18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

The Growing Hierarchical Self-Organizing Map (GHSOM) for analysing multi-dimensional stream habitat datasets

S. Bizzi[1], R.F. Harrison[2], D. N. Lerner[1]

[1] *Catchment Science Centre, the University of Sheffield*
[2] *Department of Automatic Control and Systems Engineering, the University of Sheffield*
Email: s.bizzi@sheffield.ac.uk

Abstract: River field surveys are carried out to describe biological habitats and the main geomorphic features of a river stretch. They can be extensive, expensive and time consuming campaigns sampling a high number of features. These features belong to a complex river ecosystem characterized by many different processes at various scales from simple to highly non linear. Researchers need sophisticated techniques to manage this multi-dimensional dataset which was so arduously obtained.

In this paper we apply an algorithm, the Growing Hierarchical Self-Organizing Map (GHSOM), to analyse the River Habitat Survey (RHS) database in the UK. The RHS is a system for assessing the character and quality of rivers based on their physical structure. More than one hundred variables were sampled for each survey site. They were sampled at more than 10 000 sites over the past ten years. The GHSOM is a variant of the SOM algorithm which is particularly useful for explorative data mining of multi-dimensional datasets. It produces an intuitive representation of hierarchical relations in the data.

More than seventy ordinal variables, each representing the occurrence of a feature in the river stretch, are analyzed for 7000 sites, and hierarchical patterns are obtained. The algorithm produces hierarchical structure of four layers of clusters: from a general classification of stream habitats composed of 6 clusters to a very fine one with a few hundred clusters. This complex hierarchical structure is firstly interpreted labelling the clusters with the most frequent features, i.e. using just its input variables.

We are interested to assess how closely a habitat type (as defined by a cluster) corresponds to a river type (as defined using different river classifications). A specific index able to assess the supposed link between river types and stream habitats is developed. It is able to quantify the distribution of different river types across the hierarchical clustering structure: it is calculating how much a stream habitat is common amongst different typologies or in other words how well it could be representative of a specific river type. Two different river classifications are analysed. One has been developed by UK Environment Agency (EA) to disseminate the results of the first national RHS. The other has been proposed by UK as a river typology classification to meet the requirement of the Water Framework Directive (WFD). The latter shows a weak link amongst river types and stream habitats. Instead a river classification based on natural drivers such as geology, slope, mean annual discharge and altitude developed by EA shows a much stronger link. These results draw interesting insights of the key roles of natural driving forces on the geomorphic processes responsible of stream habitat formations which deserve further analysis. The hierarchical structure allows furthermore assessing this link for different type of habitat classifications from broad to details ones. Analysing if much finer stream habitat classifications, i.e. composed of high number of clusters, allow a better link with river types creates the possibility to identify which aspects of stream habitats, e.g. very general (different eco-regions) or very detailed (various management of riparian vegetation) ones, are more sensitive to identify river types and to investigate the possible reasons for it.

The framework presented is then suitable to analyse the influences of stream habitats on a full range of environmental objectives. In the present work we analyse river classifications, but the same approach could be applied to other components of the fluvial ecosystem such as fish or invertebrates communities. This approach has capabilities to improve our understanding of the fluvial ecosystem and to bring management benefits. It gives the possibility to develop optimum habitat classifications able to meet management requirements and to minimize the number of habitat classes identified. This output could then produce management benefits addressing the characterization of habitat status and the planning of the future monitoring campaign, which could be optimised in relation to the classification adopted.

Keywords*: Self-Organizing Maps, Hierarchical Clustering, Fluvial Geomorphology, River typologies, River Habitat Survey*

1. INTRODUCTION

Field surveys are a primary resource to analyse biological and geomorphic river processes in order to further our understanding of the processes. Given the inherent complexity of these processes, surveys need to be extensive and are generally time consuming. Because of the enormous effort needed to collect these valuable datasets, which are composed of a high number of different variables characterizing different aspects of the fluvial ecosystem, the information collected has not always been exploited. Often, only selected, more relevant, variables are analysed. There is a constant effort to find an optimum trade off between complexity, i.e. a more realistic conceptualization of the system, and simplicity, i.e. a more intuitive system to interpret.

