SANDIA REPORT
SAND2009-4223
Unlimited Release
Printed July 2009

Preliminary Analysis of the International Data Centre Pipeline

John H. Gauthier

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia 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.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: 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 any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent 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, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

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

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-4223
Unlimited Release
Printed July 2009

Preliminary Analysis of the International Data Centre Pipeline

John H. Gauthier
System Readiness & Sustainment Technologies
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-1188

Abstract

The International Data Centre of the Comprehensive Nuclear-Test-Ban Treaty Organization relies on automatic data processing as the first step in identifying seismic events from seismic waveform data. However, more than half of the automatically identified seismic events are eliminated by IDC analysts. Here, an IDC dataset is analyzed to determine if the number of automatically generated false positives could be reduced. Data that could be used to distinguish false positives from analyst-accepted seismic events includes the number of stations, the number of phases, the signal-to-noise ratio, and the pick error. An empirical method is devised to determine whether an automatically identified seismic event is acceptable, and the method is found to identify a significant number of the false positives in IDC data. This work could help reduce seismic analyst workload and could help improve the calibration of seismic monitoring stations. This work could also be extended to address identification of seismic events missed by automatic processing.

ACKNOWLEDGMENTS

I would like to thank the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, Provisional Technical Secretariat, International Data Centre Division, in particular, Fil J. Filipkowski, IDC Services Officer. I would also like to thank Mark Harris, Chris Young, Mike Procopio, and David Gallegos for their support, thoughts, and guidance, which helped this work immeasurably. Sandia 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.

CONTENTS

1. Executive Summary 7
2. Introduction 9
3. Research accomplished 11
   3.1. Background 11
   3.2. Initial Assessment of the IDC Dataset 12
   3.3. Population Analysis of Bulk Data 15
   3.4. Empirical Method for Prediction of False Alarms 18
4. Conclusions and Recommendations 23
5. References 25
Distribution 27

FIGURES

Figure 1. IDC Pipeline. 12
Figure 2. Origin locations from SEL3 and REB. 14
Figure 3. IMS Primary (blue) and Auxiliary (red) stations referenced in SEL3 and REB. 15
Figure 4. Comparison snr CDFs for three populations. 17
Figure 5. CDFs (left) and PDFs plus probability of good origin (right). 19
Figure 6. CDFs (left) and PDFs plus probability of good origin (right). 19

TABLES

Table 1. Numbers of some tables in the IDC dataset. 13
Table 2. Test of the empirical method. 21

1. EXECUTIVE SUMMARY

An analysis is performed on seismic-event data processed in October 2002 by the International Data Centre (IDC) of the Preparatory Commission for 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 also be applicable to similar automatic processing systems.

Initial examination of the data shows that there are 264,676 arrivals during the interval from 2 October to 27 October (when all of the analyzed IDC data tables overlap). Automatic event processing uses 27,779 of these arrivals (11%) to generate 3758 origin hypotheses. Of these automatic origin hypotheses, 2052 (55%) do not survive analyst review, while 1706 (45%) are considered good. In addition, analysts build another 240 origins (12%) that the automatic processing missed. Thus, over the 26-day interval, a total of 1946 analyst-reviewed origins are recorded (an average of 3 per hour) using 22,350 arrivals (8% of the total number of arrivals).

To further investigate the data, the automatically generated origins and the analyst-reviewed origins are divided into five populations: (1) good automatic origins (automatically generated origins that ultimately survive analyst review); (2) isolated false origins; (3) confounded false origins (a false origin with at least one arrival later found in an analyst-built origin); (4) new analyst-built origins; (5) analyst-rebuilt origins (origins built using one or more arrivals from false origins). In this preliminary work, we concentrate on the differences between isolated false origins and good automatic origins, and the differences between analyst-built new origins and good automatic origins. With the assumption that the populations are independent, cumulative distribution functions (CDFs) are created for the mean, median, standard deviation, and sum of data entries used for each origin in these populations. The CDFs are then compared. In most cases the data are similar between populations. However, data that could be used to distinguish isolated false origins from good origins include the number of stations, the number of phases, the signal-to-noise ratio, and the pick error. Data that could be used to distinguish analyst-built new origins from good origins (i.e., data that could be used to distinguish events missed by automatic processing from events found by automatic processing) include the number of phases, the signal-to-noise ratio, the slowness, the pick error, and the azimuth error. Finding measures of travel time (pick error), slowness, and azimuth in this set of distinguishing data indicates that improved station calibration could have a significant effect on reducing the number of origins missed by automatic processing.

Based on the distinguishing data, an empirical method is devised to determine whether an automatically generated origin is good. The method is found to have a 100% success rate on the dataset used—the number of false alarms is reduced from approximately 50% to 0%. A similar empirical method could be developed to evaluate the incoming stream of arrivals and flag those that might come from actual events but were missed by the automatic

processing. These methods could be used to quantify how much improved station calibration could improve automatic processing. These methods might also eventually be used to reduce analyst workload by reducing the number of repeatedly made similar corrections that analysts must now perform.

2. INTRODUCTION

The Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) was established in 1996 by resolution of the United Nations. The CTBTO has 180 Member States, including the USA, which is a Signatory Member. The mission of the CTBTO is to implement the Comprehensive Nuclear Test Ban Treaty. To meet this mission, the CTBTO has established a global verification regime, based primarily on monitoring global seismicity, that includes 321 monitoring stations and 16 radionuclide laboratories throughout the world, preparation of on-site inspections in case of a suspected nuclear test, and provisional operation of the International Data Centre (IDC).

The IDC, established in Vienna in 1997, processes and analyzes data from the International Monitoring System (IMS) to provide Member States with basic information describing possible nuclear weapons tests. The IDC receives several gigabytes of seismic waveform data every day. These data are first automatically processed and then reviewed by specialized analysts. Analysts typically reduce the number of automatically identified events approximately by half. In 2007, IDC analysts identified and evaluated an average of 126 seismic events per day (CTBTO 2008).

