VERSITA
DOI:10.2478/bile-2013-0004

Biometrical Letters
Vol. 49 (2012), No. 1, 57-66

Profile analysis of mothers susceptible to contaminant exposure in the Algarve region: Application of the HJ-BIPLOT method

A. Serafim[1], R. Company[1], B. Lopes[1], N. Silva[2], E. Castela[2], M.J. Bebianno[1], G. Castela[2]

[1]University of Algarve, CIMA, Faculty of Marine and Environmental Sciences, Campus de Gambelas, 8005-139 Faro, Portugal, aserafim@ualg.pt
[2]University of Algarve, CIEO, Faculty of Economy, Campus de Gambelas, 8005-139 Faro, Portugal

SUMMARY

The HJ-BIPLOT method developed by Galindo (1986) was applied in order to identify and categorize mothers vulnerable to environmental contamination in the Algarve region (South Portugal). The application of the BIPLOT method made it possible to recognize the most important exposure routes for contamination, showing that workplace, diet and smoking habits seem the most significant factors contributing to maternal and foetal exposure vulnerability.

Key words: HJ-BIPLOT method, contamination exposure routes, profile analysis.

1. Introduction

The BIPLOT representation (BI referring to the simultaneous representation of individuals and variables) is more informative than any other dispersion diagram and therefore widely used in the biological sciences. Furthermore, the relations between the individuals, between the variables, and between individuals and variables are easy to interpret, making this method a useful tool to understand complex multivariate data in space.

BIPLOT methods operate as a data visualization tool two basic characteristics: inner product properties that give an exact or approximate representation of individuals in space, and the equality property between the

cosine of the angle formed by two variables and the correlation coefficient between the same variables. Due to these characteristics the BIPLOT methods allow the visualization of 1) relations between the markers in columns; 2) relations between the markers in rows; and 3) the interactions between markers in rows and columns.

In this work, the HJ-BIPLOT method developed by Galindo (1986) was applied in order to identify and categorize mothers vulnerable to environmental contamination in the Algarve region (South Portugal). This study can provide significant insight into the effects of systemic contamination with problematic chemical compounds in the Algarve region and establish background levels of contaminants that can be used by physicians and scientists to determine an above-normal degree of exposure for both individuals and populations.

2. HJ-BIPLOT

The HJ-BIPLOT for a data matrix X_{nxp}, is defined as a multivariate graphic representation of markers $j_1, j_2,\ldots,j_n$ for rows $h_1, h_2,\ldots,h_p$ and for columns X, selected in such a way that both markers can be overlapped in the same reference system, with the highest quality of representation. Rows (samples) are displayed as points, while columns (variables) are displayed as vectors (Galindo, 1986).

HJ-BIPLOT is based on the singular values decomposition (SVD) of the data matrix. Any real matrix $X_{(nxp)}$ with r ($r \leq \min(n, p)$) can be decomposed as the product of three matrices such as:

$$X_{(nxp)} = U_{(nxr)}\Lambda_{(rxr)}V'_{(rxp)} \quad \text{with} \quad U'U = V'V = I_r \tag{2.1}$$

where $U_{(nxr)}$ is the eigenvectors matrix of XX', $V_{(pxr)}$ is the eigenvectors matrix of $X'X$, and $\Lambda_{(rxr)}$ is the diagonal matrix of $\lambda_1, \lambda_2, ..., \lambda_r$, corresponding to *r* eigenvalues of XX' or $X'X$.

The elements of $X_{(nxp)}$ in (2.1) are given by:

$$x_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k} u_{ik} v_{jk} \; i = 1,2,...,n, \quad j = 1,2,...,p \tag{2.2}$$

Hence, following the SVD, the selection of markers in the dimension q for the rows and columns of the matrix X are:

$$J_{(q)} = U_{(q)}\Lambda_{(q)} \quad \text{and} \quad H_{(q)} = V_{(q)}\Lambda_{(q)} \tag{2.3}$$

The quality of representation for rows and for columns in the data matrix X is the same, and rows and columns are expressed in principal coordinates.

Since rows and columns have the same representation quality, one can interpret the position of rows, of columns and the relation between rows-columns by the relative contribution of the factor to the element, and of the element to the factor (Galindo and Cuadras, 1986).

