Hierarchical Clustering of Historic Sound Speed Profiles

Roger Meredith, Bryan Mensi
Naval Oceanographic Office
Stennis Space Center, MS 39522

Marlin Gendron
Naval Research Laboratory
Stennis Space Center, MS 39529

***Abstract*-Categorizing historical sound speed profiles stems from the desire to map spatial and temporal variability. Sound speed variability correlates with environmental phenomena and is an indicator of changes in sonar performance. Sound speed maps will assist in planning more efficient environmental survey operations such as conductivity, temperature, depth (CTD) collections. Intrinsic profile attributes estimated directly from the profile, such as the mean, variance, and derivative values, are used in the clustering process separately or in conjunction with extrinsic attributes such as location and ocean floor depth. Examples are used to demonstrate that the underlying spatial boundaries of the cluster groupings identify regions where sound speed profiles are consistent. The process is easily tailored to multiple clustering based upon the spatial and temporal scales of interest or on generic properties of the individual profile. The sensitivity of the cluster boundaries and group statistics to the addition of new profiles, or to changes in temporal and spatial scale, defines a new environmental characterization.**

I. Introduction

Sound speed varies strongly with temperature, depth, and salinity. Ocean temperature and salinity vary under many forcing functions including depth, tides, wind stress, waves, solar heating, atmospheric pressure, current, vorticity, and Earth's rotation. Thus, sound speed is constantly evolving over large spatial and temporal scales spanning turbulent scales (less than a second and smaller than a meter) to synoptic scales (days and months over distances of tens of thousands of meters). These large scales of variability make unconditional categorization improbable.

Classes or categories are used to describe similarities and differences in a way that helps understanding and describes relationships. Categorizing sound speed profiles stems from the desire to map sound speed spatial variability. The principal objective is to identify geographical regions or provinces where sound speed profiles are consistent over some defined set of attributes and to quantify sound speed variability within each group and between each group. In addition, the geographic location, extent, and separation of cluster groups would be interesting in themselves and provide new descriptions of variability that comprise the assimilated results of multiple forcing functions. Ideally these categories or cluster groups would be solely based on profile attributes; however, the dependence of sound speed profiles on so many factors and environmental effects related to location and time implies that a combination of intrinsic profile characteristics and extrinsic spatial and temporal indicators are useful for categorizing profile data. Sound speed profile clustering can provide new insight into historical sound speed variability. Maps of sound speed changes identify potential sonar performance changes. Sound speed cluster maps can be easily tailored for a wide range of temporal scales and spatial scales.

The U.S. Navy maintains a temperature–salinity–depth database entitled the Master Oceanographic Observation Data Set (MOODS), which contains more than 8 million profiles from around the world spanning 125 years. The MOODS database contains public domain profiles derived from Navy and from the National Oceanographic Data Center (NODC), a repository and dissemination facility for worldwide oceanographic data under the auspices of the National Oceanic and Atmospheric Administration (NOAA) [1]. As used herein, each profile in the MOODS database is considered to be of equal value with all other profiles. Clustering does not average, synthesize, filter, or fuse profiles, but it does identify a group of profiles that would be used for such analysis.

II. Background

In everyday experience, the spatial distance from one point to another in a three-dimensional plane has two components, one for each plane. The total distance between the points is given by an accumulation of the individual distances. This concept of distance is easily extended into n-dimensions, one difference for each dimension, followed by accumulation. In clustering, the dimensionality is determined by the number of attributes or components that are used to characterize the profile. Thus, a profile of sound speed as a function of depth is clustered based on the n-dimensional vector of attributes derived from the profile. Distance will be the metric used to assign profiles to a cluster group and assess the quality of the groupings. Important distances are from one profile to the center of a cluster, the distance between clusters, and the distance between one profile and another.

Hierarchical methods form a multilevel hierarchy, where clusters at a lower level (leaf nodes) are joined to form higher level clusters (branches). This hierarchy forms a dendrogram, a tree-like structure, that shows the links in terms of distance and provides flexibility in choosing the number of clusters

best suited for that hierarchy. The process begins by computing the distance from each profile to every other profile using the vector of attributes. The two profiles with the smallest distance are linked first to form a single cluster. The hierarchy is grown by successive linking based on the smallest separation until all clusters are linked into a single cluster (root). The separation distances reveal patterns that identify natural groupings that can be used to determine the appropriate number of clusters and identify profiles that are greatly different from the other profiles. MATLAB© [2] was used to perform the clustering.* The cluster results are also dependent upon the attributes used to compute distances. The choice of attributes is somewhat subjective. The cophenetic correlation coefficient (CCC) is the metric for the strength of the separation between clusters for the hierarchical method. The cophenetic correlation for a dendrogram is the linear correlation coefficient between the cophenetic distances obtained from the tree and the original distances used to construct the tree. The cophenetic distance is the height of the link that joins two clusters and is used to represent how well dendrograms model the variability in observations [2]. This value should be close to 1 for a high-quality solution.

