Hydrological determination of hierarchical clustering scheme by using small experimental matrix

M. Cüneyd Demirel
Istanbul Technical University, Institute of Science and Technology, 34469 Maslak Istanbul, Turkey; also at Rosenstiel School of Marine and Atmospheric Sciences, Division of Meteorology and Physical Oceanography (RSMAS/MPO), University of Miami, Miami, Florida, USA

Ercan Kahya[1]
Istanbul Technical University, Civil Engineering Department, Hydraulic Division, 34469 Maslak Istanbul, Turkey

Abstract. In this investigation we tested the performance of five available hierarchical clustering algorithms and nine distance metrics. An arbitrarily chosen experimental matrix (6x3) was used in this analysis to evaluate 45 clustering schemes using the dendrogram and cophonet coefficient index. Priori knowledge of cluster dispersion was the key element to determine non-useful cluster structures. The combination of Euclidean metric and Wards method is most preferred to define homogenous clusters in hydrological studies; however, the combination of Mahalanobis metric and Average Linkage method emerged with a higher cophonet index (0.90420). The most efficient grouping was achieved by the use of City Block and Euclidean metrics in all combinations while the other distance metrics resulted in a non-interpretable dendrogram. Major dendrogram plots and the cophonet index values are presented for visual comparison.

1. Introduction

Unsupervised learning algorithms, such as clustering and nearest neighbour classification, rely on priori definition of distance measures over the input domain (Xing *et al*., 2002). It is known that selecting a "good" metric critically affects the algorithms' performance. Distance metrics are the essential tool in different disciplinary applications ranging from multi-dimensional scaling and unsupervised learning (clustering) to probabilistic roadmap methods for local planners, name-matching tasks (Cohen *et al.* 2003), pattern recognition and even document browsing for data miners (Schultz and Joachims, 2004; Aggarwal *et al.*, 2001). Since such measures are formulated for a specific problem, it might not be accurate for all clustering cases in hydrological applications. Therefore we aimed to explore the performance of metric-clustering algorithm combinations using any priori knowledge for the metrics and hierarchical clustering algorithms.

[1] Assoc. Prof., Hydraulic Division
Civil Engineering Department
Istanbul Technical University
34469 Maslak Istanbul, Turkey
Tel: + 90 (212) 285-3002
e-mail: kahyae@itu.edu.tr

Until the 1980s, the discussion concentrated mainly on techniques that encompass different clustering algorithms. At the end of the 1980s, the whole process of clustering-starting with the selection of distance metrics and method then ending with the validation of clusters became dominant (Arabie et al., 1996). The performance of many clustering and data mining algorithms depend sensitively on their being given a good metric over the input space. This problem is particularly acute in unsupervised settings, such as hierarchical clustering, and is related to the perennial problem of there often being no "absolute right" answers for clustering data (Xing *et al.*, 2002).

In this paper, we are only interested in the performance problem by means of cophonet coefficient and interpretable tree plots (dendrogram). The linear relations are included into the matrix elements to make them link in the same cluster; hence, a visual inspection could be possible at the end of the clustering scheme in order to easily distinguish the success and failure of 45 metric-method combinations. The dendrogram structure is a strong evidence of failures in agglomerative clustering methods, which is often used in hydrological applications.

2. Data

The experimental matrix (6x3) proposed by Demirel (2004) was used in our analysis. An observation number, ranging from 1 to 6, was assigned to each entity (hydrometric station). The three columns of variables were chosen to set each station pairs in the same cluster so that the control structure will be basically 3 distinct clusters at any hierarchy tree: (1, 2), (3, 4) and (5, 6).

3. Methods

The scheme evaluation method does not require any priori assumptions about the metric of the 6x3 clustering sample. The hierarchical clustering uses a distance to the nearest neighboring entity. The available nine metrics in Matlab program are given in Table 1 (Url-1). There are five following critical steps in the analysis procedure:

(i) the choice of variables,
(ii) decision on standardization,
(iii) the choice of similarity metrics,
(iv) selection of methods, the number of clusters,
(v) test of stability (validation) in the clustering scheme.

