# Finding relevant features for the characterization of the ecological status of human altered streams using a constrained mixture model

Alfredo Vellido[1], J. Comas[3], Raúl Cruz[1], Eugenia Martí[2]
[1] LSI. Universitat Politècnica de Catalunya. Barcelona, Spain
[2] CSIC-CEAB. Blanes, Barcelona, Spain
[3] LEQUIA. Universitat de Girona. Girona, Spain. quim@lequia1.udg.es

**Abstract:** The large dimensionality of real data sets usually hampers the interpretability of the results of their analysis. In a previous study, some stream data that are part of the knowledge base of an environmental decision support system were explored through clustering and visualization. The interpretability of these clustering results would be improved by the use of a feature selection strategy based on a method capable of ranking the observed features according to their relative relevance. In this paper, we use one such a method that is an integral part of a probabilistic model for multivariate data clustering and visualization: Generative Topographic Mapping. The feature relevance determination method estimates a saliency for each feature, which is a measure of its influence on the clustering structure of the data. It is, therefore, a fully unsupervised interpretation of relevance. Its application to the available streams data shows that chemical parameters dominate the clustering structure, which is an indication that they might be also relevant for the prediction of the streams' ecological status. Furthermore, no feature is deemed irrelevant by the model, fact that supports expert decisions in the pre-processing stage of the mining of these data.

***Keywords**:* Human-altered streams; multivariate data clustering; Generative Topographic Mapping; Feature relevance determination; Ecological status.

## 1. INTRODUCTION

The data analysed in this study are part of the knowledge base of the environmental decision support system (EDSS) that was the object of the STREAMES (STream REAch Management, an Expert System) European project. The EDSS involved: a) the evaluation of water quality and, to a larger extent, the ecological status of fluvial ecosystems; b) the examination of possible causes of ecosystem impairment; and c) the proposal of ecologically sound management strategies. These strategies dealt with the concept of optimum ecological status[1] of the stream described by the Council of the European Communities (2000).

The data were obtained from streams affected to different degrees by inputs of nutrients from point or diffuse sources. Several streams were selected throughout Europe and Israel, with emphasis on streams located in Mediterranean regions, for which the effects of nutrient inputs are amplified by their usually irregular and relatively low flows.

In a previous study by Vicente *et al*. (2004), the neural network-inspired Generative Topographic Mapping (GTM: Bishop *et al*. (1998)) model was used to cluster and visualize these data (while reconstructing their missing values). We wanted to examine if the clustering was mostly controlled by the geographical distribution of the streams, or by the own physical, chemical and biological data features available. The results indicated that the differences between streams (i.e., mostly geographic) dominated, albeit not completely, the clustering distribution.

The interpretability of the GTM results, both in terms of clustering and visualization, might be rather difficult for data sets of large dimensionality, such as the one analysed here. This interpretability would be greatly improved by the use of a method capable of ranking the observed data features according to their relative relevance in generating cluster structure and, eventually, by the use of a feature selection

[1] The Water Framework Directive (WFD: Council of the European Communities, 2000) considers five categories of ecological status: bad / poor / moderate /good / high. Year 2015 was set as their target to achieve at least a category of "good" for the ecological status of freshwater and coastal ecosystems in Europe.

method based on it. Feature selection for unsupervised learning has received less attention than its supervised counterpart, where relevance is understood in relation to classification or prediction tasks. A recent main advance on feature selection in unsupervised model-based clustering was presented by Law *et al*. (2004) for mixtures of Gaussian distributions, and was extended to the GTM in Vellido *et al*. (2006). The proposed feature relevance determination (FRD) technique, embedded into GTM, allows focusing the interpretation of the clustering results only on a parsimonious subset of selected relevant features, easing considerably the interpretation of the resulting clusters.

The paper is structured as follows. First, the GTM model is introduced, and its extension for FRD is described in some detail. The analysed stream sites and data are then described. This is followed by the presentation of the experimental results and their discussion. Some brief conclusions are finally provided.

