SANDIA REPORT
SAND2007-6439
Unlimited Release
Printed October 2007

The GNEMRE Dendro Tool

B. John Merchant

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.

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

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

SAND2007-6439
Unlimited Release
Printed October 2007

The GNEMRE Dendro Tool

B. John Merchant
Next Generation Monitoring Systems
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS0404

Abstract

The GNEMRE Dendro Tool provides a previously unrealized analysis capability in the field of nuclear explosion monitoring. Dendro Tool allows analysts to quickly and easily determine the similarity between seismic events using the waveform time-series for each of the events to compute cross-correlation values. Events can then be categorized into clusters of similar events.

This analysis technique can be used to characterize historical archives of seismic events in order to determine many of the unique sources that are present. In addition, the source of any new events can be quickly identified simply by comparing the new event to the historical set.

ACKNOWLEDGMENTS

We thank all of the Dendro Tool users who have helped us to debug and improve the software, particularly our colleagues at LANL and LLNL.

CONTENTS

FIGURES

TABLES

NOMENCLATURE

DOE	Department of Energy
GNEMRE	Ground-based Nuclear Explosion Monitoring Research & Engineering (Program within NA22)
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
NA22	Office of Non-proliferation Research & Development (Office within NNSA)
NNSA	National Nuclear Security Administration (Office within DOE)
PNNL	Pacific Northwest National Laboratory
SNL	Sandia National Laboratories

1. INTRODUCTION

The fundamental problems faced in nuclear explosion monitoring are those of detecting, locating, and identifying the sources of various seismic events. Dendro Tool is specifically targeted at helping to identify seismic events.

There are many known sources of seismic events: earthquakes, mining activities, nuclear explosions, etc. These sources generate vibrations that travel through the earth that can be detected by seismic monitoring stations. The waveform observed at a seismic station from a particular source, such as a mine, is unique to the location and the type of the source. So, two similar waveforms observed at a station have a high confidence of originating from the same location and the same event type. Over time, a single repeating seismic source – e.g. a mine -- may be responsible for generating many highly similar events.

The traditional method of identifying the source type of an event coming from an area of repeating seismicity is to visually compare the waveform data to historical waveforms from other nearby events. If the waveforms appear qualitatively similar, then an analyst may infer that the events have a common source mechanism. This method of identification can be very effective, but it is very time consuming and prone to human error.

In the late 1980's and early 1990's several papers in seismology journals (e.g. Israelsson, 1990; Harris, 1991; Riviere-Barbier & Grant, 1993) began to discuss the use of waveform correlation as a measure of similarity between events. Advances in computer processing power were beginning to make such well known mathematical techniques feasible for general purpose data analysis. These papers also described the use of cluster analysis to group and identify the similar events. Unfortunately, no capability existed, outside of research articles, for an analyst to perform such a task.

Dendro Tool was introduced in the late 1990's as a prototype software application that provides the capability to easily determine the similarities between seismic events based upon their waveform data. It does so by cross correlating the waveform data from a group of events and then employing cluster analysis to find groups of similar waveforms. Dendro Tool forms a hierarchical classification tree, called a dendrogram that can be used to gauge the relative similarities between any pair of waveforms. Dendro Tool can be used to characterize the sources of seismic activity for a region by forming a dendrogram composed of the past waveforms for that region. The source of a waveform can then be identified by inserting the new waveform into the dendrogram and observing which cluster the waveform is associated with.

2. KEY TECHNICAL CONCEPTS

2.1 Waveform Correlation

Dendro Tool determines the similarity between pairs of waveforms by computing a cross correlation sequence between the waveform time series. The cross correlation operation can be represented in the time-domain as the dot product between the sections of overlapping waveform data for every possible offset between the two waveforms. Mathematically, the cross correlation is represented in the discrete-time case as follows:

$$C_{i,j}[t] = \sum_{k=0}^{N-1} w_i[k+t] * w_j[k]$$

Basically, the first waveform is shifted past the second waveform in increments of time *t*. At each time step *t* the overlapping portions are correlated. Actual computation of the cross correlation sequence is performed in the frequency domain for performance reasons. However, employing the Fourier Transform, the resulting time-domain and frequency-domain solutions are the same.

Once the cross correlation sequence has been computed, it is necessary to determine the maximum point in the sequence:

$$C_{i,j} = MAX_t\left(C_{i,j}[t]\right)$$

$$T_{i,j} = \text{value of t for which } C_{i,j}[t] = C_{i,j}$$

This maximum value, called the cross correlation, represents how similar the two waveforms are. The point at which the maximum correlation occurred indicates how much the two waveforms should be shifted relative to each other so that they best match one another.

The cross correlation values that have been computed represent how similar the two waveforms are. More similar pairs of waveforms will have larger cross correlation values and less similar pairs of waveforms will have smaller cross correlation values. However, the relative magnitude of the raw cross correlation value is dependent upon the amount of power in each of the signals. This variability in the magnitude makes it difficult to compare cross correlations between different pairs of waveforms. So, in order to come up with a consistent measure of similarity, the cross correlations must first be normalized to account for the amount of power in each of the waveforms.

The correlation coefficient, or coherency, is equal to the cross correlation value scaled by the power in each waveform, also called the auto correlation.

$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} \times C_{j,j}}}$$

Since the cross correlation values are all guaranteed to be positive, the correlation coefficient varies between 0 and 1. When the correlation coefficient is 0, the two waveforms are linearly independent of one another, and thus completely dissimilar. When the correlation coefficient is 1 the two waveforms are completely linearly dependent upon one another and are thus identical.

As an example of cross correlating waveforms, the two waveform segments displayed below both originated from the same mine in southern New Mexico, but occurred a month apart from one another.

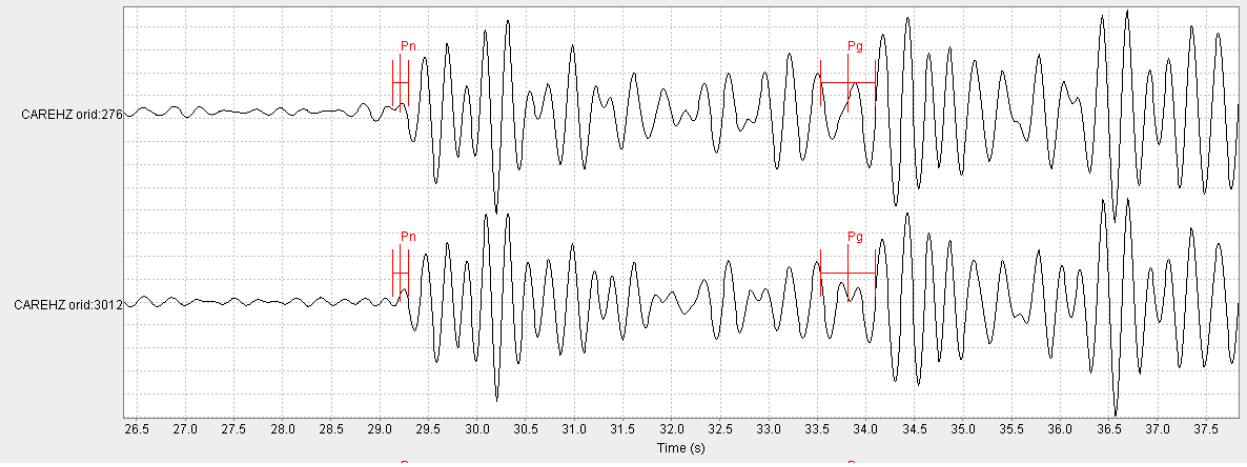

Figure 1. Similar Waveforms

The normalized cross correlation sequence for these two waveform segments is shown below:

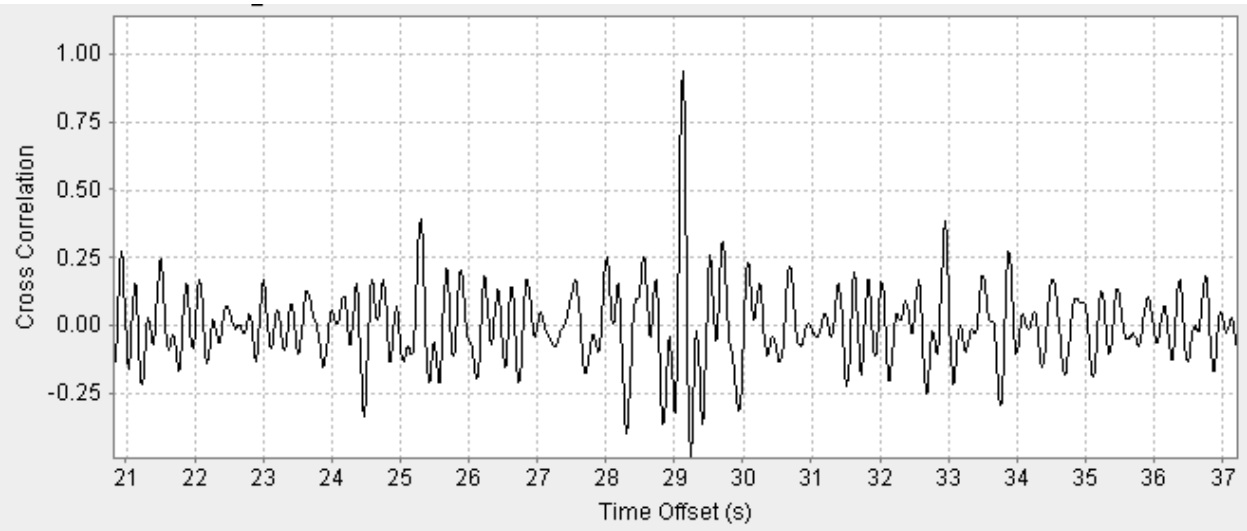

Figure 2. Cross Correlation Sequence

The peak cross correlation value for this pair of waveforms is approximately 0.935. Given the level of background noise observed at this station, you would only expect to see the random occurrence of two waveforms as similar as these two once in over 10^{44} years (Wiechecki, 2001). Needless to say, these waveforms would be considered to be highly similar and originating from the same source.

