on Knowledge-Based Intelligent Engineering Systems
& Allied Technologies (KES'2000)
Brighton, UK, 2000

Selecting Fuzzy Connectives to Represent Heavy Metal Distribution in Liverpool Bay

Ameena S. Al-Zaidan and Ludmila I. Kuncheva
School of Informatics, University of Wales, Bangor, UK
E-mail:map005@bangor.ac.uk, l.i.kuncheva@bangor.ac.uk

Abstract

It is known that metals and their combinations can be toxic to certain species in the aquatic environment. Heavy metal concentrations are measured annually in Liverpool bay. Each metal concentration is associated with a fuzzy set "contaminated", defined over the set of sites where the measurements have been taken. Fuzzy aggregating operators are used to construct a set of loading indices accounting for the overall contamination of Liverpool bay. We design and compare 10 loading indices two of which are based on fuzzy integral. To select a useful subset of loading indices we calculate the similarities between them and run a relational data clustering. The final set of indices is found by nominating one representative of each cluster. The fuzzy aggregation operations chosen for the final set of loading indices are {minimum, product, maximum, and arithmetic mean}.
Keywords: Environmental modelling; Fuzzy aggregation operations; Fuzzy measures; Fuzzy integral; Weighted average; Similarity measures; Liverpool bay; Spatial distribution of heavy metals.

1 Introduction

The dynamic nature of coastal waters presents a severe challenge to environmental assessment of disposal activities in near shore waters. Liverpool bay has received large quantities of sludge-input, best known as 'industrial waste', on regular basis since the late 1960's. To regulate the release of these chemicals, disposure places, referred to as 'dumping sites', were designated in Liverpool bay. Heavy metal concentrations are measured regularly on a grid of locations (including the dumping sites) in order to detect and monitor the changes in the ecological structure of Liverpool bay.

In our case study, knowledge of the spatial distribution of heavy metals in surface sediments of Liverpool bay is required for industrial and ecological purposes. The problem is to find an overall distribution of metal concentrations (or contamination) given that the metals have different concentration scales and the way of combining the concentrations is not prescribed. Principal component analysis (PCA) or cluster analysis have been typical choices for this kind of problems [7], but results of both methods are difficult to interpret, unless the data has favourable structure and characteristics. The difficulty in devising one single loading index comes from the fact that there is no *true* contamination distribution which we should try to match.

Fuzzy sets have been applied to various areas of environmental sciences: soil, forest and air pollution, meteorology, water resources, etc. [4, 6]. In our previous study on fuzzy modelling of Liverpool bay data [5], we analysed the overall contamination with 7 heavy metals using 6 aggregating operators.

In this paper we use fuzzy densities to account for the metal "importance", extending the set of loading indices with four additional models. Similarity measures are used to quantify the similarities between the indices. Hierarchical clustering is used to determine groups of similar indices, so that a set of representatives is selected as our final choice.

2 Liverpool Bay data

Liverpool bay receives heavy metal discharges from continuous sources (Mersey and Dee Estuaries), industrial off shore disposal grounds, as well as natural mineral background concentrations through erosion [1]. The effect of metal contamination on biota in sediments is complex; exposure of organisms to high levels of more than one metal introduces further environmental stress. It is therefore appropriate that an approach to developing patterns of contaminant distribution should include all metals. The problem from the point of view of environmental management is to develop patterns from metal data which reflect the current status of sediments.

Environmental monitoring of sludge input in Liverpool bay is conducted annually. For the present research, analysis has been carried out on 7 metal concen-

trations, collected in 1988 from Liverpool bay [1]. These metals are: x_1 = Mercury (Hg), x_2 = Cadmium (Cd), x_3 = Chromium (Cr), x_4 = Copper (Cu), x_5 = Nickel (Ni), x_6 = Lead (Pb) and x_7 = Zinc (Zn).