The purpose of the effort reported here is to better understand the automatic processing used by the IDC, in particular what seismic data most influence acceptable and unacceptable automatic processing. The IDC is being studied because it is an important international endeavor that is generally similar to other seismic monitoring systems. With the information developed here, an empirical method is derived that could substantially reduce the number of false acceptances (false alarms) generated by the automatic processing. Such a reduction in false alarms could reduce analyst workload and improve the efficiency and effectiveness of the IDC. Further, a similar method could be developed to reduce the number of false rejections (missed seismic events) that are currently overlooked by the automatic processing.

3. RESEARCH ACCOMPLISHED

Three analyses are described here. The first is an initial assessment of the behavior of the IDC automatic seismic-event processing. The second is an analysis of five populations within the IDC data, defined according to error type, to determine what differences in the data account for these populations—i.e., effectively perform a sensitivity study to discover the data that discriminate between populations. The third involves construction and testing of an empirical method to determine, a priori, whether an origin produced by the automatic processing belongs to the good-automatic-origin population or the falsely accepted population.

3.1. Background

The International Data Centre (IDC) is a worldwide seismic monitoring system, created to process and analyze seismic-event data reported by the International Monitoring System (IMS). Of interest is quantifying the efficiency and effectiveness of the automatic processing of seismic data used by the IDC. Also of interest are possible ways of improving the automatic processing—in particular, reducing the number of false acceptances and reducing the number of false rejections. A false acceptance (type 1 error) is an origin hypothesis produced by automatic event processing that is eventually rejected by a human analyst. A false rejection (type 2 error) is an origin that was missed by automatic event processing but subsequently identified by a human analyst. To study these issues, Sandia National Laboratories asked the IDC for a one-month dataset. In late 2002, SNL received IDC data tables representing event processing performed in the month of October 2002.

The IDC dataset consists of 14 different database table types for 4 different steps in the IDC pipeline. Included are both static tables (site, network, sitechan, affiliation, gregion, and sregion) and dynamic tables (arrival, assoc, event, netmag, origin, origerr, stamag, and stassoc). SNL also received the IDCX tables of incoming arrivals (idcx_arrival). The different pipeline steps included in the dataset are SEL1, SEL2, SEL3, and REB (CMRWEB 2001). Standard Event List 1 (SEL1) is a list of events and event parameters (origin and associated arrival information) obtained from automated processing of data from the IMS Primary seismic stations. SEL2 and SEL3 are lists of events and event parameters obtained from automated processing of data from the IMS Primary and Auxiliary seismic stations. SEL1 is generated about one hour after real time, SEL2 about 6 hours after, SEL3 about 12 hours after. SEL3 is the used by human analysts to produce the Reviewed Event Bulletin (REB). The REB is a list of events and event parameters obtained by analyst review of the automated processing results and, in some cases, by analysts manually building events from data from the IMS seismic (Primary and Auxiliary), hydroacoustic, and infrasonic stations. The REB is issued between two and five days after the data were originally acquired. The IDC pipeline is shown in Figure 1.

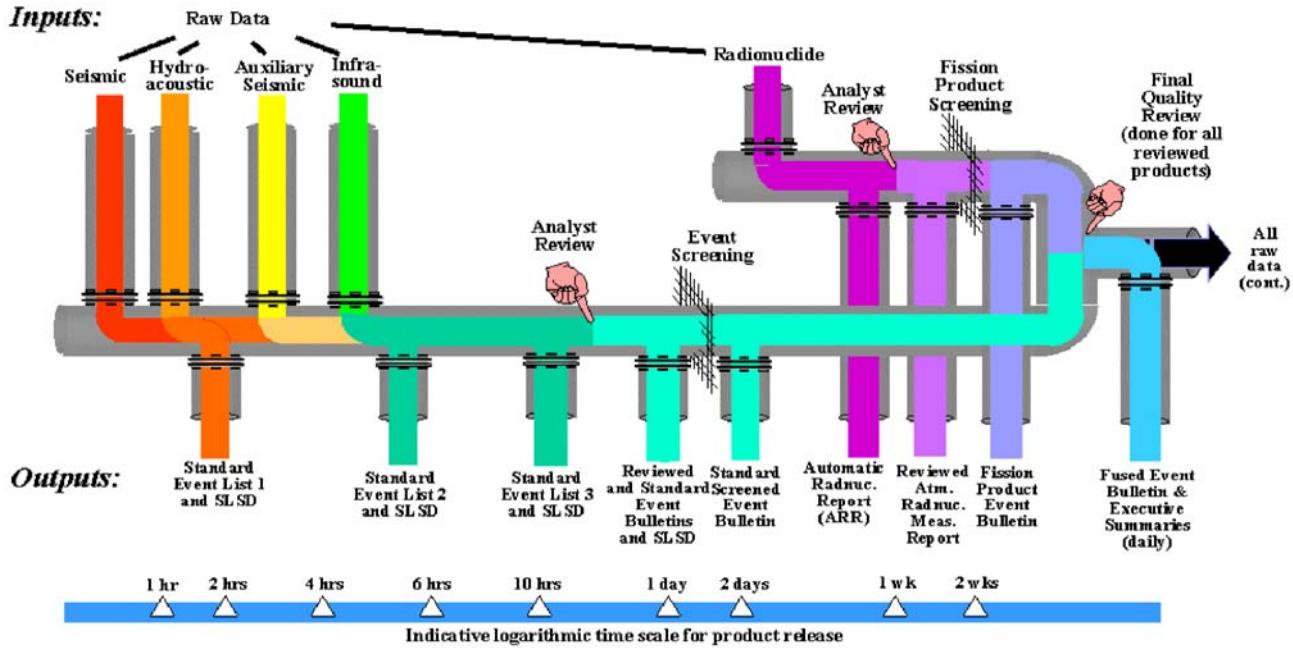

Figure 1. IDC Pipeline.
Shown with the IMS data flowing into the IDC at the top of the figure and the IDC products flowing out of the IDC at the bottom (taken from CTBTO 2000).

