SAND2009-8167C

USING MACHINE LEARNING TO IMPROVE THE EFFICIENCY AND EFFECTIVENESS OF AUTOMATIC NUCLEAR EXPLOSION MONITORING SYSTEMS

Michael J. Procopio, Christopher J. Young, and Jennifer E. Lewis

Sandia National Laboratories

Sponsored by National Nuclear Security Administration

Contract No. DE-AC04-94AL85000

ABSTRACT

An analysis is performed on the seismic-event data processed from 1999 through 2009 by the International Data Centre (IDC) of the Comprehensive Nuclear-Test-Ban Treaty Organization. One purpose of the analysis is to determine if there are characteristics of the data that could be utilized to understand and improve automatic seismic-event processing. Another purpose is to determine whether improved station calibration could improve the automatic processing. It is hoped that knowledge gained by this effort could be applicable to similar automatic processing systems.

The overall quality of the IDC bulletin is excellent, but achieving this quality requires a significant amount of analyst effort. Initial examination of the data shows that automatic processing produces 421,244 origin hypotheses, or about 118 per day over the nearly 11 year period of operation. Of these, 224,643 (53%) do not survive analyst review (i.e. false positives), while the remaining 196,601 (47%) are approved. In addition, analysts build another 30,606 origins (13% of analyst approved total) that the automatic processing missed (i.e. false negatives). Thus analyst-approved origins occur at a rate of about 64 per day. From these figures it is evident that significant improvement of the automatic system, and hence decrease in the analyst workload, is possible both by decreasing the number of false events as well as by decreasing the number of missed real events. Currently, analysts must correct non-trivial numbers of both types of errors by the automatic system.

Previous work in analyzing a much smaller portion of the IDC data from 2002 (Gauthier, 2009) found that some attributes, or *features*, appear well-suited to *discriminating* among the different families, or *classes*, of events. This evaluation used principled but ad-hoc methods to show a basic ability of certain features, such as number of stations and signal-to-noise ratio, to predict the true class of a particular detected origin, which may or may not be valid. In particular, a method was devised that identified with very high confidence many “false origins,” or *false positives*, which did not survive analyst review in the final REB (Reviewed Event Bulletin) table.

Building off the premise that there are powerful features in the origin and arrival data, this new study uses modern machine learning and pattern recognition methods to train *models* on previously labeled data in existing archives, and then uses those models to make predictions on future data whose classes are not known. Crucially, and in contrast to previous work, such methods are able to consider multiple features simultaneously, which we hypothesize will yield better prediction performance. Importantly, this new work focuses heavily on identifying these features (typically from the arrival table, and also include the number of observing stations, pick error, slowness, azimuth, etc). In particular, principled *feature selection* methods exist which will identify the most useful features for classification, while also identifying features which have little or no discriminatory power at all.

The machine learning approach, which will consider an array of supervised learning algorithms including Support Vector Machines and Random Forests, should result in a marked decrease in the *false positive rate*, that is, the number of incorrect detections with regard to the total number of detections. In parallel, this approach will enable a greater *true positive rate*, or correct detection rate. Together, such an improved system will reduce both the analyst burden in sifting through false detections, while also reducing the risk that a true event goes undetected by both the automatic system and the analyst.

The techniques presented are rigorously and quantitatively evaluated via *cross-validation*, using known, robust performance metrics in the machine learning community such as Receiver Operating Characteristic (ROC) curves and their respective summary statistic, Area Under the ROC Curve (AUC).

OBJECTIVES

The objectives of this work are three-fold. First, we perform an analysis of a ten-year archive of IDC data in order to show that a large percentage of the events produced by the automated event detection system are bogus, and manual analyst effort must be spent on screening out such events. Further, the analysis characterizes the number of events missed altogether by the automatic system.

Second, we propose the application of supervised machine learning to train predictive models on data from large, tagged event libraries, that are capable of automatically screening out false events. We give several examples of machine learning methods, ranging from simple methods yielding highly interpretable models, to more elaborate methods whose models may be more difficult to interpret.

Finally, based on the empirical results, we suggest a roadmap to help guide future research in this area in order to improve the performance of machine learning methods on screening out false events from the automatic detection system.

RESEARCH ACCOMPLISHED

Introduction

The verification regime of the Comprehensive Nuclear Test-Ban Treaty (CTBT) includes the International Monitoring System (IMS) and the International Data Center (IDC). The purpose of this paper is first, to review the quality of the IDC bulletin of waveform-technology (seismic, hydroacoustic, and infrasounic) detected events throughout the history of operations to determine if there are characteristics of the data that could be utilized to understand and improve automatic seismic-event processing, and second, to determine whether improved station calibration could improve the automatic processing.

