Using Fuzzy Similarities to Analyze Heavy Metal Distribution in a Marine Environment

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

Abstract

We use data from Liverpool bay and Morecambe bay where metal concentrations are measured annually at a set of designated sites. Each metal concentration is associated with a fuzzy set "contaminated", defined over the set of sites. Ten fuzzy aggregating operators are used to construct loading indices. To select a small set of most different (and hopefully informative) indices we first calculate the similarity between each pair of indices and then group the indices by relational clustering. Four standard measures of similarity have been tried but the results did not comply with our visual observation of similarities between the color contour plots. Therefore a new similarity measure is proposed in this study. From each cluster a loading index is picked as a representative thereby forming the final (small) subset of relevant indices. We propose a general procedure of generating and selecting loading indices and apply it to Liverpool bay and Morecambe bay data. Since there is no objective criteria or target with which we can contrast our results (no contamination pattern that we have to match), feedback has been constantly sought from a domain expert. The results from our study have been subjectively confirmed to be adequate and useful.

Keywords: Environmental modeling; Fuzzy aggregation operations; Similarity measures; Liverpool bay; Morecambe bay; Spatial distribution of heavy metals.

1 Introduction

Knowledge of the spatial distribution of heavy metals in surface sediments is required for industrial and ecological purposes in a marine environment. The problem is to find an overall distribution of metal concentrations (or contamination) given that the individual 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 [9, 10, 11], but results of both methods are not straightforwardly interpretable. The difficulty in devising one single loading index comes from the fact that there is no objective criteria or *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. [8, 10]. In our previous study on fuzzy modeling of Liverpool bay data [1, 9], we analyzed the overall concentration levels with 7 heavy metals using 6 aggregating operators corresponding to the most intuitive choices.

In this paper, we will produce a more generic methodology for generating and selecting loading indices for a marine environment. We use similarity measures between fuzzy sets to quantify the similarities between the loading indices [5, 6, 7, 12, 14, 16]. In Section 2, we describe the type of data used and in Section 3 we explain how loading indices are designed. In Section 4 we present four standard similarity measures and explain by synthetic examples why these measures are inadequate for our type of data. Hence we propose a new measure which we use further in this study. Section 5 gives the results of applying the new measure for selecting loading indices for Liverpool bay and Morecambe bay, and Section 6 offers our general methodology as the conclusions.

2 Liverpool bay and Morecambe bay data

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 (industrial waste) on regular basis since the late 1960's. Heavy metal concentrations are measured annually on a grid of locations in order to detect and monitor the changes in the ecological structure of the bay. Morecambe bay is a large macro-tidal estuary with important wildlife resources. The geographical locations of the two regions are shown in Figure 1.

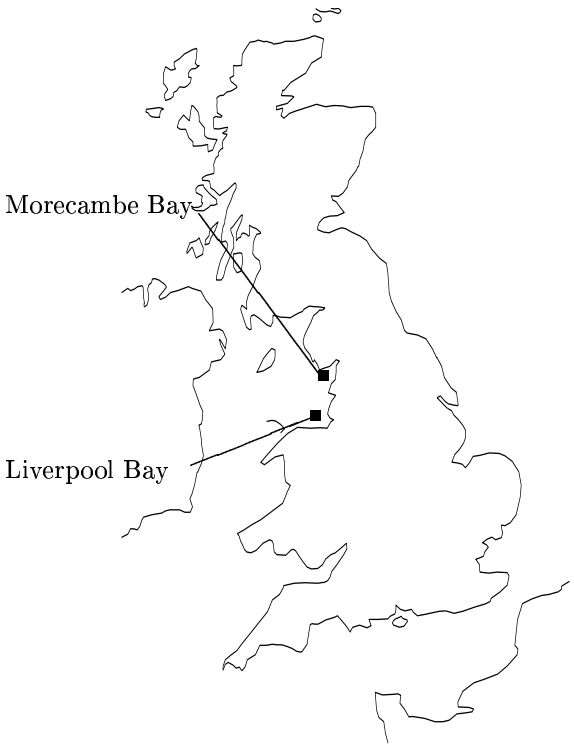

Figure 1: Liverpool Bay and Morecambe Bay location

Morecambe bay is assumed to contain lower levels of contamination compared to Liverpool bay. Therefore, Morecambe bay was suggested as a benchmark for a *relatively uncontaminated* marine region. An important difference between Morecambe bay and Liverpool bay is that the contamination within Liverpool bay is "point-wise" (i.e., there is a designated dumping site), whereas Morecambe bay obtains a diffused input (i.e., lead contamination from the nearby industrial parts of England) [4]. Liverpool bay also receives heavy metal discharges from continuous sources (Mersey and Dee Estuaries) as well as through erosion [3].

