A Strategy for Visual Structural Data Analysis of Labor Accident Data

Mateus Pinto Rodrigues

Federal University of Uberlândia
Faculty of Computing
Post-Graduation Program in Computer Science

Uberlândia
2019

Mateus Pinto Rodrigues

A Strategy for Visual Structural Data Analysis of Labor Accident Data

Master Thesis presented to the Faculty of Computing Post-Graduation Program of the Federal University of Uberlândia as part of the requirements for obtaining the title of Master of Science in Computer Science.

Concentration Area: Computer Science

Advisor: José Gustavo de Souza Paiva

Uberlândia
2019

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado Acadêmico, 20/2019, PPGCO				
Data:	27 de novembro de 2019	Hora de início:	**9h03min**	Hora de encerramento:	**11h00min**
Matrícula do Discente:	11722CCP008				
Nome do Discente:	Mateus Pinto Rodrigues				
Título do Trabalho:	A Strategy for Visual Structural Data Analysis of Labor Accident Data				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se na sala 1B132, Bloco 1B, Campus Santa Mônica, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Murillo Guimarães Carneiro - FACOM/UFU, Bianchi Serique Meiguins - ICEN/UFPA e José Gustavo de Souza Paiva - FACOM/UFU, orientador do candidato.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. José Gustavo de Souza Paiva, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **José Gustavo de Souza Paiva**, **Professor(a) do Magistério Superior**, em 28/11/2019, às 09:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bianchi Serique Meiguins**, **Usuário Externo**, em 28/11/2019, às 14:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 09/12/2019, às 09:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1714561** e o código CRC **08D02FA1**.

Referência: Processo nº 23117.101727/2019-10 SEI nº 1714561

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

R696 Rodrigues, Mateus Pinto, 1993-
2019 A strategy for visual structural data analysis of labor accident data [recurso eletrônico] / Mateus Pinto Rodrigues. - 2019.

Orientador: José Gustavo de Souza Paiva.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2019.2578
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Paiva, José Gustavo de Souza,1979-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

This work is dedicated to all those who, sometimes at great personal costs, fought for the freedom of all beings.

Acknowlegedment

Agradeço a Deus pela vida concedida todos os dias e pelas oportunidades dadas a mim.

Agradeço a minha família pelo apoio prestado em todos os momentos, principalmente os mais difíceis. Agradeço especialmente a minha mãe Maristela, minha tia Ana Cláudia, e minha avó Maria de Nazaré, sem as quais eu certamente não teria chegado até aqui. Também agradeço com muito carinho a minha esposa, Luciana, por encher minha vida de alegria, me motivar e acreditar em mim, e a minha gatinha Shizuka, que não nos deixa passar um dia sem sorrir.

Agradeço a todos os meus amigos, de longa data e os de vida acadêmica, que compartilharam momentos importantes comigo.

Agradeço ao meu orientador, José Gustavo Paiva, pelo seu ensino, dedicação e paciência.

Agradeço a CAPES pela bolsa concedida durante a duração dessa pesquisa.

"Should any political party attempt to abolish social security, unemployment insurance, and eliminate labor laws and farm programs, you would not hear of that party again in our political history."

(Dwight D. Eisenhower)

Resumo

Acidentes de trabalho são um problema social sério, que causa danos para empregados, empregadores, e governos, consumindo uma parcela significativa do PIB mundial. No Brasil, o Ministério Público do Trabalho é o órgão institucional responsável pela defesa dos direitos dos trabalhadores, e entre as suas funções estão a supervisão e controle da saúde e segurança no trabalho. Eles coletam dados sobre acidentes de trabalho no Brasil e disponibilizam esses dados publicamente. Esse processo gera um grande volume de dados contendo informações estratégicas importantes, geralmente difíceis de extrair por meio de análises manuais. Visualização da informação é uma área de pesquisa que estuda a criação de representações visuais para dados, visando ajudar pessoas a executarem tarefas mais eficientemente. Propomos uma estratégia que emprega técnicas de visualização da informação para analisar dados de acidentes de trabalho, sem ser restrita a este cenário. Desenvolvemos um sistema que implementa essa estratégia, e que compreende duas visualizações complementares, i) projeção multidimensional + mapa político, e ii) treemap + conjuntos paralelos. Realizamos diversas análises exploratórias aproveitando as capacidades complementares das visualizações em prover análise simultânea de diferentes aspectos dos dados. Identificamos perfis associados a áreas geográficas pequenas/grandes, similaridades entre localidades geograficamente distantes, padrões de ocorrência relacionados ao tamanho e desenvolvimento econômico das cidades, a distribuição de frequência dos tipos de acidentes de trabalho no Brasil, e os caracterizamos em termos de tipo de ocupação, diferenças de gênero, agente causador, entre outros aspectos. Cremos que a estratégia proposta facilita e melhora a análise de dados de acidentes de trabalho, provendo meios eficazes e eficientes para ajudar governos a avaliar políticas atuais e fomentar a criação de novas políticas para reduzir acidentes de trabalho e garantir segurança para empregados, e também encorajar a transparência em governos e a participação popular.

Palavras-chave: Acidente de Trabalho. Dados Governamentais. Visualização da Informação.

Abstract

Labor accidents are a serious social problem that results in damages to employees, employers, and governments, also consuming a significant portion of the World's GDP. In Brazil, The Brazilian Federal Labor Prosecution Office is the institutional service responsible for the defense of worker rights, and among its functions is the supervision and control of labor health and safety. They collect data on labor accidents in Brazilian territory and provide an anonymized version of this data publicly. This process generates a large volume of data containing important strategical information, which is often not straightforward to be extracted with manual analysis. Information visualization is a research area that studies the creation of visual representations for abstract structured or non-structured data, aiming to help people execute tasks more effectively. We propose a computational strategy employing a combination of Information Visualization techniques to perform a visual analysis of labor accident data, while not being restricted to this scenario. We developed a system that implements our strategy, and is comprised of two complementary visualizations, i) a multidimensional projection layout + a political map, and ii) a treemap layout + a parallel sets layout. We performed several exploratory analysis, in order to exploit the visualizations' complementary capacities in providing simultaneous analysis of different data aspects. We obtained interesting results, identifying profiles associated with small/large geographical areas, similarities among geographically distant localities, occurrence patterns related to cities' size and economic development, the frequency distribution of labor accident types in Brazil, and characterized labor accidents in terms of occupation type, gender differences, causer agent, among other aspects. We believe that the proposed strategy facilitates and enhances the analysis of labor accident data, providing effective and efficient means to help governments to evaluate current public policies and foment the creation of new ones to reduce labor accidents and grant safety to employees, and also encourage transparency in governments and citizen participation.

Keywords: Labor Accident. Governmental Data. Information Visualization.

List of Figures

List of Tables

Acronyms list

BFLPO Brazilian Federal Labor Prosecution Office

INFOVIS Information Visualization

LA Labor Accident

LAC Labor Accident Communication

ODSST Observatório Digital de Saúde e Segurança do Trabalho

Contents

I hereby certify that I have obtained all legal permissions from the owner(s) of each third-party copyrighted matter included in my thesis, and that their permissions allow availability such as being deposited in public digital libraries.

student name and signature

Chapter 1

Introduction

Labor Accidents(LAs) represent serious problems for the economy of a locality, because they result in physical and psychological disorders for the employee, loss of manpower for the employer, and generate significant expenses with compensation benefits and health related costs for the government. Every year more than 2.78 million deaths occur due to LAs directly or to illnesses related to them, in addition to other 374 million non-fatal accidents. The estimated expenses associated with these accidents is approximately 3.94% of the global yearly GDP(ILO, 2019). In Brazil, in the period of 2012 to 2018, 4.503.631 LAs were registered, with 16.455 deaths. The expenses with accidents registered in this period were of R$29.145.635.014, and when counting the expenses for accidents occurring before but still being paid the sum goes up to R$79.000.041.558(Smartlab de Trabalho Decente MPT - OIT, 2017).

The Brazilian Federal Labor Prosecution Office (BFLPO) is the institutional service responsible for the defense of worker rights, and among its functions is the supervision and control of labor health and safety. The BFLPO collects data on labor accidents in Brazilian territory using **Labor Accident Communications (LACs)**, which is a document issued to register labor, route accidents and/or occupational diseases(INSS, 2018).An anonymized version of this data is publicly available in(Smartlab de Trabalho Decente MPT - OIT, 2017). In a country of continental proportions, the volume of data generated by LA occurrences is huge, which makes manual analysis of such data tedious, difficult, and error prone. On the other hand, automatic analysis of LA data, without human intervention, employing machine learning strategies or purely statistical calculations, might limit the extraction of information from the data repository, because they do not allow the refinement of results in an intuitive way to address the particular needs of the analyst, nor do they easily accommodate interactive exploration of the data. Furthermore, those automatic strategies make difficult the understanding of the relationship between the produced results and the data structure, thus the conclusions are reached in spite of humans, which might not be the best case when public interests are at stake.

Information Visualization (INFOVIS) is a research area that studies the creation of

visual representations for datasets, in order to enhance human cognitive capacities and aid people to fulfill data exploration tasks more easily, in a form that is both efficient and effective(MUNZNER, 2014). The BFLPO provides a website named ***Observatório Digital de Saúde e Segurança do Trabalho***(ODSST)[1], in which LAs data can be analyzed using simple visual approaches. Such layouts are limited in terms of ability to highlight patterns and data aspects, and offer few or no interaction tools, which reduces the capacity of exploring the repository, and diminishes the potential of extracting strategical information. Using more sophisticated INFOVIS techniques may help BFLPO specialists in the analysis and exploration of LA data. The idea is to use a combination of several widely used techniques to create visualizations that communicate the underlying structure of those data, that is, the inherent relationship between different data instances without regarding temporal aspects, exploiting its heterogeneity, associated hierarchical organization, among other aspects.

In this context, INFOVIS techniques might be appropriate for this analysis, since they represent a potential tool for aiding in the identification of patterns and trends in the data that represent strategic information for BFLPO specialists. These visualizations may allow the identification of profiles for accidents, localities, occupations, and make correlations among several indices measured in diverse regions, highlighting similarities and differences among distinct localities, even from different hierarchy levels. An important aspect of INFOVIS is the extensive use of interactions and visual metaphors when displaying data, allowing a natural and effective exploration. While interacting with a visual representation of a dataset, new facets of the data are revealed, and this flexibility improves knowledge discovery.

1.1 Objectives

The analysis of LAs data is important to the BFLPO, as it aids them in its primary task of providing labor safety. However, important as it may be, there is still a lack of proper analysis tools addressing their needs. Thus, the objective of this project is to develop a exploratory visual analysis strategy for LA analysis and exploration, employing INFOVIS techniques, in order to highlight different aspects of the underlying structure of the data.

The specific objectives are:

- ❏ Create *layouts* that communicate the structure of the repository data, highlighting patterns, trends and correlations among measured indices;
- ❏ Create interaction tools and visual metaphors to improve data exploration, as well as to maximize information extraction;

[1] https://observatoriosst.mpt.mp.br

- ❏ Develop a system implementing the proposed strategy, for exploratory analysis of the BFLPO repository and also to evaluate the proposed strategy.

1.2 Hypothesis

A visual analysis strategy employing information visualization techniques highlights the underlying structural characteristics of the BFLPO dataset, and is effective for the analysis and exploration of governmental data, improving decision making by specialists.

1.3 Contributions

This project has the following contributions:

- ❏ A system coordinating different layouts that employ established interactive visualization techniques in the analysis and exploration of LAs data;
- ❏ The combination of several widely used INFOVIS techniques in a coherent and effective way to aid the analysis of specific data types.
- ❏ A visual analysis strategy that broadens the comprehension of analysis and exploration of LAs data, particularly in Brazil, providing a better understanding of LA profiles, by means of exploring the particular LA dynamics in the country, helping to choose appropriate policies to prevent them.

1.4 Thesis Organization

The remainder of this thesis is organized as follows: Chapter 2 presents background knowledge required as well as related work. Chapter 3 presents the proposed strategy and describes the data used . Chapter 4 describes the developed system and its implementation details. Chapter 5 discusses the results obtained. Finally, Chapter 6 concludes with an overview of what was achieved, the limitations of our approach and future work.

Chapter 2

Fundamentals

In this chapter we present some Information Visualization (INFOVIS) fundamentals and a discussion of related works. We first introduce basic concepts that are important to comprehend the work developed in this project, including concepts related to governmental data. We also present INFOVIS concepts, focusing in which techniques were employed in this project. Finally, we discuss several strategies that employ INFOVIS techniques in the analysis of governmental data.

2.1 Basic concepts

This section presents some basic concepts used in this research and in the related work.

Attribute: Also called variable, feature or characteristic, is a measured property of a data instance.

Instance: A representation of a data entry in the original space, described by all its attribute values.

Dataset: Also called data collection, is a set of several data instances.

Dimension: Represents an attribute in a dataset when its instances are to be represented as a vector in a specific space.

Layout: A single basic visual representation which can be combined with other layouts and interactions.

Visualization: A more elaborate visual representation, composed of one or more layouts and interactions.

Interaction: Actions performed by the user of a visualization, which causes an alteration in the appearance of the visualization.The modification of the visualization itself may also be called interaction in specific situations.

Overview: Visual presentation of the general context of a dataset.

Original Space: Multidimensional space composed of all the instances with all its attributes.

Visualization Space: Space with a specific number of dimensions in which a visualization is presented.

Point: A representation of an instance in the visualization space, as a result of a multidimensional projection technique application.

2.2 Governmental Data

This section presents some works that address governmental data analysis, highlighting its importance to several sectors of society. Governmental data is data of any type or format collected or used by a government from all sectors of society. Giving its nature, this data is usually voluminous and heterogeneous, which represents important challenges in dealing with it. Governmental data is not always publicly available, but usually is, which encourages transparency in public institutions and promotes innovative civic-centric services(OECD, 2019).

According to (RADL et al., 2013) the analysis of governmental data present enormous potential in two ways. First, it allows the validation of previous knowledge and the gaining of insights and new knowledge about specific fields. Second, it can also help the administration of a locality by enabling the creation of new semantic technologies for the population and government. The data provided by the government in several forms, when correctly comprehended, can aid in the creation of public policies, or in the identification of areas lacking attention, thus optimizing government expenditure.

