False Event Screening Using Data Mining in Historical Archives

Timothy J. Draelos,[1] Michael J. Procopio,[1] Jennifer E. Lewis,[1] and Christopher J. Young[1]

INTRODUCTION

Analysts working at the International Data Centre (IDC) in support of treaty monitoring through the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) spend a significant amount of time reviewing hypothesized seismic events produced by an automatic processing system to ensure a high-quality event bulletin, which is then made available to the member states of the CTBT. Such a system is characterized as forming signal detections from the waveforms recorded at the International Monitoring System (IMS) stations, performing association or grouping of the detections, and finally converging on a set of hypothesized events that can account for the groups of detections. For each detection, a suite of features is automatically measured (*e.g.*, arrival time, azimuth, signal to noise ratio [SNR], etc.) and recorded in a relational database. These features are then used by the association algorithm to form event hypotheses, where an acceptable hypothesis is one that explains a suite of features (within acceptable tolerances) across the network. Thus, the quality of the automatically built events is directly dependent on the comprehensiveness and quality of these features.

A principal characteristic of current monitoring systems is that all events produced by automatic processing must undergo a varying degree of analyst review. Due to the desire to detect any potential nuclear event, such systems are by design very sensitive, and their tuning is heavily biased toward a low probability of missing a legitimate event. This bias comes with the tradeoff of a high rate of hypothesizing an event that later is not included in the final event bulletin (*i.e.*, a "false" event). This approach reflects the inherently greater cost of missing a real event versus removing false events and results in analysts screening out approximately half of the events hypothesized during automatic association.

Data mining methods, which rely in some form on archives of historical data, are used in false event screening to correct errors in the current automated event association system prior to analyst review (Procopio *et al.* 2009; Mackey *et al.* 2009; Brogan and Misrak 2010). Machine learning models are trained on features of events recorded at monitoring stations. Ground truth event labels, required in supervised learning, are based on whether or not an event exists in the analyst-reviewed bulletin. These trained models are subsequently applied to future hypothetical events in order to make predictions that closely match those of the analysts.

The recent successful application of data mining methods to the monitoring domain is largely due to the recent confluence of three key factors: large archives of training data, new methods, and ready access to computational power. None of these components were tangibly in place when the IDC automated processing system was first envisioned and designed in the late 1980s and implemented in the mid-1990s (Ringdal 1994). First, the IDC now has more than ten years of high-quality data available as a result of routine operation of its monitoring pipeline. These archives include raw waveform data as well as automatically generated and analyst-reviewed events with many associated measurements (features). Second, algorithms for data mining and machine learning have advanced significantly in the past 15 years. In particular, ensemble learning, Bayesian networks, and the invention of support vector machines (SVMs) are significant recent advances to the machine learning field (Bishop 2007). Finally, the computing power available on typical commodity machines has drastically improved to the point where the required computationally intensive processing of these large archives of data is practical.

It is important to recognize that IDC analysts are not limited to the set of features recorded in the IDC databases. In addition, the features they select in analyzing a particular event are heavily dependent on education/training and past experience. One method often used by analysts to gauge the validity of an event involves examining the set of seismic and other stations involved in the detection of an event. In particular, an analyst—leveraging past experience—can say that an event of a certain magnitude located in a certain part of the world is usually detected by Stations A, B, and C (perhaps high-quality stations near the source of the event). Implicit in this statement is that such an event is usually *not* detected by Stations X, Y, or Z (*e.g.*, because they are lower quality stations and/or stations distant from the event). Any significant deviation from this expected set of detections and non-detections suggests

1. Sandia National Laboratories, is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

doi: 10.1785/gssrl.83.2.267

that there is a problem with an event hypothesis. As we show herein, this procedure can be captured in a new class of features based on formalizing and quantifying the difference between the observed set of stations associated in the detection of some hypothesized event versus the expected set of stations historically associated with detecting similar events.

The research presented in this paper is motivated by the need to produce an accurate event bulletin more quickly, implying less analyst review. Ideally, our methodology will decrease the number of false events that an analyst must screen out to produce a final bulletin, hence increasing the speed with which such a bulletin is produced. This improved bulletin timeliness is important for nuclear explosion monitoring, because our ability to gather and clearly identify the tell-tale radionuclide products of a nuclear explosion decreases rapidly as those products decay and are dispersed within the atmosphere (*e.g.*, Perkins *et al.* 1996). Since the number of seismic events increases exponentially as event size (magnitude) decreases, the challenge becomes greater as event detection thresholds are decreased to verify compliance with the CTBT, which has no threshold at all. Lowering detection thresholds across a network will result in a tremendous increase in the number of events that are built by the automatic system and reviewed by analysts. Therefore it is essential that minimal time is spent finding and removing false events.