However the distance metric or similarity metric selection affects the cluster structure. The major steps in a cluster analysis are outlined by Arabie et al., (1996); Everitt, (1993); Url-1, and Hair et al., (1987).

Table 1. Distance metrics (Url-1).

Squared Euclidean distance:

$$d_{rs}^{2} = (x_r - x_s)(x_r - x_s)' \qquad \text{Eq. (1)}$$

Euclidean distance:

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^{2} \right]^{1/2} \qquad \text{Eq. (2)}$$

Mahalanobis distance:

$$d_{rs}^{2} = (x_r - x_s)V^{-1}(x_r - x_s)' \qquad \text{Eq. (3)}$$

where V is the sample covariance matrix.

City Block metric:

$$d_{rs} = \sum_{j=1}^{n} \left| x_{rj} - x_{sj} \right| \qquad \text{Eq. (4)}$$

Minkowski metric:

$$d_{rs} = \left\{ \sum_{j=1}^{n} \left| x_{rj} - x_{sj} \right|^{p} \right\}^{\frac{1}{p}} \qquad \text{Eq. (5)}$$

Note that for $p = 1$, the Minkowski metric becomes the City Block metric; and for $p = 2$, the Minkowski metric is equal to the Euclidean distance.

Cosine distance:

$$d_{rs} = \left(1 - x_r x'_s / (x'_r x_r)^{\frac{1}{2}} (x'_s x_s)^{\frac{1}{2}} \right) \qquad \text{Eq. (6)}$$

Correlation distance:

$$d_{rs} = 1 - \frac{(x_r - \bar{x}_r)(x_s - \bar{x}_s)'}{[(x_r - \bar{x}_r)(x_r - \bar{x}_r)']^{\frac{1}{2}} \, [(x_s - \bar{x}_s)(x_s - \bar{x}_s)']^{\frac{1}{2}}} \qquad \text{Eq. (7)}$$

where

$$\bar{x}_r = \frac{1}{n} \sum_{j} x_{rj} \qquad \bar{x}_s = \frac{1}{n} \sum_{j} x_{sj}$$

Hamming distance:

$$d_{rs} = (\#(x_{rj} \neq x_{sj})/n) \tag{Eq. (8)}$$

Jaccard distance:

$$d_{rs} = \frac{\#[(x_{rj} \neq x_{sj}) \wedge ((x_{rj} \neq 0) \vee (x_{sj} \neq 0))]}{\#[(x_{rj} \neq 0) \vee (x_{sj} \neq 0)]} \tag{Eq. (9)}$$

4. Results:

The performance evaluation is carried out for all nine metrics given in Table 1. Only the significant tree plots will be presented here to demonstrate their clustering performance. It is interesting that, only the City Block (Eq. 4), Minkowski (Eq. 5), and both Euclidean metrics (Eqs. 1 and 2) performed well with hierarchical clustering method combinations. However Hamming and Jaccard measures (Eqs. 8 and 9) failed and resulted in the same tree structure except for the Centroid method case (Figures 1 and 2).

