Multivariate indices for analysing correlation structures in environmental datasets

Ragosta Maria
Dipartimento di Ingegneria e Fisica dell'Ambiente - Università degli Studi della Basilicata
V.le dell'Ateneo Lucano
Potenza (85100), Italy
E-mail: maria.ragosta@unibas.it

Di Leo Senatro
Istituto di Metodologie per l'Analisi Ambientale IMAA-CNR
C.da S. Loja
Tito Scalo (PZ, 85100), Italy
E-mai: senatro.dileo@imaa.cnr.it

In the most recent literature, we may find many studies concerning biogeochemical features of ecosystems, communities' ecological structure, analysis of multi-temporal and multi-scales data sets coming from remote observations, atmospheric pollution modeling and forecasting in which multivariate procedures are applied. Particularly we would highlight applications of multivariate methods aimed to compare different hierarchical classifications, to reduce model complexity, to support the fuzzy algorithms application, to identify structural changes in ecological communities, to analyze geographical and temporal distributions of measured variables for evaluating their evolution in scenario analysis (Primpas et al. 2010; Law et al. 2009; Qi et al. 2009; Solans Vila and Barbosa 2009; Zou et al. 2009; Fernandez et al. 2008; Penenko and Tsvetova 2008; Settle et al. 2007; Dawes and Goonetilleke 2006; Felipe-Sotelo et al. 2006; Raik et al. 2006). Moreover the joined application of different statistical procedures to characterize multidimensional datasets is widely used. Specifically a combination of cluster analysis (CA) and principal component analysis (PCA), in order to better characterize the data correlation structure, is currently applied (Katahira et al. 2009; Verfaillie et al. 2009; Cosmi et al. 2008; Ragosta et al. 2008; Shah and Shaheen 2008).

In many of these studies, for a better and easier analysis of the underlying correlation structure of data, it may be useful to apply recursively the multivariate data analysis procedure. The definition of new tools, able to compare different correlation structures obtained starting from a set of input matrices, becomes a crucial point. In this context we propose two new aggregated indices, the *N*ormalized *P*rincipal *C*omponent *I*ndex and the *C*luster *I*ndex, for comparing and interpreting the results of recursive multivariate procedure, based on joined application of CA and PCA methods. These indices allow evaluating, quantitatively, a standardized weight for descriptors and clusters characterizing each correlation structure.

In the large part of the multivariate studies, input data are organized in 2D-matrices [object-observations or object-samples (objects) × measured variables (descriptors)], but it may be interesting to investigate the evolution of the system throughout spatial and/or temporal horizons. In these cases the data matrices have to be organized in multiD-matrices. We introduce a layer for each spatial or temporal event describing different scenarios or characterizing system evolution. Consequently we may organize all input data in a 3D-matrix [layers × objects × descriptors].

Starting from this data matrix, [H layers × M objects × N descriptors], we may determine H 2D-sub-matrices [M objects × N descriptors] and, for each of these sub-matrices (representing the h-th layer with *h=1,...,H*), we may calculate *2H* association matrices (squared and symmetric matrices) applying different similarity measures. Particularly we obtain *H* association matrices $A^h_{[N,N]}$, evaluating the correlation coefficient among the *N* descriptors and *H* association matrices $B^h_{[M,M]}$, evaluating the distance among the *M* objects.

