Comparative Studies in Brazilian Metropolitan Cities regarding the National Consumer's Prices Index: a Multivariate Statistical Analysis from 2012 to 2014

Bruna Cristina Freitas Costa
Universidade Federal de Uberlândia, Uberlândia, Brazil brunacosta.ufu@gmail.com

Priscila Neves Faria
Universidade Federal de Uberlândia, Uberlândia, Brazil priscila@famat.ufu.br

Isabella Pimenta Rossi
Universidade Federal de Uberlândia, Uberlândia, Brazil- isabella.prossi@gmail.com

Summary

Studies on inflation intended to indicate the average effect on the economy of a complex phenomenon: the rise prices of various assets that comprise it. Specifically, inflation can be constructed from different baskets of goods as the prices relevant for groups that you wish to consider- Consumers, producers an industry or the economy in general, residents of an specific region. This study had objective to apply the exploratory analysis of multivariate datas in order to analyze inflation, but focused on grouping of metropolitan cities, aiming to contribute to research involving administrative power such as dealing with the measurement of inflation of a set of products and services sold at retail by the population of this cities. As a result, were obtained four groups of similar regions, considering the variables involved in the study, these are: food and drinks, housing, residence items, clothes, transport, health and personal care, personal expenses, education and communication. The metropolitan region of Curitiba remained isolate of the rest.

Keywords: inflação, hábitos de consumo, *cluster analysis.*

1. Introduction

Nowadays, researchers have tried to analyze data not only quantitatively, but also qualitatively, adapting the statistically based techniques, aiming a more objective case study analysis.

The multivariate technique of the grouping analysis (Cluster Analysis CA) is a technique which has as primary objective to discover the natural grouping of the variables, which is made based on the similarities or dissimilarities (characterized by different forms of distance calculus). It is a way of obtaining uniform groups, uniting the data in case in a determined number of groups, in a way that there is a great uniformity within each group, but heterogeneity between them (Jonhsone Wichern, 1992; Cruz e Carneiro, 2003).

In this way, it is possible to construct the grouping in the samples according to their similarities, using all the available variables, and then representing them in a two-dimensional way through a dendrogram graph.

Today, a lot of different dissimilarity measurements are theoretically suggested, mainly due to the great development and use of the multivariate techniques (Khattreee Naik, 2000).
possible that a considerable simplification of the original information occurs in the formation of the dendrogram, and then, generating some pattern distortions regarding dissimilarities in the elements of the study. So, it is necessary to judge the fitting of the results, what consists in the final phase of the grouping process.

2. National Index of Consumer Prices Amplo-15 (IPCA-15)

s Prices Index (ANIPC) was created in 1978, and it consists on a combination of processes for the production of price indexes to the consumer, from the

aggregation of regional results. From May 2000, the National Index of Consumer Prices Amplo 15 (IPCA-15) began to be produced and available by the IBGE. The IPCA-15 is the IPCA itself, however, considering another period of price gathering, from the 16th of the previous month to the 15th of the current month.

The IPCA-15 p of goods or services that are commercialized in retail, referent to the personal consumption of the family. The target market is represented by the resident families of urban areas whose income varies from 1 to 40 minimum wages, independently on the source of it.

The IPCA-15 covers nine metropolitan regions: Rio de Janeiro, Porto Alegre, Belo Horizonte, Recife, São Paulo, Belém, Fortaleza, Salvador, Curitiba, Brasília and Goiânia. This study the last two regions mentioned.

Generally, the methodology scope of the National Index of Consumer Prices covers the following subjects: the assembling of the general structure of weight, the definition of the subscription bases of goods, price gathering e calculus methodology.

The pondering structures are assembled by the use of a grouping code organization that was logically established in a way that the similar consumption categories sticks together, hierarchically structured in groups, subgroups, items and sub items. These last mentioned represent the most disaggregated level for which we obtain the weights that are used in the calculus of the price index. These ponderings portrait the importance and representativeness of the sub items that belong to the family consumption package, that are established from the consumption habits of the target market.

The data gathering is done from the definition of the subscription the informants and goods, following the gathering method. Two types of procedures are used in the subscription generation, according to the nature of different researched goods. The main type consists in weighing the informants through the Research of the Point of Purchase PCL, which defines where the data should be collected. The second type employs specific procedures for the sub items which peculiarities require so the so called extra-PCL sub items for which the PCL methodology is not adequate. Then special treatment is requires do determine where are the points of data gathering.

Some example of extra-PCL sub items are dwelling rent, domestic workers, public services, taxes and so on.

So, the subscription of informants is essentially formed by commercial establishments, rented dwells, dealership companies that are responsible for services, official organs and independent professionals like doctors and dentists.

To define the group of goods that compose the subscription, we consider representativeness of the goods that the population consumes. Then, the subscription is generated from the relation of the sub items that belong to the weight structure of each income zone. After that the Research of Goods and Services Specification PEPS is done, so the study has the base for the definition of the goods subscription, characterizing the specification levels used in the price gathering.

All the
Governo Federal (http://dados.gov.br/tag/IBGE)
without license restriction, patents ou control mechanism.