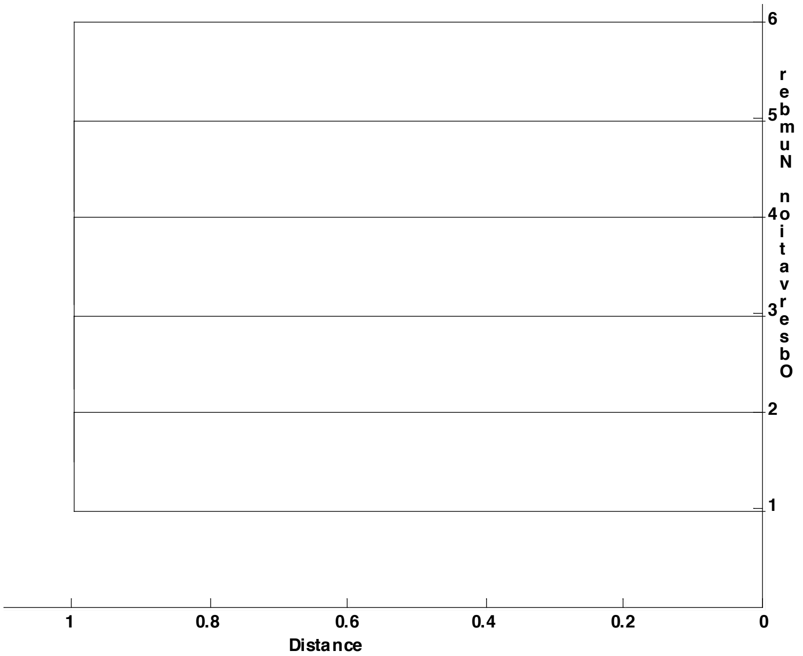

Figure 1. Hierarchy tree plot for the combination of Hamming distance metric and Single Linkage method.

The Mahalanobis distance (Eq. 3) and Wards' method combination resulted in a clear and distinctive tree plot with a high cophonet index of 0.84603, indicating a robust clustering scheme for hydrological studies (Figure 3, Table 2).

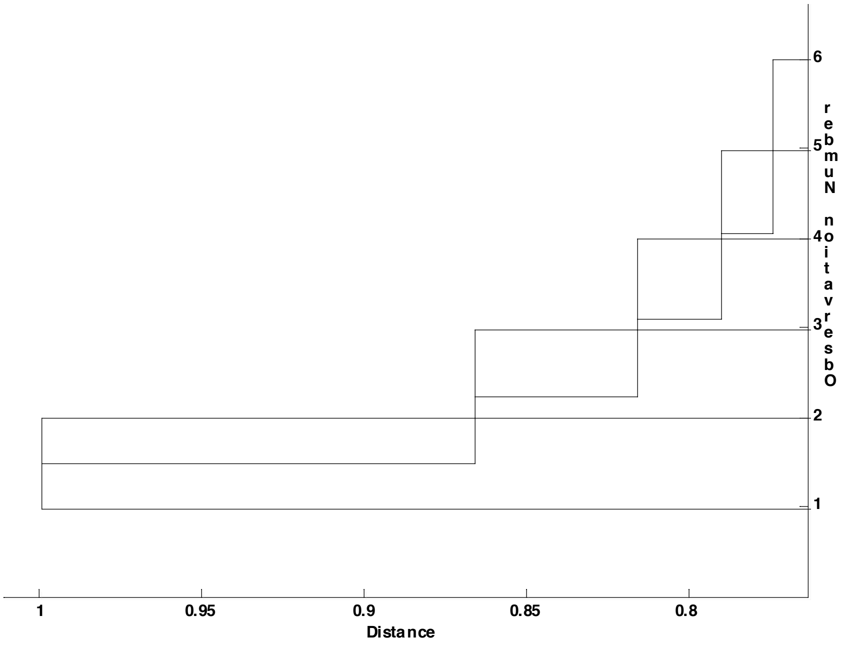

Figure 2. Hierarchy tree plot for the combination of Jaccard distance metric and Centroid method.

