Symbolic Approach to Analyzing Administrative Management

Eufrásio de A. L. Neto[1] and Francisco de A. T. de Carvalho[2] *

[1]Universidade Salgado de Oliveira / Datagro (www.datagro.com)
Rua Ernesto de Paula Santos, 187, Sl 403 - Boa Viagem
CEP: 51021-330, Recife-PE, Brazil
eufrasio@datagro.com.br

[2]Centro de Informatica, CIn/UFPE (www.cin.ufpe.br)
Av. Prof. Luiz Freire, s/n - Cidade Universitaria
CEP: 50740-540, Recife-PE, Brazil
fatc@cin.ufpe.br

Submitted: March 1, 2003; Accepted: July 3, 2003.

Abstract

The aim of this work is to identify locales with similar administrative management characteristics by using residents' opinions on essential public services, such as schools, traffic contorl, public safety, etc. Both the positive and negative aspects of each community will be enhanced by way of a suitable graphical tool. To achieve this goal, we will use clustering methods and graphical tools developed within the framework of symbolic data analysis. We will also see why the symbolic approach to data analysis is appropriate for this kind of study.

Keywords. Symbolic Data Analysis, Administrative Management Analysis, Divisive Hierarchical Clustering, Zoom-Star.

1 Introduction

This study concerns a public opinion survey held in 25 cities of the state of Pernambuco (Brazil). 5,241 residents responded to a list of questions about the administrative management of their city. The aim of this study is to partition these cities into groups with similar administration characteristics by using the residents' opinions on essential public services. In a second step, we will enhance the positive and negative aspects of each city by way of a suitable graphical tool.

In the original data table there were 5,241 rows representing the inhabitants of the 25 cities and 15 columns (variables) representing the residents' opinions. In addition, there was a supplementary column where the overall appreciation of the city's administrative

*This paper is supported by grants from CNPq, Brazil.

management was recorded. We used the SODAS software (www.ceremade.dauphine.fr/touati/sodas-pagegarde.htm) to construct a symbolic data table with 25 rows representing the cities and 15 columns representing the distribution of the residents' opinions in each city on these essential public services. Through this software, it was possible to aggregate the opinions of the respective inhabitants in each city and represent each city by a vector of distributions (symbolic data). Further details regarding symbolic data analysis can be found in Bock and Diday (2000).

The partitioning of these cities into groups with similar administration characteristics was achieved by way of a suitable divisive hierarchical clustering method. A graphical tool suitable for visualizing symbolic objects enhanced the positive and negative administration aspects of each city.

This article is organized in the following manner: Section 2 briefly describes the standard data table used in the survey. The symbolic data table, derived from this standard table, is also presented. Sections 3 and 4 respectively present the methodological aspects of the study and the corresponding results obtained. Finally, in section 5 we present our final comments and conclusions.

2 Standard and symbolic data tables.

The opinions of 5,241 residents from 25 cities in the state of Pernambuco regarding essential public services were placed into the cells of the standard data table. This table had 89,097 cells distributed among 5,241 rows (residents) and 16 columns (variables). The fifteen variable columns of this table represent the following essential public services: Street Cleaning, Garbage Collection, Street Lights, Schools, Social Services, Parks Maintenance, Public Health, Events, Water Supply, Public Sanitation, Public Safety, Street Surfacing, Employment Opportunities, Road Repairs and Traffic Control. A special column of this table contains the Administrative Balance variable, which represents the opinion of each inhabitant regarding the city's overall administrative management. All variables are qualitative ordinal variables and have the same domain formed by the following categories: Very Bad, Bad, Adequate, Good and Excellent.

The symbolic data table (Figure 1) was obtained from the standard table using the DB2SO module of the SODAS software (Stéphan et al (2000)). DB2SO enables the user to build a symbolic data table from data stored in a relational database. It is assumed that a set of items is stored in the database and that the individual items are distributed into groups. Then DB2SO can aggregate this data and construct a symbolic description for each group of individuals.

The symbolic data table has 25 rows representing the symbolic descriptions of the 25 cities, and 16 columns - 15 modal symbolic variables, relating to public services, and 1 standard dichotomous variable. Due to reasons of confidentiality, the original names of the cities are omitted. We used the following fictitious names: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Z.