## 2. Feature relevance determination for GTM

In general finite mixture of distributions models, the observed data are assumed to be generated by a combination, or finite mixture, of $k$=1,…,$K$ components, weighted by unknown priors $P(k)$. The data associated to each component can be thought of as forming a cluster. Given a $D$-dimensional data set $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, $\mathbf{X}$ is said to follow a $K$-component mixture distribution if the corresponding mixture density can be defined as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|k;\theta_k)P(k), \qquad (1)$$

where each mixture component $k$ is parameterized by $\theta_k$. For continuous data, the choice of Gaussian distributions is a straightforward option. One of the practical drawbacks of general finite mixture models is their lack of data visualization capabilities. The GTM was defined as a constrained mixture of distributions precisely to provide such visualization capabilities, akin to those of the widely used SOM by Kohonen (2001). The GTM is a constrained mixture of distributions model in the sense that all the components of the mixture are equally weighted by the constant term $p(\mathbf{u}_k) = 1/K$, and all components share a common variance $\beta^{-1}$. The GTM can also be seen as a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multivariate data space. As such, it is further constrained in that the centres of the mixture components do not move independently from each other, as they are limited by definition to reside on a low-dimensional manifold embedded in the $D$-dimensional space. This is made explicit through the definition of a prior distribution in the latent space:

$$p(\mathbf{u}) = 1/K \sum_{k=1}^{K} \delta(\mathbf{u} - \mathbf{u}_k), \qquad (2)$$

where $\delta$ is the Kronecker's delta, and the $K$ latent points $\mathbf{u}_k$ are sampled from the latent space, forming a regular grid. This latent space discretization makes the model computationally tractable and provides an alternative to the clustering and visualization space of the SOM.

For each data feature $d$, the functional form of the mapping from a low dimensional latent space onto the multivariate data space is the generalized linear regression model:

$$y_d(\mathbf{u}, \mathbf{W}) = \sum_{m}^{M} \phi_m(\mathbf{u}) w_{md}, \qquad (3)$$

where $\boldsymbol{\Phi}$ is a set of $M$ basis functions $\boldsymbol{\Phi}(\mathbf{u}) = (\phi_1(\mathbf{u}),..., \phi_M(\mathbf{u}))$, originally defined as spherically symmetric Gaussians, and $\mathbf{W}$ is the matrix of adaptive weights $w_{md}$ that specifies the mapping. The probability distribution for a data point $\mathbf{x}$, induced by the latent distribution in (2) and given the adaptive parameters of the model, which are the matrix $\mathbf{W}$ and the inverse variance of the Gaussians $\beta$, can be written as:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \, \|\mathbf{y} - \mathbf{x}\|^2\right\}, \qquad (4)$$

where the $D$ elements of $\mathbf{y}$ are given by (3). Using (2) to integrate the latent variables out, we obtain:

$$p(\mathbf{x}|\mathbf{W}, \beta) = \int p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) p(\mathbf{u}) d\mathbf{u} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \, \|\mathbf{y}_k - \mathbf{x}\|^2\right\} \qquad (5)$$

leading to the definition of the log-likelihood:

$$L(\mathbf{W}, \beta|\mathbf{X}) = \sum_{n=1}^{N} \log\left\{\frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \, \|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}\right\} \qquad (6)$$

Now we can resort to the Expectation-Maximization (EM) algorithm to obtain the Maximum Likelihood estimates of the adaptive parameters of the model: $\mathbf{W}$ and $\beta$. For details on this procedure, see Bishop *et al*. (1998).

### 2.1 FRD-GTM

The GTM was originally defined as a constrained mixture of distributions to provide the

visualization capabilities that general finite mixtures of distributions lack. Even the interpretability of the clustering results provided by the GTM through visualization can be limited for data sets of large dimensionality, such as the one analysed in this study. A method for FRD should help to alleviate this problem.