The importance of open governmental data is also reported in(GRAVES; HENDLER, 2013), arguing that simply making data open is not enough to keep the population well informed. Thus, it is necessary to create mechanisms that enable the execution of necessary operations for this data to make sense.

Governmental data is also useful to promote population participation, as shown in (DíAZ; AEDO; HERRANZ, 2014). The authors discuss how population participation is important for the development of policies for crisis management, and that for this participation to be effective the population also needs to have more confidence in the data provided by the government. It is also shown in (YING; XIALING; WEI, 2017) how a government can benefit from analyzing large volumes of data to optimize expenditure and improve the interaction between government and population.

Other data sources can also be used with governmental purposes, as in (ALOWIBDI; GHANI; MOKBEL, 2014). The authors present an application that suggests holiday locations based on twitter data via a flow map. Maps are interesting to communicate

data, since they present a familiar layout, allowing fast comprehension to the general public. The target audience of this application are citizens that want to choose a good place to spend a holiday, and the government, that can focus its tourism policies in specific places depending on the time of they year.

In Brazil, governmental data is provided by several public agencies. The Brazilian Federal Labor Prosecution Office(BFLPO) is the institution responsible for supervising work conditions, as well as intervening on labor issues to guarantee worker rights. The BFLPO counts with an ample dataset of labor accident records, composed of notifications made, preferably by the employer, via LACs. These data are of strategic importance to Brazil, since its correct comprehension can help directing public policies to improve work relationships. The BFLPO also makes these data freely available.

Governmental data can be an important source of information both for government and population in general, because they allow for greater transparency in the services provided by the government, and the monitoring of its actions by the population, thus increasing social welfare. However, the lack of methods capable of communicating this data in a comprehensive and efficient manner is noticeable, as well as the lack of tools for analyzing them, as discussed in Section 2.6.

2.3 Information Visualization

Information visualization studies the creation of visual representations for abstract structured or non-structured data. The representations are designed to help people execute tasks more effectively, with the use of the space for the visual encoding chosen by the designer, and it has as one of its primary focus the determination of which techniques are appropriate to combine a dataset with a task(MUNZNER, 2014). The tasks which concern INFOVIS are the ones that the human being is crucial, such as the ones in which the problem is not well defined, or it is not know for sure what is being searched. For these types of tasks INFOVIS harnesses human being's innate capacity for visual pattern recognition.

The vision sense has a prominent role in INFOVIS. The reason to use vision is that it is the most accurate sense in a human, capable of transmitting great quantities of data at the same time to the brain, where visual stimuli are perceived in parallel. This characteristic parallelism of vision can be contrasted to hearing, that only perceives information in a sequential manner and is much less sensitive to nuances outside the main context. As for other senses, there are limitations of technical nature, because there are no tools yet, nor even sufficient research to begin exploring their capacities(MUNZNER, 2014).

Figure 1 depicts the classic INFOVIS workflow. First data are obtained, which can be structured or not. Those data are organized and transformed to treat missing, spurious values, or to use other desired encoding, and then may be filtered, by the selection of

subsets of the data. In the mapping stage, each data subset is associated to graphic primitives (lines, circles, etc) or more complex visual metaphors, and to graphic attributes (color, size, etc). This mapping is rendered into an image, which is the main visual abstraction of the data, and set of interactions is provided to manipulate it. The last stage in the process is the visualization, where the user interacts with the rendered image by means of controls in a graphical interface, to explore the dataset(LIU et al., 2014). INFOVIS techniques are many and diverse and a good overview of them can be found in (HEER; BOSTOCK; OGIEVETSKY, 2010).

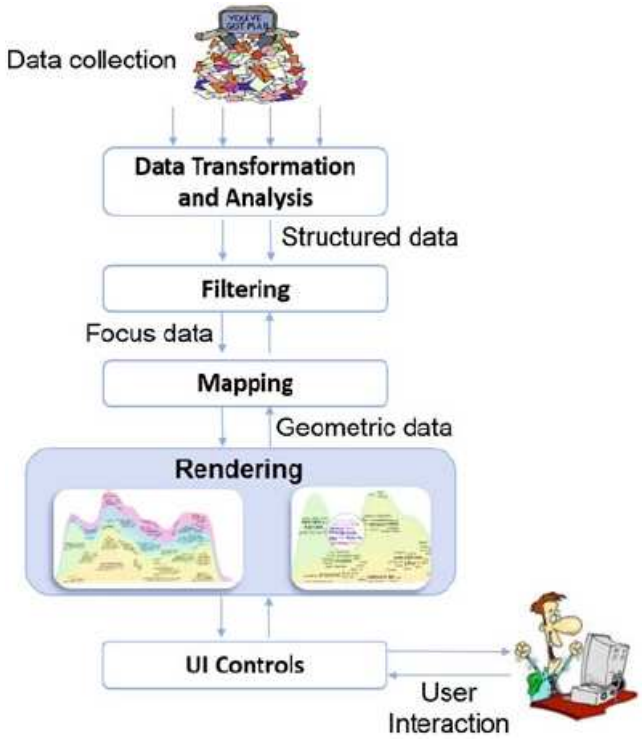


Figure 1 – Classic INFOVIS Workflow

Source: (LIU et al., 2014).

2.4 Multidimensional Projections

Multidimensional projection refers to the act of representing a n-dimensional dataset in a p-dimensional space, where $p << n$ and usually $p = 2$ for visualization purposes, while also trying to preserve the characteristics present in the many attributes in a reduced set of synthetic attributes. To reach this goal dimensionality reduction techniques are applied to the dataset. The results of these techniques is commonly visualized using a scatterplot layout. Assuming that the original relationship among instances are preserved in this layout, it is possible to visualize the general structure of the dataset, to perceive similarities between seemingly distinct instances, as well as to detect outliers and to

select subsets of instances for other specific analysis. Several multidimensional projection techniques can be found in the literature. We present here commonly used techniques.

Principal Component Analysis (PCA)(JEONG et al., 2009) is a technique that projects the instances of a dataset in a new system of coordinates based on the eigenvalues and eigenvectors of a covariance matrix of the data that minimizes the redundancy and maximizes the variance between these data instances. The eigenvectors found that have the greater eigenvalues are the ones that describe the most significant relationship among the data instances, which are called **principal components**. For visualization purposes, the first two or three principal components are used to determine the coordinates of each instance in the visualization space.

Multidimensional Scaling (MDS)(COX; COX, 2000) is a family of several linear and nonlinear techniques. The goal is to associate instances to specific positions in a space, in such a way that the Euclidean distance between them in this space best reflects their proximities, which can be similarities or dissimilarities, observed among them in the original space. To measure the correspondence between the obtained distances and the original relationship a stress function is commonly used, the smaller the obtained value the better is the correspondence.

t-SNE(MAATEN; HINTON, 2008) is a nonlinear method that tries to minimize the distance between similar data instances. A probability distribution is constructed over a set of instances, so that pairs of similar objects are chosen with high probability and dissimilar pairs are chosen with significant low probability, almost infinitesimal. Similarly, a distribution is constructed for a lower dimensional space, and the final positions are obtained by moving the points in this lower dimensional space, in a way to minimize the Kullback-Leibler divergence between the two distributions.

Least Square Projection (LSP)(PAULOVICH et al., 2008) is a method that aims to reduce the dimensionality of a set of instances, while also preserving the original neighborhood relationship among the instances. This method comprises two main steps, the first one consists in applying the MDS technique over a subset of those instances, called **control points**, to project them in the desired dimensional space. The second step consists in building a linear system using the neighborhood relationships of the original set and the Cartesian coordinates of the projected control points. The solutions to this linear system are the coordinates for the points in the new visualization space.

Local Affine Multidimensional Projection (LAMP)(JOIA et al., 2011) also employs the concept of control points, and using neighborhood information from those points orthogonal affine mappings are built, one for each instance from the dataset. As the LAMP projection depends on the control points configuration, the projection might be improved interactively by the user by manipulating these control points, and the mapping is thus able to reflect user specific perspectives of the data.

A more detailed discussion of these and other techniques can be found in(NONATO;

AUPETIT, 2018).

2.5 Strategies for Structural Data Visualization

Here we discuss research works that present ideas and visualizations that were influential to this research. For the ODSST dataset we want to highlight behavior profiles, correlations, context, among others. The visualizations discussed here are appropriate to highlight patterns which allow users to comprehend these phenomena in the structure of the datasets.

Parallel coordinates(INSELBERG, 1985) is a layout used for multivariate data, in which each variable is represented by an axis, which are drawn parallel to each other and a data instance is represented by a line that intercepts each axis in the point corresponding to the value of each variable for that instance. The axes do not impose or imply a natural order of the data, therefore they can be reordered to obtain different perspectives. Figure 2 shows a parallel coordinates layout of the Iris dataset(FISHER, 1936), in which each color represents a different species of iris and the axes represent the length and width of its petals and sepals. It can be observed that the species *setosa* presents an interesting contrast to the other two species, having, in general, sepals with bigger width and the other values smaller, while the other two species present similar values, with *virginica* species' values being proportionally larger than the *versicolor*'s values in all measures.

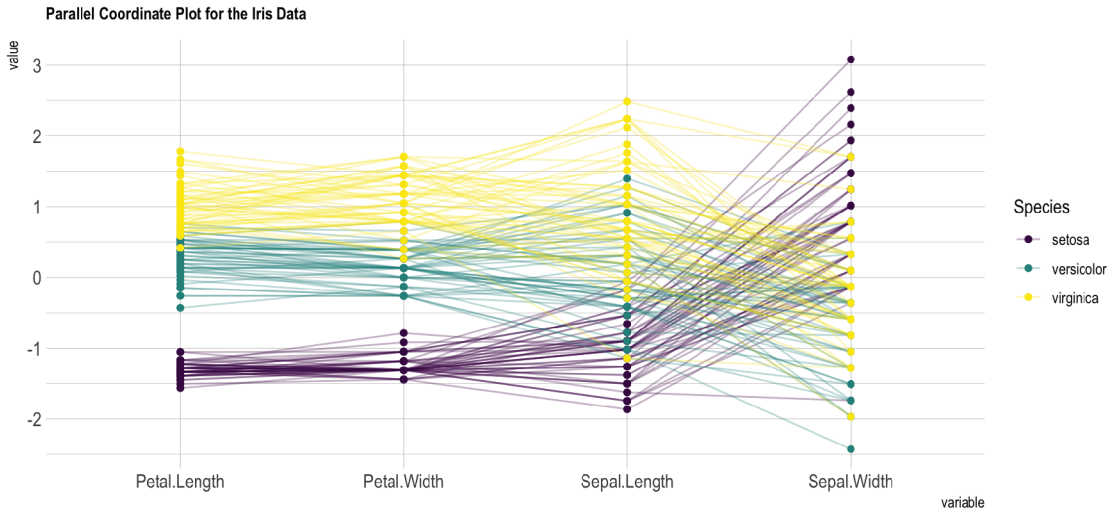


Figure 2 – Iris dataset presented in a parallel coordinates layout.

Source: https://www.data-to-viz.com/graph/parallel.html

Some research work propose improvements on the original parallel coordinates layout. Considering that one of the problems in a parallel coordinates layout is the difficult in

perceiving the volume of data that intercepts a given value in an axis, the modification proposed in(TUOR; EVéQUOZ; LALANNE, 2016) is interesting. The layout, called **Parallel bubbles**, uses circles of variable radius in the axes ticks, the bubbles, and the radius of a circle is proportional to the volume of lines intersecting each point, as can be seen in Figure 3, in which continuous values in the left axis are mapped to categorical values in the right axis. In a traditional parallel coordinates layout, a large volume of data can cause overlapping lines, and using bubbles make it easier to detect and compare the volume of data that crosses an axis in a given point.

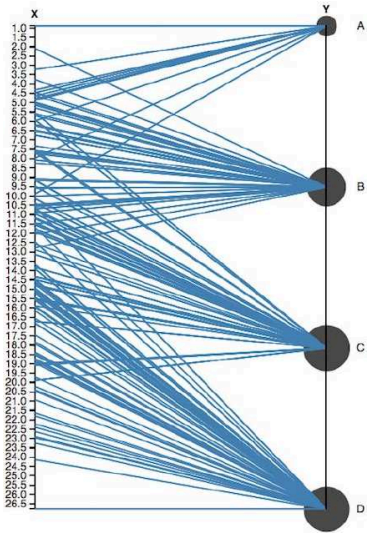


Figure 3 – Parallel Bubbles relating a continuous variable(left axis) to a categorical variable(right axis).

Source: (TUOR; EVéQUOZ; LALANNE, 2016)

A visualization called **Pivotviz**[1,2] is proposed in (NIELSEN; GRøNBæK, 2015). It combines a parallel coordinates layout with summarizations in a pivot table representing the result a specific selection. This work employs a modification in the parallel coordinates, which is the use of a single line of variable width to combine identical transactions. This approach is useful for revealing clusters, but subtle groupings might be masked. Figure 4 shows an example of this visualization, in which the value "Renew" is selected, which causes a great number of lines to be highlighted. The analysis of correlations between axes might be impaired, since it is not possible to compare the slope of a set of lines. This may occur because variable width lines might overshadow other lines or be almost imperceptible. Depicting summarizations of the data using a visualization based on a parallel coordinates strategy may be useful to reveal important information for massive datasets, such as the one provided by the ODSST.