All the standard ordinal qualitative variables from the standard data table, with the exception of the Administrative Balance variable, are now modal symbolic variables of the symbolic data table, which takes the distribution of residents' opinions on the list of

Table

	Administrative	Street Cleaning
K	n	Very Bad (0.29), Bad (0.15), Adequate (0.26), Good (0.26), Excellent (0.03)
J	n	Very Bad (0.13), Bad (0.10), Adequate (0.34), Good (0.35), Excellent (0.09)
L	n	Very Bad (0.31), Bad (0.13), Adequate (0.27), Good (0.24), Excellent (0.05)
F	p	Very Bad (0.14), Bad (0.07), Adequate (0.24), Good (0.41), Excellent (0.14)
P	n	Very Bad (0.07), Bad (0.04), Adequate (0.24), Good (0.49), Excellent (0.15)
M	n	Very Bad (0.19), Bad (0.17), Adequate (0.30), Good (0.26), Excellent (0.07)
Q	p	Very Bad (0.08), Bad (0.02), Adequate (0.22), Good (0.50), Excellent (0.18)
N	n	Very Bad (0.17), Bad (0.12), Adequate (0.25), Good (0.39), Excellent (0.07)
C	p	Very Bad (0.05), Bad (0.01), Adequate (0.13), Good (0.57), Excellent (0.25)

Figure 1: Symbolic data table

essential public services for each city (row). Consequently, in each cell of the symbolic data table there is a list of pairs formed by a category of a symbolic variable and its associated frequency. For example, in Figure 1 the modal variable Street Cleaning takes the distribution $\{VeryBad(0.29), Bad(0.15), Adequate(0.26), Good(0.26), Excellent(0.03)\}$ as the value for the city K.

The Administrative Balance variable was transformed into a dichotomous variable, called Balance, according to the following rule:

If

[(%Excellent + %Good) - (%Very Bad + %Bad)] > %Adequate

then

Balance = p (positive administrative balance)

else

Balance = n (negative administrative balance).

where %Very Bad means the percentage of city residents which have classified the city's overall administrative management as *Very Bad*, and so on.

In this way, without taking the Administrative Balance variable into consideration, each row (city) of the symbolic data table is represented as a vector of distributions:

$$\begin{aligned} s_i \quad = \quad & (\{\text{Clean} = \text{Very Bad}(p_{1i}), \ldots, \text{Clean} = \text{Excellent}(p_{5i})\}, \ldots, \\ & \{\text{Traffic} = \text{Very Bad}(q_{1i}), \ldots, \text{Traffic} = \text{Excellent}(q_{5i})\}) \end{aligned}$$

Note that $\sum_{j=1}^{5} p_{ji} = 1$ and $\sum_{j=1}^{5} q_{ji} = 1$.

3 Partitioning the cities according to similar administration characteristics for essential public services.

As previously mentioned, the aim of this work is to identify cities with similar administration characteristics according to their residents' opinions on essential public services.

To achieve this objective, we selected from the SODAS software a divisive hierarchical clustering method - DIV - suitable for grouping similar symbolic descriptions into the same class (Chavent (1997), Chavent (1998), Chavent (2000)).

This method starts with all the objects in a cluster and proceeds by successive divisions of each cluster. At each step, a cluster is divided into two clusters according to a binary question. This binary question induces the best partition into two clusters according to an extension of the inertia criterion. The algorithm stops after $K-1$ divisions, where K is the number of clusters given as input by the user. The output of this divisive clustering method is a hierarchy and a decision tree (Chavent (1998)).

In the framework of the DIV algorithm (Chavent (2000)), a cluster C is split according to a binary question in the form of *is* $Y_j \leq v$?, where v is a value of the domain of Y_j that is called the *cut value*. An object $s \in C$ answers "yes" or "no" to the binary question according to the binary function $q_j : C \longrightarrow \{true, false\}$. The bipartition (C_1, C_2) of C induced by the binary question is defined as:

$$C_1 = \{s \in C | q_j(s) = true\} \text{ and } C_2 = \{s \in C | q_j(s) = false\}$$

If Y_j is a modal symbolic variable with $Y_j(s) = \pi_s$, where π_s is a frequency distribution, the function q_j is defined as (Chavent (2000)):

$$q_j(s) = true \text{ if } \sum_{x \leq v} \pi_s(x) \geq 1/2 \text{ and } q_j(s) = false \text{ if } \sum_{x \leq v} \pi_s(x) < 1/2$$

The 25 cities were clustered on the basis of the information supplied for the following 15 public services: Street Cleaning, Garbage Collection, Street Lights, Schools, Social Services, Parks Maintenance, Public Health, Events, Water, Public Sanitation, Public Safety, Street Surfacing, Employment Opportunities, Road Repairs and Traffic Control. Note that the Balance variable has no participation in the cluster partition process. All the variables corresponding to the public services are modal and, as such, the comparison between a pair of symbolic descriptions (cities) is achieved through the Φ^2 distance (Chavent (1997)). In the DIV module of the SODAS software, we considered this distance without normalization.

The DIV module achieved the partitioning of the symbolic descriptions representing the cities according to a user selected number of classes between 2 and 10. None of these partitions were able to completely separate the cities according to the dichotomous Balance variable (positive or negative). However, among these partitions, the 5-cluster partition simultaneously had the smallest number of clusters and a high degree of homogeneity according to the positive/negative balance given by the Balance variable. We therefore selected this partition.

