_{ji}$, are replaced by nonparametric smooth functions, $s_j(x_{ji})$.[30] Although different smooth functions can be selected, we used thin plate regression splines, which are the default in the {mgcv} package. The benefit of a GAM over a GLM is that it allows more flexibility in the dependence of the response on the covariates without requiring specification of detailed parametric relationships. The challenge is that GAMs require determining how to represent the smooth functions and how smooth they should be. If any smooth functions were allowed, maximum likelihood estimation would tend to estimate ones that overfit the data. To avoid this, {mgcv} uses penalized likelihood maximization, in which a penalty for each smooth function is added to the model (negative log) likelihood to penalize its "wiggliness" (or lack of smoothness). For each term, a smoothing parameter controls the tradeoff between smoothness and goodness fit. In {mgcv}, smoothing parameters are estimated automatically so as to minimize prediction error, as measured with an un-biased risk estimator (UBRE) criterion when the scale parameter is known (or GCV if the scale parameter is not known, as in the case of the quasi-Poisson or NB model). UBRE is a linear transformation of the Akaike information criteria (AIC).[30] Based on early results, we also set $\gamma = 2$ to avoid overfitting the smooths. The parameter $\gamma \geq 1$ is a factor that inflates the degrees of freedom in the UBRE or GCV score, thus encouraging a smoother model.[30]

3.3. Boosted Regression Trees

BRTs is a newer method that is fundamentally different than GLMs and GAMs because instead of aiming to fit a single best model to relate a response variable to a set of covariates, BRTs work by fitting and combining a large number of relatively simple models. We chose to try BRTs for this application because they are able to handle interactions and nonlinear relationships well, and at least in some cases, they have been shown to provide better predictions than GLMs and GAMs.[35] BRTs combine statistical and machine learning techniques through the use of two algorithms—classification and regression trees (CART) and boosting.[36,37] The former is used to develop individual models; the latter builds and combines collections of those individual models to optimize predictive performance.

CARTs is a nonparametric, rule-based classification technique that groups observations with similar values of the response variable using a binary recursive partitioning algorithm.[37,38] Starting with the entire data set, the algorithm selects a covariate to be the basis of the first split and a value of

that covariate to be the split point that defines the two groups. Similar binary partitions are applied recursively until a stopping criterion is reached, with splitting variables and split points determined each time so as to minimize prediction errors. To avoid overfitting, a single tree is often fitted by first growing a large tree, then pruning it by collapsing branches that have the least power to classify instances. The result is a tree that specifies how to group observations based on their values on each of the splitting variables. For regression trees (which are used in this analysis), the response for each observation is assumed to be the mean response of all observations in the same terminal node of the tree. Regression trees are intuitive, unaffected by changing covariate measurement scales, and insensitive to outliers or inclusion of extra covariates, but they exhibit relatively poor predictive performance. Combining CARTs with boosting aims to overcome that limitation. Boosting is an iterative, forward, stage-wise procedure. In the first step, a regression tree is fitted to the data to minimize a selected loss function (deviance is used in the {gbm} package). In each subsequent step, a tree is fitted to the residuals of the previous trees and the new tree is added to all the previous trees (which are left unchanged). The final BRT model is a linear combination of many regression trees (usually thousands).

Because our response variable, number of ignitions in area unit i, is count data, we specified the loss function to be the Poisson deviance.[32] Although the variance of the response is greater than the mean as assumed by a Poisson (2.7 times greater for the ground motion data set), Poisson deviance is the most appropriate loss function for count data in BRTs and thus is typically used.[31,32] We set the bag fraction to the default value of 0.5, meaning that at each step, a randomly selected 50% of the training set data is used. This typically improves speed and accuracy.[39] Fitting BRTs requires setting three main parameters: (1) learning rate or shrinkage lr, which determines the contribution of each tree to the model; (2) tree complexity or interaction depth tc, which indicates the number of splits used for each tree; and (3) number of trees or iteration nt, which indicates the maximum number of trees used.

4. MODEL SELECTION PROCESS

The aim of the selection process was to identify and compare the best models based on predictive power, fit to the sample data, and ease of use. There were two main steps: (1) preliminary analyses for each model type, and (2) cross-validation analysis to compare the best models of each type, plus a few others of interest.

4.1. Preliminary Analyses for Each Model Type

For the GLMs, we first selected a covariate to represent each concept in the framework (Fig. 2). Because we wanted to include at most one of the six candidate ground motion covariates (see rules in Section 2.3), for example, we first compared models that were the same except for the covariate used to represent ground motion. We then chose the most promising models considering only that smaller set of covariates and including the possibility of interactions suggested by the framework (hazard covariate*exposure covariate, or hazard*exposure*building vulnerability covariates). In all cases, GLMs were compared based on the following goodness-of-fit metrics: (1) deviance pseudo-R^2, R^2_{dev}; (2) pseudo- R^2 based on α, R^2_α; and (3) Akaike information criteria, AIC. In linear regression models, R^2 represents goodness-of-fit as the percentage of variability in the observation y_i that a model explains. Alternative pseudo-R^2 statistics have been developed for nonlinear and discrete count models. The deviance pseudo-R^2 is defined as $R^2_{dev} = 1 - [D(y, \hat{\mu})/D(y, \bar{y})]$, where $D(y, \hat{\mu})$ is the deviance for the fit model and $D(y, \bar{y})$ is the deviance for the intercept-only model (a model with $\hat{\mu}_i = \bar{y}$ for all i). It measures reduction in deviance due to the inclusion of covariates, and generally increases with the addition of new covariates, but cannot be used to compare Poisson and NB models.[40] For NB models, $R^2_\alpha = 1 - (\alpha/\alpha_0)$, where α and α_0 are the overdispersion parameters (Equation (2)) from the fit model and the intercept-only model, respectively, is an appropriate measure as well. Although both R^2_{dev} and R^2_α are always from zero to one, the former measures the fraction of total variation in the raw counts y_i explained by the model, and the latter measures the fraction of the variation in the true Poisson means explained by the model. When the mean count μ_i is small (as in this case), most of the variability in the counts y_i can be due to the Poisson variability, and thus R^2_{dev} are much lower than R^2_α. The AIC is defined as $\text{AIC} = -2\log L + 2q$, where $\log L$ is log-likelihood, q is number of independent parameters, and a smaller AIC indicates the preferred model.[41] Unlike the pseudo-R^2 metrics, AIC penalizes more complicated models. With an aim to fit the data well without overfitting them,

we also selected the promising GLMs based on a preference for simpler models when two fit the data equally well. In some cases, models were included for purposes of comparing with other model types using the same terms. A similar process was followed with the GAMs, but using the UBRE/GCV as an additional goodness-of-fit metric (Section 3.2).

