COMPARISON OF CLUSTER ANALYSIS METHODS FOR IDENTIFYING REGIONAL SEISMIC EVENTS

Christopher J. Young[1], Bion J. Merchant[1], Richard C. Aster[2]
Sandia National Laboratories[1]
New Mexico Institute of Mining and Technology[2]

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. DE-AC04-94AL85000

ABSTRACT

In recent years, nuclear explosion and non-proliferation monitoring have focused on smaller yield events, creating two major issues. First, smaller events are typically more difficult to detect and locate. Second, as characterized by the Richter/Gutenberg frequency of occurrence relation, there are many more small events than large events so the monitoring workload is exponentially increased. These issues represent a new challenge for the nuclear explosion monitoring community, but for regional network operators they are commonplace and, to a large extent, solved. These operators routinely locate and identify large numbers of events at least as small as those of interest to the monitoring community, often on the basis of as little as one waveform. The operators accomplish this seemingly impossible task by simply viewing and recognizing the similar waveforms from repeating seismic sources such as mines. Such an analyst-intensive, subjective process is not generally appropriate for nuclear explosion monitoring, but the effectiveness of this technique suggests that automated pattern recognition techniques could have a significant impact on monitoring. In this paper we will show how cluster analysis (CA) techniques can be used to automate the waveform recognition problem and compare the performance of different CA methods.

Cluster analysis is the term for a family of techniques for aggregating similar entities into groups or clusters. In this study we compare three different CA techniques: agglomerative hierarchical clustering (represented by dendrograms), Q-mode factor analysis, and ordination. Our data set consists of 651 regional distance events recorded by the New Mexico Institute of Technology network from July 1997 through February 1998. The events are predominantly mining explosions from operations in western New Mexico, as well as southeast Arizona. Because we have full-network recordings for these events, we are able to locate events from within each cluster to tie the clusters to known mining regions. All of the cluster techniques are based on a similarity matrix formed by comparing each entity with every other entity. For this study, we base our measure of similarity on normalized waveform correlations for a single station. The resulting clusters depend strongly on the processing parameters applied to the waveforms (phase windowing, filtering, Hilbert enveloping). Our results suggest that dendrograms of the Hilbert enveloped waveforms produce the most useful results, while factor analysis may prove useful as an auxiliary technique. Ordination produced marginally useful results only after non-linear rescaling of the similarity data, and we do not think it shows promise with this type of data.

None of the CA techniques automatically determine how many clusters are represented in the data. This decision is often made subjectively, but it can be based on the application of clustering criteria. We compared several proposed clustering criteria methods to determine which work the best for our data set, but none performed well in general.

KEY WORDS: cluster analysis, dendrograms, factor analysis, ordination

OBJECTIVE

The objective of this study is to investigate the use of cluster analysis techniques to identify similar seismic events based on waveform correlation. Because of the strong dependence of seismic Green's functions on source mechanism and source location at frequencies typically observed in regional studies (> 1 Hz), a high correlation value indicates both a similar location and a similar source type, and both types of information are of great importance to the nuclear explosion and non-proliferation monitoring community.

Currently, location and identification are accomplished by minimizing the misfit between a set of low-order parameters derived from the waveforms from a network (i.e. arrivals and various measurements associated with them) and the corresponding parameters predicted by Earth models. Because this parametric misfit minimization approach is model-based, it can be used to evaluate events anywhere on the Earth, which is one of the main reasons why it is used so widely. However, the model-based approach does not work well for smaller events which are typically recorded at fewer stations. In these cases, there are often too few observations to locate and identify the events. This is a significant concern if we seek to lower the monitoring magnitude threshold.

For small events, waveform correlation can provide much better results than the model-based approach if an archived set of similar events is available. Experienced seismic network analysts often can identify an event as coming from a given region based on a single waveform. The analysts do this by matching the present waveform with memories of others which were known to have come from that region. This technique works because the seismogram resulting from a given source-to-receiver path is as unique as a fingerprint. Paradoxically, this is exactly the same reason why synthetic waveforms produced from models are not typically used in operational monitoring systems: the resolution of most models is not sufficient to match the source-receiver path and reproduce the fine details of the observed waveforms.

Our study follows the lead of Riviere-Barbier and Grant (1993), who showed how waveform correlation-based dendrograms could be used to identify regional mine blasts recorded by the FINESA array. In that study, the authors carefully determined the optimal signal processing parameters to yield the best clusters and painstakingly ground-truthed their events to tie the clusters to mines. They did not, however, focus much on cluster analysis itself. They chose one of the simplest and most reliable methods, the formation of dendrograms by complete linkage hierarchical clustering, and applied it to their data set. In this study, we provide a more thorough investigation of various cluster analysis techniques and assess their applicability to the event location and identification problem.