Samples of surface sediment in 1988 were collected and chemically analyzed for concentrations of seven heavy metals: 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). The sampling was done at 70 sites in Liverpool bay and 203 sites in Morecambe bay.

3 Constructing loading indices

Let $S = \{s_1, s_2, ..., s_m\}$ be the set of m sites and let $x_i(s_j)$ be the concentration of metal x_i measured at site s_j. To transform the metal concentrations into degrees of membership corresponding to "contamination with x_i," we need lower and upper limits on the concentration values. We shall adopt *average shale values* L_i [15] as lower limits as these are being used in the literature [13]. Upper limits M_i were calculated using upper trigger limits from the UK Department of Environment (DOE) [2].

We chose linear transformation as the simplest model. For each metal x_i we define a fuzzy set A_i over S, corresponding to "contamination with x_i", with membership function

$$\mu_{A_i}(s_j) = \frac{x_i(s_j) - L_i}{M_i - L_i}. \qquad (1)$$

Having designed the 7 fuzzy sets, we use the following fuzzy aggregation operations to define overall loading (contamination) indices

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 ($g2$)

Here $g1$ and $g2$ are two different ways of assigning weights to each metal. These weights have been used as the fuzzy densities to calculate the λ-fuzzy measure for the (Sugeno) fuzzy integrals.

Our choice of aggregation operators was guided by the following reasons:

- Most intuitive aggregation methods (which happen to be the simplest ones as well) were included, e.g., minimum, maximum, and average.

- The additional aggregation methods were needed to make sure that the simple methods had not missed important information.
- We kept the selection within the simplest aggregation operators which at the same time cover a wide range between 0 (the ultimate pessimistic aggregation) and 1 (the ultimate optimistic aggregation), and also account for the possible different weights of the metals.

Next we use similarity measures to find groups of loading indices and subsequently elect one representative from each group.

4 Similarity Measures

In the same way that fuzzy sets allow for gradual transition between full membership and non-membership, a similarity measure captures a gradual transition between equality and non-equality. A similarity measure $\mathcal{S}$ indicates the degree to which two fuzzy sets A and B on the same universal set U are equal or similar. Consider a discrete finite U, $|U| = m$. Denote by $\mathcal{P}(U)$ the class of all fuzzy sets on U and by $\mathcal{C}(U)$ the class of all crisp sets on U.

Definition 1 (Similarity Measures) *: A function $\mathcal{S} : \mathcal{P}(U) \times \mathcal{P}(U) \longrightarrow [0,1]$ is called a similarity measure on $\mathcal{P}(U)$, if $\mathcal{S}$ satisfies the following general properties:*

1. $\mathcal{S}(A,B) = \mathcal{S}(B,A), A, B \in \mathcal{P}(U)$ *(Symmetric);*
2. $\mathcal{S}(D,\bar{D}) = 0$, $D \in \mathcal{C}(U)$, *and* $\bar{D}$ *is the complement of* D*;*
3. $\mathcal{S}(E,E) = 1$, $\forall E \in \mathcal{P}(U)$ *(Reflexive);*
4. *If* $A \subseteq B \subseteq C, \forall A, B, C \in \mathcal{P}(U)$, *then* $\mathcal{S}(A,B) \geq \mathcal{S}(A,C)$ *and* $\mathcal{S}(B,C) \geq \mathcal{S}(A,C)$. *Note that,* $A \subseteq B \subseteq C$ *implies that* $\mu_A(u) \leq \mu_B(u) \leq \mu_C(u) \forall u \in U$.

In general, the larger the value of $\mathcal{S}(A,B)$, the more similar A and B are.