3 Loading indices design

Let $S = \{s_1, s_2, ..., s_{70}\}$ be the set of 70 sites in Liverpool bay, and let $x_i(s_j)$ be the concentration of metal x_i measured at site s_j. For each x_i, define a fuzzy set A_i over S, corresponding to 'contamination with x_i' with membership function

$$\mu_{A_i}(s_j) = \frac{x_i(s_j) - \min_k\{x_i(s_k)\}}{\max_k\{x_i(s_k)\} - \min_k\{x_i(s_k)\}}. \quad (1)$$

Having designed the 7 fuzzy sets, six fuzzy aggregation operations were used in [5] to define overall loading indices as fuzzy sets on S. These operations are, the Minimum, Product, Geometric mean, Maximum, Arithmetic mean, and Competition jury (drop the highest and the lowest degrees of membership and average the rest). Figure 1 presents the shaded contour-plots using these indices. High gray level intensity corresponds to high contamination, while brighter regions correspond to low contamination.

Here we add to these six indices four new ones: two based on fuzzy integrals and two, calculated as the weighted average. For all these indices we need the so called *fuzzy densities*, one value for each metal, expressing the "importance" of the metal for the overall contamination pattern. The fuzzy integral model is richer than that. It can incorporate not only the individual densities but also the importance of each subset of metals. However, determining heuristically coefficients for all possible $2^7 - 1$ combinations of metals is infeasible, even for the leading experts. This somehow devalues the fuzzy integral models and the question arises whether a simple weighted average using the individual densities would not lead to the same or a similarly useful result as the fuzzy integral. That is why we adopted in this analysis both fuzzy integral and weighter average.

To define the 7 fuzzy densities for Hg, Cd, Cr, Cu, Ni, Pb, and Zn, the so called "*target values*" were used. These were the upper guideline limits of metal contaminants in aquatic species provided in the documentation, published by the Ministry of Agriculture, Fisheries and Food (MAFF), UK. Beyond these concentrations, the metals are considered toxic (individually) to species. Another possible benchmark are the "*shale values*". These are background metal concentrations that are naturally present in sediments. In this study we used both to define 2 sets of fuzzy densities.

We assumed that the fuzzy densities should reflect the degree in which a metal concentration exceeds a benchmark value (target or shale values). The median was taken to represent the metal concentration over all sites and to be compared with the benchmarks. We chose the median to reduce the effect of outliers manifested by unrealistically high concentrations. Example of this, are the dumping of copper wires or lead batteries in the bay. Finally, two fuzzy density vectors, $g = [g^{Hg}, g^{Cd}, g^{Cr}, g^{Cu}, g^{Ni}, g^{Pb}]$ were computed as $\mu_{A_i}(med_i) - \mu_{A_i}(benchmark_i)$. These differences were scaled, so that the smallest density is 0.05 and the largest is 0.95. The results are $g1 = [0.8, 0.94, 0.3, 0.86, 0.05, 0.9, 0.95]$ for shale values, and $g2 = [0.81, 0.44, 0.92, 0.98, 0.05, 0.87, 0.95]$ for target values. Thus, four indices were calculated: fuzzy integral with $g1$ and $g2$ and weighted average with $g1$ and $g2$.

The ten loading indices are denoted as

LI_1: Minimum
LI_2: Product
LI_3: Geometric mean
LI_4: Maximum
LI_5: Arithmetic mean
LI_6: Competition jury
LI_7: Fuzzy integral ($g1$)
LI_8: Weighted average ($g1$)
LI_9: Fuzzy integral ($g2$)
LI_{10}: Weighted average ($g2$)

4 Selecting representatives

The problem now is to select a "reasonable" subset of loading indices which capture all interesting features of the contamination distribution. To base this selection on a formal criterion, we chose to use similarity measures between fuzzy sets.