The combination of proportional and nonproportional layouts is explored in (WITTENBURG; TURCHI, 2016). A proportional layout represents attributes as areas pro-

[1] http://odaa.datavis.dk/thebooksof.html
[2] http://kk.datavis.dk/thebooksof.html

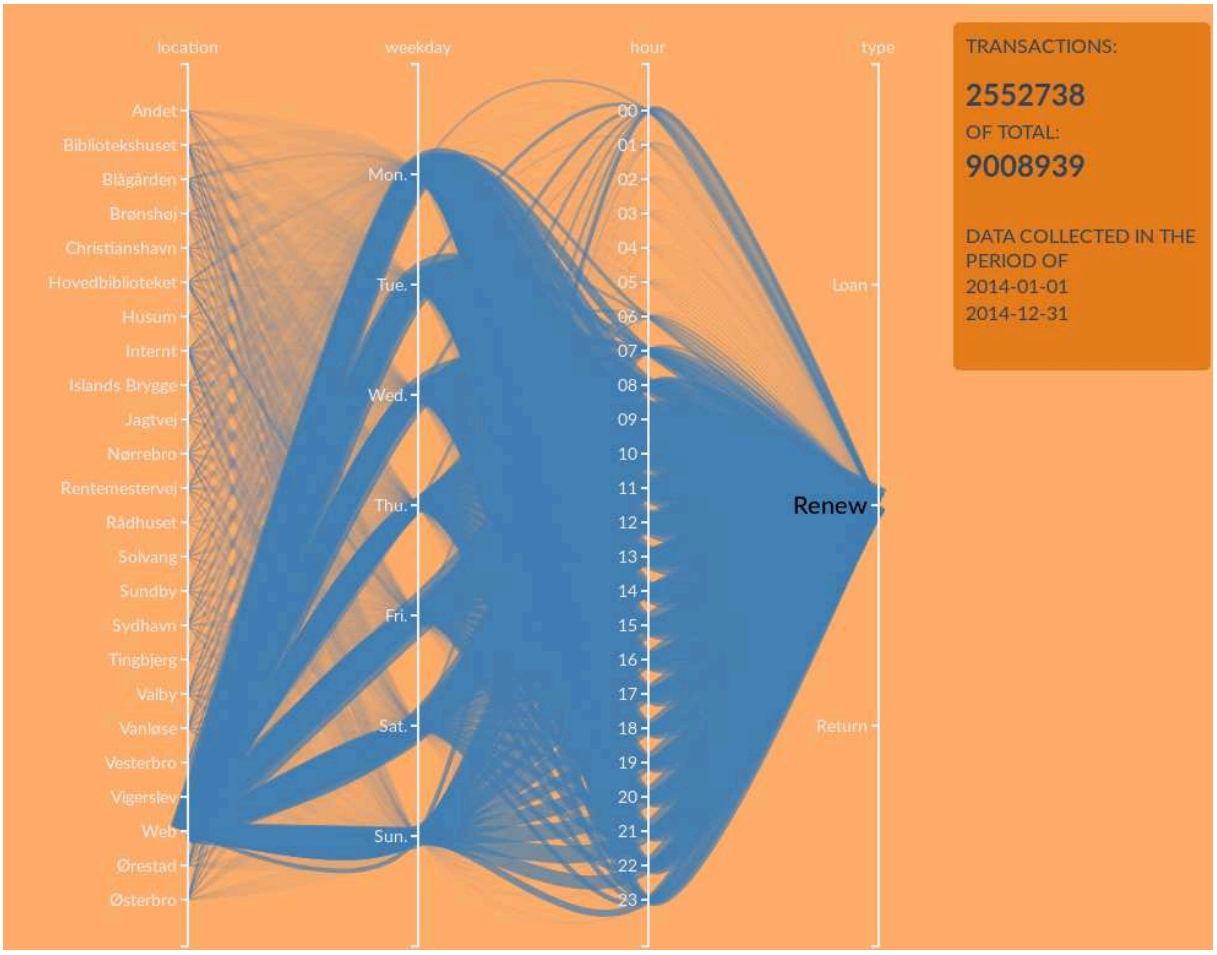


Figure 4 – Pivotviz highlighting the distribution of "renew" transactions for the municipal libraries of Copenhagen.

Source: http://kk.datavis.dk/thebooksof.html

portional to its value. This research work employs a combination of a treemap with nonproportional layouts, such as line charts and histograms. The goal is to facilitate the comparisons between parallel structures in a treemap, by embedding any nonproportional layout inside the treemap divisions. Although useful, the embedding might limit the visibility of parallel structures, as stated by the authors. We believe that the combination of a treemap with other proportional layouts may improve the visualization of multiple attributes, providing a more detailed analysis.

A visual analytics system for financial sector data, more specifically stock market, is presented in (LEI; ZHANG, 2010). According to the authors the analysis of this type of data is usually performed by examining multiple charts in conjunction, with no combination among them, consuming an excessive amount of time. In this sense, an holistic vision of the data greatly benefits the analysis, because it permits the identification of trends and to rapidly make projections. Figure 5 shows an example of two combined layouts in the system, the layout on top with a ring format shows the volatility of stocks, while

the one on bottom shows price clusters for stocks selected by clicking on the ring. This work is more concerned with analyzing data over time, however the characteristics present here, that is, the combination of layouts and interactions for the exploration and general visualization of a dataset in a visual analytics system, can be promptly transposed to a structural visualization-centric approach. It can be done using techniques such as multi-dimensional projections, parallel coordinates, and treemaps, providing similar benefits.

Figure 5 – Combined layouts for stock market analysis. Top pane is a ring visualization of stock volatility. Bottom pane is a visualization showing stock price clusters.

Source: (LEI; ZHANG, 2010)

Four visualization with different focus are shown in (SHARMIN et al., 2015), spatio-temporal, temporal, contextual, and event-centric. The visualizations are intended to aid just in time adaptative intervention, geared towards patients that suffer from stress. The idea of a visualization that uses contextual information, as shown in Figure 6, is interesting for this research, as the LACs data have several contextual data for each accident occurrence and can be exploited to identify profiles of accidents. Information such as where a LA occurred, the occupation of the injured, the economic activity, the sex of the employee, can be related in layouts such as parallel coordinates and variations, or even directly combined with multidimensional projections.

The research works cited in this section present a variety of INFOVIS techniques that help to highlight the structural characteristics of a dataset. This research explores the use of those techniques or modifications in them that may improve the analysis of the data in study, as well as to improve the extraction of information about LAs.

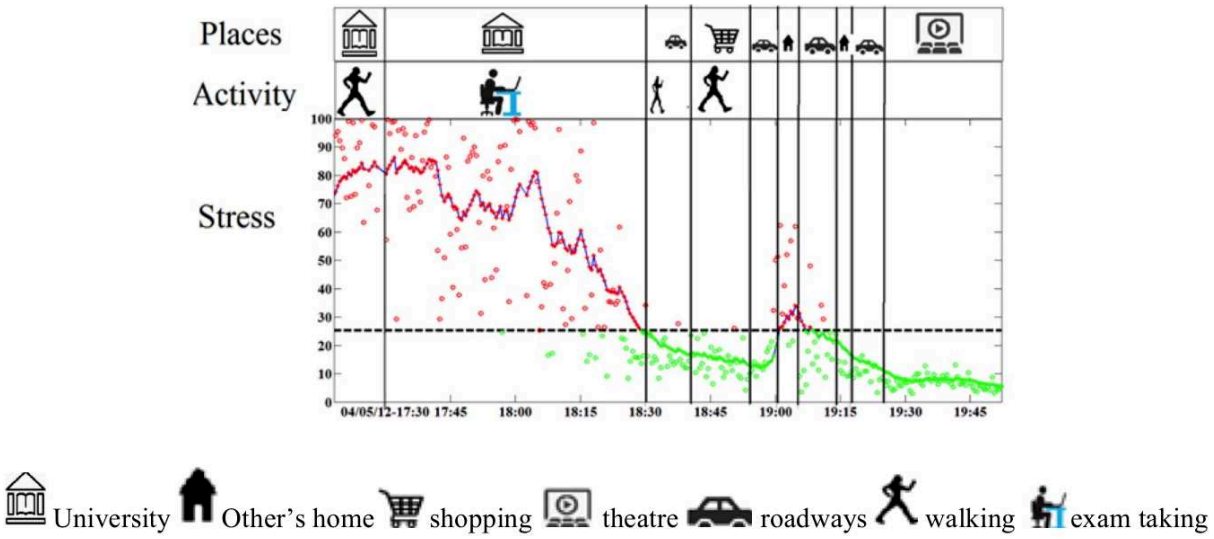


Figure 6 – Contextual visualization showing the stress profile of a participant during a day, as his activities change.

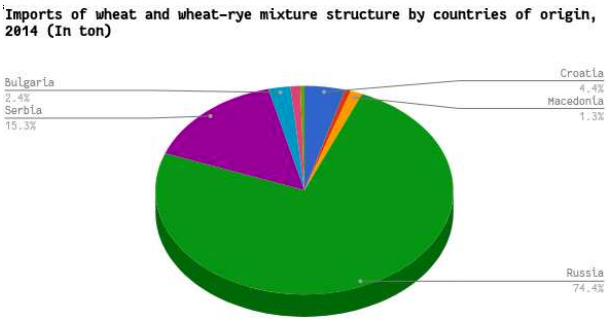
Source: (SHARMIN et al., 2015)

2.6 Governmental Data Visualization

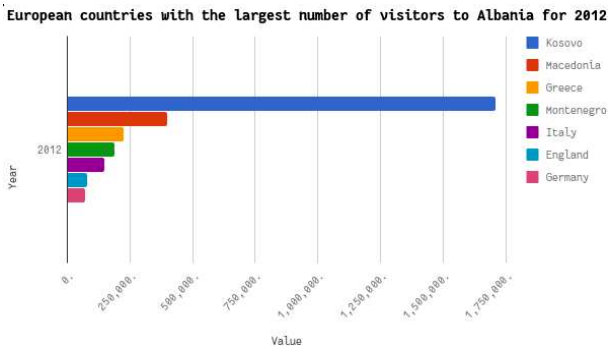
This section presents some works that make use of INFOVIS techniques for the visual communication of governmental data, showing the potential of these techniques in highlighting patterns that represent strategic information for managers and population.

In Albania a website was developed in which visualizations for their governmental data are provided(HOXHA; BRAHAJ; VRANDEčIć, 2011), in an attempt to make this data more accessible. This website presents data visualizations on education, economy, demography, poverty, science and technology, justice, tourism, agriculture and energy. While this initiative is interesting, the visualizations employed are simple and have low interactivity, basically composed of static graphs with tooltips. Figure 7 presents some examples taken from the website, showing piecharts, barcharts, and lines with bars. The layouts' simplicity may impair data analysis, because it may be difficulty to perceive complex patterns, relate different instances, identify groups, among other tasks.

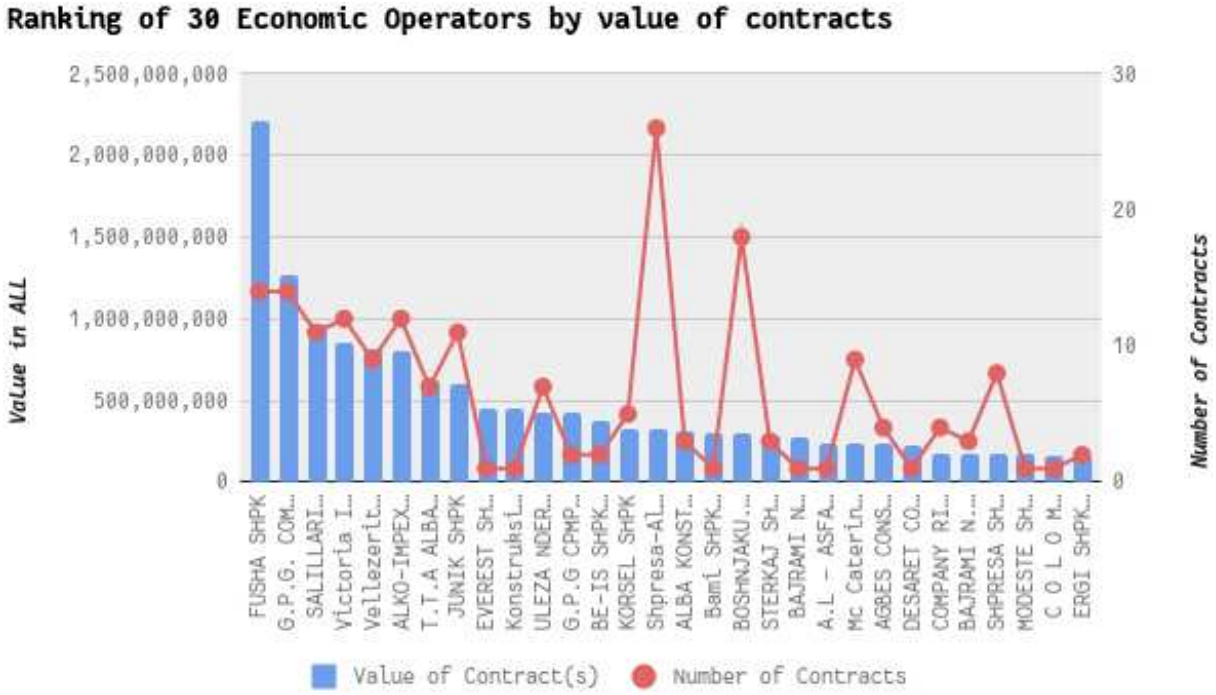
Dataviva(DATAVIVA, 2013) is a platform for visualization of Brazilian governmental data. It provides data about employment and income, commerce, education, and health, with the target audience being educators, entrepreneurs, as well as the general populace. Visualizations available include barchart, treemap, ThemeRiver, choropleth map, among others. Figure 8 shows a choropleth map , available in the platform, encoding the number of employees in higher education in Brazil in 2016. A divergent color scale is used, ranging from blue to red, in which red represents higher concentration. It is possible to notice that Sul and Sudeste regions concentrate the highest number of employees in higher education, with São Paulo being responsible for the highest concentration. However, for



(A) Piechart showing import figures for wheat and wheat and rye mix.



(B) Barchart of the number of visitors to Albania.



(C) Bar and lines chart showing 30 economic operators in Albania.

Figure 7 – Examples of charts on Open Data Albania.

Source: http://open.data.al

states that present similar hues, it is not possible to distinguish which has higher values, or what is the level of similarity between them. It is also difficult to know if there is any relationship between the states by looking only at this layout. There is also a lack of coordination among the layouts and it is not possible to compare, in a practical manner, indices without alternating between layouts. Furthermore, the layouts do not offer any interactivity except for the exhibition of tooltips. Other similar approaches can be found applied to open governmental data of other countries(International Food Policy Research Institute (IFPRI); DATAWHEEL, 2017; DATAWHEEL, 2016; DATAWHEEL, 2018).