For the BRTs, the preliminary analysis began with the smaller sets of covariates (one for each concept) determined for the GLMs and GAMs, and then focused on determining appropriate values of the three main modeling parameters—learning rate *lr*, tree complexity *tc*, and maximum number of trees *nt*. As suggested in Elith *et al.*,[36] we used cross-validation with deviance reduction as the measure of success to compare all combinations of values of learning rate (0.01, 0.005, 0.001), tree complexity (1, 2, 3, 5), and maximum number of trees (15,000), and thus choose the optimal ones for our data sets.[17] Based on that analysis, we used $lr = 0.001$, $tc = 2$ or 3, and $nt = 15{,}000$ for all ground motion runs; and $lr = 0.001$ or 0.0005, $tc = 3$, and $nt = 15{,}000$ for all tsunami runs. However, because the BRT can overfit the data when too many trees are included, within each run, the 10-fold cross-validation option of *gbm* was used to identify the optimal number of trees. Note that this cross-validation exercise to determine *lr*, *tc*, and *nt* is separate from that described in Section 4.2.

4.2. Cross-Validation Analysis

The best models of each of the three types were then compared, plus a few additional models of interest. To compare predictive power of the models, we conducted a 10-fold cross- validation (CV). The observations were partitioned into 10 randomly sampled folds. For each fold, the 90% of observations not in the fold comprised the training set. They were used to fit the models, which were then applied to predict values for each of the 10% of observations in the fold—the validation set. In this way, we developed out-of-sample predictions of ignition rate $\hat{\mu}_i$ for each observation, and estimates of mean absolute error (*MAE*), square root of the mean squared error (*RMSE*), mean error (*ME*), and error in the expected total number of ignitions for the region (*TE*) and for each prefecture *P* (TE_P) Equations (3)–(7), where y_i is the observed number of ignitions in area unit *i* and n_P is the set of observations *i* in prefecture *P*. To minimize the effect of the particular fold sample, we repeated the cross- validation 140 times, each with a different set of randomly generated folds and averaged the resulting 140 estimates of each error metric (Equations (3)–(7)) for each model.

$$MAE = \frac{1}{n}\sum_{i}^{n} |y_i - \hat{\mu}_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} (y_i - \hat{\mu}_i)^2} \tag{4}$$

$$ME = \frac{1}{n}\sum_{i}^{n} (y_i - \hat{\mu}_i) \tag{5}$$

$$TE = \sum_{i}^{n} y_i - \sum_{i}^{n} \hat{\mu}_i \tag{6}$$

$$TE_P = \sum_{i \in n_P} y_i - \sum_{i \in n_P} \hat{\mu}_i \; \forall P \tag{7}$$

For the ground motion data set, we also compared our models to four from the literature—Kawasumi, Mizuno, HAZUS, and Zhao. Using the form of each model presented in Zhao *et al.*,[14] we refit the models with our ground motion data set and computed the predictive errors. For Kawasumi, $\ln(y_g/x_h)$ is a linear function of $\ln(x_{pdam3})$. For Mizuno, $\ln(-\ln(1-(y_g/x_h)))$ is a linear function of $\ln(1-x_{pdam3})$. For HAZUS and Zhao, y_g/x_h are second-order polynomial and linear functions of x_{pga}, respectively. These analyses had to be done using the larger prefecture as the area unit rather than municipality/ku, or there would be too many zeros. As a result, there were only 17 observations and we did leave-one-out rather than 10-fold cross-validation.

5. GROUND MOTION RESULTS

Each model allows prediction of the expected number of ignitions $\hat{\mu}_i$ for each area unit *i* (and $\hat{\alpha}$ for NB models) as a function of the attributes of the hazard intensity and built environment in the area unit. The $\hat{\mu}_i$ (and $\hat{\alpha}$) defines the Poisson (or NB) distribution that describes occurrence of ignitions in that area unit. Thus, each model can be used as a predictive tool, and to understand the relative importance of variables contributing to ignitions. With these uses in mind, we first discuss the results of the preliminary analyses, and then compare the models according to: (1) predictive power (i.e., ability to predict both total number of ignitions for the whole region and

geographic distribution of ignitions); (2) goodness-of-fit to the data; and (3) relative importance of covariates. Finally, we discuss recommended models and their application as predictive tools.

5.1. Preliminary Analysis Results

For the GLMs, in the first step, instrumental intensity (x_{ii}) and area of commercial zoning (x_{com}) were clearly the preferred measures of ground motion intensity and residential/commercial exposure, respectively. The best NB GLM with x_{pga} underestimated the total count by $TE = 18.2$ compared with $TE = 4.1$ for the best NB GLM with x_{ii} (Table II). For GAMs, instrumental intensity (x_{ii}) and PGA (x_{pga}) were both promising measures of ground motion intensity, and area of commercial zoning (x_{com}) and number of business establishments (x_{estab}) were both promising measures of residential/commercial exposure. A nonlinear relationship is apparent between ignition rate and ground motion with the marginal increase in ignition rate declining with ground motion (e.g., Fig. 6). Because the GLM cannot capture that nonlinearity, instrumental intensity x_{ii}, which is defined to be more directly related to damage and has a nonlinear relation with PGA, provides a better fit to the data.[42] Because GAMs can capture the nonlinearity, the choice between x_{ii} and x_{pga} is not as pronounced. The preference of an indicator of commercial exposure (x_{com} or x_{estab}) over residential exposure (population, x_{pop}, or area of residential zoning, x_{res}) is also notable because one might expect the latter to be preferred as most postearthquake ignitions occur in residential buildings,[2] and x_{pop} would be easier data to collect for application of the model. Although they are highly correlated with x_{pop} ($\rho = 0.8$), it appears that x_{com} or x_{estab} is consistently preferred at least in part because of three observations in Tokyo (Chiyoda, Minato, and Shinjuku), which have multiple ignitions (2, 7, and 1, respectively) and higher values of x_{com} and x_{estab} than expected given the population. Thus, x_{com} and x_{estab} are better able to predict those counts than x_{pop}. In other regions, x_{pop} may be equally appropriate.

In both GLMs and GAMs, although the choice of covariates to represent the concepts of damaged buildings and vulnerability of buildings to damage did not matter greatly to the model fit, based on comparison, number of houses with minor damage (x_{dam1}) and percentage of houses that collapsed (x_{pdam3}), respectively, were selected. In the second step, we determined that the area of industrial zoning (x_{indus}), building damage (x_{dam1}), and building vulnerability to damage (x_{pdam3}) covariates did not contribute substantially to the model goodness-of-fit, and thus were not considered in the final models.

5.2. Predictive Power

Table II compares the most promising of each model type according to five measures of predictive power (defined in Section 4.2): $RMSE$, MAE, ME, TE, and $E[|TE_P|]$. Comparing the $RMSE$, MAE, and ME across models suggests that the average errors ($RMSE$ and MAE) are slightly smaller for BRTs than GLMs and GAMs, but the BRT predictions are biased a bit low (with a mean error of 0.03 for the models in Table II).

