Luiz José Schirmer Silva

CrimeVis

An Interactive Visualization System for Analyzing Criminal Data in the State of Rio de Janeiro

Dissertação de Mestrado

Departamento de Informática
Programa de Pós–graduação em Informática

Rio de Janeiro
June 2016

Pontifícia Universidade Católica
do Rio de Janeiro

Luiz José Schirmer Silva

CrimeVis

An Interactive Visualization System for Analyzing Criminal Data in the State of Rio de Janeiro

Dissertação de Mestrado

Dissertation presented to the Programa de Pós Graduação em Informática of the Departamento de Informática of PUC-Rio as partial fullfillment of the requirements for the degree of Mestre em Informática.

Advisor : Prof. Hélio Côrtes Vieira Lopes
Co–Advisor: Prof. Simone Diniz Junqueira Barbosa

Rio de Janeiro
June 2016

Pontifícia Universidade Católica
do Rio de Janeiro

Luiz José Schirmer Silva

CrimeVis

An Interactive Visualization System for Analyzing Criminal Data in the State of Rio de Janeiro

Dissertation presented to the Programa de Pós Graduação em Informática of the Departamento de Informática do Centro Técnico Científico da PUC-Rio as partial fullfillment of the requirements for the degree of Mestre.

Prof. Hélio Côrtes Vieira Lopes
Advisor
Departamento de Informática — PUC-Rio

Prof. Simone Diniz Junqueira Barbosa
Co–Advisor
Departamento de Informática — PUC-Rio

Prof. Klarissa Almeida Silva Platero
Universidade Federal Fluminense

Prof. Pedro Carvalho Loureiro de Souza
PUC-Rio

Prof. Waldemar Celes Filho
PUC-Rio

Prof. Márcio da Silveira Carvalho
Coordinator of the Centro Técnico Científico da PUC-Rio

Rio de Janeiro, June 2nd, 2016

Luiz José Schirmer Silva

The author graduated in Computer Science from Universidade Federal de Santa Maria - UFSM in 2014, he has interest in scientific visualization, computer graphics and information visualization.

Bibliographic data

Silva, Luiz José Schirmer

CrimeVis: An Interactive Visualization System for Analyzing Criminal Data in the State of Rio de Janeiro / Luiz José Schirmer Silva; advisor: Hélio Côrtes Vieira Lopes; co–advisor: Simone Diniz Junqueira Barbosa. — Rio de Janeiro : PUC-Rio, Departamento de Informática, 2016.

(em Inglês)

53 f: il. (color.); 29,7 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2016.

Inclui bibliografia.

1. Informática – Teses. 2. Vistas coordenadas. 3. Visualização científica. 4. Dados criminais. 5. Agrupamento de dados. I. Lopes, Hélio Côrtes Vieira. II. Barbosa, Simone Diniz Junqueira. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Para pessoas especiais.

Acknowledgments

First of all, i would like to thank my fiancee Letícia Fausto for her love and all the amazing moments we've gone together. Also, i would like to thank my parents for their unconditional love and support.

My deepest gratitude goes to my advisors, Hélio Côrtes Vieira Lopes and Simone D. J. Barbosa, for all patience, friendship and for the enormous contribuition to my academic growth. I would like to thank prof. Waldemar Celes for the opportunity to work in the Geresim development group.

I thank my colleagues Cassio Almeida and Sonia Gonzales for all help in this project. Your comments and discussions are always enriching.

My apreciation goes to all my friends in particular, Fabrício Cardoso, Guilherme Schardong, Renan Spencer, Laisla Ribeiro, Aline Machado and Lidiane Castagna. Also, I thank all the friends and members of the visualization group of Tecgraf/PUC-Rio.

Finally i also thank CAPES, PUC-Rio and Tecgraf for the financial aid, without which this work would not have been possible.

To all of you, my sincere thanks.

Abstract

Silva, Luiz José Schirmer; Lopes, Hélio Côrtes Vieira (Advisor); Barbosa, Simone Diniz Junqueira (Co-advisor). **CrimeVis:An Interactive Visualization System for Analyzing Criminal Data in the State of Rio de Janeiro**. Rio de Janeiro, 2016. 53p. MsC Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This work presents the development of an interactive graphic visualization system for analyzing criminal data in the State of Rio de Janeiro, provided by the Public Safety Institute from the State of Rio de Janeiro (ISP-RJ, Instituto de Segurança Pública). The system presents to the user a set of integrated tools that support visualizing and analyzing statistical data on crimes, which make it possible to infer relevant information regarding government policies on public safety and their effects. The tools allow us to visualize multidimensional data, spatiotemporal data, and multivariate data in an integrated manner using brushing and linking techniques. The work also presents a case study to evaluate the set of tools we developed.

Keywords

Coordinated views; Scientific visualization; Criminal data; Data clustering;

Resumo

Silva, Luiz José Schirmer; Lopes, Hélio Côrtes Vieira (Orientador) ; Barbosa, Simone Diniz Junqueira (Co-orientadora). **CrimeVis: Um sistema interativo de visualização para análise de dados criminais do estado do Rio de Janeiro**. Rio de Janeiro, 2016. 53p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho apresenta o desenvolvimento de um sistema gráfico e interativo para análise de dados criminais do estado do Rio de Janeiro. O objetivo é introduzir uma ferramenta de apoio à análise de dados estatísticos fornecidos pelo Instituto de Segurança Pública do Estado do Rio de Janeiro (ISP-RJ). Para o usuário é apresentado um conjunto de ferramentas integradas para visualização e através desta análise é possível obter informações relevantes a respeito das políticas públicas de segurança. A ferramenta desenvolvida permite analisar tanto dados multidimensionais, variando no espaço (2D, 3D) e tempo; quanto dados multivariados, compreendendo n variáveis, de maneira integrada através de técnicas de "brushing and linking". Ao final ainda é apresentado um estudo de caso direcionado para avaliar o conjunto de ferramentas apresentado.

Palavras–chave

Vistas coordenadas; Visualização científica; Dados criminais; Agrupamento de dados;

Contents

List of figures

List of tables

1
Introduction

As cities and their populations have grown, so has violence increased in Brazil. In Rio de Janeiro, especially, we observe that the large social and financial inequalities, as well as the regional distribution of the population may influence criminal data. The number of crimes – especially the violent ones – increases every year. Containing that increase and ensuring better quality of life have become major concerns of the government and of public safety institutions [4]. Defining efficient public policies is a challenge for any government, and devising strategies for combating criminality directly affects most of the vulnerable population.

Many areas within Rio de Janeiro are still controlled by criminals, in communities whose population, in addition to living under the domain of drug traffickers, is exposed to territorial disputes between different criminal groups or to confrontations between the criminals and the police. Despite the gravity of those confrontations, there is no systematic evidence of the short- and long-term impacts of those confrontations on the people who live and work in the affected areas [4].

This work focuses on the development of a visualization tool to support the analysis of criminal data of Rio de Janeiro, to provide researchers with an efficient means to analyze the possible hypothesis and the possible impact of criminality in the state. The research and the system development have an exploratory nature, with the goal to evaluate data provided by the state's Public Safety Institute (ISP-RJ, Instituto de Segurança Pública). In total, we analyzed statistical data of 138 police districts over 12 years. The data were built from criminal notifications (RO - *Registros de Ocorrência* recorded in the civil police stations (DPs - *Delegacias de Polícia*), in addition to complementary information from Military Police organs.

The statistics provided are based on the data where the criminal notification was recorded; they include not only the type of crime, but also geographical information of where it occurred. This way, the recorded date becomes an important variable for the consolidation of data and the production of official statistics. In addition, the geographical distribution of the criminal occurrences

is associated to the DP where the crime occurred, regardless of the DP where it was notified. The data obtained from ISP-RJ were also analyzed together with socioeconomic data collected and made available by IBGE, the Brazilian Institute of Geography and Statistics [5]. Those data include schooling, ethnicity, and average family income for the area corresponding to each DP.

Given the investigated context, it was necessary to develop a support tool to make it possible to quickly analyze hypotheses and evaluate the government safety policies. Thus, we developed CrimeVis to use georeferenced and statistical data provided by ISP-RJ and IBGE. That system provides a set of exploration tools, which allow the discovery of patterns and correlations between the analyzed data sets. The application allows us to integrate visualizations and use them synchronously. In addition, it allows the user to look for patterns using different clustering and classification methods. The main contributions of this work are:

1. The visualization of n-dimensional criminal data, making it possible to contrast them with socioeconomic variables and relate them to the distribution of the population of Rio de Janeiro.

2. The application of new forms of statistical data analysis for the studied domain.

3. The implementation of a set of tools that allow the analysis of public safety policies adopted during the investigated period of time, and the relation of these data with the social and economic distribution of the population.

This work is organized as follows. Section 2 presents some theoretical background on our work. Section 3 presents detailed information regarding the developed tool suite and the tool characteristics. Next, in Section 4 we present the results of a user study conducted with people from a variety of educational backgrounds. Finally, in Section 5 we discuss the importance of those results from the users' point of view and their impact on state public safety policies. We also point to some opportunities for improvement, from a computational standpoint.

2
Theoretical Background

2.1
Criminal Statistics and Socioeconomic Variables

In Rio de Janeiro, police districts are distributed in 138 DPs throughout the state. The state government, following a policy in favor for transparency and access to the statistics of criminal data, makes those data publically available through the website of the ISP-RJ [6]. Those data span a period from 2003 to 2015. The crimes are counted as follows: for crimes against the individual (i.e., homicides, lesions, and threats), they consider the number of victims, whereas for crimes against property (i.e., robberies and thefts), they consider the number of cases, regardless of the number of victims in each case. The case of gun seizure constitutes an exception, for which the number of weapons is counted. From those data we generate statistics according to the population distribution.

Evaluating and quantifying the impact of violence on the state of Rio de Janeiro present several challenges. First, understanding the problem is made difficult because of the inconsistency and lack of relation between the different types of data. Violence naturally varies very much both geographically and over time. The exposition to violence is different when a person lives in a conflict territory, at its surroundings, or at a five kilometer radius from its epicenter [4]. The official criminal data of Rio de Janeiro are aggregated by city regions: they do not allow us to identify either the precise location of violence epicenters or information about the population of each region. To circumvent this problem, the data obtained can be contrasted with socioeconomic data related to the region associated to each DP. Those data, obtained by the Brazilian 2010 census [5], include information on the population ethnicity and income, as well as a classification of regions considered subnormal, which are regions lacking essential public services and with buildings that were constructed irregularly. In this way, not only we can analyze criminal data as indicators of public safety policies, but also the visualization of the available socioeconomic data and complementary information can suggest possible patterns and relations

among the various data.

2.2 Statistical Data Visualization and Analysis

Regarding the analysis of criminal data, a major challenge is related to data clustering. One of the key points is the possibility to discover patterns and practical meanings for the user. The knowledge discovery process requires direct dialogue with a specialist in making decisions based on the results [7], in order to answer relevant questions. However, to help in that task, different clustering algorithms can be used and evaluated. Algorithms such as k-means [8] and k-medoids [9] can be used in an attempt to identify homogeneous groups distributed in the observed source. In this process, the objects are grouped to maximize the intra-cluster similarity and minimize the inter-cluster similarity [10]. This way we can evaluate n-variate data sets in order to uncover possible patterns and correlations. Although those algorithm subdivide the data set efficiently, their limitation in the context of criminal data analysis is due to the lack of consideration of spatial data, i.e., the spatial adjacencies of the records. Conversely, the SKATER (Spatial Kluster Analysis by Tree Edge Removal) algorithm [3] partitions a data set according to the spatial distribution of the data. It uses a connectivity graph to find adjacency relations between the analyzed objects. In this algorithm, each data point is considered a vertex that has both its attributes and its spatial location. The weight of each edge is a measure of dissimilarity, which can be the Euclidian distance between the multivariate attributes. The algorithm works as follows: first it creates a minimum spanning tree of the adjacency graph based on the Euclidean distance between the attributes. Later, the tree is partitioned in spatial clusters based on a global measure. This way, the location attributes directly affect the clustering results, making the distribution homogeneous. This is essential for representing clusters of criminal data, because we can then consider violent regions as belonging to the same cluster taking into account their geographic location.

Other algorithms for the data analysis can be adopted, such as Multidimensional Scaling (MDS) [11]. MDS is a set of techniques for analysing objects in a data set by reducing their dimensionality. The similarity measure is usually related to a distance matrix. This way, each object whose distances to the others are represented in a matrix D is projected onto an n-dimensional space in which the Euclidian distances among the resulting points are roughly the same as in D. This technique makes it easy to analyze the distance between specific objects and helps to identify outliers in clusters.

Regarding data visualization, different techniques can be applied in the context of multidimensional data visualization. Data sets that include temporal data are ubiquitous and notoriously difficult to explore visually in an efficient way, especially when they have several dimensions besides time, such as the criminal data [12].

In this context, the parallel coordinates chart can be used. It allows visualizing multidimensional data through parallel axes in a 2D chart [13][14], where each axis represents a dimension or attribute. For instance, in the case of criminal data visualization,each line represents a police station, intercepting each axis at its corresponding attribute values. It is thus possible to analyze information of n attributes in a single chart. This technique is closely related to the analysis of time series, although each axis does not represent points in space. There is no natural order of the axes, which allows the user to modify the order in which they are displayed to allow for better data analysis. Thus, it is possible to uncover characteristics which could be previously hidden by the juxtaposition of lines in the initial order.

Parallel coordinate charts have some limitations. Even with an average-sized data set, it suffers from overplotting, making it difficult to identify characteristics, trends or patterns. Moreover, as the axes do not have a single order, finding a good order requires heuristics and experimentation. Instead of visualizing each item in the data set separately, one can use the aforementioned clustering algorithms to group them. To visualize those clusters, we can associate a certain color to each one, or even apply to the lines a transfer function to highlight certain characteristics [15][16]. Some solutions adopted in CrimeVis for this problem are described later in this paper. In addition, the parallel coordinates chart can be a powerful tool when coupled with traditional visualization methods, such as scatterplots and time series.