2.1. Properties of HJ-BIPLOT

i) This method allows the best simultaneous representation. Galindo (1985, 1986) and Galindo and Cuadras (1986) showed that the relations within the scatterplot are baricentric coordinates similar to those of Correspondence Factor Analysis. Hence starting with the relations $U = XV\Lambda^{-1}$ and $V = X'U\Lambda^{-1}$, the following equations are obtained:

$$J_{(q)} = U_{(q)}\Lambda_{(q)} = XV_{(q)} = XX'U_{(q)}\Lambda_{(q)}^{-1} = XH_{(q)}\Lambda_{(q)}^{-1} \tag{2.4}$$

$$H_{(q)} = V_{(q)}\Lambda_{(q)} = X'U_{(q)} = X'XV_{(q)}\Lambda_{(q)}^{-1} = X'J_{(q)}\Lambda_{(q)}^{-1} \tag{2.5}$$

i.e. the coordinates for the rows are weighted means of the coordinates for the columns, where the relative weights are the original values on the matrix X. The same holds for the coordinates of the columns.

ii) The scalar product of the columns in matrix X is coincident with the scalar product of the H markers, i.e.:

$$X'X = (U\Lambda V')'(U\Lambda V') = (V\Lambda)(V\Lambda)' = HH' \tag{2.6}$$

iii) The square of the longitude of vectors h_j is proportional to the variance of variable x_j. This means that in an HJ-BIPLOT graphic the variables with higher variability are represented by longer vectors.

iv) The cosine of the angle between two vectors h_i ,h_j represents the correlation between variables x_i and x_j. Therefore, in an HJ-BIPLOT, if two variables attain a similar classification, the vectors representing those variables form acute angles. On the other hand, if two variables are inversely correlated, their vectors form almost straight (flat) angles. If the classifications of two variables are not related to other variable, the vectors in a BIPLOT will form a right angle.

v) The scalar product of the rows of matrix X coincide with the scalar product of the markers j, i.e.:

$$XX' = (U\Lambda V')(U\Lambda V')' = (U\Lambda)(U\Lambda)' = JJ' \tag{2.7}$$

vi) The Euclidean distance between two rows in matrix X coincides with the Euclidean distance between the markers j of BIPLOT. This means that if two individuals are represented close to the factorial graphic, those individuals have similar profiles.

vii) The markers for the rows coincide with the individual coordinates in the principal component variables. This allows identification of gradients that may correspond to trends.

viii) The markers for the columns coincide with the variable coordinates in the row components. This allows identification of homogeneity gradients.

ix) If one variable (attribute tagging) has a preponderant value for a given individual, that variable point is plotted closer to the individual.

x) The higher the variability in the study, the more distant are points that represent markers in columns. The less stable attributes are represented by longer vectors.

xi) The representation quality for the rows and columns is the same, expressed by:

$$\left(\frac{\sum_{i=1}^{d}\lambda_i^2}{\sum_{i=1}^{r}\lambda_i^2}\right)*100 \qquad (2.8)$$

This means that both individual positions as well as attribute positions in the factorial plan are reliable.

2.2. Graphic interpretation of the HJ-BIPLOT method

Gabriel (1971) shows that the cosine of the angles between vectors representing variables in a BIPLOT are the correlation coefficients between the respective variables.

The property of equality between the cosine and the correlation, assuming that the data are centred, implies that for each *x* and *y* plotted, taking the cosine q_{xy}, the angle between two variables *x* and *y* is equal to its correlation r_{xy}, according to the following equation:

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \frac{x_\bullet y}{|x||y|} = \cos\Theta_{xy} \qquad (2.9)$$

In the HJ-BIPLOT, if two attributes have similar classification, the vectors form an acute angle (e.g. a 35º angle). In this case, there is a positive correlation between attributes. If one attribute is inversely correlated to another, the vectors will form flat or straight angles (e.g. 180º) and there is a negative correlation between attributes. If the classification of a given attribute has no relation with another attribute, the vectors will form a right angle (90º) and the correlation between attributes is null.

Also, the greater the distance of the individual projection on a variable (measured from the center of the graph axis), the more preponderant is the variable in explaining the response of the individual.

In the HJ-BIPLOT method, if one variable is preponderant for one individual, the dot corresponding to that attribute will be displayed closer to the dot corresponding to the individual. Since in this method both individuals and

variables are represented in the same scale, it is possible to interpret the distances between individuals and variables as the predominance of a variable in explaining the individual, or conversely the contribution of an individual to the variable data.