The variables that are going to be used in the present study are: Foods and Beverages, Housing, Residency Articles, Clothing, Transportation, Health e Personal Care, Personal Expenses, Education and Comunication.

For each obtained variable, a descriptive statistical study was done, based on the arithmetic mean calculus, standard deviation, variation coefficient, and graph representation of the data. Afterwards the data will be analyzed in face of the grouping analysis.

3. Methodology

The multivariate technique of grouping analysis (Cluster Analysis CA) behaves as a technique of a dependent individual, in which the distance values, on the form of matrixes, are arranged. The fraction of the data group, of unities of observation or subgroup cases or uniform groups

is the goal of this analysis, defining then, a more intense uniformity within the subgroup of the biggest heterogeneity in relation to other subgroups (Mardia, Kent and Bibby, 1995). The estimation of the parameter is not required, in this case, which ratifies the non-statistical character (Chatield e Collins, 1986).

Techniques such as this have an advantage related to their capacity to reproduce a multidimentional space to a distance measuring between objects, representing this in a two-dimensional space, that is a lot simpler than a multidimensional one (Cormack, 1971; Mardia, Kent e Bibby, 1995).

The first step of the CA is the conversion of the data matrix n x p, in another n x p matrix of similarity and dissimilarity measurements, measured in relation to the pairs of n sample unities, in

of studied elements, data is presented in a symmetrical matrix (n x n), and from this point, the visualization and interpretations of the distances can be facilitated by the use of the grouping method and/or graph dispersion.

The distances are measurements that are used to represent the points in the structure of the similarity. This measurement represents the smallest space between two points, as an extension of the Pythagorean Theorem for multidimensional cases.

When the variables are defined, it is expected that they present an equivalent contribution on the CA, so the distance between th
unities of distinct measurements in the same group of data, in a way that the variables present similar discriminatory power, not based in the value amplitude.

So, if all the data is in the same pattern of measurement, the variability of each characteristic will be uniform or almost uniform. So in this case, it is possible to use the original data. However, if this does not occur, the measurement of the variables suggests a data transformation.

Taking $Y_{ij,}$ the observation of the i-th individual (clone, cultivate, linage and etc) for the j-th

$$d_{ii'} = \sqrt{\sum_{j} \left(Y_{ij} - Y_{i'j} \right)^2}$$

The Euclidian distance does not preserve the distance order with the scale change, so, It is common to do the standardization of the variables before obtaining the value of the distances. The most used way of data transformation is the standardization of the variables, according to the expression $Z_j = \frac{Y_j}{\sigma_j}$, in which σ_j corresponds to the standard deviation assossiated to the j-th characteristic.

The Hierarchy of grouping methods

There are countless grouping methods that are distinguished from each other by the type of result that will be provided e by the different ways of proximity definition between an individual and a group that is already shaped, or between two given groups. In all cases there is no knowledge about the number of groups that will be established. Besides, different methods generate different results. From the grouping methods, the one that are used the most are the hierarchy based and the optimization.

In the optimization methods, the groups are formed by the adequacy of a criteria of grouping. The goal is to reach a parting of individuals that optimizes (maximizes or minimizes) a previously defined measurement. One of the most commonly used methods is the one proposed by Tocher, quoted by Rao (1952). In the hierarchy based methods, the individuals are grouped by a process that repeats itself in a lot of levels until the dendrogram or the tree diagram is established. In this case,

there is no worry with the optimum number of groups, once that the biggest interest is in the tree and in it s ramifications.

In the hierarchy based procedures, there are the agglomerative ones, that describe the orientation of the grouping from the principle that each object é a natural grouping, afterwards gathering to others of bigger affinity, that successively are gathered until the formation of the supra-grouping, that is the group of objects as a whole.

There are a lot of ways to represent this grouping structure, such as: the method of the next neighbor, of the last neighbor, the UPGMA method, the Ward method, among others.

The Ward method, when applied to the data group of this study, is based in the reduction of the resulting information, given the inclusion of the set of objects in a group. This reduction of the information is determined by the adding of the squared error of each object, in function of the group mean in which this is included. The advantages of this agglomerative hierarchy based method are based on the fact that it uses the squared adding within the grouping, maximizing the differences between the groupings. Moreover it is less influenced by the presence of outliers in comparison with other methods. The expression is:

$$E_{(G1,G2)} = \sum_{V}^{P} \sum_{\substack{i=1 \\ i \in G1}}^{n} (x_{iv} - \bar{x}_v)^2$$

In which $\bar{x}_v$ is the group mean in each variable V.

This inclusion rule involves all the possible pairs that are defined as belonging to a given group the object that contributes the less to the increase of the adding of the squared error.

The statistical analysis were executed with the use of the software Action, a free supplement to Microsoft Excel®, based on the software R 2.6.1 for Windows®.

4. Dendrograma

The groupings are done by the use of all the available variables and representing in a two-dimensional way through the dendrogram (tree shaped two-dimensional diagram). In the dendrogram there are lines that are connected according to the similarity levels, that will set the pairs of individuals or of the variables according to Everitt (1993) e Landim (2001).