Many measures of similarity among fuzzy sets have been proposed in the literature [6, 5, 7, 12, 14]. The motivation behind these measures is either geometric or set-theoretical. In Pappis et al. [12] and Chen et al. [5], measures of similarity of fuzzy sets (coming from both approaches) are presented and compared. The authors point out that although several properties were common to these measures, there exist notable differences between the similarity measures, and care should be taken when selecting a measure for a particular application.

We consider the following four similarity measures which we believe to be the most intuitive ones

1. *Cardinality-ratio* measure based on intersection and union [6]:

$$\mathcal{S}_C(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{\sum_{i=1}^{m} \min(\mu_A(u_i), \mu_B(u_i))}{\sum_{i=1}^{m} \max(\mu_A(u_i), \mu_B(u_i))}, \qquad (2)$$

2. *Vector-product* measure based on the inter product operation between fuzzy sets [5]:

$$\mathcal{S}_V(A,B) = \frac{A \cdot B}{\max(A \cdot A, B \cdot B)}, \qquad (3)$$

where $A \cdot B$ expresses the inter-product of A and B taken as vectors of membership degrees. Then this measure can be written as,

$$= \frac{\sum_{i=1}^{m} \mu_A(u_i) \cdot \mu_B(u_i)}{\max(\sum_{i=1}^{m} \mu_A(u_i)^2, \sum_{i=1}^{m} \mu_B(u_i)^2)} \qquad (4)$$

3. *Difference-sum ratio* measure

$$\mathcal{S}_D(A,B) =$$

$$1 - \frac{\sum_{i=1}^{m} |\mu_A(u_i) - \mu_B(u_i)|}{\sum_{i=1}^{m} (\mu_A(u_i) + \mu_B(u_i))}, \qquad (5)$$

4. *Symmetrical difference* based measure [6, 7]

$$\mathcal{S}_{Sd}(A,B) = 1 - \|A \bigtriangledown B\|$$

$$= 1 - \frac{\sum_{i=1}^{m} |\mu_A(u_i) - \mu_B(u_i)|}{m} \qquad (6)$$

where $\|A \bigtriangledown B\|$ denotes the relative cardinality of the symmetrical difference $A \bigtriangledown B$.

We demonstrate by a synthetic example some disadvantages of the four measures. Let $X = \{x_1, x_2, x_3\}$ be an universal set. Six example pairs of fuzzy sets A and B are shown in Table 1. The (intuitively) reasonable value of the similarity is expressed verbally and an interval is suggested for each example.

Table 2 shows the values of the four similarity measures on the six example pairs and gives a short comment for each measure.

Since the results with the above four measures of similarity were not satisfactory, here we propose a new measure of similarity defined by

$$\mathcal{S}_*(A,B) = 1 - \frac{\parallel A \bigtriangledown B \parallel}{height(A \bigtriangledown B)}$$

$$= 1 - \frac{1}{m}\left\{\frac{\sum_{j=1}^{m} |\mu_A(u_j) - \mu_B(u_j)|}{\max_j |\mu_A(u_j) - \mu_B(u_j)|}\right\} \qquad (7)$$

As the denominator is undefined for $A = B$, we extend (7) to be

$$\mathcal{S}_*(A,B) = \begin{cases} 1, & \text{if } A = B \\ 1 - \frac{\|A \bigtriangledown B\|}{height(A \bigtriangledown B)}, & \text{otherwise.} \end{cases}$$

The bottom row in Table 2 shows that values of the proposed measure for the six examples are consistent with the desired values.

5 Results

Using (1), metal concentrations were mapped to the unit interval $[0,1]$, so that 0 corresponds to no contamination and 1 corresponds to full contamination. Since some of the concentrations were below the lower limits (the typical shale values), negative degrees of membership were re-assigned the non-membership ($\mu_A(s_j) = 0$).