The resulting 5-cluster partition is:

Cluster 1: C_1 = {K (n), J (n), L (n), M (n), N (n), O (n), Z (n)}
Cluster 2: C_2 = {F (p), G (p), H (p), I (p)}
Cluster 3: C_3 = {A (p)}
Cluster 4: C_4 = {C (p), B (p), D (p), E (p)}
Cluster 5: C_5 = {P (n), Q (p), R (p), S (p), T (p), U (n), V (n), X (n), Y (p)}

This 5-cluster partition comprises: (i) cluster(s) where all the cities have a positive administrative management (C_2, C_3, C_4); (ii) cluster(s) where all the cities have a negative administrative management (C_1) and (iii) cluster(s) with both positive and negative administrative management characteristics (C_5).

The four binary functions used to construct this 5-cluster partition are:

$q_1(s)$ = [Traffic Control(s) $\leq$ Adequate]
$q_2(s)$ = [Street Surfacing(s) $\leq$ Adequate]
$q_3(s)$ = [Street Cleaning(s) $\leq$ Adequate]
$q_4(s)$ = [Employment Opportunities(s) $\leq$ Bad]

According to the binary functions q_1, q_2, q_3 and q_4, production rules can be defined for each cluster (Chavent (2000)). Indeed, the dendrogram can be read as a decision tree and the rules can be read as classification rules that assign a new object to one of the five clusters.

Figure 2 displays the dendrogram of the hierarchy obtained with the Φ^2 distance. At each node, the right child answered *true* for the corresponding binary question and the left child answered *false* for the same binary question.

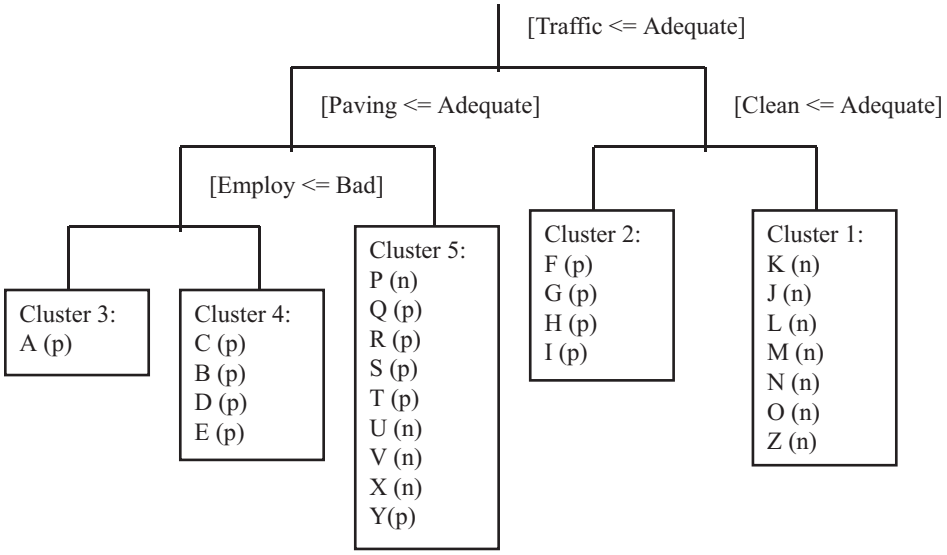

Figure 2: Dendrogram for the symbolic data set of administrative management

From this dendrogram we can see a first splitting of these cities as induced by the binary question $q_1(s)$ = [Traffic Control(s) $\leq$ Adequate], where in eleven cities - situated on the right side of the dendrogram - the majority of the inhabitants have evaluated the administrative management of services relating to traffic control as either Very Bad, Bad

or Adequate; and the majority of the inhabitants in fourteen cities - situated on the left side of the tree - have evaluated the administrative management of services relating to traffic as either Good or Excellent.

Using the information contained in the most recent demographic census taken by the Brazilian Institute of Geography and Statistics (IBGE), we observe that the average population of the cities on the right side of the dendrogram (clusters C_1 and C_2) is 300,000 inhabitants against the 94,000 inhabitants for the cities on the left side of the dendrogram (clusters C_3, C_4 and C_5). Note that the probability of traffic problems in cites with a larger number of people is higher than that in cities with fewer people. This explains the choice of the Traffic variable for inducing the initial partitioning of this data set.

Returning to the right side of the tree, the eleven cities situated on this side are split into clusters C_1 and C_2 according to the binary questions $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}]$ and $q_2(s) = [\text{Clean(s)} \leq \text{Adequate}]$.

Cluster C_1 is characterized by the following rule:

If $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}] = true$ **and** $q_3(s) = [\text{Clean(s)} \leq \text{Adequate}] = true$ **then** $s \in C_1$

According to this classification rule, the majority of the inhabitants in cities belonging to this cluster evaluated the administrative management of the services relating to traffic and street cleaning as Very Bad, Bad or Adequate. All the cities belonging to this cluster exhibit a Negative (n) for the category of the Balance variable.