This paper describes an application of the growing hierarchical self-organising map (GHSOM) (Rauber et al., 2002). The algorithm has been chosen for this application for its capability to develop a hierarchical structure of clustering and for the intuitive outputs which help the interpretation of the clusters. These capabilities allow our analysis of different stream habitat classifications from general to very detailed ones. This technique is a development of the self-organising map (SOM), a popular unsupervised neural-network model for the analysis of high dimensional input data (Kohonen, 2001). The SOM has been used in river research including, for instance, stream classification based on biotic characteristics (Kruk, 2006). This type of technique has been also compared with standard multivariate techniques in ecological data analysis literature, showing similar results constituting a validation of the method (Giraudel and Lek, 2001).

We use the River Habitat Survey (RHS) database for our analysis (Raven et al., 1998). RHS is a system for assessing the morphological character and quality of rivers based on their physical structure. It is a high dimensional dataset; over one hundred variables are collected from each site, and more than ten thousand sites have been surveyed over the last ten years in England and Wales.

As the number of variables collected is very high the dataset could not be easily interpreted. Vaughan (2005) developed an extensive analysis of the RHS at national scale. The analysis was based on a mix of clustering and PCA techniques. The difficulties of working with such multi- dimensional data were well explained, and a number of indexes summarizing key habitat features have been developed. The interpretation of these indexes is not simple and intuitive, and the amount of missing information due to dimensionality reduction is very difficult to assess. Indexes suitable for ecological modelling have been however successfully developed.

Here a GHSOM data mining algorithm is applied to the dataset with the following objectives:

- Managing the data and developing a hierarchical clustering of the dataset.
- Assessing whether or not differences in stream habitats are related to differences in river types.
- Identifying suitable stream habitat classifications in relation to the needs of management

2. DATA MANAGEMENT: GHSOM LABELLED WITH INPUT VARIABLES

This research uses the England and Wales RHS database which contains more than 10,000 sites spread throughout the country. After a clean-up of the database to guarantee a complete set of the 73 variables for each site, around 7,000 remained. The variables represent the occurrence of geomorphic features in the 10 spot-checks within each 500 m stretch, using ordinal variables ranging from zero to ten. The variables are normalized with a logistic function to ensure that all values were within the range [0-1] and are normalized on their variance. The transformation is more or less linear around the mean value, and has smooth nonlinearity at both ends which ensure that all the values are within the range.

The GHSOM (Rauber et al., 2002) algorithm is capable of learning from complex, multidimensional data without specification of what the outputs should be, and of generating a nonlinear classification of visually decipherable clusters. It is composed of a hierarchical structure of independent SOMs. It goes then beyond the SOM's limitation that a fixed number of clusters (i.e. neurons) have to be decided a priori. A graphical representation of a GHSOM is given in Fig. 1. In this picture the map in layer 1 consists of 3X2 clusters and provides a rather rough organization of the input data. The six independent maps in the second layer offer a more detailed view of the data. The input data for one map is the set of sites which is in the corresponding cluster in the upper layer. Each layer represents a certain granularity of the data. A global check in the algorithm decides whether to create a new deeper layer or to stop the algorithm. The algorithm is then able to create a non-symmetric hierarchical structure. It is a fully adaptive architecture for the inherent patterns present in the data.

Running the algorithm for the RHS sites, a hierarchical architecture of four layers is obtained. The first layer is composed of 6 different clusters of habitats. The sites are almost homogeneously spread onto the map. Thus the first layer contains hundreds of sites for each of the 6 clusters. The fourth layer creates a very detailed data representation composed of hundreds of clusters, each one containing only 15 to 30 sites. Such a hierarchical structure can detect broad habitat classification that would be able to detect macro-difference in morphological features identifying for instance different eco-regions, e.g. layer 1 composed of 6 clusters. At the same time it can produce very detailed stream habitat classifications. Layer 3 composed of 226 clusters is able to detect minor differences within similar stream habitats as for instance different approach to management of the vegetation, differences in the type of the predominant bank material or different levels of local artificial features present in the stretch.