3. Application

A cross-sectional study was conducted in one of the main public hospitals in South Portugal (Algarve region), serving a population of approximately 253,000. The study included eligible women hospitalized for childbirth at the District Hospital of Faro between 2010-11-16 and 2011-08-04. Only women who had been residents in the area for a minimum of ten years were recruited for the study. A total of 109 women agreed to participate in the study. After obtaining the patient's consent, a detailed questionnaire with 62 questions was collected by a trained interviewer. The questionnaire included among others the following information: general personal data (age, place of birth); general health status and data on the last pregnancy and delivery (parity, previous abortions or miscarriages, twins); data on occupational and/or environmental sources of contaminants (e.g. professional history – education, position, present employment); potential sources of recent environmental exposure (place of residence, vicinity of factory, smelter, highway, street, road, large crossroad, or bus terminal with heavy traffic); dietary history (self-identification of the predominant diet type; consumption of fish, canned foods, etc); alcohol consumption; smoking habit (self-classification as active smoker, former smoker, or non-smoker); active smoking: data on number of cigarettes, smoking during pregnancy; former smoking, passive smoking, data on active smokers in the family and at the workplace. The study was approved by the Ethics Committee of the Hospital of Faro, and all patient information was coded to maintain confidentiality.

Multivariate statistical analysis was performed using the software package MULTBIPLOT (MULTivariate Analysis using BIPLOT) (Vicente-Villardón, 2010) and was applied to transformed data (data divided by column means and

centre). Cluster analysis was performed using hierarchical cluster with the Euclidean distance using the HJ-BIPLOT (principal components) scores and Ward's method as the process of linkage. A value of $p<0.05$ was considered statistically significant in all statistical analyses.

Out of all the study variables, only 21 were considered active variables, i.e. with statistically significant variance (Table I), able to find a spatial structure standardized to the rows. In this configuration four clusters were detected using the Ward method (Figure 1).

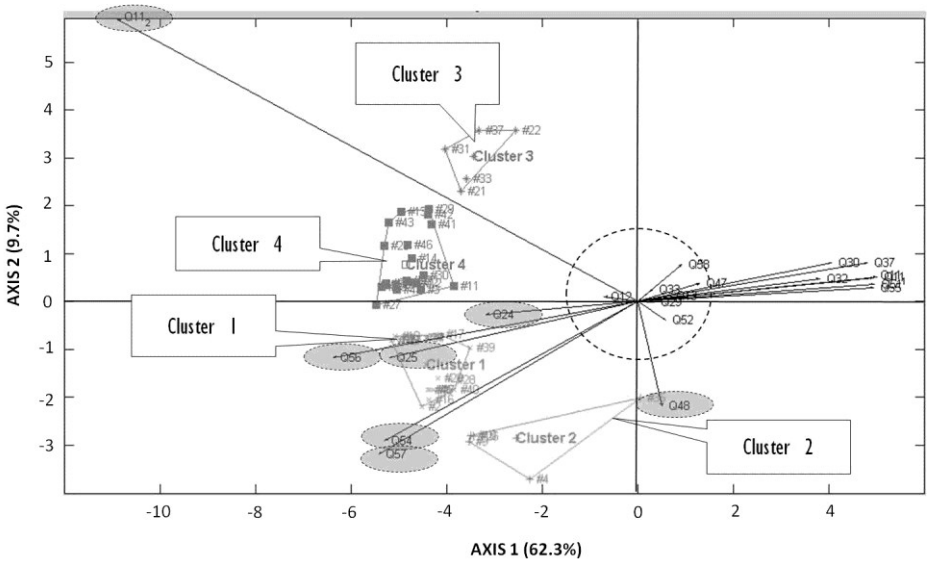

Figure 1. HJ-BIPLOT representation showing the relationships among the individuals and the active variables that determine the group structure

The first cluster is characterized by the questions Q24, Q25, Q54, Q56 and Q57 (Figure 1), and shows that the active variables that differentiate this cluster are related to education and dietary habits. This group comprises primiparous women with academic degrees, no record of smoking, low-consumption of canned foods and dried fruits, and consumption of bottled mineral water over tap water (Table 1).