To gain more insight into the practical significance of the errors, it is useful to examine the models' ability to predict the total numbers of ignitions, both for the entire study region and to get an idea of the models' ability to capture the geographic distribution—the total for each prefecture. Table II shows that although the BRTs had the best $RMSE$ and MAE values, they underestimate the total number of ignitions by almost 20. The Poisson GAMs, on the other hand, are within one ignition of the observed total of 147. Remembering that these are out-of-sample predictions, the GLMs and GAMs provide results that should be accurate enough to be of practical use. All four models from the literature provide substantially worse predictions, with errors that are two to six times larger (Table II).

Fig. 3 summarizes each model's predictive errors by prefecture, TE_P. Table II includes the average of absolute ignition count errors across prefectures (almost all are between 2 and 3). These results suggest that although the number of ignitions for the region is predicted quite well, there is some error in geographic distribution. In particular, the GLMs have somewhat higher prefecture errors than the other model types, and the BRTs underestimate prefecture counts more than overestimate. The largest errors are underestimation in Miyagi and Iwate prefectures, and overestimation in Tochigi.

5.3. Goodness of Fit

Because the models are intended for use in prediction, predictive power is arguably more important than goodness of fit, and although we seek models that fit the sample data well, we do not want to

Table II. Summary Comparison of Selected Most Promising Models of Each Model Type, for Ground Motion Data Set

	Model	Root Mean Squared Error (*RMSE*)	Mean Absolute Error (*MAE*)	Mean Error (*ME*)	Total Error (*TE*)	Total Error %	Average of Total Absolute Prefecture Errors $E[\lvert TE_p \rvert]$	α	R^2_α	R^2_{dev}	*UBRE*/*GCV*	*AIC*	Moran's *I p*-Value	Formula
1	NB GLM	0.689	0.301	0.01	4.1	3%	2.8	0.99	0.82	0.43	–	557	0.36	$x_{ii} + lx_{com}{:}x_{ii}$
2	NB GLM	0.692	0.302	0.01	4.6	3%	2.8	1.00	0.82	0.43	0	557	0.24	$x_{ii} + lx_{com}$
3	NB GLM	0.696	0.303	0.01	4.8	3%	2.9	0.99	0.82	0.43	–	559	0.28	$x_{ii} + lx_{com} + x_{pwood}$
4	P GAM	0.688	0.298	0.00	0.3	0%	2.5	–	–	0.46	–0.39	564	0.97	$s(x_{ii}) + s(lx_{com})$
5	P GAM	0.691	0.298	0.00	0.3	0%	2.4	–	–	0.44	–0.35	579	0.57	$s(x_{ii}, lx_{com}, x_{pwood})$
6	P GAM	0.693	0.299	0.00	0.6	0%	2.1	–	–	0.49	–0.38	553	0.82	$s(x_{ii}) + s(lx_{com}) + s(x_{pwood})$
7	NB GAM	0.670	0.289	0.00	0.3	0%	2.5	0.51	0.90	0.49	–0.50	542	0.61	$s(x_{pga}) + s(lx_{estab}) + s(x_{pwood}) + s(x_{pga}, lx_{estab}) + s(x_{pga}, lx_{estab}, x_{pwood})$
8	NB GAM	0.680	0.291	–0.01	–3.3	–2%	2.1	0.51	0.90	0.48	–0.53	536	0.17	$s(x_{pga}) + s(lx_{estab})$
9	NB GAM	0.688	0.298	0.00	1.6	1%	2.4	0.84	0.83	0.46	–0.51	557	0.26	$s(x_{ii}, lx_{com}, x_{pwood})$
10	BRT	0.663	0.288	0.03	18.8	13%	3.2	–	–	0.63	–	–	–	$x_{pga} + lx_{com} + x_{pwood} + lx_{indus}$
11	BRT	0.652	0.278	0.03	19.1	13%	2.4	–	–	0.62	–	–	–	$x_{pga} + lx_{estab} + lx_{indus} + x_{pdam3} + x_{pwood} + lx_{dam1}$
12	BRT	0.682	0.287	0.03	19.6	13%	2.6	–	–	0.63	–	–	–	$x_{ii} + lx_{com} + x_{pwood} + lx_{indus}$
13	HAZUS	8.7	5.6	–2.6	–44.4	–30%	5.6	–	–	–	–	–	–	x_{pga} (2nd order polynomial)
14	Zhao	9.1	5.9	–3.2	–55.2	–38%	5.9	–	–	–	–	–	–	x_{pga} (linear)
15	Kawasumi	16.8	12.4	–7.3	–124.5	–85%	12.4	–	–	–	–	–	–	$ln(x_{pdam3})$ (linear)
16	Mizuno	15.8	11.1	–7.6	–129.9	–88%	11.1	–	–	–	–	–	–	$ln(1–x_{pdam3})$ (linear)

Notes: For all smooths shown, k=20. For all BRTs shown, tc=2, lr=0.001, and nt=15,000. Observed total number of ignitions=147. Six left-most columns computed in cross-validation. Six right-most columns computed based on fitting model with the full data set. R^2_{dev} for Poisson and NB models are not directly comparable.