We simulate the small event monitoring problem using a set of regional events recorded by the New Mexico Tech Seismic Network (NMTSN). The events were generated by an automatic, grid-based system (Withers et al., 1999) running between July 2, 1997, and February 27, 1998. In all, 651 events were detected and located, but clearly there are problems with the automatically generated catalog (Figure 1, left). First, the grid itself is coarse (10-km spacing) for events outside central New Mexico, so the locations have an inherent limitation. Second, even allowing for the coarseness of the grid, the events do not exhibit the linear features (faults) and point features (mines, volcanoes, etc.) that one would expect. In fact, it is known that the bulk of the events recorded by the NMTSN come from 5 areas: 4 mining regions in western New Mexico and southeastern Arizona, and the region of the magma body near Socorro (Figure 1, right -- see Balch et al., 1997 for information about the Socorro magma body). Thus, much of the scatter in the locations of the events probably reflects gross event mis-location

We tried to classify these events using only the waveforms from a single station of the NMTSN, station CAR. However, because we have full network data for the events, we can ground truth any of them by re-timing the arrivals for the various phases, and re-locating using an appropriate regional travel time model.

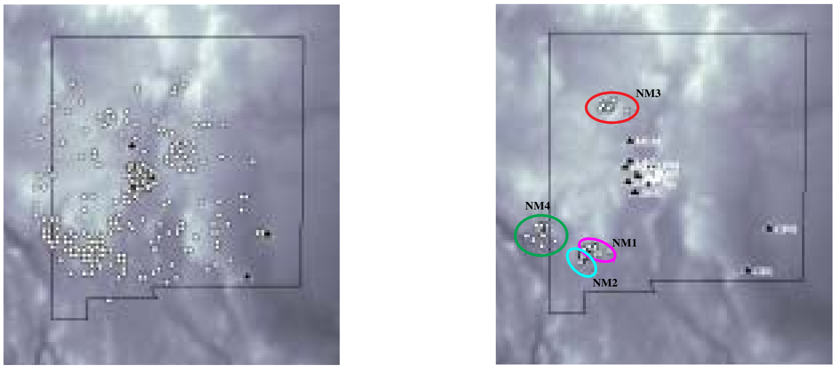

Figure 1. (left) The full 651 events detected and located by a grid-based automatic system running on data from the NMTSN for the period of July 2, 1997, through February 27, 1998. **(right)** The 60-event training subset, the 4 identified mining regions, and the NMTSN (15 stations). In both cases the background image is topography and the outline shows the New Mexico border. Because some of the information is impossible to convey without the use of color, and the printed Proceedings are in black and white as a cost-saving measure, we refer the reader to one of the electronic versions of the Proceedings for interpreting the figures in this paper. The electronic versions can be found at http://www.nemre.nn.doe.gov/review2001 or on the CD ROM version of the Proceedings. Or, readers should feel free to contact the authors.

Many of the global monitoring stations are located in areas at least as seismically active as New Mexico, so it is likely that these stations will detect a large number of regional events that will not be recorded at the other stations (given the typical global monitoring network station spacing). Typically, these events cannot be located and thus they determine the magnitude threshold for the network in the vicinity of the stations. The only way to lower the threshold for the network (other than adding more stations) is to better utilize the information in the waveform from a single station. One way to do this is by cluster analysis using full waveforms.

CLUSTER ANALYSIS

Cluster analysis (CA) is the term for a family of techniques for aggregating similar entities into groups or clusters. Much of the CA methodology comes from the field of numerical taxonomy, which involves trying to group similar organisms to establish evolutionary relationships (e.g. Ludwig and Reynolds, 1988). We compare three different CA techniques: agglomerative hierarchical clustering (represented by dendrograms), Q-mode factor analysis, and ordination. All of the cluster techniques are based on a similarity matrix formed by comparing each entity with every other entity. For this study, we use as our measure of similarity the absolute value of the normalized waveform correlation (i.e. best lagged cross-correlation divided by the autocorrelations).

Hierarchical Clustering (Dendrograms)

Hierarchical clustering is an iterative process in which similar entities are repeatedly joined to form a tree (e.g. Davis, 1986). One begins by selecting the most similar pair of entities, in our case waveforms, from the similarity matrix and joining these together on the dendrogram. All of the similarities between each of these entities and all of the other entities in the similarity matrix must now be replaced by a similarity between the newly formed group and all of the other entities (e.g. one can average the individual similarities). Doing this will shrink the dimension of the matrix by 1. Once this has been accomplished, the process is repeated until the matrix is reduced to a dimension 1, and an entire dendrogram has been built. The only trick in hierarchical clustering is in the method chosen for calculating similarities between a newly formed group and the remaining entities. There are many methods to calculate the similarity between an entity *k* and a group, *ij*, formed by the fusion of entities *i* and *j*, but all of the ones we are tested can be represented with the formula (Lance and Williams, 1967):

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|$$

Where d_{ij} is the distance (i.e. 1 - similarity) between groups *i* and *j*. By choosing different values for the weights α, β, γ this formula can represent any of the standard hierarchical formulas. For example, for single linkage $\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0$ and $\gamma = -\frac{1}{2}$. For complete linkage, the only difference is that $\gamma = \frac{1}{2}$.

