EARTH SCIENCES
RESEARCH JOURNAL

Earth Sci. Res. J. Vol. 12, No. 2 (December 2008): 265-268

DISCUSSION OF "CLUSTERING ON DISSIMILARITY REPRESENTATIONS FOR DETECTING MISLABELLED SEISMIC SIGNALS AT NEVADO DEL RUIZ VOLCANO" BY MAURICIO OROZCO-ALZATE, AND CÉSAR GERMÁN CASTELLANOS-DOMÍNGUEZ

Mehmet C. Demirel[1], Ercan Kahya[2] and Diego Rivera[3]

[1] Ph.D. Student, Department of Water Engineering and Management, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: m.c.demirel@utwente.nl

[2] Civil Engineering Department, American University of Sharjah, PO Box 26666, Sharjah, United Arab Emirates (corresponding author). E-mail: ekahya@aus.edu

[3] Professor, Department of Water Resources, University of Concepcion, Vicente Méndez 595, Chillán, Chile. E-mail: dirivera@udec.cl

The authors are to be congratulated for a systematic investigation of the accurate and non subjective classifying approach in seismic research. The authors have conducted several clustering algorithms to the seismic event records from Volcanological and Seismological Observatory at Manizales. Their objective was to improve the grouping of seismic data (i.e., volcano-tectonic earthquakes, long-period earthquakes and icequakes) digitized at 100.16 Hz sampling frequency. Their study seems adding new approach to their previouswork of Langer et al. (2006) who applied different classificationtechniques to seismic data.

The discussers have the following suggestions to improve the author's investigation on Ruiz volcano data and to be a guide for similar future studies.

1. There are five empirical steps that should be followed in the application of cluster analysis which are (i) selection of variables, (ii) selection of standardization technique (if necessary), (iii) dissimilarity metric, (iv) selection of an appropriate method, and (v) test of stability or validation (Demirel 2004; Everitt 1993; Green *et al.* 1990). These steps are difficult to distinguish in the present study by Orozco-Alzate and Castellanos-Domínguez (2007). The users` experience and preferences in these steps may radically affect the resultant cluster structures. For instance, the selected distance metric was not clearly mentioned in the text. Did the notation ρ_{kl} refer to the correlation coefficient between entities k and l? How many stations were selected near the Olleta crater and the glacier at Nevado del Ruiz volcanic complex? Were there any scale issues in the dataset which may perturb the dissimilarity matrices? Henceforth

the span of the data and some statistical information on the data structure should be explained for easy follow of readers and to avoid the aforementioned questions. Standardization priori to analysis phase is necessary when the scale differences emerged in a dataset (Demirel *et al*. 2008; Everitt 1993; Gnanadesikan *et al.* 1995; Milligan and Cooper 1988).

2. The authors applied several algorithms on their data and reasoned the following statement: "the lack of a single appropriate clustering algorithm". However most of the algorithms were already tested in the literature and the relevant shortcomings are given in many text books (Bacher 2002; Everitt 1993). Single linkage produces chain type cluster which is not be desirable for many applications, and complete linkage may create small and compact clusters (Demirel 2004; Everitt 1993). On the other hand the Wards method emerged to make more successive clusters with small inner variance. Hence it is herein suggested to use the Wards method with the squared Euclidean metric to get more distinct clusters in future investigations.

3. In the context of text indications in notation wise, at page 133: the notation "DC" was not explained in the text. At page 133: D(T,T) designates to distance/dissimilarity measure; however, the notation "d" was used for the same purpose in table 1. It is important to maintain consistent use of notations for the dissimilarity measure throughout the text.

4. The mismatches in labeling were counted for the performance comparison and number of runs was given as 10. The author also mentioned that "Hierarchical methods report the same number of mismatches over the runs". It should be noted that cluster structure in the hierarchical methods do not differ in any run as the steps in dissimilarity calculations and cluster delineation has concrete algorithm; thus, it is herein encouraged that issues similar to these unclear points should be justified in the manuscript.

5. The clustering results were not given in the text. The labels of clusters and statistics (i.e., variance, mean) of each cluster should be summarized in the result section. Only the averaged numbers of mismatches between class labels were presented but this was not adequate for the readers to have appropriate insights regarding the main objective of the study. Since the article is about signal clustering, it would have been very illustrative to put 2 figures: One graph including 3 representative signals (e.g. Langer et al., 200) and one figure representing the topological structure of the clusters, e.g. dendrogram. Both figures allow analyzing in an intuitive way dissimilarities among signals.