Mandelberg and Frizzell-Makowski used hierarchical clustering to province sound speed profiles obtained from the Naval Oceanographic Office's General Digital Environmental Model (GDEM) database in 2000 [3]. Their goal was to reduce the number of profiles needed for acoustic propagation modeling in the North Atlantic and North Pacific oceans. They correlated computed provinces with major oceanographic features and circulations. Similar large-scale correlation is used here to assess cluster results.

III. Attributes for Clustering

The multi-dimensionality used to cluster profiles results in a wide range of numerical values or a wide range of magnitude scales. Normalizing often improves the cluster results, and many normalizing methods are available. Everitt et al. [4] recommend normalizing by the range of the values based on anecdotal reports of multiple users. Our limited testing with various normalizations concurred that the range normalization yields the strongest clustering. A partial list of intrinsic attributes used in clustering includes

- Estimates of the mixed layer depth
- Estimates of surface speed
- Relative changes in sound speed
- Integrated values of sound speed
- Central moments estimated from the profile
- Autocorrelation estimates from the profile

First and second derivatives of the profile are estimated using smoothed forward and backward divided differences. From these derivatives, additional attributes are obtained including scintillation and the depth at which the maximum derivative occurs.

Adaptive Profile Clustering

One key to robust clustering is principal component analysis (PCA) [4]. PCA is a method for determining which attributes contain the greatest amount of statistically independent information. By choosing attributes that contribute more to the principal components, clustering is optimized for that set of profile data. The other key is multiple clustering and ranking results. A process for adaptive profile clustering in evolving environments (APCEE) was developed based on PCA and multiple clustering using different combinations of attributes (trials). The amount of time required to complete the process and generate maps is ill-defined due to the size of the profile data set being clustered. The overall process is generalized in Fig. 1. In this paper, five trials, some with and some without PCA, are used to determine the variability in the clustering results. These five trials are ranked to determine the best result. The trials are identified by subsets of attributes as follows:

Trial 1: all attributes (no PCA)
Trial 2: intrinsic attributes only (no PCA)
Trial 3: all attributes are used in PCA
Trial 4: only intrinsic attributes used in PCA
Trial 5: a fixed subset of attributes used in PCA

An increase in the value of the CCC indicates greater overall separation between clusters. The increase is meaningful only if it occurs above the linkage threshold value being used to cluster. Thus, the trial with the highest CCC value is normally the better choice. However, several trials may give the same general result. The choice for the number of clusters is less obvious. A process based on the nulls in the derivative of the link distances was developed to determine "natural" thresholds for the number of clusters.

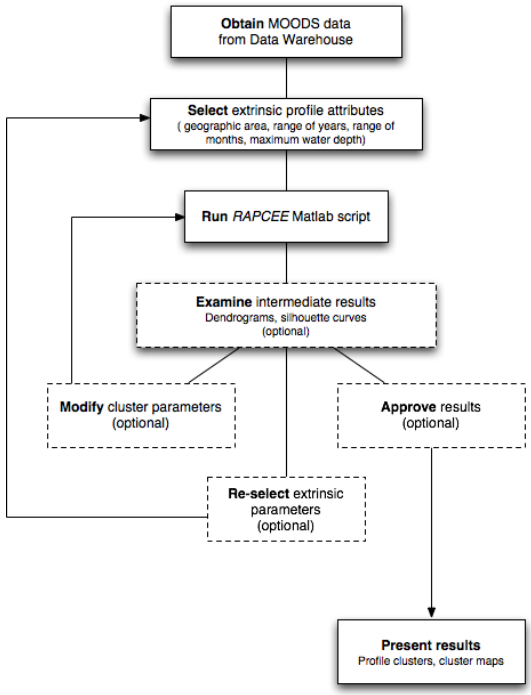

Figure 1. Process outline for clustering sound velocity profiles.

*The inclusion of commerical products does not imply endorsement by the U.S. Navy or NAVOCEANO.