Cluster C_2 is characterized by the following rule:

If $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}] = true$ **and** $q_3(s) = [\text{Clean(s)} \leq \text{Adequate}] = false$ **then** $s \in C_2$

This means that the majority of the inhabitants in the cities belonging to this cluster evaluated the services relating to traffic as Very Bad, Bad or Adequate, and those relating to street cleaning as Good and Excellent. Note that all cities belonging to this cluster indicated a positive (p) for the category of the Balance variable.

The splitting of these cities in clusters C_1 and C_2 is explained by the different way where the majority of its inhabitants evaluated the services relating to street cleaning.

Cluster C_3 is characterized by the following rule:

If $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}] = false$ **and** $q_2(s) = [\text{Paving(s)} \leq \text{Adequate}] = false$ **and** $q_4(s) = [\text{Employ(s)} \leq \text{Bad }] = false$ **then** $s \in C_3$

City A is alone in cluster 3. For this city, the Balance variable indicates the positive (p) category. The majority of its inhabitants evaluated the administrative management of the services relating to traffic and street surfacing as either Good or Excellent, and those relating to employment opportunities as Adequate, Good or Excellent.

Cluster C_4 is characterized by the following rule:

If $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}] = false$ **and** $q_2(s) = [\text{Paving(s)} \leq \text{Adequate}] = false$ **and** $q_4(s) = [\text{Employ(s)} \leq \text{Bad }] = true$ **then** $s \in C_4$

The majority of the inhabitants evaluated the services relating to the traffic and street surfacing as Good or Excellent, and those relating to employment opportunities as Very bad, Bad or Adequate. Note that all the cities belonging to this cluster have the Balance variable taking the positive (p) category.

From the dendrogram (Figure 2), we can conclude that the existence of cluster C_3 is explained by at least an Adequate perception on the part of the majority of inhabitants in city A concerning the management services relating to employment opportunities.

Finally, cluster C_5 is characterized by the following rule:

If $q_1(s) = [\text{Traffic(s)} \leq \text{Adequate}] = false$ **and** $q_2(s) = [\text{Paving(s)} \leq \text{Adequate}] = true$ **then** $s \in C_5$

The majority of inhabitants evaluated the services relating to traffic as either Good or Excellent, and those relating to street surfacing as Very Bad, Bad or Adequate. Note that only cluster 5 contains cities where the Balance variable indicates the negative (n) category and other cities where this variable indicates the positive (p) category.

4 Visualizing cities according to the administrative management of essential public services.

In this section, we will use a graphical tool suitable for visualizing symbolic data. In Noirhomme-Fraiture (2002), visualization techniques for data in aggregate forms are presented, particularly a type of graphic called Zoom Star. With this graphical representation, the diameter dots on the axes representing modal categorical variables are proportional to the frequency of the category in question for the symbolic object. To lend shape, the categories with the highest frequency are joined together (Noirhomme-Fraiture (2002)).

The Zoom Star graphic is used in our study with the aim of constructing an administrative evaluation summary of each city represented by a symbolic object. With this graphical tool, we will have a better visualization of each city, thus easily identifying the positive and negative aspects of each administration. Moreover, it will be possible to verify if the cities belonging to the same cluster are similar or not.

All the modal symbolic variables in the study share the categories Very Bad, Bad, Adequate, Good and Excellent. In this particular situation we can state that a Zoom-Star 2D graphic from a city with a predominantly circular shape and the border of which is far from the origin will characterize a mainly approved administration. A Zoom-Star 2D graphic with a predominantly circular shape and with the border close to the origin will characterize a mainly disapproved administration. A Zoom-Star 2D graphic with a predominantly non-circular shape indicates divided opinions regarding the evaluation of the city's administrative management. These trends will reflect the opinion of a majority of people who approve of the city administration if the Zoom-Star 2D graphic border is far from the origin and it will reflect a majority of people that disapprove of the city administration if the Zoom-Star 2D graphic border is close to the origin.

Figures 3 and 4 show the Zoom-Star graphical representation of the administrative management in cities K, J, L, M, N, O and Z, which belong to cluster 1.

