STATISTICAL CLASSIFICATION AND PREDICTION OF SEISMIC EVENT LOCATION ACCURACY

Gardar Johannesson and Jerry J. Sweeney

Lawrence Livermore National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Development
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-48

ABSTRACT

In this paper we address the question of how to assign a statistically meaningful error estimate to the location accuracy of an individual seismic event contained in an international seismic bulletin—in this case the U.S. Geological Survey National Earthquake Information Center Reviewed Event Bulletin (NEIC REB). Our approach is to use a training dataset of ground-truth events, for this study obtained from the Middle East-North Africa region, and apply a wide variety of statistical methods. Using the distance between the catalog location and the ground-truth (GT) location of an event as the dependent variable, we examine the influence of the available catalog variables: distance to nearest station, number of detecting phases, primary azimuthal gap, secondary azimuthal gap, and magnitude. We examine several different regression and classification methods and look at their success rate and efficiency as location classifiers for GT5 to GT25. As expected, the classification error rate is inversely related to GT level; but we found it varied only slightly between the different methods. The analysis reveals that representative GT levels that can be obtained from the catalog for one type of regression model are $GT25_{95}$, $GT20_{90}$, $GT15_{80}$, $GT10_{70}$, and $GT5_{50}$ (using the nomenclature of Bondar et al., 2004). We also present an approach for using the statistical model to determine the probability that a given event is at a given level of ground truth, i.e., determine P for GTn_P where n is the level of ground truth. Advantages of regression methods over classification methods are that they make better use of the information in the data and they allow one to construct a classifier based simply on the predicted accuracy of the location, which can then be weighted. We hope that this study will inspire further investigation of the use of formal statistical regression and classification methods to quantify the accuracy of reported seismic locations in bulletins.

OBJECTIVES

Use of GT seismic events to calibrate seismic travel times has become an important way of calibrating seismic stations and regions for improved location accuracy. Statistical methods, such as those introduced by Myers and Schultz (2000), produce spatial travel time corrections based on calibration to ground truth data which itself can have different degrees of uncertainty. Obviously, earthquake and explosion sources will be of variable quality due to different uncertainties in origin time and hypocenter; but as long as the statistical uncertainty in the location of a source event is known or can be estimated, this possible variance can be incorporated in the analysis. In order to calibrate stations over wide areas of the globe, evenly spaced distributions of large numbers of ground truth events are desired. One means of acquiring a large number of ground truth events is to extract them from published catalogs of events, such as the International Seismological Centre (ISC) and the U. S. National Earthquake Information Center (NEIC). The key to being able to use these sources of ground truth is knowing what criteria to use for selection and how to estimate the uncertainty in the accuracy of the ground truth; i.e., what is the level of ground truth (5, 10, 15, 20 km error, etc.) and level of certainty about that level (95%, 90%, 67%, 50% confidence, etc.). Other considerations also come into play, such as how efficient is the selection process.

The objective of this study is to provide a careful statistical assessment of the whole range of parameters available from a published catalog (the NEIC reviewed event bulletin—REB), examine the relative merits of individual or combinations of parameters, examine different types of statistical classification schemes, and assess the ability of these schemes to successfully classify ground-truth events in an international bulletin.

RESEARCH ACCOMPLISHED

Data Used in the Study

Ground-truth events used in this study are those used in an initial study by Sweeney (1998), with five additional event collections added. The seismic reference events are from earthquakes and aftershocks that have been accurately located with temporary arrays of instruments in a local network or with a relatively dense regional array with an accuracy of at least 5 km and selected explosions. While the number of reference events used in this study is statistically significant, it is not as large as that used in other studies, such as Bondar et al. (2004). However, the purpose of this study is not to come up with a definitive answer to the question of event selection, but rather to illustrate a rigorous statistical approach that can be used with any reference data.

For the study we focus on six parameters available from published bulletins. The dependent variable (the variable of interest) is the distance between the catalog location and the ground truth location (which we denote by "**dist**"). The independent (explanatory) variables we consider are the number of defining phases ("**N**"), the largest azimuthal gap between stations recording the event ("**gap**"), the largest azimuthal gap filled by a single station (Bondar et al., 2004—"**gap2**"), the smallest distance between the event location and a recording station ("**sta**"), and the magnitude (typically m_b) of the event ("**mag**"). We use the open-source statistical computing language R (R Development Core Team, 2004) for our analysis. The R environment has emerged as the tool of choice in the statistical community for developing and testing new statistical procedures.