Increasing the number of analysts reviewing the automatic results is not a practical option. The analyst role takes time to learn and has a high turnover rate, so the most efficient highly trained analysts are rare, while operational budgets place firm financial constraints on the number of analysts able to engage in event review. Thus, the approach presented in this paper is to leverage data mining techniques to improve the quality of the automatically produced results that require review by analysts (*i.e.*, to make better use of the analysts' time). Since 1999, there have been two nuclear explosions and on the order of 400,000 other events, mostly earthquakes. The bulk of the analyst burden at the IDC is in handling these earthquakes and making sure each is not a nuclear explosion (which, if the pattern of activity since 1999 continues, it won't be 99.999% of the time). Our work has the potential to help these analysts by limiting the number of bogus events they have to examine.

DATA MINING FOR FALSE EVENT SCREENING

The false event screening problem is framed as a binary (2-class) classification task. The classifier is presented with a set of domain-specific features that describe a hypothesized event (*e.g.*, time residual, slowness uncertainty, SNR, etc.), and makes a decision on that data. The decision is either "Classifier Predicts False (Bogus) Event" or "Classifier Predicts True (Legitimate) Event." Features useful for false event classification are extracted from the IDC database of past data processing results. These are the same features that the association algorithm used to build the event hypotheses that our system will attempt to classify. At the IDC, seismic event data are organized in a sequence of lists of increasing completeness and quality. The final human analyst-reviewed bulletins offer ground truth that is necessary for supervised machine learning. The event lists available at the IDC are as follows.

Standard Event Lists (SEL1, SEL2, and SEL3)—progressively enhanced lists of computer-generated events (both false and true seismic events). SEL1 events include detections from primary seismic stations only, while SEL2 events may include detections from auxilliary stations as well (data that were automatically requested and processed based on the SEL1 event list to corroborate those events). SEL3 events may include detections from any late arriving seismic data that were not available at the time of SEL1 or SEL2 processing. SEL3 is the final automated event list for review by analysts.

Late Event Bulletin (LEB)—a list of SEL3 events that are reviewed by analysts and judged to be legitimate, plus additional events the analysts constructed.

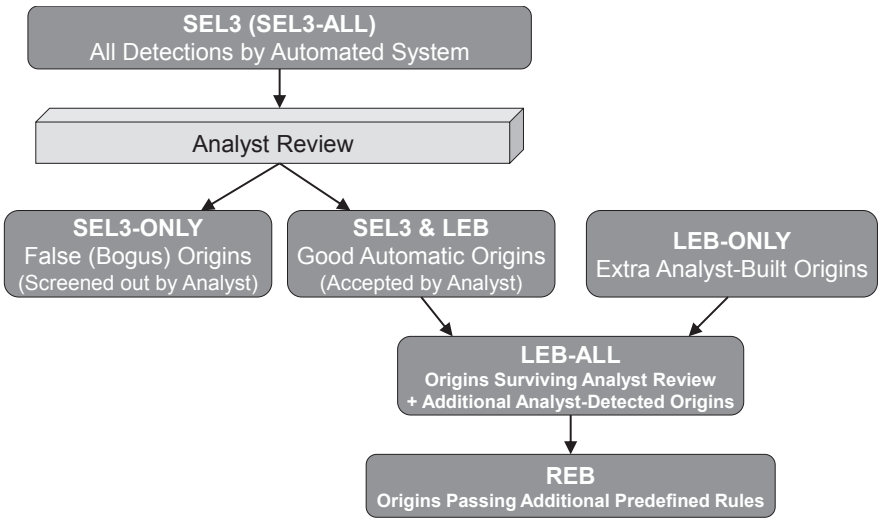
Reviewed Event Bulletin (REB)—a subset list of events from the LEB that pass specific quantitative criteria related to observations used to define event location and origin time. For seismic events, three or more first-arriving *P*-type phases at primary network stations are required (if only three, at least two must be arrays and the first arriving *P* phases must include time, azimuth, and slowness measurements (Le Bras and Wuster 2002).

Figure 1 illustrates the process of establishing the IDC SEL3, LEB, and REB event lists. We focus on the transition between SEL3 and LEB, as this is where the analyst effort is applied.