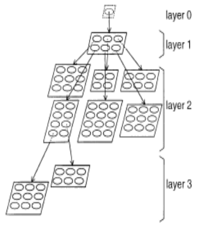
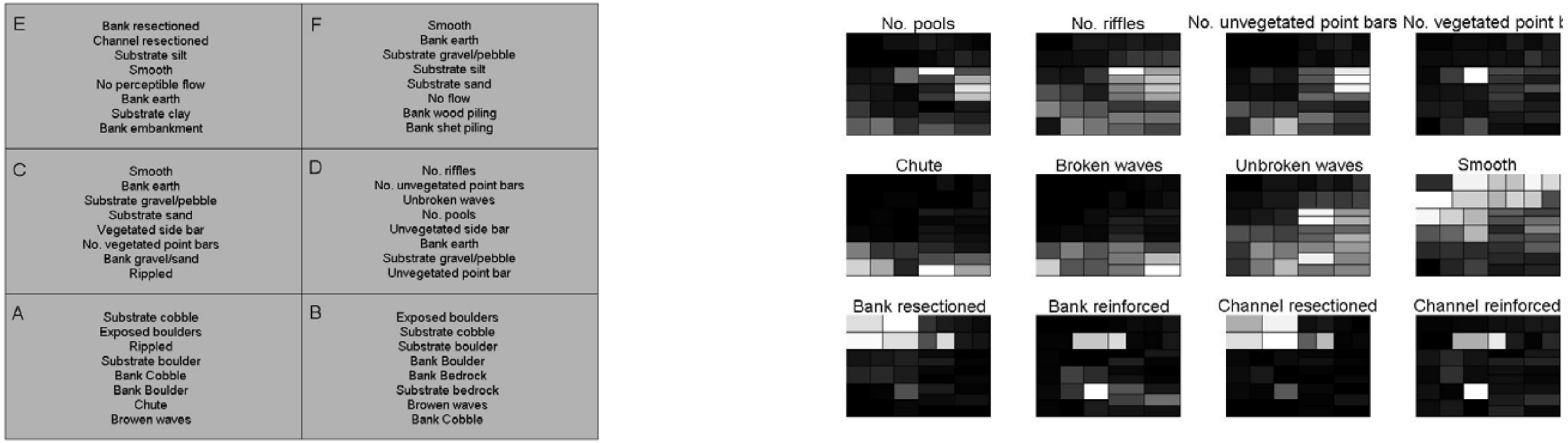


Figure 1. Graphical representation of the hierarchical structure of a GHSOM

The algorithm produces a clustering of stream habitats which broadly confirms the expectations of river type classifications discerning from up-stream to low-land situations given different geologic constraints. The hierarchical structure of clusters is mainly driven by hydrologic, geologic and topographic constraints. As an example we can interpret the meaning of the clusters using the input variables to label the map. Figure 2a shows the first layer labelled with the 8 most frequent input variables. These indicate the frequent and relevant physical features of the clusters, as follows.

- Cluster A - Upstream stream habitats characterized by *exposed boulders* and associated with predominant substrates *as boulders and cobble*. Flow types are mainly *rippled, chute* and *broken waves*, typical respectively of runs, cascades and rapids.
- Cluster "B" – Upstream stream habitats characterized by *exposed boulders* associated with predominant substrates of *boulders, cobble* and *bedrock*. Frequent flow type is *broken waves* mainly associated with rapids.
- Cluster "C" – middle and low-land stream habitats where predominant substrates and bank materials are a mix of *sand, gravel and pebble*. Frequent presence of *vegetated and unvegetated bars*.
- Cluster "D": stream habitats typical of natural alluvial rivers dominated by pool and riffles sequences. *Riffles and pools* are frequent and consequently the *unbroken waves* flow type characterises this cluster. The predominant substrates are *gravel and pebble*. *Unvegetated bars* are frequent.
- Cluster "E": low-land stream habitats characterized by high degree of engineering modifications such as *bank* and *channel resections* and *embankments*. The predominant substrates are *silt* and *clay*. The most frequent flow types are *smooth* and *not perceptible*.
- Cluster "F": low-land stream habitats characterized by a considerable level of engineering modifications. The most frequent intervention is *bank protection* mainly with *wood and sheet piling*. Predominant substrates range from *gravel* and *pebble* to *sand* and *silt*. Frequent flow types vary from *smooth* to *no flow*.