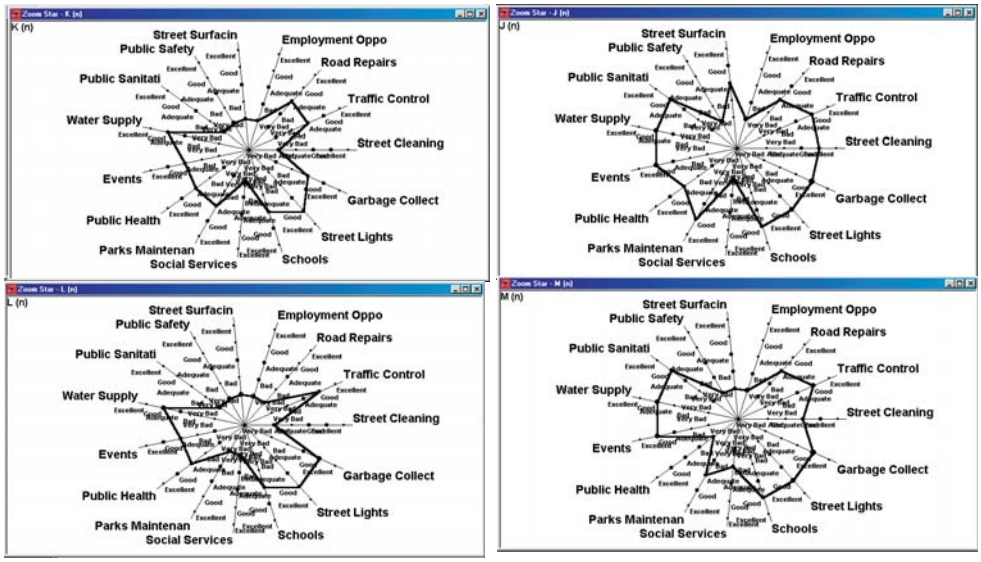

Figure 3: Zoom Stars of the cities K, J, L and M from cluster 1

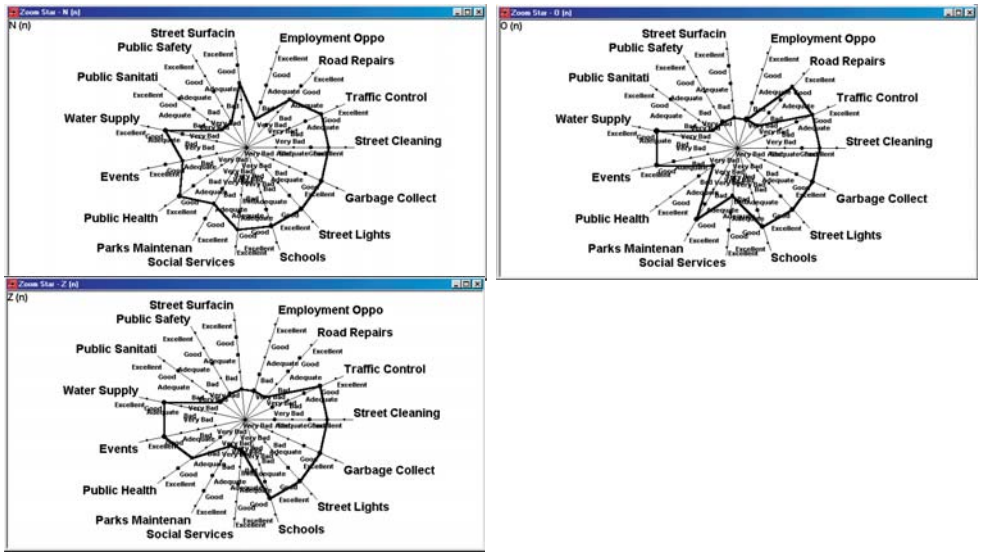

Figure 4: Zoom Stars of the cities N, O and Z from cluster 1

The Zoom-Stars representing these cities present a predominantly non-circular shape and a border close to the graphic origin. Thus, they characterize a predominantly disapproved administration. Indeed, these cities have the Balance variable in the category negative (n). It is possible to identify serious deficiencies in the following public services offered by these cities: Public Sanitation, Public Safety, Street Surfacing, Employment Opportunities, Social Services, Street Cleaning and Public Health Services - classified as Very Bad in one or more cities. However, it must be registered that services regarding Water, Schools, Street Lights and Garbage Collection were classified as Adequate or Good.

Figure 5 shows the Zoom-Star graphical representation of the administrative management in cities F, G, H and I from Cluster 2.