IV. MAY– GULF OF MEXICO EXAMPLE

Data from the MOODS database were extracted under the dissemination heading *Unclassified Public Release*. The 1284 profiles span the upper-half of the Gulf of Mexico for the month of May between the years 1971 and 2006. The clustering results from all trials with CCC values greater than an arbitrary threshold of 0.85 were kept for further examination. For this set of profiles, clustering with Trial 1 and Trial 4 did not yield CCC values above 0.85. Thus, based on the trials for this subset of profiles, metadata play an important role in creating clusters, and PCA analysis improved the strength of the clustering. The CCC value depends only on linkage distances and is independent of the number of clusters. Trial 5 gave the overall highest CCC; however, the span of CCC values was small, less than 10%, indicating that any of these five results could give similar clustering. Two discussions ensue: (1) using the dendrogram to determine an appropriate number of natural clusters within a single trial, and (2) comparing the clusters from different trials for a near-constant number of clusters.

Fig. 2 shows a dendrogram for the trial with the highest CCC value, representing the strongest clustering (the greatest overall distances between linkages). Each profile is represented along the x-axis of the dendrogram at the link distance of zero (y-axis). The separation between any two profiles (or clusters) is the sum of the two cophenetic (or link) distances (one up and one down), where the profiles (or clusters) are joined by a horizontal bar. The larger the sum, the more separated the profiles (or clusters).

In Fig. 2, the dendrogram is replicated four times and color-coded based on different linkage threshold values (the dotted horizontal black lines) that determine the sorting of profiles into cluster groups. The number of resulting clusters is given at the top of graph next to the keyword "mynoc." The first threshold (leftmost) gives 5 clusters; the fourth gives 40 clusters. Due to the size of the plot, not every cluster is easily visible, nor are the individual 1284 profiles that appear at the very bottom of each dendrogram.

The leftmost dendrogram has five cluster groups; however, 79% of the profiles falls into one cluster group (the red one on the right-hand side of the leftmost dendrogram). Approximately 19% of the profiles falls in the second largest group, and the other three groups together represent about 2.5% of the profiles and are too small to be seen clearly at this scale. The three groups of the profiles are not outliers, only profiles separated by larger link distances to the other profiles. These low-density groups may indicate unique variability in the sound speed profiles or the need for more or fewer cluster groups. The links above the threshold are also revealing. The solid horizontal bars indicate the threshold value that will separate the profiles into distinct clusters. Where vertical distances are larger, profiles are more strongly separated. Where vertical links are short, cascading consistently with little vertical separation, profiles are weakly separated. Thus, many deep vertical nulls are a sign of cluster group separation,

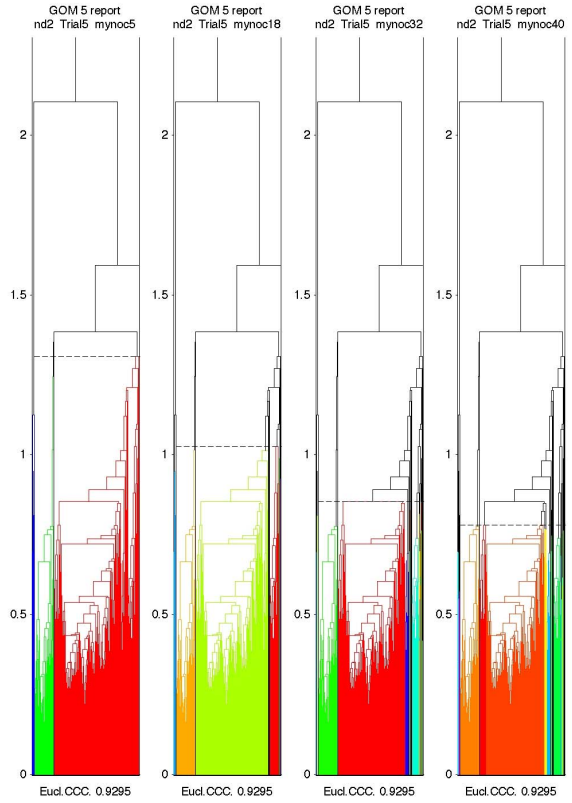

Figure 2. Identical dendrograms color-coded by threshold link values that determine the number of clusters.

and short vertical links indicate a condition of diminishing returns on the optimal number of cluster groups.