Let A and B be fuzzy sets on $U = \{u_1, \ldots, u_m\}$. A similarity measure $\mathcal{S}(A, B) \in [0, 1]$ indicates the degree to which A and B are equal or similar. The larger the value of $\mathcal{S}(A, B)$, the more similar A and B are. Many measures of similarity have been proposed in the literature [2, 3]. The Vector-product measure [2] was employed to quantify the similarity between the 10 loading indices

$$\mathcal{S}_2(A, B) = \frac{\sum_{i=1}^m \mu_A(u_i) \cdot \mu_B(u_i)}{\max(\sum_{i=1}^m \mu_A(u_i)^2, \sum_{i=1}^m \mu_B(u_i)^2)}$$

Similarities between all pairs of loading indices were calculated, shown in Table 1. The following observations can be made

[1] Although more recent data has been measured, we chose to use the 1988 data because both literature analysis and human expertise were available for that.

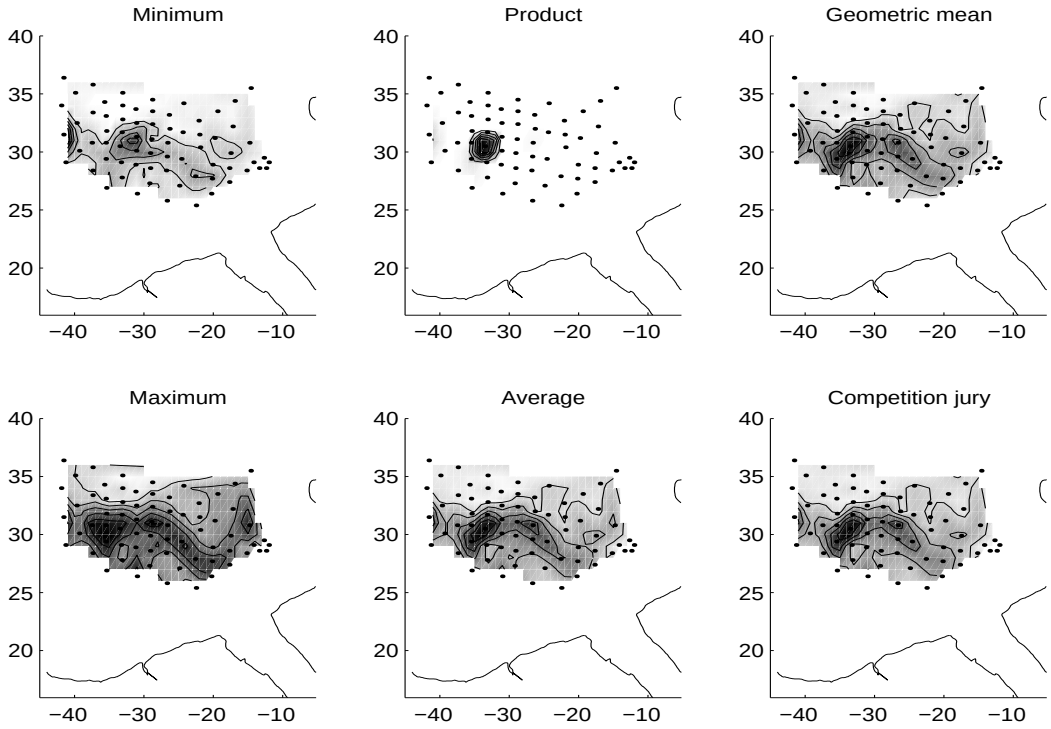

Figure 1: Shaded Contour plots of the six Loading indices.

1. The highest measures of similarity were found between the mean aggregating operators, (especially, LI_5 and LI_6) and the weighted average connectives (LI_8 and LI_{10}) for both shale and target fuzzy densities. These values are shown in boldface in Table 1.

2. Similarities between LI_{10} (Weighted average with $g2$) and all other loading indices, are equivalent to the similarities between LI_6 (Competition jury) and all other indices, i.e., $S_2(LI_{10}, LI_i) \simeq S_2(LI_6, LI_i)$.

3. Loading indices produced by the Minimum and Product operators reflect unmatched information (no similarity with any of the other indices).

Hierarchical relational data clustering, based on the Single Linkage method was applied to identify clusters of loading indices. This method groups objects starting with the set of indices as 10 separate clusters and joining the "most similar" 2 clusters at each step. The algorithm ends with one cluster containing all 10 indices. We chose to classify the 10 loading indices into 2, 3, 4 and 5 clusters and observe the outcome. To make an overall judgement of the results, the Complete and Average Linkage methods were also used. All three hierarchical clustering methods produced the same grouping in our experiments.