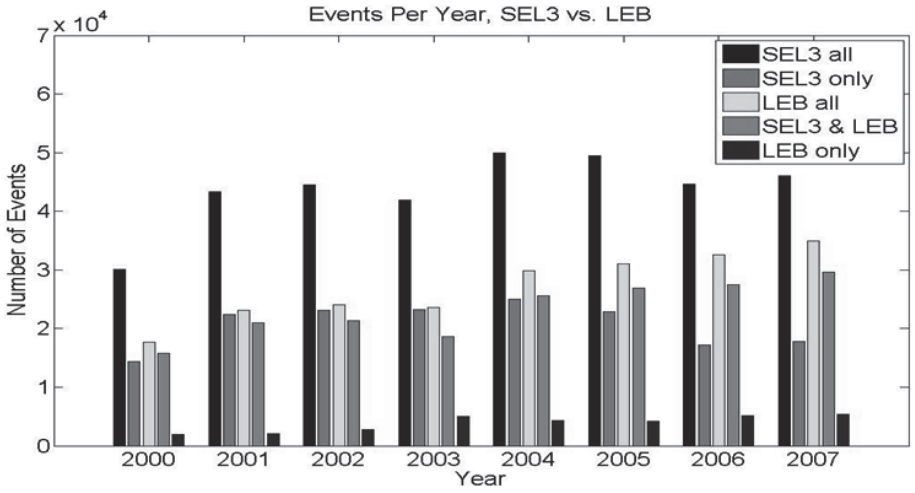
The primary goal of the current work is to minimize the number of SEL3 events requiring screening by analysts. This is done by identifying false events and removing them from the SEL3 list presented to the analysts for review. Figure 2 shows summary statistics of SEL3 and LEB events over an eight-year period. The second bar for each year represents the number of events screened out. None of these events end up in the LEB or REB, yet analysts reviewed each one, a significant time commitment. The proportion of screened events is roughly half and relatively flat over the course of this analysis, indicating that the situation has not significantly improved.

Considerations for Data Mining in the Monitoring Domain

It is important to consider the applicability of learning-based statistical data mining methods in two distinct contexts: the current generation of monitoring systems in active use and the next generation of monitoring systems in active development. Current monitoring systems, such as the one at the IDC, play a crucial role in ongoing monitoring operations. Data mining approaches are in a strong position to benefit these systems, but their use must be tailored to the architecture of current production systems, carefully augmenting them where they stand to benefit the most. The false event screening problem addressed in this research is a canonical example. Here, data mining algorithms can operate in conjunction with the existing monitoring pipeline and function as an automated screening layer in between association and analyst review.



▲ **Figure 1.** Process of establishing seismic events at the IDC.



▲ **Figure 2.** Event list quantity and composition, by year. "LEB only" are events that do not correspond to SEL3 events; these are the events manually constructed by the analyst.

The work presented in this paper can augment an existing system with an additional module to correct system shortcomings, and we show that the reduction in analyst workload can be significant. However, a more complex, bottom-up solution to replace the current signal association system may lead to more significant gains in monitoring systems. Two such systems driven heavily by both historical, analyst-corrected data and real-time data are under research and development (Arora *et al*. 2010; Draelos *et al*. 2011). Rather than using data mining approaches to clean up signal association mistakes, the ultimate goal of those studies is to use those approaches to make the association better so fewer mistakes are made in the first place.

CLASSIFICATION USING IDC DATABASE FEATURES

Many features are available from the IDC database for hypothesized events (origins) and their associated signals (detections, also known as arrivals) seen at the individual sensors. The SEL3 event bulletin provides a list of all events requiring classification, while the LEB event list provides the subset of these events that pass analyst review. Hence the difference between these lists is the set of events that we seek to learn to screen. Of potential use for this study is the LEB analyst discard table, which lists one of five reasons given by analysts for screening a false event. Although the information in the table is useful in understanding the problem of event characterization, none of the information is available in the SEL3 and no features were identified for use in false event screening. Other features available in IDC databases include the following.

Features closely tied to the event hypothesis: event location, time, and magnitude along with associated uncertainty information.

Features generated when arrivals are associated with and used to locate an event: residuals associated with time, azimuth, and slowness.

