Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1679-9291
ISSN on-line: 1807-8648
Doi: 10.4025/actascihealthsci.v34ispec.10055.

Viability of using multidimensional graphics in epidemiological data analysis: case study with respiratory diseases

Elisa Norberto Ferreira Santos[1*] and Marcelo Ângelo Cirillo[2]

[1]*Instituto Federal do Triângulo Mineiro, Rua João Batista Ribeiro, 4000, 38064-900, Campus Uberaba, Uberaba, Minas Gerais, Brazil.*
[2]*Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil. *Author for correspondence: E-mail: norelisa@hotmail.com*

ABSTRACT. The statistical methodology to be used in epidemiological data analysis is extremely important for obtaining reliable and plausible results that can be interpreted in the epidemiological context. In order to carry out a case study about the incidence of the main respiratory diseases in some cities with different climatic seasons, the present work aimed at studying the viability for applying the multidimensional graphic technique known as "h-plot" to identify the main variables that resulted in a higher contribution to the data dispersal. In accordance with the results obtained and discussed, we state that the h-plot graphics are viable to be applied as an alternative method as regards to the identification and collection of the variables.

Keywords: clusters of variables, covariance matrix, h-plot.

Viabilidade do uso gráfico multidimensional para análise de dados epidemiológicos: estudo de caso com doenças respiratórias

RESUMO. A metodologia estatística a ser utilizada na análise dos dados epidemiológicos é de suma importância para que os resultados obtidos sejam confiáveis e plausíveis de serem interpretados no contexto epidemiológico. Com o propósito de realizar um estudo de caso sobre a incidência das principais doenças respiratórias em algumas cidades com estações climáticas bem diferenciadas, o presente trabalho tem por objetivo estudar a viabilidade da aplicação da técnica gráfica multidimensional conhecida por "h-plot" para identificar principais variáveis que resultaram em maior contribuição na dispersão dos dados. Em consonância com os resultados obtidos e discutidos neste trabalho, recomendou-se que os gráficos h-plots são viáveis de serem aplicados como um método alternativo no que se refere à identificação e agrupamento das variáveis.

Palavras-chave: agrupamentos de variáveis, matriz de covariância, h-plot.

Introduction

Conducting an accurate epidemiological survey or a research in health area depends on the statistical methodology employed in collection, analysis and interpretation of data. Other factors like the lack of consistent epidemiological data contribute to the planning and execution of preventive measures is made with better efficiency in health-related issues.

Specifically in the interpretation of results, Godoy et al. (2001) warn that caution is needed for two reasons: (a) the underreporting is still high in several regions; and (b) certain diagnoses provide better remuneration to the hospital.

In this way, the epidemiological research along with the search for statistical methods that promote greater ease in performing analyses and interpretation of results become a wide field for research, the most common statistical methods are elaborated by means of established assumptions, such as normality of data or homoscedasticity of population variances. If we consider a study that encompasses the Brazilian regions, certainly the data will present heterogeneous variability.

In a comparative study, in the different Brazilian regions, performed by Chiesa et al. (2008), which affirm that respiratory problems are the second cause of death in South and Southeast regions, and the third cause in the other regions. In São Paulo State, the records of mortality from acute respiratory infections (colds, ear infection, sinusitis, tonsillitis, lower respiratory tract problems, epiglottitis, bronchitis, and pneumonia) are also significant, occupying the third place in pediatric population. Despite being the state with the best survival rates for children from zero to six years, according to a study of the *Instituto Brasileiro de Geografia e Estatística* (IBGE) and *Fundo das Nações Unidas para a Infância* (Unicef).