Previous work analyzing a month of IDC data from 2002 (Gauthier 2009) found that some attributes, or features, appear well-suited to discriminating among the different families, or classes, of events. That evaluation used principled but ad-hoc methods to show a basic ability of certain features, to identify the true class of a particular automatically detected origins, which may or may not be valid. In particular, a method was devised that identified with very high confidence many “false origins,” or false positives, which did not survive analyst review in the final REB (Reviewed Event Bulletin) table.

This new study uses modern machine learning and pattern recognition methods to train models on labeled data drawn from the full IDC data set (over 10 years), and then use those models to make predictions on withheld data whose classes are not known.

IMS Network and IDC Processing

The current IMS network of seismic (primary and auxilliary), hydroacoustic, and infrasonic sensors provides good overall global coverage, given the limited distribution of land area. Station density is significantly higher in some regions, though not necessarily in proportion to monitoring priority.

The IDC pipeline processes waveforms to produce detections, which are then associated into events. The pipeline is designed to meet requirements for both responsiveness and quality of event bulletin. Three automatic event lists are produced with increasing quality as more data becomes available. The final of these (SEL3) is then reviewed by analysts to produce a high-quality reviewed event bulletin (REB). For this study, we focus on differences between SEL3 and REB.

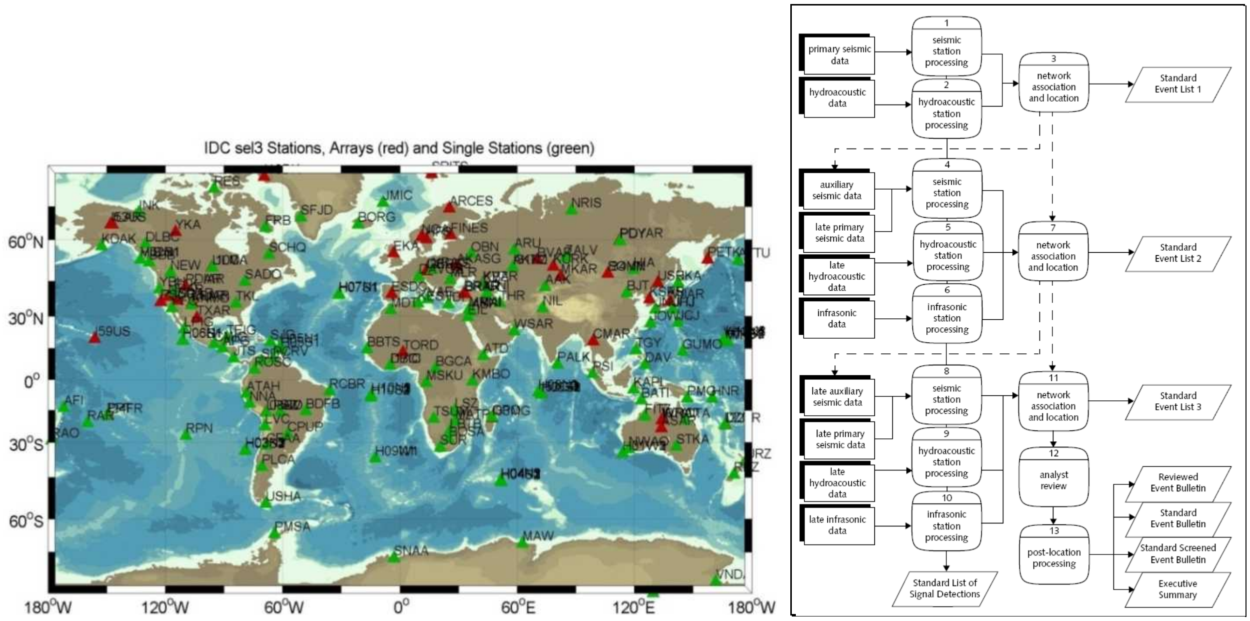

IDC sel3 Stations, Arrays (red) and Single Stations (green)
60°N
30°N
0°
30°S
60°S
180°W
120°W
60°W
0°
60°E
120°E
180°W
1
seismic station processing
2
hydroacoustic station processing
3
network association and location
Standard Event List 1
primary seismic data
hydroacoustic data
4
seismic station processing
5
hydroacoustic station processing
6
infrasonic station processing
7
network association and location
Standard Event List 2
auxiliary seismic data
late primary seismic data
late hydroacoustic data
infrasonic data
8
seismic station processing
9
hydroacoustic station processing
10
infrasonic station processing
11
network association and location
Standard Event List 3
late auxiliary seismic data
late primary seismic data
late hydroacoustic data
late infrasonic data
12
analyst review
13
post-location processing
Standard List of Signal Detections
Reviewed Event Bulletin
Standard Event Bulletin
Standard Screened Event Bulletin
Executive Summary