Events in SEL1, SEL2, SEL3, and REB must meet specific event-definition criteria. The criteria are largely those set during the GSETT-3 experiment. The main difference between the automatic lists and the REB is that events in the REB must also be located using first-arriving (P-type) phases at three IMS Primary stations. In addition to the event-definition criteria, events in the automatic lists must also pass a screening test that is based upon the results of an error-ellipse test, a probability-of-detection test, and an arrival-quality test. The threshold for the last of these tests is set slightly lower for the SEL1 than for the subsequent SEL2 and SEL3 because the later lists include Auxiliary as well as Primary stations. Thus small events that are given in SEL1 may not appear in SEL2 or SEL3 because of these higher thresholds (CMRWEB 2001).

3.2. Initial Assessment of the IDC Dataset

The IDC data tables overlap between epoch times 1033516800 and 1035759600, corresponding to the time interval between 2 October and 27 October 2002. Table 1 presents the sizes of some of the IDC data tables. In this time interval, 264,676 arrivals occur. Based on comparison of the sizes of idcx_arrival and reb_arrival, only about 8% of the arrivals are ever used. IDCX arrivals include the following phases: Lg (4%), N (27%), P (22%), Pg (2%), Pn (10%), Px (2%), Rg (1%), Sn (<1%), Sx (22%), tx (7%), T (2%), and Tx (<1%). Arrivals referenced by REB add 17 new P phases, 5 new S phases, and an LR phase. By far the most numerous phase in REB is P (64%), followed by LR (8%), tx (7%), Pn (6%), Sx (3%), N (3%), Pg (2%), T (2%), and the rest

(<2%). S-waves are under-represented in the REB, probably because the event-selection criterion of three P-phases puts a premium on detecting and using good P arrivals.

Table 1. Numbers of some tables in the IDC dataset.
(Tables are named idcx_arrival, sel1_assoc, etc.)

	arrival tables	assoc tables	origin tables
idcx	264,676	—	—
sel1	—	27630	4396
sel2	—	27600	3739
sel3	—	27779	3758
reb	22350	22341	1946

Of most interest here are the automatic processing steps, especially SEL3, which describes the culmination of the automatic processing, and the analyst-review step, REB. SEL1 associates 27,630 arrivals into 4396 origin hypotheses. By SEL3, 27,779 arrivals are associated in 3758 origin hypotheses. Upon review by analysts, 1706 of the SEL3 origin hypotheses (45%) are accepted and 2052 (55%) are rejected. The accepted origin hypotheses are here called good automatic origins; the rejected origin hypotheses are called false origins. Analysts build another 240 origins that were missed by the automatic processing. These origins are here called analyst-built origins. Thus, of the 1946 origins listed in the REB, 1706 (88%) are good automatic origins and 240 (12%) are analyst-built origins.

In terms of average behavior, a new arrival is added to IDCX about every 8 seconds. A SEL3 origin hypothesis is generated about every 10 minutes. A REB origin is generated about every 19 minutes, i.e., events average 3 per hour or 75 per day. Note that a P-wave can take 30 minutes to traverse the interior of the Earth, and therefore many signals from different events can be expected to cross.

Figure 2 presents a map showing the locations of origins in SEL3 and REB. The origins are color-coded such that green circles indicate good automatic origins, blue crosses indicate false origins, and red diamonds indicate analyst-built origins. The origins tend to follow the plate boundaries of the Earth as expected. The analyst-built origins occur mostly in areas where there are many good automatic origins. The false origins are more randomly located.

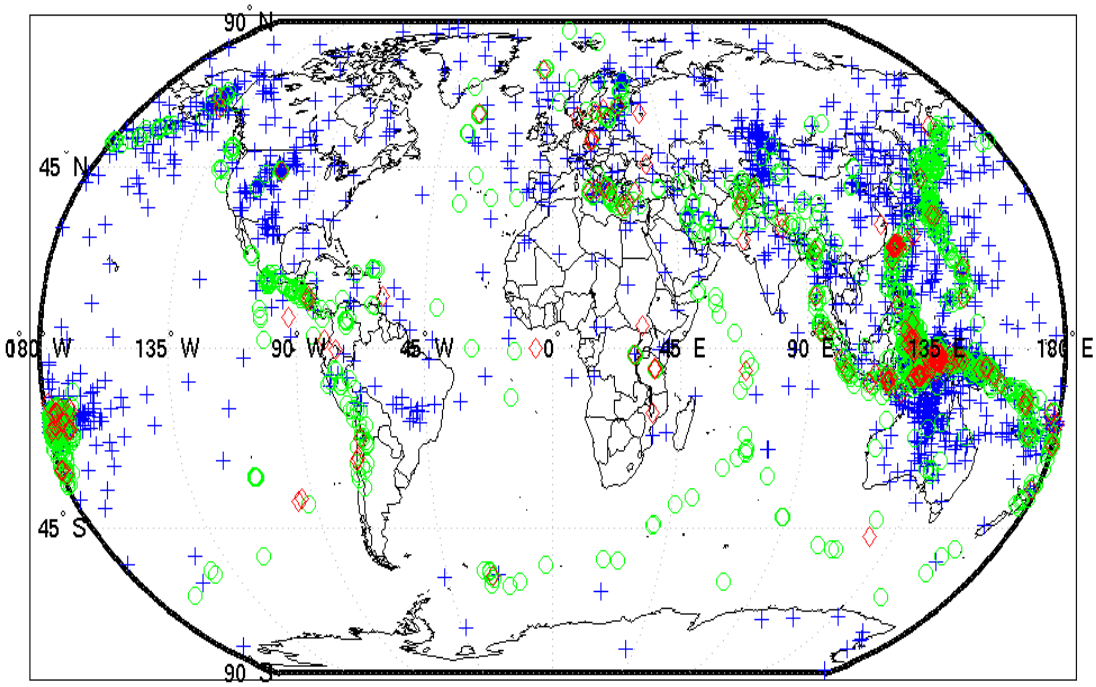

Figure 2. Origin locations from SEL3 and REB.
Green indicates good automatic origins, blue indicates false origins, and red indicates analyst-built origins. The colors overlap with red on top, then green, then blue.