A real situation and easy to understand this heterogeneity in the data can be exemplified by studies on respiratory diseases, since in addition to air pollution, inherent to each region, several reasons can be attributed to data heterogeneity. Prietsch et al. (2003) indicated the exposure to household pollutants, and environmental attributes such as: type of flooring used in the residence, pets, passive smoking, in short many variables related to climatic variations and social conditions may undermine the interpretation of the results when they are derived from a statistical method based on the assumption of variance homogeneity. Evidently, statistical methods in literature are proposed for cases in which the variance homogeneity is detected. Among them, the transformation of the data. Nevertheless, the choice to transform the data may difficult the interpretation of the results, since the original scale is changed and the residuals obtained between the difference of the covariance matrix and the principal coordinate matrix from the two first singular values. In this context, the goodness of fit criteria, in particular the value of the coefficient R^2 is influenced.

Due to all above mentioned, the comparative study related to the incidence of the main respiratory diseases performed in this study, was made with the purpose to add information in the epidemiological and statistical areas. In an epidemiological context, this contribution verifies the existence of some heterogeneity in the incidence of spread of these diseases, based on information of the different cities with levels of industrialization and seasons well defined as a case study. The contribution for statistical area is that the use of h-plot technique (CORSTEN; GABRIEL, 1976) when applied to multidimensional data identifies the variables responsible for this supposed heterogeneity.

In this way, the goal of this study was to investigate the feasibility of applying a multidimensional graphical technique (h-plot) as an alternative method to identify which variable(s) resulted in a greater contribution to the dispersion of data relative to the records of hospitalizations for some respiratory diseases. With this purpose we considered a sample as a case study, in which the observations were collected from a database of the health care system, in different cities with well differentiated seasons.

Material and methods

Acquisition of data

For this study, it was used the records of hospitalizations (Table 1) in the period from 2003 to 2008, referent to the incidence of respiratory diseases: pneumonia, bronchitis, asthma and other diseases, in the cities of São Paulo, Curitiba, Recife and Brasília. These data were provided by the Ministry of Health that contain information of all hospital admissions made through the Hospital Admission Authorizations (AIH) of the Unified National Health System (SUS).

Table 1. Records of hospitalizations for respiratory diseases (pneumonia, bronchitis, asthma and other diseases) in the period between 2003 and 2008, in the cities of Curitiba, Recife, Brasília and São Paulo.

Cities	Period	Pneumonia	Bronchitis	Asthma	Others
Curitiba	2003	5,478	1,652	955	2,273
	2004	5,299	1,182	934	2,175
	2005	4,900	764	825	1,937
	2006	4,604	772	586	2,014
	2007	4,832	782	677	2,284
	2008	5,200	591	473	1,954
Recife	2003	7,826	748	4,929	2,313
	2004	8,199	683	3,888	2,278
	2005	7,682	702	3,689	2,565
	2006	7,737	774	3,993	2,856
	2007	8,352	830	4,042	3,274
	2008	6,818	1,499	2,717	2,039
Brasília	2003	8,170	1,166	3,162	1,564
	2004	10,520	1,165	3,485	2,023
	2005	9,015	1,025	3,273	1,875
	2006	9,811	993	2,786	1,932
	2007	9,999	1,170	2,361	1,983
	2008	9,782	862	1,983	1,722
São Paulo	2003	26,616	2,888	8,241	8,718
	2004	26,821	2,953	8,141	8,513
	2005	26,499	2,928	8,170	8,796
	2006	28,061	2,882	7,956	8,320
	2007	28,978	3,130	8,315	9,076
	2008	27,152	2,756	5,674	6,357

Constructing the h-plot

With this information, following the methodology proposed by Corsten and Gabriel (1976), we construct the graphs so-called 'h-plots' according to the steps:

1º - For each city, it was estimated the covariance matrix relative to the k-th city S_k (k=1, ... 4).

2º - Once fixed a covariance matrix S_k, it was computed the first two eigenvalues λ_1^2 and λ_2^2 and the respective normalized eigenvectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$.

3º - Then the coordinates were obtained in order to represent the vectors (diseases) in the two-dimensional graph. These coordinates were specified in matrix H (1), given by:

$$\boldsymbol{H}_{(p\times 2)} = (\lambda_1 \boldsymbol{q}_1, \lambda_2 \boldsymbol{q}_2) \quad (1)$$

where,

p is the number of variables, i.e., equal to the number of respiratory diseases, and each row, represented by the vector $\boldsymbol{h}_i$ refers to the coordinates used to represent the vector (disease) in the graph. The plot of this vector was provided by a line

segment between the origin (0;0) and the point (h_{i1}, h_{i2}) $(i = 1,...,p)$.