In order to highlight the geographical distribution of governmental data, maps with pins are used in (MENDONçA; MACIEL; FILHO, 2014) to monitor the incidence of the *Aedes Aegypti* mosquito in the city of Cuiabá, using data from the Mato Grosso

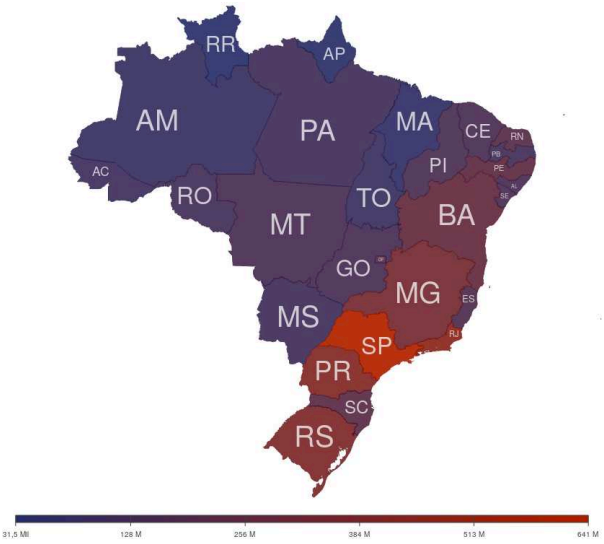


Figure 8 – Choropletic maps showing number of employees in Higher Education in Brazil in 2016

State Secretary of Health. However, the visualization has also low interactivity and exploration capacity and it is only accessible to authorized personnel, which restricts the potential to make this data transparent, limiting its analysis and popular participation. Also using maps to show incidence indices there is a website named *Observatório da Intervenção* (Intervention Observatory), that monitors the cases of violence and rights violation resulting from the military intervention in Rio de Janeiro. In this website there is a bubble map, in which each bubble represents a value corresponding to the number of violations in the region. By clicking on bubbles zooming is done on the map region and the bubbles are split to give greater precision on the locality of the occurrence, up to street level. An example of this map can be seen in Figure 9. There are also some tables available containing information about the occurrences.

Figure 9 – Bubble map showing number of violence/rights violation due to the military intervention in Rio de Janeiro.

Source: http://observatoriodaintervencao.com.br/dados/mapa-da-intervencao/

The works previously discussed fulfill more of an informative than analytical role, but there are other works that employ governmental data for analytical purposes, such as (ZHIYUAN et al., 2017), which makes use of several different visualizations to analyze the flow of passengers in the Shanghai Metro. This work shows that it is important

to offer different perspectives of the data over different layouts to understand big and complex datasets. The system uses bubble, bar, line charts, heat and flow maps, and starplot, highlighting different aspects of the data, such as passenger movements, how full is an station, paths most used, among others. In (RODRIGUES et al., 2017) a web application was developed to provide information about energy production in Germany. The application combines maps with glyphs and uses ThemeRiver to show the evolution of energy production in the power plants. An example of this application is shown in Figure 10, which shows the integration between the two layouts. Selections and filters can be used on the map, which in turn modify the ThemeRiver visualization. The focus of this research is to foment political discussion, nudging, and story telling, having an ample target audience, including the general population. In our research work we also combine different layouts and interaction tools to enable the analysis of different data aspects, thus aiding specialist analysis and the population's comprehension of the data.

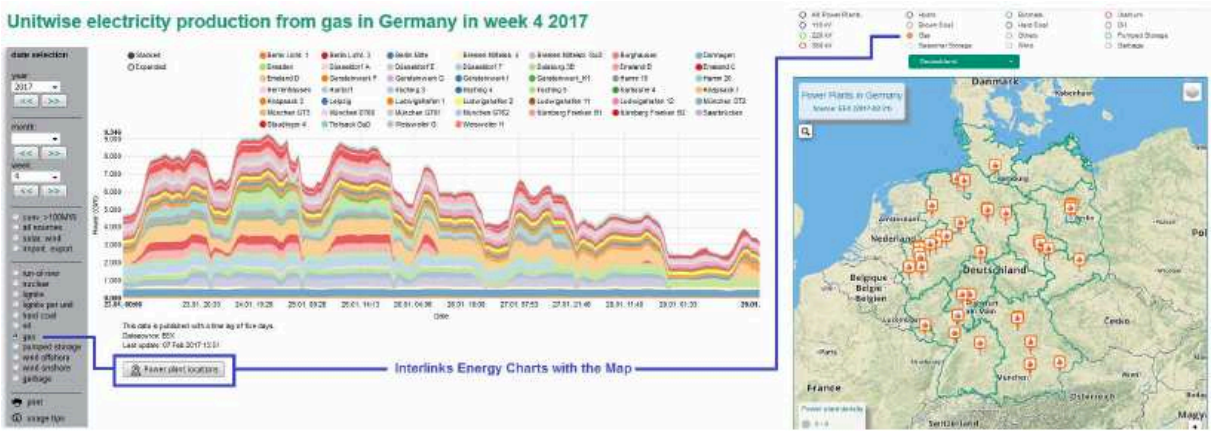


Figure 10 – Web application to visualize energy production in Germany.
Source: (RODRIGUES et al., 2017)

In the specific case of LAs, Groeger, Grabell e Cotts (2015) presents a visualization that shows the amount of compensation received, in the USA, by works that lost limbs due to LAs. The visualization consists of pictograms representing the human body, one for each American state, with the value of the compensation encoded as the size of each limb, as shown in Figure 11. It can be noticed that the state of Alabama pays the lowest compensations in general, while Georgia, a bordering state, grants much bigger compensations. This visualization works better as informative and not for deep analysis, but shows how intuitive visual metaphors can facilitate the comprehension of data and the similar behavior between geographically distinct regions.

In Brazil, the ODSST(Smartlab de Trabalho Decente MPT - OIT, 2017) is the platform currently provided by the BFLPO to communicate data about LAs. In Figure 12 some examples of available layouts in the website can be seen, and it is noticeable how the visualizations are simple. Figure 12A shows a non interactive piechart that shows only

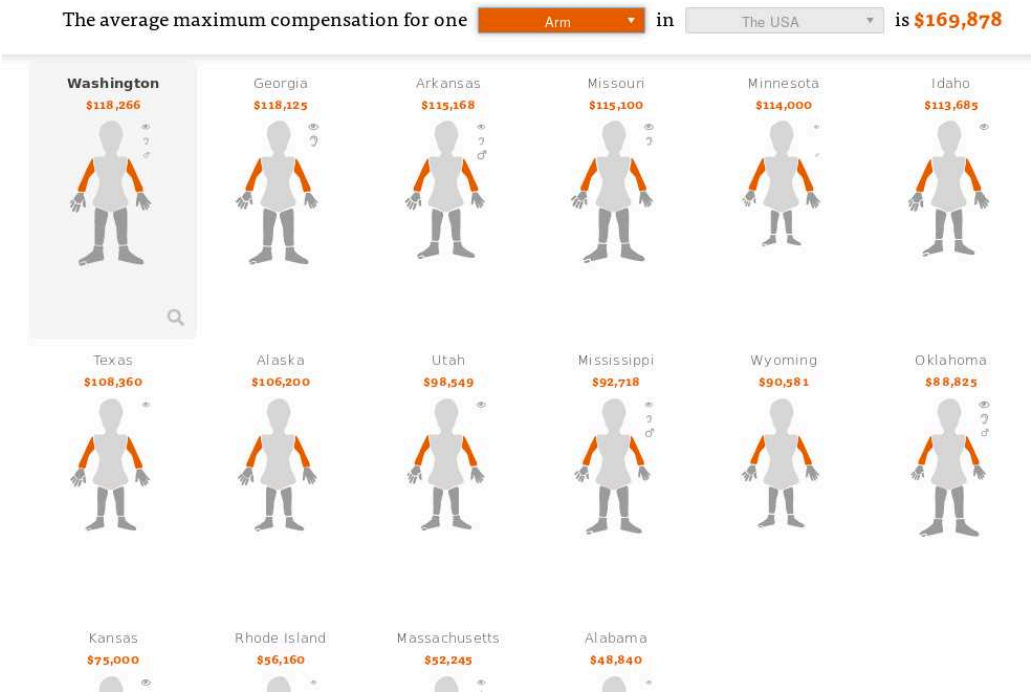
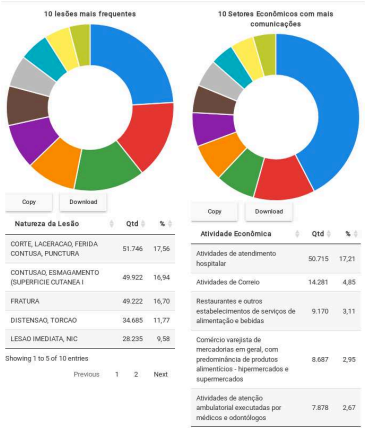


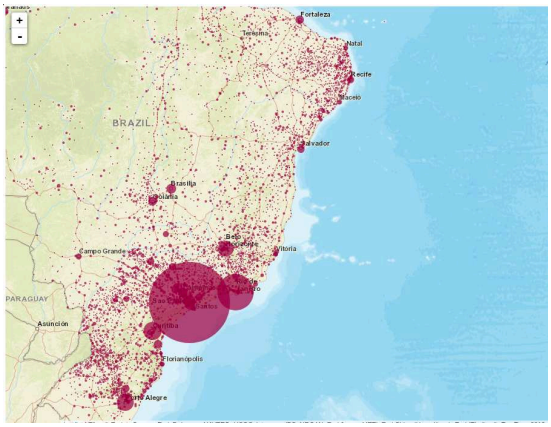
Figure 11 – Visualizations that show the values of compensations for LAs

Source: (GROEGER; GRABELL; COTTS, 2015)

a very limited part of the data, impairing its comprehension. This also occurs in the example of Figure 12B, in which the use of bubbles to encode the number of accidents might not be appropriate, since bigger bubbles might limit the visibility of smaller ones.



(A) Non interactive piechart showing quantities of LAs for injure nature and economic activity attributes.



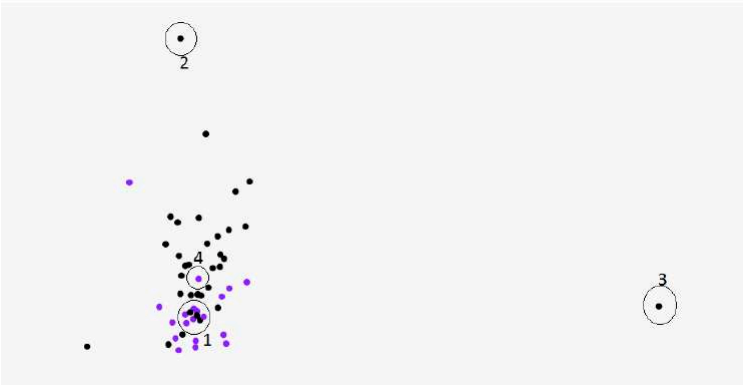
(B) Bubble map showing the incidence of LAs in Brazil.

Figure 12 – Visualization tools provided by the ODSST.

In (LIMA; PAIVA, 2017) data provided by DataViva are displayed using the PCA and MDS techniques. In the generated layouts it was possible to identify groups of localities with similar behavior profiles, as well as some localities with peculiar behaviors in relation to what was expected with the considered set of indices. Figure 13 shows

PCA and MDS scatterplots. Figure 13A shows the PCA layout of municipalities from the mesoregions of Triângulo Mineiro and Norte de Minas with respect to international commerce in which 4 groups can be noticed, 3 of them containing cities from different mesoregions, denoted by colors, presenting similar international commerce profiles. The fourth group is composed of a single outlier, the city of Uberaba, which shows commerce characteristics totally different from the others. Figure 13B shows a MDS layout regarding higher education in microregions from Santa Catarina and Rio Grande do Sul states. The distinct colors represent the state to which a microregions belongs, and it can be noticed some intersection among localities from different geographic regions, as well as two distinct outliers. Group 2 represents a microregion, Sananduva, with weak educational infrastructure and group 3 the microregion of Porto Alegre, which stands out from the rest because of the large values of the indices measured there. This research work shows how analysis strategies focused on structural characteristic of the data can contribute with the capacity of extracting relevant information.

(A) International commerce PCA layout of municipalities from the Triângulo Mineiro and Norte de Minas mesoregions.



(B) Higher education MDS layout of microregions of Santa Catarina and Rio Grande do Sul states.

Figure 13 – Multidimensional projection techniques applied to data from DataViva.

Source: (LIMA; PAIVA, 2017)

INFOVIS techniques have been seldom employed for the visualization of governmental data, with interesting results. There is still a lack of visual analytics systems for governmental data and none of them, to the extent of our knowledge and up to the writing of this document, was employed in the analysis of LAs data.

2.7 Final Considerations

Adequate analysis of governmental data has enormous potential to improve government policies and population welfare. However there is a lack of appropriate tools to facilitate this analysis and communicate such data. We showed that the use of INFOVIS techniques might be a solution to this problem. There are interesting research works capable of capturing the underlying structural characteristics of a dataset. However, the application of such techniques to the analysis of governmental data, specifically LAs data, is still lacking. This research aims to create a strategy for the analysis of governmental data, focusing on LACs, combining structural visualization techniques, such as maps, treemap and multidimensional projections, associated with sophisticated interactions to help in the efficient coordination of those layouts.

Chapter 3

Proposal

This research work proposes a strategy for visual structural analysis of LA occurrences data. The developed strategy intends to facilitate the comprehension of the structure of those data, by identifying accident behavior profiles, correlations between data measured in different geographic regions, among other tasks. We propose two visualizations, i) a combination of a multidimensional projection with a political map, and ii) a combination of a treemap with a parallel sets layout.

While this strategy is not restricted to a particular dataset, it was conceived to address a lack of analysis strategies capable of fulfilling the necessities of the BFLPO data analysis and should fulfill the following requirements:

r1: identify work profiles;

r2: identify areas lacking attention;

r3: characterize localities, independent of its geographical position;

r4: characterize wide geographical areas.