The synchronous use of multiple views can allow interactive exploration across them, through the techniques of "brushing and linking"[17], where the selection of an attribute in one view is used to highlight some characteristics in another view. The coordinated use of various views can provide the specialist with a rich strategy to analyze patterns in a chart and their projection onto another one. An analyst can use such views in various ways, in order to, according to some criteria, filter information, analyze relations among data, and answer several relevant questions. Analyzing multiple views requires a certain effort from the user to interpret patterns, but this technique is more efficient than presenting a single, independent but confusing visualization portraying multiple dimensions.

The visual analysis of data usually requires several different graphical

tools to be successful. In our system these tools must work in conjunction to help the user draw feasible conclusions from the data. Brushing and Linking [18, 19] has proved to be and invaluable technique to achieve this goal. As a result, several frameworks for data analysis have been proposed with a brushing and linking component as its central piece. Demir et al. [20] makes use of this technique by allowing the user to select interesting regions in a line/bar chart view while highlighting the corresponding attributes in a tridimensional view of the datasets. Chen et al. [21] uses brushing and linking for selecting points of interest in a bidimensional projection of a ensemble data and showing the geo-location, uncertainty histogram and parameters of the brushed points. Potter et al. [22] allows the brushing of geographical regions in a bidimensional view and shows quartiles and filmstrip summary views of the ensemble.In our work, this technique is used do highlight criminal attributes in a dataset and show this data in a set of coordinated views, such as parallel coordinates and tematic maps.It's expected that this technique may help experts to identify structures and trends in the analysed data.

2.3 Related Work

To justify the implementation of the proposed software and the adopted visualization tools we conducted a comparative study of recent solutions for the specific domain and for statistical data visual analysis.

To make an efficient use of the application, users should be able to easily answer questions about the data, discover interesting patterns and identify abnormalities in the data [17]. A contextualized analysis made by specialists can assign meaning to trends, dicover relations, and identify outliers in the data set under investigation. In the literature we find several softwares implemented for such purpose, and applied to different domains, such as: criminal, meteorological, or electoral funding data, for example.

Law et al. [23] proposed a tool to analyze patterns and evolution of criminality using a spatio-temporal Bayesian analysis.Their work has the objective to explore Bayesian spatio-temporal methods to analyze local patterns of crime change over time at the small-area level. They build a visual application to analyze the property crime data in the Regional Municipality of York, Ontario, Canada. However, regarding visualization, their software fails to present interactive tools to manipulate and evaluate the data. They show only maps of the investigated area.

Chainey et al. [24] proposed a system that uses the Hotspot Mapping technique to analyze spatial characteristics of criminality. The system maps

criminal data according to where the crimes occurred using a geographic information system (GIS), which allows the analyst to identify patterns and trends for the given areas. Although, the system is efficient with respect to the distribution of crime occurrences as related to the population distribution, it does not consider any socioeconomic characteristics of the area.

Arietta et. al. [25] present a method for automatically identifying and validating predictive relationships between the visual appearance of a city and its non-visual attributes (e.g. crime statistics, housing prices, population density etc.). To do so, they combined Support Vector Regression [26] with Data Mining techniques [27]. Since each such city attribute is associated with a location (latitude, longitude), they typically visualize them as thematic maps. In our system thematic maps are also used to visualize the distribution of an attribute over the DPs.

Similar to our approach, *Crime in Chicago* [28] is a data visualization that lets users explore crime trends in Chicago's 50 wards. It was built using open data about Chicago crimes released by the Chicago Police Department. With this website, users can compare crime levels over the years and across city wards. Although it does not consider socioeconomic attributes and its not possible to link this data with the population only by analysing the charts displayed. In addition, they only consider crimes in the city and not the effects of the public polices over others regions.

Some applications allow filtering information according to certain data, either considering a period of time or specific attributes. CommonGIS [12] is a powerful system for analyzing spatiotemporal data. It encompasses a wide range of tools, such as animated charts, time series, maps related to spatial attributes, among others. It also allows us to use coordinated views and space-time cubes, a technique drawn from cartography, in which the visualization of spatiotemporal data is performed through a cube, presenting both geographic information in two dimensions together with time in a third dimension. The system was created to be general, so it could be used in different contexts, including criminal data.

Connect 2 Congress (C2C) [29] presents a bidimensional political spectrum for the American Congress. Its goal are to analyze voting of projects and map the political profile of congressmen. Data can be filtered by name, state, political party, religion, and gender. The analyzed period can be dynamically changed, resulting in an animation where the representatives are continually organized according to their voting behavior. However, C2C does not present data from different terms; it only allows to analyze data within a 2-year period. Likewise, CivisAnalysis [1] was developed to analyze the Brazilian Congress.

That web-based tool allows the population to evaluate the behavior or political parties and congressmen in different terms. It also overcomes some of C2C's limitations, as it contains a set of tools that can be used in a coordinated manner to filter data.The interface of CivisAnalysis is presented in the figure 2.1.

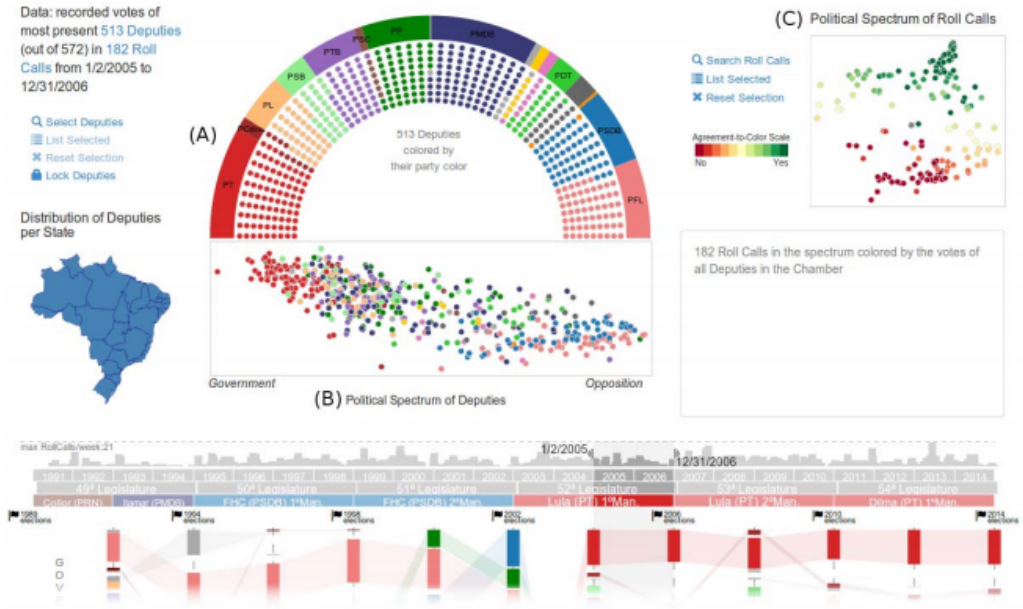


Figura 2.1 – CivisAnalysis's Inspection module, depicting the Brazilian Chamber of Deputies in 2005-2006 [1]

Regarding coordinated views, Ensemble-Vis [2] was developed to analyze simulations of meteorological data. It provides a set of coordinated, interactive tools for statistical analysis. Their goal was to provide scientists with a practical way to analyze uncertainty associated with the data under investigation. The interface of this software is presented in the Figure 2.2.

Our proposal incorporates several tools described in this section in order to provide more flexibility and analytical power to the analysis of the specific domain of criminal data. In the next section we describe the visualization system named CrimeVis.

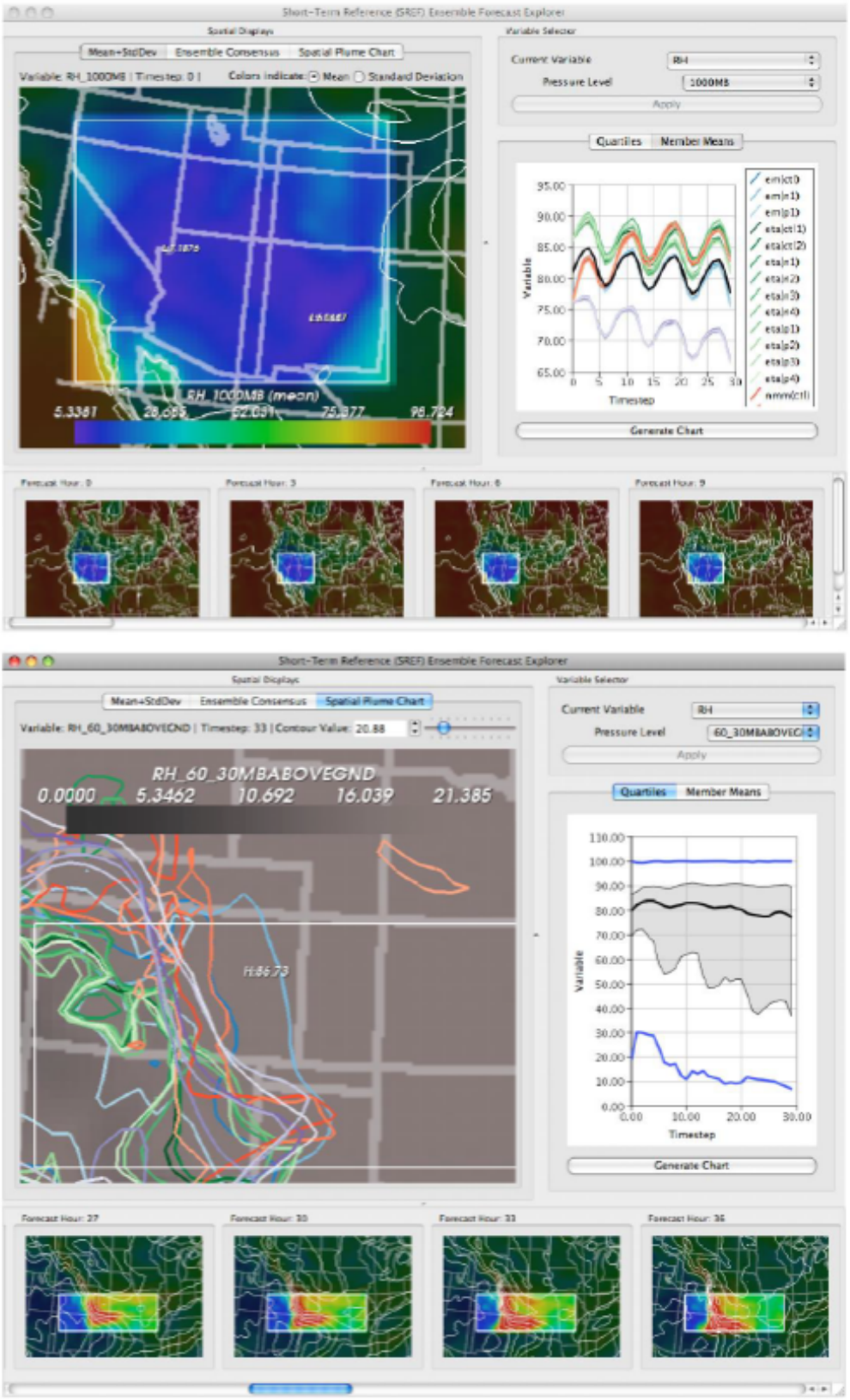


Figura 2.2 – Two coordinated views in Ensemble-Vis [2]

3
CrimeVis

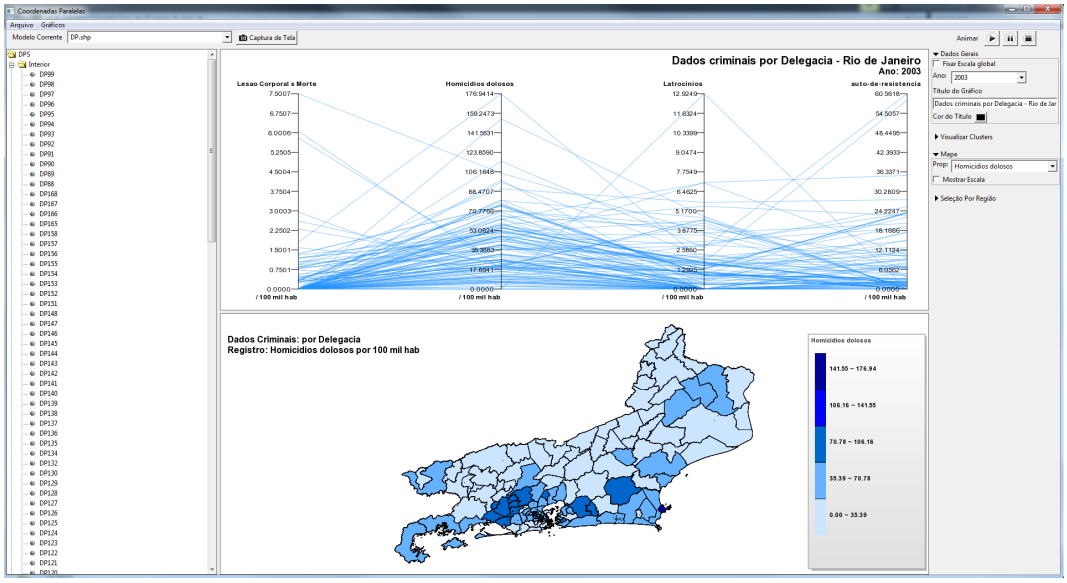


Figura 3.1 – Initial prototype screen of the CrimeVis visualization system.