4º - After plotting all the vectors, it was verified the goodness of fit $\left(R^2\right)$ in the two-dimensional graph (2). Where,

$$R^2 = 1 - \left\| \boldsymbol{S_k} - \boldsymbol{H_k}\boldsymbol{H_k^t} \right\|^2 \Big/ \left\| \boldsymbol{S_k} \right\|^2 = (\lambda_1^4 + \lambda_2^4) \Big/ \sum_{i=1}^{p} \lambda_i^4 \quad (2)$$

λ_i^4 $(i = 1,...,p)$ referred to each eigenvalue of S_k, and $\left\| \boldsymbol{S_k} \right\|^2 = traço(\boldsymbol{S_k^t}\boldsymbol{S_k})$ (GABRIEL, 1971).

For this, we used the software R version 2.13.0 (R DEVELOPMENT CORE TEAM, 2011), to estimate the eigenvalues and eigenvectors, as well as for obtaining the matrix $\boldsymbol{H}$ and determine the goodness of fit.

Interpreting the graph h-plot

After building up the graph h-plot for each city, according to previous procedure, the interpretation of these graphs is made by assuming that the covariances matrices S_k and $S_{k'}$ (k=1,...,4 and k'≠ k) are mutually different. Nevertheless, with the purpose to answer which variable (disease) contributes more markedly for this differentiation, emerges the interpretation of the vectors plotted in the graph h-plot. In this way, Corsten and Gabriel (1976) suggested that the greater the angle between any two vectors, better the indication that these variables contributed to the distinction of the covariance matrices. Particular cases, pointing equality between these matrices, may occur; an example can be given when observing that the vectors of both populations are plotted in the same quadrant.

Results and discussion

Prior the construction of h-plots, considering the multivariate information represented by the number of hospitalizations due to respiratory diseases (Table 1) for each city, we estimated the covariance matrix. The results are described in Table 2.

The construction of the h-plots could be done using the correlation matrix, but in this study we chose to use the covariance matrix because the data relative to the records of hospitalizations (Table 1) are in the same scale and unit of measure. In this way, the coordinates were calculated (x, y). Each coordinate allowed representing the respiratory diseases of each city, whose representation is characterized by the vectors illustrated in Figure 1.

Table 2. Covariance matrix and sample correlation for the number of hospitalizations referent to respiratory diseases in the cities of Curitiba, Recife, Brasília and São Paulo.

	Covariance Matrix			
Curitiba	107,319.37	89,638.57	33,984.47	18,686.57
	89,638.57	15,394.97	62,781.07	41,141.77
	33,984.47	62,781.07	38,008.67	15,692.47
	18,686.57	41,141.77	15,692.47	24,892.57
Recife	288,206.40	-138,235.00	255,338.60	153,297.40
	-138,235.00	96,898.27	-169,236.47	-60,512.73
	255,338.60	-169,236.47	505,689.44	98,519.33
	153,297.40	-60,512.73	98,519.33	201,841.38
Brasília	691,013.90	-7,945.50	-95,669.80	120,107.50
	-7,945.50	15,841.10	41,456.20	4,291.50
	-95,669.80	41,456.20	334,797.47	14,376.13
	120,107.50	4,291.50	14,376.13	30,629.37
São Paulo	691,013.90	23,644.50	-139,438.50	-146,546.50
	23,644.50	14,913.77	90,369.10	95,705.93
	-139,438.50	90,369.10	1,048,358.50	994,611.80
	-146,546.50	95,705.93	994,611.80	968,773.47

As regard to the quality of this representation into a two-dimensional space, the summarization of the total variability in the two first components led to a loss of information of about 1%. This fact was observed for all cities studied, according to the results of the goodness of fit (R^2) found in Table 3.