Figure 1. IMS network, including seismic, hydroacoustic, and infrasonic stations (left).
IDC pipeline, from "IDC Proc. of Seism., Hydro., and Infra. Data", Le Bras et al., 2002 (right).

Analysis of IDC Event Catalogs

On average, only 47% of the automatically built events in SEL3 survive analyst review, and this proportion seems to be fairly stable over the operating history of the IDC, despite the addition of new stations. Those events that are approved by analysts form some 87% of the REB. The additional 13% are newly built by the analysts, though they may make use of arrivals that were included in SEL3 events.

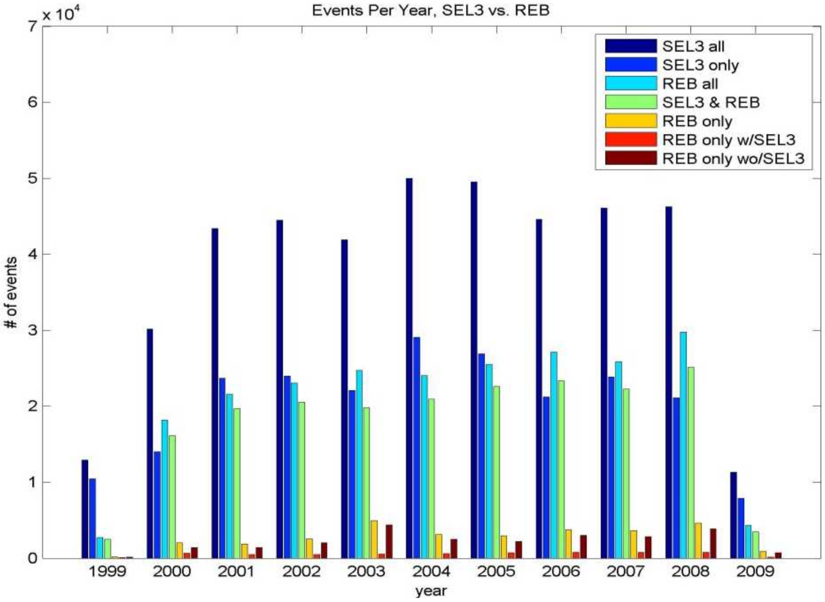


Figure 2. Categorization of automatic and analyst reviewed events for 1999-2009.

REB events occur almost entirely along the plate boundaries, as would be expected for earthquakes. SEL3 events that end up in the REB follow the plate boundaries, but not as reliably. Bogus SEL3 events can be very far from plate boundaries indicating false detections and/or false phase identifications.

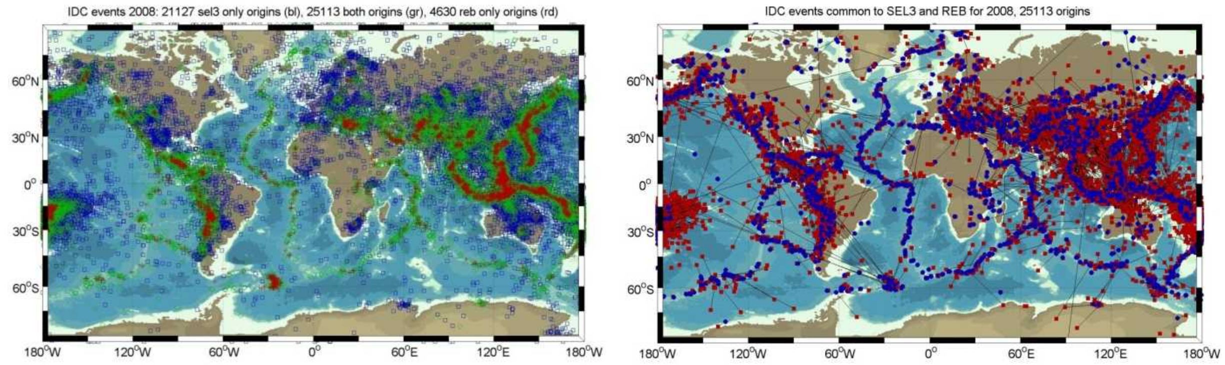

IDC events 2008: 21127 sel3 only origins (bl), 25113 both origins (gr), 4630 reb only origins (rd)
IDC events common to SEL3 and REB for 2008, 25113 origins
60°N
30°N
0°
30°S
60°S
180°W
120°W
60°W
0°
60°E
120°E
180°W