TABLE 1
IDC Database Features and Decision Tree Classification Percentages of Surviving and Screened Out Events

Feature Name	Description	Events Classified Correctly	Surviving Events Classified as Screened
ndef	Number of time-defining phases associated with an event	76.2 %	11.9 %
Lowest deltim	Lowest time uncertainty from all the time-defining arrivals for an event	74.5 %	5.7 %
Uncertainty ellipse area	Size of the error ellipse for an event's location (smajax * sminax * pi)	67.7 %	6.8 %
Lowest slores	Lowest slowness residual from all the time-defining arrivals for an event	72.7 %	6.3 %
Lowest delslo	Lowest slowness uncertainty from all the time-defining arrivals for an event	71.3 %	8.2 %
Best SNR	Best (highest) SNR from all the time-defining arrivals for an event	74.5 %	5.7 %
2nd best SNR	2nd best (highest) SNR from all the time-defining arrivals for an event	78.6 %	8.5 %
3rd best SNR	3rd best (highest) SNR from all the time-defining arrivals for an event	80.7 %	9.4 %
4th best SNR	4th best (highest) SNR from all the time-defining arrivals for an event	78.4 %	16.6 %
All Features	9-dimensional feature vector	82.0 %	9.7 %

Features from individual arrivals at stations independent of event hypotheses: SNR and uncertainties associated with arrival time, azimuth, and slowness measurements.

In order to use a consistent number of features but not use any more than are necessary, decisions were made regarding how to handle the surplus of observables for large events and the limited number of observables for small events. For example, if an event is associated with 50 arrivals at 10 stations, then hundreds of measured observables are available as features. However, smaller events may be associated with only five arrivals at two stations. One way to manage the number of features is to average observables across all arrivals, resulting in a single feature value for an event (*e.g.*, the average SNR). Not surprisingly, we found that averaging observables does not work well for discriminating between screened out and surviving events. Instead we chose to use maxima or minima of observables across all arrivals, which were more effective ways to reduce to a single feature for a set of arrivals. For example, the best SNR among all arrivals will represent reasonably well the station or arrival quality contributing to the detection of an event.

We used standard feature selection and transformation techniques, such as principal components analysis (PCA), as well as statistical inspection of the data and seismological domain knowledge to choose the features most likely to contribute to a classifier's ability to differentiate between those automatically generated SEL3 events that survive analyst review and those that are screened out after review (Guyon and Elisseeff 2003). The features we chose to use are listed in Table 1, which also shows the classification results of using a simple decision tree classifier with the Weka machine learning software toolkit (Hall *et al.* 2009) on each feature individually and with all features together. Events randomly sampled from the 2004–2007 SEL3 lists were used to test and train the classifier, with true/false ground truth labels based on the LEB list. Classification was performed using 10-fold cross-validation on the data in which a random selection of 90% of the data is used to train the classifier and the remaining 10% is used to test performance for each fold; the final result is the average of the test performances over all 10 folds. For each test sample, the feature vector is presented to the trained classifier and the output is compared to the sample's ground truth label. Errors occur when bogus events are classified as surviving and when surviving events are classified as screened. We wish to minimize all errors, but particularly the latter one.

Results illustrated in Table 1 provide a number of key insights. First, classification based on single features can be statistically very effective. For example, the number of location-defining phases (*ndef*) alone can correctly classify 76% of the events in the dataset. However, the resulting 24% of events misclassified includes incorrectly screening out a sizable percentage of legitimate events (~12% of all events), an unacceptable number of discarded events. Further, it is this very class of poorly detected events that are of most interest for CTBT monitoring. Other single features performed similarly. The best single feature result (80.7%) was achieved using *3rd best SNR*. Apparently both true and false events may have good (high) *1st* and *2nd best SNR*s, but for false events, the *3rd best SNR* is typically poor (low) compared to true events. The lack of a good third arrival is a strong signal that the event may not survive analyst review.

Using all nine features together results in the best classification accuracy, 82%. Of the 18% of events that are incorrectly classified, approximately 10% are events listed in the LEB but screened out by our classifier. This is the classification error that must be minimized even at the expense of reducing the overall performance. Although using all features together improves performance over single features, the improvement in correct classification accuracy is only 1.6% better than the accuracy with the *3rd best SNR* feature alone. This indicates that it is dif-

ficult to find features that will dramatically improve results. In other words, there is a set of events for which extra information is necessary—information not available in the IDC database.

While we did notice a slight variation in performance across different classification algorithms, the choice of algorithm is not considered a central question in this research. In contrast to further attention to classification algorithms, the above results identified improved features as the primary area for future work most likely to result in improved performance. We found that database features can be used to differentiate between events that are obviously false or obviously legitimate. The difficulty lies with events with similar features that sometimes survive and sometimes are screened out. Analysts can benefit from knowing where to focus their attention and decide between these events, but unless a difference is reflected in feature values for screened out events versus surviving events, it may be necessary to make additional measurements directly from the waveforms to classify correctly these more ambiguous events. It is possible that new, valuable features related to inherent analyst knowledge derived from experience (in terms of the likelihood of detecting/non-detecting stations) can be extracted from database archives and applied to the classification problem.