Using the cross correlation methods described above, Dendro Tool computes and stores correlation coefficients for every pair of waveforms that it is analyzing

2.1 Cluster Analysis

Dendro Tool groups waveforms together using a cluster analysis technique called agglomerative hierarchical clustering (Everitt, 1993). Basically, each of the waveforms is represented as a leaf node in the dendrogram and the similarities between all of the nodes are stored as a matrix of correlation coefficients. The nodes are iteratively grouped together by identifying the two most similar nodes. These two similar nodes are merged into a single node. For the purposes of clustering, the new merged node is treated from then on as a single entity, replacing the two nodes that it makes up. The clustering operation then repeats, merging the next two most similar nodes. This grouping process continues until all of the nodes have been combined into a single tree-like structure called a dendrogram.

The difficult part of clustering involves determining, at each step, how to combine the two nodes into a single composite node. The correlations between all of the pairs of waveforms have already been computed as described in the section on cross-correlations. However, the correlations between the node representing the two linked waveforms and the remainder of the waveforms cannot be computed in that way since the node doesn't actually have any time series data associated with it. Instead, the correlations are derived as a weighted combination of the correlations that correspond to the waveforms that went into the node. As can be expected, there are many different ways to perform this clustering process. However, all of the clustering methods used by Dendro can be performed using a distance weighting function:

$$D_{i,j} = 1 - R_{i,j}$$

$$D_{k,ij} = \alpha_1 * D_{k,i} + \alpha_2 * D_{k,j} + \beta * D_{i,j} + \gamma * \left| D_{k,i} - D_{k,j} \right|$$

R_{ij} is the correlation coefficient between groups i and j. $D_{i,j}$ is the distance between groups i and j. $D_{k,ij}$ is the distance between group k and the union of groups i and j.

The clustering methods supported by Dendro are as follows:

The **Single Link**, or nearest neighbor, method determines the distance between two groups of waveforms to be equal to the distance between the two nearest (most similar) waveforms from each of the groups.

$$\alpha_1 = 0.5$$
$$\alpha_2 = 0.5$$
$$\beta = 0$$
$$\gamma = -0.5$$

The **Complete Link**, or furthest neighbor, method determines the distance between two groups of waveforms to be equal to the distance between the two furthest (least similar) waveforms from each of the groups.

$$\alpha_1 = 0.5$$
$$\alpha_2 = 0.5$$
$$\beta = 0$$
$$\gamma = 0.5$$

The **Median** method determines the distance to be the distance between the medians of each of the two groups of waveforms.

$$\alpha_1 = 0.5$$
$$\alpha_2 = 0.5$$
$$\beta = -0.25$$
$$\gamma = 0$$

The **Group Mean** method determines the distance between two groups of waveforms to be the average of the all the distances between each pair of waveforms from each of the two groups of waveforms.

$$\alpha_1 = n_i/(n_i + n_j)$$
$$\alpha_2 = n_j/(n_i + n_j)$$
$$\beta = 0$$
$$\gamma = 0$$

The **Centroid** method determines the distance to be the squared euclidan distance between the centroids of the two groups of waveforms.

$$\alpha_1 = n_i/(n_i + n_j)$$
$$\alpha_2 = n_j/(n_i + n_j)$$
$$\beta = -n_i * n_j/(n_i + n_j)^2$$
$$\gamma = 0$$

The **Minimum Variance** method determines the distance to be

$$\alpha_1 = (n_i + n_k)/(n_i + n_j + n_k)$$
$$\alpha_2 = (n_j + n_k)/(n_i + n_j + n_k)$$
$$\beta = -n_k/(n_i + n_j + n_k)$$
$$\gamma = 0$$

The **Flexible** method allows the user to manually specify the clustering parameters. A typical set of values that will result in visually appealing groups are provided below.

$$\alpha_1 = 0.625$$
$$\alpha_2 = 0.625$$
$$\beta = -0.25$$
$$\gamma = 0$$

Note that dendrograms built using these values will tend to form groupings of low correlated waveforms prior to grouping other more highly correlated waveforms.

Any of these clustering methods may be used in forming a dendrogram. They will all successfully group highly correlated waveforms. Where they differ is in how they arrange lower correlated waveforms. It is important that the user choose a method of clustering that makes sense for the analysis being performed prior to any computation. Otherwise, there may be a tendency to choose a clustering method that makes the dendrogram "look good" and fit some preconceived notion as to how the waveforms should be grouped. If in doubt, the Single Link method produces the most simple and consistent results.

As an example of the clustering process, a contrived set of correlation coefficients for five waveforms have been assembled into the matrix shown below. The waveforms are arrayed across the rows and columns of the matrix. Each entry in the matrix contains the correlation coefficient that corresponds to the pair of waveforms represented by that row and column. Note that the matrix is symmetric about the diagonal and that each waveform entry is inherently identical to itself. Each of the waveforms has also been positioned as a leaf node along the y axis of the plot that will display the dendrogram.

	WFM1	WFM2	WFM3	WFM4	WFM5
WFM1	**1.0**	**0.95**	**0.25**	**0.35**	**0.5**
WFM2	**0.95**	**1.0**	**0.3**	**0.2**	**0.45**
WFM3	**0.25**	**0.3**	**1.0**	**0.9**	**0.8**
WFM4	**0.35**	**0.2**	**0.9**	**1.0**	**0.75**
WFM5	**0.5**	**0.45**	**0.8**	**0.75**	**1.0**

Table 1. Clustering Example, Step 1

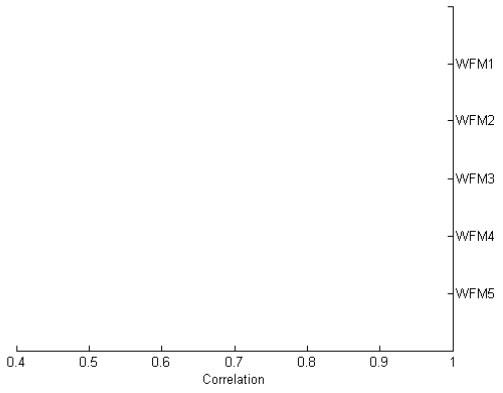

Figure 3. Clustering Example, Step 1

Searching the correlation matrix, waveforms 1 and 2 are identified as being the most similar with a correlation coefficient of 0.95. These two entries are merged into Node 1 and the correlation coefficient values between node 1 and waveforms 3, 4, and 5 are assigned using the single link clustering method.

The updated matrix and plot are shown below.

	NODE1	WFM3	WFM4	WFM5
NODE1	1.0	0.3	0.35	0.5
WFM3	0.3	1.0	0.9	0.8
WFM4	0.35	0.9	1.0	0.75
WFM5	0.5	0.8	0.75	1.0

Table 2. Clustering Example, Step 2

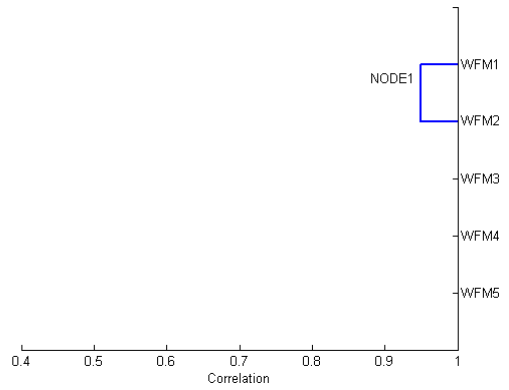

Figure 4. Clustering Example, Step 2

The next two most similar entries are waveforms 3 and 4 with a correlation coefficient value of 0.9. Again, these two entries are joined and the updated matrix and plot are shown below.

	NODE1	NODE2	WFM5
NODE1	1.0	0.35	0.5
NODE2	0.35	1.0	0.8
WFM5	0.5	0.8	1.0

Table 3. Clustering Example, Step 3

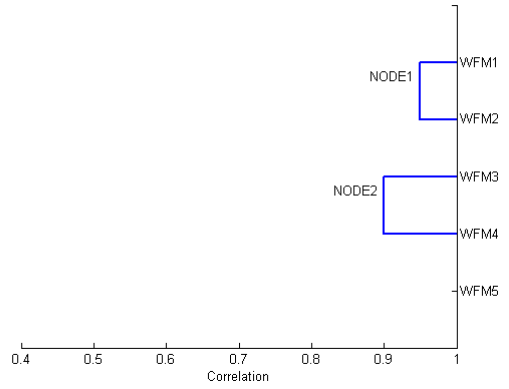

Figure 5. Clustering Example, Step 3

The next two most similar entries are node 2 and waveform 5 with a correlation coefficient value of 0.8. Again, these two entries are joined and the updated matrix and plot are shown below.

	NODE1	NODE3
NODE1	**1.0**	**0.5**
NODE3	**0.5**	**1.0**

Table 4. Clustering Example, Step 4

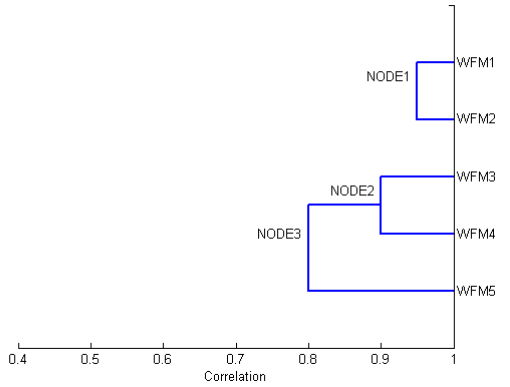

Figure 6. Clustering Example, Step 4

The final entry remaining is between nodes 1 and 3 with a value of 0.5. These two nodes are joined together and with no remaining pairs, the dendrogram is now complete.