Figure 3. Comparison of automatic and analyst review events, 2008 (left). Relocation vectors for SEL3 events in REB, 2008 (right).

SEL3 events that are approved by analysts can have large relocation vectors suggesting problems with probability of detection calculations in the Global Associator (GA) software. Predicted azimuth and slowness should show large variations between SEL3 and REB locations, with the smaller residuals corresponding to the REB-adjusted locations. GA grid points nearby the REB-locations would have been considered during the association step, but further ones were preferred. This problem may indicate the need for better azimuth and slowness calibration.

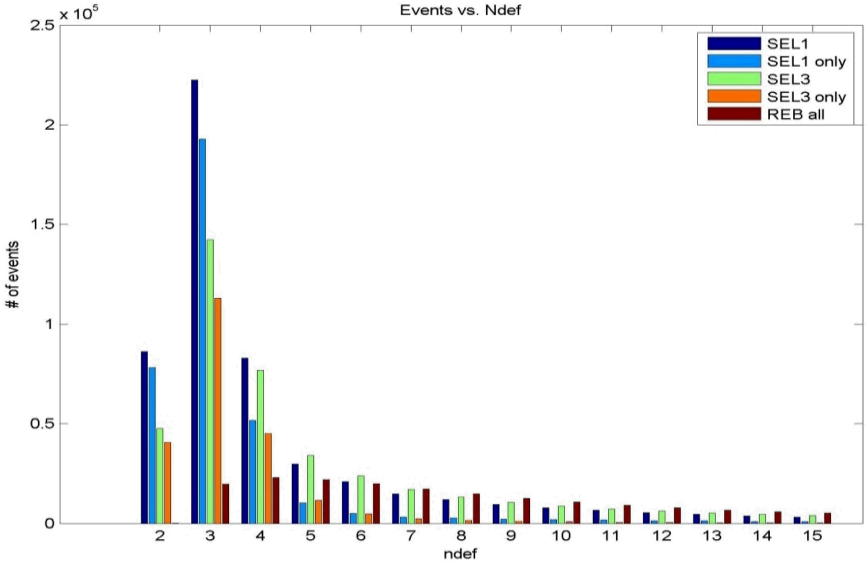


Figure 4. Number of events vs. number of location-defining phases, automatic vs. analyst reviewed, 1999-2009.

The event definition criteria for automatic events is a combined weight of 3.55 (seismic travel time has a weight of 1.0), implying that an event can be built with time-defining phases from as few as two stations if azimuth and slowness are used. By comparison, analyst approved events must have a weight count of 4.6 or more and must have defining phases from 3 or more primary seismic stations (for terrestrial events). SEL1, with no auxilliary or late-arriving seismic data is dominated by 2 and 3 station events. However, 91% of 2 station and 87% of the 3 station SEL1 events do not result in REB events, so the payoff for building these events is small. SEL3 reduces those percentages somewhat to 85% and 79%, presumably due to the additional data that becomes available, but clearly a large amount of analyst effort is being used to screen these marginal events, suggesting that the automatic system may be building too many of them.

Of the 44,993,580 IDCX arrivals, only 4,049,843 (9%) are associated with SEL3 events. The total number of REB associated arrivals is 3,651,404, but this includes new arrivals added by the analysts. Thus, the vast majority of automatic detections are never used to produce REB catalog events. This disproportion is intentional because the consequences of missing an important detection are far more costly than those of producing a false detection. However, stations with anomalously high proportions of automatic arrivals to REB arrivals may be good candidates for further tuning.