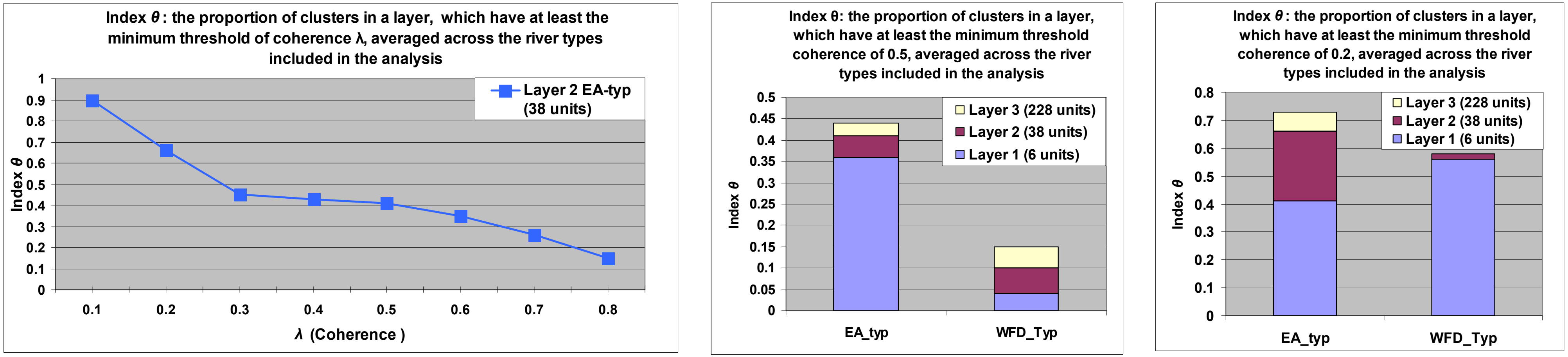
This first layer is just a broad classification of stream habitats into six categories. The 2nd layer shows a much more detailed granularity, as 38 clusters are produced. Each of these can be analysed. Figure 2b shows 12 relevant variables picked up from the 73 input variables mapped onto the 2nd layer. Darker colour means lower value, i.e. lower occurrence of that feature. It is possible then to obtain a comprehensive picture regarding all the variables not labelled in the map and to analyse their patterns in greater detail for the clusters previously discussed. Focusing on the clusters A and B in the 1st layer for instance we can observe, see Figure 2b, how they have been divided in 2x3 and 3x2 clusters respectively in the 2nd layer. The flow type chute often associated with cascades is very frequent especially in the bottom left of both clusters. These areas are then particularly characterized by this specific habitat configuration. Resectioning and reinforcing variables show very low values in clusters A and B, which is coherent with the fact that they represent mainly upstream habitats. Nevertheless for instance at the top right of cluster A, see Figure 2b, there is a high presence of channels and banks reinforced and the related stream habitats are strongly characterized by these impacts. This hierarchical clustering of stream habitats is a precious source of information which allows us to analyse the influences of stream habitats on key characteristics of the fluvial system with extended benefits to academics and management personnel. The next paragraphs will show similar examples.