The indices appear to be naturally grouped, e.g., mean type aggregation connectives are placed in the same group. The results from the relational clustering are displayed in Table 2 where the clusters are enclosed in brackets. The product (L_2) and the minimum (L_1) are significantly different and both should be retained in the final set of loading indices. From the "mean" group we can pick the average because it is most intuitive and simple. Fuzzy integral and weighted average did not introduce a new cluster, and it seems that the spectrum has been covered by $L_1, \ldots, L_6$. Therefore, the final set can be $\{LI_1$ (minimum), LI_2 (product), LI_4 (maximum) and LI_5 (arithmetic mean)$\}$.

The main difficulty in this type of analysis is that there is no golden standard or target with which we can contrast our results. This gives us the freedom to choose from a large toolbox of ideas and methods but also poses the questions of plausibility or usefulness of the results. Therefore, in the course of the work, we constantly sought feedback from our expert user. An important message re-comfirmed in this study is that the simplest formulas and techniques should be the first choice as they many times suffice for solving the problem.

Table 1: Measures of similarity of all 10 paired loading indices. $(S2)$

	LI_2	LI_3	LI_4	LI_5	LI_6	LI_7	LI_8	LI_9	LI_{10}
LI_1	0.02	0.23	0.08	0.16	0.16	0.09	0.18	0.08	0.17
LI_2		0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01
LI_3			0.33	0.71	0.74	0.41	0.82	0.36	0.73
LI_4				0.47	0.44	0.77	0.40	0.91	0.45
LI_5					**0.97**	0.58	**0.87**	0.50	**0.97**
LI_6						0.56	**0.90**	0.48	**0.99**
LI_7							0.51	0.84	0.57
LI_8								0.43	**0.89**
LI_9									0.49

Table 2: Hierarchical clustering of the 10 loading indices

Number of Clusters	Grouping
2	(LI_2), $(LI_1, LI_3, LI_4, LI_5, LI_6, LI_7, LI_8, LI_9, LI_{10})$
3	(LI_1), (LI_2), $(LI_3, LI_4, LI_5, LI_6, LI_7, LI_8, LI_9, LI_{10})$
4	(LI_1), (LI_2), $(LI_3, LI_5, LI_6, LI_8, LI_{10})$, (LI_4, LI_7, LI_9)
5	(LI_1), (LI_2), (LI_3), $(LI_5, LI_6, LI_8, LI_{10})$, (LI_4, LI_7, LI_9)

5 Acknowledgements

The authors wish to thank Dr. John Wrench from the Ocean Science Department, University of Wales, Bangor for providing the data sets and helping us with valuable consultations throughout this study.

References

[1] V.F. Camacho-Ibar. *Trace elements and polychlorinated biphenyls (PCB) congeners in Liverpool bay sediments*. PhD thesis, University of Wales, Bangor, 1991.

[2] S-M Chen, M-S Yeh, and P-Y Hsiao. A comparison of similarity measures of fuzzy values. *Fuzzy Sets and Systems*, 72:79–89, 1995.

[3] D. Dubois and H. Prade. *Fuzzy Sets and Systems*. Academic press INC., 1980.

[4] J.J. De Gruijter, A.B. McBratney, and K. McSweeney. Special issue on Fuzzy Sets in Soil Sciences. *Geoderma*, 77(2–4), 1997.

[5] L.I. Kuncheva, J. Wrench, L.C. Jain, and A.S. Al-Zaidan. A fuzzy model of heavy metal loadings in Liverpool bay. *Environmental Modelling and Software*, 15(2):161–167, 2000.

[6] K. Lehn and K-M Temme. Fuzzy classification of sites suspected of being contaminated. *Ecological Modelling*, 85:51–58, 1996.

[7] J.A. Markus and A.B. McBratney. An urban soil study: Heavy metals in Glebe, Australia. *Australian Journal of Soil Research*, 34:453–465, 1996.