STATION SET RESIDUAL (SSR)—A NEW FEATURE

We propose a new class of feature, named Station Set Residual (SSR), motivated by our observations of the decision-making process used by analysts during event review and related inputs into that process. We found that the set of detecting stations involved in the creation of a hypothesized event is correlated with its location and magnitude and plays an important role in the analysts' review process. Implicit in this finding, the comparison of the current set of detecting (or observing) stations versus the analysts' expected set of stations was heavily considered as evidence when determining a hypothesized event's legitimacy. Importantly, the notion of expected stations is not necessarily binary but is tied to varying degrees of expectation.

We propose that the analysts' expected set of stations can be modeled probabilistically by statistical consideration of analyst-reviewed archive data and that the resulting models can be conditioned on the hypothesized event's location and magnitude. Our hypothesis is that the difference (residual) between the stations expected to see an event (based on historically generated probabilities) and the stations that actually saw the event is meaningful and correlated with event survival to the final event bulletin produced by the monitoring process.

Example—Surviving Event

Consider an event occurring on 12/3/2008 with Origin ID (orid) = 5049761, 4.47° south latitude, 146.06° longitude, and $m_b = 3.6$ from the SEL3 that survives analyst review and is listed in the LEB. This event was observed by five stations out of a total of 12 stations that historically observed events of similar magnitude originating in a similar location. We use "observed" to mean that a station provided a detection, or arrival, that was associated with the event. Figure 3A shows a map of the event with observing and non-observing stations (ILAR not shown). The color and size of the symbols plotted at each IMS station make it easy to assess how well the set of observations matches the expected set.

The historical station probabilities of event 5049761 are summarized in Figure 3B. The height of each bar represents the historical probability that a station was involved in the detection of a surviving event. The probability, conditioned on latitude, longitude, and magnitude, is simply the ratio of historical events observed by a station to total number of historical events, and hence the range of values is 0 to 1. The difference between this set of historical probabilities for a given location and magnitude and an event proposed to have occurred with those coordinates is the SSR. We next show how to form quantitative features based on SSR.