Figure 3 shows the locations of the 48 stations referenced in SEL3 and REB. 27 stations belong to the IMS Primary network; the additional 21 stations belong to the Auxiliary network. (The 11 hydroacoustic stations in the IMS Primary network are not shown.) The Primary stations are spread relatively uniformly across the globe. The Auxiliary stations are mostly in North America, Northern Europe, and Japan. Clusters of false origins shown in Figure 2 occur in North America, Northern Europe, Central Asia, and Eastern Asia, near the location of major seismic arrays shown in Figure 3. The stations, even those in the IMS Primary network, do not see equal usage. IDCX arrivals are divided between both arrays and 3-component stations; however, arrays tend to be referenced more often in the REB. Sixteen stations account for more than half of the IDCX arrivals: ASAR, CMAR, FINES, ILAR, INK, JNU, JOW, KMBO, MJAR, MKAR, NOA, NVAR, PDAR, STKA, TXAR, and WRA. Ten stations account for more than half of the arrivals referenced in REB: ASAR, CMAR, FINES, ILAR, MKAR, NOA, NVAR, PDAR, STKA, and WRA.

This initial assessment of the IDC dataset shows that the automatic processing is relatively inefficient (over half of the automatically generated origins are subsequently rejected by analysts as false) but for the most part effective (automatic processing misses 12% of good origins). Over 90% of the IMS arrivals reported to the IDC are unused. P-phase arrivals are most often used to produce good automatic origins.

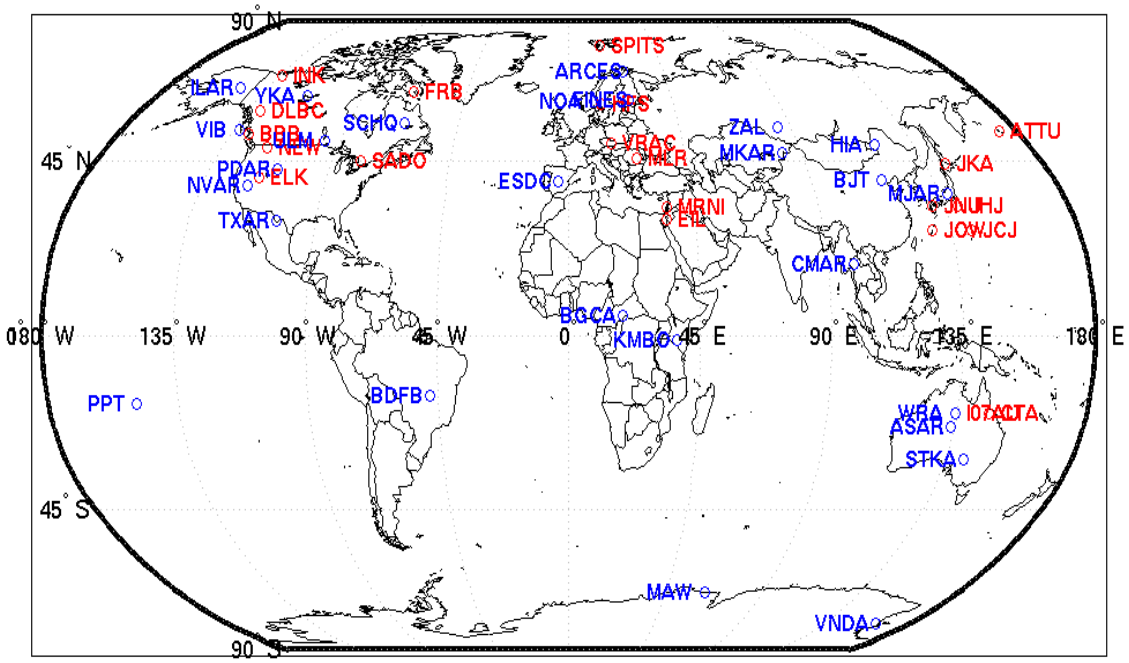

Figure 3. IMS Primary (blue) and Auxiliary (red) stations referenced in SEL3 and REB.

3.3. Population Analysis of Bulk Data

Here origins in the IDC dataset are separated into different populations according to whether the origins are correct or erroneous. The data referenced by these populations are examined, and it is found that certain data can be used to distinguish between populations. This information is valuable because it directly indicates what improvements to the IDC should reduce errors, and thus improve performance.

Above, the IDC dataset is divided into three, ad hoc populations: good automatic origins, false origins, and analyst-built origins. Here populations are defined more rigorously using their database relationships. The false-origin and analyst-built-origin populations are further subdivided according to whether they interfere with each other or not.

Good-automatic origins. These 1706 origins are built by the automatic processing and survive analyst review. The population is defined as origins in sel3_origin tables that point to the same evid (event identifier) as origins in reb_origin tables.

False origins. These 2052 origins in SEL3 are incorrectly reported by the automatic processing. They are defined as origins in sel3_origin tables that point to an evid that is not pointed to by any origins in the reb_origin tables. False origins can be divided into two subsets: *isolated false origins* are 1768 false origins referencing none of the arrivals that are referenced by reb_origin tables, and *confounded false origins* are 284 false origins referencing arrivals that are subsequently referenced by analyst-built origins.

