ELSEVIER

Marine Geology 232 (2006) 111 – 113

MARINE GEOLOGY
INTERNATIONAL JOURNAL OF MARINE GEOLOGY, GEOCHEMISTRY AND GEOPHYSICS

www.elsevier.com/locate/margeo

Discussion

Reply to L.J. Hamilton's comment regarding Orpin and Kostylev (2006)—Towards a statistically valid method of textural sea floor characterization of benthic habitats. Marine Geology 225(1–4), 209–222

A.R. Orpin *, V.E. Kostylev

Geological Survey of Canada (Atlantic), PO Box 1006, Dartmouth, Nova Scotia, Canada B2Y 4A2

Accepted 1 August 2006

We appreciate the interest and welcome the criticism of Orpin and Kostylev (2006; herein O & K) and the numerical approach it adopted. Unquestionably, there are statistical methods that can better mine existing data and explore the statistical basis to grouping geological and environmental data. In benthic habitat mapping these methodologies can have a profound effect on boundaries and their validity. More closely linked with benthic habitat mapping today is the requirement to characterise very large areas of the shelf. In an ideal world samples will be taken in a prescribed grid to optimise sampling density and the range of sea floor characteristics. It comes as no surprise that historical data rarely follow this trend. Hence, there is a need for tools that can contribute to sea floor characterisation in a repeatable, transferable, and statistically meaningful way.

Our study primarily examined textural properties, but not exclusively (e.g. O & K Fig. 2d, Fig. 6). This fact appears to have been lost in Hamilton's discussion of texture only, but it is critical to our thesis. At its core is an appreciation that natural variability can be continuous with no clear discontinuities, identifiable as groups or clusters in the data. The goal of clustering is to discover group structure in the data and whether discovered groups are meaningful (Zhao and Karypis, 2005). Classification of measurements is fundamentally based on either goodness-of-fit to a postulated model, or natural groupings revealed through analysis (Jain et al., 1999). The objective is to establish if objects fall into a number of groups, so that objects within a group are more similar to each other than to objects in a different group. Similarly, a key question is whether clusters exist in the data at all? Hence, with our methodology an equally valid result as several clusters is an optimal solution of only 2 clusters with no clear discontinuities. According to Hamilton our failure to find clusters reads as a failure of the approach: "*Their methodology fails when clusters are not found, or when clusters are indicated not to be statistically valid…*". When placed in the context of our objectives it does not. The essence of the problem is outlined in O & K (page 211): "*Habitat classification is aimed at boundary definition, but the classification process is problematic because most statistical clustering techniques will, by their very nature, form clusters, which may or may not represent a meaningful difference between groups*". Here, O & K outline a method which allows repeatable and statistically meaningful partitioning of data into natural groups,

DOI of original article: 10.1016/j.margeo.2006.08.001.

* Corresponding author. Present address: National Institute of Water and Atmospheric Research (NIWA), Private Bag 14-901, Kilbirnie, Wellington, New Zealand. Tel.: +64 4 386 0300; fax: +64 4 386 2153.

E-mail address: a.orpin@niwa.co.nz (A.R. Orpin).

0025-3227/$ - see front matter
doi:10.1016/j.margeo.2006.08.002

which removes subjectivity of "required solutions" in Hamilton's sense.

Field-based study requires that inferences obtained from the classification of a subsample are projected onto large unsampled areas: an age-old practice in geological mapping. This necessitates an estimation of the possible extent of data variability (at a given resolution) in the sampled population, defined by Sokal and Rolf (1969) as "the totality of individual observations about which individual inferences are to be made, existing anywhere in the world or at least within a definitely specified sampling area limited in space and time". On this issue Hamilton argues that "*no matter how many grain size classes are used, and no matter how many values can be assigned to a grain size class, the maximum number of different curve shapes for a set of m seabed samples is m*". Hence, O & K attempted to address the question from a combinatorial approach. In general, combinatorial mathematics is used to calculate the number of combinations of k elements within a larger set (n). Here, we iteratively (from $k=1$ to $n-1$) consider all the possible combinations of k non-zero classes in an n class system. The implicit assumption is that each combination pertains to a distinct state of the n class system where k classes are non zero and $n–k$ classes are zero. Hamilton's appraisal of our combinatorics (points 3 and 7) highlights some ambiguities with our analytical expression. Hamilton arrives with an algebraic expression (2^n-1) that differs from the results of our calculation only by the value of 1, which we should have stated explicitly as 2^n-2. In this case the negative 2 stands for all n cases being 0 or 1, which is logically the same (i.e. no shape information of the distribution curve). Irrespective, both approaches are mathematically sound in terms of the end result, but the overriding issue raised by O & K remains outstanding; complex trends and sub-populations mapped-out from high-resolution data require large sample sets to be statistically meaningful. Ultimately, data resolution and sea-floor sampling strategies should be intimately linked and at a resolution meaningful for statistical analysis. On this point we believe there is universal agreement.