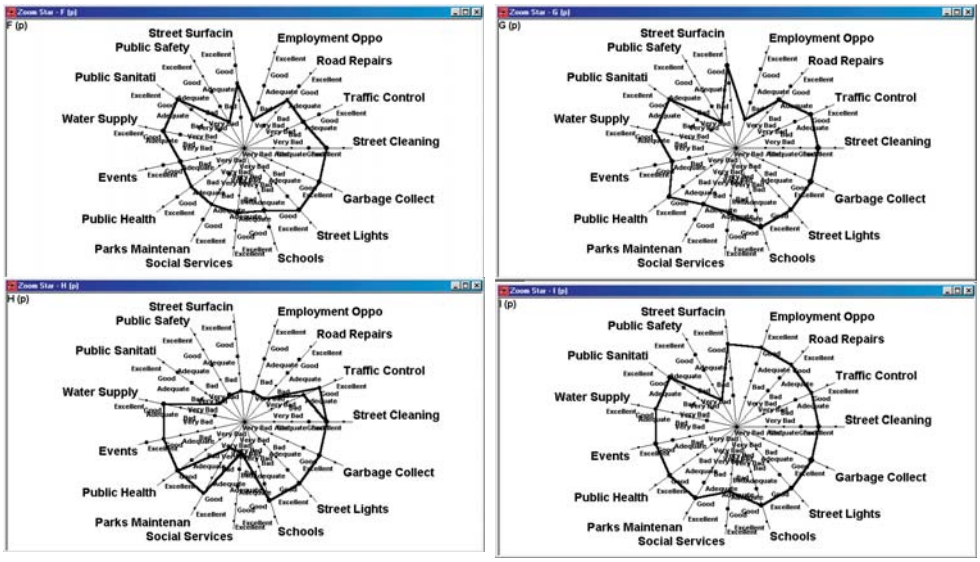

Figure 5: Zoom Stars of the cities belonging to cluster 2

The cities in cluster 2 have the following services classified as Adequate: Social Services, Parks Maintenance, Road Repairs and Street Surfacing. Also, the services of Public Safety and Employment Opportunities were classified as Very Bad. However, the services regarding Street Cleaning, Garbage Collection, Street Lights, Schools and Water Supply were classified as Good. This is why the administrative management of these cities was mainly approved.

Figure 6 shows the Zoom-Star graphical representation of the administrative management in city A from Cluster 3.

In this city, the majority of the population classified the public services as Good on a scale that varies from Very Bad to Excellent. This city was isolated in a separated cluster from the cities of cluster 4 because of the Adequate category regarding the Employment Opportunities variable.

Figure 7 shows the Zoom-Star graphical representation of the administrative management in cities B, C, D and E from Cluster 4.

We can see that the Zoom-Stars of the cities' administrative management in this cluster are quite similar: the shape is predominantly circular and the border far from the origin of the graphic. We conclude that the great majority of the essential public services have