CrimeVis was developed to help researchers to visually explore criminal and socioeconomic data obtained from ISP-RJ and IBGE. The criminal data are public and can be accessed through a web application made available by the state government [6]. The socioeconomic data are also public and made available by IBGE [5]. We considered as relevant attributes: education, ethnicity and family income. To use this data, we created a database to be used for analysis in conjunction with our software.

Our application was designed as a interactive graphical system. It integrates several clustering techniques to help the users to discover relations in the data. This is a free desktop application developed with the aid of experts in the field, where the design was driven by their needs. Furthermore, the tool architecture has been developed to be extensible, where is possible to add new forms of visualization and data analysis. This tool has a minimalist interface on which the users are free to interact with the charts just by simple clicks or by the selection of a view. Basically any interaction with the graphs can be done by mouse commands.

Each adopted visualization chart has been chosen to answer important specialist questions efciently. Nevertheless, the difference of this application to others is the ability to combine different visualization techniques to present data synchronously between different graphical environments. Our system has the capability of allowing experts to make a deeper analysis of the data, facilitating the discovery of patterns or the identificaton of abnormalities in the data. Our prototype features sophisticated interaction, including automatic brushing and linking of the data and animated transitions in thematic maps. All data are loaded on demand, thus our system is free from CPU and memory intensive processes.

The program was modeled to answer several questions, such as distribution by police station and its relation with the social characteristics of the population. The system was developed using iterative prototyping with a simple lifecycle [30]. The first cycles involved evaluation with potential users, who are researchers of the criminality domain. To satisfy the user's needs, the system was iteratively refined according to their feedback.

With the set of tools we developed it is possible to analyze patterns of criminal distribution along a certain period of time, as well as their social implications, in addition to answering complex questions (Research Tasks - RT), such as:

1. What locations can be considered concentrations of certain types of crime?

2. How have criminality rates evolved over time? And what is their relation with social characteristics of the population?

3. How can we subdivide the state areas according to socioeconomic and criminality criteria?

4. What are the practical effects of the policy of deploying Pacifying Police Units (Unidades de Polícia Pacificadora - UPPs)?

5. Is there any inconsistency in the data made available by the government?

In the next section we present the software in detail, discussing the forms of interaction and how we can answer each of those question .

3.1 Overview

The initial screen of CrimeVis presents a module for data inspection in a period of time. The available options allow us to inspect data over a period of 12 years (2003-2015), the evolution of criminality rates in the state, as well as socioeconomic data (RT-2).

The goal is to give to the users an overview of how criminality behaved. To obtain specific details about a period, users can select arbitrary time frames on the a panel in the right and clicking on a specific time range. The system initially presents to users a parallel coordinates chart in which each line represents the information of a DP and a map with the geographic distribution of the DPs, as can be seen in Figure 3.1. This can be used to answer the question RT-1. The system also offers other views, such as time series charts, MDS projection, scatterplot and 3D parallel coordinates chart. All the views are synchronized, i.e., the user actions in one view are reflected in all the others. In this way, it is possible to identify relations in the data and filter specific information simply by clicking on the attributes of one of the charts and observing how these data are projected on another.

CrimeVis provides interactive filtering, letting users create selections and compare them using the coordinated views.A single DP or groups of DPs can be highlighted in all views by left-clicking. Also through a parameter tree, the users can remove or add data from the visualization just by right clicking over the options. For example a parameter such as homicides or scholarship can be add or removed from the visualization and all the views are modified on-the-fly. Our system provides others filters as selection by region, group by crime rate or socioecnomic data or by a user approach. To load user data, we allows the user to use the results from R scripts created by a preprocessing step.

It is also possible to group the data set in clusters to present them visually. For this purpose, we have used three techniques: K-Medoids, SKATER, and a combination of MDS and K-Medoids. When visualizing a set of data that considers homicides, thefts, missing persons, and population income, we can group the data according to all the attributes (RT-3) together or to each one separately. By knowing in which group each DP is located, we can analyze properties specific to each group, as well outliers and anomalies (RT-5) in the input data.

To conclude, CrimeVis provides a simple user interface, which allows us to explore multivariate spatio-temporal data sets, to easily investigate hypotheses and to identify patterns.

3.1.1 Data Clustering

The main goal of data clustering is to discover the natural grouping(s) of a set of patterns, points, or objects. Cluster analysis can be defined as a non-supervised machine learning task, which is used to discover whether the elements of a set fall into different groups by making quantitative comparisons of their multiple characteristics.

DPs are clustered considering a given period of time and the set of attributes chosen by the user. With this data input, the system allows us to adopt one of three strategies for data clustering: K-Medoids[10], SKATER[3], and MDS[31] + K-Medoids. This strategy help us to support questions as the RT-3.

The objective of this approach is to automatically discover groups in the analysed data. Clustering data is an important step for knowledge discovery in data, which allows the users to easily identify patterns and tendencies, helping in decisions based on the data analysis.

In our application, the data are clustered for a specific period of time and the set of attributes of each DP chosen by the user. Given a period of time, and considering the n dimensions of the data, the system allows us to adopt one of three strategies for data clustering, as mentioned before: K-Medoids, SKATER, and MDS + K-Medoids. The three clustering strategies described in the next sections can be used and evaluated according to the set of tools available in the system, combined with charts and aiming to extract relevant patterns for the researchers.

3.1.2 K-Medoids

The k-medoids clustering algorithm [10] is a variation of the k-means[8] algorithm. The k-means is a partition algorithm that search directly for optimal division (or approximately optimal) of n elements using an interactive process to group them in k clusters based on a dissimilarity function. The k-medoids algorithm is also partitional and attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers instead of artificial centroids.

The most common implementation of the k-medoids clustering is the Partitioning Around Medoids (PAM) algorithm [10], which is:

1. Randomly select k of the n data points as the medoids;

2. Associate each data point to the closest medoid;

3. For each medoid m and each data point o associated to m swap m and o and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated to m);

4. Select the medoid o with the lowest cost of the configuration;

5. Repeat the steps 2 and 3 until there is no change in the assignments.

In our context, this algorithm was used to select a set of clusters of police stations with similar criminal rates. In our system we uses the Euclidian distance as a measure of dissimilarity. Users can also set the number of clusters to be created, but through some experimentation we have concluded that, for this context, a good number of clusters is five. We got to this number by estimating an optimal average silhouette width [32] and the Calinski-Harabasz index [33].

Unfortunately, this algorithm does not consider the spatial adjacency relations of police stations. It groups the DPs in clusters based only in the crime rates and socioeconomic variables, disregarding the influence of the dispersion of crimes according to a neighborhood region.

3.1.3 SKATER: Spatial 'K'luster Analysis by Tree Edge Removal

To overcome the recently mentioned limitation of k-medoids, another clustering strategy called SKATER has been proposed [3]. It creates spatial clusters based on the data and on a dissimilarity function.

The aim of traditional unsupervised classification is to partition a set of n items into k clusters such that items within the same cluster should be similar to each other and items in different clusters should be dissimilar from each other. The dissimilarity refers to a set of attributes measured in each item. When these items have a spatial location determining a neighborhood structure, it is usual to be interested in their clustering constrained by spatial contiguity. That is, we want to partition the n items into internally homogeneous clusters with respect to the attributes but the items within a given cluster should be constrained by the spatial neighborhood structure, normally defined by geographical adjacency.

SKATER first constructs a spatially contiguous graph by removing edges that do not connect spatial neighbors and then builds a minimum spanning tree (MST) [34] from the graph based on pairwise dissimilarities between the

nodes. The spanning tree is then recursively partitioned to generate a given number of regions based on a global heterogeneity objective function.

In this technique using a connectivity graph, we capture the adjacency relations between regions. In the graph, each police station is considered as a vertex and linked by edges to its neighbours. The cost of each edge is proportional to the dissimilarity between the objects, where we measure dissimilarity using the values of the crime rates of the neighboring pair. By cutting the graph at suitable places, we get connected clusters. In this way, we transform the regionalization problem in an optimal graph partitioning problem. This proposal, first presented by Assunção et. al.[3], is to limit the complexity of the graph by pruning edges with high dissimilarity. The pruning produces a reduced graph that defines a small number of possible partitions and whose edges join similar areas considering their attributes. In the reduced graph, further removal of any edge splits the graph into two unconnected subgraphs. To carry out this idea, the reduced graph is take to be a minimum spanning tree (MST). After creating the MST, we obtain clusters by partitioning this the tree.

To build the graph consider contiguous geographic units, such as the 138 DPs shown in 3.1, as a set of spatial objects O with a set of attributes (homicide, thefts, missing persons, etc.) p associated to a object O_i. The arrangement of the set determines a connectivity graph $G = (V, L)$ with a set of vertices V and a set of edges L. There is an edge connecting vertices v_i and v_j if areas i and j are spatially adjacent. A $costd(i, j)$ is associated to a edge (v_i, v_j) by measuring the dissimilarity between the objects i and j. An usual choice for the dissimilarity measure is the square of the Euclidean distance between the attribute vectors pi and pj. A path from node v_i to v_k is a sequence of nodes $(v_i, ..., v_k)$ connected by edges $(v_i, v_{i+1})...(v_{k-1}, v_k)$. A spatial graph G is connected if for any pair of nodes there at least one path connecting them. So, a spatial cluster is a connected subset of nodes. The main objective is to accomplish the partition of the graph G into C spatial clusters where their union is G, and each one is a connected subgraph. A spanning tree T of a graph G is a tree containing all n nodes of G, where any two nodes are connected by a unique path, and the number of edges in T is $n - 1$. The removal of any edge from T results in two disconnected subgraphs that are spatial clusters candidates. This is a spanning tree with minimum cost, where the cost is measured the sum of the dissimilarities over all the edges of the tree. To generate this so called minimum spanning tree (MST), we use a recursive implementation based on Prim's algorithm [34].

To generate the partition of n objects in k regions, it is necessary to

remove $k-1$ edges from the MST. Each resulting cluster will be a tree with all vertices connected. To make the partition, the algorithm uses a hierarchical division strategy. In the first step, a single tree contains all the objects. When edges are removed from the MST, we have a set of disconnected trees, and each tree correspond to a region. At each iteration, one of the trees is subdivided into two others trees by cutting an edge, until we have the number os cluster previously set. The partitioning algorithm produces a graph G^* that contains a set of trees. Initially, we have just one tree in G^*, but at each iteration, the graph is examined, and we cut one edge that will divide the tree T int two other trees. The selected edge have the largest increase in the overall quality of the resulting clusters. For this, a quality measure is defined by the sum of the intracluster square deviations, and this measure is the one that needs to be minimized:

$$Q(\alpha) = \sum_{i=0}^{k} SSD_i, \tag{3.1}$$

where α is a partition in K trees; $Q(\alpha)$ is the value associated with the partition and SSD is the sum of square deviations of a region. The intracluster square deviation SSD is a measure of dispersion of attribute values for the objects in a region, and homogeneous regions have small SSDs values. The intracluster square deviations is defined by 3.4:

$$SSD_k = \sum_{j=1}^{m} \sum_{i=1}^{n_k} (x_{ij} - avg(x_j))^2, \tag{3.2}$$

where n_k is the number of objects in the tree k; x_{ij} is the j^{th} attribute of a object i, m is the number of attributes, and $avg(x_j)$ is the average value of the j^{th} attribute for all objects in the tree.

At each iteration, we need to subdivide the graph G^*, i.e, remove and edge of the graph that contains a set of trees $T_1, ...T_n$. The solution for a best subdivision of a tree T is defined by the objective function defined by the equation 3.3

$$f(S_l^T) = SSD_T - (SSD_{Ta} + SSD_{Tb}), \tag{3.3}$$

where S_l^T is the arrangement produced by cutting out the edge l from the tree T, and Ta and Tb are the two trees produced by dividing T after cutting the edge l. In this algorithm we divide the tree that has the highest value for the objective function. Starting from the MST, the clusters are produced as follows:

1. Start the graph $G^* = MST$;

2. Identify the edge that has the highest objective function;

3. While $size(G^*) < k$(desired number os clusters), repeat the next two steps;

4. For all trees in the graph, select the tree T_i with the best objective function

5. Split T_i into two new subtrees and update the graph.

Figure 3.2 from Assunção et. al [3] shows the method with the first 3 iterations.