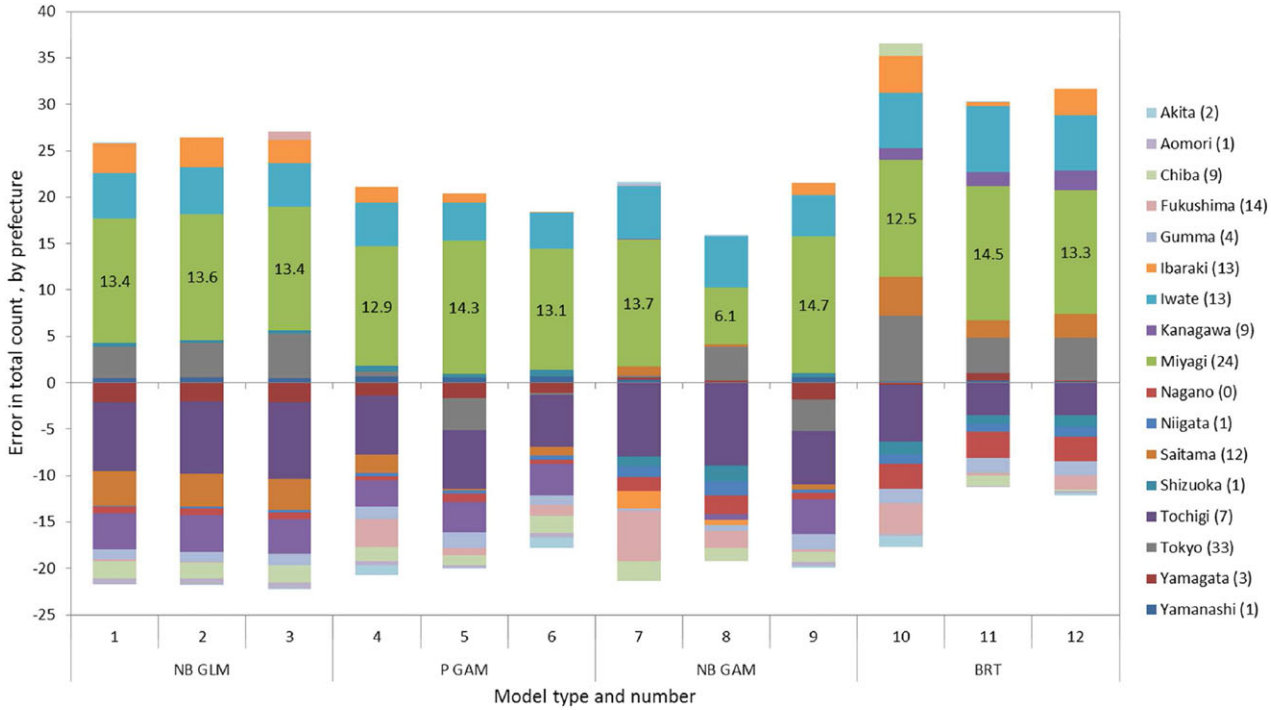

Fig. 3. Errors in total number of ignitions TE_p (observed-predicted) by prefecture and model, for ground motion models. Labeled column layers are for Miyagi prefecture. Each prefecture name includes the total number of observed ground motion ignitions.

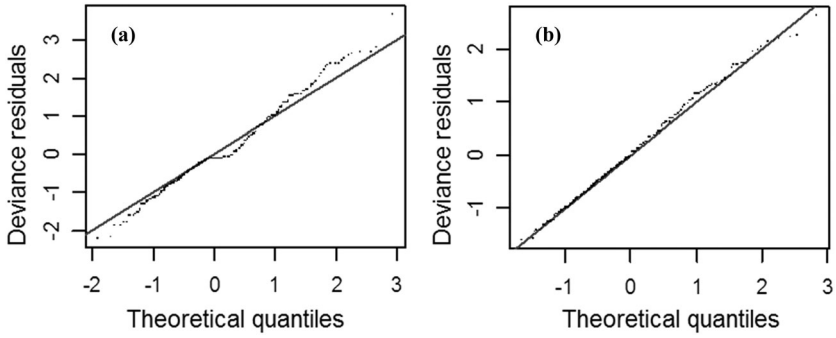

Fig. 4. *qq* plots of deviance residuals for (a) Poisson GAM Model 5 and (b) NB GAM Model 9, for ground motion data set.

overfit them. Nevertheless, goodness-of-fit measures were useful in the preliminary analyses to identify promising models for the CV analysis and can help identify model misspecification; thus we include them (computed using the full data set) in Table II. First, we consider the assumption of a Poisson versus NB distribution for the ignition counts. A quantile–quantile (*qq*) plot of deviance residuals provides an indication of the appropriateness of the assumed distribution. The closer the residuals are to the 45° line, the more appropriate the model. As an example, Fig. 4 shows *qq* plots for Models 5 and 9, which are the same except the former is a Poisson GAM and the latter is an NB GAM. It suggests that the NB is a more appropriate assumption in this case, and the same conclusion holds for most similar models we tested. The overdispersion parameter values $\alpha > 0$ suggest some preference for an NB model as well. Nevertheless, the predictive errors are larger for the NB GAM ($TE = 4.5$ compared with 0.3 for the Poisson).

Second, in this analysis, we implicitly assume that conditional on the covariate values, the observations are independent. Because the observations are actually derived from spatial data, we check the reasonableness of that assumption by determining if there

is spatial correlation in the residuals. If there is, that would suggest we should use a more complicated model that assumes spatially correlated residuals.[43] For each model, we used the package {ape} in *R* to compute the Moran's *I* for the residuals.[44] Moran's *I* is a commonly used measure of spatial autocorrelation. For each model, the *p*-value of the null hypothesis that there is no spatial autocorrelation is included in Table II, showing in each case no evidence to suggest rejecting that hypothesis and moving to a more complicated model.

5.4. Relative Importance of Covariates

Refitting each model with the full set of data, we can examine how they compare in terms of the relative importance they assign to each concept in the framework—ground motion, exposure, vulnerability, and damage. For the BRTs, the {gbm} package provides a measure of the relative importance of each covariate. It is "based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees."[36] It is then scaled so the contributions of all covariates sum to 100, with higher numbers indicating greater contributions. For each GLM and GAM, we estimated the relative importance of each covariate by computing the range of the response $\hat{\mu}$ over the observed range of the covariate, then normalizing those values so that the normalized values over all covariates sum to 100. For the same models in Table II, Fig. 5 shows the relative importance of each concept by model (when there were multiple covariates for a concept, we added their contributions). It shows that ground motion and exposure are most important, with vulnerability covariates never exceeding a contribution of 14 out of 100, and damage not contributing to most models. For comparison, we included some models with only a ground motion covariate in the CV analysis, but found that although they could estimate the total ignition count well, they did not capture the geographic distribution well. For example, a Poisson GAM with only the covariate x_{ii} had $TE = -0.2$, but the average absolute value of the prefecture error was $E[|TE_P|] = 4.2$ with much higher errors for Tokyo, for example, which experienced relatively low ground motion intensity but has almost twice the exposure as the next largest prefecture. Comparing Models 4 and 6 (Poisson GAMs without and with x_{pwood} included) suggests that including the vulnerability covariate x_{pwood} can reduce prefecture errors from $E[|TE_P|] = 2.5$ to 2.1. Looking at the smooth for x_{pwood} in Model 6, however, shows that it is not monotonically increasing, but rather "wiggles" up and down, which has no ready interpretation and may indicate overfitting. The results also show that the NB GLM assigns a much higher contribution to ground motion than exposure, whereas the reverse is true for the GAMs and BRTs. This is likely related to the inability of the NB GLM to capture the nonlinearity in the ground motion intensity.