Principal Component Analysis (PCA) is applied to *A*-matrices. For each matrix $A^h_{[N,N]}$ (with *h=1,...,H*), we calculate the eigenvalues $(\lambda^h_1,...,\lambda^h_N)$, with $\lambda^h_1 > ... > \lambda^h_N$, and the corresponding eigenvectors $(a^h_1,...,a^h_N)$. Eigenvectors represent mutually orthogonal linear combinations of the original descriptors ($X_1,\dots, X_N$), $\{a^h_n = l^h_{n1}X_1 + ... + l^h_{nN}X_N\}$ (with *h=1,...,H* and *n = 1,...N*), and each of them may be considered a new independent variable (Principal Component). Their associate eigenvalues represent the amount of total variance explained by each of the new variables. For each eigenvalue λ^h_n (with *h=1,...,H* and *n = 1,...N*), the percentage of variance explained is $(p^h_n)\% = \lambda^h_n \Big/ \sum_n \lambda^h_n * 100$. For each layer, in order to investigate the nature of the new variables $(a^h_1,...,a^h_N)$, we take into account the loading matrix $L^h_{[N,N]}$, the coordinate matrix $L^{*h}_{[N,N]}$ and the percentage weight matrix $W^h_{[N,N]}$.

The loading matrix is

$$L^h_{[N,N]} = \begin{pmatrix} l^h_{1,1} & \cdots & l^h_{1,N} \\ \cdots & l^h_{n,r} & \cdots \\ l^h_{N,1} & \cdots & l^h_{N,N} \end{pmatrix}$$

in which $l^h_{n,r}$ represents the loading of *n*-th descriptor in the *r*-th principal component (for each component $\lambda_r = \sum_j l_{i,r}$); descriptors with loading ≥ 0.5 are considered to be significant for the principal component and can give us information about the physical nature of the component.

The coordinate matrix is

$$L^{*h}_{[N,N]} = \begin{pmatrix} l^{*h}_{1,1} & \cdots & l^{*h}_{1,N} \\ \cdots & l^{*h}_{n,r} & \cdots \\ l^{*h}_{N,1} & \cdots & l^{*h}_{N,N} \end{pmatrix}$$

in which $l^{*h}_{n,r}$ represents the coordinate of *n*-th descriptor in the *r*-th principal component and it is $l^{*h}_{n,r} = \sqrt{l^h_{n,r} \Big/ \lambda^h_r}$ with $\sum_r l^{*h}_{n,r} = 1$; in this case, the component interpretation is aimed by the sign of the co-ordinates.

The percentage weight matrix is

$$W^h_{[N,N]} = \begin{pmatrix} w^h_{1,1} & \cdots & w^h_{1,N} \\ \cdots & w^h_{n,r} & \cdots \\ w^h_{N,1} & \cdots & w^h_{N,N} \end{pmatrix}$$

in which $w_{n,r}^{h}$ represents the percentage weight of *n*-th descriptor in the *r*-th principal component and it is $w_{n,r}^{h} = (l_{n,r}^{*h})^{2}\%$ with $\sum_{r} w_{n,r}^{h} = \sum_{n} w_{n,r}^{h} = 100\%$.

The *B*-matrices are used for clustering procedure. For each input matrix $B_{[M,M]}^{h}$ we group the *M* objects in homogeneous sub-groups. For the *h*-th layer, we have C^{h} clusters. The cluster test and the cluster interpretation may be carried out by means of endogenous variables (centroids). The centroids method allows relating clusters and descriptors and it may simplify the characterization and the interpretation of the grouping. For the *h*-th layer (*h = 1,…,H*) and for the *j*-th cluster (*j = 1,…,C*^{h}) of the matrix $B_{[M,M]}^{h}$, the percentage centroid is

$$(ct_{j,n}^{h})\% = \left(\frac{V_{j,n}^{h} - V_{n}^{h}}{V_{n}^{h}} \right)\%$$

in which $V_{j,n}^{h}$ is the mean values of the *n*-th descriptors (*n = 1,…,N*) calculated on the objects included in the *j*-th cluster and V_{n}^{h} is the mean values of the *n*-th descriptors (*n = 1,…,N*) calculated on all the *M* objects. In this way for the *j*-cluster we have a centroids vector $\overline{ct}_{j}^{h} = (ct_{j,1}^{h},...,ct_{j,N}^{h})$ and for each layer we have a centroid matrix

$$\begin{pmatrix} ct_{1,1}^{h} & ... & ct_{1,N}^{h} \\ ... & .ct_{j,n}^{h}.. & ... \\ ct_{C1}^{h} & ... & ct_{CN}^{h} \end{pmatrix}$$

characterizing each run of recursive clustering procedure.