Table 3. Coordinates for the h-plot about respiratory diseases in the cities of Curitiba, São Paulo, Brasília and Recife, from 2003 to 2008.

Respiratory diseases	Curitiba		São Paulo		Brasília		Recife	
Pneumonia (x_1)	277.0	174.1	253.4	939.8	822.7	-117.7	277.0	174.1
Bronchitis (x_2)	379.8	-85.8	99.5	47.3	19.5	-71.8	379.8	-85.7
Asthma (x_3)	159.4	-60.0	1009.3	-155.1	194.4	-544.7	159.4	-60.0
Others (x_4)	100.6	-60.6	977.4	-88.3	136.5	-76.6	100.6	-60.6
R^2	99.60%		99.99%		99.98%		99.29%	

Evidently, a high explanation of total variability is given by the information of the number of variables and the number of components extracted, i.e., in all cities there are four variables and two extracted components. In other situations where the number of variables is relatively high, for the construction of the h-plots, the researcher will focus on only three components, due to the representativeness and geometric interpretation of the plans generated by the vectors.

For didatic purposes, the Figure 1 shows how to interpret the vectors in the graph h-plot. Note that the vectors x_1 and x_2 are correlated, as well as x_3 and x_4, but once the vector x_3 has greater amplitude in relation to the vector x_4 it cannot be overlapped on x_4, so, there is an indication that the vectors do not differ, moreover the specific values of the contributions of both vectors in relation to the estimated components are well distinct.

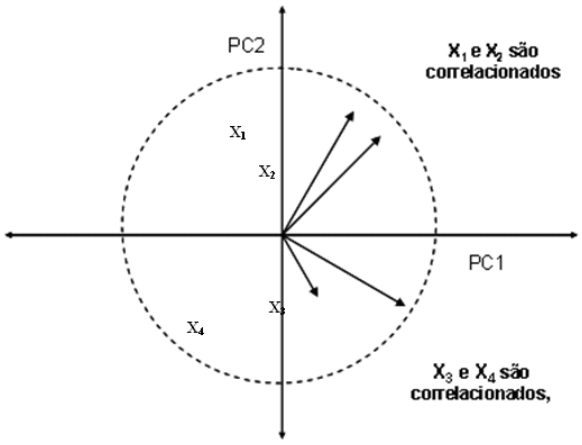

Figure 1. Ilustração da interpretação dos vetores do gráfico h-plot.

It is important to emphasize that the interest in interpreting the h-plot is to analyze which vectors contribute to the differentiation of the covariance matrices. In this study it corresponds to show evidences of which variables (diseases) are responsible for the heterogeneity in the number of hospital admissions specific to these respiratory diseases in the studied cities.

By analyzing the vectors illustrated in Figure 2, when comparing the results of Curitiba and Recife, there are statistical evidences to state that the vectors that represent the variable pneumonia, in both cities, resulted in similar contributions in relation to the formation of the two components. Nevertheless, when analyzing the vectors referent to the other diseases and asthma, in both cities the results were more discrepant, so that for the disease Asthma the vectors are correlated but with quite different contributions to the formation of the components. On the other hand, the results for the other diseases allowed detecting that the vectors were not correlated and with distinct contributions, due to the different amplitudes observed in each vector, therefore, a preliminary statistical evidence to affirm that supposedly the records of other respiratory diseases with lower frequency in relation to the pneumonia, asthma and bronchitis were responsible for causing heterogeneity in the incidence of hospitalizations in the cities of Recife and Curitiba. However both cities, Recife and Curitiba, present well differentiated seasons, but epidemiological or climatic causes that could justify these results cannot be detected in the h-plot analysis, since are attributed to several other factors. As an example, we can cite Prietsch et al. (2002) that studied the prevalence of acute diseases in the lower respiratory tract and the influence of the factors related to the housing conditions and maternal smoking. The authors reported that the general prevalence of the acute respiratory disease was 23.9%, and as consequence, they suggested the implementation of specific programs to control acute respiratory diseases in the studied population.