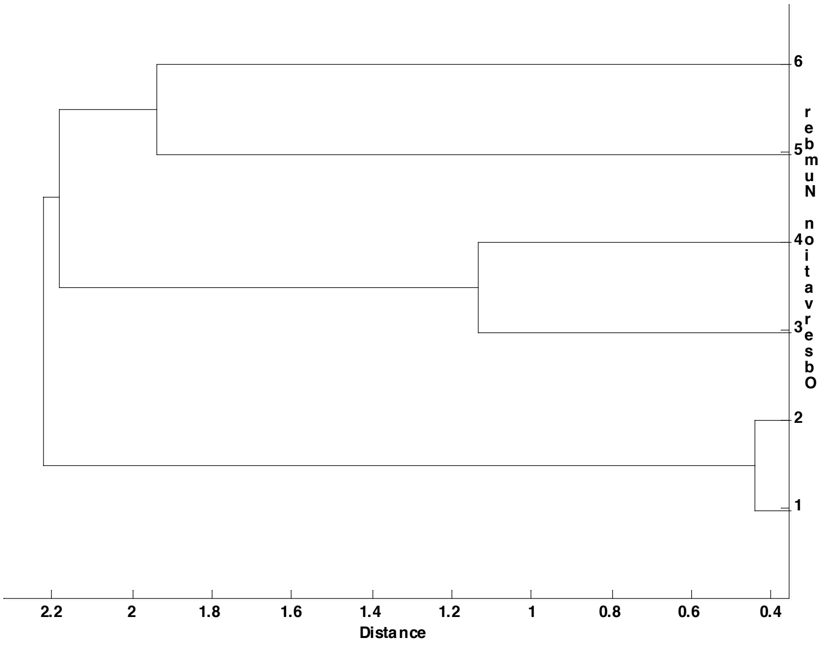

Figure 3. Hierarchy tree plot for the combination of Mahalanobis distance metric and Wards' method.

Euclidean distance is the most commonly used dissimilarity measure in cluster analysis. The literature reviews provided by Gong and Richman (1995) shows that the large majority (85%) of investigators applied this metric in their hydrology based papers. The results of this study verify this common usage too with a well-structured tree plot and a high value of cophonet index as 0.88114 (Figure 4, Table 2). Cosine metrics (Eq. 6) with Single Linkage (SL) method failed in the visual inspection due to well-known chaining affect of SL (Everitt, 1993, Figure 4).

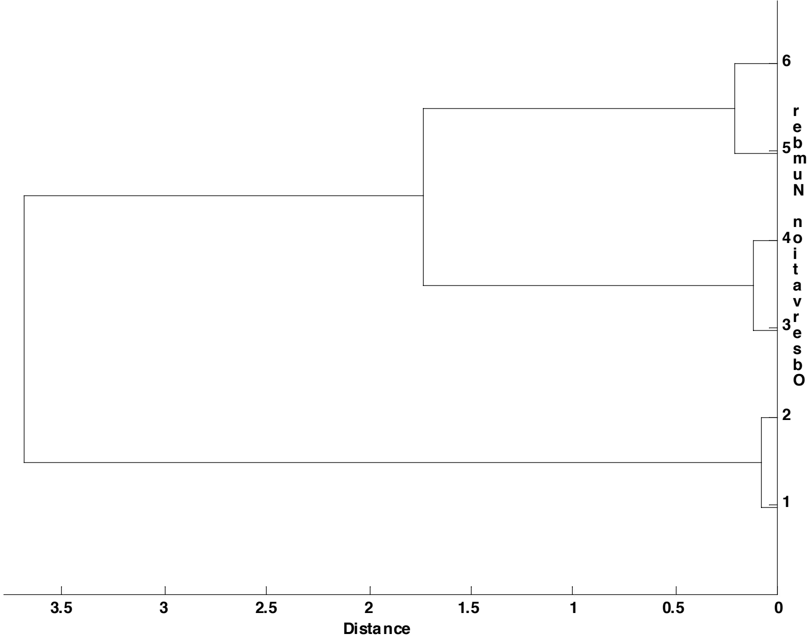

Figure 4. Hierarchy tree plot for the combination of Euclidean distance metric and Wards' method.