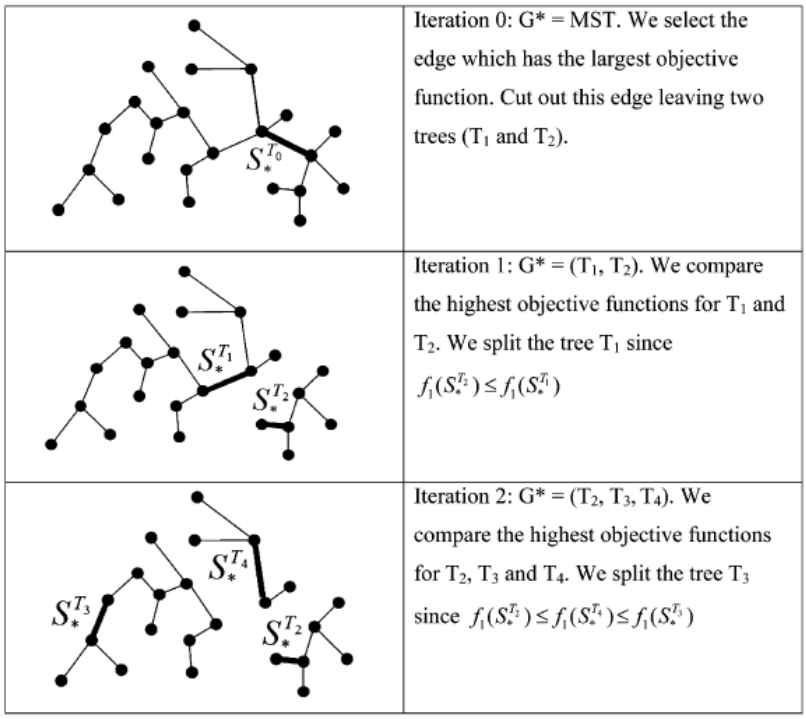
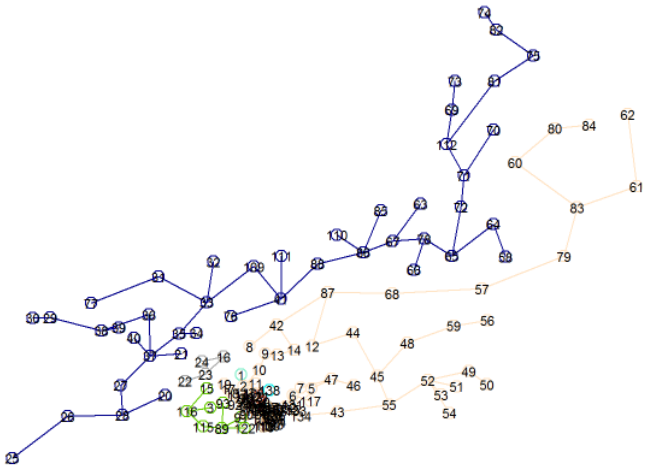


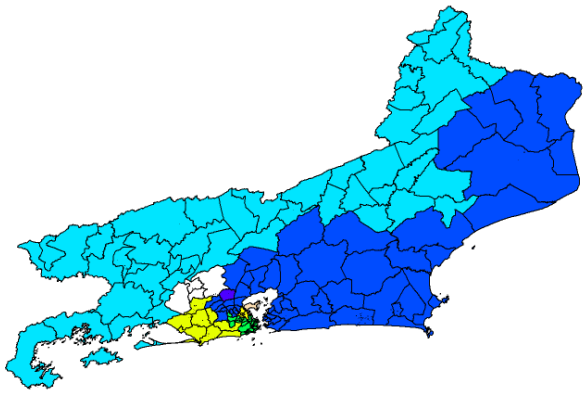
Figura 3.2 – Partitioning of the MST [3]

This algorithm has the advantage of maintaining the spatial relation instead of considering only the data attributes, resulting in better data analysis. With this technique, the researchers can evaluate the evolution of crime rates over a region and their influence in others areas. They can analyze the dispersion or changes in the criminal activities and identify the possible evidence of its causes. In other words, identify a possible set of standards and structures in the data considering the geographical distribution of police stations.

In the context of the evolution of criminality, this technique shows to be more efficient than traditional clustering algorithms. Figure 3.3 shows the algorithm applied to Rio de Janeiro's crime rates considering the distribution of police stations.



(a) Clusters generated by the algorithm



(b) SKATER clustering for Rio de Janeiro Police Stations

Figura 3.3 – Clusters of DPs in Rio the Janeiro using the Skater Algorithm. The first figure shows the partition of the MSP and the second, the final result.

3.1.4 Multidimesnional Scaling

Finally, data can also be clustered by combining the MDS[31] and the K-Medoids[10] algorithm. MDS aims to project the data in an n-dimensional space so that the distances between data points remain approximately the same [11]. To achieve this, the algorithm processes a distance (or dissimilarity) matrix between every data pair and searches for a projection that minimizes the cost function. The primary outcome of an MDS analysis is a spatial configuration, in which the objects are represented as points. The points in this spatial representation are arranged in such a way, that their distances correspond to the similarities of the objects: similar object are represented by points that are close to each other, dissimilar objects by points that are far apart [35]. One of the applications of MDS is to use it for visualizing correlational data. Even

considering few objects, such a matrix become complex, and it is hard to detect patterns of correlation. The MDS solution plots the objects on a map, with their correlational structure is accessible by visual inspection. In other words, considering our contexts of criminality, with this technique we can easily found correlations between the crime rates. From the data alone it is not easily to see which crime rates are related [35]. Figure 3.4 shows a MDS projection which simplifies this task. The distances between points in the figure correspond to the correlation coefficients, so that a high correlation is represented by a small distance, and vice versa.

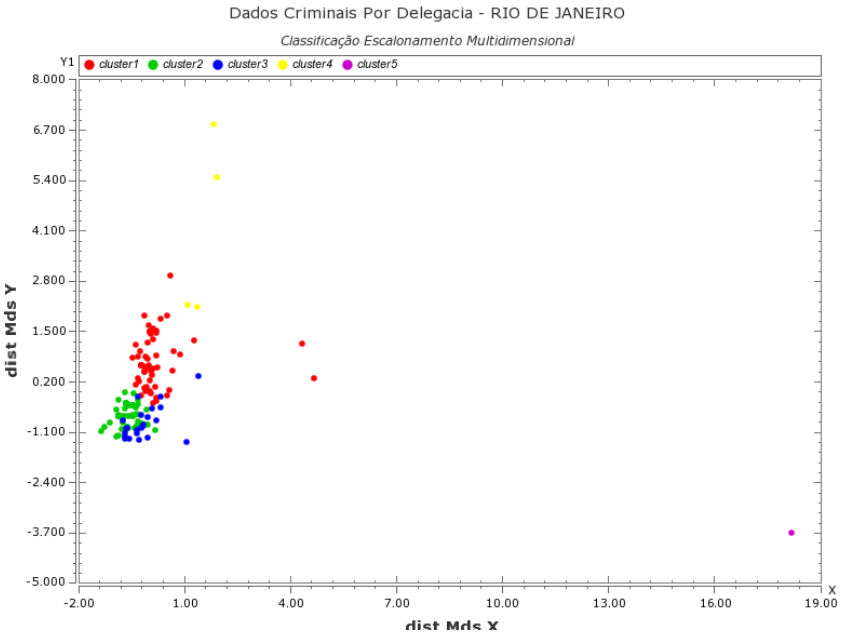


Figura 3.4 – : A two-dimensional MDS representation of the correlations of crime rates in Rio de Janeiro

Consider the classical MDS approach. Given a set I of objects for which there is a distance metric *delta* defined, the dissimilarity matrix between the objects is given by Equation 3.4.

$$A_{m,n} = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix} \tag{3.4}$$

This technique can be also modeled as an optimization problem where the algorithm attempts to minimize a strain, or stress function S contained in terms of D and the euclidean distance between a set of projected points X. S is given by the equation 3.5.

$$S = \min_{x1,\ldots,x_I} \sum_{i<j} (||x_i - x_j|| - \delta_{i,j})^2 \tag{3.5}$$

In the classical MDS approach, the algorithm provides a mean to visualize high dimensional data in a lower dimensional space preserving the distances between entities. This facilitates the analysis of relationships between individual data points in a chart, for example.

One important remark is that the number of dimensions has a direct effect on the preservation of the original distances between the points. In general, larger values of dimensions produce lesser errors on the final projection. Another important property of the projected points is that they can be rotated and translated freely in a 3D chart, as long as the distances between them are preserved. Some approaches, like ours, are often employed to fit the points in a set of meaningful axes to aid the interpretation of results.

The classical MDS algorithm can be described by the following steps:

1. Let D_2 be the squared distances $[d^2]$ in D;
2. Let $B = -\frac{1}{2}JD_2J$ be a double centered version of D_2 using the centering matrix $J = I - \frac{1}{n}\mathbf{1}$;
3. Extract the N largest positive eigenvalues of B and their corresponding eigenvectors;
4. Let Λ be the diagonal matrix with the N larges eigenvalues of B and let E be the matrix with the corresponding eigenvectors. Let $P = E \times \Lambda^{\frac{1}{2}}$ be the projected coordinates.

To cluster the data, the k-medoids algorithm is modified to consider not only the attributes of each data, but also to contemplate the output values of MDS containing an associated weight. These combined techniques allow us to easily identify outliers and inconsistencies int the analyzed data.

In conclusion, with these methods, researchers can recognize dissonant standards in the general context of the data. In particular, they can find police stations who have the highest crime rates and are considered to distant from the others. In that way, anomalous cases can be analyzed in a more detailed way after the identification.

3.2 Visualization and Inspection Module

The visualization module provides a set of graphical tools which include 2D and 3D parallel coordinates charts, scatterplots, time series, and MDS projections. The views present important variables that should be considered in data analysis: criminality attributes, socioeconomic attributes, clusters of

DPs and their spatial distribution. These variables are interrelated: a point in the scatterplot represents a DP, which corresponds to a set of attributes in the parallel coordinates chart. All views are synchronized to express subattributes in the data set, in which any user action in a view affects all the others.

3.2.1 Parallel Coordinates

The parallel coordinates chart became a very important visualization tool offered by the system. It organizes several data dimensions (attributes) as parallel axes next to one another in a plane (see Figure 3.5). This chart provides an overview of the relations between different attributes. Given the context, through this visualization, criminality data can be related to socioeconomic data, achieving one of the key goals to answer important questions posed by most researchers in the domain.

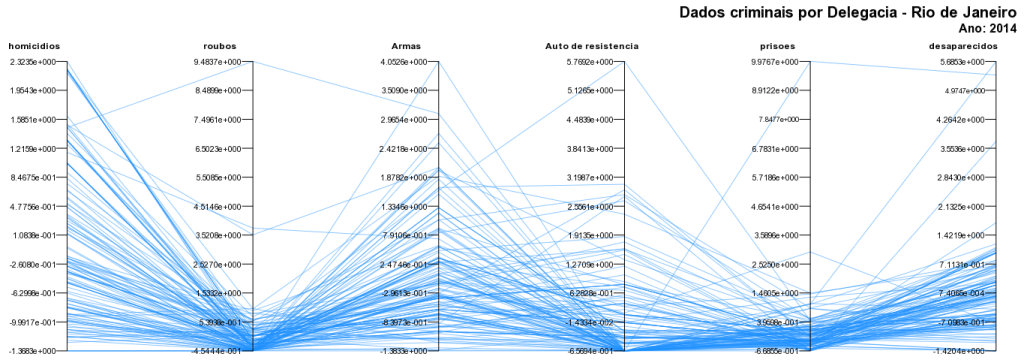


Figura 3.5 – Traditional Parallel Coordinates

The basic interacion with parallel coordinates includes the possibility to get tooltip information, re-order the axes by dragging them with mouse click, the brush mode, which allows the user to select records and to highlight them. The linking mode shows the corresponding information of a DP or a group of DPs in other synchronized chart. Users can select a DP by clicking and hovering over them displays more informations in a tooltip. Figure 3.6 shows a selected line representing the DP 58 in the city of Nova Iguaçu.

Still regarding the visual analysis of data, the traditional parallel coordinates chart makes it difficult to identify some characteristics due to the juxtaposition of the curves. To resolve this, we implemented a bundled curve model [36] that uses curve geometry to improve the visibility of structure in the data over multiple axes. To do so, we replaced the polyline approach by Bézier curves. When the curves are bundled based on cluster membership, structure within clusters can be compared. The curve construction guarantees the continuity of the curves, particularly at the coordinate axes, alleviating

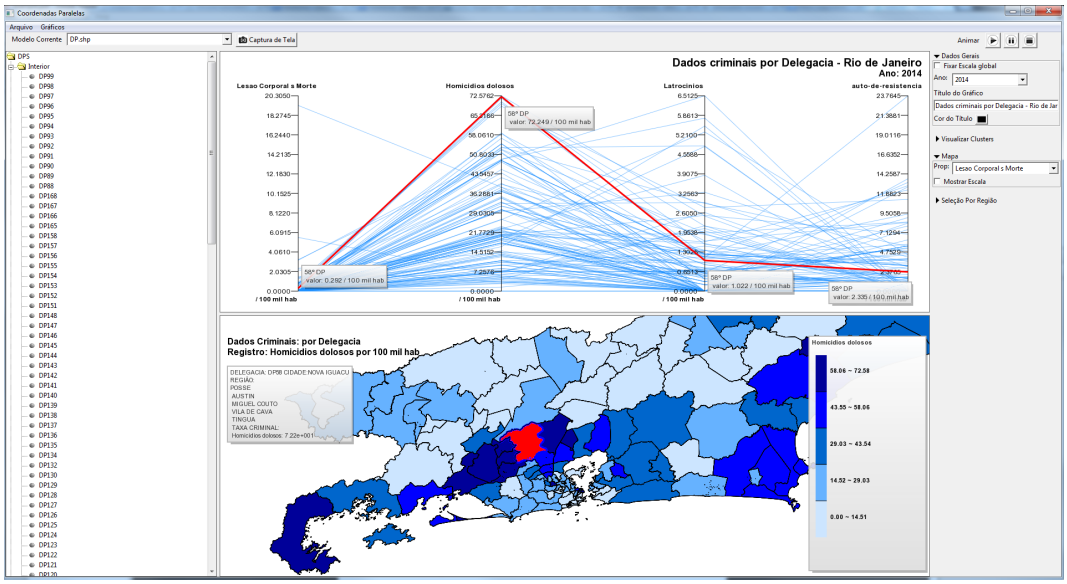


Figura 3.6 – The selection mode of parallel coordinates working with the "Brushing and Linking"mode. The highlighted DP in the first view is also selected in the map.

the overploting problem. At the same time, curve control points are chosen to obtain good approximation to the original polylines. The bundled curves model is designed in order to maintain not only the desirable characteristics of the polyline plots used in the original Parallel Coordinate chart, but also its ability to reveal correlations between variables.