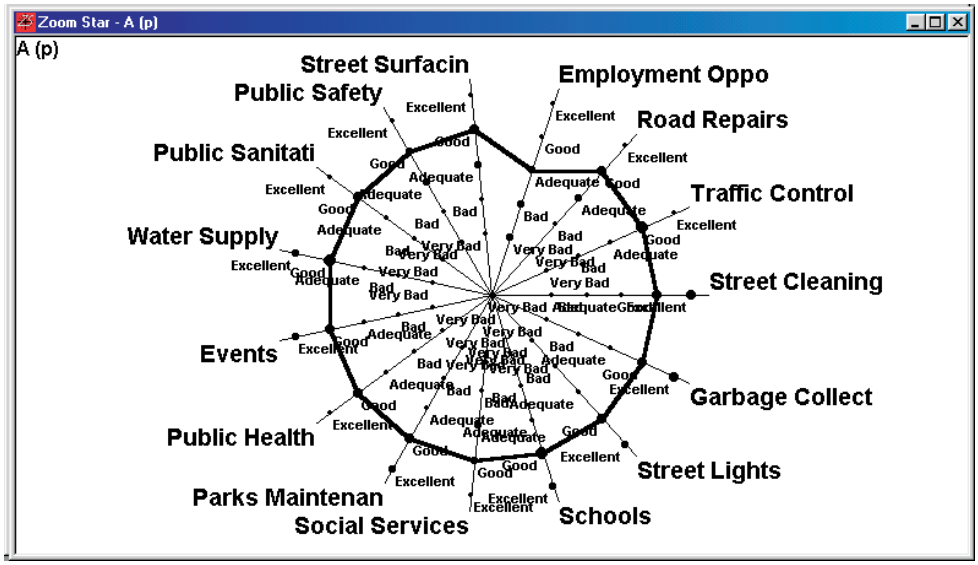

Figure 6: Zoom Stars of the city A belonging to cluster 3

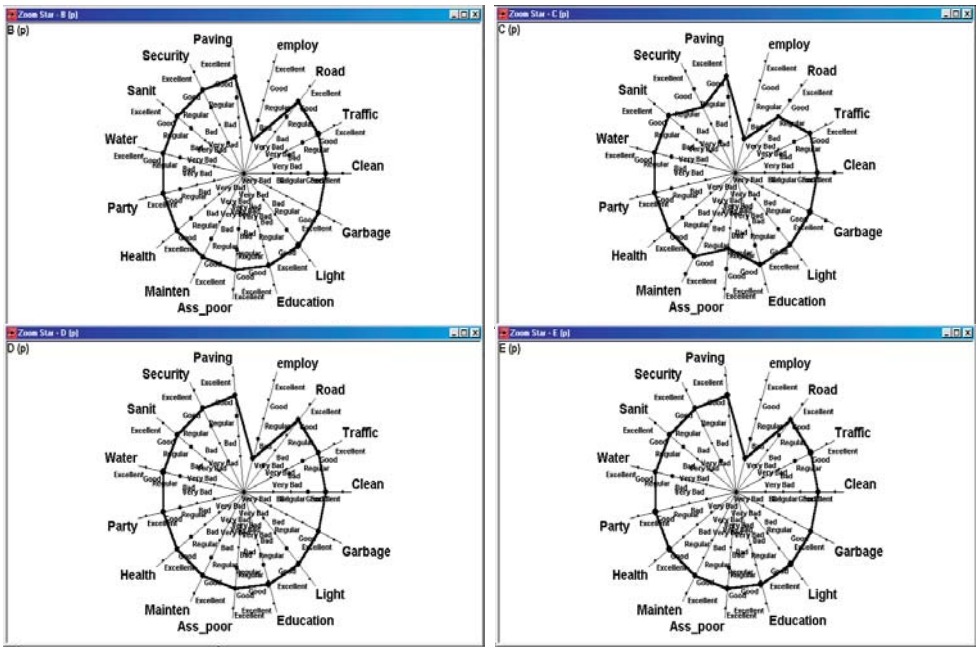

Figure 7: Zoom Stars of the cities B, C, D and E belonging to cluster 4

been given a satisfactory evaluation by their residents, with the exception of the Public Safety and Social Services in city C and Employment Opportunities in all the cities of this cluster.

Figures 8 and 9 show the Zoom-Star graphical representation of the administrative management in cities P, Q, R, S, T, U, V, X and Y that belong to cluster 5.

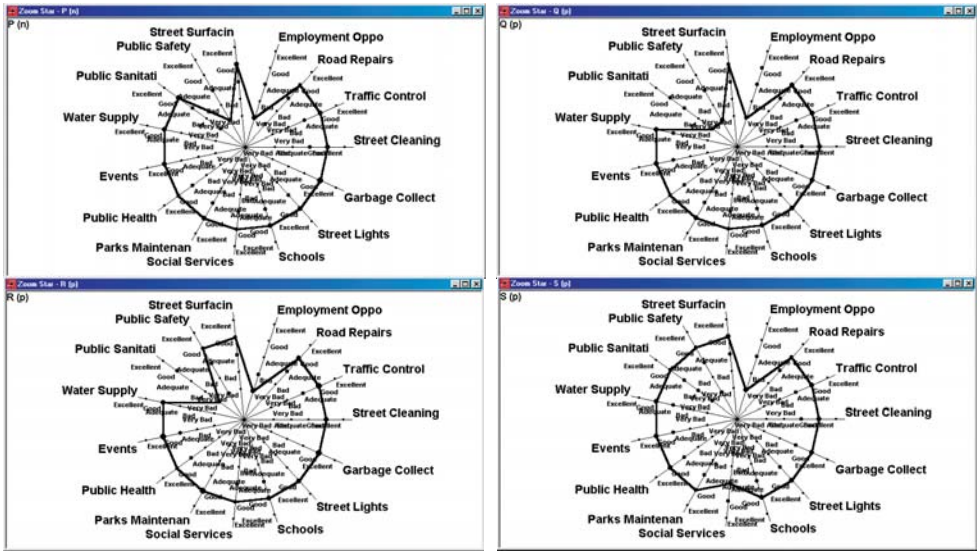

Figure 8: Zoom Stars of the cities P, Q, R, and S from cluster 5

As we have already remarked, this cluster is only heterogeneous in the partition (cities that have the Balance variable in the positive (p) category and cities that have this variable in the negative (n) category). From Figures 8 and 9, we can see that the following public services are classified as Good: Schools, Street Lights, Garbage Collection and Street Cleaning. On the other hand, Employment Opportunities and Public Safety are classified as Very Bad.

The services regarding Public Sanitation and Social Services were classified as Very Bad in some cities and Good in others, confirming the differences between the residents' evaluations of the administrative managements in the cities of this cluster.

It is important to notice that the shape of the Zoom-Stars representing the cities in clusters 3 and 4 is predominantly circular and that their border is far from the origin when compared to the Zoom-Stars representing the cities in cluster 2. By contrast, the Zoom-Stars representing the cities in cluster 1 present a predominantly non-circular shape and a border close to the graphic origin. Thus, they characterize a mainly disapproved administration.

5 Final Comments and Conclusions

This study concerned a public opinion survey from 25 cities in the state of Pernambuco (Brazil) where 5,241 residents responded to a list of questions on the administrative management of these cities. To analyze these responses, we used several tools available in the SODAS software developed within the framework of the symbolic approach to data