Thus, this chapter details the data from this repository, all necessary preprocessing steps, as well as the design decisions taken to produce the visualizations for the analysis.

3.1 Data Description

LAs data is freely available in the BFLPO repository[1]. Table 1 describes all the attributes used in this research. We decided to not consider the temporal attributes, as individually they do not add any strategic information to the analysis. The exception is the attribute "time", which was transformed in a binary attribute indicating if the accident occurred during the day or night shift. We refer to this new attribute as "shift".

[1] https://observatoriosst.mpt.mp.br/

Table 1 – Brief description of all attributes in the BFLPO dataset.

Attribute	Type	Description
Causer Agent	Nominal	Type of object that caused the LA.
Shift	Nominal	Work shift.
Death	Nominal	Indicates if the injured died or not.
Injury Nature	Nominal	Indicates the type of injury suffered.
Occupation	Nominal	Indicates the function developed by the employee.
Accident Type	Nominal	Indicates the situation of the LA occurrence.
Accident Locality Type	Nominal	Indicates the locality where the LA occurred.
Injured Body Part	Nominal	Indicates the part of the body injured.
Economic Activity	Nominal	Indicates the labor economic sector.
Sex	Nominal	Indicates the sex of the employee.
Age	Numeric	Indicates the age of the employee.
State	Nominal	Indicates the state (geographical locality) where the LA happened.
City	Nominal	Indicates the city where the LA happened.
LAC Emitter	Nominal	Indicates who reported the LA occurrence to the authorities through the LAC.

Only the state and city where an accident occurred is present in the original BFLPO data, so we added the complete Brazilian territorial division provided by the IBGE[2] (see Table 2), in order to provide detailed analysis for all geographic hierarchy levels. However, this information was not taken into account when generating multidimensional projections, as the aim here is to identify behavior profiles that are independent from geographical location.

3.2 Data preprocessing

The main idea of the proposed strategy is to allow the identification and characterization of groups of localities based on its structural characteristics, that is, the inherent relationships among data instances, manifested via correlations among measured attributes, behavior profiles, among others.

As can be seen in Table 1, the majority of the attributes are categorical, and some of them present a large number of categories. Thus, we decided to group them into fewer and broader categories, keeping their original meaning. We manually grouped the attribute "ds_parte_corpo_atingida", and grouped the attributes "ds_cbo", "ds_cnae_classe_cat"

[2] https://www.ibge.gov.br/

Table 2 – Brazil's geographical division, comprehending five hierarchical levels, from largest to smallest. Only states and cities have political autonomy, with their own laws and constitution but subordinated to federal laws and constitution.

Geographical Division	**Hierarchy Level**	**Description**	**# of Categories**
Region	1	Division of Brazil's states in groups based on similarities to help the interpretation of statistics, without political autonomy.	5
State	2	Subdivisions of the country with political autonomy but with some subordination to the country's government.	27
Mesoregion	3	Subdivision of the states to help the interpretation of statistics, without political autonomy.	137
Microregion	4	Subdivision of the mesoregions to help the interpretation of statistics, without political autonomy.	554
City	5	Administrative subdivision of the states with political autonomy but subordinate to the power of the states and the country.	5570

and ds_agente_causador" according to external official criteria. The sources used for each grouping and the resulting reduction in categories are presented in Table 3.

Table 3 – Attribute grouping in categories, as defined by the correspondent external sources.

Attribute	Categories before grouping	Categories after grouping	Grouping source
ds_cbo	2270	9	Brazilian National Classification of Economic Activities
ds_cnae_classe_cat	670	21	Brazilian Classification of Occupations
ds_parte_corpo_atingida	44	6	Author
ds_agente_causador	302	21	Ceded by BFLPO specialist

It is possible to use multidimensional projections for the analysis of a variety of data aspects, from different perspectives. We decided to perform the analysis from the cities perspective. We then summarized all LACs for each city, and used these summaries to represent each of them. However, effective summarization of categorical data is not trivial. In this work, we transformed all categorical attributes into numeric attributes, creating dummy variables. The summarization is then performed by summing up all the values for each city and calculating the mean of all occurrences.

3.3 Visual Analysis Strategy

Our strategy employs multidimensional projections in a two-dimensional scatterplot layout to reveal the underlying structure of the data, highlighting the similarities of instances, as well as outliers. Any projection technique can be used. As we are dealing with data containing associated geographical positioning, our strategy also uses a layout capable of showing the geographical location of a data instance, in such a way that they can be related to a projected instance.

For the second visualization we employ a hierarchical layout, that exploits the natural organization of the data to show instances distribution while also providing numerical context and filtering capabilities to another layout that shows relationships among several categorical attributes at once.

Figure 14 shows the architecture of our strategy. The following subsections detail the resulting visualizations on this architecture.

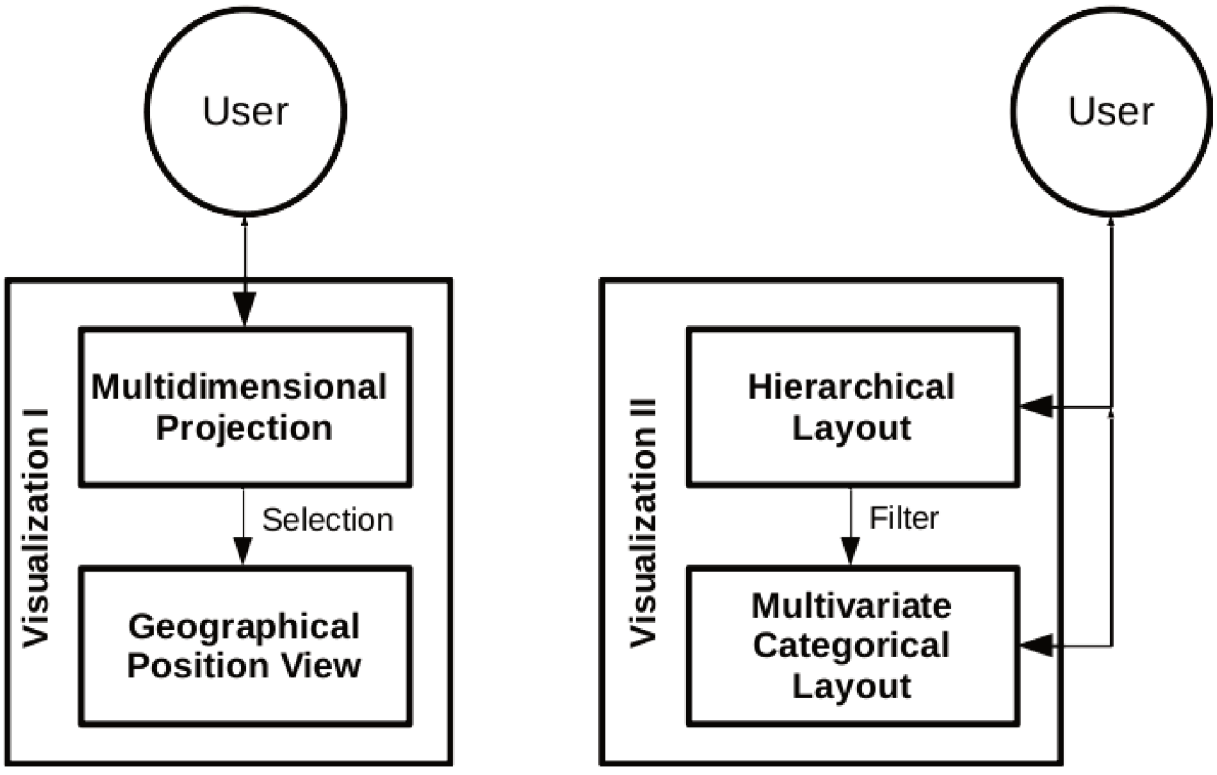


Figure 14 – Strategy Architecture.

3.3.1 Visualization I: Multidimensional Projections + Political map

As described in Chapter 2, multidimensional projections have been successfully used to analyze governmental data(LIMA; PAIVA, 2017). We decided to apply them in the

analysis of LA data, in order to reveal and highlight the behavior profile of the cities, independent of their geographical location, as well as their relationship regarding these profiles.

The scatterplot layout constructed from the multidimensional projection is presented in Figure 15. Each circle represents a city, and their color reflect a chosen category to which the corresponding city is assigned. As the considered dummy variables do not have any intrinsic value, we used the mode category of an attribute to assign the color to a city according to that attribute. We also associated a political map to the projections, so that a user can readily identify the geographical location of the cities under analysis. Figure 15 shows the resulting composition.

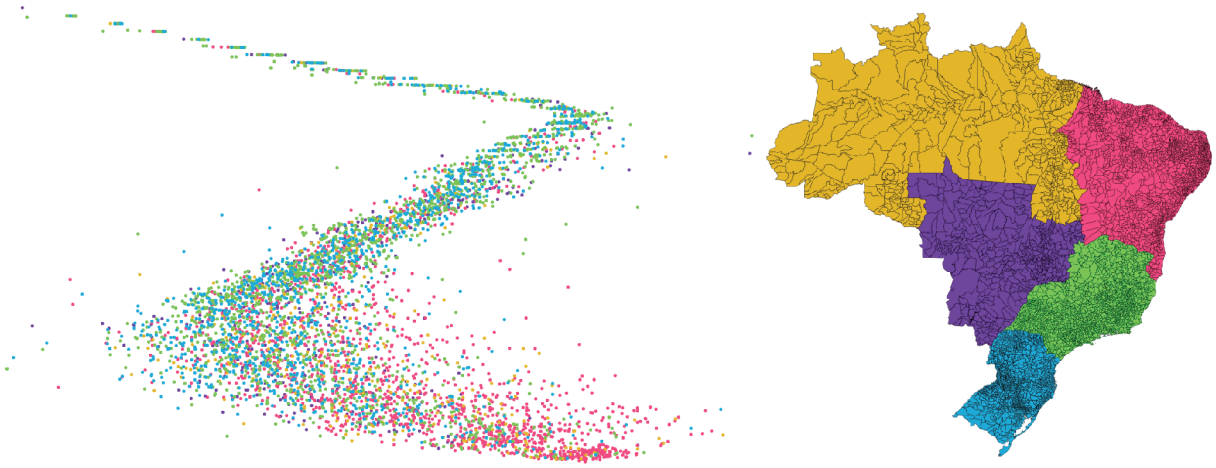


Figure 15 – Visualization I: Multidimensional Projection + Political Map.

3.3.2 Visualization II: Treemap + Parallel Sets

Treemap(SHNEIDERMAN, 1992) is a space filling layout, in which geometric figures, usually rectangles, are nested and the size of each rectangle encodes a value. These rectangles are organized with sizes proportional to a specific measured attribute and nested according to the hierarchy present in the data. Figure 16 shows a treemap of the quantity of LACs in Brazil between 2012 and 2017. Observing the organization and the size of the rectangles it is easy to perceive the quantity of LACs by region and the relative order between them. Inside each rectangle the next hierarchy level, considering the information of that rectangle, can also be visualized, allowing the same analysis strategy but considering a specific data portion. The LAC data from the BFLPO have attributes that can be organized hierarchically, making the treemap a natural layout for the representation of such information. Furthermore, the treemap also provides exploration of subsets of the data and communicates the distribution of LACs in different localities, providing numerical context for accompanying categorical analysis.

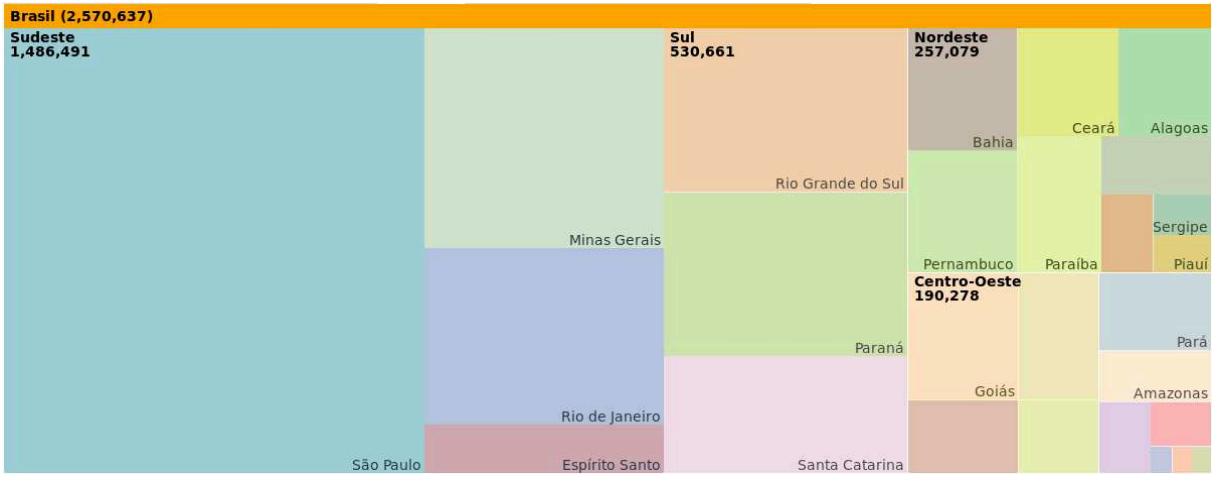


Figure 16 – Labor Accidents in Brazil

Parallel coordinates is a widely used layout to show multivariate data, as discussed in Chapter 2 Section 2.5. However it is not appropriate for massively categorical datasets, which is the case of the BFLPO repository. Parallel sets(KOSARA; BENDIX; HAUSER, 2006) is a layout derived from the parallel coordinates, specially designed for categorical data visualization. This layout uses the parallel axes, to visualize several attributes simultaneously, but focuses on showing sets and subsets of items, instead of individual data points, and on showing how these different sets relate to each other. Each attribute in the dataset is represented by an axis, divided in pieces (categories), whose sizes are proportional to the frequency of each category. A "ribbon" connects a category in an axis to a category in another axis, if they both occur simultaneously, meaning that the width of a ribbon is the intersection of the sets up to that point. Thus, "paths" are formed by these ribbons, which can be changed by reordering the axes/categories, resulting in different visual patterns.