Q-mode Factor Analysis

Q-mode factor analysis is essentially a principal components technique. Again, one begins with the similarity matrix. In this case, however, we simply calculate the eigenvectors of this matrix and then plot these against each other, scaled by the eigenvalues, to identify the groups. As would be expected, the eigenvectors corresponding to the highest eigenvalues will be the most significant, but it turns out that the first eigenvector contains the centroid position and thus should be ignored (Davis, 1986), so we plot the second and the third eigenvectors.

The Q-mode refers to the fact that we solve for the eigenvectors of the Q matrix, which is defined to be the matrix of data vectors dotted with each other. The alternative is R-mode in which we would decompose the R matrix, which is defined to be the matrix of variable vectors dotted with each other. R-mode factor analysis is used to establish correlations between variables (as opposed to between entities). This is not the goal of our study, so we do not investigate R-mode factor analysis.

Ordination

The final CA process we investigate is ordination. This technique has proved very successful in mining text-based data sets (e.g. newspaper and magazine articles) and has had only a limited number of applications to other types of data. However, ordination is ultimately based on the same sort of similarity matrix, so it can readily be applied to our data set. To ordinate a set of entities, the entities are placed in a plane such that their relative positions agree as closely as possible with the similarity information in the similarity matrix. This is done through a computationally intensive, iterative Monte Carlo process in which the entities are assigned initial positions and then moved randomly to see if their new positions agree better with the similarity matrix information. Interested readers should refer to Wylie et al. (2000) for more information.

RESEARCH ACCOMPLISHED

Preliminary Data Analysis

As stated above, the NMTSN events are expected to be dominated by 5 sources: 4 mining regions and the local events associated with the magma body near Socorro. The mining events are typically much more common than the magma body events. To create a training set to test our cluster analysis techniques, we had an experienced NMTSN analyst scan three months of our data (from August through October) to identify representative events known to come from the mining regions (60 events in all). These events were then hand-picked and re-located using the one-dimensional regional travel-time model that New Mexico Tech has developed for the area. The locations of these events are shown in Figure 1 (left). Hereafter we will refer to this as the training set. Though there is still some scatter in the locations of these events, the 4 mining regions are clear. Because these events lie outside the NMTSN, significant further improvement of these locations is unlikely without adding data from other stations in the region.

Representative waveforms for NMTSN station CAR for each mining region are shown in Figure 2. The distance to NM1 is ~190 km, to NM2 is ~ 220 km, to NM3 is ~190 km, and to NM4 is ~260 km. The NMTSN analysts try to pick Pn, Pg, Sn, and Sg for these events. For the Socorro magma body events, they also try to pick magma body converted phases (PzP and SzP). See Withers et al. (1999) for examples of waveforms and discussions of the regional phases.

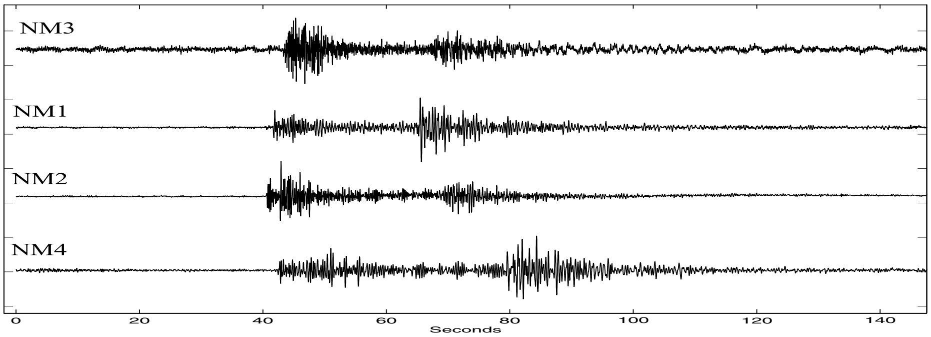

Figure 2. Typical waveforms for station CAR for each of the 4 mining regions.

The data are sampled at 100 sps, but we find that the usable frequency content is generally below 15 Hz. Lowering the sample rate improves the speed of the waveform correlations, so we down-sampled the data to 25 sps (an anti-alias filter was applied). All of our clustering was done using the waveforms for station CAR, which was chosen simply because it is a typical station near the center of the network.