(a)

(b)

Figure 2. (a) 1st layer (6 clusters) labelled with the 8 most frequent morphological features present at the site (b) 12 input-variables mapped onto the 2nd layer (38 clusters): darker box means lower values of the variables



(a)

(b)

(c)

Figure 3. Definition of coherence (λ): the proportion of a river type present in a cluster (a) The proportion of clusters in a layer which have at least the minimum threshold of coherence λ (averaged across the river types included in the analysis) for 2nd layer labelled with EA-typ. (b-c) The proportion of clusters in a layer which can guarantee at least the minimum amount of coherence of 0.5(b) and 0.2(c), for EA-typ and WFD-typ for the 1st 2nd and 3rd layer

3. CORRESPONDANCE OF STREAM HABITATS AND RIVER TYPES

River type classifications are often based on natural catchment characteristics, such as flow, catchment area, geology and altitude. These variables are amongst the main drivers of the fluvial geomorphic processes. A wide range of river classification schemes have been proposed since the 19th century. Geomorphological classifications aims to discern typologies based on different geomorphic processes. Other ecology-driven classifications are more oriented to identify ecological habitats. However, the two are correlated. It is perhaps reasonable to assume that similar river types have similar stream habitats. This supposed link of habitat type to river type could vary notably due to the complexity of fluvial processes and the high variety of natural and anthropogenic factors which are influencing in habitat formations. Assessing this link is the key to further understand the meaning of a river classification. From this, observations can be gathered on the criteria used to define the river type in order to understand their implication for habitat creation and eventually to identify possible and suitable uses of these classifications for management.

As an example we analyse this link between habitat type and river type for two river classifications. The first one was developed by Environment Agency (EA) to summarize the status of UK's rivers after the first national RHS (Raven et al., 1998). The second is the UK classification of water bodies (UKTAG, 2004) developed to meet the requirements of the European Water Framework Directives (EC, 2000) where 18 types have been identified. For clarity in the comparison with the EA typology we analyse just four of them which together accounts for more than the 50% of UK's rivers. Table 1 summarises the criteria used to define these river typologies.

EA typology (Raven et al., 1998)	***Selection of WFD typology (UKTAG, 2003)***
Steep Streams: Gradient > 40 m/km; Mean annual discharge < 1.25 m^3/s	*River Type 1*: small catchment area (10-100km^2); mean catchment altitude- low (<200m); with a predominantly siliceous geology
Mountain valley rivers: Altitude > 100m; Vertical height between source and site > 300m; Gradient < 10m/km; Channel not constrained by a gorge	*River Type 2*: small catchment area (10-100km^2); mean catchment altitude- low (<200m); with a predominantly calcareous geology.
Chalks rivers: Sites on Cretaceous upper chalk, not influenced by overlying glacial clays	*River Type 5*: medium size catchment area (100-1000 km^2); mean catchment altitude- low altitude (<200m); with a predominantly calcareous geology
Small, lowland riffle-dominated rivers: Site altitude between 20m and 200m; Height of source < 200m; Bankfull width between 2m and 15m; Bed slope > 5m/km; Not on chalk geology	*River Type 10*: small catchment area (10-100km^2); mean catchment altitude- medium (200-800m); with a predominantly siliceous geology

Table 1. River type classifications

For this analysis, we are interested in studying how closely a habitat type (as defined by a cluster) corresponds to a river type (using one of the definitions above). To get an assessment of this link we need to calculate the distribution of the river types amongst the clusters and see whether some specific stream habitats are mainly associated with one river type – the coherence between them. We are also interested to know the degree of scattering of a particular river type between various clusters to assess how many different habitats can be associated with it, and which habitats can be found in a variety of river types. We define a summary index, θ, to represent the average degree of coherence between river types and habitats as follows:

$$\theta = \frac{\sum_{g=1}^{n\_typ} \left(\frac{\sum_{i=1}^{n\_clust} \delta_{g,i}}{n\_clust} \right)}{n\_typ}$$

$$\delta_{g,i} = \left\{ \begin{array}{l} 0 \;\; if \;\; \rho_{g,i} < \lambda \quad \lambda \in \{0-1\} \\ 1 \;\; if \;\; \rho_{g,i} \geq \lambda \end{array} \right\}$$

where $\rho_{g,i}$ is the proportion of all sites in cluster i which are of river type g, and can be seen as the coherence between a river type and a cluster. λ is a coherence threshold so that θ is calculated for levels of coherence at or above those defined by λ. *n_clust* is the number of clusters present in the layer analysed. *n_typ* are the number of river types analysed. θ is then the proportion of clusters in a layer, which have at least the minimum threshold of coherence defined by λ, averaged across the river types included in the analysis.