The dendrogram provides the physical oceanographer with the visual clues for setting the thresholds based on experience and intuition. The four dendrograms in Fig. 2 are color-coded at lower thresholds (dashed horizontal lines) that separate profiles into larger numbers of cluster groups. Some groups have many profiles and some have few. The threshold value of the fourth dendrogram results in 40 distinct cluster groups; however, 54% of the profiles still fall into one cluster group near the middle. The second largest cluster has about 18% of the profiles, and there are 5 groups with 3 to 10% of the total profiles. Several more groups contain less that 1% of the profiles each. Multiple thresholding reveals that although many groups can be easily assigned, the bulk of the profiles are still assigned to one group. This distribution of profiles among cluster groups tends to support the conviction that larger numbers of cluster groups are not necessary. A higher value (between 20 and 40) is more appropriate for identifying anomalous profiles and allows the analyst to identify extreme profile variance. A lower value (between 10 and 20) is useful for identifying major profile trends and characteristics. Dendrograms provide valuable clues to the upper and lower limits of the natural numbers of clusters that are appropriate to these profiles. Different trials or cluster parameters will provide different dendrograms with different CCC values.

A. Profile Cluster Groupings

Comparisons of dendrograms for different trials will follow plots of the group profiles and a map of cluster groups. Fig. 3 shows the profiles belonging to the four largest cluster groups from the dendrogram thresholded to yield 18 cluster group (mynoc 18 in Fig. 2). Cluster 9 contains approximately 68% of the profiles. Cluster 14 profiles are nearly isovelocity even though speeds range from 1500 to 1535 m/s. A positive result is that although the speeds and depth vary and overlap to some degree, the distinct shapes of the profiles are accurately separated. The profiles in cluster 4 exhibit more curvature and are visibly different from other clusters. The profiles in cluster group 11 are shallow, less than 12 m, and although few, they too are visibly different from the other profile groups. These are encouraging results.

Variability still exists in each cluster group in Fig. 3. This is seen in the sound speed at zero depth and in thermocline structures of cluster 4. The variability within a single cluster group can be increased by raising the link threshold (reducing the total number of clusters), or it can be decreased by adding more cluster groups (reducing the link threshold). Thus, if the analyst has a priori requirements for sound speed variability, dendrogram thresholding can accommodate such constraints.

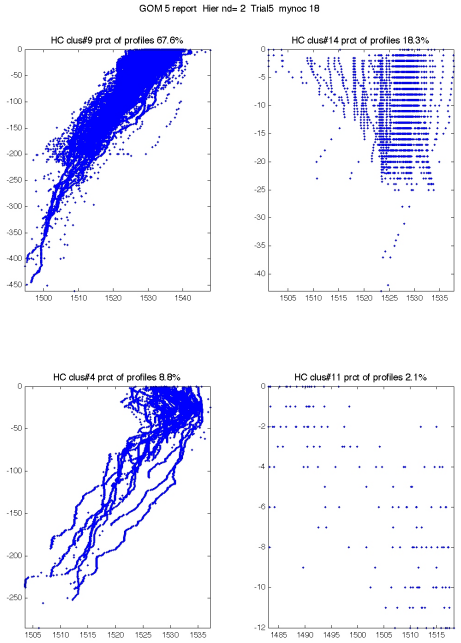

Figure 3. Cluster groups from the second dendrogram in Fig. 2.

B. Sound Velocity Profile Maps

Fig. 4 is a profile map color-coded by cluster group (top)and a bathymetry map from NOAA's web site (bottom) [5]. Only clusters containing more than 1% of the total number of profiles are included. For this example, that cluster group density restriction limits the map to the four clusters shown in Fig. 3. The map is intuitively reasonable: the profiles in Mobile Bay form the smallest group (cluster 11), and the profiles near the continental shelf edge form the largest group (cluster 9). The second largest cluster is scattered about the region in longitude, but always closer to the coastline. A distinct cluster group or province appears off the coast of Texas. This map has intuitive sensibility in regard to the bathymetry, the differing bottom types in Florida and Texas, and the different profiles typical of near shore and offshore due to temperature and salinity. Similar banding is often seen in satellite sea surface temperature imagery.

Determining the number of cluster groups depends somewhat on the rationale for clustering, the number of profiles, the geospatial scale, and time extent of the data being clustered. Larger linkage threshold values yield the smaller numbers of cluster groups. Fig. 4 (top) and Fig. 5 show three clear baseline groups: one for large bays, one for inshore, and one near the edge of the continental shelf. Dendrogram thresholding provides flexibility in determining the level of profile variability and accommodates clustering in relation to spatial or temporal extent or mission objectives like counter-detection or active searching. However, this flexibility is limited by the number of profiles available for clustering. In Fig. 5, the number of high-density cluster groups has been reduced from three to two and fails to resolve variability along the Texas and Florida coasts.