In this approach, control points of each curve are influenced by each cluster's centroid, obtained by one of the three aforementioned strategies, and by the points in which each curve intercepts each axis. To build a curve between two adjacent axes x_i and x_{i+1}, we inserted an imaginary central axis between axes x_i and x_{i+1} in which point C_i is the point where the cluster centroid intercepts this axis. This point is used to attract the curves of the corresponding cluster and, as a result, the original line, which would pass by Q_i now passes by Q'_i. Auxiliary axes are also added to smooth the curve drawing in a distance d of each axis. A polyline is replaced by a Bézier curve with the following properties:

1. The curve interpolates the points $P1, P2, ..., PN$ at the value axes;
2. The curve is continnuous throughout;
3. Curves corresponding to data points that belong to the same cluster are bundled between adjacent axes. This is accomplished by inserting a bundle axis midway between the axes and by appropriately positioning the Bezier control points;

4. The cluster centroid is the projection of the centroid on the plane defined by the axes, intersected with the bundle axis.

As an illustration, Figure 3.7 presents a more clearly defined view of the sets created by the process of clusterization when compared to the traditional parallel coordinates chart in the Figure 3.5, in which each group can be selected and analyzed separately.

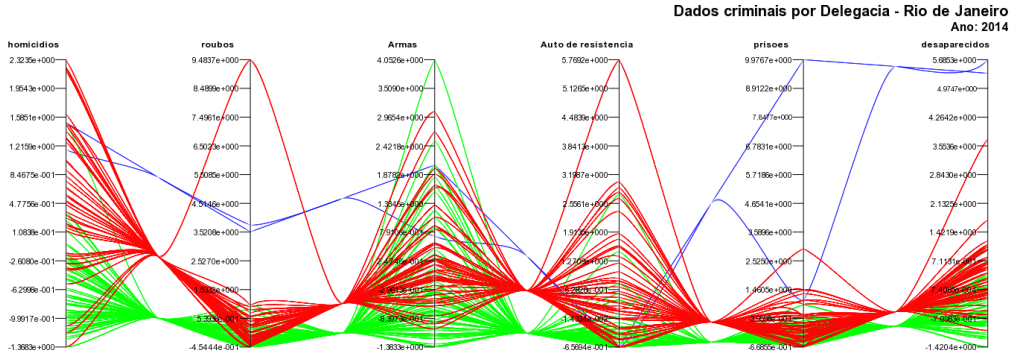


Figura 3.7 – Clusters in Parallel Coordinates.

CrimeVis also provides the 3D parallel coordinates chart, which is a straightforward extension of the traditional 2D chart. The idea to extend the parallel coordinates plot in 3D space is natural, so some author explore different implementations of it [37].

Our basic idea is use the third dimension to explore some patterns that went unnoticed in the original chart. In the Z axis the data are sorted by his geographical region or a value property defined by the user. Each parallel axis of the original chart is extended in a third dimension, forming a plane that represents a 2D scatterplot relating properties A and B. As with 2D parallel coordinates, the points corresponding to the neighbor axis are, then, connected by line segments.

For the cluster analysis, the 2D parallel coordinates chart, even when using Bézier curves, often obscures relevant characteristics of the data, as in some structures we have investigated. Thus, the 3D chart aims to support the identification of relevant details that would not be easily noted in the 2D chart. For each scatterplot, the values in Z represent an attribute chosen by the user, and in the Y axis the values are maintained from the original chart. Figure 3.8 presents the proposed 3D chart, in which the geographic distribution of DPs in the state is chosen for the Z axis. Given the geographic location of each DP, they were sorted according to the following distribution: Baixada Fluminense, Interior, Grande Niterói, and Capital. The user can filter data, either by selecting a specific set of lines or a single line. Some degree of

transparency is applied to the lines that were not selected so as to highlight the data of interest to the user.

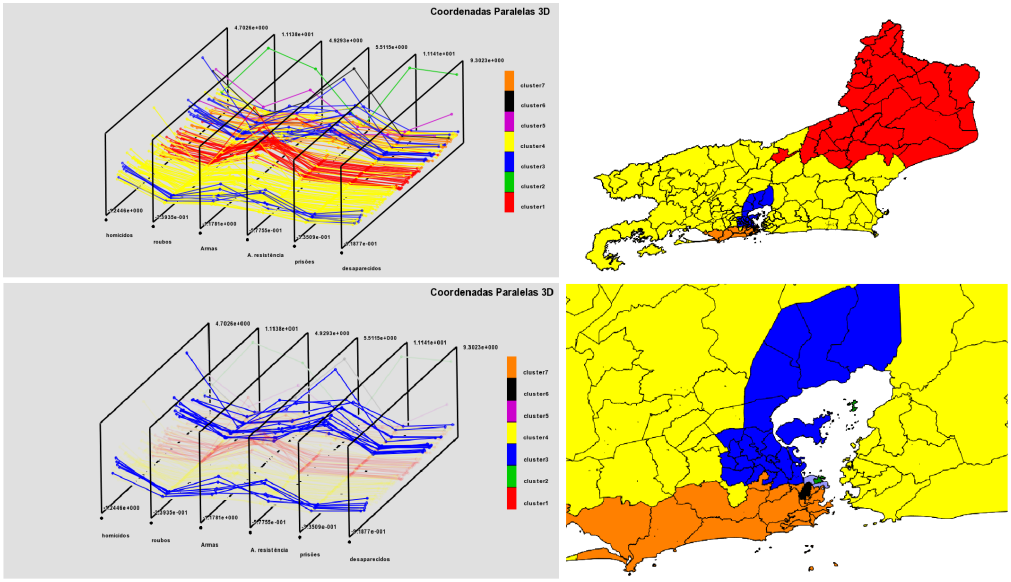


Figura 3.8 – 3D parallel coordinates and their relation with the spatial distribution.

3.2.2 Map Analysis

Thematic maps can also be used to observe a certain variable in a particular period of time. This kind of visualization allows us to answer several questions related to the specified period of time (RT-2, for example), because the map as a whole aims to present all the distribution of values of the selected attribute in that time period. Digital maps are the quickest means of visualising the entire crime scenario. The locations of crime events, arrests, etc. can be routinely displayed on maps. This provides an easy method of viewing activities in an area rather than searching through a listing of events [38].

This kind of visualization allows us to answer several questions related to the specified period of time, because the map as a whole aims to present all the distribution of values of the selected attribute in that time period. To analyze the values in a certain location, the user may simply click with the left mouse button over that area. This forces the system to present the corresponding attribute values, such as the DP name, the city, the neighborhoods, socioeconomic and criminality variables as a tip located on the left-hand side of the map. In addition, users can select groups of DP by pressing shift and clicking with the left button over them or brushing and can hover over them with the cursor to display correspondent information in other view. Also it is possible to show a time lapse animation to visualize the crime evolution over time and to zoom in the specified area. This animation is demonstrated

in the Figure 3.9. The selection panel provides a map animation, in which it is possible to see the evolution over time of the spatial distribution of the selected attribute. Figure 3.1 contains the view map for the homicide attributes for the year 2003. Maps offer crime analysts graphic representations of crime related issues. This type of visualization can help researchers to understand where and why crimes occur, which is an important tool for the analysis of crime rates evolution and the analysis of public security polices.

3.2.3 Scatterplots, time series charts and MDS projections

CrimeVis also provides three other charts: scatterplots (in 2D or in 3D), time series, and MDS projection. The scatterplots can be used to analyze the correspondence between two variables for one DP and the distribution of DPs in the generated clusters. It allows us to find direct relations between two attributes and easily identify outliers. To help the interpretation, this chart also offers a linear regression analysis. It consists in performing a statistical analysis in order to verify the existence of a functional linear relationship between a dependent variable with one or more independent variables. In other words, it consists in to obtain a linear equation which attempts to explain the variation of the dependent variable through the variation level of the independent variables.

In our application, the simple linear regression model was adjusted using the Least Squares method [39]. Figure 3.10 presents the scatterplot 2D view.

We also offers to the user a scatter plot 3D component for the visualization of multivariate data in a three dimensional space. Basically this graph generates a scatter plot in the 3D space using a parallel projection. To visualize higher dimensions (fourth, fifth, etc) of the data, some extent using, e.g. different colors, symbol types or symbols sizes can be used.

This component is fully interactive, where the user can rotate it freely and select different sets of points. This view is also integrated into the brush and linking system, working in coordination with the other components. Figure 3.11 presents an example of use of the scatter plot 3D view.

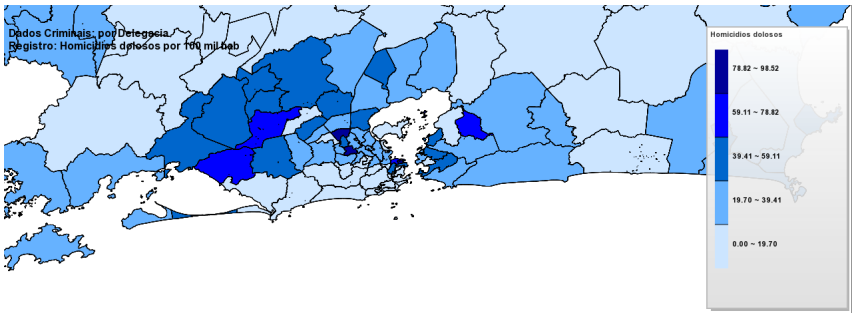
The time series chart is used to analyze an attribute in the whole time range of the data set. It makes it possible to create clusters for a specific variable and observe its evolution over time. Groups can be generated according to the algorithms previously specified. It is up to the user to choose the most suitable for your analysis. Additionally, the user can view data set according to the geographical distribution of the state, that is, according to distribution of DPs in the capital, interior, Baixada Fluminense and Grande Niterói. The

user can also select only one police station to analyse by its selection in the component tree located in the left corner of the window. Also is possible to select a specific curve and highlight it when analyzing a data set. The police station which has automatically highlighted its curve will also be highlighted in other views. The time series plot is presented in Figure 3.12.

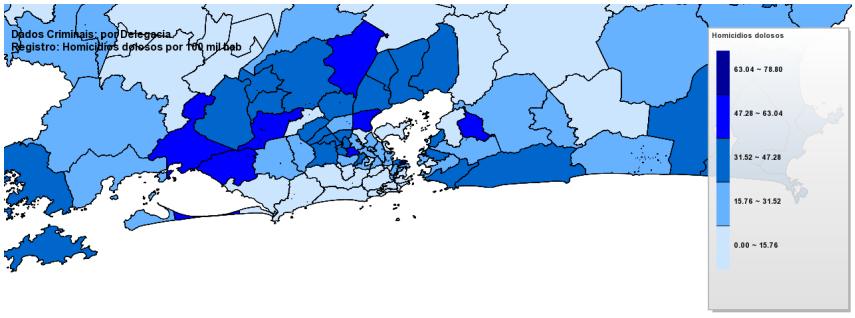
Finally, we have the MDS projection, which is nothing more than a scatterplot in which it is possible to see the data distribution according to the MDS algorithm and its combination with the K-Medoids algorithm for clustering. Figure 3.4 presents the MDS chart in 2D with 5 clusters and the Figure 3.13 a 3D chart with 3 clusters. Through this visualization, we can easily identify outliers. For the 3D chart, the user also has the zoom tool to get a more detailed view. All these views are interactive and allow the user to change them dynamically, either by selecting a specific data set or by altering the selected attributes. In addition, they are all synchronized, reflecting all user actions. For all charts, the users can always select or hover over a DP to get tooltip information, and they can always synchronize this charts with other views.

3.2.4 Filtering Data

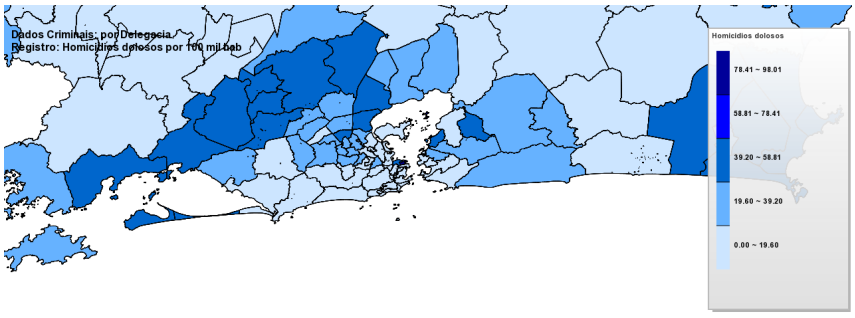
CrimeVis also provides interactive filters through which the user can select a data set and observe its correspondence in another synchronized view. For parallel coordinates, DPs and clusters can be selected by clicking the right mouse button and dragging the cursor. Other selections can also be made through a set of options presented in the user actions panel, such as year and area, besides removing and adding attributes to the views. Each filter operation modifies the views on-the-fly. For instance, when selecting a set of lines in a parallel coordinates chart, the corresponding DPs can also be selected in another active view, such as the map of DPs.