Analyst-built origins. These 240 origins in REB are missed by the automatic processing and built by analysts. They are defined as origins from reb_origin tables that reference an evid that is not referenced by any origins in the sel3_origin tables. Analyst-built origins can be divided into two subsets: *new analyst-built origins* are 169 origins referencing none of the arrivals that are referenced by false origins, and *analyst-rebuilt origins* are 71 origins referencing arrivals that have been referenced by false origins. Note that analyst-rebuilt origins and confounded false origins are related. Examination of the time sequence of origins suggests that confounded false origins are created when two actual origins occur close together in time; the automatic processing occasionally mixes the signals, misses one of them, and creates a false origin instead.

The populations can be compared by assuming that the origin data are independent and calculating cumulative distribution functions (CDFs) for each datum type and each population type. For example, CDFs can be calculated for the signal-to-noise ratio (snr) for each population. To calculate the CDFs for snr, a list of orids (unique origin identifiers) is selected from one of the populations, and then snr data are selected for each orid. The mean, median, standard deviation, and sum of the snr data are calculated for each orid. Now consider only the set of mean-snr-per-orid data (the other sets are treated similarly). The mean-snr set is sorted and the cumulative probability for each value is calculated as i/N, where i is the position in the sorted list and N is the length of the list. The CDFs for the different populations are then plotted together and compared. If the probabilities are substantially different between any of the populations, the data can *distinguish* between the populations. This process is repeated for the sets of medians, standard deviations, and sums. Then it is repeated for all the data. In this manner, depth, nass, ndef, grn, srn, depdp, mb, ms, ml, and time from the origin tables are compared for the different populations; belief, delta, seaz, esaz, timers, azres, slores, emares, wgt, and the number of stations from the assoc tables are compared; deltim, azimuth, delaz, slow, delslo, ema, rest, amp, per, logat, and snr from the arrival tables are compared.

Most data are found to be similar between populations. For example, isolated false origins tend to have slightly smaller body magnitude (mb) than the other populations, but the distributions overlap so much that one cannot say with any certainty that any given magnitude belongs to a particular population. Some data, however, are found to be distinguishing. Based on this analysis, data that might be used to distinguish between isolated false origins and good automatic origins are the number of observing stations, the number of associating phases (nass), signal-to-noise-ratio (snr), and pick error (deltim). (The number of observing stations and snr can also be used to distinguish between all false origins and good automatic origins, a fact that is used in the final analysis below.) Data that might be used to distinguish between new analyst-built origins and good automatic origins are the number of phases (nass), signal-to-noise-ratio (snr), pick error (deltim), slowness (slow), and azimuth error (delaz).

As an example of distinguishing data, Figure 4 presents plots of the CDFs for the mean and median snr per orid for new analyst-built origins, isolated false origins, and good automatic origins. These plots demonstrate how the populations differ in terms of their signal-to-noise ratio. They also demonstrate how different methods of examining the signal-to-noise ratio for each orid—i.e., mean and median—give different results. Mean snr per orid can be used to distinguish between the isolated-false-origin population and the good-automatic-origin population, while median snr per orid distinguishes between the new-analyst-built-origin

population and the good-automatic-origin population. Another interesting feature of these plots is that new analyst-built origins typically have as good or better signal-to-noise ratios than good-automatic origins; thus, automatic processing is missing them for some other reason than signal-to-noise ratio.

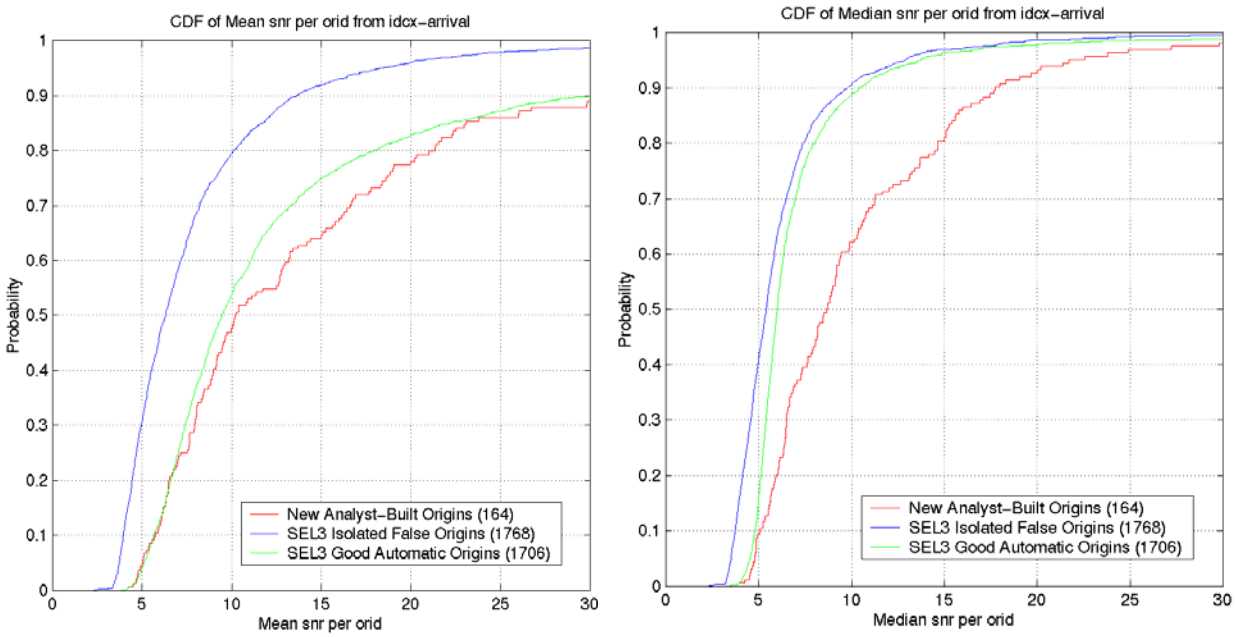

Figure 4. Comparison snr CDFs for three populations.
Mean signal-to-noise ratio (snr) per orid is shown at left; median snr per orid is shown at right.