	NODE4
NODE4	**1.0**

Table 5. Clustering Example, Step 5

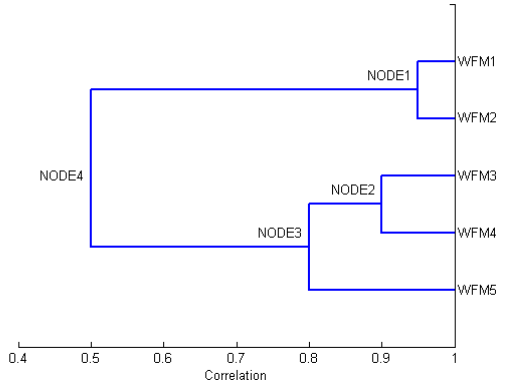

Figure 7. Clustering Example, Step 5

Now that the dendrogram has been completed, it may be used to visually infer what groupings of similar waveforms are present. Also, as a first-order approximation, the similarity between any two waveforms can be estimated by tracing the two waveforms to a common parent node.

3. SOFTWARE REQUIREMENTS

Hardware/OS -- to be useful for the primary targeted user group, it is essential that the Dendro Tool application be able to run on Sun workstations using the Solaris operating systems. Being able to run the software on other systems such as PC/Windows, Mac/OSX, or PC/Linux would be beneficial, but is not of primary importance.

Data Format -- the data that Dendro Tool must read in is stored as a set of database tables that conform to the NNSA core schema (Carr, 2006). These database tables are stored in either an SQL database such as Oracle or as a set of ASCII text files. In addition, Dendro Tool must be able to export any data it has read in to in either of these two data formats.

Performance -- our experience with the targeted group of users is that this sort of tool must be able to perform clustering of a reasonable set of waveforms – say up to several hundred – in less than at most a few tens of seconds, preferably much quicker. Use of the tool will depend directly on how quickly the clustering is done.

User Interfaces -- the tool will be used by users with excellent knowledge of the seismic waveforms, but little familiarity with cluster analysis, and so must present the clusters in a way that is as simple as possible. Further, the tool must provide ready access to the waveforms that are being clustered with standard review capabilities (zooming, panning). Also, it is essential for the user to be able to compare the map locations of the events with the clustering results.

4. SOFTWARE DESIGN

Dendro Tool was developed primarily as a Graphical User Interface (GUI). It is an application that provides analysts with the ability to easily read in data from their database and examine any waveform, spatial, or temporal similarities between the seismic events.

The initial development of Dendro Tool was performed using Matlab. Matlab is a numerical computing environment and programming language that provides many built-in capabilities for data analysis, processing, and display. Use of Matlab allowed for rapid experimentation and development using various methods of waveform processing, correlation, and clustering.

Once Dendro Tool had been in use for some time the key technical concepts became well vetted and the number of requests for new features began to subside. At that point, it became advantageous to redevelop Dendro Tool as a stand-alone application that would be better able to address customer requirements. The goals for re-developing Dendro Tool were to:

- **Eliminate dependence on Matlab** – Matlab's strengths as a prototyping platform became weaknesses as Dendro Tool began to take on a more operational role. Matlab is primarily a scripting language and so it lacks many of the capabilities found in modern programming languages such as strong typing, compile time checking, and object oriented design.

- **Improve user interfaces** – Dendro Tool's user interfaces were initially developed as functional prototypes. While these interfaces allowed for a great deal of flexibility in researching various configuration parameters, they were ill-suited to provide a stream-lined and easy to use application. In addition, the style and responsiveness were limited by Matlab's Graphical User Interface (GUI) widgets. Redeveloping the user interface with a richer set of GUI widgets and a focus on simplifying common user tasks allowed the application to become more beneficial to its users.

- **Improve computational performance -** Matlab's data processing libraries and scripting language implementation tend to experience a performance hit when dealing with large volumes of data. Implementing Dendro's core processing algorithms directly allowed the algorithms to be streamlined and executed much faster. This becomes especially important when consideration of our target users' mixed computing environment (some new machines, but also several older, less-powerful machines) is taken into account

Dendro Tool is currently written in Java, a cross platform software language developed by Sun Microsystems. Some of the benefits of using the Java programming language are:

- Cross platform. Java applications can be written and compiled once and then run it on any platform that has a Java Virtual Machine (Sun/Solaris, Unix, Linux, Mac, PC/Windows)

- Rich and powerful language with many features that simplify software design and development.

- Availability of 3rd party libraries.

- Fast computational performance equivalent to native C/C++ implementations.

- Quickly becoming the preferred development language on the SNL GNEM R&E project.

The re-development of Dendro Tool as a Java application made use of several third party libraries. In addition, Dendro Tool makes use of several libraries that have been developed internally to the GNEM project:

- Colt – a Java library for high performance scientific and technical computing. Dendro Tool uses Colt for solving systems of linear equations.

- JFreeChart – a Java library for generating charts and plots. Dendro uses JFreeChart for generating all of its plots such as the dendrogram plot, waveform time series, histograms, etc.

- DBToolBox – A Sandia/GNEM developed java library for the access and management of complex database schemas. Dendro uses DBToolBox to interface with both SQL and Flatfile databases.

- PGL (Parametric Grid Library) – A Sandia/GNEM developed C++ library (with a Java Interface) for the access and management of geophysical models. Dendro Tool uses PGL for making predictions of seismic propagation time through the earth.

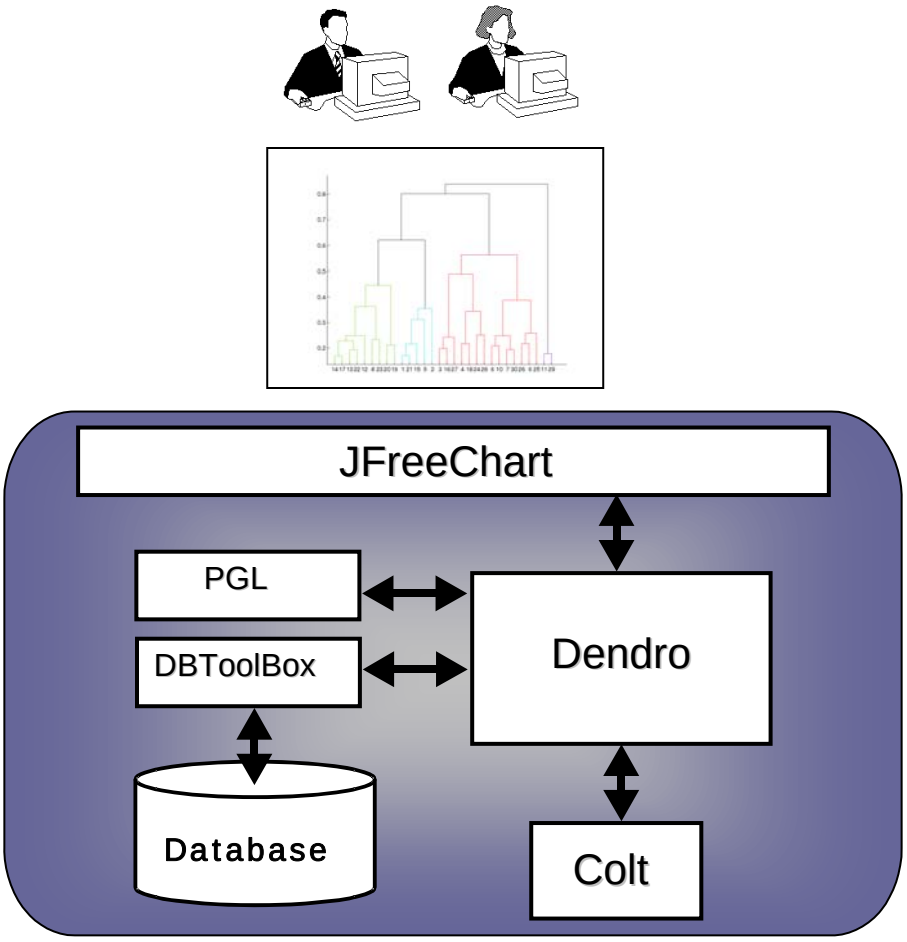

Figure 8. Dendro Tool Design

5. USING THE DENDRO TOOL

5.1 Overview

The intent of this section is not to provide a User's Manual for the Dendro Tool – such a manual already exists and is part of the software release – but to give a general sense to the reader of how the tool works and how it can be used.

5.1.1 Dendrogram Window

The **Dendrogram Window** is the primary interface for Dendro Tool in which the dendrogram is displayed. The correlation level is indicated on the x-axis and the waveforms are ordered along the y-axis. Each waveform is labeled with its station, channel, and origin identification number (if known) or date. The labels are normally colored black. Waveforms that are selected are colored blue.