5.5. Recommended Models

The results show that the best specific model depends to some extent on what one considers the most important criteria. No one model is best across all metrics and the models differ in ease of use as well. Although all model types are intended to be used in a predictive mode, it is more straightforward for GLMs, which result in a closed- form expression, than GAMs or BRTs, which do not. Some specific models also require data for more covariates or for covariates that are more difficult to measure consistently (e.g., percentage of buildings damaged). Nevertheless, taken together, the results suggest that a GAM provides the best predictive power and Model 4 would be a reasonable choice. It achieves low predictive errors (0.3 in the total number of ignitions and an average of 2.5 for each prefecture). It is a relatively simple model with just two covariates, and the smooths for Model 4 (Fig. 6) make sense intuitively, with larger values of ground motion intensity and exposure leading monotonically to higher ignition rates. Model 4 would be straightforward to implement in that it requires data for only two covariates that are relatively easy to find. If one required the closed- form equation provided by an NB GLM, Model 1 would be a reasonable choice, with final form as given in Equation (8).

$$ln(\mu_i) = -10.075 + 0.874x_{ii} + 0.066x_{ii}\ln(x_{com}) \quad (8)$$

5.6. Future Application of Models

The selected model can be used to compute the expected number of ignitions $\hat{\mu}_i$ for each area unit i, and those $\hat{\mu}_i$ can then be used for prediction for a specified historical or hypothetical earthquake scenario. To obtain the expected number of ignitions for the total area or a specified prefecture, one can just sum the expected number of area unit ignitions

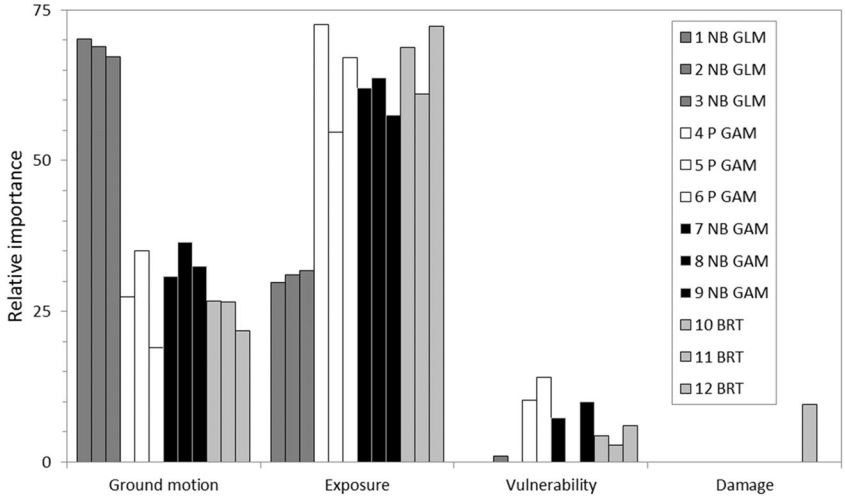

Fig. 5. Relative contribution of variable representing each concept by model type and number, for ground motion data set.

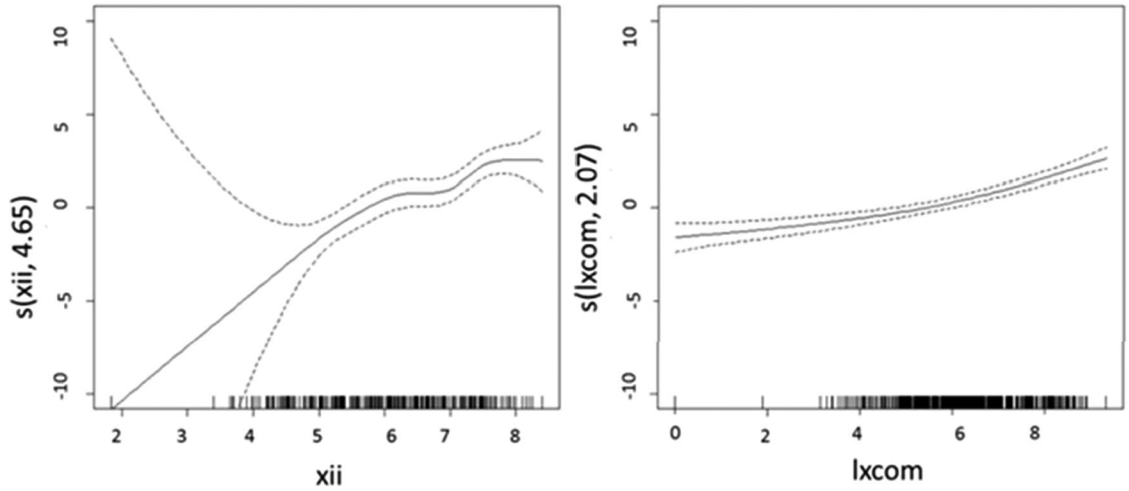

Fig. 6. Smooths for P GAM Model 4, for ground motion data set (y-axis labels include the estimated degrees of freedom of the smooth).

$\hat{\mu}_i$ as we did in this study (Equations (6) and (7)). One can also use the Poisson or NB distribution with those estimated parameters to simulate many (say, 1,000) realizations of ignition maps, each of which can be used as input for fire spread models to estimate damage and losses. Summing the ignitions in each map allows one to make a histogram of total number of ignitions, providing a description of the variability around the expected total regional and prefecture ignition counts. Fig. 7 shows an example of two such histograms developed for Models 5 and 9 using the out-of-sample predictions for the Tōhoku earthquake (and common random numbers to reduce sampling variability when comparing models). Note that although both are centered on the mean of 147 ignitions (the observed value), there is still a great deal of variability around that expected value, so any one of the thousand realizations of ignition maps from the simulation may have many fewer or more than 147. Note also that, as expected, the assumption of an NB distribution results in greater variability than Poisson (Equation (2)). This is consistent with the R^2_{dev} and R^2_{α} values (Table II), which suggest that although a large proportion of the variation in the true Poisson means is explained by the models (high R^2_{α} values), a much smaller proportion of the variation in the raw ignition counts y_i is (lower R^2_{dev} values).

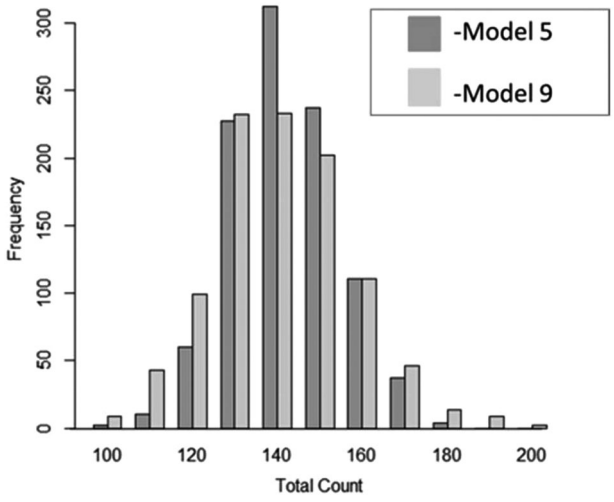

Fig. 7. Histograms of total number of ignitions simulated using out-of-sample predictions from Models 5 (Poisson GAM) and 9 (NB GAM), for ground motion data set.