This population analysis indicates that certain data can characterize origins in the IDC dataset according to whether they are good automatic origins, false origins (isolated or confounded), or analyst-built origins (new or rebuilt). The data that distinguish a population tell us something important about that population. Number of stations, nass, snr, and deltim can distinguish isolated false origins. Increasing the number of stations or associating phases, increasing the signal-to-noise ratio, or decreasing the pick error should reduce the number of isolated false origins. Indeed, increasing the number of stations from SEL1 to SEL3 by inclusion of the IMS Auxiliary network does reduce the number of isolated false origins by over 300 and the accompanying increase in the number of phases reduces the number of analyst-built origins by over 340. Improving station calibration, per se, might not necessarily reduce the number of isolated false origins, unless it acts to increase the number of stations that see a given event. nass, snr, deltim, slow, and delaz can distinguish new analyst-built origins. Several of these data would be improved by station calibration, and thus improved station calibration should increase the number of good automatic origins and reduce the number of new analyst-built origins.

3.4. Empirical Method for Prediction of False Alarms

The fact that certain data can distinguish different populations within the IDC dataset suggests the possibility of quantifying the correctness of an origin produced by automatic processing. Here an empirical method is developed and shown to be able to predict whether an automatically processed origin is good or a false alarm.

The number of observing stations and the mean signal-to-noise ratio are noted above to be able to distinguish between the good-automatic-origin population and the false-origin population (the union of the isolated and confounded false-origin populations). An algorithm to determine whether a given origin produced by SEL3 belongs to the good-automatic-origin population is as follows.

1. Calculate CDFs for number of stations and signal-to-noise ratio for the isolated-false-origin, confounded-false-origin, and good-automatic-origin populations (as above).
2. Calculate probability density functions (PDFs) for these data and populations from their CDFs. For the number of stations, the slope of the CDF is used as the PDF. For signal-to-noise ratio, an error function is hand-fitted to a CDF, then the fitted curve is numerically differentiated. Note that it should be possible to produce an automatic method for creating these PDFs. Each PDF is confirmed by checking that its integral is approximately equal to 1.
3. The PDFs are scaled by their population sizes to make frequency distributions.
4. Then for each value, x, of the number of stations or the signal-to-noise ratio, the probability that it belongs to the good-automatic-origin population (P_{sta} or P_{snr}) is the ratio of the frequency of a good automatic origin (F_{sta} or F_{snr}) at x, to the sum of the frequencies of all three populations (F_{all_sta} or F_{all_snr}) at x: $P_{sta}(x) = F_{sta}(x)/F_{all_sta}(x)$ and $P_{snr}(x) = F_{snr}(x)/F_{all_snr}(x)$. P_{sta} and P_{snr} are conditional on the probability that x is selected.
5. Once P_{sta} and P_{snr} are calculated, combine the probabilities as follows: $P_{good} = 1 - (1 - P_{sta}) \times (1 - P_{snr})$, where P_{good} is the probability that the event will survive analyst review. $P_{good} = 0.75$ is here used as the dividing line between good events and false alarms.

Figure 5 shows plots of the CDFs, PDFs, and P_{sta} curves for the number of stations; Figure 6 shows these curves for signal-to-noise ratio. (In Figure 6a, the number of points for SEL3 isolated false origins and SEL3 good automatic origins have been reduced by a factor of 20 for better visibility.)

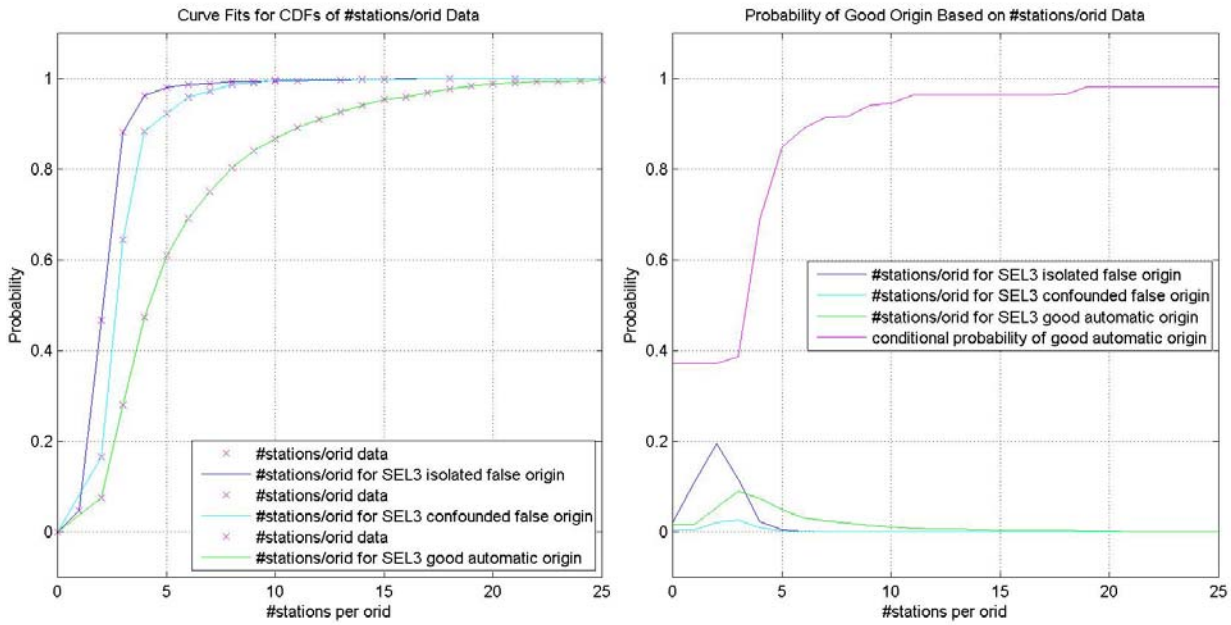

Figure 5. CDFs (left) and PDFs plus probability of good origin (right).
Plots are shown as a function of the number of observing stations.