Clustering of the Training Set

To assess the effectiveness of the CA techniques, we began by clustering the training set, with the expectation that we should identify the 4 mining regions. Based on trial and error, we determined that a 55-second-duration waveform window, with a 5-second lead time relative to the theoretical Pn arrival time were effective parameters for this data. We clustered this data three times: first, with no signal processing; second, with a 2- to 10-Hz, bandpass filter; and finally using a Hilbert envelope on the waveforms.

Dendrograms

We began with unfiltered waveforms, and tried building dendrograms using several of the cluster methods (flexible, median, group mean, centroid, single linkage, complete linkage) which are described by the Lance and Williams equation. To summarize the results, we found that most of the methods were equally effective in clustering the waveforms from the 4 mining regions. However, the flexible method consistently produced the clearest dendrograms. This result contradicts Jardine and Sibson (1971), who showed that the single linkage method is the only method that satisfies a set of necessary simple conditions (continuity, minimum distortion, etc.). However, we are in good company; many other researchers have found that while single linkage may be mathematically superior, it does not necessarily produce the best solutions (e.g. Williams et al, 1971; Gower, 1988). Our experience is nicely summed up by Davis (1986):

> *Most researchers who use clustering methods experiment with a variety of similarity measures and clustering techniques, and they choose the combination that yields the most satisfactory results with their data.*

The flexible method produced the best results for us, so we use it as our default choice for producing dendrograms. The flexible method properly refers to any choice of weights such that $\alpha_i = \alpha_j$ and $\alpha_i + \alpha_j + \beta = 1$, but we mean the particular set of weights suggested by Ludwig and Reynolds (1988), i.e. $\alpha_i = \alpha_j = 0.625$ and $\beta = -0.25$. This method can join groups at similarity values less than 0, as can be seen in some of our dendrograms. Such similarity values are artifacts of the method and are not physically meaningful, but they do not compromise the usefulness of the flexible method for identifying groups, which is our goal.

Deciding where to cut the stems of a dendrogram is a subjective process. In general, one should seek to set the level at the highest possible similarity value which still yields distinct groups. For the unfiltered data, 4 obvious groups were apparent and an appropriate threshold was easy to set. Conveniently, the 4 groups turn out to represent the 4 mining regions very well: the NM1 group has 3 mis-identified (2 NM4, 1 NM2), 1 in NM3; the NM2 group has 0 mis-identified, 1 in NM1; the NM3 group has 1 mis-identified (NM1), 0 in other groups; and the NM4 has 0 mis-identified, 2 in NM1. In all, we have only 4 errors out of 60 events identified, or a 7% error rate.

If the misidentifications are due to poor signal-to-noise ratios (SNR s), then an obvious way to improve the result is to apply a filter to improve the SNR. Based on an examination of spectrograms of the waveforms, we determined that the best SNR is in the 2- to 10-Hz band. Thus we applied a 2- to 10-Hz band pass filter to all of the waveforms and re-clustered. With the band pass filtering, the two NM4 events which were in the NM1 group move to the NM4 group, improving our identification results to 58 out of 60, or a 3% error rate.

We next tried applying a Hilbert envelope to the waveforms, the method recommended by Riviere-Barbier and Grant (1993). This is a form of non-linear low-pass filtering, so the waveforms are in effect smoothed. This leads to better correlation within the mining region clusters. The dendrogram and corresponding enveloped waveforms are shown in Figure 3. Each waveform is lagged for maximum correlation with the one directly above it. The results are the same as for the bandpass case (58 out of 60 identified correctly), but the tightness of the clusters and the separation between the clusters are better. Further, because the enveloping smooths the high frequencies, we can down-sample further before applying the enveloping (we down-sampled to 10 sps), which can dramatically speed up the processing time for large data sets.