6. TSUNAMI RESULTS

For the tsunami data set, the preliminary GLM and GAM analyses suggested that x_{pga}, x_{pgv}, and x_{ii} were candidates for best ground motion covariate. Both $ln(x_{area})$ and x_{depth} were potentially useful covariates. For exposure, $ln(x_{estab})$ again appeared to give the best fits, although for GAMs, the superiority was less pronounced than for the ground motion data set, so $ln(x_{pop})$ was considered as well. The preference for commercial covariates was likely due to the fact that a few area units had no residential zoning and very small population. Of the damage and vulnerability covariates, only x_{pdam3} appeared likely to improve the fit substantially and was considered in the CV analysis.

Table III compares some of the most promising models of each model type for the tsunami-generated ignition data. Again, although the BRT average prediction errors (*RMSE* and *MAE*) are comparably small, they are biased low compared with the other model types with $ME > 0.2$. In terms of error in the predicted total number of ignitions for the region, again, the BRTs are the worst, underestimating the total number of 119 by at least 15 ignitions (12%). Interestingly, although Poisson GAMs had excellent predictive power for the ground motion models, they perform relatively poorly for tsunami models. The two P GAMs shown in Table III are the best of the 13 included in the CV analysis, and they still overestimate the total number of ignitions by 7.1 and 10.5 ignitions, and have the highest prefecture-level errors (see $E[|TE_p|]$). Qq plots of deviance residuals and α values (Table II) provide further evidence that the NB distribution is a better fit for the tsunami ignitions than the Poisson distribution.[17] Comparing NB GLMs and NB GAMs suggests comparable performance in terms of predictive power, so the simpler NB GLMs are preferred. For all model types, the Moran's *I* values do not provide evidence for spatial correlation in the residuals for any models.

Fig. 8, which shows the errors in prefecture-level ignition counts, TE_p, indicates that as with the ground motion models, the tsunami models underestimate the number of ignitions in Miyagi and Iwate prefectures, and overestimate the number in Fukushima. This suggests the errors are not due to misidentifying some ignitions as tsunami versus ground motion generated, but rather that there may be additional covariates that are important for determining the number of ignitions that are not captured in these models. Some ignitions in Fukushima also may not have been reported because many in that area evacuated due to the nuclear power plant threat. BRTs again tend to underestimate more than overestimate prefecture-level ignition totals.

Finally, Fig. 9 shows the relative importance of the concepts in each tsunami model. In the best model types (NB GLMs and NB GAMs), inundation covariates are most important; followed by ground motion covariates, which are about half as important; then exposure. This is in contrast to the ground motion models, in which exposure was the most important concept (Fig. 5). Comparing Models 1–4, for example, suggests that while including x_{depth} and $\ln(x_{estab})$ reduce the *RMSE* and *MAE*, they do not improve the prefecture-level errors.

Again, more than one model is promising depending on the relative importance afforded the different criteria. Nevertheless, Model 1 could be one reasonable choice, offering a relatively simple model—an NB GLM with just two covariates (x_{pga} and $\ln(x_{area})$), and relatively small predictive errors (Equation (9)).

$$ln(\mu_i) = -11.556 + 1.680x_{pga} + 0.726\ln(x_{area}) \quad (9)$$

The tsunami models can be used for predicting the number and geographic distribution of ignitions in any future, historical, or hypothetical tsunami. Application of the models in a predictive mode works just as it does for the ground motion models, as described in Section 5.6.

Table III. Summary Comparison of Selected Most Promising Models of Each Model Type, for Tsunami Data Set

	Model	Root Mean Squared Error (*RMSE*)	Mean Absolute Error (*MAE*)	Mean Error (*ME*)	Total Error (*TE*)	Total Error %	Average of Total Errors $E[\lvert TE_p \rvert]$ Errors $E[\lvert TE_p \rvert]$	α	$R^2{}_{\alpha}$	$R^2{}_{dev}$	*UBRE/GCV*	*AIC*	Moran's Value	Formula
1	NB GLM	3.11	1.81	–0.02	–1.4	–1%	3.5	1.08	0.62	0.44	–	199	0.54	$x_{pga} + lx_{area}$
2	NB GLM	3.11	1.82	0.00	–0.2	0%	3.5	1.04	0.63	0.45	–	200	0.63	$x_{pga} + lx_{area} + x_{depth}$
3	NB GLM	2.99	1.72	0.02	1.3	1%	3.7	0.92	0.67	0.47	–	200	0.55	$x_{pga} + lx_{area} + x_{depth} + lx_{estab}$
4	NB GLM	3.04	1.77	0.01	0.5	0%	3.8	1.02	0.64	0.44	–	200	0.53	$x_{pga} + lx_{area} + lx_{estab}$
5	P GAM	3.72	2.03	–0.11	–7.1	–6%	9.6	–	–	0.47	1.691	248	0.53	$s(lx_{area}) + s(lx_{estab})$
6	P GAM	3.76	2.02	–0.16	–10.5	–9%	10.1	–	–	0.53	1.500	235	0.62	$s(lx_{area}) + s(lx_{pop})$
7	NB GAM	3.38	1.91	0.01	0.7	1%	7.8	1.11	0.61	0.43	0.115	199	0.86	$s(lx_{area}) + s(lx_{estab})$
8	NB GAM	3.92	1.96	–0.03	–2.0	–2%	3.3	0.98	0.65	0.46	0.110	196	0.56	$s(x_{pga}) + s(lx_{area})$
9	NB GAM	3.85	1.92	0.01	0.9	1%	3.8	0.90	0.68	0.47	0.192	197	0.54	$s(x_{pga}) + s(lx_{area}) + s(lx_{estab})$
10	NB GAM	3.23	1.87	0.09	5.5	5%	4.9	0.83	0.71	0.48	0.234	197	0.56	$s(x_{pga}) + s(lx_{area}) + s(lx_{pop})$
11	BRT	3.46	1.82	0.24	15.1	12.6%	6.4	–	–	0.78	–	–	–	$x_{pgv} + lx_{area} + x_{depth} + lx_{estab} + x_{pdam3} + lx_{dam1}$
12	BRT	3.32	1.76	0.27	17.1	14.4%	6.1	–	–	0.79	–	–	–	$x_{pga} + lx_{area} + x_{depth} + lx_{estab} + x_{pdam3} + lx_{dam1}$

Notes: For all smooths shown, k=5. For all BRTs shown, tc=3, lr=0.001, and nt=15,000. Observed total number of ignitions=119. Six left-most columns computed in cross-validation. Six right-most columns computed based on fitting model with the full data set. $R^2{}_{dev}$ for Poisson and NB models are not directly comparable.

