PNNL-12031

Pacific Northwest National Laboratory
Operated by Battelle for the
U.S. Department of Energy

Logistic Regression Applied to Seismic Discrimination

B. G. Amidan
D. N. Hagedorn

September 1998

Prepared for the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831;
prices available from (615) 576-8401.

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

PNNL-12031

Logistic Regression Applied to Seismic Discrimination

B. G. Amidan
D. N. Hagedorn

September 1998

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest National Laboratory
Richland, Washington 99352

Abstract

The usefulness of logistic discrimination was examined in an effort to learn how it performs in a regional seismic setting. Logistic discrimination provides an easily understood method, works with user-defined models and few assumptions about the population distributions, and handles both continuous and discrete data. Seismic event measurements from a data set compiled by Los Alamos National Laboratory (LANL) of Chinese events recorded at station WMQ were used in this demonstration study. PNNL applied logistic regression techniques to the data. All possible combinations of the Lg and Pg measurements were tried, and a best-fit logistic model was created. The best combination of Lg and Pg frequencies for predicting the source of a seismic event (earthquake or explosion) used $Lg_{3.0\text{-}6.0}$ and $Pg_{3.0\text{-}6.0}$ as the predictor variables. A cross-validation test was run, which showed that this model was able to correctly predict 99.7% earthquakes and 98.0% explosions for this given data set. Two other models were identified that used Pg and Lg measurements from the 1.5 to 3.0 Hz frequency range. Although these other models did a good job of correctly predicting the earthquakes, they were not as effective at predicting the explosions. Two possible biases were discovered which affect the predicted probabilities for each outcome. The first bias was due to this being a case-controlled study. The sampling fractions caused a bias in the probabilities that were calculated using the models. The second bias is caused by a change in the proportions for each event. If at a later date the proportions (a priori probabilities) of explosions versus earthquakes change, this would cause a bias in the predicted probability for an event. When using logistic regression, the user needs to be aware of the possible biases and what affect they will have on the predicted probabilities.

Introduction

In 1996 PNNL examined the theoretical applicability of various statistical classification methods for seismic discrimination (Anderson, et al., 1996). The methods tested were: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Variably Regularized Discriminant Analysis (VRDA), Flexible Discriminant Analysis (FDA), Logistic discrimination, Kth Nearest Neighbor (KNN), Kernal discrimination, and Classification and Regression Tree (CART). The methods all possessed different strengths and weaknesses.

This study takes another look at logistic discrimination using seismic event measurements from a data set compiled by Los Alamos National Laboratory (LANL) of Chinese events, recorded at station WMQ. Logistic discrimination provides a fairly easily understood method, works with a user-defined model, works with few assumptions being made concerning the population distributions, and handles both continuous and discrete data. The need for analysts to specify the model and the number of terms (linear, quadratic, etc.) is a desirable level of control over event identification. The drawbacks to logistic discrimination are: it becomes unstable when populations are widely separated; sequential decisions are not allowed; and it does not handle missing values.

PNNL applied logistic regression techniques to the LANL seismic data measurements in an effort to learn how logistic discrimination performed in a regional seismic setting. All possible combinations of the Lg and Pg measurements were tried and a best fit logistic model created. Cross-validation was then used to test the efficiency of the model. The method of cross-validation used is further explained in the Statistical Methodology section later in this report.

Seismic Background

Historically, seismic event identification has been accomplished using measurements from seismic events at teleseismic distances. Effective discrimination has been done using a linear division of earthquakes and explosions, based on the ratio of P-wave to S-wave measurements. Seismic ray paths for events at teleseismic distances extend down into the mantle and even into the core. The travel time in which the signals spend in the crust, relative to the time spent in the mantle and core, is relatively short. The Earth's core and mantle are much more homogeneous than the Earth's crust, resulting in a "smoothing filter" being applied to the seismic signals. Another factor in teleseismic discrimination is that at these long distances, only the larger events were generally seen. Empirical studies have shown that the separation of large earthquakes and explosions is better defined than for smaller events (Taylor, 1996; Taylor, et al., 1989).

As closer study is made about the signatures of small seismic events measured at regional distances, the linear division between earthquakes and explosions becomes questionable. At regional distances, many of the seismic signals used for event identification travel entirely within the crust. However, because the earth's crust exhibits a great degree of variation, the signals from which the discriminants are measured also exhibit more variability and identification becomes more challenging. Monitoring smaller magnitude events results in fewer stations recording useable signals, thus compounding an already problematic situation. Discrimination methods that are effective for these small, variable events must be able to handle complex population distributions, and must be able to correctly estimate the error for any given event.