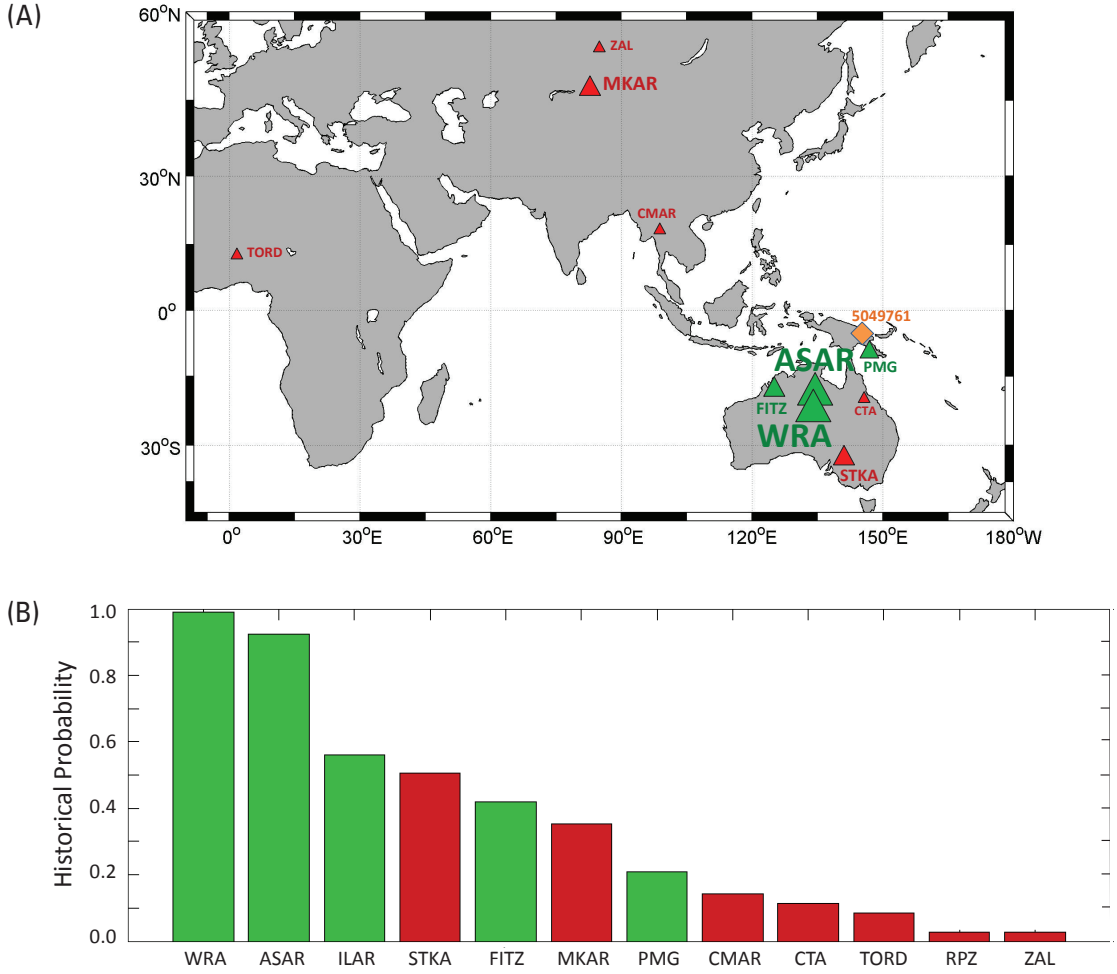
Feature Extraction from SSR

To complement currently available features from the IDC database, we extract a novel numerical feature from SSR, which we refer to as the Contiguous Score Set (CSS). This feature is influenced by the domain requirement that a minimum number of observing stations is required in order for an event to survive analyst review. The CSS feature focuses on relative differences of probabilities among the stations and quantifies how much "difference" there is between historically expected stations versus those currently observing an event. The basic concept is to transform the ordering of observing stations such that there are no non-observing stations between observing stations and to calculate the cost involved in doing this. For example, for the bar chart in Figure 3B, CSS2 = 0 and CSS3 = 0 because no adjustments are needed to achieve two and three consecutive green bars on the left, respectively. To achieve a contiguous set of four, the non-observing STKA station is swapped with the observing station FITZ to its right, incurring a penalty equal to the probability difference between them, so CSS4 = 0.08. CSS5 is computed similarly, but with more swaps of bigger differences. Note that there is little penalty incurred by swapping stations with comparable probabilities, but a significant penalty when swapping dissimilar probabilities. Not observing an event at a 0.5 probability station, but observing it at a 0.47 probability station, is not surprising. Conversely, not observing an event at a 0.9 probability station but observing it at a 0.4 probability station is suspicious.

Classification with SSR

We conduct our analysis using an eight-year archive of IDC data, from 2000 to 2007. Events from 2000 to 2003, inclusive, represent the data from which the historically expected station distributions are computed. Events from 2004 to 2007, inclusive, represent the data on which machine learning models are trained and evaluated.

Events in SEL3 that survive to the LEB are used to compute the historically expected station distributions (as a function of latitude/longitude and magnitude). Importantly, this implies that the historically expected station distributions are



▲ **Figure 3.** (A) Map of event 5049761 (orange diamond) and stations (triangles) that observed the event (green) and did not (red). The size of the triangle and its accompanying text is proportional to that station's degree of historical expectation for the event: bigger means more likely. (B) Historical station probabilities related to event 5049761 (green indicates the station did observe the event and red indicates it did not).

determined (1) solely from surviving events and (2) from SEL3 (not LEB) latitude/longitude and magnitude values. We tried several CSS features (CSS3, CSS4, etc.) and found that CSS3 performed the best: 77.8% classified correct decision tree accuracy with 11.7% surviving events classified as screened. However, this classification result is similar to those shown in Table 1 for the standard database features. This is a surprising result and we discuss it in the next section.

FINDINGS AND DISCUSSION

There are two options for establishing historical probabilities of stations observing events, neither of which resolves the difficulty of distinguishing between problematic events.

Using SEL3 information for historical data. In this case, station probabilities are incorrectly biased by badly located/measured events because we know that the automatically produced events in SEL3 suffer from both problems. This can explain why classification with an SSR feature performs no better than classification with some database features.

Using REB/LEB for historical data. In this case, the historical station probabilities will be accurate, using correct location and magnitude values for events, but badly located/measured events in SEL3 will access the wrong station probabilities. For this reason and since mislocated events likely had few stations observing the event, the SSR feature subsequently computed does not improve classification performance.