Figure 3a shows how the index θ varies with the coherence threshold λ examining the EA typology in the 2nd layer (38 clusters). On the left, a higher proportion of the layer (i.e. high value of θ) guarantee low value of coherence, as expected, and *vice versa*. We discuss the graph being split into two parts. On the left are the low values of coherence where we can measure the "rate of scattering": we are interested to assess the proportion of the layer that does not guarantee low value of coherence. The 0.2 value of coherence is guaranteed by around 65% of the layer. It means that around 35% of the map is composed of clusters which contain at least one river type that represents less than the 20% of the clusters it belongs to. On the right we can identify the percentage of clusters highly representative of a river type: 26% of the layer has coherence higher than 0.7 and 15% higher than 0.8.

We have developed similar results for the first three layers of the hierarchical structure and for both the river classifications. They all produced parallel lines with different values of the index θ for the respective values of coherence thresholds λ. For reasons of clarity we report two histograms (Figure 3b-c) comparing these different cases for two values of λ. The first is 0.5 representing the percentage of clusters in a layer which are composed of only one river type for at least their 50%. The other value is 0.2 describing the degree of scattering.

Figure 3b shows that EA-typ for a coherence threshold equal to 0.5 gets the index θ value of 0.36, 0.41 and 0.45 for 1st 2nd and 3rd layers respectively. The θ values obtained for these layers mean that more than one third of the clusters are mainly represented by one river type. Those habitats then can be considered highly representative of the related river type. Furthermore it is expected that a percentage of stream habitats will be shared amongst similar river typologies as for instance between steep-streams and mountain valley rivers. Then the percentage of clusters which are represented mainly by one river type cannot be the majority, especially in the first layer composed only of six clusters. The results confirmed then a remarkable capability to identify river habitats representative of specific river types. WFD-typ has much lower link to river habitats getting an index θ of 0.15 for the 3rd layer and being close to zero for the other layers. Looking at the scattering rates (Figure 3c) the differences between the classifications is smaller especially regarding the 1st and the 2nd layer. It accounts an index θ of 0.66 for Ea-typ 2nd layer and 0.58 for WFD-typ 2nd layer, it means that 34% and 42% of the respective clusters in the layer contain river types which represent less than 20% of the clusters they belong to. It means that, though a cluster can be composed mainly of one river type, it is likely to be composed of small percentage of other river types as well. There is a high heterogeneity of river habitats which are shared between river types. Nevertheless EA-typ shows a constant improvement from the 1st to the 3rd layer, proving how the link between stream habitats and EA-typ is getting better defining finer habitat classifications. There is indeed a major improvement in EA-typ 3rd layer which has an index θ of 0.73 for a coherence threshold λ of 0.2.

We can observe that the criteria used to define EA-typ seems to be very relevant to the geomorphic processes responsible for habitats creation, whereas the ones used in WFD-typ seems to be much more general. It reveals more insights into how the typing factors and ranges given by System A of the Annex II WFD, namely catchment geology (siliceous, calcareous and organic), mean catchment altitude and size, are poorly sensitive to gradients of different stream habitats. It should be assessed that if this is consistent with the aims of this classification or not. Whereas geology, slope, mean annual discharge and site altitude used as criteria to define EA-typ seems to have a relevant role in driving the habitat formation processes. We encourage further analysis on assessing how much stronger the link between river types and stream habitats would become if we take into account any additional natural and anthropogenic (such as land-use or level of engineering modifications) factors. This should bring a much better understanding of the driving forces of the geomorphic processes responsible for habitat formations.