Recently, a method for feature selection in unsupervised model-based clustering with Gaussian mixture models was presented by Law *et al*. (2004) and extended to GTM (FRD-GTM) by Vellido *et al*. (2006). This method estimates an unsupervised saliency as part of the EM algorithm. The saliency measures the importance of each feature on the definition of the cluster structure yielded by the model. Formally, the saliency of feature *d* can be defined as $\rho_d = P(\eta_d = 1)$, where $\boldsymbol{\eta} = (\eta_1,...,\eta_D)$ is a set of binary indicators that can be integrated in the EM algorithm as missing variables. A value of $\eta_d = 1$ ($\rho_d = 1$) corresponds to the maximum relevance of feature *d*. According

$$p(\mathbf{x}|\mathbf{W},\beta,\mathbf{w}_o,\boldsymbol{\beta}_o,\boldsymbol{\rho}) = \sum_{k=1}^{K}\frac{1}{K}\prod_{d=1}^{D}\left\{\rho_d\, p(x_d|\mathbf{u}_k,\mathbf{w}_d,\beta)+(1-\rho_d)q(x_d|\mathbf{u}_o,w_{o,d},\beta_{o,d})\right\} \quad (7)$$

to this definition, we can define a mixture density for FRD-GTM, similar to that in (5), as:

where $\mathbf{w}_d$ is the vector of $\mathbf{W}$ corresponding to feature *d* and $\boldsymbol{\rho} \equiv \{\rho_1,...,\rho_D\}$. The distribution *p* is a feature- and component-specific version of (4). A feature *d* will be considered irrelevant if $p(x_d|\mathbf{u}_k,\mathbf{w}_d,\beta) = q(x_d|\mathbf{u}_o,w_{o,d},\beta_{o,d})$ for all the mixture components *k*, where $q(x_d|\mathbf{u}_o,w_{o,d},\beta_{o,d})$ is a common density followed by feature *d*. Notice that this is the same as saying that the distribution for feature *d* does not follow the cluster structure defined by the model. This common component requires the definition of two extra adaptive parameters: $\mathbf{w}_o \equiv \{w_{o1},...,w_{o,D}\}$ (so that $\mathbf{y}_o = \phi_o(\mathbf{u}_o)\mathbf{w}_o$) and $\boldsymbol{\beta}_o \equiv \{\beta_{o1},...,\beta_{o,D}\}$), and it should reflect any prior knowledge we might have regarding irrelevant features. It accounts for data observations that the GTM constrained mixture components cannot explain well; in other words, data observations that do not fit with the cluster structure described by the model.

The Maximum Likelihood criterion can now be stated as the estimation of those model parameters that maximize the log-likelihood:

$$L(\mathbf{W},\beta,\mathbf{w}_o,\boldsymbol{\beta}_o,\boldsymbol{\rho}|\mathbf{X}) = \sum_{n=1}^{N}\log\left\{\frac{1}{K}\sum_{k=1}^{K}\prod_{d=1}^{D}(a_{knd}+b_{knd})\right\}, \quad (8)$$

where

$$a_{knd} = \rho_d(\beta/2\pi)^{1/2}\exp\left(-\frac{\beta}{2}\left(\sum_m \phi_m(\mathbf{u}_k)w_{md} - x_{nd}\right)^2\right) \quad (9)$$

and

$$b_{knd} = (1-\rho_d)(\beta_{o,d}/2\pi)^{1/2}\exp\left(-\frac{\beta_{o,d}}{2}\left(\phi_o(\mathbf{u}_o)w_{o,d} - x_{nd}\right)^2\right) \quad (10)$$

We can resort again to the EM algorithm to calculate the model parameters. The complete log-likelihood of the model can be written as:

$$L_c(\mathbf{W},\beta,\mathbf{w}_o,\boldsymbol{\beta}_o,\boldsymbol{\rho}|\mathbf{X},\mathbf{Z}) = \sum_{n,k} r_{kn}\sum_{d=1}^{D}\log(a_{knd}+b_{knd}) \quad (11)$$

where the *responsibility* $r_{kn}$ is defined as:

$$r_{kn} = p(k|\mathbf{x}_n,\mathbf{W},\beta,\mathbf{w}_o,\boldsymbol{\beta}_o,\boldsymbol{\rho}) = \frac{\prod_{d=1}^{D}(a_{knd}+b_{knd})}{\sum_{k'=1}^{K}\prod_{d=1}^{D}(a_{k'nd}+b_{k'nd})} \quad (12)$$