In this methodological context, two indices are proposed: *N*ormalized *P*rincipal *C*omponent *I*ndex (*NPCI*) and the *C*luster *I*ndex.(*CI*) For the *h*-th layer, starting from PCA, we take into account the Q^{h} eingenvalues higher than 0.5 $(\lambda_{1}^{h},...,\lambda_{Q}^{h})$ with $\lambda_{1}^{h} > ... > \lambda_{Q}^{h}$, the corresponding percentages of explained variance $(p_{1}^{h}\%,...,p_{Q}^{h}\%)$, with $P^{h} = \sum_{q=1}^{Q} p_{q}^{h}\%$, and the corresponding eigenvectors $(a_{1}^{h},...,a_{Q}^{h})$. Starting from the reduced percentage weight matrix

$$W_{[N,Q]}^{h} = \begin{pmatrix} w_{1,1}^{h} & \cdots & w_{1,Q}^{h} \\ \cdots & w_{n,q}^{h} & \cdots \\ w_{N,1}^{h} & \cdots & w_{N,Q}^{h} \end{pmatrix}$$

in which $\sum_{n=1}^{N} w_{q,n}^{h} = 100\%$, we may calculate for the *n*-th descriptor, the *P*rincipal *C*omponent *I*ndex (*PCI*) and the corresponding normalized value (*NPCI*), following the formulas

$$PCI_{n}^{h} = \frac{1}{P^{h}} \left[1 - \left(\frac{R_{n}^{h} - 1}{Q^{h}} \right) \right] \frac{w_{n,\max}^{h} p_{n}^{*h}}{wq_{n}^{h}}$$

and

$$NPCI_n^h = \frac{PCI_n^h}{\sum_n PCI_n^h}$$

in which Q^h is the number of retained eigenvalues, P^h is the corresponding percentage of explained variance, $wq_n^h = \sum_{q=1}^{Q} w_{q,n}^h < 100\%$ is the cumulative percentage weight of the n-th descriptor in the Q^h principal components; $w_{n\max}^h$ is the maximum percentage weight $w_{n\max}^h = \max(w_{n,1}^h,...,w_{n,Q}^h)$; R_n^h is the rank of the principal component in which $(w_{n,q}^h)_{q=1}^Q$ show the maximum value and p_n^{*h} is the percentage of explained variance by this component.

NPCI is able to evaluate a standardized weight for each descriptor in a correlation structure. It represents a quantitative tool to compare the role of each descriptor in different layers and contemporaneously, the role of different descriptors in each layer. *NPCI* low values indicate descriptors with a marginal role in the correlation structure; *NPCI* high values indicate dominant descriptors.

Starting from the centroid vector calculated for the *j*-th cluster $\overline{ct}_j^h = (ct_{j,1}^h,...,ct_{j,N}^h)$ and the values of the index *NPCI* for the *N* descriptors in the *h*-th layer ($NPCI_1^h,...,NPCI_N^h$), we may calculate also the *C*luster *I*ndex as

$$CI_j^h = \sum_{n=1}^{N} (ct_{j,n}^h)(NPCI_n^h)$$

This index allows evaluating the role of different clusters in the correlation structure. In fact, the index formulation gives a greater weight to the cluster in which a dominant descriptor shows high centroid. In the analysis and in the interpretation of the underlying structure correlation, this cluster index is able to simplify the cluster identification and the cluster test.

In conclusion *NPCI* and *CI* allow to evaluate and to compare descriptors and clusters role in different correlation structures. These indices are very effective to interpret the role of different variables in scenario analysis. Particularly we suggest their application in environmental decision-making processes for sustainability polices that requires the handling of multi-dimensional and multi-scale datasets.