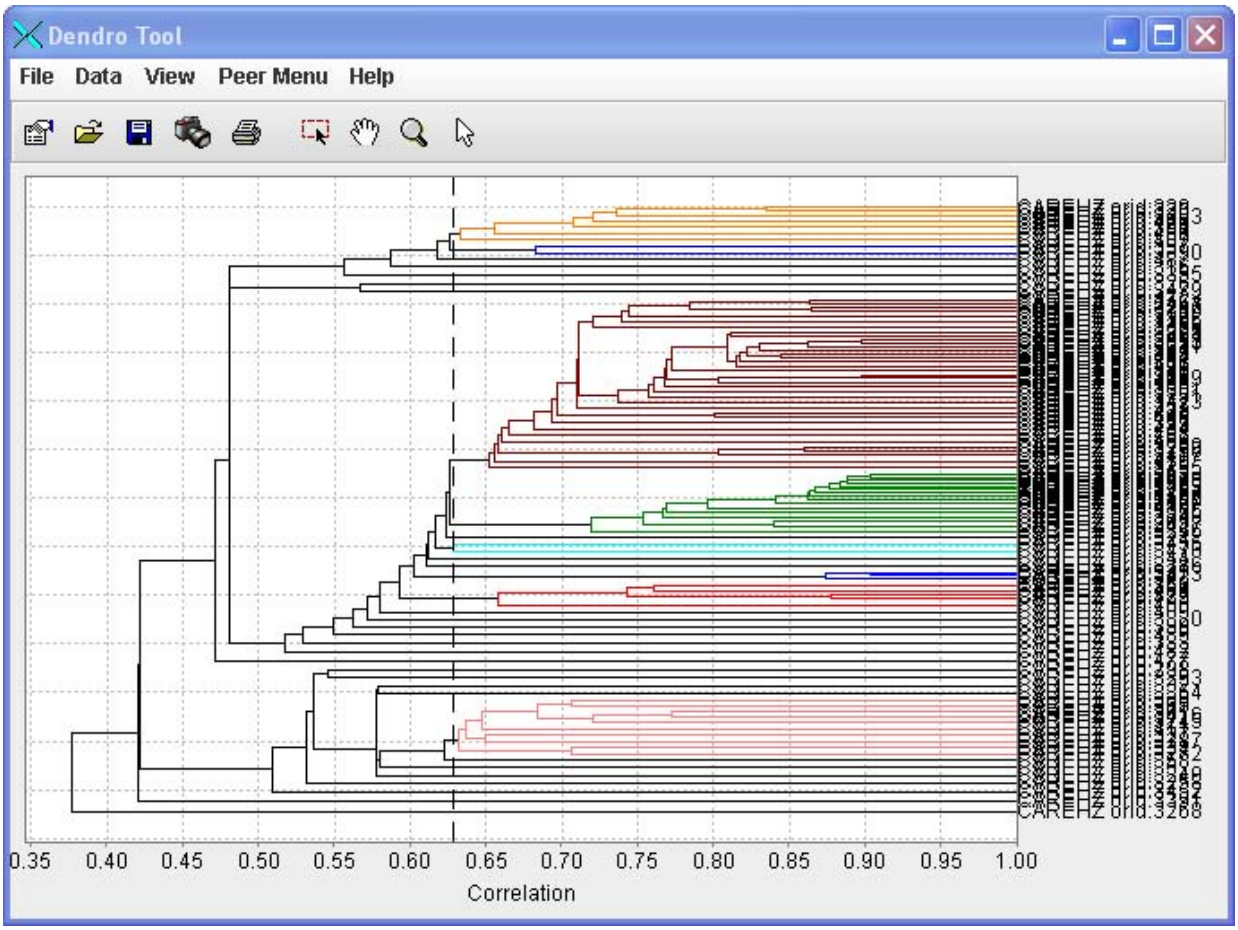

Figure 9. Dendrogram Window

Within the dendrogram window is a **Threshold Line** that allows the dendrogram to be divided into clusters. The threshold level can be adjusted by the user with the mouse. Each cluster within the dendrogram has its own color that can be used to help differentiate the clusters.

The Dendrogram Window supports many methods of user interaction including zooming, panning, node selection, and setting the threshold. These controls allow the user to easily navigate through the dendrogram and analyze the data within it.

5.1.2 Dendrogram Properties

The Dendrogram Properties dialog controls various aspects of how the dendrogram is computed and displayed.

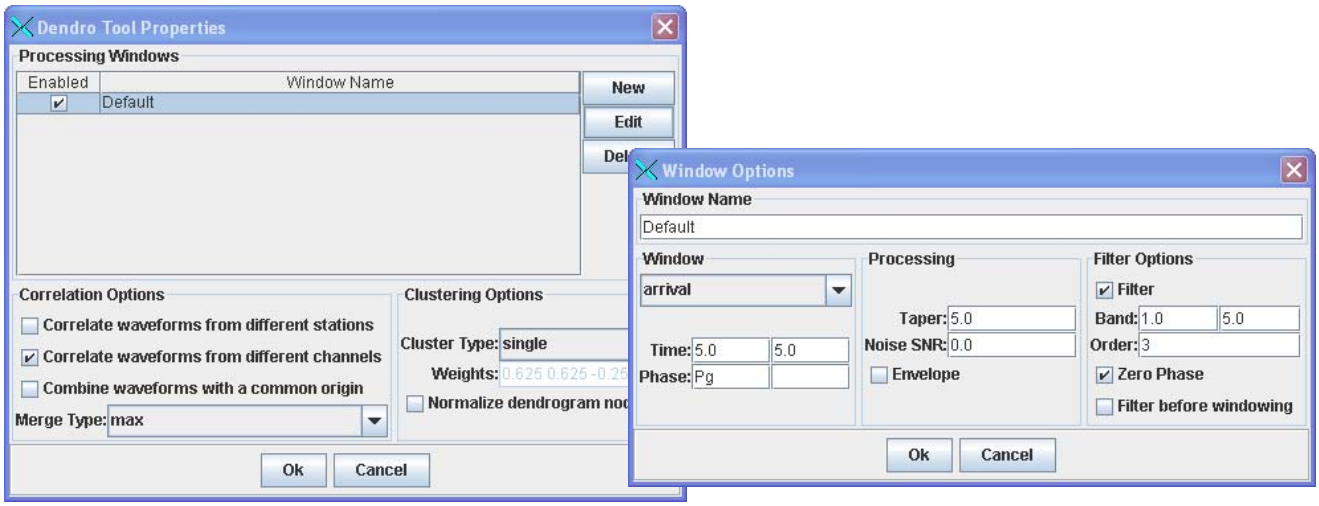

Figure 10. Dendrogram Properties

The user may define one or more windows for processing the waveforms prior to correlation. The processing window may include configurations such as a time segment to cut, tapering the ends, and applying a band-pass filter. The user may also specify various options that control how the correlations and clustering are performed.

5.1.3 Waveform Viewer

The user may display the waveforms for a selected subset of the dendrogram. The waveforms are positioned along the Y-Axis in the same order as they appear in the dendrogram. In addition, the waveforms are displayed with the same color as the corresponding node entry in the dendrogram to make identification between the windows easier. The X-Axis is the relative time, in seconds, since the start of each of the waveforms. The waveforms are initially optimally lagged so that they are all aligned to the point of maximum correlation. Arrivals, which indicate the estimated time at which a unique portion of the seismic energy reached the station, are also displayed alongside the waveforms.

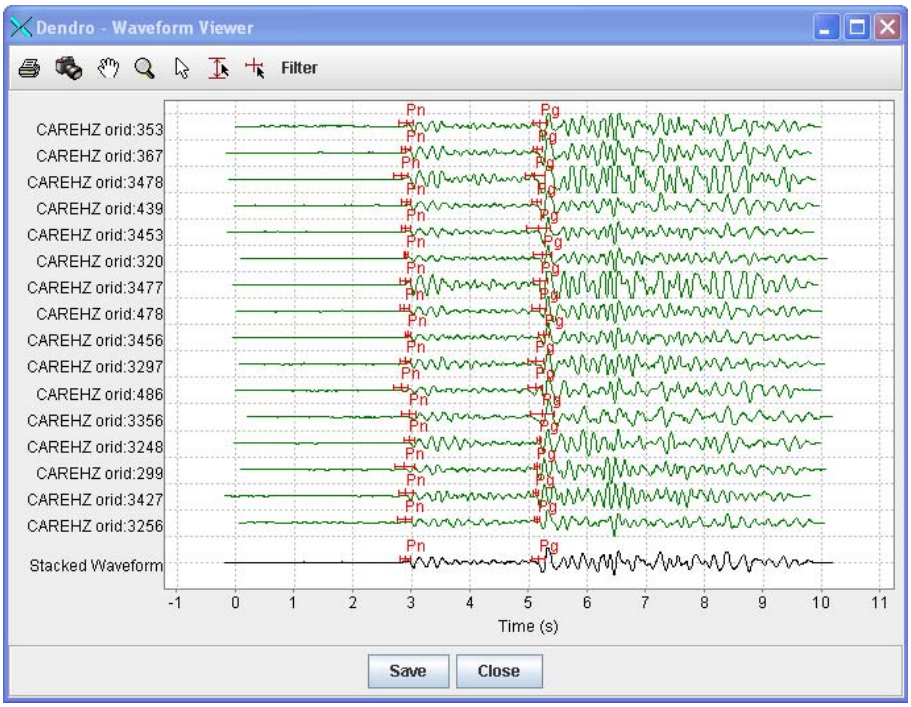

Figure 11. Waveform Viewer

The Waveform Viewer also contains other controls allowing the user to zoom and pan the display. Waveforms can be filtered and dragged across the window to allow the user to easily compare and overlay the waveforms.

The stacked waveform displayed on the bottom of the diagram is an average of all of the shifted waveforms in the viewer. Dragging any of the waveforms causes the stacked waveform to be dynamically updated. Because waveforms from the same source are generally coherent and noise sources are generally incoherent, the stacked waveform can exhibit up to a sqrt(N) increase in the signal to noise ratio, where N is the number of waveforms.

Finally, the Waveform Viewer also allows the user to make changes to the arrival times for each of the waveforms. Having multiple similar waveforms present allows the user to easily make more consistent and accurate arrival time predictions. Increasing the accuracy of the arrival time predictions can result in computing more accurate locations for the events of interest.

5.1.4 Correlation Matrix Viewer

The user may display the correlation matrix that was used in the clustering process to create the dendrogram. The waveform labels are displayed along the Y-Axis and assigned the same colors as in the dendrogram. Note that the correlation matrix is symmetric about the diagonal. In addition, white boxes have been drawn around the sections of the correlation matrix that have been identified as clusters by the threshold line in the dendrogram.