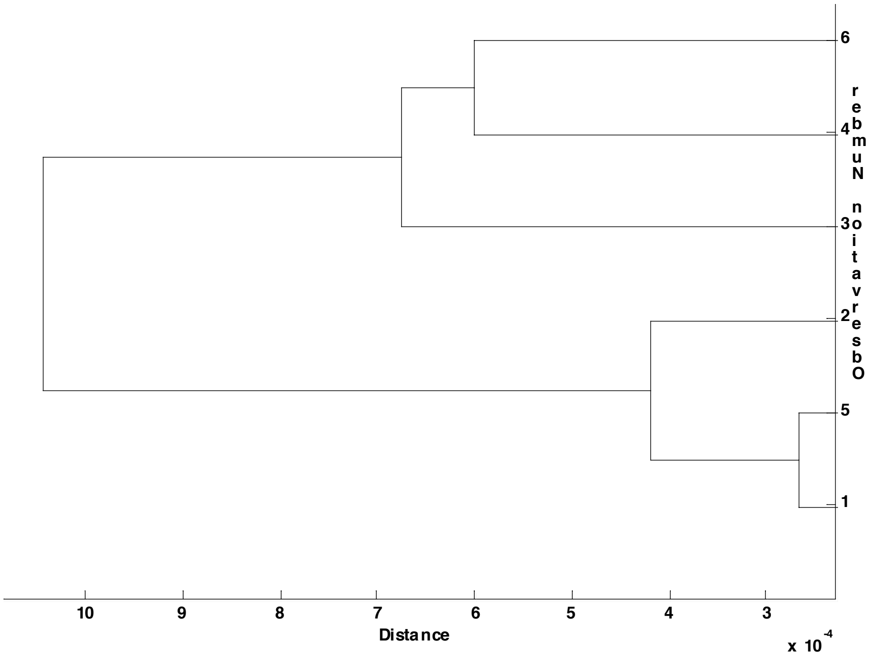

Figure 5. Hierarchy tree plot for the combination of Cosine distance metric and Single Linkage method combination.

It should be noted that Mahalanobis distance metrics can be used to significantly improve clustering performance when using it with Wards' method.

Table 2. Distance metric and clustering method combinations adapted from Demirel (2004).

Method Combinations	Cluster Membership of Observation						**Cophonet Coefficient**
	Obs1	**Obs2**	**Obs3**	**Obs4**	**Obs5**	**Obs6**	
Euclidean and Single Linkage	3	3	1	1	2	2	0.86004
Euclidean and Complete Linkage	3	3	1	1	2	2	0.87795
Euclidean and Average Linkage	3	3	1	1	2	2	0.88335
Euclidean and Centroid	3	3	1	1	2	2	0.88335
Euclidean and Ward	3	3	1	1	2	2	0.88114
Squared Euclidean and Single Linkage	3	3	1	1	2	2	0.84092
Squared Euclidean and Complete Linkage	3	3	1	1	2	2	0.86627
Squared Euclidean and Average Linkage	3	3	1	1	2	2	0.87320
Squared Euclidean and Centroid	3	3	1	1	2	2	0.87320
Squared Euclidean and Ward	**3**	**3**	**1**	**1**	**2**	**2**	**0.87074**
Cityblock and Single Linkage	3	3	1	1	2	2	0.85142
Cityblock and Complete Linkage	3	3	1	1	2	2	0.87174
Cityblock and Average Linkage	3	3	1	1	2	2	0.87788
Cityblock and Centroid	3	3	1	1	2	2	0.87787
Cityblock and Ward	3	3	1	1	2	2	0.87554
Mahalanobis and Single Linkage	1	1	2	2	3	2	0.86177
Mahalanobis and Complete Linkage	1	1	3	3	1	2	0.81073
Mahalanobis and Average Linkage	**1**	**1**	**1**	**1**	**3**	**2**	**0.90420**
Mahalanobis and Centroid	1	1	1	1	3	2	0.88764
Mahalanobis and Ward	3	3	1	1	2	2	0.84603
Minkowski and Single Linkage	3	3	1	1	2	2	0.86004
Minkowski and Complete Linkage	3	3	1	1	2	2	0.87795
Minkowski and Average Linkage	3	3	1	1	2	2	0.88335
Minkowski and Centroid	3	3	1	1	2	2	0.88335
Minkowski and Ward	3	3	1	1	2	2	0.88114
Cosine and Single Linkage	3	3	1	2	3	2	0.68007
Cosine and Complete Linkage	3	3	1	2	3	2	0.68965
Cosine and Average Linkage	3	3	1	2	3	2	0.69095
Cosine and Centroid	3	3	1	2	3	2	0.69094
Cosine and Ward	3	3	1	2	3	2	0.68807
Correlation and Single Linkage	1	1	2	3	2	3	0.69398
Correlation and Complete Linkage	1	1	2	3	2	3	0.74243
Correlation and Average Linkage	1	1	2	3	2	3	0.74406
Correlation and Centroid	1	1	2	3	2	3	0.74393
Correlation and Ward	1	1	2	3	2	3	0.74352
Hamming and Single Linkage	1	1	1	1	2	3	-
Hamming and Complete Linkage	1	1	1	1	2	3	-
Hamming and Average Linkage	1	1	1	1	2	3	-
Hamming and Centroid	1	1	1	1	2	3	-
Hamming and Ward	1	1	1	1	2	3	-
Jaccard and Single Linkage	1	1	1	1	2	3	-
Jaccard and Complete Linkage	1	1	1	1	2	3	-
Jaccard and Average Linkage	1	1	1	1	2	3	-
Jaccard and Centroid	1	1	1	1	2	3	-
Jaccard and Ward	1	1	1	1	2	3	-