REFERENCES

Cosmi, C., Di Leo, S., Macchiato, M., Ragosta, M., (2008). Multivariate techniques for the analysis of partial equilibrium energy models results. *Fresenius Environmental Bulletin*, 7, 1391-1402.

Dawes, L., Goonetilleke, A., (2006). Using multivariate analysis to predict the behaviour of soils under effluent irrigation. *Water Air and Soil Pollution*, 172, 109-127.

Felipe-Sotelo, M., Gustems, L., Hernandez, I., Terrado M., Taule, R., (2006). Investigation of geographical and temporal distribution of tropospheric ozone in Catalonia (North-East Spain) during the period 2000–2004 using multivariate data analysis methods. *Atmospheric Environment*, 40, 7421–7436.

Fernandez, N., Bellas, J., Lorenzo, J.I., Beiras, R., (2008). Complementary approach to assess the environmental quality of estuarine sediments. *Water Air and Soil Pollution*, 189, 163-177.

Katahira, K., Ishitake, M., Moriwaki, H., Yamamoto, O., Fujita T., Yamazaki, H., Yoshikawa, S., (2009). Statistical analysis of metal concentrations in a sediment core to reveal influences of human activities on atmospheric environment for 200 years. *Water Air and Soil Pollution*, 204, 215-225.

Law, T., Zhang, W., Zhao, J., Arhonditsis, G.B., (2009). Structural changes in lake functioning induced from nutrient loading and climate variability. *Ecological Modelling*, 220, 979–997.

Penenko, V., Tsvetova, E., (2008). Orthogonal decomposition methods for inclusion of climatic data into environmental studies. *Ecological Modelling*, 217, 279–291.

Primpas, I., Tsirtsis, G., Karydis, M., Kokkoris, G.D., (2010). Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. *Ecological Indicators,* 10, 178–183.

Qi, Z., Dierckens, K., Defoirdt, T., Sorgeloos, P., Boon, N., Bao, Z., Bossier P., (2009). Analysis of the evolution of microbial communities associated with different cultures of rotifer strains belonging to different cryptic species of the Brachionus plicatilis. species complex. *Aquaculture*, 292, 23–29.

Ragosta, M., Caggiano, R., Macchiato, M., Sabia, S., Trippetta, S., (2008). Trace elements in daily collected aerosol: levels characterization and source identification in a four-year study. *Atmospheric Research*, 89, 206-217.

Raick, C., Beckers, J.M., Soetaert, K., Gregoire, M., (2006). Can principal component analysis be used to predict the dynamics of a strongly non-linear marine biogeochemical model? *Ecological Modelling*, 196, 345–364.

Settle, S., Goonetilleke, A., Ayoko, G.A., (2007). Determination of surrogate indicators for phosphorus and solids in urban stormwater: application of multivariate data analysis techniques. *Water Air and Soil Pollution*, 182, 149-161.

Shah, M.H., Shaheen, N., (2008). Annual and seasonal variations of trace metals in atmospheric suspended particulate matter in Islamabad, Pakistan. *Water Air and Soil Pollution*, 190, 13-25.

Solans Vila, J.P., Barbosa, P., (2010). Post-fire vegetation regrowth detection in the Deiva Marina region (Liguria-Italy) using Landsat TM and ETM+ data. *Ecological Modelling*, 221, 75–84.

Verfaillie, E., Degraer, S., Schelfaut, K., Willems, W., Van Lancker, V., (2009). A protocol for classifying ecologically relevant marine zones, a statistical approach. *Estuarine Coastal and Shelf Science*, 83, 175–185.

Zou, Y., Huang, G.H., Nie, X., (2009). Filtered stepwise clustering method for predicting fate of contaminants in groundwater remediation systems: a case study in Western Canada. *Water Air and Soil Pollution*, 199, 389-405.