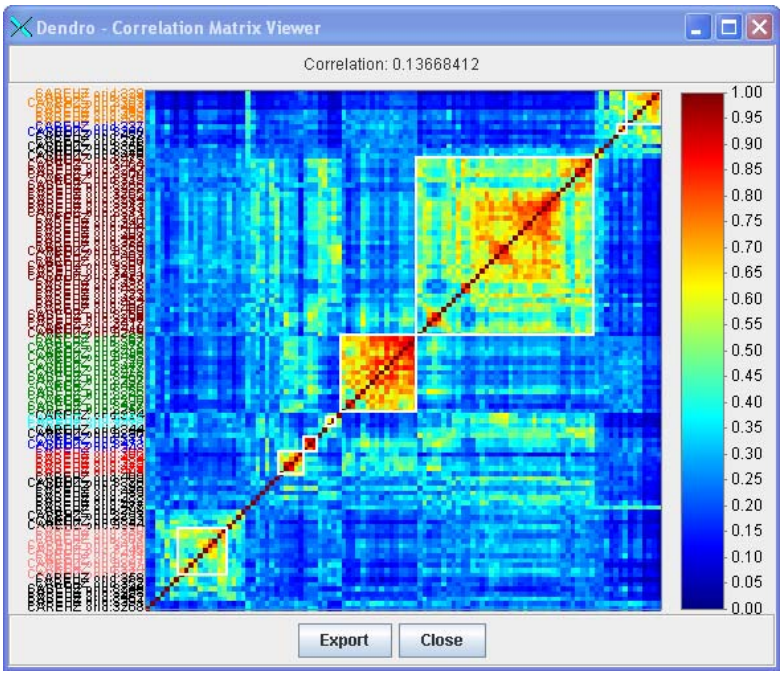

Figure 12. Correlation Matrix

The Correlation Matrix Viewer is a tool that can be useful to the user in setting the threshold level since ideally the boxes representing clusters would include regions of high correlations and exclude regions of low correlations. The user can change the threshold level within the Dendrogram Window and the Correlation Matrix Viewer will automatically update the white boxes to reflect the new clusters.

5.1.5 Clustering Metrics

Dendro Tool contains a number of clustering metrics that measure various statistics about the clustering process. These metrics can be used to identify key transition points in the clustering that tend to indicate a good place to set the dendrogram threshold line that determines the number of clusters.

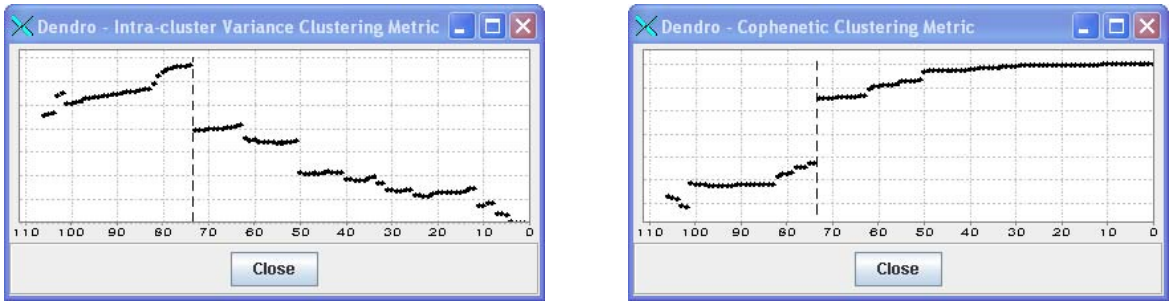

Figure 13. Clustering Metrics

The two clustering metrics shown above are the Intra-Cluster Variance and Cophenetic Clustering Metrics. Both of the metrics exhibit a sharp transition at the same point in the

clustering. The threshold line has been assigned to that point as a good place to divide the dendrogram into separate clusters.

5.1.6 Time Histograms

In addition to displaying the waveform data, users may also wish to display information about when then seismic events occurred. Man made seismic sources tend to fall into a typical Monday to Friday, 9 am to 5 pm work schedule.

Dendro Tool can display temporal histograms such as Hour of Day and Day of Week histograms shown below. The individual entries in the histogram are assigned the same color as in the dendrogram window.

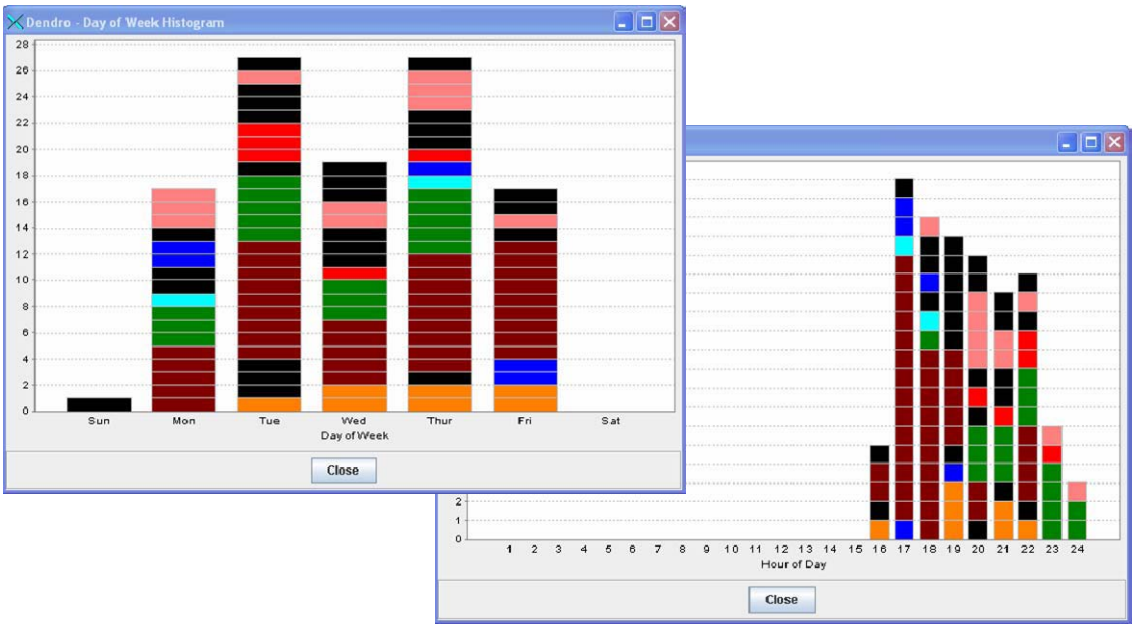

Figure 14. Time Histograms

5.1.7 Map

The location of the events in Dendro Tool can be plotted on a map. The color displayed at each of event location corresponds to the color assigned in the dendrogram. Using the map, the user can determine whether the events within a cluster have near-by computed locations.

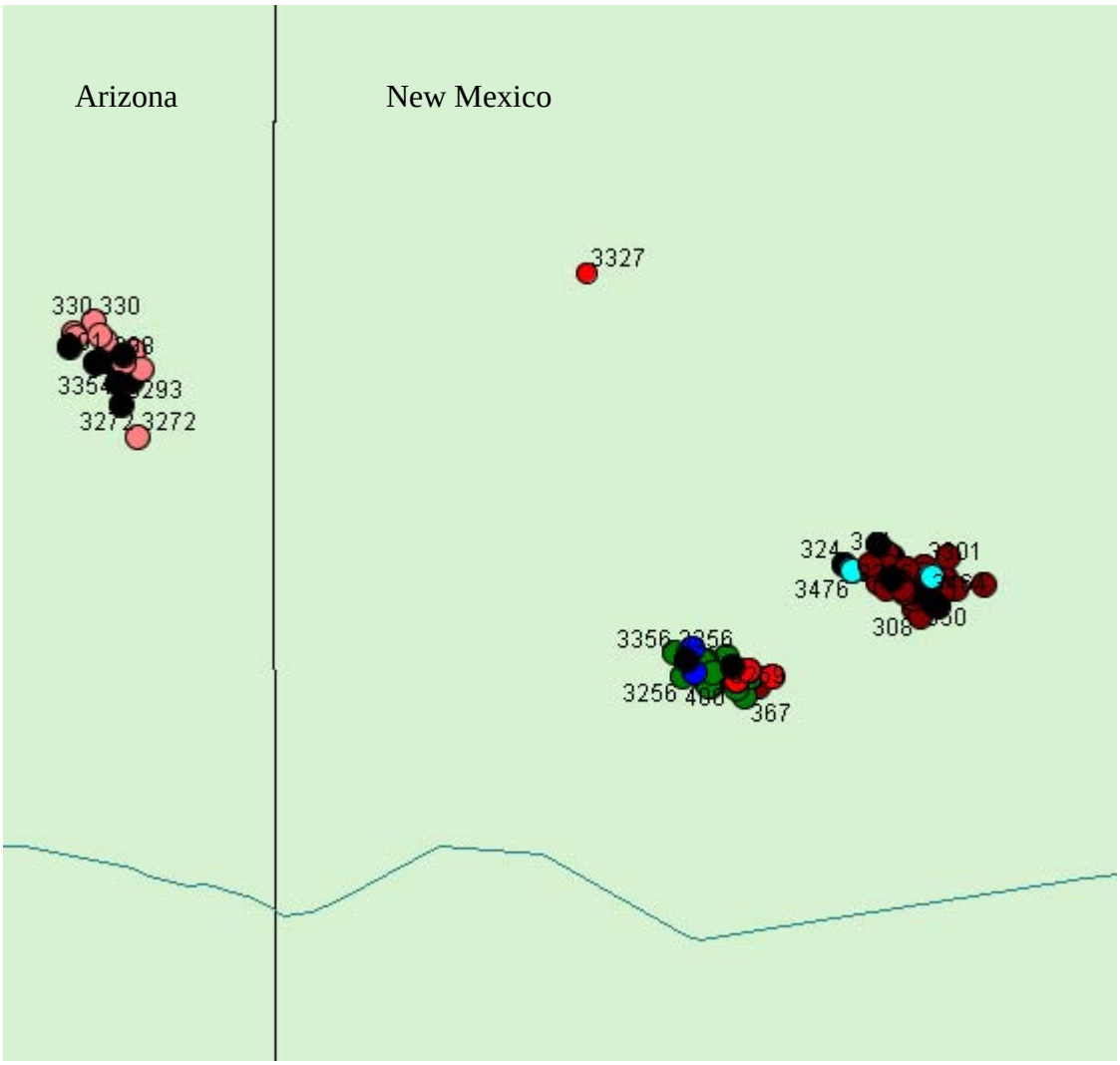

Figure 15. Map with Clustered Events

5.2 Intended Uses

5.2.1 Building a Dendrogram from the Map (requires GNEMRE KBNav software)

Dendro has the ability to create a dendrogram from events displayed on a map. The first step in creating a dendrogram from the map is to read in the events from the database.

