Combining Correspondence Analysis And Spatial Statistics For River Water Quality Assessment And Prediction

Henrique Garcia Pereira[1] and Jorge Ribeiro[1,2]

[1] CVRM - GeoSystems Centre of IST
Av. Rovisco Pais, 1
1049-001 Lisboa
(e-mail: hpereira@alfa.ist.utl.pt)
[2] Faculdade de Arquitectura da Universidade Técnica de Lisboa
Rua Prof. Cid dos Santos
Pólo Universitário do Alto da Ajuda
1349-055 Lisboa
(e-mail: jribeiro@fa.utl.pt or jribeiro@alfa.ist.utl.pt)

Abstract. In the context of a practical case study regarding an environment application, a methodology for river water quality assessment and prediction was developed. Such a methodology consists of calculating a quality index by correspondence analysis and predicting its value at non-sampled locations by spatial statistics.
Keywords: Environment, Water Quality Index, Correspondence Analysis, Cumulative Variogram, Kriging.

1 Introduction

When a river is submitted to anthropic environmental stress, *e.g.*, an industrial discharge, a variety of physical-chemical-biological variables are to be monitored in a series of downstream stations in order to guarantee the quality of its water.
For the sake of control by Environment concerned agencies (both official and NGOs), this disparate set of variables should be summarized in some kind of a global straightforward quality index, easy to be appreciated by public opinion and regulatory institutions. On the other hand, this summary measure should also account for all the available information related to the influence of the discharge onto the river water. Once calculated this index, an assessment can be made on the river water quality. But, obviously, if no modelling procedure is applied in order to provide some sort of prediction, this assessment refers only to the sampled points (the stations where the basic measurements are made).
Since no dispersion deterministic model is prone to be applied to the quality index (no mechanism can be assigned to the dynamics of such a hybrid combination of parameters), a stochastic forecasting methodology should be devised in order to predict the index at any non-sampled point (or domain),

as required by the above mentioned Environment concerned agencies. Aiming at approaching this issue from the standpoint of spatial statistics, the standard estimation methodology should be adjusted in order to cope with the specific characteristics of such a problem, where geometry and dynamics play a determinant role. This entails the calculation of a non-Euclidean distance along the river and the development of a non-stationary estimation approach, adjusted to the river flow characteristics.

2 Methodology

The proposed methodology to address this two-fold problem consists of two steps:
In the first step, the barycentric affectation procedure put forward by Benzécri [Benzécri, 1980], and modified by Pereira [Pereira, 1988], was applied in order to produce a comprehensive quality index, ranging from -1 to +1, and accounting for the entire set of variables available at all monitoring stations. For this end, it is required that a panel of experts scrutinizes all measured parameters, split their range into p significant classes, and create two vectors in the variable classes space, designated by the 'GOOD' and 'BAD' poles. These poles represent, respectively, the 'ideal' water quality in its two extremes: pure and polluted water (according to the expert panel). These two 'ideal' vectors are arranged in a 2 x p matrix and submitted to Correspondence Analysis, providing an axis, onto the empirical samples (coded in complete disjunctive form) are projected as supplementary lines. The co-ordinate of each sample in this axis is the required index.
In the second step, the kriging technique, developed in [Matheron, 1965] for the case of space-stationary random functions, was adjusted to the specific features of river water flow according to the guidelines provided in Pereira *et al.* [Pereira *et al.*, 2000]. In particular, the lag for calculation of spatial auto-correlation function - Matheron's variogram - was not measured as an Euclidean distance, but as a 'meandric' one, which is the analogue, for the case of rivers, of the well known 'block distance', used in urban applications. Also, the variogram function and the resultant kriging system were modified to account for the fact that the index at a given point of space along the river depends only on the corresponding upstream values. Hence, a new auto-correlation tool - the cumulative variogram, as proposed by Sen [Sen, 1989] in a different context - was developed in order to avoid any stationarity assumption. This tool - which stands for the Probability Cumulative Function, as the "usual" variogram stands for the Probability Density Function, is defined by:

$$ {}^{a}_{w}\gamma\,[d(i)] \;=\; \sum_{i=1}^{m} (z_w - z_i)^2 \tag{1} $$

where $d(i)$ is the "meandric" distance between w and the station i ($i = 1, ..., m$) and z_w is the index at the point w, to which the cumulative variogram refers. This tool allows to respect the practical order relationships between stations and points (or domains) to be predicted, as given by the river flow. Based on its auto-correlation with upstream values, the proposed index, viewed as a Regionalized Variable, can be estimated at any downstream non-sampled domain by the modified kriging system given below:

$$ \begin{bmatrix} 0 & 0 & ... & 0 & 0 & 0\ 1 \\ \gamma_{1w} & 0 & & 0 & 0 & 0\ 1 \\ \gamma_{2w} & \gamma_{21} & & 0 & 0 & 0\ 1 \\ \vdots & \vdots & & 0 & 0 & 0\ 1 \\ \gamma_{iw} & \gamma_{i1} & ... & 0 & 0 & 0\ 1 \\ \vdots & \vdots & & \gamma_{(m-1)(m-2)} & 0 & 0\ 1 \\ \gamma_{mw} & \gamma_{m1} & & \gamma_{m(m-2)} & \gamma_{m(m-1)} & 0\ 1 \\ 1 & 1 & ... & 1 & 1 & 1\ 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_w \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_i \\ \vdots \\ \lambda_m \\ \mu \end{bmatrix} = \begin{bmatrix} \overline{\gamma}_w \\ \overline{\gamma}_1 \\ \overline{\gamma}_2 \\ \vdots \\ \overline{\gamma}_i \\ \vdots \\ \overline{\gamma}_m \\ 1 \end{bmatrix} \tag{2} $$

where w is the central point of the domain to be estimated on the grounds of i upstream stations ($i = 1, ..., m$), λ are the kriging weights to be assigned to each sample value to predict the average index in the required domain, γ is the usual variogram deduced from the cumulative one by differentiation, and μ is the Lagrange parameter.
Details of the methodological framework where this step relies are given in Ribeiro [Ribeiro, 1999].

3 Case Study

In order to illustrate the above proposed methodology, a case study referring to the Oeiras River (south of Portugal, Fig. 1) is presented. The river is submitted to an industrial discharge and the non-sampled domain of concern on its water quality is located just before the junction with the main Guadiana River (domain W in Fig. 1). Along Oeiras River, water quality is monitored in a series of sampling stations (Fig. 1), for the variables given in the first column of Table 1. The second column of Table 1 contains the classes constructed by the panel of experts for each one of the variables, in order to define the 'Good' and 'Bad' poles, on which the index calculation relies. In the third and forth columns of Table 1, the weights assigned by experts to each variable modality are given.

Variables	Classes	Good Pole	Bad Pole
Biotic diversity based on macro-invertebrate *taxa*	1	0.00	0.80
	2	0.01	0.14
	3	0.05	0.05
	4	0.14	0.01
	5	0.80	0.00
Dissolved Oxygen (%)	[0 ; 50]	0.00	0.91
	] 50 ; 90]	0.09	0.09
	> 90	0.91	0.00
Temperature (^oC)	[0 ; 20]	0.50	0.02
	] 20 ; 30]	0.50	0.98
pH	[0 ; 6]	0.00	0.50
	] 6 ; 9]	1.00	0.00
	] 9 ; 14]	0.00	0.50
Conductivity (μS/cm)	[0 ; 400]	0.95	0.00
	] 400 ; 1500]	0.05	0.05
	> 1500	0.00	0.95
Chemical Oxygen Deficiency (mg/l)	[0 ; 10]	0.90	0.00
	] 10 ; 40]	0.10	0.10
	> 40	0.00	0.90
Sulphates (mg/l)	[0 ; 400]	0.99	0.01
	] 400 ; 3200]	0.01	0.99
Nitrates (mg/l)	[0 ; 25]	0.97	0.01
	] 25 ; 50]	0.02	0.02
	> 50	0.01	0.97
Phosphates (mg/l)	[0 ; 0.54]	0.96	0.01
	] 0.54 ; 0.94]	0.03	0.03
	> 0.94	0.01	0.96
Cu (mg/l)	[0 ; 0.005]	0.98	0.03
	> 0.005	0.02	0.97
Fe (mg/l)	[0 ; 0.3]	0.96	0.03
	> 0.3	0.04	0.97

Table 1. Weights defining the 'GOOD' and 'BAD' poles for river water quality.