For the data set of 69 records, 48 events (69.6%) have **dist** less than 15km, 54 events (78.3%) have **dist** less than 20 km, and 56 events (81.2%) have **dist** less than 25km. Histograms of the variables (Figure 1) generally show a left tail tendency, in particular for **dist**, **N**, and **sta**. To reduce the impact of extreme values (e.g., large values) in our analysis, the **dist**, **N**, and **sta (in degrees)** variables were logarithmic (base-10) transformed. Scatter plots of log(**dist**) versus the explanatory variables are shown in Figure 2. We observe (1) an expected positive relationship between log(**dist**) and **gap**, **gap2**, and log(**sta**), (2) an expected negative relationship between log(**dist**) and log(**N**), and (3) a somewhat surprising positive-then-negative relationship between log(**dist**) and **mag**. Finally, Figure 3 shows a scatter plot matrix of the explanatory variables. We observe a strong correlation between log(**N**), **gap**, **gap2**, and **mag**, as might be expected since a large magnitude (**mag**) event is typically observed at many stations (**N**) yielding a small azimuthal gap (**gap** and **gap2**). However, the distance to the closest station (**sta**) is observed to have little correlation to the remaining four explanatory variables.

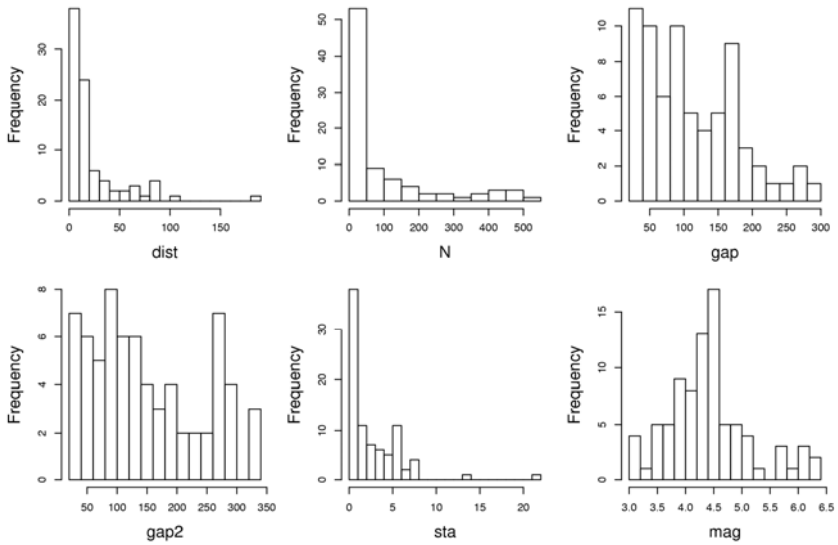

Figure 1. Histograms of the six variables of interest in the NEIC dataset.

The result of using the GT20 and GT25 classification criteria given by Bonder et al. (2004) to the 69 events in the full data sets is given in Table 1. The Bonder et al. classification rule has a very low error rate for the events that are predicted to be GT20 or GT25 events, 3.2% for GT20 events, and 0.0% for GT25 events. We refer to this error rate as type-I error rate; that is an event is classified as a GT event, but it is not a GT event. However, the Bonder et al. classification rule has a very high, what we refer to as type-II, error rate: an event is classified as not being a GT event, but it is a GT event. As a result, the total classification error rate is high (36.2% for both GT20 and GT25 events). This makes the classification rule very inefficient at classifying GT events. For example, of the 54 GT20 events in our set of data, 30 are classified as GT20 events, while 24 are not; we also note that the type-I error for the Bondar et al. GT5 criteria is 50% (2 events misclassified out of 4 predicted) for our GT reference set.