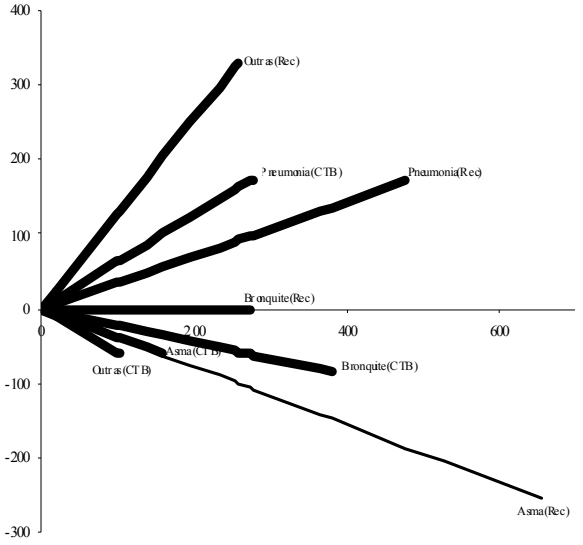

Figure 2. H-plot of the covariances of respiratory diseases in Curitiba (CTB) and Recife (Rec) (2003 to 2008).

Through the Figure 3, by comparing the vectors representing the diseases in the cities of São Paulo and Brasília that in general it had contributions well differentiated in relation to the principal components estimated, based on the respective covariance matrices (Table 2). It can be noted that the vector that represents the records of hospitalizations for Asthma, in both cities, evidenced that the cases of hospitalizations were not correlated. The same result was found for Pneumonia and Bronchitis. Since these cities are characterized by high levels of pollution, the lack of correlation apparently counteracts the statement of Rezende et al.(2009) that reports that in big cities these diseases are increasingly common, especially as a function of the air pollution. Carbon monoxide and carbon dioxide are pollutant gases derived from burning fossil fuels (gasoline and diesel) and are very harmful to the human respiratory tract. The inhalation of these gases may cause the onset of some of these diseases.

In this way, the exploratory results evidenced by the construction of the graphs h-plots are of paramount importance to identify trends and possible groupings. Thus the use of this technique along with other statistical methodologies will

provide additional results that facilitate the interpretation of the results. Regarding the inference and modeling, this technique can be used as a preliminary study on data covering seasonality as for example the study performed by Francisco et al. (2003), which considered the analysis of temporal trend by means of scatterplots, whose results allowed observing the relationship between death rates for respiratory disease and the years of study on elderly population in the state of São Paulo, between the years 1980 and 1998.

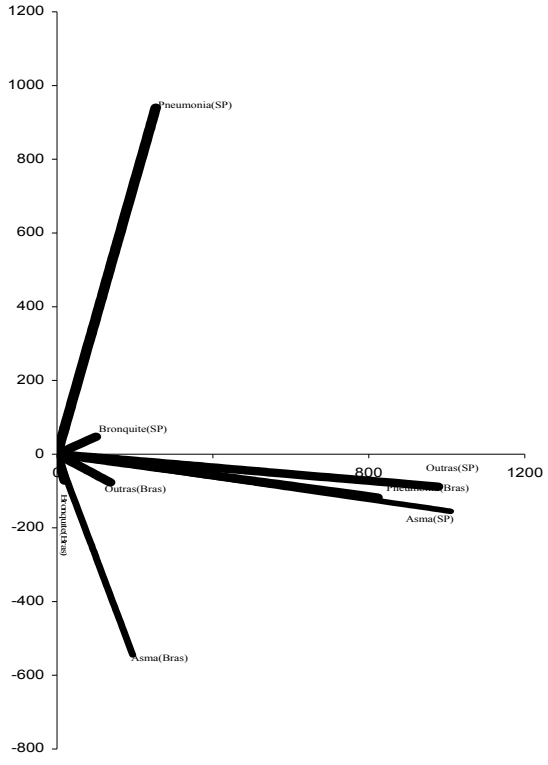

Figure 3. H-plot of the covariances of respiratory diseases in Brasília (Bras) and São Paulo (SP) (2003 to 2008).