The 10 loading indices were calculated for the data in Liverpool bay and Morecambe bay. For our study, we wish to reduce the number of loading indices by replacing a group of indices containing similar information with one single index. Having picked the similarity measure, the main question is "How many different groups of loading indices shall we consider"? Results of applying the proposed similarity measure $\mathcal{S}_*$ (4) to Liverpool and Morecambe bay data are listed in Tables 3 and 4, respectively. The similarities for the Morecambe bay are higher than these for the Liverpool bay. This matches our expectation, knowing that the Morecambe bay has a uniformly lower contamination: Since there are large areas in the bay which are not contaminated, and hence all loading indices have low values across these areas, the similarity measure takes high values.

Hierarchical relational clustering, using single linkage, complete linkage and average linkage were applied to identify clusters among the ten loading indices. The results with 2, 3, 4 and 5 clusters are listed in Table 5 for Liverpool Bay and and Table 6 for Morecambe Bay.

Comparing the results in Tables 5 and 6,

- Clustering into three clusters produced the same grouping of indices for both regions
- Relational clustering was able to identify the group consisting of $\{LI_{max},\ LI_{FI(g1)},\ LI_{FI(g2)}\}$, and keep the three together for 3,4 and 5 clusters, for both regions.

This shows that the proposed measure gives stable results, and also indicates a natural grouping of the indices based mainly on the "degree of optimism" of the aggregation.

6 Conclusions

As a conclusion of our study, we suggest the following general methodology for constructing and selecting loading indices in marine environment.

1. Identify the contaminants and collect a data set of measurements of these contaminants across the region of interest.
2. Specify the lower and upper limits for each contaminant.
3. Calculate the membership degrees of the fuzzy sets over the set of sites.
4. Using fuzzy aggregation operators, calculate a set of loading indices. Many operators are applicable at this stage, not only the 10 reported in this study.
5. Calculate pairwise similarities between the loading indices using a similarity measure (we recommend S_* for the reasons explained in the text).
6. Run relational clustering for several number of clusters to find groups of similar indices. (We found out that all three hierarchical clustering procedures led to the same grouping.)
7. Display the results, e.g., by color contour plots, and select one loading index from each group.
8. Use the selected indices to characterize the region of interest. We believe that the indices will be grouped according to the "level of optimism" involved in the aggregation. It is important to select indices which are interpretable in terms of the domain context. For example, the meaning of minimum, maximum and average aggregation can be explained to the environmentalist user.

We applied the above methodology to Liverpool bay and Morecambe bay data and propose the product, maximum and average as the most informative different loading indices.

7 Acknowledgements

The authors are grateful to Dr. John Wrench, our marine environment expert, for supplying the data and for providing constant help, support, and expertise throughout this study.

References

[1] A.S. Al-Zaidan and L.I. Kuncheva. Selecting fuzzy connectives to represent heavy metal distribution in Liverpool Bay. In *Proc. Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pages 602–605, Brighton, UK, 2000.

[2] B.J. Alloway and D.C. Ayres. *Chemical principles of environmental pollution.* Blackie Academic and Professional, 1993. pp.140, 158–160, 232–234.

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

[4] CEFAS. Monitoring and surveillance of non-radioactive contaminants in the aquatic environment and activities regulating the disposal waste at sea, 1995 and 1996. Science series, Aquatic environment monitoring report 51, The Centre for Environment, Fisheries and Aquaculture Science, 1996.

[5] 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.

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

[7] J. Fan and W. Xie. Some notes on similarity measure and proximity measure. *Fuzzy Sets and Systems*, 101:403–412, 1999.

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

[9] 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.

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

[11] 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.

[12] C.P. Pappis and N.I. Karacapilidis. A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, 56:171–174, 1993.

[13] C.L. Macleod R.H.C. Emmerson, S.B. O'reilly-Wiese and J.N. Lester. A multivariate assessment of metal distribution in inter-tidal sediments of the black estuary, uk. *Marine Pollution Bulletin*, 34(11):960–968, 1997.

[14] M. Setnes, R. Babuška, U. Kaymak, and H.R. Lemke. Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics*, 28(3):376–386, 1998.

[15] K.K. Turekian and K.H. Wedepohl. Distribution of the elemnts in some major units of the earth's crust. *Geological Society of America Bulletin*, 72:175–192, 1961. Table 2.

[16] Wen-June Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85:305–309, 1997.