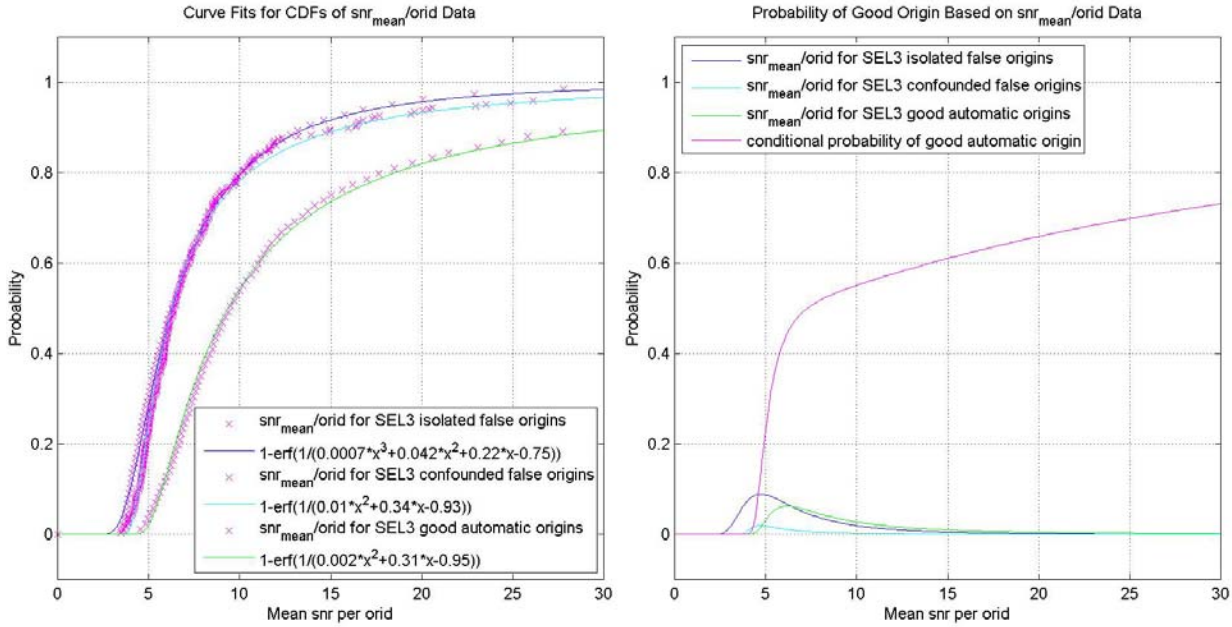

Figure 6. CDFs (left) and PDFs plus probability of good origin (right).
Plots are shown as a function of the mean signal-to-noise ratio (mean snr).

This empirical method is tested against 33 origins from SEL3 that escaped being included in the initial definition of populations. (They occurred before epoch time 1033516800.) The results of the test are presented in Table 2. For the dataset being used, the empirical method correctly predicts good origins and false alarms every time. Thus, the origins produced by the IDC automatic processing went from approximately 50% false alarms to 0% false alarms. Given that the empirical method was not tuned (e.g., the probability based on signal-to-noise ratio could have been weighted differently than the probability based on number of stations), this 100% success rate is surprising and encouraging.

Although this empirical method shows promise, it would be premature to use it immediately to improve IDC performance. The method should be redeveloped and tested with a larger dataset. In this regard, calculation of PDFs should be automated. Also, as presented here, the method does not learn and thus cannot adjust to changes in the IDC or the IMS network; these issues should also be addressed. The method does, however, demonstrate that certain characteristics of the IMS data could be exploited to improve both the efficiency and effectiveness of the IDC.

Table 2. Test of the empirical method.

Orid	**Actual Event** No=0;Yes=1	**#Stations** (P_{sta})	**Signal-to-Noise Ratio** (P_{snr})	P_{good}	**Prediction** P_{good} > 0.75
1321700	0	2 (0.37)	3.2 (0.00)	0.37	0
1321729	1	9 (0.94)	8.0 (0.52)	0.97	1
1321580	1	6 (0.89)	7.4 (0.50)	0.95	1
1321767	1	4 (0.69)	7.4 (0.50)	0.85	1
1321612	1	2 (0.37)	31.3 (0.74)	0.84	1
1321632	1	2 (0.37)	37.8 (0.77)	0.86	1
1321757	1	8 (0.92)	5.5 (0.37)	0.95	1
1321665	0	2 (0.37)	5.2 (0.29)	0.55	0
1321835	0	3 (0.39)	5.9 (0.42)	0.65	0
1321694	1	7 (0.91)	21.1 (0.67)	0.97	1
1321849	0	2 (0.37)	2.1 (0.00)	0.37	0
1321876	1	3 (0.39)	23.6 (0.69)	0.81	1
1321707	0	2 (0.37)	6.6 (0.47)	0.67	0
1321850	0	2 (0.37)	2.3 (0.00)	0.37	0
1321880	1	13 (0.96)	11.3 (0.57)	0.98	1
1322210	1	8 (0.92)	8.0 (0.52)	0.96	1
1321915	1	4 (0.69)	6.9 (0.49)	0.84	1
1321918	0	3 (0.39)	5.8 (0.40)	0.63	0
1321745	0	1 (0.37)	13.1 (0.59)	0.74	0
1321938	1	8 (0.92)	5.6 (0.38)	0.95	1
1321801	1	11 (0.96)	42.2 (0.79)	0.99	1
1321807	1	8 (0.92)	6.7 (0.48)	0.96	1
1321818	0	3 (0.39)	10.0 (0.55)	0.73	0
1321821	0	2 (0.37)	8.4 (0.53)	0.70	0
1321843	1	2 (0.37)	13.9 (0.60)	0.75	1
1321993	0	3 (0.39)	10.1 (0.55)	0.73	0
1322037	1	8 (0.92)	12.1 (0.58)	0.97	1
1321895	0	2 (0.37)	4.7 (0.14)	0.46	0
1321912	0	2 (0.37)	3.7 (0.00)	0.37	0
1321911	0	3 (0.39)	4.0 (0.00)	0.39	0
1321914	1	5 (0.85)	18.9 (0.65)	0.95	1
1321946	0	1 (0.37)	10.0 (0.55)	0.72	0
1321966	0	2 (0.37)	9.9 (0.55)	0.72	0