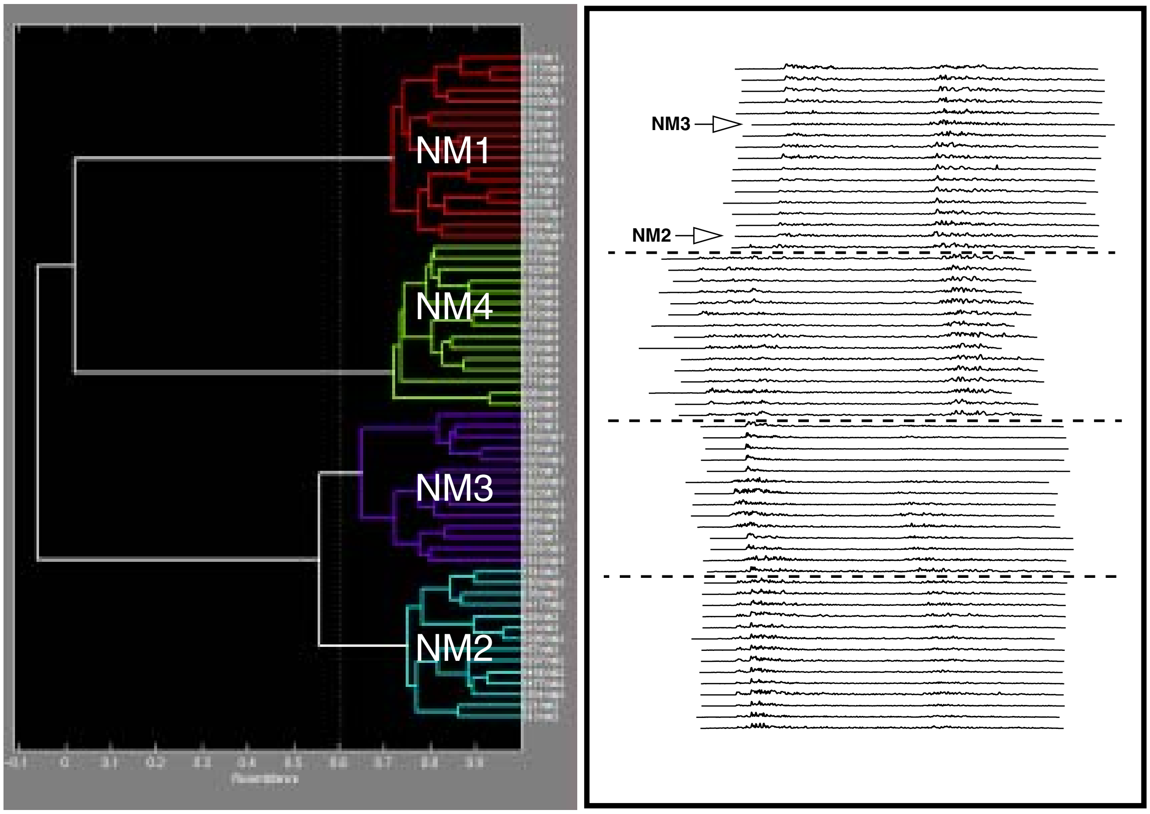

Figure 3. Dendrogram and corresponding waveforms for Hilbert enveloped training set. NM1 group has 2 mis-identified (1 NM2, 1 NM3), 0 in other groups; NM2 has 0 mis-identified, 1 in NM1; NM3 has 0 mis-identified, 1 in NM1; NM4 has 0 mis-identified, 0 in other groups.

Factor Analysis

Next, we applied factor analysis to the training set. For brevity, we show only the results for the Hilbert enveloped training set (Figure 4, left). We color the entities to identify the corresponding clusters in the dendrogram, but it is important to note that this information is not derived from the factor analysis. With the colors, it is apparent that the locations of the different mining region events are segregated on the factor analysis plot, but only the NM4 cluster (green) is well-isolated. This result is typical of many applications of factor analysis that we tried, and it illustrates two key problems with the method. First, as the clusters are not necessarily simple geometric shapes, one would have to use some sort of generalized shape to lasso the entities within them. Secondly, unless the clusters are well-separated, and they generally are not, assigning entities in overlapping regions may be difficult. Note that these problems do not occur with the dendrogram. Admittedly, choosing the threshold at which to cut the tree can be difficult, but once it is done, grouping the entities is trivial.