An example of parallel sets showing data on survivors of the Titanic shipwreck can be seen in Figure 17. In this figure one notices that even though the number of survivors were almost evenly split between women and men, only a small fraction of men survived, while the majority of women have survived. The further ribbon paths help to narrow down on the survivor profile, allowing to notice, for instance, that approximately half of the surviving men were part of the ship's crew.

Parallel sets is appropriate to show the relationship between the different attributes characterizing LAs, as well as to permit the exploration of this data from different perspectives, by starting with different attributes such as gender or economic activity. The treemap provides the exploration of different granularities of the data in the parallel sets and contextualizes the amount of LACs in a locality. The resulting visualization can be seen in Figure 18.

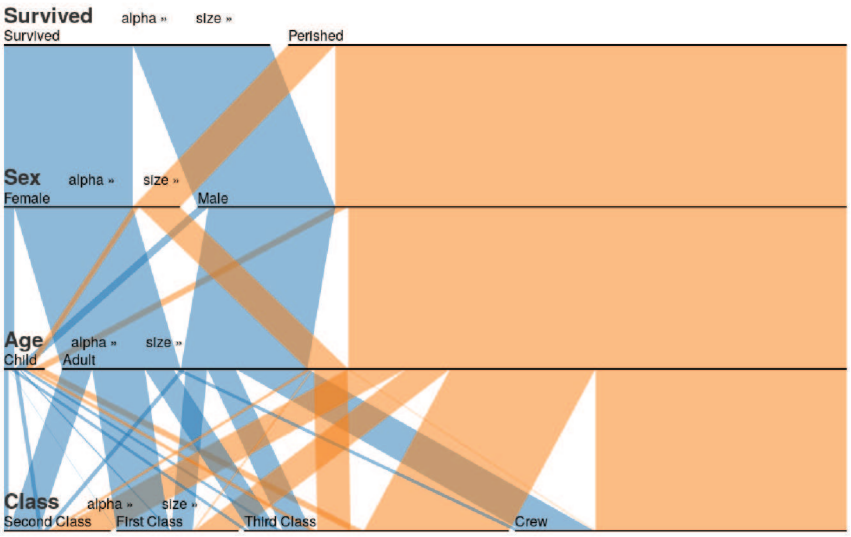


Figure 17 – Classic example of Parallel sets depicting the breakdown of survivors of the Titanic shipwreck.

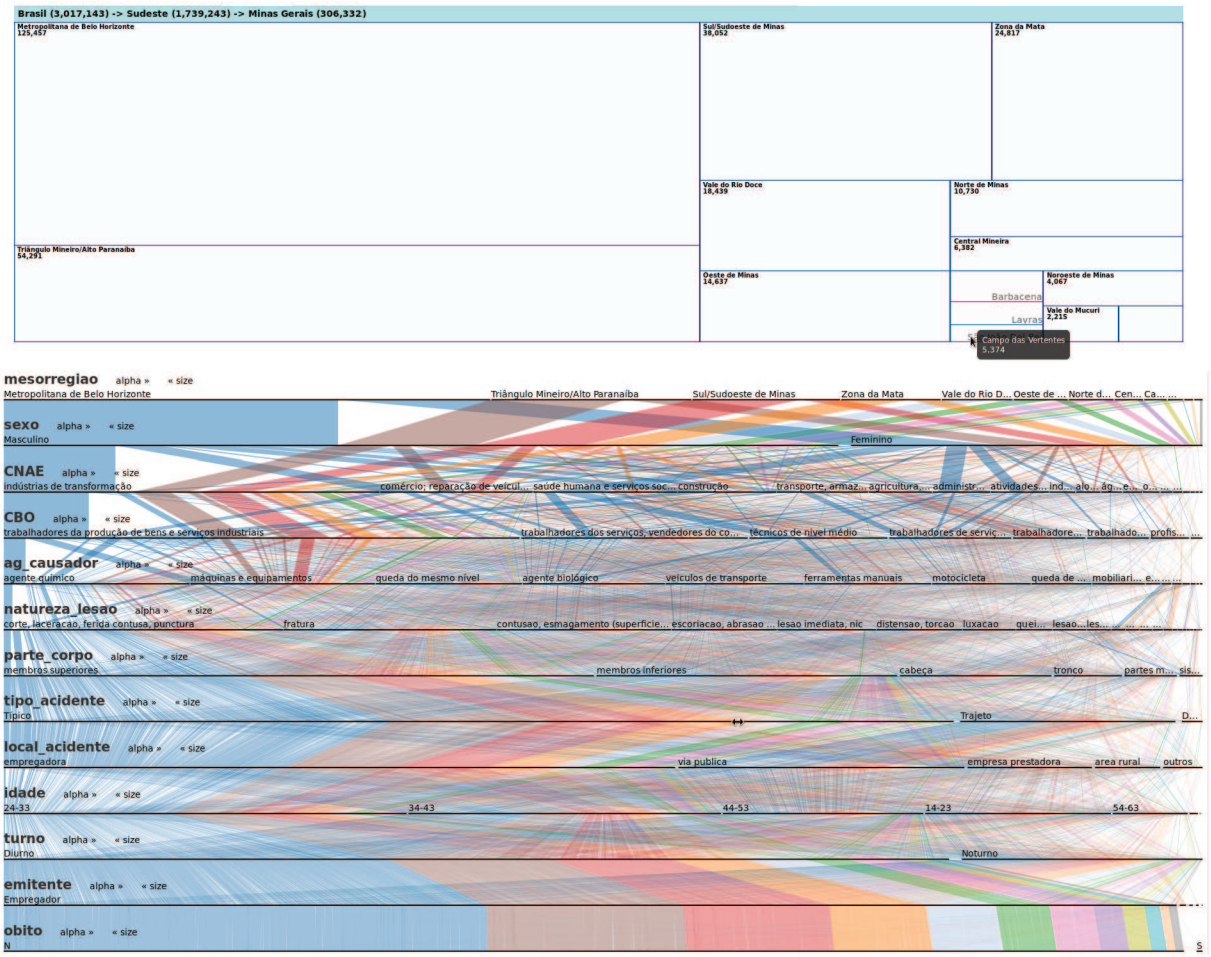


Figure 18 – Visualization II: Treemap + Parallel Sets.

Chapter 4

System Description

This chapter presents a developed system that implements the proposed analysis method described in Chapter 3. It details each visualization and its accompanying interactions, as well as how we combine them to provide several data perspectives. This chapter also presents the materials used for the development of the system.

4.1 Visualization I: Multidimensional Projection + Political map

We employ a scatterplot layout to present the results of the multidimensional projection. As discussed in Chapter 3, we expect that the layout will reveal patterns that communicate the behavior profile of the cities, as well as how they relate to each other regarding these profiles. Each point in the scatterplot represents a city, using information from 2012 to 2017, colored according to the mode of a specific categorical attribute. The categorical attributes available for coloring are: Region, State, Mesoregion, Microregion, Causer Agent, Injury Nature, Occupation, Accident Type, Accident Locality Type, Injured Body Part, Economic Activity, Shift, and LAC Issuer.

After choosing a coloring attribute, another dropdown menu is used to select multiple categories in that attribute, which results in toggling the color of all the points having that category as mode, that is, points not currently highlighted are colored black, and the others receive a color corresponding to its mode category. Finally further to the left there is a checkbox which allows the user to toggle the color of all points at once, that is they can be all colored black, or each receive the color of its category. Figure 19 shows an example in which the colors of all cities were toggled off with the checkbox, and subsequently the category “Minas Gerais” of the Attribute “uf” was seleted.

A point or a group of points can be freely selected using a polygonal selection tool, and the selected points are highlighted by coloring the circle borders in red. Figure 20 shows an example of a possible selection with this tool. The formed selection shapes can

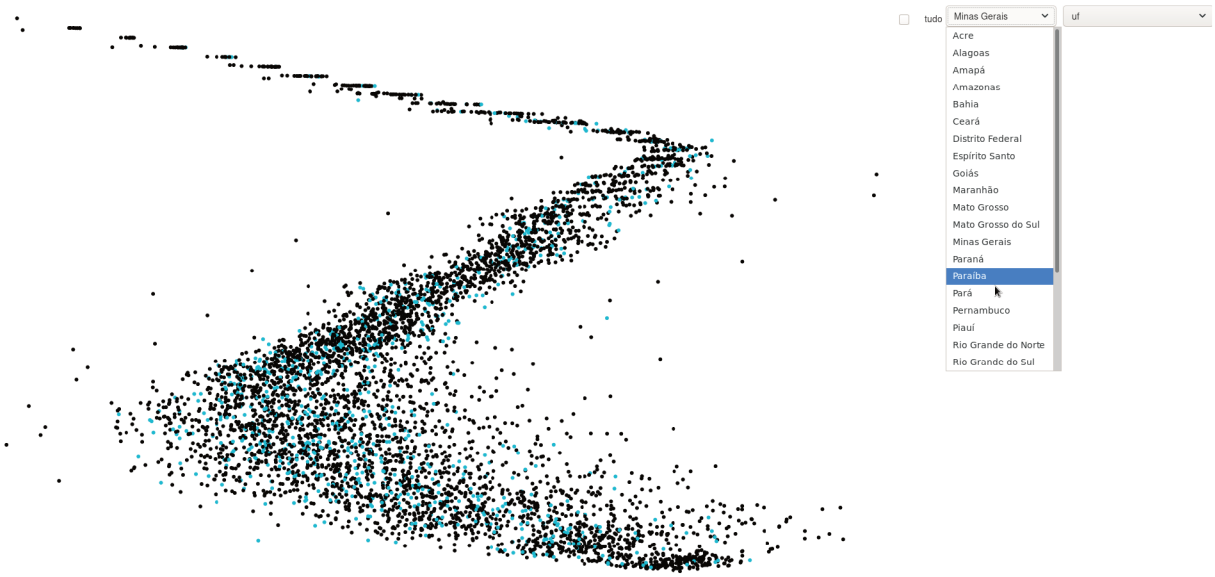


Figure 19 – Attribute "uf" is chosen for coloring. The checkbox is unmarked, which turns all colors off. The blue points represent all cities of the Minas Gerais state, which was chosen manually for coloring. The dropdown shows all categories in the attribute "uf".

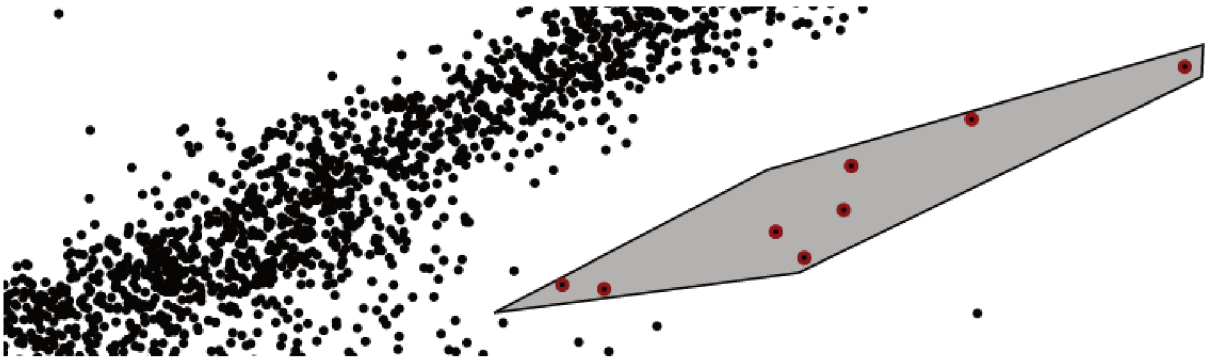


Figure 20 – Arbitrary shapes can be created and moved around to select points.

also be translated to perform different selections. By hovering over the selection a tooltip displays the name of the selected city(ies).

All the selected cities in the scatterplot are highlighted in a political map in the visualization, in order to contextualize user exploration. By showing where each selected city is located, the visualization gives insight about the geographical distribution of cities with similar behavior in the country, and helps users to identify all the cities in a selection, specially in situations in which points occlusion occur, enhancing the analysis and possibly improving the comprehension of several associated phenomena. Figure 21 shows the resulting visualization.

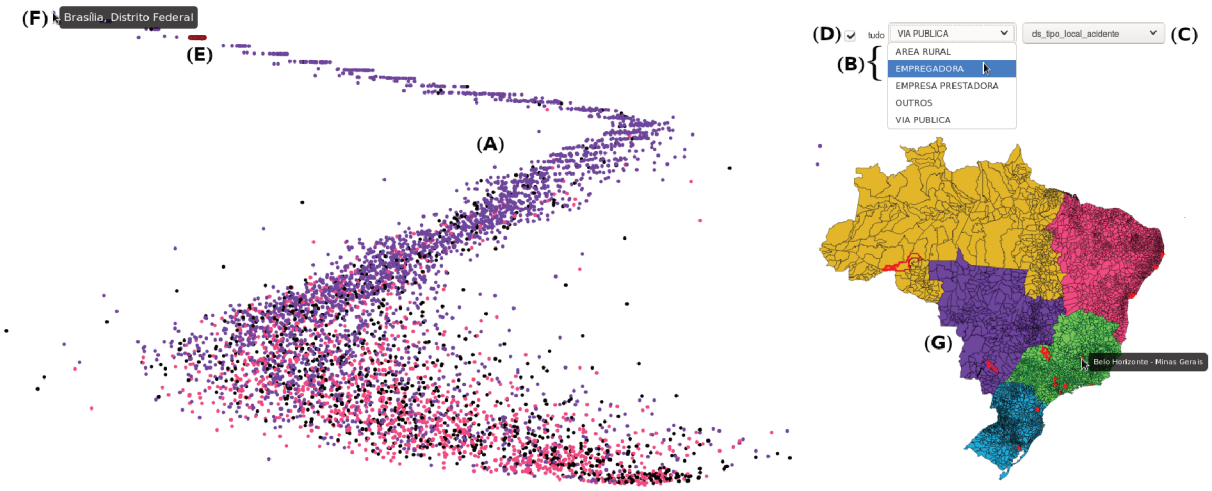


Figure 21 – (A) Scatterplot layout with LAMP projection results. (B) Dropdown menu to select which points to toggle the color based on the most frequent category. (C) Dropdown menu to select coloring attribute. (D) Checkbox to toggle color of all points. (E) Points highlighted by user. (F) Tooltip showing a town name. (G) Cities present in the highlighted points are also highlighted in the map.