4. CONCLUSIONS AND RECOMMENDATIONS

Several caveats apply to this effort. The data are limited to 26 days and might not be representative. The data are dependent on the configuration of the IDC and the IMS network and might not apply to other systems. Not all issues associated with automatic processing are considered. For instance, the IMS network is heterogeneous and individual station characteristics are not considered. Phase misidentification could also play an important role in erroneous behavior and it is not considered.

With acknowledgement of the caveats, the preliminary analysis of the IDC dataset presented here supports the following conclusions: (1) IDC automatic event processing is relatively inefficient in that approximately 55% of its results are false origins. (Although many of these origins might be categorized as false only because they do not meet the selection criteria required for the REB.) IDC automatic event processing is for the most part effective, in that it catches 88% of actual origins. (2) Performance of the IDC automatic event processing depends most on the number of stations observing an event and the signal-to-noise ratio of arrivals from an event. (3) Improved station calibration should help reduce the number of origins missed by automatic processing and could help reduce false origins. (4) Empirical methods that exploit data characteristics show promise in increasing the efficiency and effectiveness of the IDC. In particular, an empirical method is presented that shows promise in reducing the number of false alarms generated by automatic event processing. Potential value added by using this method includes the following: analysts can have another way to prioritize events for examination; the system can be tuned to greater sensitivity without overwhelming the analysts; and if the system is kept at the same sensitivity, analyst intervention could be reduced, and thus costs could be reduced.

Another opportunity suggested here is that this empirical method could be extended to find events missed by automatic processing. Analysts at the IDC are required to build more than 10% of all events—meaning that automatic processing missed those events. Many events are missed because they occur almost at the same time as other events. As timing a weapon test with an earthquake is one possible way of hiding it, these events are important to identify. As discussed above, the number of phases, snr, deltim, slow, and delaz can distinguish between the good-automatic-origin population and the analyst-built-origin population. Using a similar empirical method with these new data, unused arrivals could be scanned to see if a collection of them within a reasonable time interval might show indications of belonging to the analyst-built-origin population. These arrivals could be flagged and a probability of correctness could be presented to the analyst. Further, more than 90% of arrivals are never associated with an event. Such a method might be able to find events that were even missed by analysts.

5. REFERENCES

1. CTBTO (2008), http://www.ctbto.org.

2. CTBTO (2000), IDC Users' Guide, Version 2.1, Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, Provisional Technical Secretariat, International Data Centre Division, Monitoring Section, Fusion, Review and Services Unit

3. CMRWEB (2003), IDC Documents, http://cmrweb.cmr.gov/librarybox/idcdocs/idcdocs.html.

4. CRMWEB (2001), Standard event list 1, http://cmrweb.cmr.gov/pidc/helpbox/sel1.html, Standard event list 2, http: http://cmrweb.cmr.gov/pidc/helpbox/sel2.html, Standard event list 3, http://cmrweb.cmr.gov/pidc/helpbox/sel3.html, Reviewed event bulletin, http://cmrweb.cmr.gov/pidc/helpbox/reb.html, Specific criteria, http://cmrweb.cmr.gov/pidc/helpbox/spec_crit.html.

DISTRIBUTION

1 Leslie A. Casey
NNSA Office of Nuclear Detonation
Detection/NA-222
1000 Independence Avenue SW
Washington, DC 20585

1 John J. Dwyer
Air Force Technical Applications
Center/TTR
1030 S. Hwy A1A
Patrick AFB, 32925

1 Dave Schmitt
Air Force Technical Applications
Center/TTR
1030 S. Hwy A1A
Patrick AFB, 32925

1 Jorge I. Roman-Nieves
Air Force Technical Applications
Center/TTR/MTC Technologies
Command Technologies Inc.
1030 S. Hwy A1A
Patrick AFB, FL 32925-3002

1 Jeff Miller
Air Force Technical Applications
Center/TTR/MTC Technologies
1030 S. Hwy A1A
Patrick AFB, FL 32925-3002

1 Gregory S. Wagner
Air Force Technical Applications
Center/TTR
1030 S. Hwy A1A
Patrick AFB, FL 32925-3002

1 Dr. Mark T. Woods
Air Force Technical Applications
Center/TTR
1030 S. Hwy A1A
Patrick AFB, FL 32925-3002

1 William Junek
Air Force Technical Applications
Center/TTR
1030 S. Hwy A1A
Patrick AFB, FL 32925-3002

1	MS0401	Dorthe B. Carr	05527
1	MS0401	David P. Gallegos	05527
1	MS0401	Jennifer E. Lewis	05527
1	MS0401	Michael J. Procopio	05527
1	MS0401	Christopher J. Young	05527
1	MS0401	Timothy McDonald	05736
1	MS0627	J. Mark Harris	05627
5	MS1188	John H. Gauthier	06342
1	MS1188	Gerard E. Sleefe	06340
1	MS1188	Bruce M. Thompson	06342

1	MS0899	Technical Library	9536 (electronic copy)

Sandia National Laboratories