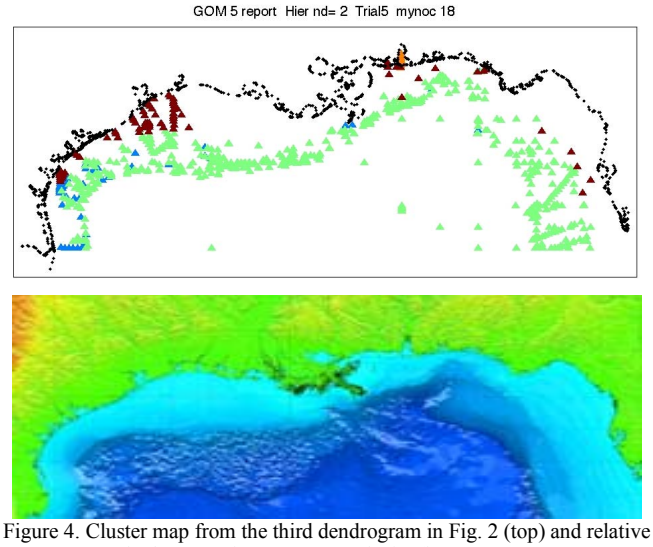

Figure 4. Cluster map from the third dendrogram in Fig. 2 (top) and relative bathymetry from NOAA web site (bottom) [5].

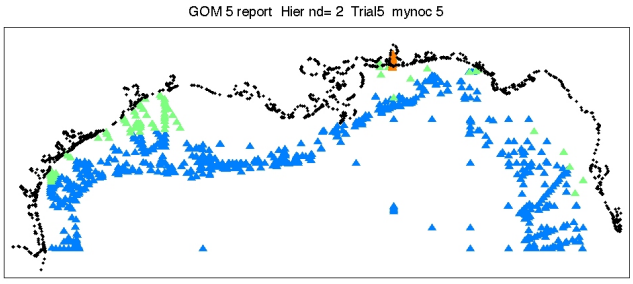

Figure 5. Cluster map from the second dendrogram in Fig. 2.

V. Multiple Trial Comparisons

Results using different profile attributes are compared and delineated by a trial number. Each dendrogram was thresholded to yield approximately 18 groups to match Fig. 4. To visibly compare the dendrograms shown in Fig. 6, start at the top, where all the profile groups form one cluster, and compare how groups separate while moving down the vertical link scale (y-axis). The color codes are assigned based on cluster group number and not on cluster group density. For each of the dendrograms, individual profiles are not in the same horizontal order and the linkages to other profiles are not the same. Differences in symmetry and cascading links above the threshold (dashed horizontal line) are important effects from trials using differing attributes. Differences below the threshold, the color-coded structures, show less change. A color-coded structure in one trial often shows similar symmetry and cascading to a different colored structure in another trial. Comparing the linkage between any two profiles cannot be done at this plotting scale; however, that information is available in the links database generated by the clustering algorithm. Higher CCC values require lower threshold values, which creates more cluster groups. If one were to keep threshold values relatively constant, for example around 1.2, then the number of clusters for each of the dendrogram would vary (~8, ~41, and ~28 in left-to-right order for Fig. 6). Thus, profile attributes used for clustering significantly affect the dendrogram linkage structure and the order of the cluster groups. Each dendrogram arranges the individual profiles in a unique horizontal order; however, substructures in one dendrogram may be similar to substructures in another. In Fig. 6 three major categories arise for each dendrogram. For the sake of brevity, the comparisons of the profiles in each individual cluster group and cluster maps are omitted. The four largest cluster groups from each dendrogram have different percentages of the profiles, given in Table I, but the profile shapes are similar to those in Fig. 3.

TABLE I
Comparison of Profile Densities for the Four Highest Populated Cluster Groups from Three Different Trials for May Profiles.