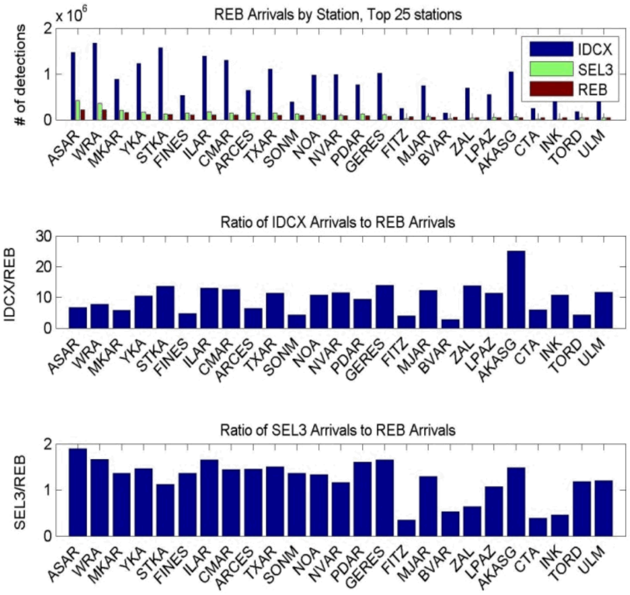


Figure 5. Comparison of all automatic arrivals with SEL3-associated and REB-associated arrivals, 1999-2009.

Automatic Detection System Event Classes and Proposed Evaluation

The automatic detection system produces an event bulletin that gets filtered several times before finally resulting in the final list, SEL3 (referred to as the set SEL3-ALL in the results). The events in SEL3-ALL undergo analyst review, which divides the events produced by the automatic bulletin into two categories. The first category comprises bogus events, known as False Automatic Origins. Here, because they were incorrectly predicted by the automatic system, they are labeled as False Positives. These events go on to get screened out by the analyst.

The remaining events that do not get screened out are known as Good Automatic Origins (Gauthier 2009), and are labeled as True Positives. Because these events fall in both SEL3 and endure through to the REB, they are labeled as SEL3&REB, i.e., those events lying at the intersection of the sets SEL3 and REB.

There also exists a set of events that *should* have been detected by the automatic system, but were not. These events are ones manually identified by an analyst and added to those surviving analyst review from the automatic system. These events, known as "Extra Analyst-Built Origins," are labeled as False Negatives (or missed detections), because the automatic system incorrectly failed to identify them as true events.

The events surviving analyst review are combined with the additional ones manually identified by an analyst; these are combined together to form the REB, and this set of events is referred to as REB-ALL. We note that both sets SEL3&REB and REB-ONLY can be further divided into subcategories (Gauthier 2009); this is not represented here but will be considered in future work.

There exists a fourth category of events, i.e., all of those possible events that could have been detected but were not, and if so, would have been incorrect. Hence, this "Universe" of events can be considered to be correct non-detections, labeled as True Negatives, but can never directly be measured.

Such scenarios where valid measures exist for true positives, true negatives, false positives, but not true negatives are common in information retrieval. Traditionally, such scenarios are evaluated in terms of *precision* and *recall*, and a *precision-recall* curve can be generated. Finally, the associated summary statistic, the F-Score, can be used to evaluate the overall effectiveness of the automatic detection system. In future work, the performance of the automatic system can be evaluated in these terms.