The maximization of this expected log-likelihood yields the following update formulae for the model parameters:

$$\rho_d^{new} = \frac{1}{N}\sum_{n,k} r_{kn}u_{knd}\,, \quad (13)$$

where

$$u_{knd} = \frac{a_{knd}}{a_{knd}+b_{knd}}\,, \quad (14)$$

$$\beta^{new} = \frac{\sum_{n,k} r_{kn}\sum_d u_{knd}}{\sum_{n,k} r_{kn}\sum_d u_{knd}\left(\sum_m \phi_m(\mathbf{u}_k)w_{md} - x_{nd}\right)^2} \quad (15)$$

$$\beta_{o,d}^{new} = \frac{\sum_{n,k} r_{kn}v_{knd}}{\sum_{n,k} r_{kn}v_{knd}\left(\phi_o(\mathbf{u}_o)w_{o,d} - x_{nd}\right)^2}\,, \quad (16)$$

where

$$v_{knd} = \frac{b_{knd}}{a_{knd}+b_{knd}}\,. \quad (17)$$

The maximum relevance ($\rho_d \to 1$) of a feature, makes the corresponding common component variance vanish: $(\beta_{o,d})^{-1} \to 0$. The elements of matrix $\mathbf{W}^{new}$, for each feature *d*, are obtained as the solution of the following system of equations:

$$\boldsymbol{\Phi}^T\mathbf{G}^*\boldsymbol{\Phi}\mathbf{W}_d^{new} - \boldsymbol{\Phi}^T\mathbf{R}^*\mathbf{X}_d = 0\,, \quad (18)$$

where $\mathbf{R}^*$ has elements $r_{kn}^* = u_{knd^*}r_{kn}$ for a given feature $d^*$ with $r_{kn}$ given by (12), and $\mathbf{G}^*$ has elements $g_{kk'}^* = \begin{cases}\sum_{n=1}^{N} r_{kn}^*, & k = k' \\ 0 & k \neq k'\end{cases}$. Similarly, we obtain $\mathbf{w}_o^{new}$, for each feature, as the solution of:

$$\phi_o{}^T g^*\phi_o\mathbf{w}_{o,d}^{new} - \phi_o{}^T\mathbf{r}^*\mathbf{X}_d = 0\,, \quad (19)$$

where $\mathbf{r}^*$ has elements $r_n^* = \sum_k r_{kn}^* = \sum_k v_{knd^*}r_{kn}$ for a given feature

$d^*$, and $g^* = \sum_{n,k} r_{kn}^*$. Further details of all these calculations can be found in Vellido *et al*. (2006).

### 2.2 GTM visualization

As previously mentioned, GTM was explicitly defined as a constrained mixture model in order to provide simultaneous data clustering and visualization. Each of the points $\mathbf{u}_k$ in the GTM latent space (or *visualization space*) can be considered as a representative of a cluster containing the subset of observed data assigned to it. The *responsibility* in (12) can be used to assign each data record to a cluster. For simplicity, we assign $\mathbf{x}_n$ to the cluster representative $k^*$ that takes maximum *responsibility* for it:

$$\mathbf{u}_{k^*,n} = \arg\max_{\mathbf{u}_k} r_{kn}, \qquad (20)$$

The centres of the GTM mixture components $\mathbf{y}_k$, are usually known as *reference vectors* or *prototypes* of a cluster. Each of the *D* components of these vectors corresponds to one of the features of the observed data and, given their one-to-one relation to the latent points $\mathbf{u}_k$, their values over the visualization space can be plotted using colour-coding. These plots are known as *reference maps* and they provide intuitive visual information on the behaviour of each feature and its influence on the clustering results.