4.2 Visualization II: Treemap + Parallel Sets

The treemap is useful to highlight aspects related to the hierarchical disposition of the data. In the system, each cell of the treemap represents a locality, and double encodes the amount of LACs for a locality in its size and hue. Darker hues represent a greater absolute value of accidents. The number of LACs in a locality is also displayed in its corresponding cell. The treemap implemented is an expandable treemap, which means a cell is expanded when selected. At the top of the treemap users can see the full navigation path and the total amount of LACs for the current level of exploration: region, state, mesoregion, microrregion, or city, and users can click on the path to return to higher hierarchical levels. Finally, hovering over a treemap cell shows a preview of the subsequent hierarchy associated with that cell, if applicable. Figure 22, shows an example of these interactions.

The implemented parallel sets provide an initial suggested configuration, which can be changed by the user, with the localities in the first axis. The axes/attributes can be reordered by dragging, and the attributes can also be automatically sorted by size/name. Hovering over a ribbon highlights the full path up to that point and shows the absolute number of LACs in that path, and the proportion to the total of all localities. This helps the user to see the proportions of attribute values in each axis, and highlight the relationship between categories in different attributes.

The treemap is coordinated with the parallel sets layout, providing navigation on the hierarchy and more context to explorations in the parallel sets. The current level shown in the treemap is also shown in the parallel sets, allowing the exploration of localities in both perspectives. Figure 23 shows the complete visualization.

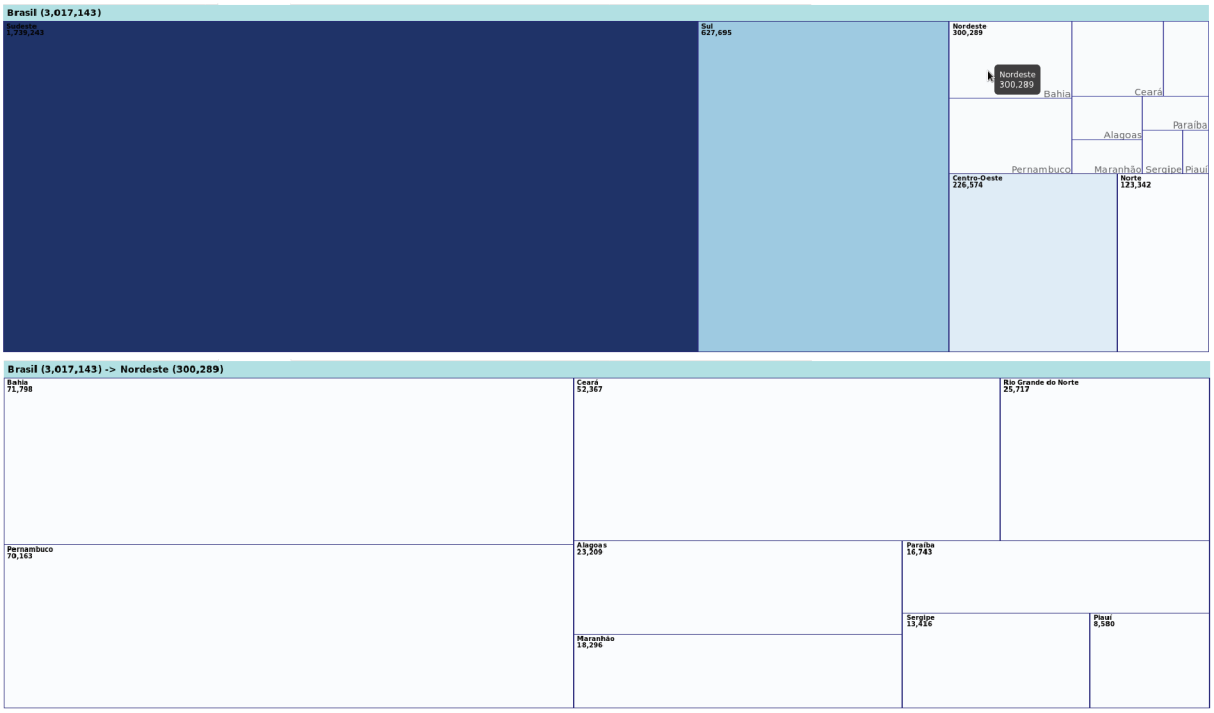


Figure 22 – The treemap at the top is at the highest level of the hierarchy, the regions level. Hovering over Nordeste region shows a preview of its states. The treemap at the bottom shows the result of expanding the Nordeste cell. At the top of each treemap the current level and amount of LACs for each level are shown.

Both visualizations presented in this chapter are independent of one another, but they provide complimentary exploration. In this sense, both visualizations provide interesting means to determine the profile of the localities, even if they are geographically distant, and the results of one can give insights to analysis in the other.

4.3 Implementation details

For the development of this system two software development tools were used, the programming language R(R Core Team, 2018) with the Shiny(CHANG et al., 2017), mp(FATORE; FADEL, 2016), and tidyverse(WICKHAM, 2017) packages, and the javascript programming language with the D3 library(BOSTOCK; OGIEVETSKY; HEER, 2011).

R is a programming language focused on data analysis and statistical calculations, and it was used to preprocess the data, as described in Chapter 3, generate the projections, transmit data to D3, and generate the web layout for the visualization.

Tidyverse is a collection of R packages for data science. It provides several amenities for flexible and fast data processing, and was used extensively to process our dataset to be used by D3. The mp package provides several functions that generate multidimensional projections, and was used to generate different projections for our dataset. The Shiny

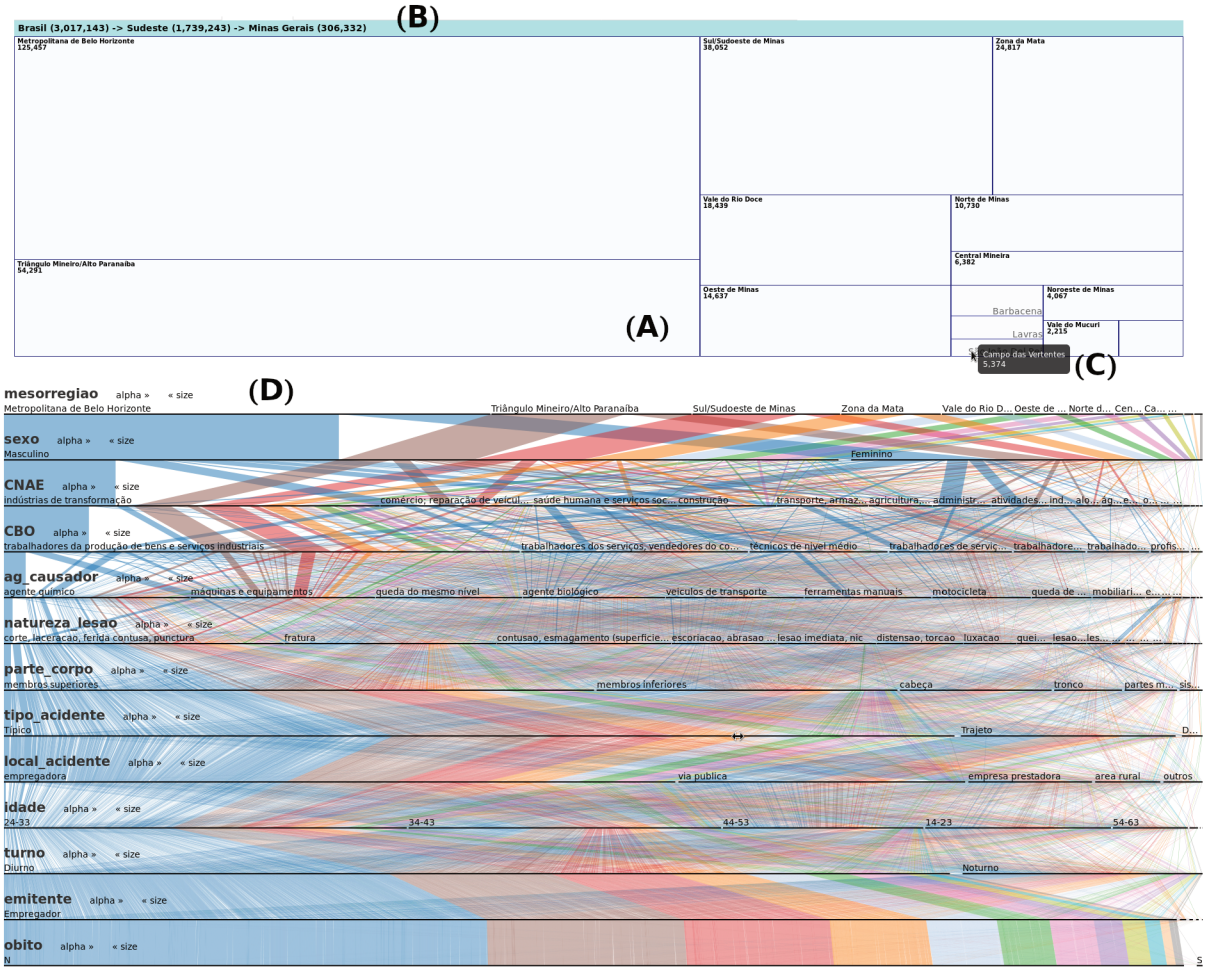


Figure 23 – (A) Treemap with Brazil's hierarchy. (B) a bar with the path into the hierarchy, showing the name of the locality and the amount of LACs in each hierarchical level. (C) Cell hovering showing the subsequent hierarchy and a tooltip with the name of the locality and the number of LACs. (D) parallel sets showing the locality selected in the treemap.

package aids in the creation of web applications and offers an API for efficient interaction with the javascript language. It basically generates boiler plate html code and provides some web layouts ready for use. It also establishes a two-way communication between R and javascript, allowing the sending of data from R to javascript and vice-versa, using the JSON format.

Javascript is a widely used programming language for the web, and its choice was motivated by its portability, requiring just a modern web browser. We specifically used the D3 library, which is a library frequently used in visualization projects, including some of the presented ones in Chapter 2, because it allows the practical manipulation of web elements and makes the development of layouts and interactions faster and more flexible.

Finally a collection of GeoJSON files containing geographical information of Brazil(LUIZ, 2014) was also used to create the political map.

Figure 24 shows the interaction between these parts.

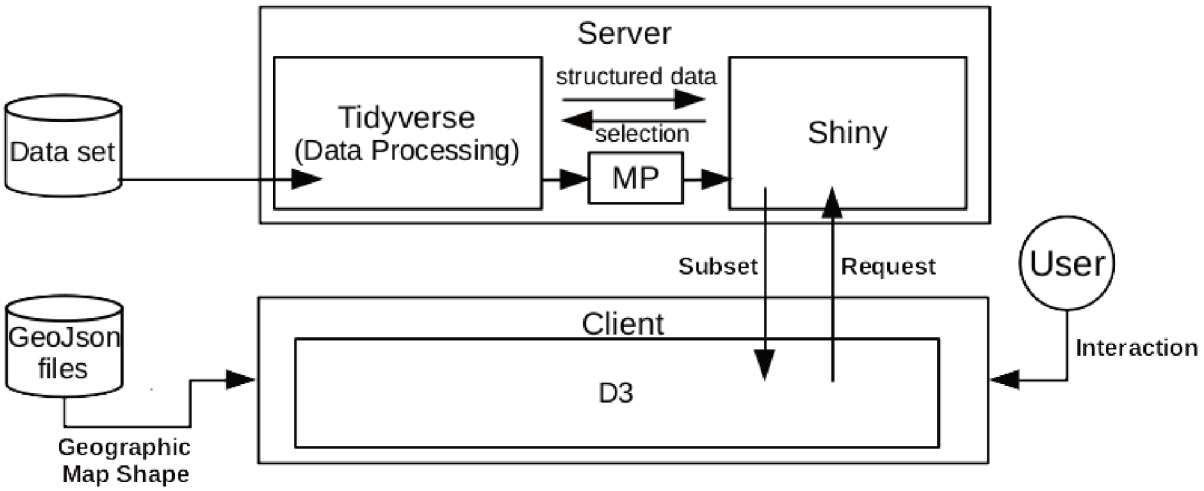


Figure 24 – System overview.

Chapter 5

Results

This chapter presents the results of applying our proposed strategy in the analysis of the BFLPO data set. We start by detailing the analysis procedure adopted in the experiments, and then discuss our findings, highlighting the importance of each layout for the analysis, as well as how the complementarity of the visualizations may improve the analysis.

5.1 Analysis Procedure

We performed several analysis using the visualizations independently, in order to explore their individual capabilities. However, we also explored how each visualization may shed more light on the findings from the other, to produce a more complete analysis.

Using the multidimensional projection + political map we first focused in analyzing the points distribution in the layout, identifying groups of cities, as well as isolated cities that may present anomalous behavior. We also investigated the categorizations produced by each coloring attribute in order to understand how they are related to the LAs behavior. The geographical map was used to better identify/understand profiles that are independent of the cities' geographical location.

We freely explored the parallel sets + treemap visualization to characterize broader locality groups, defined by their geographic location, such as regions and states, by investigating the role of each measured variable in their characterization, or in distinguishing them, as well as if/how they correlate to each other. This visualization was also used to further explore some interesting findings of the multidimensional projection layout, to help in the comprehension of how each measured variable influences the characterization of the groups present there, and in the identification of correlations among the involved attributes, the distribution of instances among attributes, behavior trends, among other tasks.

In each analysis we highlight elements in the proposed layouts that contributed to reveal each pattern, and to lead to a judicious decision making, and associate these

analysis with the requirements presented in Chapter 3, in order to highlight how these requirements were fulfilled. Finally, whenever we felt necessary, to explain some behavior, we consulted external sources, such as news or the raw data.