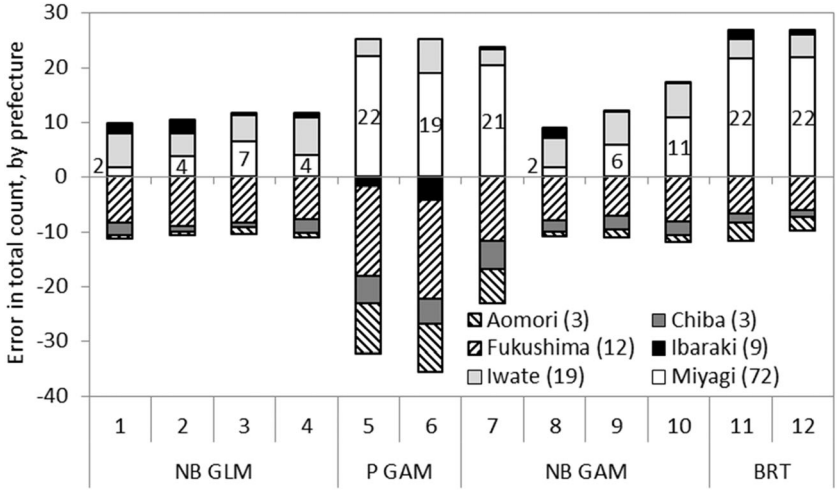

Fig. 8. Errors in total number of ignitions TE_P (observed-predicted) by prefecture and model, for tsunami data set. Labeled column layers are for Miyagi prefecture. Each prefecture name includes the total number of observed tsunami-ignitions.

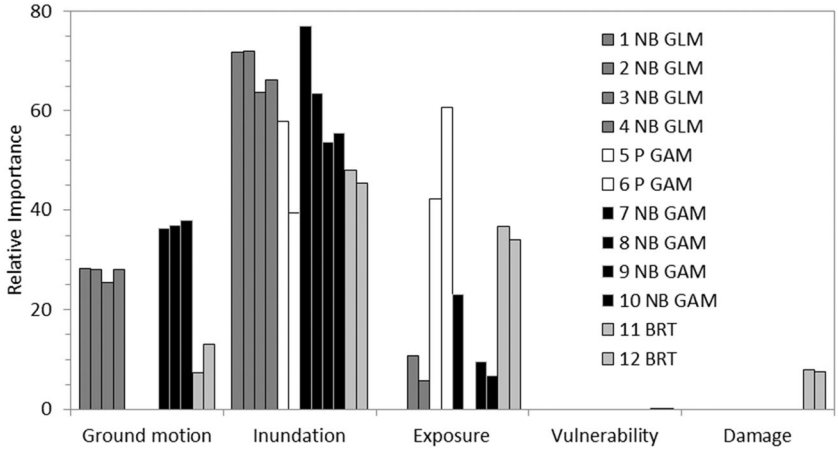

Fig. 9. Relative contribution of variable representing each concept by model type and number, for tsunami data set.

7. CONCLUSIONS

In this article, we developed new ignition models using data from the Tōhoku earthquake and tsunami. We compared three different model types (GLMs, GAMs, and BRTs) for two data sets (ground motion- and tsunami-generated ignitions). Results of a cross-validation analysis show all models demonstrate predictive power far superior to those that were tested from the literature. In general, the BRTs result in small mean errors, but are not as good as the other model types in predicting the total number of ignitions for the region. For the ground motion data set, a GAM is recommended; for the tsunami data set, an NB GLM or NB GAM is preferred. Out-of-sample predictions by the best models predicted the total number of ignitions in the region within one or two. At the prefecture level, however, errors were greater (approximately three on average), underpredicting the number in Miyagi and Iwate prefectures in both cases, overpredicting in Tochigi for the ground motion data set and Fukushima for the tsunami data set. The analysis suggests exposure then ground motion had the greatest contributions in the ground motion ignition models; and inundation, then ground motion in the tsunami ignition models. A nonlinear relationship was apparent between ignitions and ground motion, so for GLMs, which assume a linear response-covariate relationship, instrumental intensity was preferred over other possible ground motion covariates because it captures part of that nonlinearity. Measures of commercial exposure were preferred over measures of residential exposure for both ground motion and tsunami ignition models. This may vary in other study regions, but does highlight the value of testing

alternative measures for each concept. The models with the best predictive power included two or three covariates. Those with just one were not able to capture as much variability; those with more did not improve the predictive ability and in some cases overfit the data.

Though the proposed models are able to capture a great deal of the variability in the mean number of ignitions per municipality/ku, it is important to remember that substantial Poisson (or NB) variability remains in predicting the specific ignition map that will occur for a given earthquake or tsunami. Furthermore, although the data sets enjoy several strengths (Section 1), it is important to remember that they come entirely from a single earthquake in Japan and thus may not be applicable in other countries and may not capture features associated with the time of occurrence (day of the year or time of the day), which is thought to affect the occurrence of postearthquake ignitions.[2]

The analysis also suggests a few notes about the process. In any statistical analysis like this, it is important to ensure that the proposed model makes sense given what is known about the phenomenon—postearthquake ignitions in this case. Thus, the conceptual framework is a useful tool for guiding the choice of covariates and ensuring that the contribution of each is reasonable in terms of parameter sign for GLMs or smooth shape for GAMs. In particular, although GAMs provide an excellent option, care must be taken to ensure the smooths are not overfitted. The analysis also shows the value of comparing models across multiple criteria—total region and prefecture-level predictive error, ease of application and use, and goodness of fit to the sample data. No single model performs best across all, and thus the choice will depend on the intended use.

In this article, we have applied two modeling techniques—GAMs and BRTs—not previously used in this application area. Using a newly available, unique data set, we have introduced new models of ground- motion-generated ignitions for Japan that are a substantial improvement over available models, both in terms of the statistical methods used and the demonstrated predictive power. We have also introduced the first statistical models of tsunami-generated ignitions. These new models can be used for prediction of the number and geographic distribution of ignitions in future earthquakes and tsunamis.

ACKNOWLEDGMENTS

The authors thank the National Science Foundation (CMMI-1138675) for financial support of this research, and also thank the Committee for Postearthquake Fire Research of the Japan Association for Fire Science and Engineering, which provided ignition data. The author Himoto worked at Kyoto University for part of the time during which the research was conducted.