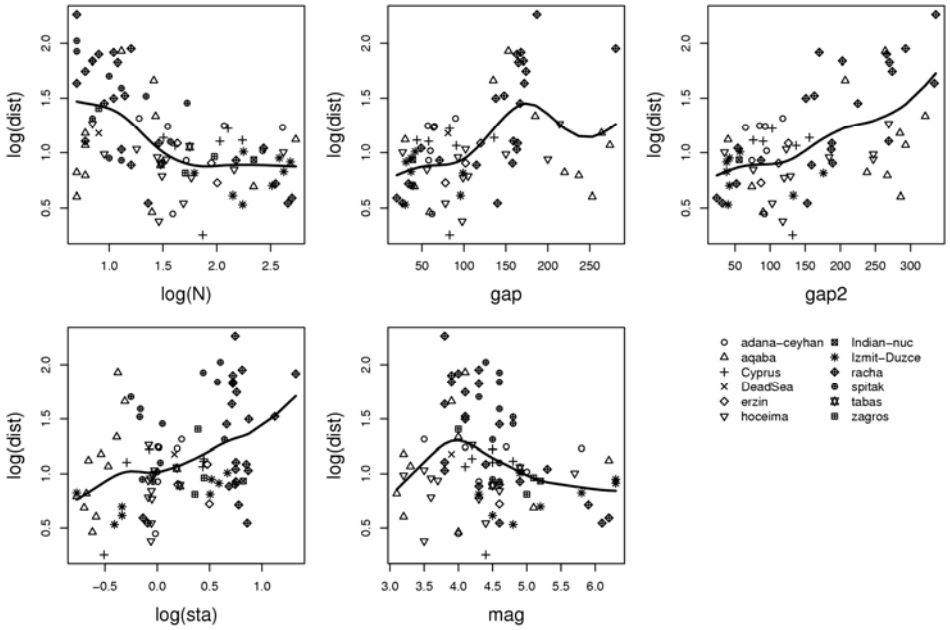

Figure 2. Scatter plot of log(dist) versus the five potential explanatory variables (with N and sta log10 transformed). A superimposed smoothing-spline trend line is also shown on each panel.

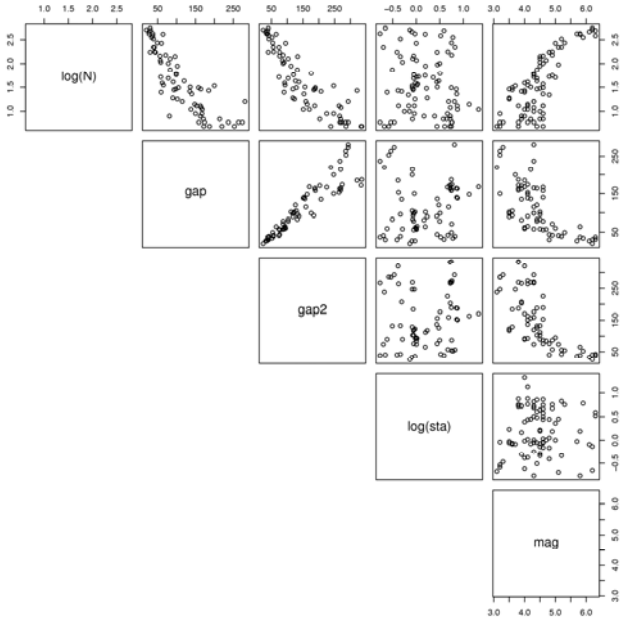

Figure 3. A scatter plot matrix of the potential explanatory variables.

Table 1. Classification results of GT20 and GT25 events using the classification rules of Bondar et al. (2004). The column headings "P>20" and "P<20" denote events predicted (classified) to be non-GT20 and GT20, respectively, and the row headings "O>20" and "O<20" denote events observed to be non-GT20 and GT20 events, respectively. The entry 30 in the left table shows the number of events that are both observed and predicted to be GT20 events, and the number 31 shows the total (Tot) number of predicted GT20 events. The entry "P<20 Err" gives the classification error in predicting GT20 event (given by 1/31), and the entry "Tot Err" gives the total classification error [given by (1+24)/69].

GT20	GT25
P>20 P<20 Tot	**P>25 P<25 Tot**
O>20 14 1 15	**O>25** 13 0 13
O<20 24 30 54	**O<25** 25 31 56
Tot 38 31 69	**Tot** 38 31 69
P<20 Err = 3.2%	**P<25 Err** = 0.0%
Tot Err = 36.2%	**Tot Err** = 36.2%

Analysis: Regression Methods

In our analysis, we considered two groups of statistics-based classification schemes: those based on regression techniques and those based on classification techniques. We first discuss the regression methods.

Following a standard regression notation (see, e.g., Hastie et al., 2001, Chapter 3), let Y = the output/response variable, which is equal to log(**dist**) in our case, and let

$$X = (X_1, ..., X_p) \qquad (1)$$

be a vector of *p* input variables. The set of input variables does not only include the available explanatory variables, but can also include transformations of those and in general any known function of the available explanatory variables.