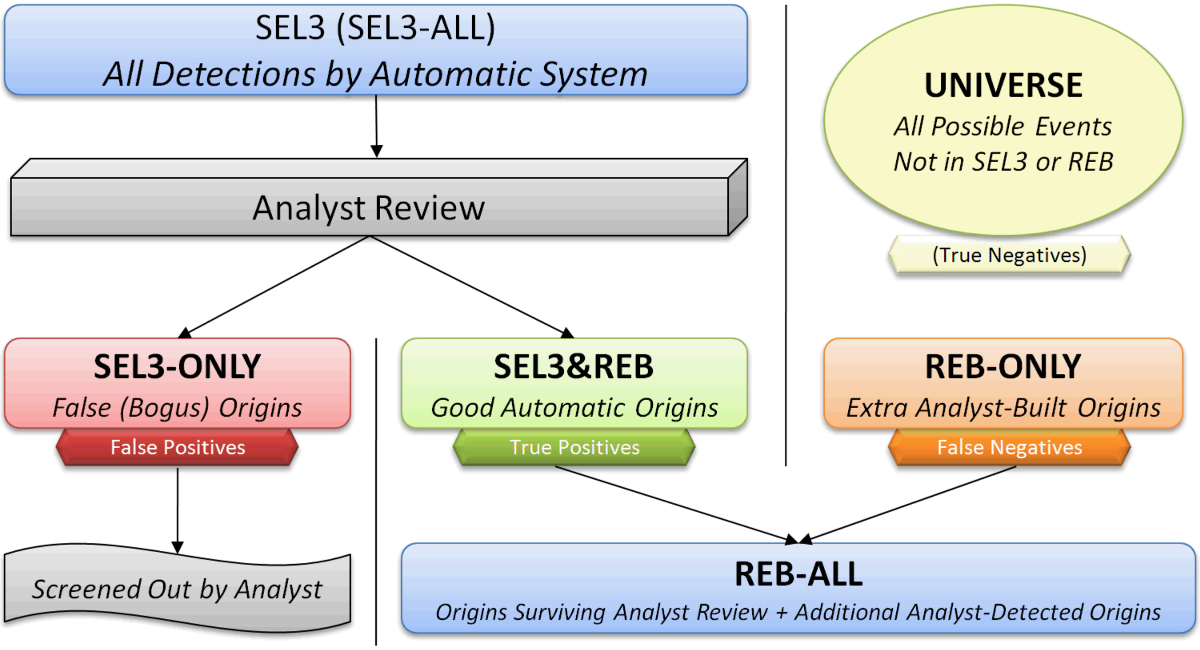


Figure 6. Overview of Automatic Detection System Outcomes

Application of Machine Learning: Screening False SEL3 Events

We use machine learning algorithms to process event features to try to reduce the number of false automatic events that IDC analysts must screen. ML can accommodate as many features as the researcher wants to evaluate and will find and properly weight the optimal combination to provide the best classification.

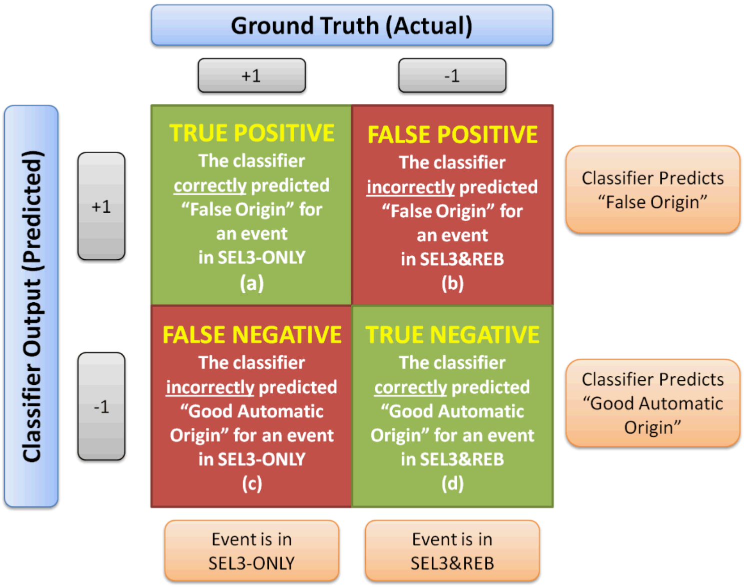
For this initial empirical study to determining the feasibility of the machine learning approach on reducing analyst burden in the automatic detection system, we concentrate specifically on the problem to screen out events (False Origins) that an analyst would otherwise have to screen out. Thus, the problem we address is the *automatic discrimination,* or separation, of events which would eventually be assigned to the set SEL3-ONLY from those in SEL3&REB (Figure 6). Thus, the problem as framed is a two-class or *binary* problem, where the classifier (or model) predicts only one of two possible class labels for a given event in SEL3-ALL.

In our approach, we train supervised machine learning models using various algorithms on half of the tagged origins from SEL3-ALL, i.e., origins whose class is known. The model is then *evaluated* over the remaining *test* data set, disjoint from the training data set used to train the models initially. The goal is for the classifier to correctly predict the true class (SEL3-ONLY or SEL3&REB) for as many test origins in SEL3-ALL as possible. Because the ground truth class for each of these test origins is known, we can evaluate performance of any particular classification method by comparing the classifier's predicted output with the known ground-truth labels.

Evaluation Approach: Confusion Matrix (Contingency Table)

We evaluate our ML models using the well-established confusion matrix approach, seeking to reduce the overall error rate (combination of false positive and false negative rates). Figure 7 showns the contingency table, known as a *confusion matrix* in the machine learning / pattern recognition literature, which characterizes all four possible outcomes in a two-class problem.