Trial	1st highest % of profiles	2nd highest % of profiles	3rd highest % of profiles	4th highest % of profiles
2 5	68%	18%	9%	2%
2 3	55%	38%	2%	2%
3 2	57%	31%	6%	2%

The cluster maps are not shown, but are similar to Fig. 4. All three trials cluster the Mobile Bay profiles uniquely. All three trials delineate the inshore profiles from the shelf-edge profiles. The profiles nearest the shelf edge are nearly identical for all three trials, as are the inshore profiles. Differences among trials occur primarily in the boundary of the near-shore and shelf-edge profiles. The consistency in the inshore and shelf-edge cluster groups by multiple trials

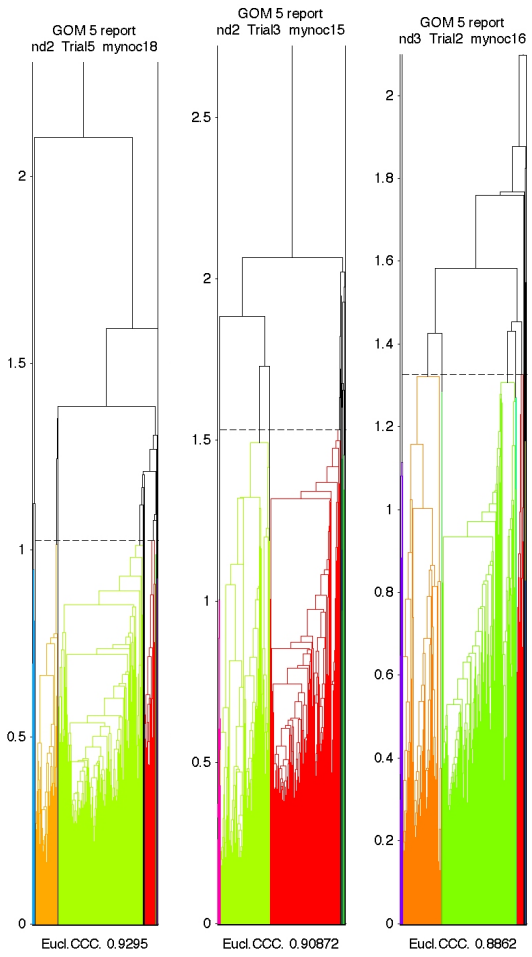

Figure 6. Dendrograms from three different cluster trials.

indicates that sound speed profile clustering is reliable and consistent.

VI. November– Multiple Trial Comparisons

Cluster maps can reveal sound velocity changes over varying time scales. This section examines cluster results for sound speed profiles for the month of November. Color codes identify cluster group number for a single trial. The color palette may match other trials, but no correspondence exists between cluster groups from different trials. The trial numbers correspond to those previously defined. The top map (thresholded at eight cluster groups) shown in Fig. 7 is very similar to the May maps (Figs. 4 and 5). The Mobile Bay data are clustered separately, and the inshore profile group is separated from the shelf-edge group. The same general trends are apparent in the lower map (20 cluster groups). As expected, with more clusters comes more variability, especially near the shelf edge. It also shows more sound speed variability for the Florida coast than previous examples with fewer clusters, but similar variability on the Texas coast.

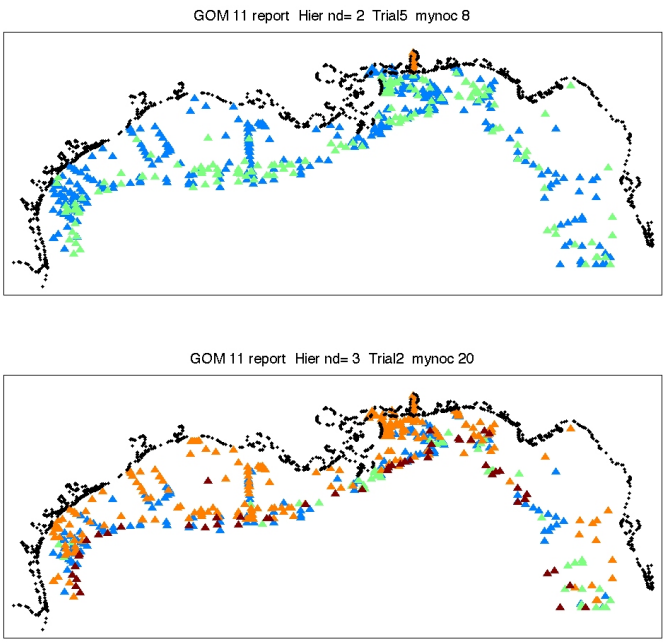

Figure 7. November cluster maps for multiple trials. The top map is thresholded to eight cluster groups; the bottom map to 20 groups. (Compare with Figs. 4 and 5.)