Table 1. Significant variables considered in the application of the HJ-BIPLOT method

	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
Q11=Age (years)	Mean = 32 ± 5.7	Mean = 31 ± 5.2	Mean = 29 ± 4.8	Mean = 28±2.3
Q11.1=Local	Mode=1 (City)	Mode= 1 (City)	Mode= 1 (City)	Mode= 2 (Rural)
Q11.2= Exposure in workplace	Mode = 1 (Yes)	Mode = 1 (Yes)	Mode = 2 (No)	Mode =1 + 2
Q12= Height (cm)	Mean = 164±4.1	Mean = 164±5.6	Mean = 164±5.5	Mean = 162±4.5
Q13=Weight before pregnancy (Kg)	Mean = 61.7±6.7	Mean = 64.5±11.5	Mean = 64.7±11.9	Mean = 68.7±10.9
Q24= Education	Mode = 5 (University)	Mode = 4 (Secondary)	Mode = 4	Mode = 4+5
Q25= Previous pregnancies	Mode = 2 (No)	Mode = 1 (Yes)	Mode = 1 (Yes)	Mode =1 + 2
Q29= Dental restoration during pregnancy	Mode = 2 (No)	Mode = 2 (No)	Mode = 2 (No)	Mode = 1+2
Q30= Hair dyes during pregnancy	Mode = 1 (Yes)	Mode = 2 (No)	Mode = 1 + 2	Mode =1 + 2
Q32= Type of house	Mode = 2 (apartment)	Mode = 2	Mode = 1 (independent house)	Mode = 1+2
Q33= Age of house	Mode = 1 (< 5 years)	Mode = 2 (5-14 years)	Mode = 4 (>29 years)	Mode = 4+2
Q37= House repairs in the last 6 months	Mode = 2 (No)	Mode = 2 (No)	Mode = 1+2	Mode = 2 (No)
Q47= Passive smoker	Mode = 1 (Yes)	Mode = 1 (Yes)	Mode = 1+2	Mode = 1+2
Q48= Smoker or ex-smoker	Mode = 2 (No)	Mode = 1 (Yes)	Mode = 1+2	Mode = 1+2
Q51= Type of contraceptives	Mode = 1 (hormonal)	Mode = 2 (non-hormonal)	Mode = 1+2	Mode = 1+2
Q52= Medication used during pregnancy	Mode = 1 (Yes)	Mode = 1 (Yes)	Mode = 1+2	Mode = 1+2
Q54= Main water consumption	Mode = 1 (Bottled)	Mode = 1+2	Mode = 1+2	Mode = 2 (Tap)
Q55= Fresh vegetables*	Mode = 5	Mode = 4	Mode = 3	Mode = 3
Q56=Canned vegetables*	Mode = 1	Mode = 2	Mode = 3	Mode = 3
Q57=Canned fruit*	Mode = 1	Mode = 2	Mode = 3	Mode = 2
Q58= Dried fruits*	Mode = 2 + 3	Mode = 1	Mode = 1	Mode = 1

* mode 1: (never); mode 2: (1–3 times/month); mode 3: (1–3 times/week); mode 4: (4–6 times/week); mode 5: (every day)

On the other hand, cluster 2 is characterized by the active variables (Q48) related to smoking habits. The women in cluster 2 are active or ex-smokers, with secondary school education; the majority live in urban areas in relatively older houses (15 to 30 years) and have had more than one child (Table 1).

Finally, clusters 3 and 4 are mainly characterized by the Q11 variable, related to exposure to contaminants in the workplace. This group comprises women with a wide range of secondary education level, parity and smoking habits (Table 1).

The variables plotted in the centre of the spatial representation (Figure 1), although they are active variables, do not have enough variability to characterize the clusters; this applies as well to the variables plotted in the right axis (1st and 4th quadrant).

4. Conclusion

This study shows that application of the multivariate analysis HJ-BIPLOT makes it possible to discriminate different clusters that represent mothers with distinct susceptibility to environmental contaminants. Workplace, diet and smoking habits seem the most significant factors contributing to maternal and foetal exposures. Even in non-industrialized areas like the South of Portugal there is a need to implement measures to eliminate or minimize the risk of contaminant exposure during pregnancy.

REFERENCES

Gabriel K.R. (1971): The BIPLOT display of matrices with application to principal components analysis. Biometrika 58: 453–467.

Galindo M.P. (1985): Contribuciones a la Representación Simultánea de Datos Multidimensionales. Tesis Doctoral. Universidad de Salamanca.

Galindo M.P. (1986): Una alternativa de representación simultánea: HJ-BIPLOT. Questíio 10 (1): 13–23.

Galindo P.E., Cuadras C. (1986): Una extensión del método Biplot y su relación con otras técnicas. Publicaciones de Bioestadística y Biomatemática. Universidad de Barcelona 17.

Vicente-Villardón J.L. (2010): MULTBIPLOT: A package for Multivariate Analysis using Biplots. Departamento de Estadística. Universidad de Salamanca. (http://biplot.usal.es/multbiplot).