DOE's current CTBT Research & Development monitoring program is working to better define the critical corrections to the recorded seismic signals by better modeling the earth's crust. A complimentary effort is also underway - acquiring as much regional seismic data as possible, so as to better estimate the variability of the discriminant populations and to allow for empirical analysis of potential discriminants. This report makes use of preliminary results from one region being studied by LANL seismologists.

The LANL seismologists found that two discriminants were most useful: a high frequency ($f > 4.0$ Hz) P/S ratio vs. mb, and a short period ($f > 1.0$ Hz) P to long period ($0.05 < f < 0.1$ Hz) Rayleigh wave ratio vs. mb (Hartse, et al., 1996). Predictably, the ratios show the least amount of variability when the measurements were taken as windowed averages, rather than as discrete seismic peak measurements.

In this study, PNNL used Pg and Lg amplitudes at the three windowed frequency ranges supplied by LANL, 0.75 to 1.5 Hz, 1.5 to 3.0 Hz, and 3.0 to 6.0 Hz. The ratios formed by the Pg and Lg measurements provide a good example of how logistic analysis performs in a regional seismic setting.

Data

Seismic phase measurements for 409 events recorded at the Chinese seismographic station WMQ, from the period 1988-1996, were made by LANL staff and provided to PNNL. The data set was composed of 243 known earthquakes, 139 probable earthquakes, 26 nuclear explosions, and 1 mine collapse.

The measured phases used in this study were Pg, Lg, and the pre-event (Pn) noise level, taken at three windowed frequency ranges (0.75 to 1.5 Hz, 1.5 to 3.0 Hz, and 3.0 to 6.0 Hz). Throughout the remainder of this report Pg (3.0 to 6.0 Hz) will be referred to as $Pg_{3.0\text{-}6.0}$ and the other measurements referred to likewise as: $\textit{seismic phase}_{\textit{frequency} \text{ range}}$. The pre-event noise measurements were made in window lengths comparable to the window length for the respective signal measurements (personal communication with Dr. Hans Hartse). In some cases, this resulted in quite different noise levels for the two phases. Since these noise levels were pre-event, and not pre-phase, and in order to simplify this demonstration of logistic analysis, we chose to use the noise measurement associated with the Pn measurement for both phases.

No distance corrections were applied to these data, since low-frequency Pg and Lg amplitude ratios show very little dependence on distance in Western China (Hartse, et al., 1996). Data measurements are in m/s (velocity). Missing Lg measurements may be due to short waveform records, and do not necessarily indicate censoring. However, in this study any missing measurement is treated as censored.

Statistical Methodology

Logistic discrimination uses logistic regression to find predicting variables to help forecast an outcome. The principles used in logistic regression are similar to those used in linear regression. The major difference with logistic regression is that the outcome (response) variable is binary instead of continuous, which causes the assumed distribution of the errors to be binomial, not normal as in linear regression.

In this case, the outcome variable is the seismic event, earthquake or explosion. The possible predictor variables in this study are the following seismic discriminants: $Pg_{0.75\text{-}1.5}$, $Pg_{1.5\text{-}3.0}$, $Pg_{3.0\text{-}6.0}$, $Lg_{0.75\text{-}1.5}$, $Lg_{1.5\text{-}3.0}$, and $Lg_{3.0\text{-}6.0}$.

The logistic regression model is:

$$\text{Pr (earthquake} \mid \text{x)} = \frac{e^{\beta' x}}{1 + e^{\beta' x}} \tag{1}$$

$$\text{Pr (explosion} \mid \text{x)} = 1 - \text{Pr (earthquake} \mid x) \tag{2}$$

where x represents the chosen predictor variables (discriminants), and β is a vector of regression coefficients. Logistic regression can be viewed as tossing an earthquake/explosion coin, where the probability of an earthquake is a function of β'x. An event with seismic measurements **x** could be classified as an earthquake if Pr (earthquake | **x**) > Pr (explosion | **x**), e.g., Pr (earthquake | **x**) > ½. The use of ½ as the "discrimination" rule is not mandated, and this "discrimination" rule could be adjusted according to the error and biases in the predictions.

If we form a log-odds ratio, we have,

$$Log\left(\frac{\text{Pr (earthquake} \mid \text{x)}}{\text{Pr (explosion} \mid \text{x)}}\right) = g\ (x) = \beta' x \tag{3}$$

which is the rationale for the term "logistic" regression. The log-odds ratio between

Pr (earthquake | **x**) and Pr (explosion | **x**) is a model reformulation known as a logit transformation (McLachlan, 1992). The regression coefficients, β, are calculated using an iterative maximum likelihood estimation procedure.