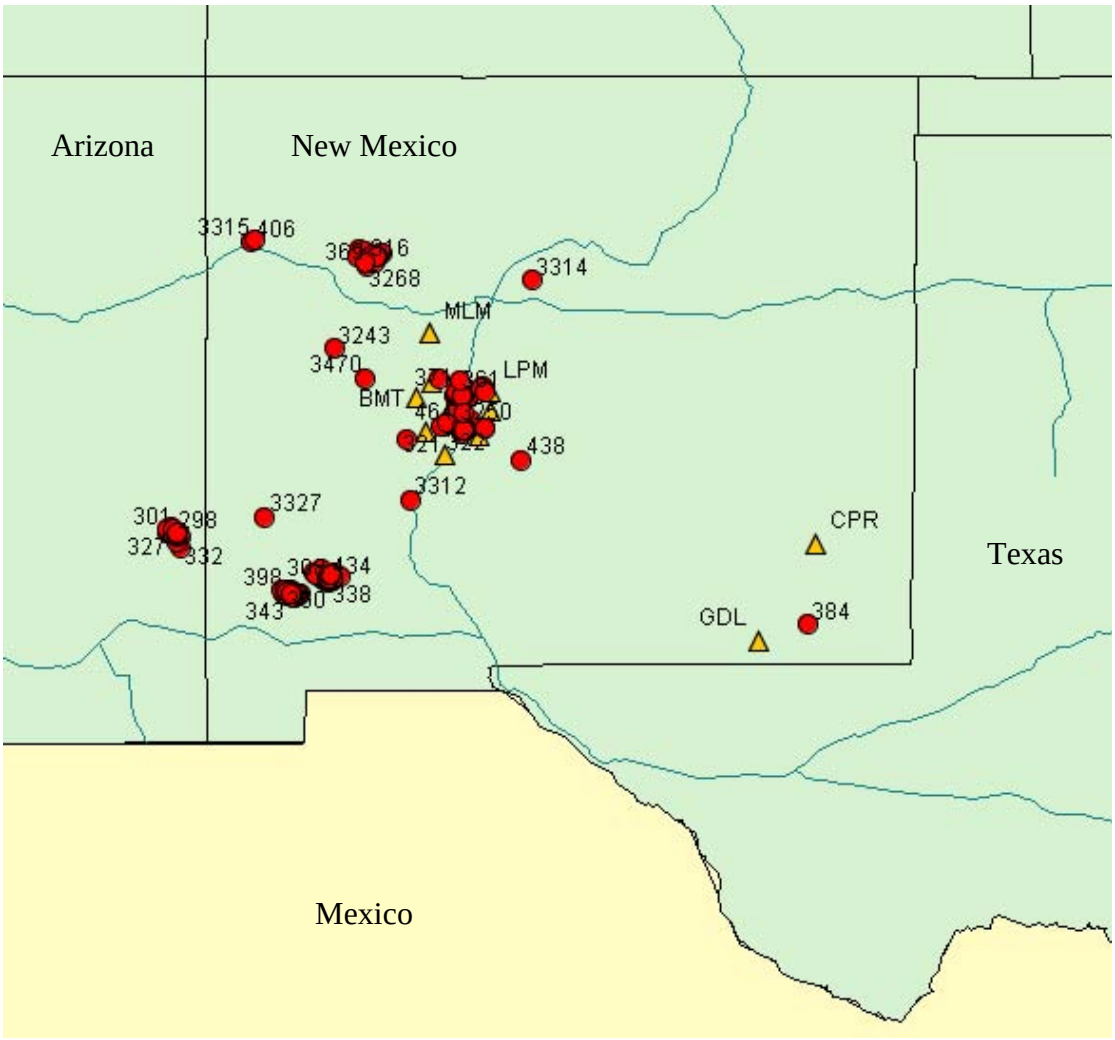

Figure 16. Building a Dendrogram using the Map

If the user also has available the GNEMRE KBNav software application, then the user can construct a database query to read in and select a large set of events to display on the KBNav map, and then graphically select the set of interest for clustering. The user can then launch the Dendro database read window to read in the waveforms for the events that have been selected in KBNav.

This capability allows the user to focus more easily on building a dendrogram for a precise geographical region.

Without the KBNav map, a user can still read data directly into Dendro Tool, but geographic constraint is limited to min/max latitude and longitude.

5.2.2 Dividing a dendrogram into individual clusters

After the events and waveforms have been read from the database and used to form a dendrogram, the next step in the analysis is to specify a threshold level to divide the dendrogram into clusters.

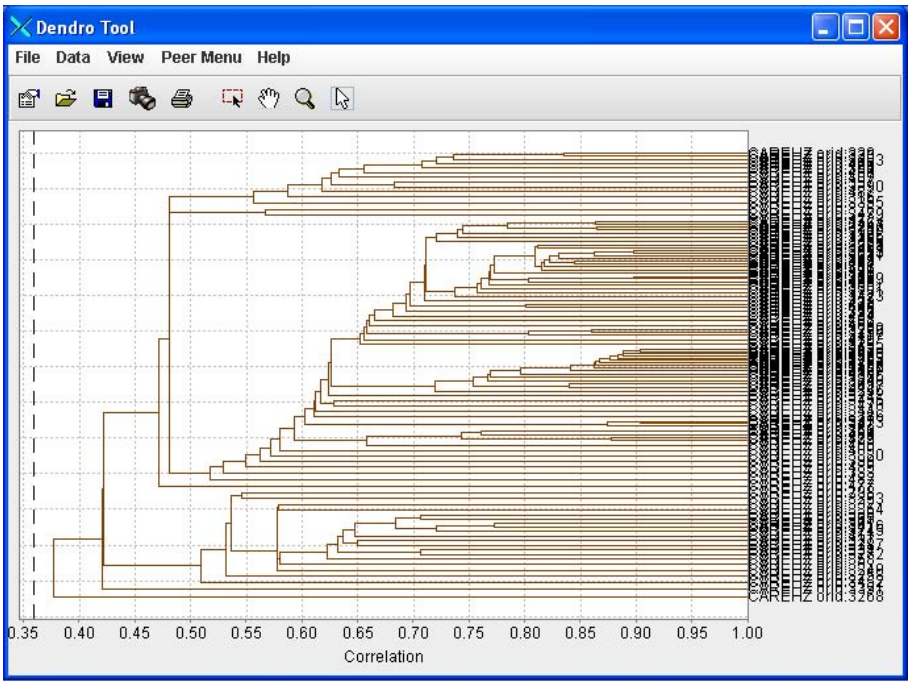

Figure 17. Dendrogram before Clustering

Dendro Tool allows the Analyst to adjust the threshold level simply by clicking on the threshold line in the dendrogram window and drag the line to its desired location. As discussed above, Dendro Tool also provides several utilities to assist the Analyst in determining an appropriate location for the threshold line.

Some of these utilities include the Correlation Matrix Viewer and several Clustering Metrics. The utilities all update dynamically as the Analyst repositions the threshold level.

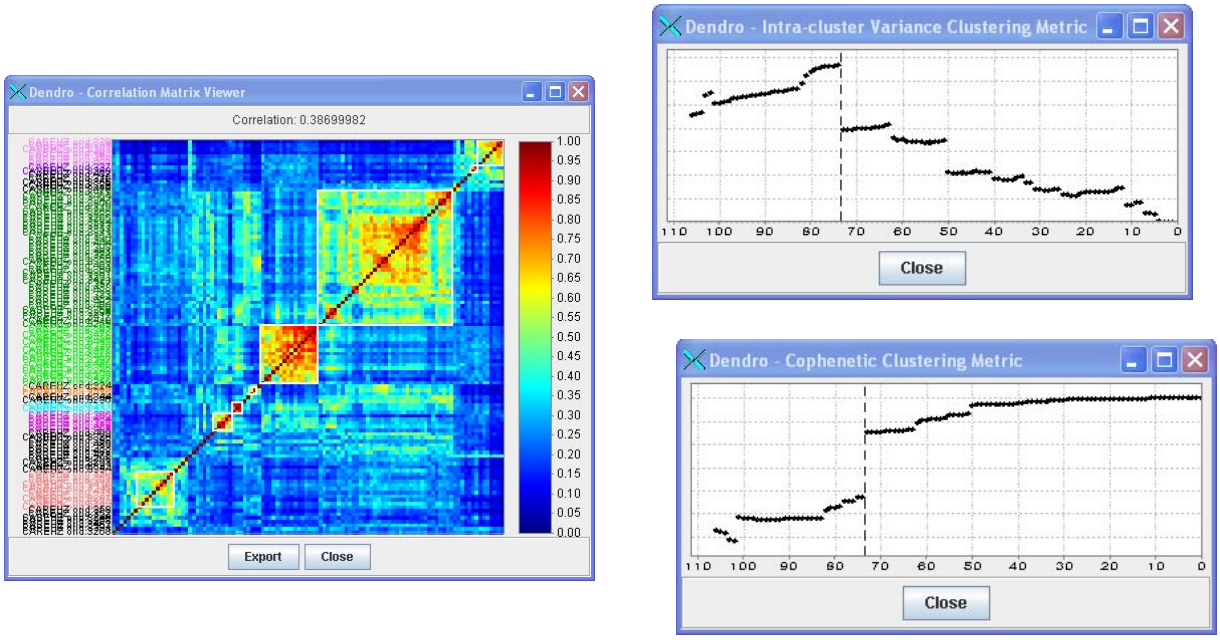

Figure 18. Clustering Threshold Level Determination Utilities

Once the threshold level has been specified, the level is used to cut the dendrogram into individual clusters. Each cluster is assigned a unique identification color as shown below.