REFERENCES

1. Murata A, Iwami T, Hokugo A, Murosaki Y. Mechanism of the outbreak of fire in the 1995 Hyogo-ken Nambu earthquake—In comparison with past earthquake fire cases. Journal of Architecture, Planning and Environmental Engineering, Architectural Institute of Japan, 2001; 548:1–8.
2. Scawthorn C, Eidinger J, Schiff A. Fire Following Earthquake, Technical Council on Lifeline Earthquake Engineering Monograph No. 26. Reston, VA: American Society of Civil Engineers, 2005.
3. Lee S, Davidson R, Ohnishi N, Scawthorn C. Fire following earthquake—Reviewing the state-of-the-art of modeling. Earthquake Spectra, 2008; 24(4):1–35.
4. Mohammadi J, Alyasin S, Bak D. Investigation of Cause and Effects of Fires Following the Loma Prieta Earthquake. IIT-CE-92-01. Chicago, IL: Illinois Institute of Technology Civil Engineering, 1992.
5. Williamson R, Groner N. Ignition of Fires Following Earthquakes Associated with Natural Gas and Electric Distribution Systems. Berkeley, CA: Pacific Earthquake Engineering Research Center Report, University of California, 2000.
6. Zolfaghari MR, Peyghaleh E, Nasirzadeh G. Fire following earthquake, infrastructure ignition modeling, Journal of Fire Sciences, 2009; 27:45–79.
7. Yildiz S, Karaman H. Post-earthquake ignition vulnerability assessment of Küçükçekmece District. Natural Hazards and Earth System Sciences, 2013; 13:3357–3368.
8. Kawasumi H. Examination of Earthquake-Fire Damage in Tokyo Metropolis, Technical Report. Tokyo: Tokyo Fire Department, 1961.
9. Trifunac MD, Todorovaska MI. The Northridge, California, Earthquake of 1994: Fire ignition by strong shaking. Soil Dynamics and Earthquake Engineering, 1998; 17:165–175.
10. Architectural Institute of Japan (AIJ). Report on the Hanshin-Awaji Earthquake Disaster, Vol. 6. Tokyo, Japan: Maruzen Publishing, 1998 (in Japanese).
11. Himoto K, Yamada M, Nishino T. Analysis of ignitions following the 2011 Tohoku earthquake using Kawasumi model. Proceedings of the 11th International Symposium on Fire Safety Science, February 10—14, Christchurch, New Zealand, 2014. Available at: www.iafss.org/publications/fss/11/24.
12. Ren A, Xie X. The simulation of post-earthquake fire-prone area based on GIS. Journal of Fire Sciences, 2004; 22(5):421–439.
13. Li J, Jiang J, Li M. Hazard analysis system of urban post-earthquake fire based in GIS. ActaSeismologicaSinica, 2001; 14(4):448–455.
14. Zhao S, Xiong L, Ren A. A spatial-temporal stochastic simulation of fire outbreaks following earthquake based on GIS. Journal of Fire Sciences, 2006; 24:313–339.
15. Cousins WJ, Smith W. Estimated losses due to post-earthquake fire in three New Zealand cities. Proceedings of the 2004 New Zealand Society of Earthquake Engineering

(NZSEE) Technical Conference, March 19–21, Rotorua, New Zealand, 2004. NZSEE, Paper 28.
16. Davidson R. Modeling post-earthquake fire ignitions using generalized linear (mixed) models. Journal of Infrastructure Systems, 2009; 15(4):351–360.
17. Anderson D. Statistical models of post-earthquake ignitions based on data from the Tohoku, Japan earthquake and tsunami [master's thesis]. Newark, DE: University of Delaware, 2014.
18. Murata A, Hokugo A. Questionnaires survey on fires after the Great East Japan Earthquake 2011—Part 1: Summary of questionnaires survey and the cause of the fire. Kasai, 2013; 63(1):1–6.
19. Baraldi A, Enders C. An introduction to modern missing data analyses. Journal of School Psychology, 2010; 48:5–37.
20. Hokugo A. Mechanism of tsunami fires after the Great East Japan Earthquake 2011 and evacuation from the tsunami fires. Procedia Engineering, 2013;62:140–153.
21. Tanaka T. Characteristics and problems of fires following the Great East Japan earthquake in March 2011. Fire Safety Journal, 2012; 54:197–202.
22. Wald D, Worden B, Quitoriano V, Pankow K. ShakeMap Manual: Technical Manual, Users' Guide, and Software Guide, version 1.0. Reston, VA: U.S. Geological Survey, 2006.
23. Geospatial Information Authority of Japan (GIAJ). Provision of Information on the 2011 Off the Pacific Coast of Japan Earthquake, 1/25000 Scale Inundation Area, 2011. Available at: www.gsi.go.jp/kikaku/kikaku40014.html, Accessed April 2012.
24. Japan Statistics Bureau, Ministry of Internal Affairs and Communications, 2009 and 2010. Available at: www.stat.go.jp/english/, Accessed November 2013.
25. Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Housing and Land Survey, 2008. Available at: www.stat.go.jp/data/jyutaku/2008/, Accessed January 2013.
26. Ministry of Land, Infrastructure, Transport and Tourism (MLIT). National Land Numerical Information Download Service, 2011. Available at: http://nlftp.mlit.go.jp/ksj-e/, Accessed January 2013.
27. Fire and Disaster Management Agency (FDMA). 148th Report on the 2011 Off the Pacific Coast of Tohoku Earthquake, 2013. Available at: www.fdma.go.jp/bn/higaihou/pdf/jishin/148.pdf, Accessed September 2013.
28. RCore Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2013, Available at: www.R-project.org.
29. Venables WN, Ripley BD. Modern Applied Statistics with S, 4th ed. New York: Springer, 2002.
30. Wood S. Generalized Additive Models: An Introduction with R. Boca Raton, FL: Chapmen & Hall, 2006.
31. Ridgeway G with contributions from others. GBM: Generalized Boosted Regression Models. R package version 2.1-0.3, 2013. Available at: http://code.google.com/p/gradientboostedmodels/.
32. Ridgeway G. 2012. Generalized Boosted Models: A Guide to the GBM Package. R Package Vignette. Available at: http://CRAN.R-project.org/package=gbm.
33. McCullagh P, Nelder J. Generalized Linear Models, 2nd ed. London: Chapman & Hall, 1989.
34. Cameron A, Trivedi P. Regression Analysis of Count Data. Cambridge: Cambridge University, 1998.
35. Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton J, Peterson AT, Phillips SJ, Richardson KS, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 2006; 29:129–151.
36. Elith J, Leathwick J, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology, 2008; 77:802–813.
37. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, 2001.
38. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.
39. Friedman JH. Stochastic gradient boosting. Computational Statistics and Data Analysis, 2002; 38:367–378.
40. Cameron A, Windmeijer F. R-squared measures for count data regression models with applications to health-care utilization. Journal of Business and Economic Statistics, 1996; 14(2):209–220.
41. Akaike H. A new look at statistical model identification. IEEE Transactions on Automatic Control, 1974; AU-19:716–722.
42. Wald DJ, Quitoriano V, Heaton TH, Kanamori H. Relationship between peak ground acceleration, peak ground velocity, and modified Mercalli intensity for earthquakes in California. Earthquake Spectra, 1999; 3 (15):557–564.
43. Liu H, Davidson R, Apanasovich T. Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. Reliability Engineering and System Safety, 2008; 93(6):897–912.
44. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics, 2004; 20:289–290.