This finding illustrates the difficulty in trying to use incorrect SEL3 values (*e.g.*, lat/lon and magnitude) for establishing SSR probabilities and using them for classifying hypothetical SEL3 events. The actual values used by analysts, as reflected in the LEB event list, are often significantly different from those available in the SEL3 event list. Regardless of the value of SSR as a discriminator in false event screening, the SSR bar charts and maps are useful as visual aids to analysts, allowing them to quickly see inconsistencies with expected station responses to

TABLE 2
Results of Cost-Sensitive Machine Learning on False Event Screening

Penalty for Classifying Surviving Events as Screened	Events Classified Correctly	Reduction to Analyst Workload (Events Classified as Screened)	Workload Reduction Cost (Surviving Classified as Screened)
1 x	83.6 %	52 % (~62 / day)	9 % (~11 / day)
10 x	73.5 %	26 % (~31 / day)	1 % (~1.2 / day)

an event. These visualizations are a powerful means of capturing the historically most probable stations to detect a specific event, the expected station probabilities, and their locations. The contrast between expected and observed results can help an analyst develop an early sense for event legitimacy. It can also help identify the most promising candidate stations to inspect if a historically highly probable station is absent from an event that appears otherwise legitimate. This is a particularly effective tool for new analysts who have not yet learned the expected patterns, a process that can take many months.

COST-SENSITIVE CLASSIFICATION USING ALL FEATURES

Despite the fact that using SSR-based features on their own did not improve performance, it still may be the case that in combination with the other features, they might improve classification capability. Combining the direct database features with the newly developed SSR features, we explored an assortment of classification algorithms such as support vector machines, random forests, decision trees, and multilayer perceptron neural networks. All of the algorithms resulted in similar accuracy results regarding correct classifications. By utilizing the entire suite of features presented in Table 1 along with the preferred SSR feature (CSS3), 83.6% correct classification was achieved, a slight improvement over the use of a decision tree on all the database features. Assuming a typical rate of approximately 120 events (60 false and 60 true) automatically detected per day, this results in the screening of about 62 events per day that analysts would not need to review, a significant benefit to the IDC. However, the classifier also screened out approximately 11 legitimate events per day, which is unacceptable.

A cost-sensitive machine learning approach was developed to reduce the number of legitimate events classified as screened since this is the scenario that we must avoid in nuclear explosion monitoring. During training, a penalty for classifying true events as false is imposed that is higher than the penalty for classifying false events as true. Different penalty ratios were evaluated. Table 2 contains classification results from using cost-sensitive training with a 100-tree random forest classifier. Column 1 lists the penalty ratios for two different experiments (1x indicating that the penalty for classifying true events as false is the same as the penalty for classifying false events as true; and 10x indicating that the penalty for classifying true events as false is 10 times that of classifying false events as true). Column 2 lists the events classified correctly, both surviving and screened out. Column 3 lists events classified as screened, correctly and incorrectly. Having been screened, these events may not be necessary to review, thereby reducing the analyst workload. Of course the analyst workload would go to zero if all events are screened out, so the numbers in Column 3 must be compared with the numbers in Column 4, which lists the number of incorrectly screened true events. The Column 4 events represent a cost incurred for our attempt to reduce analyst workload by screening bogus events: sometimes we screen real events.

Classification was performed using the same cross-validation methodology described in the section on classification using IDC database features. By using a penalty ratio of 10, we were able to reduce the rate of events falsely screened out down to about one per day while still screening about 31 bogus events. The important implication of this result is that it is possible to use machine learning classification to eliminate a substantial number of events from the analyst workload without incurring a high probability of missing a legitimate event.

Classification with LEB Data—The Issue with Changing Features

A question arises about the quality of the features selected for classification and the feature values represented in SEL3 data. One way to test the general quality of the features is to perform classification between SEL3 screened out events and SEL3 surviving events using LEB (as opposed to SEL3) feature values for these surviving events. The classification result on this dataset is nearly 100% correct, suggesting that the feature set in general is fine, but that the feature values available in SEL3 data are not. The feature values of those events in the SEL3 data that survive analyst review change enough by the time they end up in the LEB to make discriminating between screened out and surviving events highly reliable. During their review, analysts producing the LEB often adjust and refine observables after inspecting waveforms using various filter settings and leveraging other resources. In addition, the interactive software used to produce the analyst-reviewed LEB sets some observables differently than the automatic software used to produce SEL3. We conclude that for a significant subset of events, the SEL3 feature values are so poorly determined that they are not effective for false event screening.

CONCLUSIONS

In this study, we outlined the false event screening problem in which approximately half the events produced by automatic processing at the IDC do not survive analyst review, suggesting that a great deal of analyst effort is needed to correct the auto-

matic results. The analyst's time is both limited and valuable, and improving analyst productivity by identifying false events via automated statistical data mining methods can lead to more efficient use of their time. We proposed a framework for false event screening based on supervised machine learning in which models are trained on historical data and then used to make predictions on future hypothesized events whose legitimacy has not yet been established. This approach can easily integrate with existing monitoring pipelines, such as the one in place at the IDC, as an external module or layer that requires minor changes to the underlying pipeline currently in use.