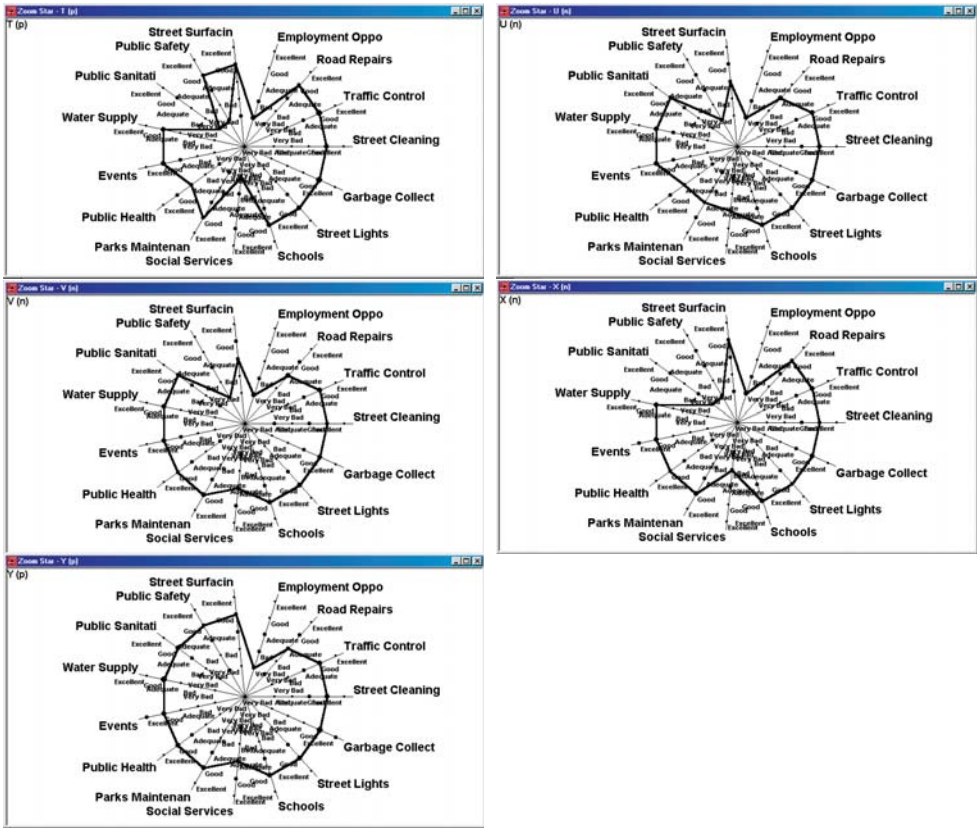

Figure 9: Zoom Stars of the cities T, U, V, X and Y from cluster 5

analysis. This choice allowed us to represent the cities by the aggregated data obtained from the residents' responses to this list of questions and then to compare them directly by way of a divisive clustering method along with a visualization tool suitable for analyzing aggregated (symbolic) data. We can now formulate the main statements concerning our study:

i) Due to the divisive clustering method, it was possible to identify homogeneous clusters of cities, which presented similar administrative management patterns. Moreover, this method also furnished the variables that best explained the partitioning of the cities into homogeneous clusters: Traffic Control, Street Surfacing, Street Cleaning and Employment Opportunities;

ii) The Zoom-Star 2D visualization method allowed us to observe the details regarding the opinions of city residents on the administrative management performance of the authorities and to verify the degree of homogeneity in the clusters as given by the divisive method. This visualization method also allowed for more detailed analysis of the administrative management for each city and the identification of its positive and negative points;

iii) Cities that have been positively classified with respect to the city authorities' administrative management performance have stars with a predominantly circular shape

with borders far from the origin. Cities that have been negatively classified present stars with a predominantly circular shape and borders close to the origin;

iv) Cities with Adequate, Bad or Very Bad classification in the variables of Traffic Control and Street Cleaning were negatively evaluated in regards to the administrative management; Cities with Good or Excellent classification in Traffic Control and Street Surfacing were positively evaluated in regards to the administrative management.

We believe that the methodological approach used in this work can easily be applied to a number of other fields to evaluate the performance of organizations and assist professionals from diverse areas in the process of decision-making, such as in:

a) Business - identify competitive and not competitive factories and the main factors explaining their degree of competitiveness;

b) Commercial - identify similarities and dissimilarities between distinct consumer markets;

c) Politics - prepare electoral profile of politicians and later identify their similarities and dissimilarities.

References

[1] Bock, H. -H. and Diday, E. (2000). *Analysis of Symbolic Data*, Springer, Heidelberg.

[2] Chavent, M. (1997). *Analyse des Données Symbolique, une méthode divisive de classification*. Thèse de l'Université Paris-IX Dauphine.

[3] Chavent, M. (1998). *A monothetic clustering method*, Pattern Recognition Letters, 19, 989-996.

[4] Chavent, M. (2000). Criterion-Based Divisive Clustering for Symbolic Data. In: H. -H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*, Springer, Heidelberg, 299-311.

[5] Marcelo, C. (2002). *Quality Control in the Portuguese Labour Force Survey*, Electronic Journal of Symbolic Data Analysis, 0, 40-46.

[6] Noirhomme-Fraiture, M. (2002). *Visualization of Large Data Sets: The Zoom Star Solution*, Electronic Journal of Symbolic Data Analysis, 0, 26-39.

[7] Stéphan, V., Hébrail, G. and Lechevallier, Y. (2000). Generation of Symbolic Objects from Relational Data Bases. In: H. -H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*, Springer, Heidelberg, 78-105.