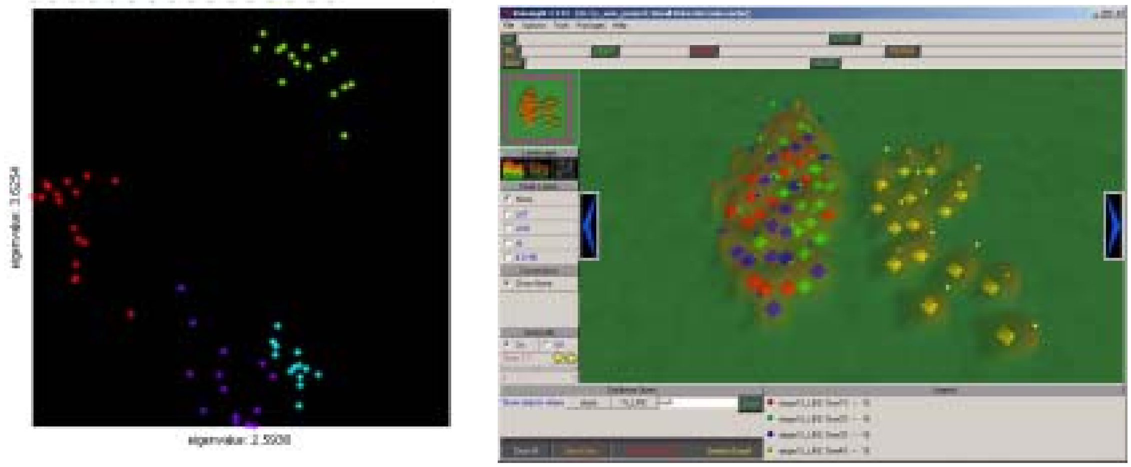

Figure 4. (left) Factor analysis for the Hilbert enveloped training set. Color coding corresponds to the dendrogram in Figure 4, i.e. NM1 = red, NM2 = cyan, NM3 = purple, NM4 = green. **(right)** VxInsight ordination plot for the Hilbert enveloped training set. NM1 = red, NM2 = green, NM3 = blue, NM4 = gold.

Based on this result and others from our analysis of the training set, we conclude that factor analysis may be useful as an auxiliary tool to verify the clustering derived from a dendrogram, but that for identifying similar waveforms it is not by itself a sufficient method.

<u>Ordination</u>

Again, we show only the results for the Hilbert enveloped case for brevity. In this case, we use the Sandia National Laboratories-developed VxInsight software to make a landscape of the ordination results. This is in essence a 2-D histogram. Ordination of the similarity matrix of the 60-event training set led to a plot with no separation of the mining regions at all. Given the high quality of this training set, and the impressive results provided by the dendrograms, this result was surprising. By investigating the software using some synthetic similarity matrices, we were able to establish that the range of similarities for the Hilbert enveloped test set was too small. To improve the ordination performance, we then tried to rescale the similarities in the matrix. A simple linear scaling (i.e. min gets 0, max gets 1) did not improve the results at all. Application of a non-linear scaling using a simple S curve shape did, however, lead to an improved result (Figure 4, right). The topology is similar to that for factor analysis, though the groups are even less distinct. The NM4 group is well separated, while the other groups are not.

Overall, we did not find that ordination worked well for identifying the mining regions in the training set. At best it seems to yield results that are similar to those derived from factor analysis, but this is only with some additional data processing. Without the additional processing, ordination did not separate the mining regions at all.

Clustering of the Full Set

The full set of 651 events includes the 60-event training set analyzed in the previous section. This is advantageous in that we can track where these events end up in the clusters produced for the full event clustering and use this information to make inferences about the clusters. It is also representative of a real monitoring scenario where an analyst would have a set of identified reference events which would be used to identify new events.

The dendrogram and corresponding color coded map for the 651-event set with Hilbert envelope method applied are shown in Figure 5.