4. IDENTIFYING SUITABLE STREAM HABITAT CLASSIFICATIONS FOR MANAGEMENT

It is worth noting on Figure 3b one aspect of the results for EA-typ: the link between river type and stream habitats does not get much stronger, especially in identifying clusters with high coherence, between the 1st and 2nd layers as expected when increasing of the number of clusters. This suggests that habitat differences identified by the hierarchical algorithm to define a finer classification going from the 6 clusters of the 1st layer to the 38 clusters of the 2nd layer are not as relevant as the first rougher classification to identify river types. The rate of scattering is notably decreasing though (see Figure 3c). Those natural drivers used to define EA-typ seem to be mainly able to detect macro habitat differences. This is an expected outcome. Those natural variables used are characterizing mainly large scale catchment features defining rough boundary conditions for habitat formations. The 3rd layer is able to detect small differences in stream habitats (226 clusters) which seem to be not relevant to river types. They are likely to be a consequence of very specific local conditions not directly related to the geomorphic processes responsible for the characterization of different river types.

We have then a suitable framework to identify efficient stream habitat classifications to support management. We can develop habitat classifications for instance to highlight key habitats for each river type. Having this aims we could discover, as we did, that the capacity to distinguish between river types offered by 6 habitat clusters it is not much higher than if we define 38 habitat clusters. We could investigate the reasons for that, as we did explain it is due to the rough kind of natural variables used to define river types mainly sensitive to major changes in stream habitats, and we could evaluate the costs and benefits to define a higher number of classes given the relative benefits brought by the 2nd layer. Similar examples could be carried out analysing biotic indexes instead of river typologies, representing fish or benthic communities, allowing the identification of efficient habitat classifications able to identify only the differences in stream habitats which are particularly sensitive to the specific management objectives. Unfortunately very often monitoring campaigns, especially at national, scale are not integrated and physical, chemical and biological habitat information is rarely jointly surveyed. RHS database for instance does not contain any biological information. Hence, we strongly encourage in the future to integrate monitoring campaigns to improve our capacity of understanding the system and to address efficient management strategies.

5. CONCLUSIONS

The application of GHSOM to the high dimensional RHS dataset has shown the capability to create flexible hierarchical clustering fully adaptive to the inherent data patterns. The framework can be used to measure how closely a habitat type corresponds to a river type. The WFD-typ showed to be poorly linked with stream habitats. Relevant insights into the driving forces of geomorphic processes responsible for habitat formations could be achieved in future applications adding new natural and anthropogenic variables defining more elaborated river typologies. Furthermore, the framework presented herein offers capabilities to analyse the influences of habitat types on a full range of environmental objectives, such as fish fauna or benthic communities, which could be labelled on the layers as we did for two river classifications. Finally, it gives the possibility to develop optimum habitat classifications that are able to meet management requirements and minimize the number of habitat classes. This output could then produce management benefits addressing the characterization of habitat status and the planning of any future monitoring campaigns, which could be optimised in relation to the classification adopted.

ACKNOWLEDGMENTS

The data for this research was released to our office from Environment Agency. Thanks to Mark Diamond and all his staff for their useful lessons about the RHS database. Thanks to the CatSci project (Early Stage Training in Catchment Science, founded by the European Commission, Marie Curie Actions Project No.: 21149).

REFERENCES

EC, 2000. Directive 2000/60/ec of the European parliament and of the council: establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327 , 22/12/2000 pag. 0001 – 0073.

Giraudel, J.L. and Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*, 146: 329-339.

Kohonen, T., 2001. Self-organizing Maps. Springer.

Kruk, A., 2006. Self-organizing maps in revealing variation in non-obligatory riverine fish in long-term data. *Hydrobiologia*, 553: 43-57.

Rauber, A., Merkl, D. and Dittenbach, M., 2002. The Growing Hierarchical Self-Organizing Map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks, 13(6): 1331-1341.

Raven, P.J. et al., 1998. River Habitat Quality: the physical character of rivers and streams in the UK and Isle of Man. Environment Agency report pag 1-86.

UKTAG, 2004. Type specific reference condition descriptions for rivers in Great Britain. TAG Work Programme 8a (02) Reference conditions for Rivers pag 1-21.

Vaughan, I.P. and Ormerod, S.J., 2005. Increasing the value of principal components analysis for simplifying ecological data: a case study with rivers and river birds. *Journal of Applied Ecology*, 42: 487-497.