The cluster group profiles for November highlight one utility of sound speed clustering. The profile groups in Fig. 8 exhibit large significant differences compared to the May profiles (Fig. 3). The May profiles are more linear with depth, and groups have varying slopes, including the isovelocity. November profiles all show more curvature. Group 2 in Fig. 8 is somewhat misleading because shallow profiles are combined with deeper profiles that shape a stronger appearance of concavity near the surface than may actually exist. In contrast to the May profiles, there is also a lack of an isovelocity profile group. This finding may indicate that eight cluster groups may be insufficient.

The lower map in Fig. 7 stems from a separate trial with slightly different attributes and a different threshold. The four largest profile groups from the lower map in Fig. 7 are presented in Fig. 9 and provide a different segregation and order of profile groups than in Fig. 8. Both represent natural thresholds based on the dendrogram linkage, and both are correct (they are derived from the different dendrograms). Choosing between dendrograms can be automated simply based on CCC value. A physical oceanographer may choose based on external factors, based on the other types of oceanographic data, based on the distribution of profiles among the four groups, or based on intracluster variability for each group. Cluster groups can be used to provide planners information other than geographical cluster maps. Each group can be further processed to provide an expected sound speed profile for each cluster group, a measure of the variability of sound speed in each cluster group, a range of expected sound speed profiles, and the variability between geospatial cluster groups.

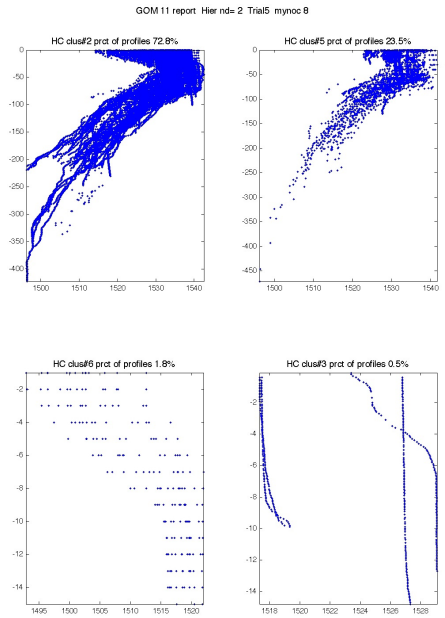

Figure 8. Hierarchical profiles of cluster groups corresponding to upper map in Fig. 7. (Compare with Fig. 3.)

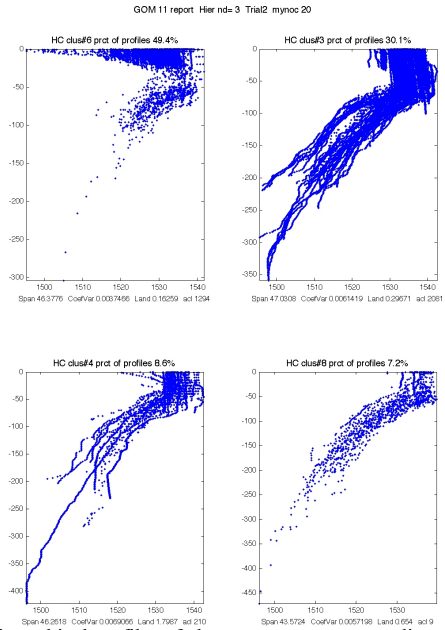

Figure 9. Hierarchical profiles of cluster groups corresponding to lower map in Fig. 7. (Compare with Fig. 3 and Fig. 8.)

VII. Seasonal Cluster Group Comparisons

Until now, clustering took place after profiles were culled for a particular region and time in order to demonstrate spatial clustering and reasonable agreement with general oceanographic features. Culling allows an analyst to select the data most appropriate to mission objectives.

This last section shows that clustering performs well, even when the data set is both spatially and temporally diverse and even if longer term oceanographic processes are involved. Sound speed profiles are clustered as in previous examples but include the entire data set spanning all months and years.

Cluster groups are formed and color-coded independent of month. Cluster groups are then plotted as a function of latitude, longitude, and month. This allows easier comparison of profile changes during a one-year cycle. Other time cycles are easily accommodated. The dendrogram threshold yielded 17 total cluster groups, as shown in Fig. 10, resulting in 11 groups with significant profile density. Fig. 11 shows the cluster map as a function of month. Each color is a profile group so that changes as a function of month and location are readily visible. The Texas coast shows more temporal variability than does the Florida coast. June seems to be the most stable month.