The predictor variables, **x**, can consist of any of the seismic discriminants, and/or any interaction terms. These predictor variables are analyzed to determine which combination of them are most related to the outcome. Only those variables that are significantly related to the outcome variable are kept in the model. Keeping variables in the model that are not helping with the prediction can lead to a model that may be overfitted and is too dependent on the observed data. Moreover, it may produce numerically unstable estimates (Hosmer and Lemeshow, 1989). A stepwise method using forward selection with a test for backward elimination is employed to add or remove variables sequentially to the model. These tests are based on statistical procedures and criteria.

In order to test the model, cross-validation is used to estimate the error rates. This is done by separating the data randomly into two sets, the training set and the prediction (holdout) set. The training set consists of 80% of the data for each outcome, while the prediction set contains the other 20% for each outcome. The training set is used to calculate the regression coefficients for the model and then each prediction set data is plugged into the model and the prediction made. The prediction is then compared to the actual outcome, and the percent of correct and erroneous predictions calculated. The whole process of randomly separating the data, calculating regression coefficients, and making predictions was simulated 100 times. The overall estimated correct and error rates are reported in the following section.

Results

A forward stepwise selection procedure was followed, with all six seismic discriminants and all interactions considered for inclusion into the logistic regression model. This procedure indicated the best-fit model was as follows:

$$\text{(Model 1)} \qquad g(x) = \beta' x = -16.74 - 84.97\, Pg_{3.0-6.0} + 79.27\, Lg_{3.0-6.0} \qquad (4)$$

Because this model depended only on the higher frequency measurements (3.0 to 6.0 Hz) and there may be times when higher frequency measurements are not possible, another selection process was run that did not allow the model to pick $Pg_{3.0\text{-}6.0}$ and $Lg_{3.0\text{-}6.0}$ together. This selection chose the following two models:

$$\text{(Model 2)} \qquad g(x) = \beta' x = 9.35 - 17.47\, Pg_{1.5-3.0} + 17.45\, Lg_{3.0-6.0} \text{ and} \qquad (5)$$

$$\text{(Model 3)} \qquad g(x) = \beta' x = -1.60 - 13.33\, Pg_{1.5-3.0} + 12.65\, Lg_{1.5-3.0} \qquad (6)$$

Plots from the three models are shown in Figure 1. These plots show the separation between the different events. The separation between the explosions (plotted in black) and the earthquakes (plotted in gray) is best shown with the 3.0 to 6.0 Hz frequencies (Figure 1a). Figures 1b and 1c illustrate that the other two models are less capable in separating the events in all instances. Figure 1d shows the discrimination at the 0.75 to 1.5 Hz frequency window and the difficulty to distinguish events at that level.

As previously mentioned, cross-validation was used to test the effectiveness of the model in correctly predicting the events. Table 1 lists the percent of correct predictions from the 100 simulations using holdout data. Any probability of over 0.5 was classified as an earthquake, and probabilities less than 0.5 were classified as explosions. All three models were effective at predicting earthquakes correctly. They differed in their abilities to correctly predict the explosions. The 3.0 to 6.0 Hz frequency windows were effective in predicting explosions correctly (98%). However, the models using the 1.5 to 3.0 frequencies were not nearly as effective (65.8% and 73.0%).