The dendrogram illustrates the fusions or partings that are done in each successive level of the grouping process, in which the X axis represents the individuals, and the Y axis represents the obtained distances after the using of the grouping methodology. The ramifications of the trees provide the order of the (n-1) connections, in which the first level represents the first connection, the second, the following connection, and so on until all of them gather.

Results and Discussion

The cluster analysis was applied aiming the identification of which metropolitan regions would present similar characteristics related to the studied variables, once that the regions are located in the same group. So, there has been the formation of four groups, and the cluster of each group of cities has similar aspects compared to each other, and different aspects compared to the other groups (Tabela 1).

Tabela 1 – Subdivision of the obtained groups using the Cluster Analysis methodology.

Group 1	**Group 2**	**Group 3**	**Group 4**
Belém	Fortaleza	Curitiba	Recife
Rio de Janeiro	Salvador		Belo Horizonte

São Paulo

Porto Alegre

The dendogram (Figura 2), portraits the formation of four groups, in which those cities are the ones that showed the smallest distance between the studied data and that have similar consumption habits. Then, it is possible to see that Belém and Rio de Janeiro are cities that have uniform characteristics, just as Salvador, Fortaleza, São Paulo and Porto Alegre, and this also happens with Belo Horizonte and Recife.

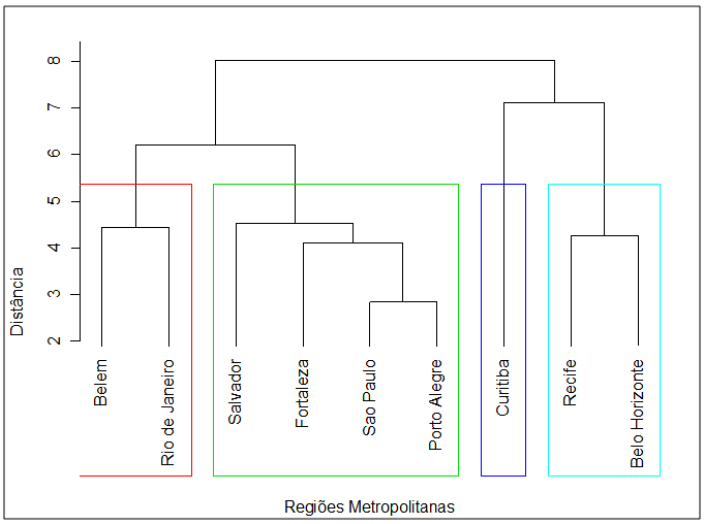

Figura 2 Resulting dendrogram of the application of the Standardized Euclidian Distance and the Ward method.

By the observation of the researched data it was possible to verify that the regions of the first group, Belém and Rio de Janeiro, have expenses with household articles and education that are very similar. On the contrary, the consumption and housing expenses that were different.

The cities that belong to the second group present significant similarities, once that Fortaleza and Salvador have equivalent expenses with food and beverages, just like São Paulo and Salvafor in housing. Fortaleza and Salvador, however, had identical expenses with health, but Salvador is the one that spends the less with personal expenses. Porto Alegre is the region that spent the most with communication.

The third group is constituted of only Curitiba, and a possible reasons for this result would be the fact that it is the only one that showed a negative result regarding one of the variables. The city exhibited the smallest expense with transportation and was the one that spent the most with clothing.

In the fourth group, Recife showed the biggest expenses with food and beverages, communication, transportation, personal expenses and household articles, while Belo Horizonte, spent the most with health and education.

5. Conclusion

From the researched data, it was possible to observe that all the metropolitan cities that were analyzed, demonstrated a growth in the price index throughout the year, and a significant decrease in the consumption at the beginning of the year. It was also possible to observe that the inflation index has

fall occurred on the following month, when the expenses with obligations increase and the buying capacity Is reduced. Moreover, in all the metropolitan regions the biggest consumption mean happened to food and beverages, personal expenses, education and health. So, in all the considered cities, the value in these aspects in bigger.

Referências

CHATFIELD, C. and COLLINS, A.J. (1986) **Introduction to multivariate analysis.** London: Chapman & Hall, 246p.

CORMACK, R.M (1971) A Review of classification. **Journal of Royal Statistical Society:** 321-367.

CRUZ, C. D. and CARNEIRO, P. C. S. (2003) **Modelos Biométricos aplicados ao melhoramento genetic**: 585p.

EVERITT, B.S (1993). **Cluster analysis**:136 p.

JOHNSON, R.A. and WICHEM, D.W. (1992) **Applied Multivariate Statistical Analysis:** 642 p.

KHATTREE, R AND NAIK, D.N. Khattree, R. and Naik, D.N. (2000) **Multivariate data reduction and discrimination with SAS software**: 558 p.

LANDIM, P. M. B. Geologia Quantitativa: Introdução à análise estatística de dados geológicos multivariados. Rio Claro - SP, 2001. (Livro em CD-ROM).

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. (1995) **Multivariate analysis:** 518.

RAO, C. R. (1952) **An advanced statistical method in biometric research:** 390.