For example,

$$X_1 = \log(\mathbf{N}),\ X_2 = \log(\mathbf{sta}),\ X_3 = \log(\mathbf{N}) \times \log(\mathbf{sta}) \quad (2)$$

is a set of $p = 3$ input variables. Using the input variables, let

$$f(X) = \text{a given model (function) for predicting } Y \text{ given the input } X, \quad (3)$$

and note that $f(X)$ typically depends on some unknown parameters that need to be specified. The unknown parameters are estimated from observed data that consists of responses, $y_1, \ldots, y_N$,with associated inputs, $x_1, \ldots, x_N$, where $x_i = (x_{i1}, \ldots, x_{ip})$. We denote by

$$\hat{f}(X) = \text{ the model prediction } f(X) \text{ with parameters obtained from the data.} \quad (4)$$

That is, an estimate of the unknown parameters is obtained from the data and simply "plugged" into the model $f(X)$.

The predictor $f(X)$ can be used to create a simple threshold classifier of "is $Y < y$?" versus "is $Y \geq y$?," where y is a given threshold number. Such classifiers can be written as

$$G(X) = \begin{cases} 1 & \text{if } f(X) < y, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Given the estimated predictor $\hat{f}(X)$, one can derive the classifier $\hat{G}(X)$ by simply "plugging in" $\hat{f}(X)$ instead of $f(X)$ in the expression for $G(X)$ above.

In accessing the accuracy of the classifier $\hat{G}(X)$ one can apply it to the data used to estimate the unknown parameters associated with $f(X)$. However, that is in general not considered a good practice as the same data set is used to both estimate (train) the classifier and validate it. An alternative and widely used approach is cross validation (see, e.g., Hastie et al., 2001, Chapter 7.10). An *M*-fold cross validation splits the available data into *M* parts of roughly equal size and for each part uses the remaining *M-1* parts to re-estimate $f(X)$ which is then used to predict the data part left out; this requires *M* re-estimations of the model. If each data part consists of a single observation (i.e., *M* is equal to the number of observations in the data set), then it is often referred to as leave-one-out cross-validation. We applied two regression methods to the NEIC data set: first a classical linear regression and a more adaptive and flexible regression approach.

In linear regression we seek a predictor of *Y* of the linear form,

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \quad (6)$$

where the β_j's are unknown parameters to be estimated from the data. Here we shall consider least squares regression where an estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is obtained by minimizing the residual sum of squares,

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2. \quad (7)$$

For details, see, e.g., Hastie et al. (2001).

Of the five explanatory variables we have available, we have the option of including them as main-effects (e.g., $X_j = \log(\mathbf{N})$), as 2-way interactions (e.g., $X_j = \log(\mathbf{N}) \times \log(\mathbf{sta})$), as 3-way interactions (e.g., $X_j = \log(\mathbf{N}) \times \log(\mathbf{sta}) \times \mathbf{mag}$), all the way up to a 5-way interaction. We used an automatic stepwise selection using the Akaike information criterion (AIC) via the stepAIC function available in the MASS package of R; see Venables and Ripley (2002, page 172) and Hastie et al. (2001, Chapter 7).

The model with the lowest AIC statistics selected by stepAIC from the multiple starting models has four input variables and is given by

$$\hat{f}(X) = 0.0121 + 0.00314 \times \mathbf{gap2} + 1.35 \times \log(\mathbf{sta}) + 0.120 \times \mathbf{mag} - 0.240 \times \log(\mathbf{sta}) \times \mathbf{mag}. \quad (8)$$

A scatter plot of the observed **dist** versus the predicted **dist** from the above model has an $R^2 = 0.53$. The term **gap2** has an expected positive contribution to the predictor, while the joint contribution of log(**sta**) and **mag** is harder to judge. A classification of GT15, GT20, and GT25 events was carried out using the threshold classifier described above and a leave-one-out cross-validation; results are given in Table 2.

Table 2. Cross-validation classification results for the linear regression model in classifying GT15, GT20, and GT25 events.

GT15	GT20	GT25
P>15 P<15 Tot	**P>20 P<20 Tot**	**P>25 P<25 Tot**
O>15 13 8 21	**O>20** 11 4 15	**O>25** 10 3 13
O<15 8 40 48	**O<20** 5 49 54	**O<25** 2 54 56
Tot 21 48 69	**Tot** 16 53 69	**Tot** 12 57 69
P<15 Err = 16.7%	**P<20 Err** = 7.5%	**P<25 Err** = 5.3%
Tot Err = 23.2%	**Tot Err** = 13.0%	**Tot Err** = 7.2%