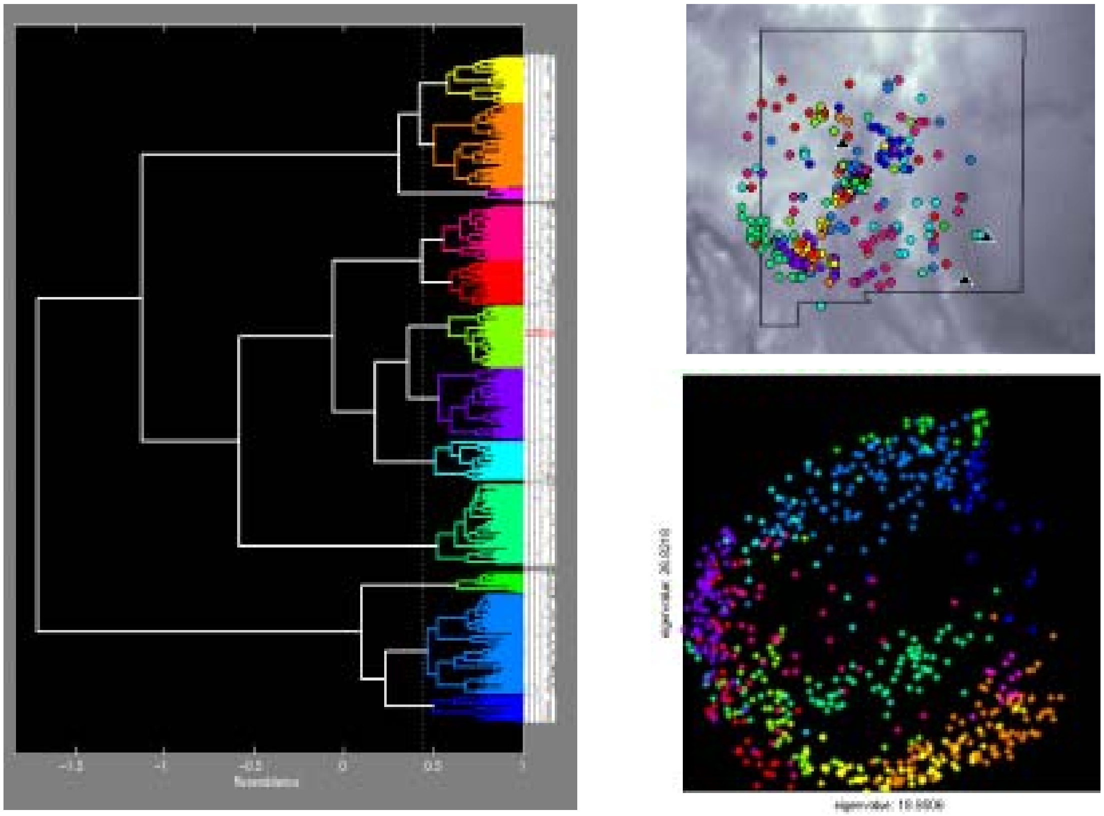

Figure 5. Dendrogram, color coded event map, and factor analysis plot for Hilbert enveloped full data set (651 events).

For the discussion which follows, we will refer to the colors of the dendrogram groupings as (from top to bottom): yellow, orange, pink, rose, red, spring green1, purple, cyan, sea green, spring green2, light blue, dark blue.

In this case, the decision of where to set the threshold to define the groups was less clear than for the training set, so we investigated several numerical methods. Cophenetic correlation refers to the correlation between the actual entity correlations in the similarity matrix and the predicted values from the dendrogram. Ideally, these would correlate perfectly, but with real data sets, they seldom do and the cophenetic correlation tends to decrease as the dendrogram is built. Sneath and Sokal (1973) suggested that a large drop in cophenetic correlation as a new group is formed might indicate a good place to set the threshold, but we seldom found this to be true. Other methods involve tracking the within-group and between-group variances as the dendrogram is built (Everitt, 1993). In the CA lingo, the W matrix contains all of the within-group variance information, and the trace of W gives the overall sum of the within-group variances. Theoretically, the minimum of trace(W) should occur at the proper threshold. Similarly, the B matrix contains all of the between-group variances and the trace of B gives the overall between-group variance. Thus, the maximum of trace(B) should occur at the proper threshold. We made plots of trace(W), trace(B), and trace(B)/trace(W) for many of the dendrograms formed for this study (training and full sets), but did not find the results useful. Ultimately, the threshold used for Figure 5 was determined by eye, with the goal of placing the training set events in separate groups. We were able to accomplish this by setting the threshold at 0.44. All NM4 test events are in the sea green group, all NM2 test events are in the purple group, all NM3 events are in the spring green1 group, and the NM1 events spread through the yellow (3), orange (7), and pink (3) subgroups of a higher order group. The scatter of the corresponding colored points on the map as compared to those in Figure 1 give some idea of the mislocations for each of the mining regions.

Even when the locations of the events are poorly known, sometimes the timing of the events can still offer important information which can be used to identify the events, particularly when combined with the groups identified by CA. As an example of this, we show histograms by day of week and hour of data for the clustering shown in Figure 6.

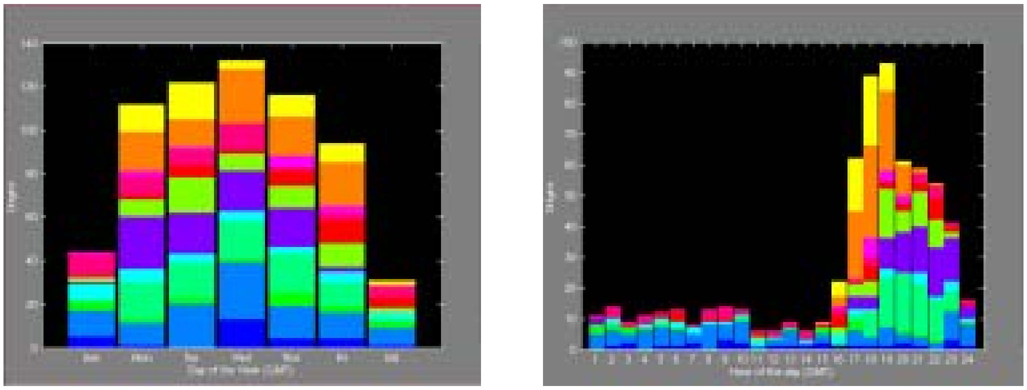

Figure 6. Histograms for Hilbert enveloped full data set. **(left)** By day of the week. **(right)** By hour of the day in GMT (subtract 6 hours to get local time). The colors correspond to the groups identified in Figure 5.