Nevertheless this distance measure did not perform well with other clustering algorithms such as Complete Linkage or Centroid. The cluster membership of entities 1, 2, 3 and 4 were defined correctly but the pair 5 and 6 was not merged in the same cluster.

4. Conclusions

In this paper, we presented a performance assessment for different distance metrics and clustering methods. This was accomplished by solving an experimental matrix clustering by 45 scenarios. We evaluated the metrics-method combinations on a collection of hierarchy tree plots and related cophonet index. The diagrams showed that City Block, Minkowski, and both Euclidean distance metrics can be successfully used with any hierarchical clustering methods. The combination of Mahalanobis metric and Average Linkage method emerged with a higher cophonet index value of 0.90420; however, this metric performed best in the dendrogram structure with Wards' method. Hence this combination is recommended for hydrology based clustering studies. Future work is needed both with respect to small matrix and clustering methods. In particular, we do not yet know if the generalization is possible for large dataset. Furthermore, the power of the assessment would be increased, if it was possible to include more complex metrics exist in the comparative clustering problems.

References

Aggarwal, C. C., A. Hinneburg, and D. A. Keim, 2001: On the surprising behavior of distance metrics in high dimensional space. In Proceedings of the International Conference on Database Theory (ICDT 2001), no. 1973 in Lecture Notes in Computer Science, Springer-Verlag, London, England.

Arabie, D., L. J. Hubert, and G. De Soete, 1996: *Clustering and classification*. World Scientific Publ., River Edge, NJ.

Cohen, W. W., R. Ravikumar, and S. E. Fienberg, 2003: A comparison of string distance metrics for name-matching tasks. In Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03).

Demirel, M. C., 2004: Cluster analysis of streamflow data over Turkey. *M.Sc. Thesis*, Istanbul Technical University, Istanbul.

Everitt, B., 1993: *Cluster analysis*. 3rd edn. Halsted Press, Division of Wiley, New York.

Gong, X., and Richman M. B., 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate*, **8**, 897-931.

Hair, J. F., R. E. Anderson, and R. L., Tatham,1987: Multivariate data analysis with readings. Macmillan, New York; Collier Macmillan, London.

Url-1 <http://www.mathworks.com/>, accessed at 19.01.2007.

Xing, E. P., A. Y. Ng, M. I. Jordan, and S. Russell, 2003: Distance metric learning, with application to clustering with side-information. *Advances in Neural Information Processing Systems*, 15, 505-512.