As an alternative to the above, multiple adaptive regression splines (MARS) expresses the impact of each explanatory variable via a combination of piecewise linear terms; see Friedman (1991) for details and Hastie et al. (2001), Chapter 9.4, for a short overview. We let $X_1, X_2, \ldots, X_p$ be our set of explanatory variables. MARS performs a linear regression with input variables selected from the set

$$\{ (X_j - k)_+, (k - X_j)_+ \} \text{ for } j = 1, \ldots, p \text{ and } k \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\}, \quad (9)$$

where the piecewise linear terms $(X_j - k)_+$ and $(k - X_j)_+$ are given by

$$(X_j - k)_+ = \begin{cases} X_j - k & \text{if } x > k, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad (k - X_j)_+ = \begin{cases} k - X_j & \text{if k} < x, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Figure 4 shows the two piecewise linear functions above in the case when X_j = **gap2** and k = 150. If all the observed values of the explanatory variables (i.e., the x_{ij}'s for each j) are different, this results in a set of $2Np$ possible input variables; that is 690 input variables when p = 5 and N = 69 as in our case. To select among these input variables and all possible interactions, MARS uses a forward stepwise selection of variables (i.e., starting with an empty set of variables and then adding variables in a stepwise fashion), which is then followed by a backward elimination of variables. The final model is of the former,

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X), \quad (11)$$

in the case when M input terms are selected, where each $h_m(X)$ is one of the input variables or a product of two or more input variables. The parameters $b_0, b_1, \ldots, b_M$ are estimated by minimizing the residual sum of squares, as in the previous case of linear regression. The size (M) of the final model selected is automatically decided on using the generalized cross-validation (GCV) criterion (e.g., Hastie et al., 2001, Chapter 9.4).

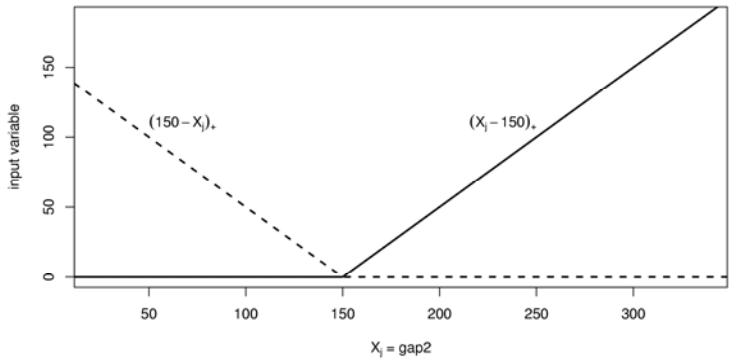

Figure 4. A pair of piecewise linear terms with a knot at 150 used by MARS for the explanatory variable gap2.

The final model selected when we applied MARS to our data with a main-effects-only restriction had only two input variables, and is given by

$$\hat{f}(X) = 1.10 + 0.00361 \times (\mathbf{gap2} - 132)_+ - 0.347 \times (0.723 - \log(\mathbf{sta}))_+ . \qquad (12)$$

The two input variables reflect the marginal relationship seen in Figure 2. The selected MARS model with main-effects and two-way interactions ended up incorporating six input variables and is given by,

$$\begin{aligned} \hat{f}(X) = {} & 0.843 + 0.00894 \times (\mathbf{gap2} - 132)_+ \times (0.723 - \log(\mathbf{sta}))_+ \\ & - 0.00985 \times (\log(\mathbf{N}) - 1.11)_+ \times (\mathbf{gap2} - 132)_+ \\ & - 0.0239 \times (\mathbf{gap} - 135)_+ \times (0.723 - \log(\mathbf{sta}))_+ \\ & + 0.0202 \times (\mathbf{gap} - 153)_+ - 0.169 \times (\mathbf{gap} - 167)_+ \times (\log(\mathbf{sta}) - 0.723)_+ \\ & + 0.510 \times (2.07 - \log(\mathbf{N}))_+ \times (\log(\mathbf{sta}) + 0.0655)_+ . \end{aligned} \qquad (13)$$

Note that of the six input variables there is only one main-effect.

The MARS model with six input variables is seen to perform better in classifying GT15 and GT20 events then the linear regression model presented in Table 2, but is slightly worse at classifying GT25 events. Even with just two input variables (i.e., the simpler MARS model), one is able to extract a considerable amount of information about the GT accuracy of the events.