In order to investigate the effects caused by air pollution on morbidity for respiratory diseases in children between 1999 and 2000, Bakonyi et al. (2004) used the Poisson generalized additive model and concluded that the air pollution promotes adverse effect for children's health, even when the pollutants are below that required by law.

Conclusion

Through the results obtained and discussed in this study, we recommend the construction of h-plots as a sorting method to identify and/or group the samples for later inference, using a confirmatory statistics.

When addressing the results related to the incidence of respiratory diseases, since it is a case study, the conclusions are limited to affirm that for the cities of Recife and Curitiba, the records of hospitalizations due to other respiratory diseases with lower frequency, in relation to pneumonia, asthma, and bronchitis, represented by the variable Others, resulted in a greater contribution to the heterogeneity in the data of hospitalizations incidence. Considering the cities of São Paulo and Brasília, the records of hospitalizations referent to Asthma, Pneumonia and Bronchitis did not evidence a possible correlation.

Acknowledgements

This study was supported by the Minas Gerais State Research Foundation (Fapemig).

References

BAKONYI, C. M. S.; DANNI-OLIVEIRA, M. I.; MARTINS, C. L.; BRAGA, F. L. A. Poluição atmosférica e doenças respiratórias em crianças na cidade de Curitiba, PR. **Revista de Saúde Pública**, v. 38, n. 5, p. 695-700, 2004.

CHIESA, A. M.; WESTPHAL, M. F.; AKERMAN, M. Doenças respiratórias agudas: um estudo das desigualdades em saúde. **Caderno de Saúde Pública**, v. 24, n. 1, p.55-69, 2008.

CORSTEN, L. C. A.; GABRIEL, K. R. Graphical exploration in comparing variance matrices. **Biometrics**, v. 32, n. 4, p. 851-863, 1976.

FRANCISCO, B. S. M. P.; DONALISIO, C. R. M.; LATTORRE, O. D. R. M. Tendência da mortalidade por doenças respiratórias em idosos do Estado de São Paulo, 1980 a 1998. **Revista de Saúde Pública**, v. 37, n. 2, p. 191-196, 2003.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, v. 58, n. 3, p. 453-467, 1971.

GODOY, D. V.; ZOTTO, C. D.; BELLICANTA, J.; WESCHENFELDER, R. F.; NACIF, S. B. Doenças respiratórias como causa de internações hospitalares de pacientes do Sistema Único de Saúde num serviço terciário de clínica médica na região nordeste do Rio Grande do Sul. **Jornal de Pneumologia**, v. 27, n. 4,p. 193-198, 2001.

PRIETSCH, O. M. S.; FISCHER, B. G.; CESAR, J.; FABRIS, R. A.; MEHANNA, H.; FERREIRA, H. P. T.; SCHEIFER, A. L. Doença aguda das vias aéreas inferiores em menores de cinco anos: influência do ambiente doméstico e do tabagismo materno. **Jornal de Pediatria**, v. 78, n. 5, p. 415-422, 2002.

PRIETSCH, S.; FISCHER, G.; CESAR, J.; LEMPEK, B.; BARBOSA, L.; DIAS, L.Z.; ZOGBI, L.; CARDOSO, O.; SANTOS, A. Doença respiratória em menores de 5 anos no sul do Brasil: influência do ambiente doméstico. **Revista Panamericana de Salud Pública**, v. 13, n. 5, p. 303-310, 2003.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2011.

REZENDE, L. G.; GRANJEIRO, R. C.; FURTADO, P. L.; PINHEIRO, G. B.; NAKANISHI, M. Is Dry Climate Related to Hospital Admission for Epistaxis? **International Archives of Otorhinolaryngol**, v. 13, n. 2, p. 172-177, 2009.

Received on April 30, 2010.
Accepted on June 15, 2011.