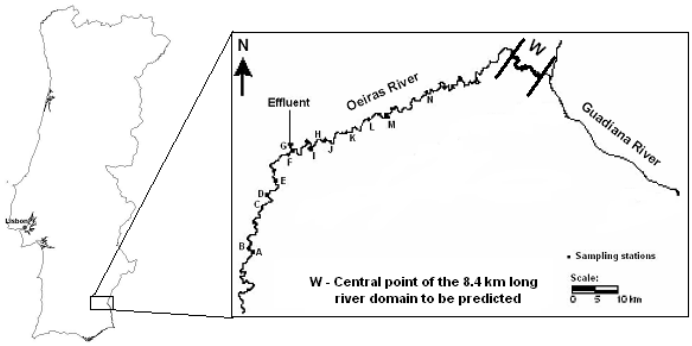

Fig. 1. Location of Oeiras River, sampling stations and domain of concern.

The results of the index calculation for each station according to the first step of the above described methodology are summarized in the histogram of Fig. 2.

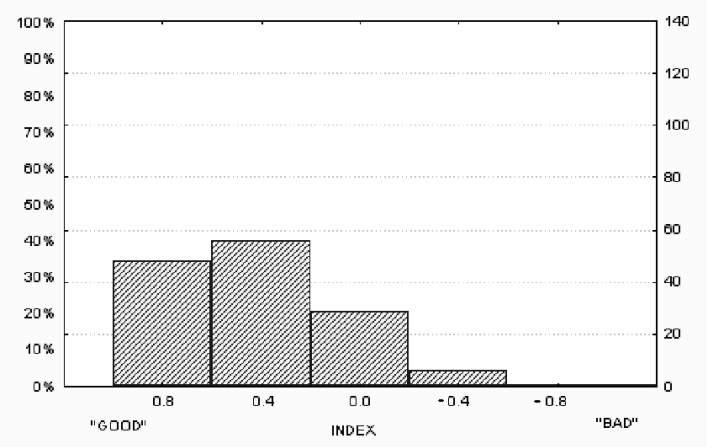

Fig. 2. Histogram summarizing the assignment of the index to the stations.

Since pluviometry can influence the dispersion of pollutants, the index was arranged in each station for the "wet" and "dry" months. The evolution of the average index along the river is shown in Fig. 3.

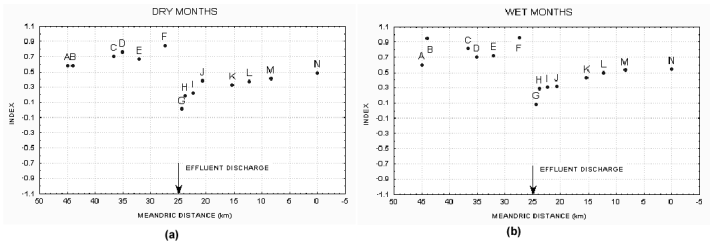

Fig. 3. Schematic representation of the average index for each station (a)-Dry, (b)-Wet months.

Regarding the second step of the methodology, the first point is to calculate the cumulative variogram for each sampling station according to equation 1, as given in Fig. 4.