Alternative Classification Rules

In addition to the classification methods above, we want take into account the uncertainty associated with the accuracy of the predictor $\hat{f}(X)$ (i.e., a predictor of log(**dist)** in our case). In addition, we might want to "penalize" differently for misclassifying a non-GT event as a GT event (i.e., a type I error) versus misclassifying a GT event as non-GT event (i.e., a type II error). Regression models produce a single best prediction and also yield a probability distribution for the predictor. This in turn allows for the computation of

$$\Pr(\hat{f}(X) < d) \text{ = the probability that } \hat{f}(X) < d \text{ for a given threshold value } d. \qquad (14)$$

The analysis revealed that, of the 69 events, 54 (78%) have established distance below the 75% threshold, 61 (88%) below the 90% threshold, and all 69 events are below the 95% threshold. Hence, the confidence intervals are in reasonable agreement with the empirical results given the size of the data set (however, given a sufficiently large data set, one might prefer to use the empirically derived confidence probabilities rather than the theoretical ones if there is a large discrepancy between the two).

Confidence levels may seem to be picked arbitrarily, but there is a connection between the confidence level and the cost associated with misclassification that might help in selecting an appropriate confidence level for classification.

Let

c_I = the cost of misclassifying a non-GT event as a GT event (type I error). (15)

and

c_{II} = the cost of misclassifying a GT event as a non-GT event (type II error). (16)

A classification rule can then be based on minimizing the expected loss of the action taken. The resulting classifier can be shown to be (see e.g., Hastie et al., 2001, Chapter 2.4)

$$G(X) = \begin{cases} 1 & \text{if } \Pr(\hat{f}(X) < d) > c_I / (c_I + c_{II}), \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

Hence, a classifier that classifies an event as a GT15 event with 90% confidence can be seen as a classifier where c_I is assumed to be nine times higher than c_{II}.

Table 3 shows the number of events classified as GT5, 10, 15, 20, and 25 events at various confidence level using the MARS 2 model. The top section of Table 4 shows the classification results when using the MARS 2 model to classify GT15 events at three different confidence levels (i.e., with three different misclassification cost assumptions). As expected, the (empirical) accuracy of the classifier increases with increasing confidence, but its efficiency (in number of events labeled as GT15) decreases. The table also shows that, using this model, it is not possible to obtain better than $GT5_{50}$, $GT10_{70}$, $GT15_{80}$ events from the NEIC REB.

Table 3. Number of events (out of 69) classified as GT5, 10, 15, 20, and 25 events at five different confidence levels using the MARS 2 model and the classifier in (2). The bottom row shows the actual (established) number of events at each GT level.

	GT5	GT10	GT15	GT20	GT25
50%	3	45	51	52	54
60%	0	39	48	51	53
70%	0	35	44	48	51
80%	0	2	40	45	46
90%	0	0	2	39	44
Est.	12	33	48	54	56

Table 4. Cross-validation classification results for classifying GT15, 20, and 25 events using different methods.

Method	GT15 at 50%	GT15 at 80%	GT15 at 90%
MARS 2	**P>15 P<15 Tot** **O>15** 15 6 21 **O<15** 3 45 48 **Tot** 18 51 69 **P<15 Err** = 11.8% **Tot Err** = 13.0%	**P>15 P<15 Tot** **O>15** 17 4 21 **O<15** 12 36 48 **Tot** 29 40 69 **P<15 Err** = 10.0% Tot Err **= 23.2%**	**P>15 P<15 Tot** **O>15** 21 0 21 **O<15** 46 2 48 **Tot** 67 2 69 **P<15 Err** = 0.0% **Tot Err** = 66.7%
	GT15	**GT20**	**GT25**
LDA	**P>15 P<15 Tot** **O>15** 13 8 21 **O<15** 5 43 48 **Tot** 18 51 69 **P<15 Err** = 15.7% **Tot Err** = 18.8%	**P>20 P<20 Tot** **O>20** 11 4 15 **O<20** 3 51 54 **Tot** 14 55 69 **P<20 Err** = 7.3% **Tot Err** = 10.1%	**P>25 P<25 Tot** **O>25** 11 2 13 **O<25** 2 54 56 **Tot** 13 56 69 **P<25 Err** = 3.6% **Tot Err** = 5.8%
Classification Tree	**P>15 P<15 Tot** **O>15** 14 7 21 **O<15** 7 41 48 **Tot** 21 48 69 **P<15 Err** = 14.6% **Tot Err** = 20.3%	**P>20 P<20 Tot** **O>20** 11 4 15 **O<20** 7 47 54 **Tot** 18 51 69 **P<20 Err** = 7.8% **Tot Err** = 15.9%	**P>25 P<25 Tot** **O>25** 9 4 13 **O<25** 6 50 56 **Tot** 15 54 69 **P<25 Err** = 7.4% **Tot Err** = 14.5%