Our work has shown that, through the use of a cost-sensitive event classifier, it is possible to reduce by roughly one quarter (~31) the number of events an analyst must consider on a daily basis (~90 instead of ~120 currently) while missing very few legitimate events (~1). This approach is, therefore, viable for inclusion in current monitoring systems, as well as in next-generation monitoring systems in development. We discussed features available via IDC databases that are useful for event discrimination and a new type of derived feature, the Station Set Residual (SSR), which represents the conceptual difference between the currently observing set of stations associated with a hypothesized event and the historically expected set of stations for events similar in location and magnitude to the hypothesized event. A visual representation of SSR is useful as an analyst aid, which can help an analyst develop an early sense for event legitimacy and help identify the most promising candidate stations to investigate further. We also proposed a novel series of features for quantifying the SSR, the Contiguous Score Set (CSS), which relies on relative differences among the expected probabilities of the historically observing set of stations. We showed that the inclusion of CSS features improved classification performance overall when considered along with the existing feature set.

Finally, our findings support the notion that there is a significant number of events, typically those having two observing stations, that have varying or ambiguous outcomes that are not well predicted by our machine learning screening algorithms. These events are generally very close to the monitoring authority threshold required for having sufficient evidence needed to permit an event to survive review and endure to the final event bulletin. Systematic or natural measurement variation present in the automated processing system, including the associator, as well as human factors, are possible explanations for such problems.

A significant finding suggested by the difficulty of improving the performance of false event screening beyond 85% without cost-sensitive learning, and 74% with cost-sensitive learning, is that feature values from SEL3 surviving events are sometimes incorrect when compared to their eventual values in the LEB. In particular, we suggest that for low-magnitude events, classification of hypothesized SEL3 events is often based on badly formed data. Developing novel, more robust features extracted from waveforms, such as those that capture the frequency content of the arrivals associated with events, may improve false event screening, as may employing other machine learning approaches, such as reinforcement learning, which can utilize feedback from human analysts or other corrective sources. However, given that event screening is needed only because the automatic system is producing so many false events, an alternative option to correcting misclassifications of current associators is improving the event association element of the pipeline. This is the subject of ongoing research, in particular, Vertical Integrated Seismic Analysis (VISA) (Arora *et al.* 2010) and Probabilistic Event Detection, Association, and Location (PEDAL) (Draelos *et al.* 2011).

REFERENCES

Arora, N., S. J. Russell, P. Kidwell, and E. Sudderth (2010). *Global Seismic Monitoring as Probabilistic Inference*. Technical Report No. UCB/EECS-2010-108, Electrical Engineering and Computer Sciences, University of California at Berkeley.

Bishop, C. (2007). *Pattern Recognition and Machine Learning*. New York: Springer.

Brogan, R., and F. Misrak (2010). Identifying SEL3 false events: performance testing results of the SVM-based False Event Identification program. CTBTO Workshop on Machine Learning and Earth Structure, September 2010.

Draelos, T., S. Ballard, M. Gonzales, and C. Young (2011). Probabilistic event detection and signal association. *Proceedings of the 2011 Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, 227–236.

Guyon, I., and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1,157–1,182.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA data mining software: An update. *SIGKDD Explorations* **11** (1), 10–18.

Le Bras, R., and J. Wuster (2002). *IDC Processing of Seismic, Hydroacoustic, and Infrasonic Data,* Revision 1. IDC Documentation User Guides (May 2002).

Mackey, L., A. Kleiner, and M. I. Jordan (2009). Improved automated seismic event extraction using machine learning. *Eos, Transactions of the American Geophysical Union* **90** (50) Fall Meeting Supplement, abstract #S31B-1714.

Perkins, R. W., H. S. Miley, W. K. Hensley, and K. H. Abel (1996). Airborne radionuclides of concern and their measurement in monitoring a comprehensive test ban treaty. In *Monitoring a Comprehensive Test Ban Treaty*, ed. E. S. Husebye and A. M. Dainty, 143–155. Dordrecht and Boston: Kluwer Academic Publishers, 836 pp.

Procopio, M., C. Young, and J. Lewis (2009). Using machine learning to improve the efficiency and effectiveness of automatic nuclear explosion monitoring systems. *Proceedings of the 2009 Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, 788–797.

Ringdal, F. (1994). GSETT-3: A test of an experimental international seismic monitoring system. *Annali di Geofisica* **37** (3), 241–245.

Sandia National Laboratories
P.O. Box 5800, MS 0401
Albuquerque, New Mexico, 87111 U.S.A.
tjdrael@sandia.gov
(T. J. D.)