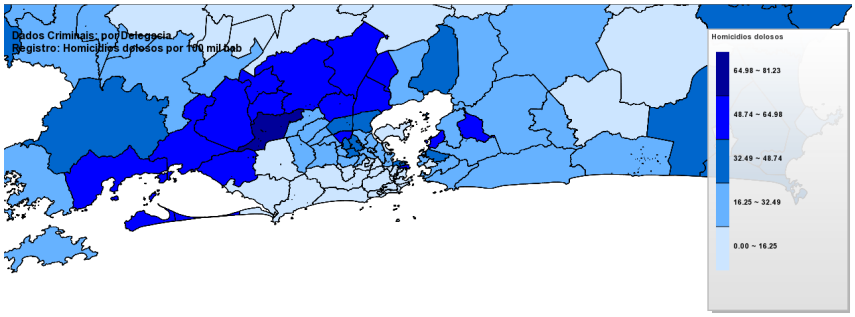
(a) Homicides in 2010



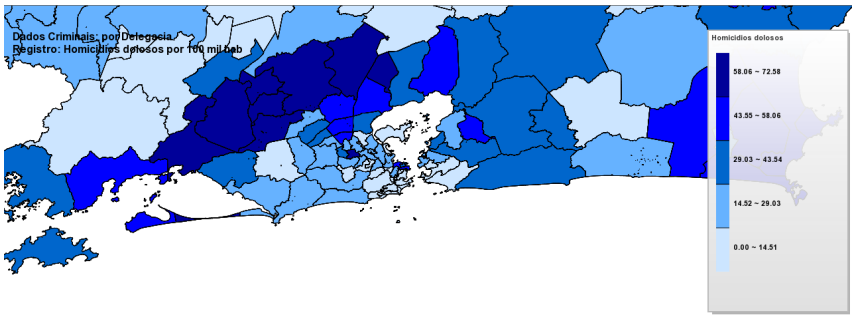
(b) Homicides in 2011



(c) Homicides in 2012



(d) Homicides in 2013



(e) Homicides in 2014

Figura 3.9 – Time lapse animation for 2010 to 2014

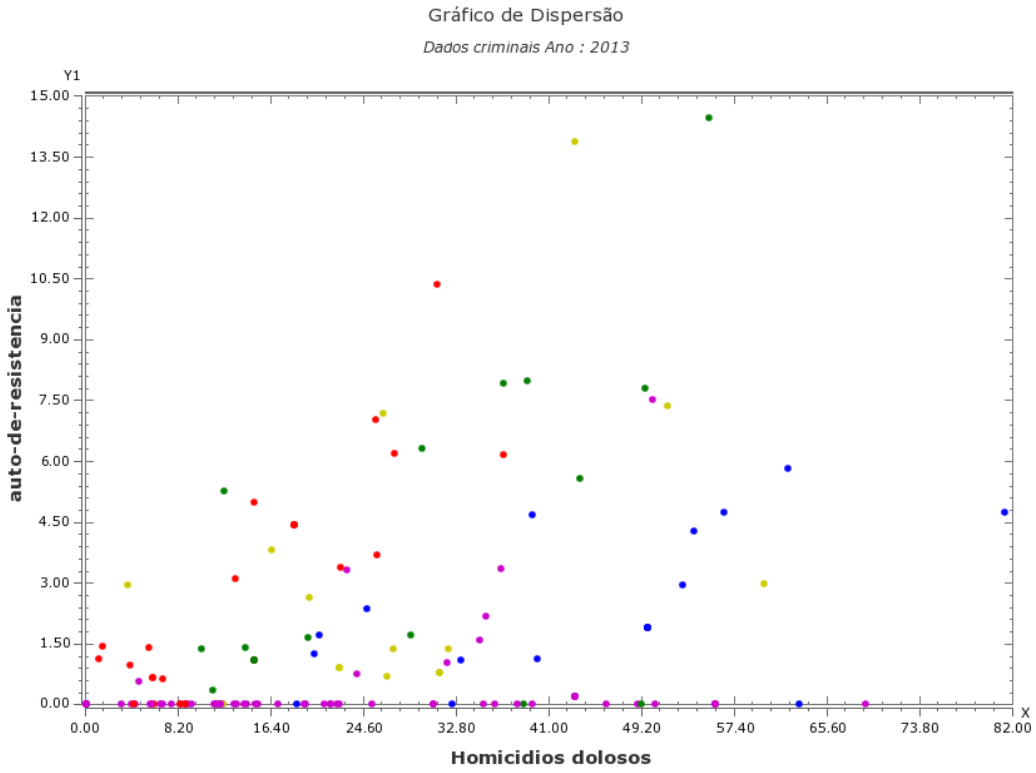


Figura 3.10 – Scatter plot view for the 138 DPs

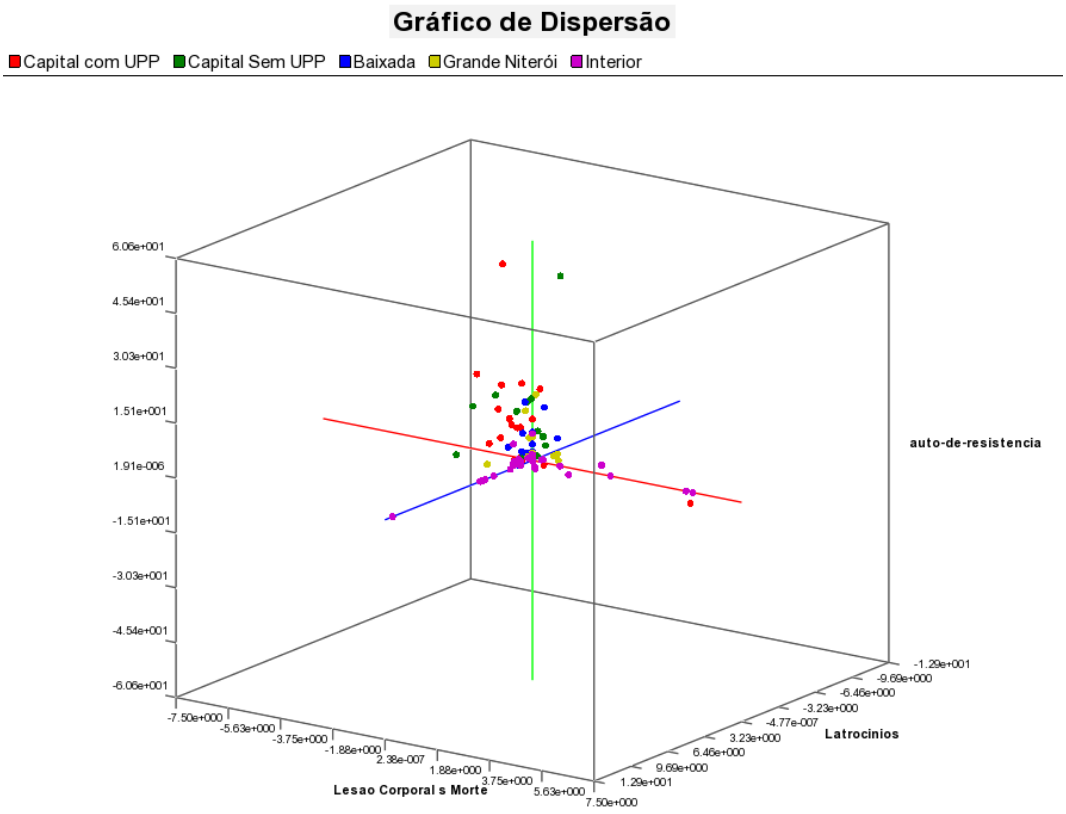


Figura 3.11 – Scatter plot 3D for 138 DPs of Rio de Janeiro

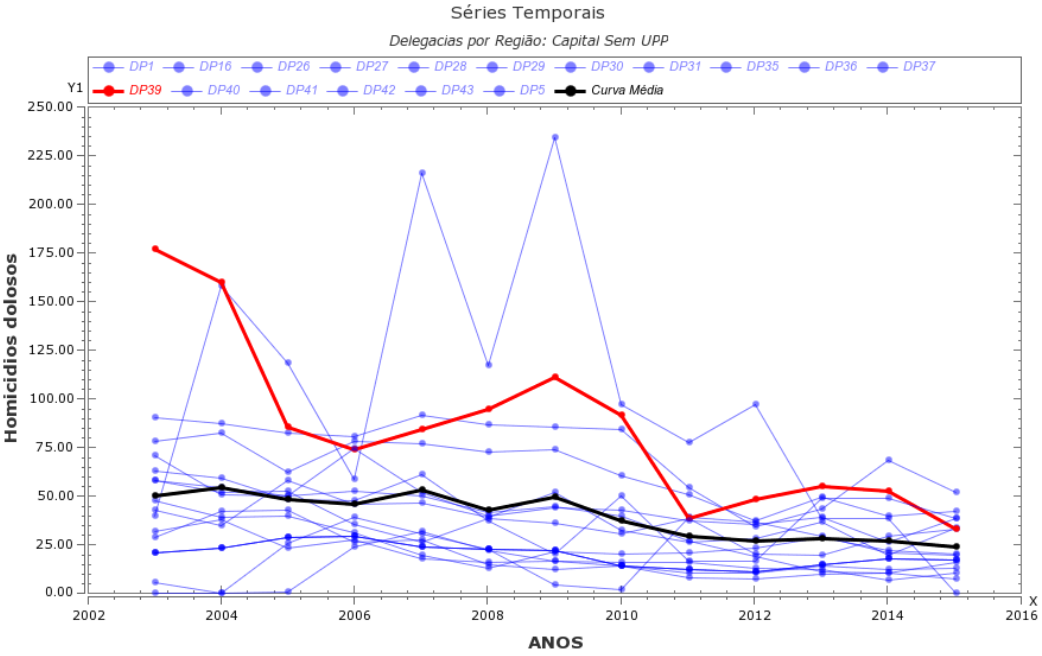


Figura 3.12 – Time Series Plot for murders in regions without upp

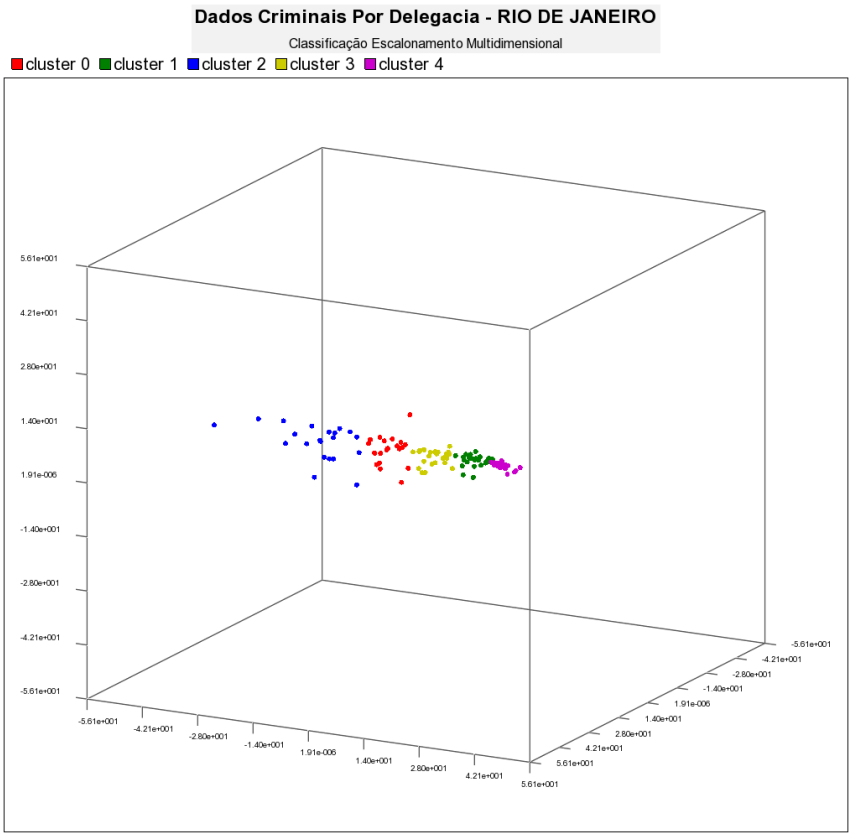


Figura 3.13 – Multidimensional Scaling tool in 3D with 5 selected clusters

4
Evaluation and Discussion

CrimeVis was evaluated throughout its development by users expert on criminality. Its implementation started with the quest for answers to questions posed by researchers on criminality and of the ISP-RJ. Initially we conducted meetings with researchers of ISP-RJ and in academia for gathering requirements for the system. A preliminary study was conducted to evaluate the proposed software, as well as the strategies adopted for analyzing the data. CrimeVis was evaluated by a group of 12 people. The group is formed by students and researchers in the area of criminality and scientific visualization experts. Among participants, 1 of them is expert in statistics, 3 specialist in the field of humanities and 8 professionals are computer scientists and scientific visualization experts. No introduction was given of the system in this test, the users interacted freely with the program for 30 minutes and them answered a questionary with 14 questions. The following sections report on the evaluation and its results.

4.1
Preliminary Study and Evaluation

We have conducted an analysis of the visualization tools, as well as their usability, in order to answer the questions posed before. The study was based on a 5-point Likert scale questionnaire (1 = completely disagree to 7 = completely agree) as int the Table 4.1.

In general, we obtained a positive result in the questionnaire, especially with respect to ease of use and interactivity with 60% positive responses. The main goal of the application, which involves the analysis of patterns and the understanding of the visualizations, was the main point investigated. As expected, the negative feedback was related to the lack of information on the techniques used and the lack of a tutorial. Besides, the 3D parallel coordinates chart was considered in some cases as redundant, because in most cases the original chart was capable of satisfying the users' doubts.

Another point raised by the study participants was the efficiency of the data selection techniques and the combination of interactive visualization with

Tabela 4.1 – Questionnaire for UI Evaluation.

	completely desagree	partially disagree	disagree slightly	neutral	agree slightly	partially agree	completely agree
It is easy to select a time interval using the map component panel							
It is easy to sort the data by using the options for grouping							
I could understood the filtered data with the options offered in each component							
I understand the list of filtered data in a component with the others							
The selection controls for each station are efficient							
The selection controls for each DP are easy to use							
I understood the data presented in 3D parallel coordinates graph							
The results of clustering algorithms were adequate							
I understood the evolution and distribution of criminal data over time							
It was easy to identify patterns and trends when visualizing clusters in parallel coordinates							
It is useful to group data by properties							
I think is useful the way the data are presented on the map component							
It is easy to understand how the date is displayed on the map component							

clustering algorithms. Some users highlighted this characteristic as a strong point of the system when compared to other systems and to R implementations. Moreover, according to the study, CrimeVis achieved the purpose of making it efficient to discover patterns and correlations in the studied data, which allowed researchers to answer questions such as the ones posed in Section 3. This results are presented in the next section.