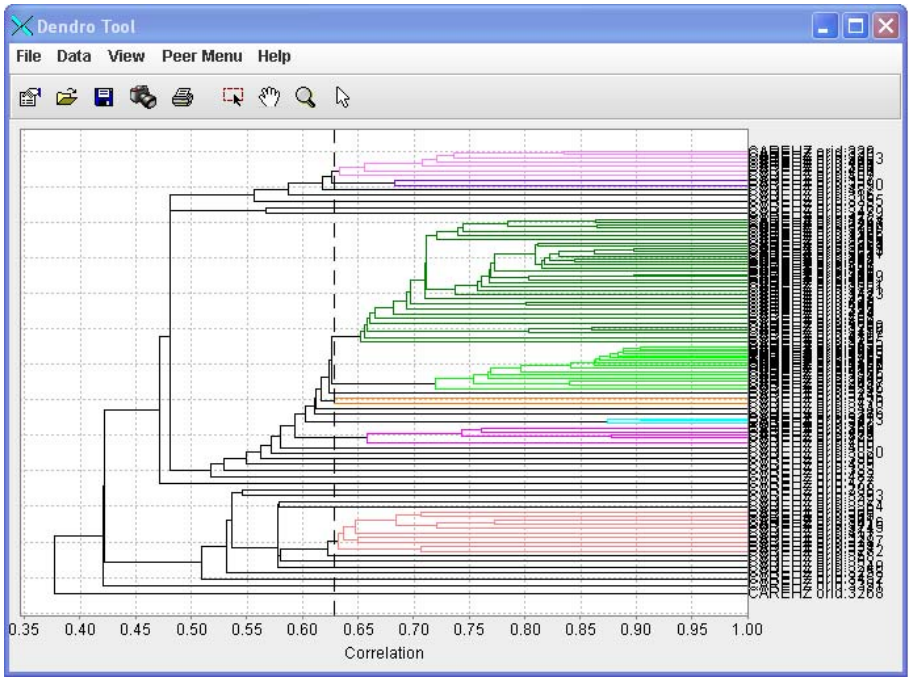

Figure 19. Dendrogram after Clustering

5.2.3 Re-timing arrivals using the waveform correlations

A common task when using the Dendro Tool is to display the waveforms for a group of similar events. Once the waveforms are displayed, as shown below, an Analyst can use the arrival

controls in the Waveform Viewer to make improvements to the consistency and accuracy of the arrival times.

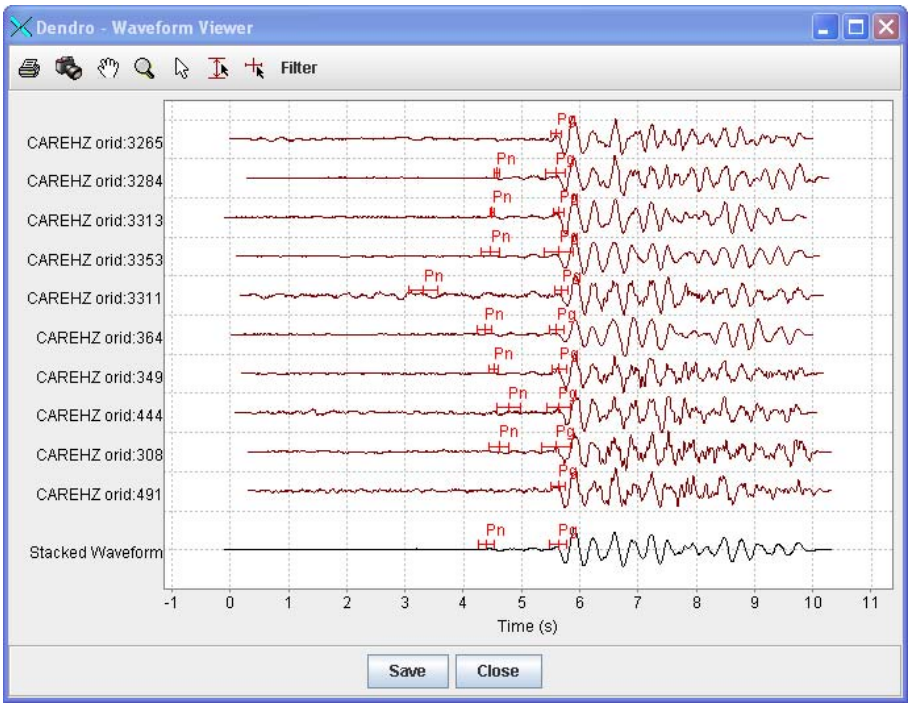

Figure 20. Waveforms Before Arrival Re-timing

Each arrival is represented as a vertical line with a phase name label indicating the arrival time along with one standard deviation error bars on either side. Arrivals can be added, removed, or adjusted as need be using the mouse.

Figure 21. Arrival Closeup Showing Pick Time and Error Bar

The quality of arrival times is a direct result of the analyst who is making the picks as well as the quality of the data that they have to work with. Having multiple waveforms visible side by side allows the analyst to come up with a more accurate estimate of the arrival times for all of the waveforms in question (Rowe, 2001). Mapping arrivals will transfer the relative arrival time and

error from one waveform to all of the other waveforms. So, the arrivals will be made to line up perfectly within the waveform viewer.

To map an arrival, enable arrival mapping using the menu controls in the Waveform Viewer. Then, click on the desired reference arrival within the waveform plot. That reference arrival time is mapped to all the arrivals with the same name for the other waveforms. Often the arrivals on the Stacked Waveform are used as reference since the Stacked Waveform typically has a higher signal to noise ratio than any of the other waveforms.

Once arrivals have been re-timed, a much more consistent set of arrival times are generated as shown below.

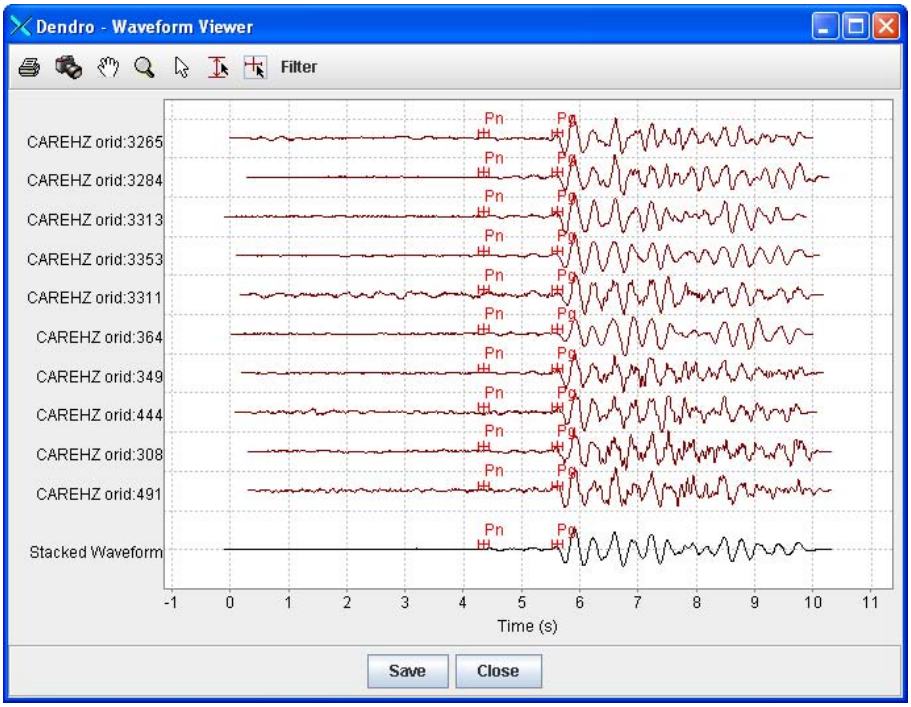

Figure 22. Waveforms After Arrival Re-timing

5.2.4 Determining the statistical significance of two similar waveforms

Dendro Tool provides analysts with the ability to determine a measure of similarity between seismic events. This measure of similarity is computed as a correlation coefficient value between two waveform time series. However, a particular correlation coefficient value does not provide any information on the statistical significance of that similarity.

Dendro Tool provides the capability to assess the statistical significance, represented as a false alarm rate, of a particular correlation coefficient value when a defined background noise level is present (Wiechecki, 2001).

By first selecting two event waveforms in the dendrogram, a master and a candidate waveform, the analyst can make use of the False Alarm Rate utility. The master waveform is used to define a noise window and a signal window. The noise and signal waveform segments are cross-correlated to generate a set of random correlation coefficient values. The distribution of these values can be used to compute a false alarm rate curve for a variety of correlation coefficients, as shown below.