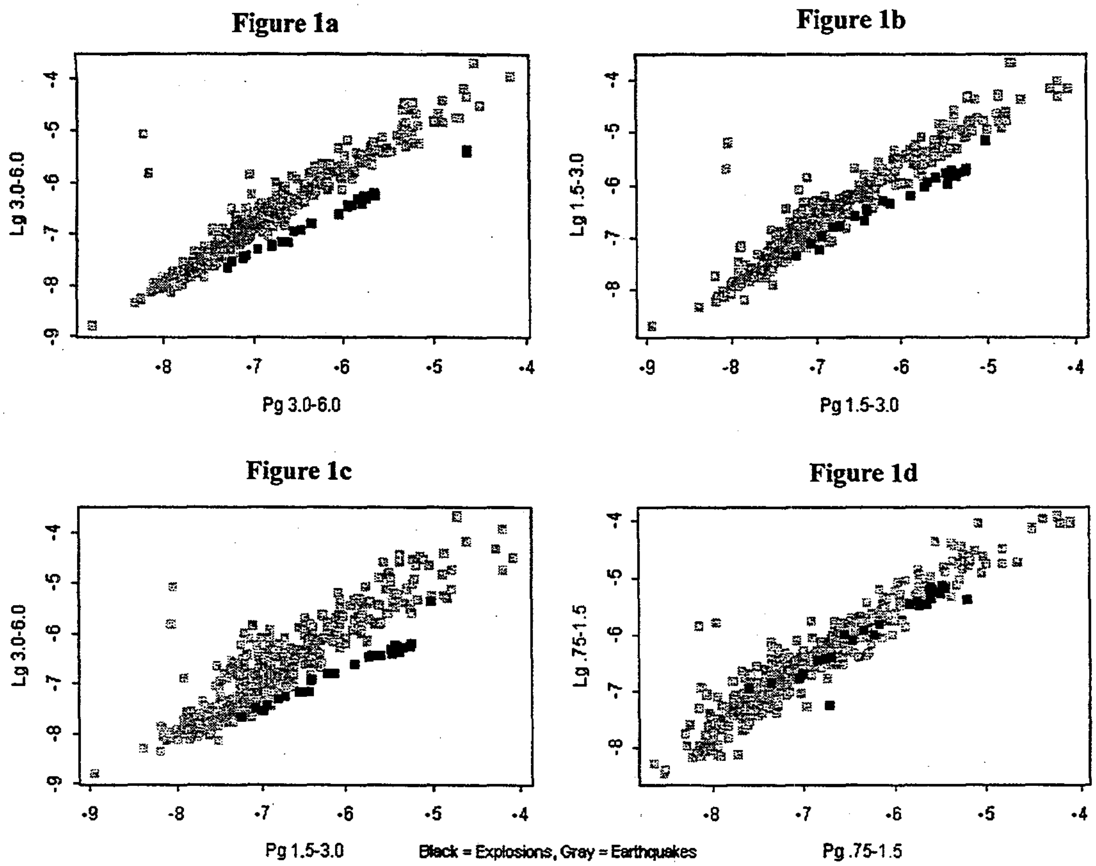

Figure 1. Earthquakes (shown in gray) and Explosions (shown in black) Plotted Against Seismic Discriminants

Table 1. Percentage of Correct Predictions from Cross-Validation Testing

	Percentage of Correct Predictions for each Outcome	
Best Predicting Variables	**Earthquakes**	**Explosions**
Model 1: $Pg_{3.0-6.0}$ & $Lg_{3.0-6.0}$	99.7%	98.0%
Model 2: $Pg_{1.5-3.0}$ & $Lg_{3.0-6.0}$	98.7%	73.0%
Model 3: $Pg_{1.5-3.0}$ & $Lg_{1.5-3.0}$	98.4%	65.8%

One concern about this analysis is that the coefficients and standard errors for the $Pg_{3.0-6.0}$ and $Lg_{3.0-6.0}$ model were large in comparison to the other models. Table 2 shows that the coefficients for Model 1 were considerably larger than the coefficients for the other models, and the standard errors were about 10 times larger. A possible explanation for this is that each of the seismic discriminants were highly correlated with one other, and therefore, the model may be experiencing instability due to the multicollinearity among the predictor variables. This multicollinearity can be viewed in the information matrix with correlations close to one off the diagonal. The information matrix was then used to construct the coefficient estimates and the standard errors of the estimates. This, in turn, could be causing an inflation in the resulting estimates.

Table 2. Regression Coefficients (β) and Standard Errors (σ_β) for Each Model

Model 1			**Model 2**			**Model 3**		
Term	**β's**	σ_β	**Term**	**β's**	σ_β	**Term**	**β's**	σ_β
Intercept	-16.74	21.68	Intercept	9.35	3.69	Intercept	-1.60	1.98
$Pg_{3.0\text{-}6.0}$	-84.97	35.22	$Pg_{1.5\text{-}3.0}$	-17.47	3.75	$Pg_{1.5\text{-}3.0}$	-13.33	2.11
$Lg_{3.0\text{-}6.0}$	79.27	32.10	$Lg_{3.0\text{-}6.0}$	17.45	3.79	$Lg_{1.5\text{-}3.0}$	12.65	2.11

Another concern is that this study is a case-controlled study (Hosmer and Lemeshow, 1989). Because the sampling design could not allow for the events to be sampled randomly, a bias is introduced in the regression equation. Thus, the predicted probability of an event is biased from the actual probability of the event. This bias is found in the intercept, β_o, and affects any inferences made using the β_o.

Corrections can be made to β_o when certain information about the sampling fractions for both events is known. This correction of β_o is done with the following equation:

$$\beta_o * = \beta_o - \ln(\tau_1 / \tau_0) \quad (7)$$

where: β_o^* is the new, corrected intercept;

τ_1 is the probability of the sampling design selecting an earthquake, given that the event was an earthquake and given x;

τ_0 is the probability of the sampling design selecting an explosion, given that the event was an explosion and given x; and

β_o is the biased intercept in the calculated regression equation.