The preliminary evaluation of CrimeVis suggests that the set of tools have achieved its purpose to support researchers on public safety. Most of the evaluated components were deemed easy to understand, with the exception of some particular issues. Regarding the interactive controls, the level of understanding was high for the configuration controls and the selection controls, 70% and 80%, respectively. Two users considered the cluster visualization in parallel coordinates plot 2D and 3D difficult to analyze in certain situations, in which there is juxtaposition of lines, even when they are replaced by Bézier curves. In adidition, the researcher needs to have some familiarity with the chart in order to easily interpret it and find an organization of the axes that more clearly reveal data attributes. One user reported difficulties to understand how to manipiulated de coordinates views without a tutorial. Two users reported difficulties to select or follow specific DPs when analyzing the parallel coordinates with clustering visualization. One user reported that in the MDS plot, is difficult to select and visualize a single DP when we have a large number of points and one reported that the MDS technique is difficult to understand and difficult to understand and to be applied in the context, since the relationship of the coordinates of the points do not match the values of the variables but the projection of MDS. Regarding data selection, two users reported that some legends and information on the data could be clearer. Despite these issues, CrimeVis was positively accepted by the users, being considered by some as a strong tool to support research.

4.2 Results

In this section we describe some discoveries in the data made visually by researchers in the domain using the CrimeVis system. These discoveries are related to the pacification program created by the State, with the so called UPPs. It is well known that the UPPs had a huge influence on the criminal activity in the recent years. Since the beginning of this program in 2008, the govern of the State created 37 UPPs in the city of Rio de Janeiro. Using CrimeVis, the specialists have found some patterns in the data. To better analyse the UPPs result, they divide the set of DPs into groups representing

the sub-regions of the State (RT-3). Figure 4.1 shows these groups of DPs. In the capital of the State, there is a group colored in red, which represents the DPs which received at least one UPP, and the group in green, representing DPs without UPPs. The other groups are: the group named Baixada Fluminense that is colored in blue, the group named Grande Niterói that is colored in yellow, and the Interior of the State colored in purple.

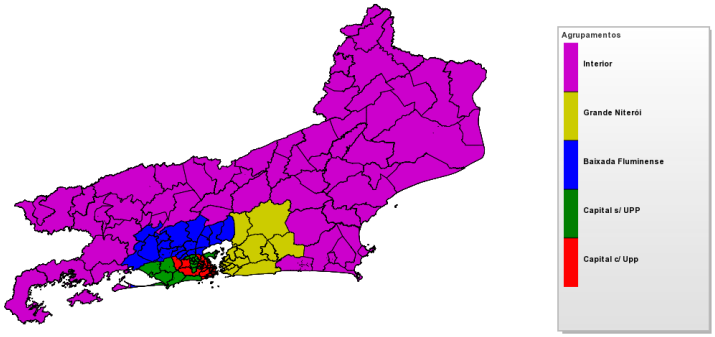
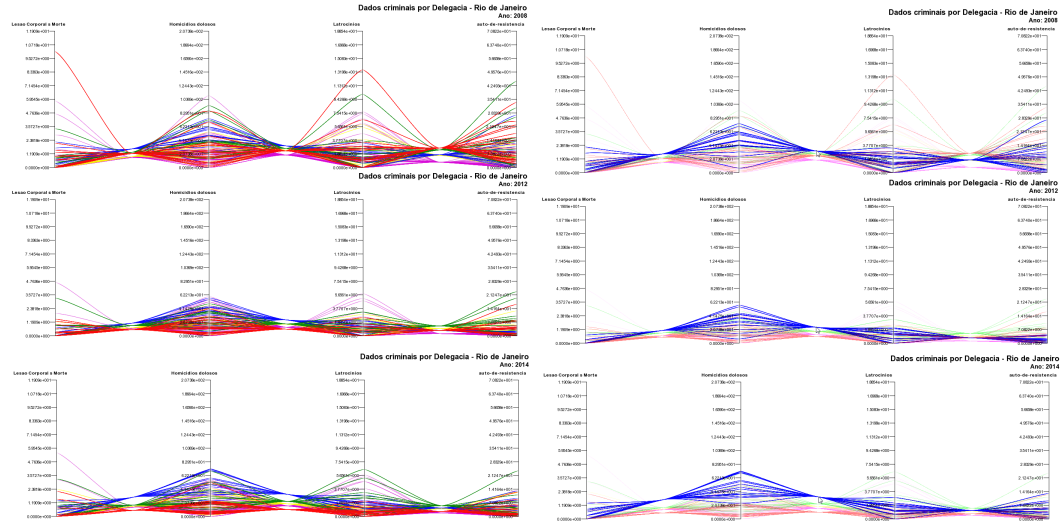


Figura 4.1 – Sub-regions of the State of Rio de Janeiro: DPs which received at least one UPP in red, DPs without UPPs in green, Baixada Fluminense in blue, Grande Niterói in yellow, and the Interior of the State in purple.

For the analysis, the researchers consider violent mortality crimes in each group. The violent mortality is composed by the body injury crimes followed by death, homicides, larceny and police killings. Figure 4.2(a) shows that the crime rates do not differ much from each other in different regions. For homicides, they noticed a predominance of Baixada Fluminense (RT-1) with the highest rates. They also noticed that, after some years, the violent mortality rates had increased and disaggregation appears in the great majority of the groups, where the the region Baixada Fluminense is the more prominent. Figure 4.2(b) shows that the Baixada Fluminense DPs were highlighted and with this highlight it is possible to notice that in 2012 the homicides taxes are distributed with less dispersion and there is a small reduction in average. Analysing how crime rates evolved over the time (RT-2) ,we notice that in 2014 the violent mortality rates are still with low dispersion, but with higher values. In general, the violent deaths increased in this region in a period of 2012 to 2014.

The first UPP was deployed in December 2008. In 2014 there were 37 in the city of Rio de Janeiro, covering over 200 communities and an estimated population of 562,691 inhabitants[5]. The expansion of the program raised several criticisms questioning the effectiveness of the UPPs in reducing criminality. In 2015, the Public Safety Institute (ISP), published a report which showed a decrease in criminality within the UPPs. However, it did not examine the areas neighboring areas that received UPPs. A program of such nature may influence criminality in the surrounding areas as well, and thus deserves a



(a) State of Rio de Janeiro (b) Baixada Fluminense

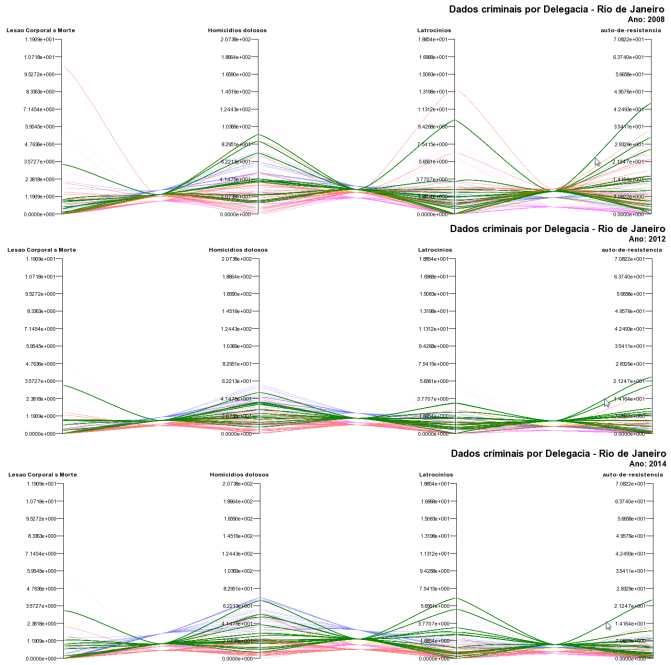
Figura 4.2 – Evolution of lethality in Baixada Fluminense and in the state of Rio de Janeiro.

broader evaluation. In Figures 4.3(a) and 4.3(b), the researchers noticed that in 2008 the regions in the capital with and the neighboring regions without UPPs had similar criminal rates. In 2012, the rates are reduced in two regions, showing that the UPPs possibly influenced the areas of the capital who don't have any UPP. Although, at the end of 2014, this two regions no longer have the same behavior. They noticed that in the regions where we don't have a UPP, we have a higher dispersion when compared with the regions with a UPP. In the regions of UPPs, intentional homicide rates splits into two groups: those with lower rates correspond mostly to the south zone's DPs and the other to the police stations in the northern and western areas of the State's capital. Stands out with the highest rate the 4th DP, which is located in the downtown area of the capital. The observed statistics show that the UPPs program possibly fails to curb the criminality. While we have been the regions with lower crime rates, there is a significant increase in the western of the capital and in the Baixada Fluminense from the year 2012. One of the practical effects of the installation of UPPs was the spread of crime (RT-4). However, despite these indications, another study done by experts is still needed to find new evidence to support this hypothesis.

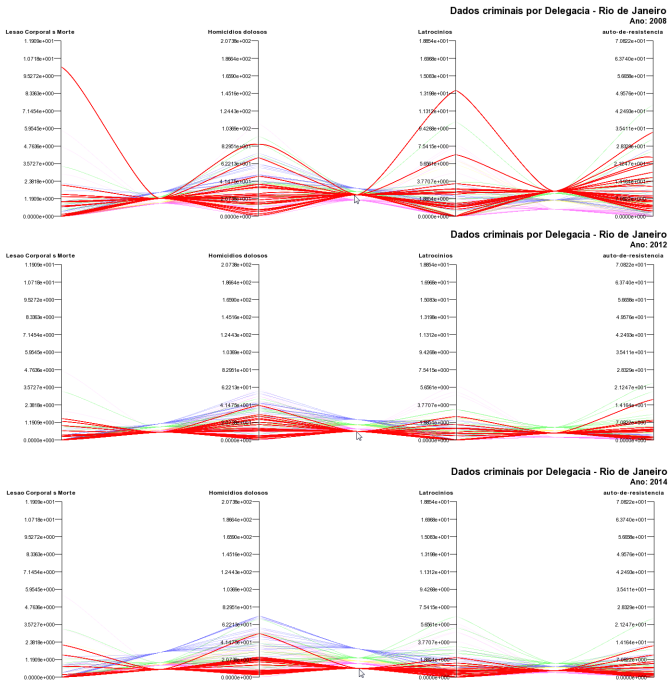
Also we found some inconsistencies in the data analyzed (RT-5). Analysing the cluster quality of the groups for each clustering algorithm used, we concluded that in some cases is difficult to identify some clusters structures. We use as measure the silhouette [32] and the Dunn index [40]. In General considering 3, 5, 7 and 10 clusters, the average silhouette width vary from 0.3 to 0.41 for all techniques and the Dunn index from 0.08 to 0.37. Figure 4.4

shows the silhouette plot for a K-Medoids instance.

These values can be related to identified outliers in the data that have a very different behavior over the time related to others DPs, such as the DP1. The 1st DP located in the central area of the city, covering part of the central region and the island of Paquetá, has high rates of intentional homicide and police murders due his small number of residents and the large floating population (the taxes are calculated by the occurrences per 100 thousand habitants). But this is not sufficient to explain why this taxes are so high in relation to other DPs considering the gravity of this kinds of crimes. A further study is needed to understand the behavior of this region.



(a) Regions without UPPs



(b) Regions with UPPs

Figura 4.3 – Lethality evolution in areas of DPs with and without UPP.