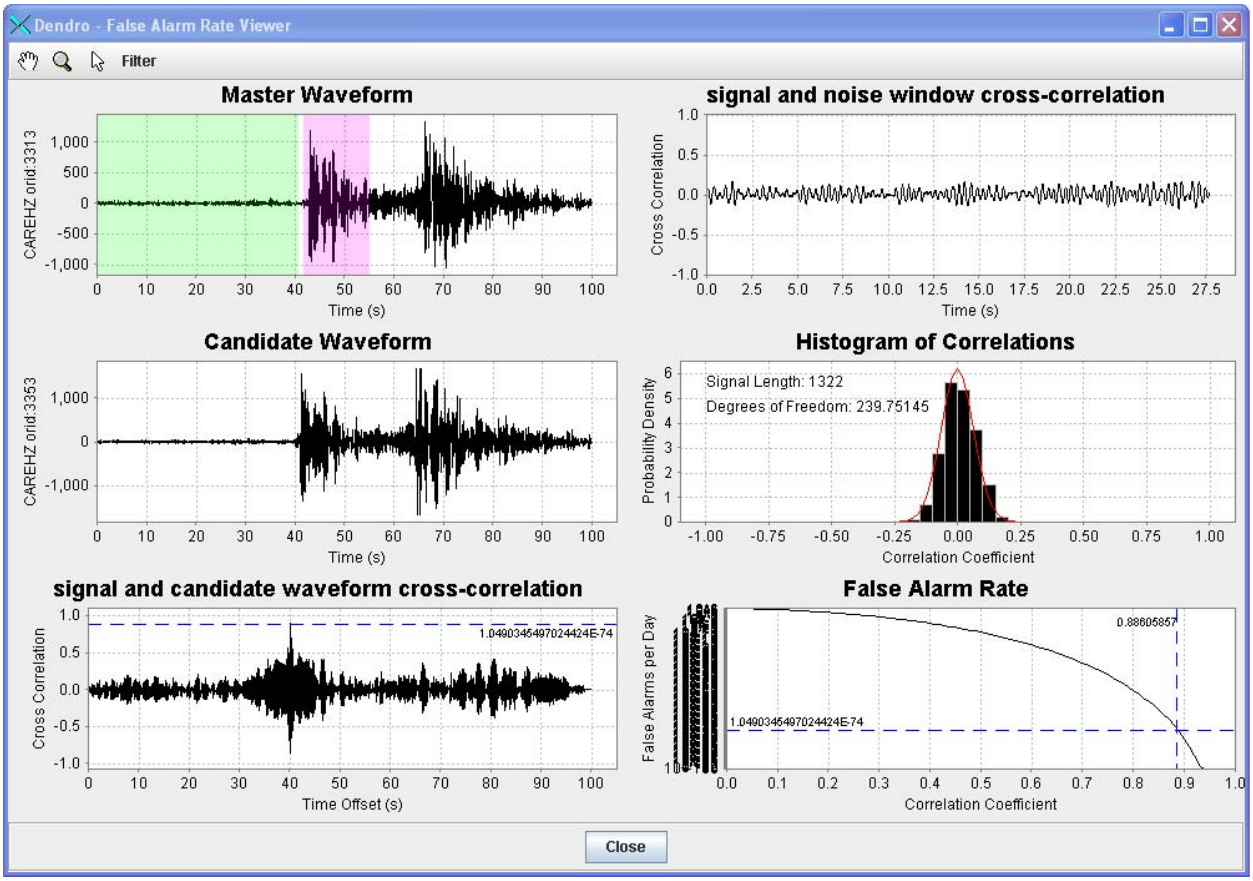

Figure 23. False Alarm Rate

A particular correlation coefficient value can then be used to lookup its corresponding false alarm rate. In the example shown above, the master and candidate waveform exhibit a peak correlation coefficient value of approximately 0.886. Looking at the False Alarm Rate curve indicates that value corresponds to a false alarm rate of 1.05×10^{-74} false alarms per day. Or, by inverting that value, approximately 10^{74} days between false alarms.

So, now an analyst has some measure of what meaning a particular correlation coefficient value has.

6. FUTURE DEVELOPMENT

Dendro Tool has reached a level of stability and maturity such that there are few requests for additional features. Continuing development consists primarily of providing basic support to fix any problems that may arise. Dendro Tool is a successful example of bridging research ideas into an operational tool. It has validated several research concepts introduced by the seismic monitoring research community, such as waveform correlation for seismic event identification, cluster analysis, and automated arrival re-timing in an operational analysis environment. It has become a mechanism for readily applying knowledge of past waveform signals and arrival time picks to current seismic events.

A possible direction for future work is to examine applying these techniques in a less user-driven manner. Waveform correlation based detection, identification, and arrival timing could be incorporated into an automated pipeline processing prior to any analyst involvement. There has already been significant work performed in developing waveform correlation based detectors (Harris, 1991). However, these applications have been fairly static and limited to small regional spotlight areas.

Making use of existing catalogues of reference events to perform automated cross-correlation detections of new events and assign arrival times would help to greatly reduce the need for the analysts to handle recurring events on an individual basis. Such a system would be able to easily process after-shocks seen from large earthquakes where more traditional analyst processing would fall behind. In addition to the time savings, the detection rate and location accuracy would also be improved.

Based on operational experience gained using Dendro Tool for analysis, we believe such a system should be able to perform well. Additional work would need to be performed in order to further validate the concepts in an autonomous environment as well as to determine the computational requirements.

7. REFERENCES

Carr, D.B. (2006). National Nuclear Security Administration Knowledge Base Core Table Schema Document, Sandia National Laboratories Report #SAND2002-3055.

Davis, J.C. (1986). *Statistics and Data Analysis in Geology, 2nd Edition*, JohnWiley & Sons, New York.

Everitt, B.S. (1993). Cluster Analysis, Edward Arnold, London.

Harris, D. B. (1991). A waveform correlation method for identifying quarry explosions, Bull. Seism. Soc. Am., vol. 81, no. 6, p. 2935-2418.

Israelsson, H. (1990). Correlation of waveforms from closely spaced regional events, Bull. Seism. Soc. Am., Dec. 1990; vol. 80, no.6, pt. B, p. 2177-93.

Kraznowski, W.J. (1988). Principles of Multivariate Analysis: A User's Perspective, Oxford University Press, Oxford.

Lance, G.N. and Williams, W.T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. Comp. J., 9, 373-380.

Riviere-Barbier, F. and L. T. Grant (1993). Identification and location of closely spaced mining events, Bull. Seism. Soc. Am., vol. 83, no. 5, p.1527-46.

Rowe, C.A., R.C. Aster, B. Borchers and C.J. Young (2002). An Automatic, Adaptive Algorithm for Refining Phase Picks in Large Seismic Data Sets. Bull. Seism. Soc. Am., vol. 92, no. 5, p. 1660-1674.

Wiechecki, S., H.L. Gray, W.A. Woodward (2001). Statistical Development in Support of CTBT Monitoring. DTRA Report # DSWA01-98-C-0131.

DISTRIBUTION

1 Leslie Casey
NNSA Office of Nonproliferation Research and Development/NA-22
1000 Independence Avenue SW
Washington, DC 20585

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

1 Dean Clauter
Air Force Technical Applications Center/TTR
1030 S. Highway A1A
Patrick AFB, FL 32925-3002

1 Jon Creasey
Air Force Technical Applications Center/TTR
1030 S. Highway A1A
Patrick AFB, FL 32925-3002

1 Michelle Crown
Air Force Technical Applications Center/TTR
1030 S. Highway A1A
Patrick AFB, FL 32925-3002

1 John Dwyer
Air Force Technical Applications Center/TTR
1030 S. Highway A1A
Patrick AFB, FL 32925-3002

1 Jeff Miller
Air Force Technical Applications Center/TTR
1030 South Highway A1A
Patrick AFB, FL 32925-3002

1 Jorge Roman-Nieves
Air Force Technical Applications Center/TTR
1030 South Highway A1A
Patrick AFB, FL 32925-3002

1 Kevin Hutchenson
Air Force Technical Applications Center/TTR/TNDC/QTSI
Suite 514

1980 N. Atlantic Avenue
Cocoa Beach, FL 32931

1 Michael Begnaud
Los Alamos National Laboratory
MS F6665
P.O. Box 1663
Los Alamos, NM 87545

1 W. Scott Phillips
Los Alamos National Laboratory
MS D408
EES-11 P.O. Box 1663
Los Alamos, NM 87545

1 George Randall
Los Alamos National Laboratory
MS D408
EES-11 P.O. Box 1663
Los Alamos, NM 87545

1 Charlotte Rowe
Los Alamos National Laboratory
MS D408
EES-11 P.O. Box 1663
Los Alamos, NM 87545

1 Richard Stead
Los Alamos National Laboratory
MS D408
EES-11 P.O. Box 1663
Los Alamos, NM 87545

1 Lee Steck
Los Alamos National Laboratory
MS D408
EES-11 P.O. Box 1663
Los Alamos, NM 87545

1 Rodney Whitaker
Los Alamos National Laboratory
MS J577
EES-2 P.O. Box 1663
Los Alamos, NM 87545

1 Leigh House

NNSA Office of Nonproliferation Research & Development/NA-22/LANL
GH-068
1000 Independence Ave SW
Washington, DC 20585

1 David Harris
Lawrence Livermore National Laboratory
L-205
P.O. Box 808
Livermore, CA 94551

1 Stephen Myers
Lawrence Livermore National Laboratory
L-205
P.O. Box 808
Livermore, CA 94551

1 Anton Dainty
Air Force Research Laboratory/VSBYE
29 Randolph Road
Hanscom AFB, MA 01731-3010

1 Rick Schult
Air Force Research Laboratory/VSBYE
29 Randolph Road
Hanscom AFB, MA 01731-3010

1 Harley Benz
USGS National Earthquake Information Center
Box 25046 MS966
Denver Federal Center
Denver, CO 80225-0046

1 Paul Richards
Columbia University
Lamont Doherty Earth Observatory
61 Route 9W
Palisades, NY 10964

1 Clifford Thurber
University of Wisconsin
Dept. of Geology and Geophysics
1215 W. Dayton St.
Madison, SI 53706-1600

1 Aaron Velasco

University of Texas El Paso
Geological Sciences
El Paso, TX 79968-0555

1 Steven Taylor
Rocky Mountain Geophysics
167 Piedra Loop
Los Alamos, NM 87544

1	MS0401	Jake Jones	05533
2	MS0401	Chris Young	05533
1	MS0404	Eric Chael	05714
1	MS0404	John Merchant	05736
1	MS0404	Mark Harris	05736
1	MS0404	Darren Hart	05736
1	MS0750	Greg Elbring	06314
1	MS0750	Neill Symons	06314
1	MS0750	David Aldridge	06314
1	MS1168	Robert Abbott	01647
2	MS9018	Central Technical Files	8944
2	MS0899	Technical Library	9536


Sandia National Laboratories