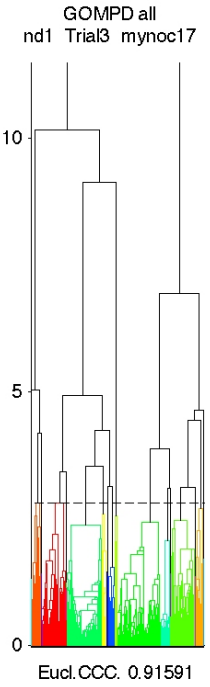

Figure 10. Hierarchical linkages of cluster groups for all 12 months.

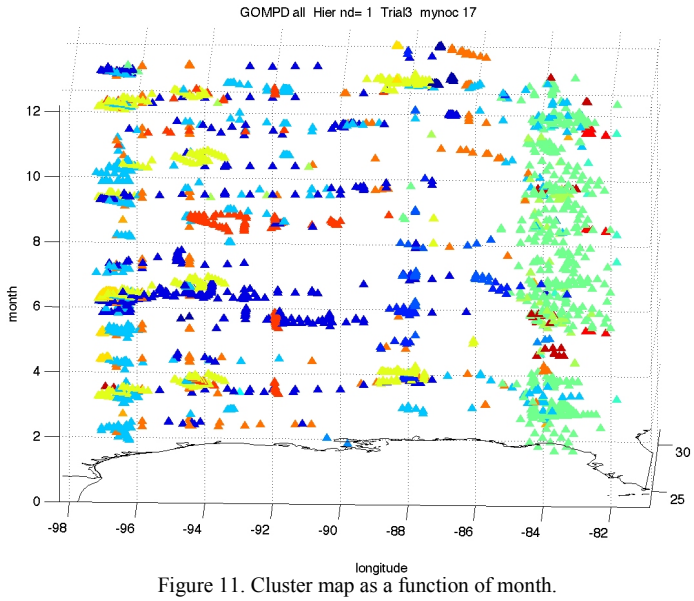

Figure 11. Cluster map as a function of month.

VIII. Summary

Hierarchical clustering methods were applied to NAVOCEANO's MOODS database to categorize sound speed profiles and use group boundaries to identify regions of significant sound speed changes. Categories were determined from a combination of intrinsic profile characteristics and metadata from each profile. The APCEE process uses clustering with different combinations of attributes and ranks results to generate a cluster map. The process allowed clustering to be optimized to the spatial range and temporal scales of interest. The process proved to be adaptive and robust due to the use of principal component analysis and clustering with multiple combinations of attributes (however, some combinations did give low CCC values).

Cluster maps can help survey planners optimize ship CTD sampling operations by indicating the location, time frame, and magnitude of historic sound speed changes. Once identified, the profiles from each cluster group can be further processed to provide statistical information such as an expected sound speed profile, a measure of the variability in each cluster group, and the variability between cluster groups. Clustering also identifies unique individual profiles.

Although PCA is used to improve the linear independence among the attributes used for clustering, linear independence is not assured, and such affects were not pursued. Multiple clustering with different profile attributes gave reasonably consistent CCC values, and maps showed correlation with local oceanographic conditions. No attempt was made to determine the minimum number of profiles needed to obtain a strong cluster result. One limitation encountered was the requirement for a multiple number of trials and attributes to ensure a high-quality cluster result. Certainly part of that is the inherent variability in sound speed profiles; however, improvements in defining attributes should improve results and reduce the number of trials. High variability in either surface temperature or water column depth always yields loose clusters; therefore, additional attributes to mitigate this issue are needed. One possible attribute would relate the time of the profile to the time of the local tidal cycle.

Acknowledgment

The authors gratefully acknowledge the helpful comments and suggestions of Betty Howell, Eric Singer, Casey Taylor, John Dubberley, James Hammack, Ken Grembowicz, James Rigney, Veronica Ladner, and Shannon Breland.

References

[1] National Oceanographic Data Center, Silver Spring MD, http://www.nodc.noaa.gov/GTSPP/gtspp-home.html
[2] The Mathworks Inc., "Statistics Toolbox," version R2006a, 2006.
[3] Mandelberg, M.D. and L.J. Frizzell-Makowski, "Acoustic Provincing of Ocean Basins," Oceans 2000 Conference and Exhibition, Vol I. pp 105-108. 2000.
[4] Everitt, Brian, Sabine Landau, and Morven Leese, *Cluster Analysis*, 4th ed., Oxford University Press, New York, 2001.
[5] National Oceanographic Data Center, Silver Spring, MD map.ngdc.noaa.gov/website/mgg/multibeam