## 3. STREAMS DATA

The STREAMES project focussed on the effects of high nutrient loads on low-order streams. Eleven third-order streams were selected across seven European countries plus Israel. Two of them were discarded for this study due to extreme data incompleteness for the data features selected in this study. Sites were selected to cover a broad range of climate, geomorphology and environmental conditions. Scenarios were differentiated according to hydrologic conditions (mesic and xeric regions) and the dominant land-use within the selected water catchment (agriculture-dominated and non-agriculture dominated). In addition, and in order to estimate the effect of nutrient inputs from point sources on the structure and function of the streams, two reaches located upstream and downstream of a wastewater treatment plant (WWTP) effluent input were selected for each stream. Further details can be found at www.streames.org.

In every reach, six (on average) experimental campaigns were conducted over a year to cover a wide range of environmental conditions. In each reach and on each date, physical (hydrology, hydraulics, morphology), chemical (nutrient and major ions concentrations), and biological (both structural: biofilm biomass and chlorophyll; and functional: nutrient retention and ecosystem metabolism) parameters were measured. In summary, the available data set for this study comprises 11 sites × 2 reaches × 6 (on average) sampling dates.

The original records-to-features ratio was far too low to implement any reliable analytical model. Therefore, experts in the areas of chemistry, biogeochemistry and stream ecology agreed on a far more parsimonious dataset, consisting on 110 records and 22 descriptive features, detailed in table 1.

| TYPE | FEATURE |
|---|---|
| Ion Concentrations (chemical) | Cations ($Na^+ + K^+ + Mg^{2+} + Ca^{2+} + NH_4^+$) |
| | Anions ($Cl^- + SO_4^{2-} + NO_3^-$) |
| | Alkalinity |
| Nutrient Concentrations (chemical) | $NH_4^+$-N |
| | $NO_3^-$-N |
| | $PO_4^{3-}$-P |
| | Dissolved Organic Carbon (DOC) |
| | Conductivity |
| | Dissolved Inorganic Nitrogen (DIN) |
| Hydrological, Hydraulic & Morphologic (physical) | Depth (Wet channel average depth) |
| | Wet Perimeter |
| | Substrate Ratio (Percentage of {Cobbles + Pebbles} substrata, divided by percentage of {Gravel + Sand + Silt} substrata) |
| | Wet Perimeter / Depth Ratio |
| | K1 (Water transient storage exchange coefficient: from water column to transient storage zone) |
| | K2 (Water transient storage exchange coefficient: from transient storage zone to water column) |
| Stream Metabolism & Biofilm (biological) | Respiration (Daily rate of ecosystem respiration) |
| | G.P.P. (Daily rate of gross primary production) |
| | G.P.P.:R. (G.P.P. to Respiration ratio per day) |
| | Daily Light (P.A.R.) |
| | Temperature |
| | Chlorophylla |
| | Biomass |

**Table 1.** List of the 22 features selected for this study, grouped by their typology.

## 4. EXPERIMENTAL RESULTS AND DISCUSION

All the GTM adaptive parameters were initialized, following a standard procedure (see Bishop *et al.* (1998)), as to minimize the difference between the reference vectors $\mathbf{y}_k = \boldsymbol{\Phi}(\mathbf{u}_k)\mathbf{W}$ and the projections into data space that would be generated by a partial PCA, $\mathbf{y'}_k = \mathbf{V}_2\mathbf{u}_m$, where the columns of matrix $\mathbf{V}_2$ are the two principal eigenvectors (given that the latent space considered in this study is 2-dimensional). The grid of latent points $\mathbf{u}_k$ was fixed to a square 10x10 layout and the corresponding square grid of basis functions $\boldsymbol{\Phi}(\mathbf{u})$ was fixed to a 5x5 layout.

Figure 1 provides the saliency results ($\boldsymbol{\rho} \equiv \{\rho_1,...,\rho_D\}$) for the 22 features of the data set.