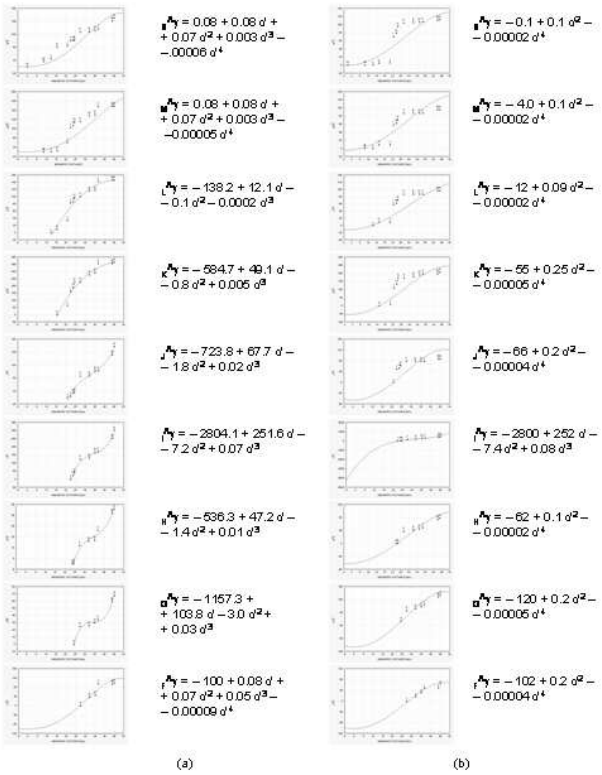

Fig. 4. Cumulative variogram for (a)-Dry and (b)-Wet months.

Differentiating the functions fitted to the curves of Fig. 4, the usual variogram is obtained per station and the system 2 is solved for obtaining the set of λ that permit to predict the value of the index in the non-sampled domain W of Fig. 1 by summing, for all stations, the product of for each λ by the corresponding average index (for dry and wet months).
The results of this calculation are given in Table 2, where the average value of the index in the domain W is compared with the corresponding values in the upstream A-B domain (before the effluent discharge, see Fig. 1).

	Average Index in the Domain W	Average Index in the Domain A-B
Wet Months	0.641	0.788
Dry Months	0.535	0.601

Table 2. Prediction of the Index after and before the effluent discharge.

Table 2 shows that, even though a small decrease in the quality index occurs from A-B to W, the contamination does not reach the Guadiana River, especially in the wet months.

4 Conclusions and Further Work

The proposed methodology allows the estimation of a river water quality index in a non-sampled spatial domain, using all upstream available information.
The point to be developed at this regard is the automatic selection of positive definite functions for the used variogram, obtained by differentiation of the empirical cumulative variogram.
In what concerns the forecasting of the index in time, the length of the available time series (7 years, two samples by year) does not allow any deeper approach than the split into "wet" and "dry" months. Nevertheless, when the time series will have some statistical significance, the parameters of their fitted models can be identified. Then, the same spatial estimation methodology can be applied to these parameters, allowing to predict their values in a non-sampled domain. Finally, the time series at this domain is simulated for future values and a space-time estimation is provided.

5 Aknowledgments

The authors are indebted to UE DG XII, that funded this research in the scope of the RIVERMOD project BE 97-4148 (BRITE - Industrial & Material Technologies).

References

[Benzécri, 1980]J.-P. Benzécri. *Pratique de l'analyse des données. Abrégé théorique - Cas modèle.* Dunod, Paris, 1980.

[Matheron, 1965]G. Matheron. *Les variables regionalisées et leur estimation.* Ed. Masson, Paris, 1965.

[Pereira *et al.*, 2000]H.G. Pereira, J. Ribeiro, A.J. Sousa, L. Ribeiro, A. Lopes, and J. Serôdio. Forecasting river water quality indices. In W.J. Kleingeld and D.G. Krige, editors, *Geostatistics 2000 - Cape Town, Proceedings of 6th International Geostatistics Congress*, pages 591–604. Geostatistical Association of Southern Africa, Cape Town, South Africa, 2000.

[Pereira, 1988]H.G. Pereira. Case study on application of qualitative data analysis to an uranium mineralization. *Quantitative Analysis of Mineral and Energy Resources*, pages 617–624, 1988.

[Ribeiro, 1999]J. Ribeiro. *Formulação de Índices Quantitativos com Base na Descriminação Baricêntrica. PhD Thesis.* IST, Lisbon, 1999.

[Sen, 1989]Z. Sen. Cumulative semivariogram models of regionalized variables. *Math. Geol.*, 21(8):891–903, 1989.