Data reduction is suggested by Hamilton to be a key component to our clustering approach (point 6), in an effort to achieve the desired result—"*O & K sought to reduce two data sets… Since they could not cluster the 32 sizes, this aim was not achieved. … Their reduction methodology is unsupported by practical demonstrations on real data sets*". On the contrary, data reduction was investigated to explore the effect of resolution on output. Recall the conclusions "*Data should be collected at the highest practical resolution, but be reduced to a resolution meaningful for statistical analysis in accordance with the total sample population*". This process was deliberately targeted at boundary recognition and the range of data types typically stored in archives globally. Thematically, the comparison between historical (typically low resolution) and modern laser-derived (high resolution) size data seems entirely appropriate. Hamilton's latter assertion that O & K's findings were "*unsupported by practical demonstrations on real data sets*" is perhaps overstated given the context of the discussion. The text simply concludes that "*reduction to 4-class data appears to have the most profound effect on the shape of the C–H curve for 1–4 group solutions and on the amplitude of C–H values*". Fair comparison was duly made to the original field-based studies from which some of these data were generated (page 219) and our argument follows a logical line of enquiry, albeit not exhaustive. Here, the C–H statistic suggests no clear clusters (2 groups), whereas a field classification suggested 8 textural groups (Orpin et al., 2004). This disparity reflects the scientific question being addressed. Orpin et al. (2004) aimed to make a geological map of the surficial sediment facies, whereas O & K sought a statistical basis to textural classification. With a field-based methodology there is a requirement to resolve complexity down to a manageable level (e.g. relative proportions of clay, silt, sand, and gravel) where relationships and trends can be identified and mapped at an appropriate scale. In practice, this result is similar to that investigated through data reduction by O & K, which reduced complexity and resulted in an optimal statistical solution of 6 groups.

The abundance of clustering packages and algorithms available today only goes to highlight that there are many statistical approaches to many types of classification and data-mining problems. However, the larger scientific goal must be kept in focus. Entropy has been applied successfully in several published field-based studies and provided sound scientific outcomes, examples of which were cited in O & K. The fact that Hamilton prefers a different algorithm, and is happy to argue the case based on limited desktop analysis, is a moot point for criticism of O & K's study. It is difficult to qualify this point further or assess how an alternative clustering methodology might improve the scientific outcomes of O & K as Hamilton does not provide any figures to support his assertion that "*CLARA clusters* [of two unrelated data sets] *were visibly more coherent and better formed than those of the Entropy algorithm*". We acknowledge that the pursuit of outliers rather than broad trends may well give clues of infrequent and/or localized processes or highly specific biological associations, i.e. legitimate data behaving abnormally. However, outliers might equally reflect measurement

artifacts or noise in the data, which is of real concern when working with historical data archives. Hence, unless compelling environmental data exist to corroborate outlier data the overarching benefits of such an approach to characterise benthic habitats appears limited.

Lastly, in terms of information theory, standardization is valuable in many empirical investigations where the absolute totals are meaningless, such as individual grain counts from the respective size classes (cf. Johnston and Semple, 1983). Hamilton's promotion of cumulative curves rather than normalized frequency distributions (referred to as "raw" data by Hamilton) raises an interesting point. Here, Hamilton notes that "*clustering of the cumulative form is more influenced by the overall distribution than clustering of raw data, which is more influenced by the presence of modes*". Given that many of the grain size data discussed in O & K are multimodal perhaps this favours the normalized distribution form. As Hamilton does not provide supporting figures or references it is difficult to qualify the purported advantages of using cumulative percent curves as another form of standardization or its relevance to the hypotheses suggested by O & K. Critical assessment of data transformation and standardization in textural sediment classification is deserving of a separate investigation.

In conclusion, we suggest that Hamilton's criticism of the entropy algorithm (Johnston and Semple, 1983) and the C–H criterion (Calinski and Harabasz, 1974) is not sufficiently substantiated by statistical evidence or corroborating published studies. Entropy as a classifier has been successfully used in earth sciences since Sharp (1973) and has been shown elsewhere to offer benefits over other classification approaches such as *k*-means/medians, hierarchical clustering, and self organized maps (e.g. Li et al., 2004). A critical review by Milligan and Cooper (1985) demonstrated the value of the C–H criterion, showing it to be superior to 30 other statistical estimators commonly used to indicate the statistically meaningful number of groups in cluster analysis. O & K offer a practical demonstration of a repeatable and statistically valid technique to define clusters in large data sets, which removes subjectivity of classification and offers insights for habitat classification.

References

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data Clustering: A Review. ACM Comput. Surv. 31 (3) September.

Johnston, R.J., Semple, R.K., 1983. Classification Using Information Statistics. Concepts and Techniques in Modern Geography, vol. 37. Geobooks, Norwich.

Li, H., Zhang, K., Jiang, T., 2004. Minimum Entropy Clustering and Applications to Gene Expression Analysis. Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004). Stanford, CA, pp. 142–151.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Orpin, A.R., Kostylev, V.E., 2006. Towards a statistically valid method of textural sea floor characterization of benthic habitats. Mar. Geol. 225, 209–222.

Orpin, A.R., Brunskill, G.J., Zagorskis, I., Woolfe, K.J., 2004. Patterns of mixed siliciclastic–carbonate sedimentation adjacent to a large dry-tropics river on the central Great Barrier Reef shelf, Australia. Aust. J. Earth Sci. 51, 665–683.

Sharp, W.E., 1973. Entropy as parity check. Earth Res. 1, 27–30.

Sokal, R.R., Rolf, F.J., 1969. Biometry. The Principles and Practice of Statistics in Biological Research. W.H. Freeman and Co., San Francisco. 776 pp.

Zhao, Y., Karypis, G., 2005. Data clustering in life sciences. Mol. Biotechnol. 31, 55–80.