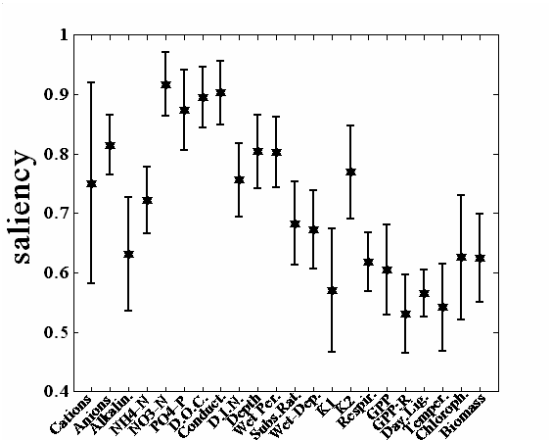


**Figure 1.** Saliency (13) results for the 22 features in the order they are listed in table 1. Bars stretching from the mean (stars) -over 30 runs of the algorithm using different random initializations- plus one standard deviation, to mean minus one standard deviation.

Several conclusions can be drawn from this figure. The first, and most general, is that the FRD-GTM model estimates that none of the features is too irrelevant: in fact, the mean saliency is not lower than 0.5 for any of them. To some extent, this validates the preliminary selection of features carried out by experts (as explained in section 3) in the pre-processing stage of the mining of these data.

All features seem to have a reasonable contribution to the cluster structure of the data. Nevertheless, only a few show consistently high relevance: The two features with $\overline{\rho}_D > 0.9$ ($NO_3^-$-N: nitrate concentration, and Conductivity) belong to the chemical features typology. These are followed in relevance by yet another pair of chemical features (D.O.C. and $PO_4^{3-}$-P: phosphate concentrations), anions concentration, and a couple of physical features (depth and wet perimeter), all of them with $\overline{\rho}_D > 0.8$. Note that these are the features that contribute most to the cluster structure of the data.

At the other end of the relevance range, amongst the least relevant, we find all the biological features from table 1, as well as the alkalinity. Previous results, in Vicente *et al.* (2004), using GTM, indicated that clustering was dominated by stream geographic distribution. The current FRD results help refining this interpretation, and suggest that these geographic differences are linked to the amount of nutrients, in particular through the $NO_3^-$ and D.O.C. concentrations for each stream. Many previous studies have shown that variation in these nutrients among human-altered streams is ultimately caused by catchment land use composition. In addition, the lowest saliency found for biological parameters indicates that ecological controls beyond nutrient availability may constrain variability in metabolic responses among streams.

The feature relevance ranking in Figure 1 can be used as the basis for feature selection, which will ease the interpretation of the clustering results. To illustrate this, the clustering results are displayed, in Figure 2 (left), on the visualization space, according to the cluster membership attribution procedure described in section 2.2. We would like to interpret the clusters according only to the most relevant features. For illustrative purposes, a restrictive selection threshold might be set at $\overline{\rho}_D = 0.89$; this way, the clusters will be interpreted using the reference maps (see section 2.2) of a selection of three features: $NO_3^-$-N and D.O.C. concentrations, and Conductivity, as seen in Figure 2 (right), instead of using the whole set of 22 reference maps available. As an example, three individual clusters (left) are selected and interpreted: large cluster '1' is characterized by very low levels of $NO_3^-$-N, and low levels of D.O.C. and Conductivity. Cluster '2' is characterized by high levels of $NO_3^-$-N, and medium-to-high levels of D.O.C. and Conductivity. Finally, cluster '3' is characterized by medium levels of $NO_3^-$-N, high levels of D.O.C., and very high levels of Conductivity.

## 5. CONCLUSION

The interpretation of the clustering results for large-dimensional data sets is usually difficult or, at least, cumbersome. The data analysed in this study are part of the empirical information of the knowledge base of the EDSS that was the object of the STREAMES European project. Even after a pre-selection carried out by experts, the dimension of the resulting data set makes the interpretation of the clustering results complicated. A method, based on the GTM model, capable of ranking the

observed features according to their relative relevance to explain the data cluster structure, has been introduced. This approach allows focusing the interpretation of the clustering results only on a parsimonious subset of selected relevant features. The proposed FRD-GTM has shown that, although none of the pre-selected features is irrelevant, most of the relevance is conveyed by chemical features. This result suggests that chemical features might be also relevant for the prediction of the streams' ecological status, if we understand this according to functional attributes, such as stream nutrient retention metrics.