If the proportion of earthquakes sampled was equal to the proportion of explosions sampled during the sampling time frame, then $\tau_1 = \tau_0$ and $\beta_o^* = \beta_o$, meaning that there would be no need for a correction in the intercept.

Further investigation of the τ's show that $\tau_1 = n_1 / N_1$, where n_1 is the number of earthquakes sampled for a given sampling time frame, and N_1 is the number of total earthquakes during the same time frame. Other equations demonstrate that $N_1 = N * \pi_1$, where N is the total number of events (earthquakes and explosions) for the given time frame, and π_1 is the proportion of all events that were earthquakes during the same time frame. Putting these equations together shows that:

$$\tau_1 = n_1 /(N * \pi_1) \tag{8}$$

Likewise, the equation for τ_0 is:

$$\tau_0 = n_0 /(N * \pi_0) \tag{9}$$

where: n_0 is the number of explosions sampled during a given sampling time frame;
N is the total number of events (earthquakes and explosions) for the same time frame;
and π_0 is the proportion of all events that were explosions during that same time frame.

These equations show that τ_1 and τ_0 are directly influenced by the prior proportions, π_1 and π_0, and that a change in the proportions will directly affect τ_1 and τ_0. A change to τ_1 and τ_0 will change the intercept, β_o, which, in turn, changes the predicted probability from the logistic regression model. This is a key concept for this study, because if at a later date the proportions of explosions change, this would cause a bias in the predicted probability for an event.

Although the above concerns show that the predicted probability could be biased, the prediction can still be useful, relative to other predictions. If Event 1 has a 0.60 probability of being an earthquake and Event 2 has a 0.90 probability of being an earthquake, we know that Event 2 has a better chance of actually being an earthquake. However, we need to remember that each probability has a bias attached to it, and with certain information that bias can be estimated.

Conclusions

Logistic discrimination has an easily understood algorithm and few assumptions. Under the right circumstances, it is a viable discrimination method. For this data, the best logistic regression model used the high frequency window $Pg_{3.0\text{-}6.0}$ and $Lg_{3.0\text{-}6.0}$ predictor variables. Cross-validation methods showed that this model was able to correctly predict the earthquakes 99.7% of the time and the explosions 98.0% of the time. Although this appears to be extremely effective, the model did show characteristics of being unstable. Because the model is dependent on the data, caution should be used when making decisions based on this model.

Two biases were also introduced with the predictions as possible concerns in using logistic regression in this particular problem. The first bias was due to the sampling design. The case-controlled study allowed for sampling that was conditional on the outcome variables, not a random sampling. The second bias would only be introduced when the proportion of each outcome changes. This is only pertinent when the percentage of explosions in the area of interest were to significantly increase or decrease, relative to the number of earthquakes. These two biases are important and need to be accounted for when using this data for predicting the source of future events. With the right information about the proportions for each event, and/or information about the sampling fractions, these biases can be estimated and accounted for. Without this correction, the predicted probabilities should not be considered as actual probabilities, but more as a ranking score that can be useful as a ranking of events. Scores on one end of the scale would more likely be an earthquake, and scores on the other end more likely to be an explosion. Even if the biases cannot be estimated, relative comparisons can be made between the scores.

References

Anderson, D.N., K.K. Anderson, D.N. Hagedorn, K.T. Higbee, N.E. Miller, T. Redgate, and A.C. Rohay. *Statistical Classification Methods Applied to Seismic Discrimination.* PNNL-11192, June 11, 1996.

Hartse, H. Los Alamos National Laboratory, (personal communication) June 1998.

Hartse, H., S.R. Taylor, W.S. Phillips, and G.E. Randall. *A Preliminary Study of Regional Seismic Discrimination in Central Asia With Emphasis on Western China*, LAUR-96-2002, June 12 1996.

Hosmer, David W. and Stanley Lemeshow (1989). *Applied Logistic Regression.* John Wiley and Sons, Inc., New York, NY.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley and Sons, Inc., New York, NY.

Taylor, S.R., M.D. Denny, E.S. Vergino, and R.E. Glaser. *Regional Discrimination Between NTS Explosions and Western U.S. Earthquakes.* BSSA, v.79, no. 4, pp. 1142-1176, August 1989.

Taylor, S.R. *Analysis of High-Frequency Pg/Lg Ratios from NTS Explosions and Western U.S. Earthquakes.* BSSA v.86, no. 4, pp. 1042-1053, August 1996.