Classification Methods

An alternative to the regression approach is to perform a direct classification of a derived group variable. In this case the response variable is not the recorded log(**dist**) variable, as in case of regression, but simply an indicator variable, indicating if an observed event is, for example, a GT15 event or not. In a more general notation, denote by

$$G = \text{a reponse group variable (indicator)} \tag{18}$$

that is either equal to 0 or 1, corresponding to two possible outcomes or groups (this can be easily extended to more than two groups). As in regression, associated with G is a vector of input variables $X = (X_1, \ldots, X_p)$. Our goal is then to derive a model for

$$\Pr(G = 1 \mid X = x) = \text{the probabilit y that } G = 1, \text{ given that } X = x, \tag{19}$$

and then classify the (unknown) event associated with the input x to group 1 ($G = 1$) if $\Pr(G = 1 \mid X = x) > 0.5$, otherwise classify to group 0 ($G = 0$); see, for example Hastie et al. (2001, chapter 2.4). This classification rule assumes an equal loss in misclassifying a group 1 event as a group 0 event and a group 0 event as a group 1 event. An unequal loss results in a threshold value for $\Pr(G = 1 \mid X = x)$ that is different from 0.5 (one that is identical to the threshold value for the regression-derived classifier in above).

In deriving the probability model $\Pr(G = 1 \mid X = x)$ we have available a *training* data set that consists of the 69 records in the NEIC data set, where the group variable is, for example, in the case of GT15 events, given by

$$g_i = 1 \text{ if } \mathbf{dist}_i < 15\text{km}, \; g_i = 0 \text{ otherwise} \tag{20}$$

for the $i = 1, \ldots, 69$ observations in the data set. Note that a separate classifier is needed to classify GT20 and GT25 events. We applied two classification procedures to the NEIC data, a linear discrimination analysis (LDA) and a binary classification tree.

LDA (see, e.g., Hastie et al., 2001, Chapter 4.3) models the conditional probability of $G = 1$, given $X = x$, as

$$\Pr(G = 1 \mid X = x) = \frac{\pi_1 \varphi(x; \mu_1, \Sigma)}{\pi_0 \varphi(x; \mu_0, \Sigma) + \pi_1 \varphi(x; \mu_1, \Sigma)} \tag{21}$$

where $\varphi(x; \mu, \Sigma)$ denotes the probability density of a multivariate normal distribution evaluated at x with mean vector μ and variance-covariance matrix Σ and $\pi_0 + \pi_1 = 1$ ($0 \le \pi_g \le 1$, $g = 1,2$). The π_0 and π_1 are the *prior* probabilities that $G = 1$ or $G = 0$, respectively (i.e., our prior believes that the event to be classified belongs to group 1 or 0 typically based on the frequency of the two groups in the available data). The two multivariate normal distributions specify the distribution of the input variable X conditional on knowing the group it is associated with. Note, we assume that the two groups have different mean vectors (μ_0 and μ_1) but have the same variance-covariance matrix Σ. Quadratic discrimination analysis (QDA) is an extension of LDA where the two groups are allowed to have different variance-covariance matrices.

We conducted (three) LDA of the NEIC data set to classify GT15, GT20, and GT25 events using the 1DA function in the MASS package of R (Venables and Ripley, 2002, Chapter 12). After exploring a number of models using the leave-one-out cross-validation process, we selected the following set of input variables:

For GT15: log(**N**), **mag**, log(**N**)×**mag**
For GT20 and GT25: log(**N**), **gap2**, log(**sta**), log(**N**)×log(**sta**)

A number of other models come close to the above selected models, so these are not unique. The cross-validation results for these three classifiers are shown in the middle of Table 4. The LDA does well when compared to the classifier derived from the linear regression model, but is slightly worse than the MARS-derived classifier for GT15 and GT25 events.

In its most common form, a tree-based classification (see, e.g., Hastie et al., 2001, Chapter 9.2) consists of a sequence of binary decision rules (binary splits), where each decision rule is based on a feedback from a single input variable. For a continuous input variable X, a binary split takes (loosely) the form "if $X < a$, do this …, else, do that …". One can visualize this sequence of binary splits as a tree, where the decision process starts at the root-node of the tree. The first split creates two branches off the root-node connecting to two children nodes, which again have their own branches connecting them to their own children nodes, etc. At the terminal nodes of the tree (at the leaves), a classification decision is made as to if $G = 1$ or $G = 0$ based on the proportion of group 1 versus group 0 training data assigned to each terminal node. Using the training data set, the tree is grown in a sequential fashion, starting at the root-node. At each node, an "impurity measure" is computed to judge among the many splits possible (i.e., which input variable to use and where to threshold); see Hastie et al. (2001), Chapter 9.2. The tree is grown this way to a large size, typically until there are only very few training observations left in each terminal node. This large tree is then "pruned" using a "cost-complexity" measure