The current study should be considered as work-in-progress, and the conclusions drawn in the previous section should be considered as preliminary. An extension of this work would benefit from comparative experiments using alternative unsupervised feature selection methods.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the European Union through the EVK1-CT-2000-00081 STREAMES project. Alfredo Vellido is a research fellow within the Ramón y Cajal program of the Spanish Ministry of Education and Science.

## REFERENCES

Bishop, C.M., Svensén, M., and C.K.I. Williams, GTM: The Generative Topographic Mapping, *Neural Computation*, *10*(1), 215-234, 1998.

Council of the European Communities, Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, 2000.

Kohonen, T., *Self-organizing maps* (3rd edition), Springer-Verlag, 501 pp., Berlin, 2001.

Law, M.H.C., Figueiredo, M.A.T., and A.K. Jain, Simultaneous feature selection and clustering using mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1154-1166, 2004.

Vellido, A., Lisboa, P.J.G., and D. Vicente, Robust analysis of MRS brain tumour data using *t*-GTM, *Neurocomputing*, in press, 2006.

Vicente, D., Vellido, A., Martí, E., Comas, J., and I. Rodríguez-Roda, Exploration of the ecological status of Mediterranean rivers: Clustering, visualizing and reconstructing streams data using Generative Topographic Mapping. In *W.I.T. Trans. on Information and Communication Technologies*, Vol.33, W.I.T. Press, Southampton, 121-130, 2004.

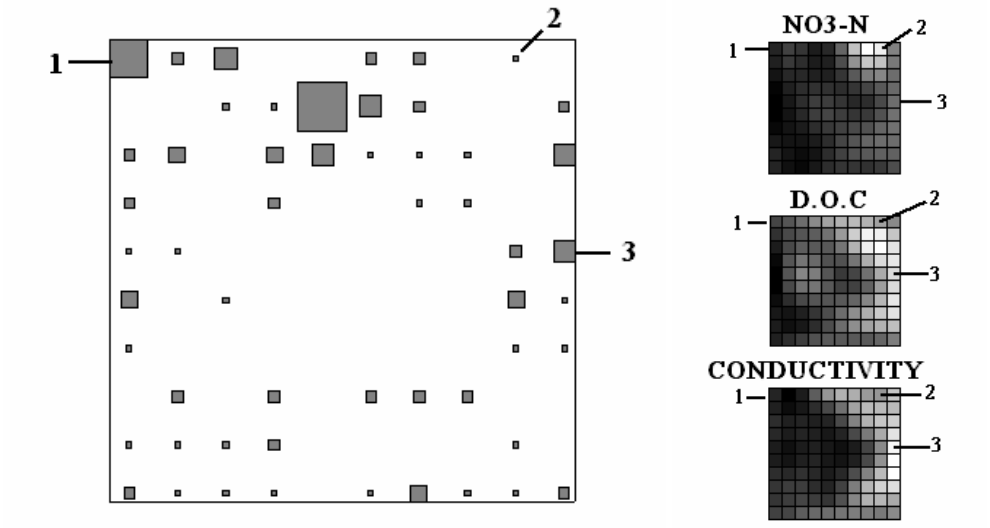


**Figure 2.** (left): GTM 10×10 cluster map in the square *visualization space* where all 110 data records have been mapped onto. The relative size of each cluster (square) indicates the ratio of records assigned to it. As explained in section 2.2, such assignment is based on (12) and (20). The axes of the plot are the elements of the latent vector **u** and convey no meaning by themselves. For that reason, axes are kept unlabeled. Three clusters, labelled as '1', '2', and '3', are selected to illustrate their interpretation using (right): the 10×10 *reference maps* of the data features with highest saliency according to Figure 1. The reference maps are

coded in grey-scale, from black (lowest values) to white (highest values) and, therefore, any cluster can be interpreted using the values of the reference maps corresponding to its location.