Table 1: Six examples of pairs of fuzzy sets and the desired similarity values

Example	$\mu(x_1)$	$\mu(x_2)$	$\mu(x_3)$	Similarity
T1	1.0	0.2	0.8	Moderate
	1.0	0.1	0.2	$0.50 \leq \mathcal{S}^{T1}(A,B) \leq 0.55$
T2	0.0	0.1	0.0	Moderately high
	0.0	0.9	0.0	$0.65 \leq \mathcal{S}^{T2}(A,B) \leq 0.75$
T3	1.0	0.1	1.0	Moderately high
	1.0	0.9	1.0	$0.65 \leq \mathcal{S}^{T3}(A,B) \leq 0.75$
T4	0.8	0.5	0.2	Low
	0.1	0.6	0.4	$\mathcal{S}^{T4}(A,B) \leq 0.50$
T5	0.0	0.1	0.3	Low
	0.0	0.9	0.8	$\mathcal{S}^{T5}(A,B) \leq 0.50$
T6	0.0	0.0	0.0	Low
	0.1	0.3	0.2	$\mathcal{S}^{T6}(A,B) < \mathcal{S}^{T5}(A,B) \leq 0.50$

Table 2: The values of the four similarity measures for the six examples from Table 1

Similarity measures	T1	T2	T3	T4	T5	T6	Remarks
$\mathcal{S}_C$, Cardinality-ratio, (2)	0.65	0.1	0.72	0.44	0.24	0	Does not comply with the desired intervals; poor result on T2.
$\mathcal{S}_V$, Vector-product, (3)	0.7	0.1	0.74	0.5	0.23	0	Similar to $\mathcal{S}_C$.
$\mathcal{S}_D$, Difference-sum ratio, (5)	0.79	0.2	0.84	0.62	0.38	0	Similar to $\mathcal{S}_C$.
$\mathcal{S}_{Sd}$, Symmetrical difference, (6)	0.77	0.73	0.73	0.7	0.57	0.8	Indistinguishable values, not corresponding to the desired ones.
$\mathcal{S}_*$, **The proposed measure**, (7)	0.61	0.67	0.67	0.52	0.46	0.33	Consistent with the desired intervals and order of preference.

Table 3: Measures of similarity $\mathcal{S}_*(A,B)$, of all 10 paired loading indices of Liverpool bay.

	LI_2	LI_3	LI_4	LI_5	LI_6	LI_7	LI_8	LI_9	LI_{10}
LI_1	1	1	0.27	0.58	0.76	0.27	0.57	0.28	0.57
LI_2	1	1	0.27	0.58	0.76	0.27	0.57	0.28	0.57
LI_3		1.00	0.27	0.58	0.76	0.27	0.57	0.28	0.57
LI_4			1.00	0.31	0.32	0.8	0.32	0.82	0.32
LI_5				1.00	0.37	0.28	0.54	0.28	0.56
LI_6					1.00	0.29	0.29	0.29	0.29
LI_7						1.00	0.3	0.84	0.3
LI_8							1.00	0.3	0.55
LI_9								1.00	0.3

Table 4: Measures of similarity $\mathcal{S}_*(A, B)$, of all 10 paired loading indices of Morecambe bay.

	LI_2	LI_3	LI_4	LI_5	LI_6	LI_7	LI_8	LI_9	LI_{10}
LI_1	1	1	0.9	0.91	0.98	0.89	0.91	0.89	0.91
LI_2	1	1	0.9	0.91	0.98	0.89	0.91	0.89	0.91
LI_3		1	0.9	0.91	0.98	0.89	0.91	0.89	0.91
LI_4			1	0.9	0.9	0.97	0.9	0.97	0.9
LI_5				1	0.9	0.9	0.91	0.89	0.92
LI_6					1	0.9	0.9	0.9	0.9
LI_7						1	0.9	0.96	0.9
LI_8							1	0.9	0.93
LI_9								1	0.9

Table 5: Hierarchical clustering of the 10 loading indices using calculated similarities $\mathcal{S}_*(A, B)$ for Liverpool bay.

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

Table 6: Hierarchical clustering of the 10 loading indices using calculated similarities $\mathcal{S}_*(A, B)$ for Morecambe bay

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