Essentially, the classifier can predict one of two classes, and the ground truth is one of two classes, resulting in four possible scenarios. These scenarios form the cells in the matrix in Figure 7 and are referred to as true positives, false positives, false negatives, and true negatives, accordingly. It is crucial to differentiate this evaluation approach from that in Figure 6, as the evaluation for a *detection* problem with unquantifiable true negatives is fundamentally different from that of a *two-class prediction* problem where correct class labels are fully known for a finite set of test data.



		Ground Truth (Actual)		
		+1	-1	
Classifier Output (Predicted)	+1	TRUE POSITIVE The classifier correctly predicted “False Origin” for an event in SEL3-ONLY (a)	FALSE POSITIVE The classifier incorrectly predicted “False Origin” for an event in SEL3&REB (b)	Classifier Predicts “False Origin”
	-1	FALSE NEGATIVE The classifier incorrectly predicted “Good Automatic Origin” for an event in SEL3-ONLY (c)	TRUE NEGATIVE The classifier correctly predicted “Good Automatic Origin” for an event in SEL3&REB (d)	Classifier Predicts “Good Automatic Origin”
		Event is in SEL3-ONLY	Event is in SEL3&REB	

Figure 7. Confusion Matrix, or Contingency Table, for comparing four possible outcomes in a binary problem

Features Used

Any Machine Learning approach begins with identifying potentially useful features. We use 18 features available from the IDC database tables for SEL3 events. These consist of 9 origin-based features (ndef, nass, depth, sdobs, smajax, sminax, sdepth, stime, numsta) and 9 arrival-based features, which we average by event (avg(snr), avg(amp), avg(rect), avg(deltim), avg(abs(timeres)), avg(delaz), avg(abs(azres)), avg(delslo), avg(abs(slores)).

Table 1. Preliminary features considered in the study.

Origin Features			**Arrival Features, Grouped by and Averaged by Origin**	
1	ndef: *number of locating phases*		**10**	avg(snr): signal-*to-nose ratio*
2	nass: *number of associated phases*		**11**	avg(amp): *amplitude, instrument corrected*
3	depth: *estimated depth*		**12**	avg(rect): *rectilinearity*
4	sdobs: *standard error of observations*		**13**	avg(deltim): *time uncertainty*
5	smajax: *semi-major axis of error*		**14**	avg(abs(timeres)): *time residual*
6	sminax: *semi-minor axis of error*		**15**	avg(delaz): *azimuth uncertainty*
7	sdepth: *depth error*		**16**	avg(abs(azres)): azimuth residual
8	stime: *origin time error*		**17**	avg(delslo): *slowness uncertainty*
9	numsta: *number of observing stations*		**18**	avg(abs(slores)): *slowness residual*

Features Used

In this initial study, we consider *decision tree* models of increasing sophistication. In our preliminary sensitivity analysis, decision-tree based machine learning classification algorithms performed best on the features listed in Table 1, exceeding that of other common methods such as the Support Vector Machine.

We compared the performance of a decision tree which *split* on a *single attribute*, then a similar tree splitting on two attributes, and finally, a sophisticated tree-based approach involving the fusion of multiple simple trees.

Single-Feature Models (Single Split)

For comparison, we first develop classifiers using single attributes. ndef and avg(SNR) are obvious choices. Applying ML to ndef alone yields the following model, based on a single split of an attribute: ndef <= 4 is a false automatic origin, ndef >= 5 is a good automatic origin. For avg(SNR) the threshold value is 5.5. Results for both of these models, as well as the more complex models, are shown in a summary table at right.

Two-Feature Model (Decision Tree)

Combining the two features, ML develops the more complex model (decision tree), learned by the Classification and Regression Tree (CART) algorithm, shown below:

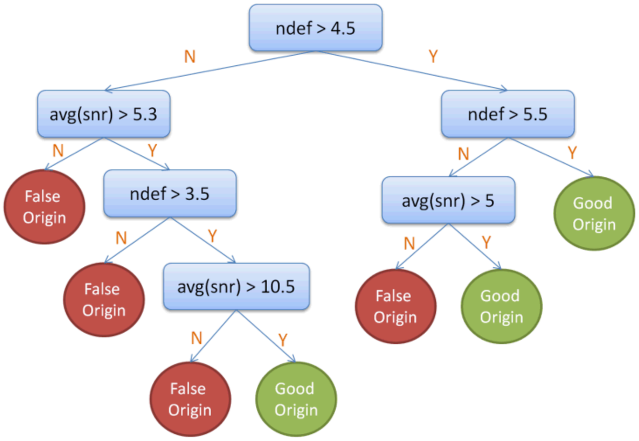


Figure 8. Example of an interpretable decision tree model providing non-linear two-class classification for two input attributes.