6. At page 135: the authors mentioned that "even though the number of cluster is fixed, single linkage and average linkage find second and third clusters of a few objects only". The single linkage and average linkage methods are in the group of unsupervised clustering techniques which has no priori knowledge on number of clusters as partitioning methods; therefore, a justification should have been indicated for that matter (Demirel and Kahya 2007; Kahya *et al*. 2007).

7. As was noted by Morlet *et al.* (1982), seismic signal does vary in amplitude, shape, frequency and phase, versus propagation time. Therefore, for clustering it is necessary to analyze signal's frequency content, as well as to localize in time changes in both, frequency and amplitude. For this task, Wavelet transform is a joint time-frequency signal representation that can give the frequency content of the signal at a particular instant of time by filtering (Sheikholeslami *et al.*, 1998). It is well suited for signal whose frequencies change with time, but also for signal containing noise and transients (Rouyer *et al*., 2008). Also, its multi-resolution property can help detecting the clusters at different levels of accuracy (Sheikholeslami *et al*., 1998). We

propose for further research to apply this technique to Ruiz volcano data. A good reference are Kumar and Foufoula-Georgiou (1997) and Torrence and Compo (1998) for methods. Indeed, Arciniega-Ceballos *et al.* (2008) applied bandpass filters before clustering in seismic data and Rouyer *et al*., (2008) applied a wavelet-based clustering technique.

References

Arciniega-Ceballos A, Chouet B, Dawson Ph. and G Asch (2008) Broadband seismic measurements of degassing activity asociated with lava effusion at Popocatépetl Volcano, Mexico. *Journal of Volcanology and Geothermal Research*, 170: 12-23.

Bacher, J. (2002). "Cluster Analysis." Lecture Notes, Nuremberg.

Demirel, M. C. (2004). "Cluster Analysis of Streamflow Data over Turkey," Istanbul Technical University, Istanbul.

Demirel, M. C., and Kahya, E. "Hydrological determination of hierarchical clustering scheme by using small experimental matrix." *27th AGU Hydrology Days*, Fort Collins, Colorado, 161-168.

Demirel, M. C., Kahya, E., and Rivera, D. (2008). Discussion of "Hydrologic Regionalization of Watersheds in Turkey" by Sabahattin Isik; and Vijay P. Singh in ASCE Journal of Hydrologic Engineering, Sep 2008, Vol. 13, No. 9, pp. 824-834. DOI: 10.1061/(asce) 1084-0699 (2008) 13:9(824) (accepted for publication).

Everitt, B. (1993). *Cluster Analysis. 3rd edn.* , Halsted Press, Division of Wiley, New York.

Gnanadesikan, R., Kettenring, J. R., and Tsao, S. L. (1995). "Weighting and selection of variables for cluster analysis." *Journal of Classification*, 12(1), 113-136.

Green, P. E., Kim, J., and Carmone, F. J. (1990). "A preliminary study of optimal variable weighting in k-means clustering". *Journal of Classification*, 7(2), 271-285.

Kahya, E., Demirel, M. C., and Piechota, T. C. "Spatial grouping of annual streamflow patterns in Turkey " *27th AGU Hydrology Days*, Fort Collins, Colorado, 169-176.

Kumar, P. and Foufoula-Georgiou, E. (1997). Wavelet analysis for geophysical applications. *Reviews of Geophysics*, 35:385–412.

Langer, H., Falsaperla, S., Powell, T., and Thompson, G. (2006). "Automatic classification and a-posteriori analysis of seismic event identification at Soufri?re Hills volcano, Montserrat." *Journal of Volcanology and Geothermal Research*, 153(1-2), 1-10.

Milligan, G. W., and Cooper, M. C. (1988). "A study of standardization of variables in cluster analysis." *Journal of Classification*, 5(2), 181-204.

Morlet, J., Arens, G., Fourgeau, E. and Glard, D. (1982) Wave propagation and sampling theory-Part I: Complex signal and scattering in multilayered media. *Geophysics* 47, 203-221.

Orozco-Alzate, M., and Castellanos-Domínguez, C. G. (2007). "Clustering On Dissimilarity Representations For Detecting Mislabelled Seismic Signals At Nevado Del Ruiz Volcano." *Earth Sci. Res. J.*, 11(2), 131-138

Rouyer T, Fromentin J-M, Stenseth N and B Cazelles (2008) Analysing multiple tiem series and extending significance testing in wavelet analysis. *Marine Ecology Progress Series*, 359:11-23.

Sheikholeslami, Gh., Chatterjee, S. and Zhang, A. (1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24th VLDB conference.*

Torrence, C. and Compo, G. (1998). A practical guide to wavelet analysis. *The Bulletin of the American Meteorological Society*, 79:61–78.