We have found these types of diagrams to be useful in separating man-made and naturally occurring events. A careful examination of Figure 6 shows that the NM mining region groups occur during the work-week and the local work-day time only, indicative of man-made events. If it is known that the data set under analysis contains groups of mining events, these types of histograms could be used to set the thresholds to cut the dendrogram.

Examining the colored bars for the remaining groups, we first note that all show no work-week or work-day dependence, suggesting that they are dominated by earthquakes. The spring green2 group are all are from within the network, and may be Socorro magma body events. The rose, red, cyan, and blue groups all plot diffusely throughout the state and each group contains a variety of waveforms, so it is likely that each group does contain events from all over the state.

We also applied factor analysis to the full event set (Figure 5), and found the results to be consistent with those for the training set. Again, the factor analysis plot of the eigenvectors corresponding to the second and third greatest eigenvalues does tend to put each of the groups in different areas, but the clusters are not well-separated, and there are many areas of overlap between the groups. Using this plot, it would be difficult to identify the entities within each cluster, though the plot might be useful for establishing how distinct the dendrogram clusters are (e.g. calculating distances between the centroids).

We might be able to improve this result by using the information in some of the other eigenvectors, and we are currently pursuing this idea. There are as many eigenvectors as there are events in the set being clustered (651 for our full event set). We found that other eigenvector combinations are more effective for identifying certain clusters, but that no single pair works best for all clusters. One can easily add another dimension and make a 3-D plot to evaluate eigenvector triplets, but we found these plots to be difficult to interpret.

CONCLUSIONS AND RECOMMENDATIONS

In this study, we have extended the waveform correlation-based cluster analysis (CA) work of Riviere-Barbier and Grant (1993), which used only complete linkage dendrograms, by evaluating several additional CA techniques to identify clusters of similar waveforms. We examined several alternative agglomerative hierarchical clustering methods (i.e. dendrogram producing methods), as well as factor analysis and ordination. We used a data set of 651 regional events recorded by the New Mexico Tech Seismic Network (NMTSN), from which we extracted a training set of 60 well-located, ground-truthed events coming from the 4 known mining regions in western New Mexico and southeastern Arizona.

Our results suggest that dendrograms seem to be the most effective cluster analysis technique, but that the complete linkage method used by Riviere-Barbier and Grant is not the best choice. We found the flexible method of Ludwig and Reynolds (1988) produces dendrograms which are at least as effective in grouping like waveforms and which make the groupings more apparent. Choosing the threshold level to identify the groups for the dendrograms remains problematic. We investigated a variety of numerical techniques and found none of them to be generally effective. Thus, we ultimately were forced to subjectively choose thresholds which looked reasonable on the dendrograms and isolated the mining regions.

The other CA techniques we tried were less effective. Ordination failed outright at first, but we were able to make it produce marginally useful results by a non-linear scaling of the correlation information. Factor analysis provided better results than ordination without any need for the rescaling, but still does not work well for assigning the individual events to clusters. We do think, however, that factor analysis might prove useful as an auxiliary tool to assess the quality of the groupings identified with the dendrograms, and intend to further investigate this possible use.

ACKNOWLEDGEMENTS

We thank Helen Koller for her assistance in using the VxInsight software to produce the ordination plots and Jeff Hampton and Neill Symons for their thorough reviews of the manuscript.

REFERENCES

Balch, B., Hartse, H., Sanford, A., and K. Lin (1997). A new map of the geographic extent of the Socorro midcrustal magma body, *Bull. Seismol. Soc. Amer.*, 87, 174-182.

Davis, J. C. (1986). Statistics and data analysis in geology, J. Wiley & Sons, New York.

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

Gower, J. C. (1988). Classification, geometry and data analysis, In Classification and Related Methods of Data Analysis (H. H. Bock, ed), Elsevier, North-Holland.

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

Ludwig, J. A. and J. F. Reynolds (1988). Statistical ecology: a primer on methods and computing, J. Wiley & Sons, New York.

Riviere-Barbier, F. and L. T. Grant (1993). Identification and location of closely spaced mining events, *Bull. Seismol. Soc. Amer.*, 83, 1527-1546.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical taxonomy, Freeman, San Francisco.

Williams, W. T., Lance, G. N., Dale, M. B., and H. T. Clifford (1971). Controversy concerning the criteria for taxomonometric strategies, *J. Comp.*, 14, 162-165.

Withers, M., R. Aster, and C. Young (1999). An automated local and regional seismic event detection and location system using waveform correlation, *Bull. Seismol. Soc. Amer.*, 89, 657-669.

Wylie, B. N., Boyack, K. W., Davidson, G. S., and D. K. Johnson (2000). Visualization of information spaces with VxInsight, Sandia Labs Report #SAND2000-3100.