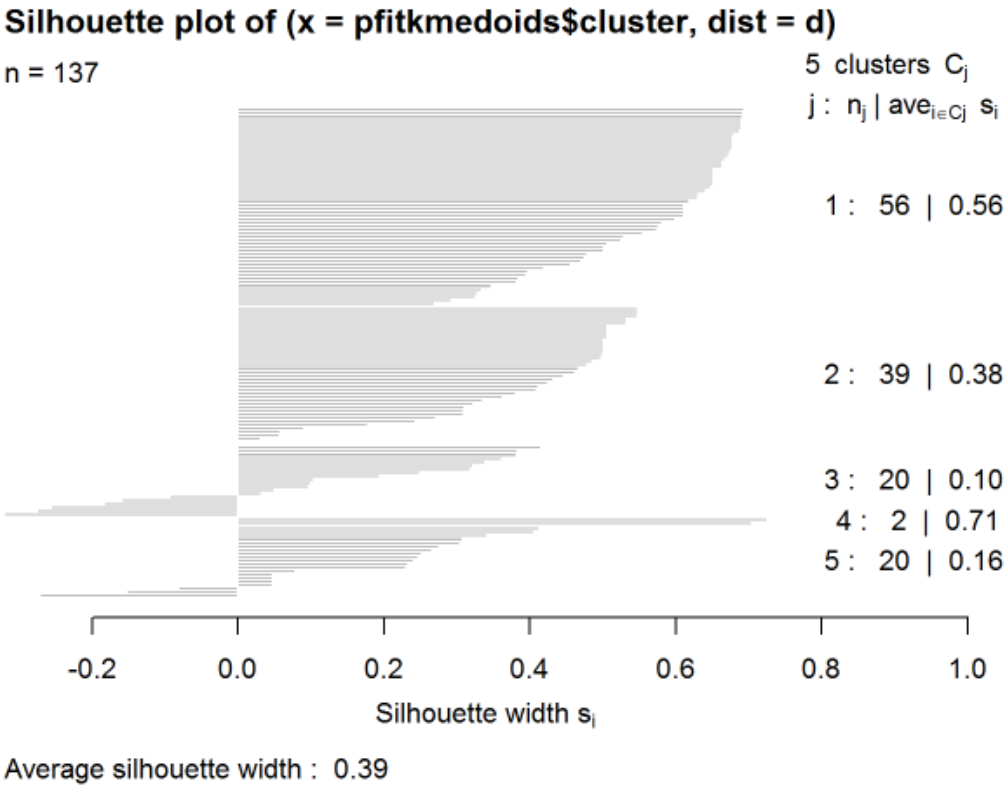
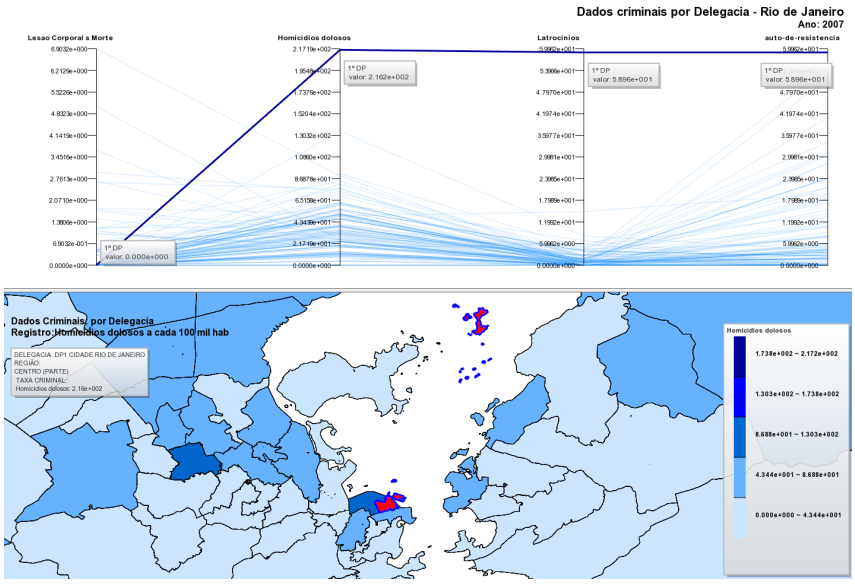
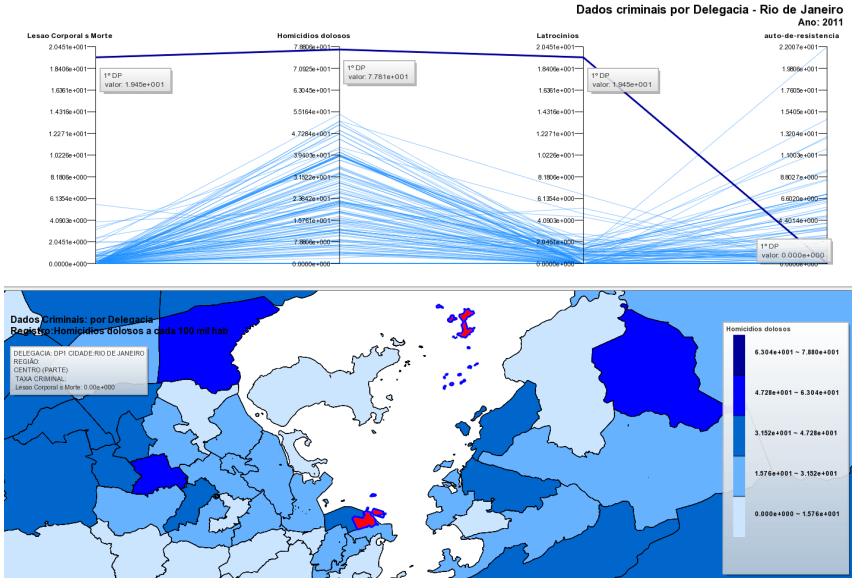


Figura 4.4 – silhouette plot for the K-Medoids Algorithm



(a) crime rates for 2007



(b) crime rates for 2011

Figura 4.5 – Crime rates of 1th DP located in the central area of the Capital

5
Conclusion

CrimeVis offers a set of efficient tools to support researchers on public safety. The overview provided by the tool allows users to easily discover patterns and analyze trends in the data being investigated. The software is still undergoing testing to be deployed and widely used by researchers in the field. Our preliminary studies showed that CrimeVis is efficient when it is necessary to analyze a data set for a specific time period. The users could easily establish relations between the data and identify trends and patterns through interactive analysis of the data. Moreover, the brushing and linking technique allows us to select and filter informations more easily, being a powerful technique to answer questions relevant to how the relation between different data attributes. With CrimeVis, one can analyze not only groups, but also individual areas using the map of DPs, in which it is possible to interpret the evolution of a certain attribute over time.Our study suggests that the set of tools have achieved its purpose to support researchers on public safety.

As for future work, new tools can be aggregated to the software to allow for deeper investigation of the data, such as including histograms and allowing users to locate and relate the evolution of a certain kind of crime in an area, split the visualization between the areas defined in the state, and navigate between tabulated data, besides building infographics.We also intend to create a module which we can use alogorithms in R language to evaluate to evaluate different types of data. Furthermore, we intend to transform the software into a web application and make it available to the general public.

Referências Bibliográficas

[1] DE BORJA, F. G.; FREITAS, C. M.. **Civisanalysis: Interactive visualization for exploring roll call data and representatives' voting behaviour**. In: GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), 2015 28TH SIBGRAPI CONFERENCE ON, p. 257–264. IEEE, 2015.

[2] POTTER, K.; WILSON, A.; BREMER, P.-T.; WILLIAMS, D.; DOUTRIAUX, C.; PASCUCCI, V. ; JOHNSON, C. R.. **Ensemble-vis: A framework for the statistical visualization of ensemble data**. In: DATA MINING WORKSHOPS, 2009. ICDMW'09. IEEE INTERNATIONAL CONFERENCE ON, p. 233–240. IEEE, 2009.

[3] ASSUNÇÃO, R. M.; NEVES, M. C.; CÂMARA, G. ; DA COSTA FREITAS, C.. **Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees**. International Journal of Geographical Information Science, 20(7):797–811, 2006.

[4] MONTEIRO, J.; ROCHA, R.. **Tráfico de drogas e desempenho escolar no rio de janeiro**. IBRE - Notas técnicas, 2013.

[5] IBGE. **Censo demográfico 2010: Características da população e dos domicílios: resultados do universo**. IBGE: Indicadores Sociais, 2011.

[6] ISP-RJ. **Notas metodólogicas do instituto de segurança pública do rio de janeiro**. 2013.

[7] DE MELO, B. M.; GUIMARAES, J. L.; DE CASTRO, A. S.; SANTOS, C. A.; NASCIMENTO, D. M. ; DEL PINO LINO, A.. **Criminal data mining: A case studyin criminal observatory tapajós**. In: INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 2015 10TH IBERIAN CONFERENCE ON, p. 1–6. IEEE, 2015.

[8] HARTIGAN, J. A.; WONG, M. A.. **Algorithm as 136: A k-means clustering algorithm**. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100–108, 1979.

[9] PARK, H.-S.; JUN, C.-H.. **A simple and fast algorithm for k-medoids clustering**. Expert Systems with Applications, 36(2):3336–3341, 2009.

[10] KAUFMAN, L.; ROUSSEEUW, P. J.. **Finding groups in data: an introduction to cluster analysis**, volumen 344. John Wiley & Sons, 2009.

[11] LEE, J. H.; MCDONNELL, K. T.; ZELENYUK, A.; IMRE, D. ; MUELLER, K.. **A structure-based distance metric for high-dimensional space exploration with multidimensional scaling**. Visualization and Computer Graphics, IEEE Transactions on, 20(3):351–364, 2014.

[12] BACH, B.; DRAGICEVIC, P.; ARCHAMBAULT, D.; HURTER, C. ; CARPENDALE, S.. **A review of temporal data visualizations based on space-time cube operations**. In: EUROGRAPHICS CONFERENCE ON VISUALIZATION, 2014.

[13] PALMAS, G.; BACHYNSKYI, M.; OULASVIRTA, A.; SEIDEL, H. P. ; WEINKAUF, T.. **An edge-bundling layout for interactive parallel coordinates**. In: VISUALIZATION SYMPOSIUM (PACIFICVIS), 2014 IEEE PACIFIC, p. 57–64. IEEE, 2014.

[14] ZHOU, H.; YUAN, X.; QU, H.; CUI, W. ; CHEN, B.. **Visual clustering in parallel coordinates**. In: COMPUTER GRAPHICS FORUM, volumen 27, p. 1047–1054. Wiley Online Library, 2008.

[15] HEINRICH, J.; WEISKOPF, D.. **State of the art of parallel coordinates**. STAR Proceedings of Eurographics, 2013:95–116, 2013.

[16] JOHANSSON, J.; LJUNG, P.; JERN, M. ; COOPER, M.. **Revealing structure within clustered parallel coordinates displays**. In: INFORMATION VISUALIZATION, 2005. INFOVIS 2005. IEEE SYMPOSIUM ON, p. 125–132. IEEE, 2005.

[17] HEER, J.; SHNEIDERMAN, B.. **Interactive dynamics for visual analysis**. Proceedings of the 26th ACM Conference on Hypertext and Social Media, 10(2):30, 2012.

[18] BECKER, R. A.; CLEVELAND, W. S.. **Brushing scatterplots**. Technometrics, 29(2):127–142, 1987.

[19] BUJA, A.; MCDONALD, J. A.; MICHALAK, J. ; STUETZLE, W.. **Interactive data visualization using focusing and linking**. In: VISUALIZATION, 1991. VISUALIZATION'91, PROCEEDINGS., IEEE CONFERENCE ON, p. 156–163. IEEE, 1991.

[20] DEMIR, I.; DICK, C. ; WESTERMANN, R.. **Multi-charts for comparative 3d ensemble visualization**. Visualization and Computer Graphics, IEEE Transactions on, 20(12):2694–2703, 2014.

[21] CHEN, H.; ZHANG, S.; CHEN, W.; MEI, H.; ZHANG, J.; MERCER, A.; LIANG, R. ; QU, H.. **Uncertainty-aware multidimensional ensemble data visualization and exploration**. Visualization and Computer Graphics, IEEE Transactions on, 21(9):1072–1086, 2015.

[22] POTTER, K.; WILSON, A.; BREMER, P.-T.; WILLIAMS, D.; DOUTRIAUX, C.; PASCUCCI, V. ; JOHNSON, C. R.. **Ensemble-vis: A framework for the statistical visualization of ensemble data**. In: DATA MINING WORKSHOPS, 2009. ICDMW'09. IEEE INTERNATIONAL CONFERENCE ON, p. 233–240. IEEE, 2009.

[23] LAW, J.; QUICK, M. ; CHAN, P.. **Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level**. Journal of quantitative criminology, 30(1):57–78, 2014.

[24] CHAINEY, S.; TOMPSON, L. ; UHLIG, S.. **The utility of hotspot mapping for predicting spatial patterns of crime**. Security Journal, 21(1):4–28, 2008.

[25] ARIETTA, S. M.; EFROS, A. A.; RAMAMOORTHI, R. ; AGRAWALA, M.. **City forensics: Using visual elements to predict non-visual city attributes**. Visualization and Computer Graphics, IEEE Transactions on, 20(12):2624–2633, 2014.

[26] BASAK, D.; PAL, S. ; PATRANABIS, D. C.. **Support vector regression**. Neural Information Processing-Letters and Reviews, 11(10):203–224, 2007.

[27] HAN, J.; KAMBER, M. ; PEI, J.. **Data mining: concepts and techniques**. Elsevier, 2011.

[28] ROUGEUX, N.; PRY, C.; EDER, D.; BAKER, PAUL, B. J. ; VELEZ, J. P.. **Crime in chicago: An interactive analysis of crime in chicago's 50 wards**.

[29] KINNAIRD, P.; ROMERO, M. ; ABOWD, G.. **Connect 2 congress: visual analytics for civic oversight**. In: CHI'10 EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS, p. 2853–2862. ACM, 2010.

[30] YVONNE, R.; SHARP, H. ; PREECE, J.. **Interaction design: Beyond human-computer interaction**, 2011.

[31] KRUSKAL, J. B.; WISH, M.. **Multidimensional scaling**, volumen 11. Sage, 1978.

[32] ROUSSEEUW, P. J.. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. Journal of computational and applied mathematics, 20:53–65, 1987.

[33] MAULIK, U.; BANDYOPADHYAY, S.. **Performance evaluation of some clustering algorithms and validity indices**. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(12):1650–1654, 2002.

[34] JUNGNICKEL, D.; SCHADE, T.. **Graphs, networks and algorithms**. Springer, 2005.

[35] WICKELMAIER, F.. **An introduction to mds**. Sound Quality Research Unit, Aalborg University, Denmark, p. 46, 2003.

[36] HEINRICH, J.; LUO, Y.; KIRKPATRICK, A. E.; ZHANG, H. ; WEISKOPF, D.. **Evaluation of a bundling technique for parallel coordinates**. arXiv preprint arXiv:1109.6073, 2011.

[37] RÜBEL, O.; WEBER, G. H.; KERÄNEN, S. V.; FOWLKES, C. C.; HENDRIKS, C. L. L.; SIMIRENKO, L.; SHAH, N.; EISEN, M. B.; BIGGIN, M. D.; HAGEN, H. ; OTHERS. **Pointcloudxplore: Visual analysis of 3d gene expression data using physical views and parallel coordinates.** In: EUROVIS, p. 203–210, 2006.

[38] LAW, J.; QUICK, M. ; CHAN, P. W.. **Analyzing hotspots of crime using a bayesian spatiotemporal modeling approach: a case study of violent crime in the greater toronto area**. Geographical Analysis, 47(1):1–19, 2015.

[39] LAWSON, C. L.; HANSON, R. J.. **Solving least squares problems**, volumen 161. SIAM, 1974.

[40] DUNN, J. C.. **Well separated clusters and optimal fuzzy partitions**. Journal of cybernetics, 4(1):95–104, 1974.