All-Features Model (Random Forests)
The Random Forests method is a state-of-the-art machine learning ensemble method due to Breiman and Cutler that combines bagging with random trees. This method is robust to noise, requires minimal parameterization and tuning, and performs well in the presence of possibly irrelevant features. Moreover, this method has a heavy theoretical basis. Here, we apply it to all 18 features simultaneously; the resulting model, comprising many simple decision trees, can be readily evaluated quantitatively but not easily visualized.

A summary of experimental results is given in Table 2.

Table 2. Experimental results summary.

Features Used	FNR	FPR	Accuracy
avg(SNR)	59.5%	9.0%	65.7%
ndef	11.6%	34.5%	76.9%
ndef & avg(SNR)	13.8%	29.8%	78.2%
All feautures	15.8%	18.2%	83.1%

CONCLUSIONS AND RECOMMENDATIONS

Applying Machine Learning to utilize a large number of readily available features to screen events that analysts will reject can improve the results, over a simplistic classifier, but the results are still not good enough to use because of the possibility of screening real events of interest. We believe that significantly better results can be achieved by use of new/enhanced features, and especially by use cost-sensitive learning to bias the classifier towards lowering the FPR at the expense of the FNR (reflecting actual monitoring system goals).

An important objective from this study was to inform future research directions in order to make the machine learning approach more effective. Future work will fall under three main areas: identification of additional features to help inform classification; improved classification algorithms; and the use of cost-sensitive learning to help improve results.

Future Work: Features
Useful features are the most important part of any machine learning classification task. Poor classification performance will occur if features that do not inform the discrimination task are not used. There are a number of ways to improve the features for this problem.

First, the power of existing features can be improved by applying special scalings and transforms. Second, new features can be computed as functions of existing features (e.g. distance between first and second detecting stations). Finally, entirely new features can be extracted from raw waveforms; this is an area of ongoing research (Meyer 2009).

Future Work: Machine Learning Algorithms
Future work will involve experimentation with various approaches and the latest algorithms in order to improve performance. An important consideration is that once the best general type of classifier is identified, improvements will probably be incremental among best classifiers, for the same feature set.

Presently, decision trees and related decision tree ensemble methods such as random forests appear to be giving the best performance. Other classes will also be investigated, including mathematical/function-based classifiers (SVM), probabilistic classifiers (Naïve Bayes), neural networks, and K-nearest-neighbor approaches. We are cautious to not draw strong conclusions in regards to best algorithms for this task, as this will be very heavily feature-dependent.

Future Work: Cost-Sensitive Learning

In this domain, like many domains, the penalty for one type of error (incorrectly predicting a “False Origin”) is much greater than the other type of error (incorrectly predicting a “Good Automatic Origin”). Machine Learning has *cost-sensitive* methods for handling unequal costs.

In particular, different error costs can be explicitly accounted for when training models in order to bias the model towards one type of error. Moreover, unequal error costs can also be considered during classifier evaluation. For example, if specific costs are known, a weighted average of the errors can be made, where one type of error receives more weight in the final score. We note that traditional classifier accuracy is a special case of this scenario, where error costs (weights) are equal.

A more elaborate approach will use ROC, or Receiver Operating Characteristic, analysis. In this scenario, all possible error costs are considered to generate a curve by varying a threshold, and the resulting Area Under the ROC Curve (AUC) summary statistic is a very robust evaluation metric. Importantly, it is becoming increasingly common in scenarios where one type of error is much more costly than the other to only consider ROC area under a certain *portion* of the curve. Such approaches will be important to use in this domain where the cost for having an analyst review one bogus event is dwarfed by the penalty of missing an event of interest.

ACKNOWLEDGEMENTS

TO BE DONE - Acknowledgements

REFERENCES

Gauthier, John H (2009). Preliminary Analysis of the International Data Centre Pipeline, Sandia National Laboratories Technical Report, SAND-!!!!.

Vapnik, Vladimir. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Breiman, Friedman, Olshen, and Stone (1984). *Classification and Regression Trees*. Chapman and Hall.

Meyer, F., Taylor, K.M., Kaslowsky, D., Procopio, M. and Young, C (2009). Evaluation of Empirical Mode Decomposition and Chirplet Transform for Regional Seismic Phase Detection and Identification [abstract]. Seismological Research Letters.

Breiman, L.(2001). *Random Forests*. Machine Learning 45(1):5-32.