$$R_\alpha = R + \alpha \times (\text{tree size}), \tag{22}$$

where R is a measure on the fidelity to the data (e.g., classification error rate), the tree size is given by the number of splits, and $a > 0$ is a control parameter. The pruning process consists of joining nonterminal nodes together in an optimal way, as guided by the cost-complexity function. Depending on how severely the cost-complexity function penalizes for the tree size (i.e., the value of a), the pruning process yields a subset of trees, each one optimal for a given range of a. The final three selected among those are then carried out via cross-validation; see Hastie et al. (2001), Chapter 9.2, for further details.

We conducted tree-based classification using the rpart function in the rpart package of R; see Therneau and Atkinsson (1997) and Venables and Ripley (2002), Chapter 9. In our application we followed closely the approach outlined in Venables and Ripley (2002, pages 261–266), which uses cross-validation for model selection. The final (pruned) classification tree for GT15 events is very simple, with two splits; the first one on "**N < 27.5**," and if that is true, a second split is carried out using "**sta ≥ 0.275**." The final classification tree for GT20 events (not shown) was identical to the final GT15 classification tree, except the second split was on "**sta ≥ 0.375**." For GT25 events, the first split was on "**N < 17**," and the second split on "**sta ≥ 0.275**." The leave-one-out cross-validation classification results for the three trees are given in bottom of Table 4. These simple classification trees are seen to perform worse than LDA, particularly for GT20 and GT25 events.

CONCLUSIONS AND RECOMMENDATIONS

The detailed application of the four classification methods presented in this study yielded cross-validated total classification error-rates ranging from 13.0%–27.5% for GT15 events, 8.7%–15.9% for GT20 events, and 5.8%–14.5% for GT25 events (see Tables 1–4). The total classification error-rate accounts both for misclassifying non-GT event as GT event and for misclassifying a GT event as non-GT event. However, we are more concerned with misclassifying a non-GT event as GT, since the selected GT events might be used to calibrate event location procedures. It is thus more appropriate to use a classifier that penalizes more for misclassifying non-GT events as GT events, such as the regression-based classifier. Such a classifier would yield a higher total classification error-rate, but a smaller GT classification error-rate and fewer events classified as GT events. The balance between GT classification accuracy and efficiency (i.e., number of GT events classified) has to be decided by the user with the usage of the resulting classified GT events in mind.

In addition to the four classification methods discussed here, we also tested other methods including support vector machine, nearest neighbor, and neural network classifiers (see, e.g., Hastie et al., 2001). These methods yielded classification error-rates in the range reported above for the classifiers presented in this study. The four methods presented here were selected as representatives for their class of techniques (broadly described as "linear" and "non-linear" regression and classification). One advantage of carrying out regression is that a single model is estimated and used to predict the location accuracy (expected distance from true location) of various events. In addition, the regression approach takes better advantage of the information in the data as it models directly the distance between the established true location of the event and the predicted one versus simply carrying out classification on a binary GT variable derived by thresholding the established distance. An alternative is to simply report the predicted accuracy (**dist**) of each event. When those events are used, for example, to calibrate a seismic event location procedure, the predicted accuracy of the events can be used to construct weights that reflect their accuracy (e.g., one over **dist** squared).

It is the authors' hope that this study will inspire further investigation of the use of formal statistical regression and classification methods to quantify the accuracy of reported seismic location in international bulletins and the real benefits and performance of such methods.

REFERENCES

Bondar, I., S. C. Myers, E. R. Engdahl, and E. A. Bergman (2004). Epicentre accuracy based on seismic network criteria, *Geophys. J. Int.* 156: 483–496.

Friedmann, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* 19: 1–141.

Hastie T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Myers, S. C. and C. A. Schultz (2000). Improving sparse network seismic location with Bayesian kriging and teleseismically constrained calibration events, *Bull. Seism. Soc. Am.* 90: 199–211.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org).

Sweeney, J. J. (1998). Criteria for selecting accurate event locations from the NEIC and ISC Bulletins, Lawrence Livermore National Laboratory report UCRL-JC-130655.

Therneau, T. M. and E. J. Atkinson (1997). An introduction to recursive partitioning using RPART routines, Mayo Foundation technical report.

Venables, W. N. and R. D. Ripley (2002). *Modern Applied Statistics with S*, Fourth Ed. New York: Springer.