5.2 Choosing a projection technique

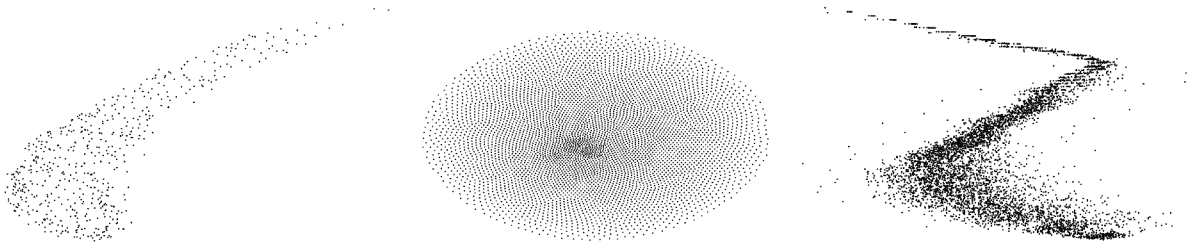
In our experiments, one of the visualizations employs a scatterplot which represents the results of a multidimensional projection technique applied to the data. Thus, it is important to choose an adequate technique, which results in a layout that is capable of highlighting patterns representing important aspects related to the LA information. Several multidimensional projection techniques exist, as shown in Chapter 2, and we investigated three state of the art techniques, implemented in the mp package described in Chapter 4, in order to decide which was best for our dataset: LSP(PAULOVICH et al., 2008), t-SNE(MAATEN; HINTON, 2008) and LAMP(JOIA et al., 2011). The parameters used for each technique are summarized in Table 4, and were suggested by the authors.

Table 4 – Main parameters used for multidimensional projection techniques.

LSP	
Sample Indices	random
Neighbors	15
Target Dimensionality	2
t-SNE	
Perplexity	30
Iterations	1000
Target Dimensionality	2
LAMP	
Sample Indices	random
Proportion of Nearest Control Points	1

Figure 25 shows the three layouts generated by LSP, t-SNE and LAMP, respectively. LSP depicted huge numbers of localities (> 1000) overlapped in single points and no distinguishable groups. t-SNE created an oval shape also with no distinguishable groups. LAMP generated a layout with better group separation for the BFLPO dataset. Even though some groups were found in LSP and t-SNE layouts, upon investigation they did not help in revealing any information leading to interesting results. In order to confirm how LAMP better represents the relationship observed in the original space, we employed the Neighborhood Preservation(PAULOVICH; MINGHIM, 2008) quality measure on the three layouts. This quality assessment measures the proportion of k nearest neighbors of all data instances that are also nearest neighbors in the visualization space, and is used to evaluate how the projection preserves the relationships observed in the original space. The results are shown in Figure 26. It is possible to notice that, t-SNE provides better results for up to 7 neighbors, due to its local preservation capabilities. However, both

LAMP and LSP perform better for more than 7 neighbors, with LAMP being the best overall. For these reasons LAMP was the technique chosen.



(A) LSP (B) t-SNE (C) LAMP

Figure 25 – Layouts of the BFPLO dataset using three state of the art multidimensional projection techniques.

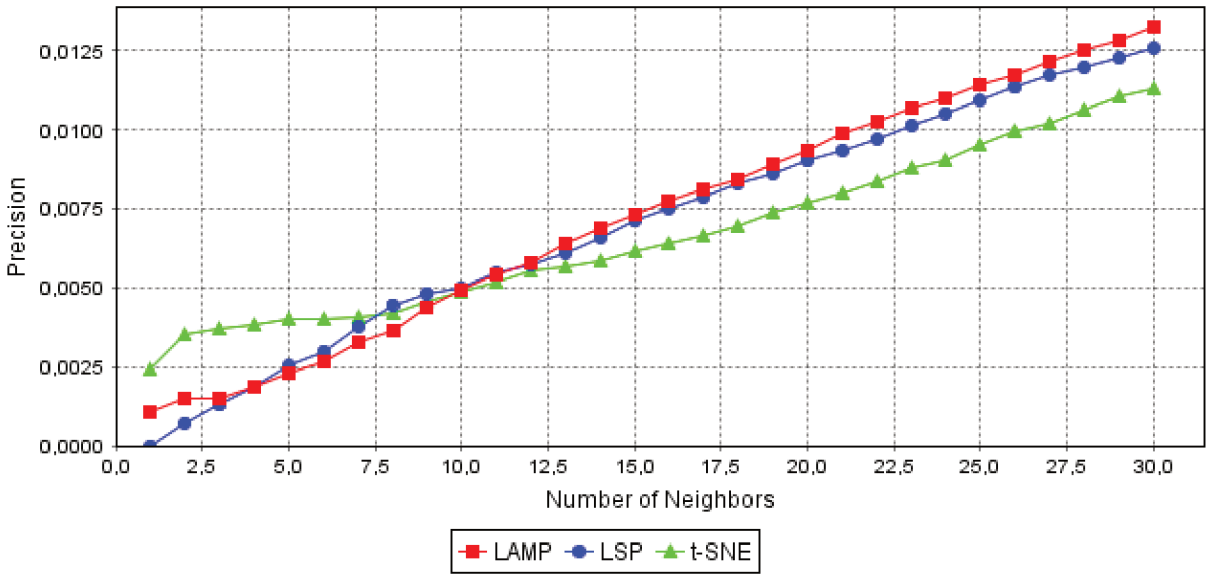


Figure 26 – Neighborhood preservation comparison of the layouts presented in Figure 25.

5.3 Results

Our goal is to identify, in localities, behavior profiles associated with LACs. Figure 27 shows the initial scatterplot of the LAMP projection, in which cities are colored according to the region they belong to. One can notice in this figure that the regions are well distributed in the layout, except for the pink points, which correspond to localities belonging to the Nordeste region. The layout produced a "Z" shape, with small groups present at the top. We noticed that these small, more coherent groups are composed, in general, of big cities and regional centers. This suggest that this point distribution might be related to the cities' size and economic development.

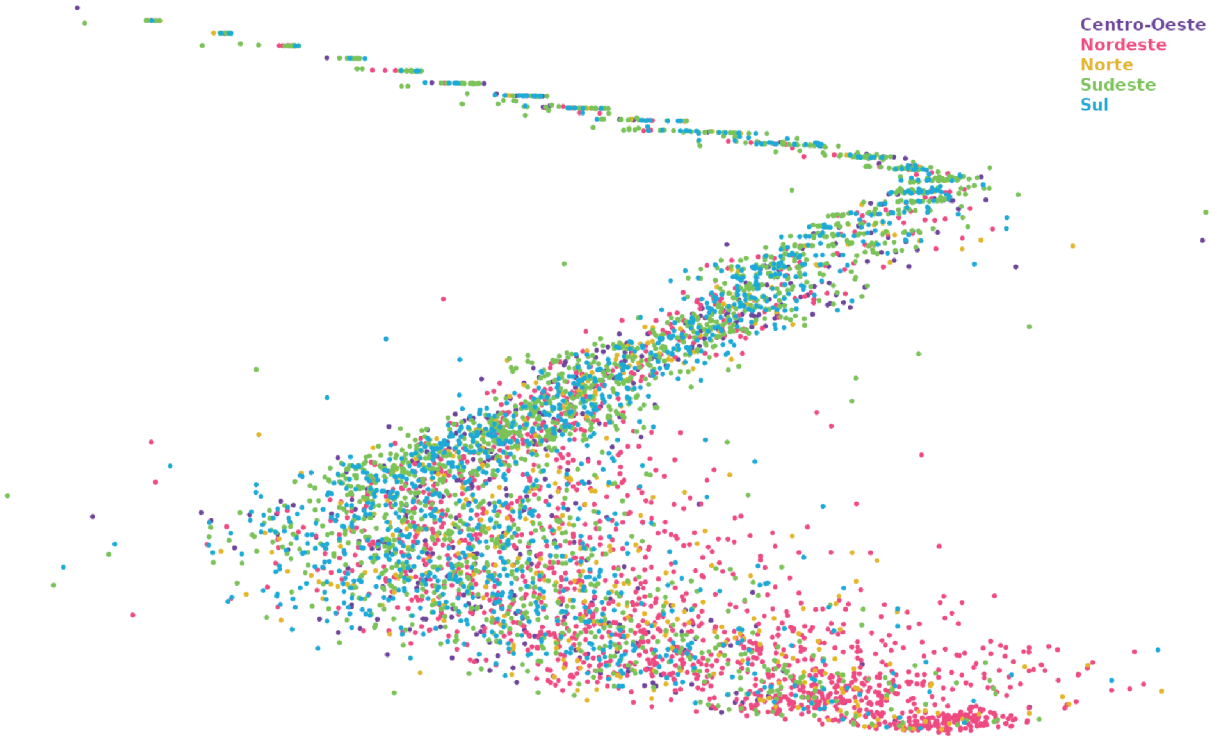


Figure 27 – Scatterplot corresponding to the results of applying the LAMP projection technique to the BFLPO dataset, in which a "Z" shape can be noticed. Cities are colored according to the region they belong.

Using place of occurrence of an accident as coloring attribute we have further indication that the points distribution reflect cities' size and development, as it is possible to notice that rural areas are located mostly at the bottom, and by inspecting some points we see that is the case for smaller cities too. Figure 28A presents a scatterplot in which only the cities where the majority of accidents occurred in rural areas were colored. Figure 28B, in turn, presents a scatterplot in which only the cities of the Nordeste region were colored. It is possible to notice, that the cities from Nordeste, which is one of the poorest regions, are mostly positioned at the bottom of the layout, and greatly overlap with the positioning of rural and small cities.

Also corroborating this analysis, one can notice that Sul and Sudeste regions' cities are homogeneously distributed over the layout, as shown in Figure 29, reflecting the economic development diversity of cities from these regions.

Using the economic activity as coloring attribute it is possible to notice that two economic activities are reported as most significant in the majority of cities, they are "transformation industries" (2021 cities), and "commerce and repair of vehicles" (708 cities) and can be seen on Figure 30. The average LACs in those categories is 450 and 600, respectively.

Although "transformation industries" and "commerce and repair of vehicles" are the most reported economic activities, there is a set of cities of particular interest that report

(A) Rural areas. (B) Nordeste cities.

Figure 28 – Scatterplots showing the positioning of rural cities and Nordeste cities. It can be noticed that rural areas largely overlap with Nordeste cities.

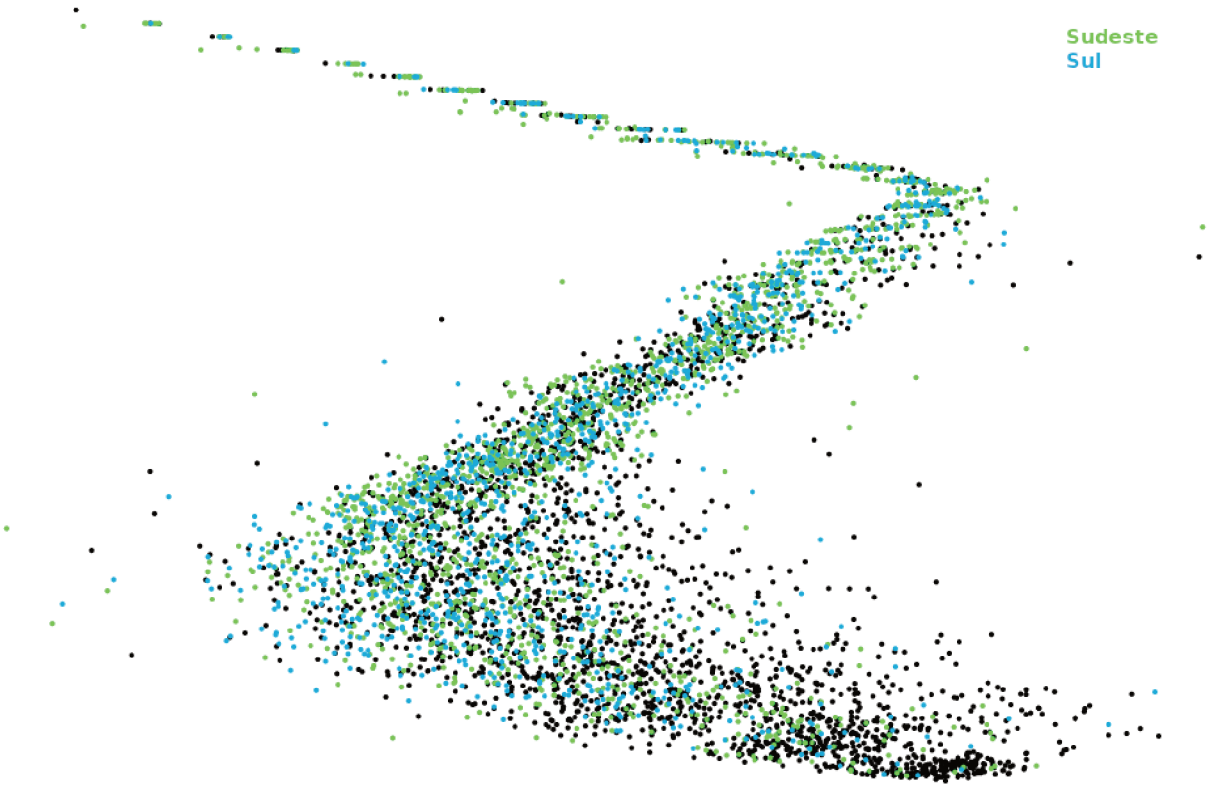


Figure 29 – Scatterplots showing the positioning of cities from the Sul and Sudeste regions. Those cities are well distributed over the layout.

the majority of its LAs as "human health and social services", shown in Figure 31. This set is composed of 127 cities, comprising many capitals/big cities, including all 7 capital cities from Sudeste and Sul regions, 3 from Centro-Oeste region, among them the Federal capital city Brasília, and 4 from Nordeste region, which amounts to 14 out of the 27 capitals. Each city in this group reported on average 3000 accidents related to this economic activity, much higher than any other category. This analysis shows how economic activity and related accidents are different in rich/big cities from the rest of the country, independent

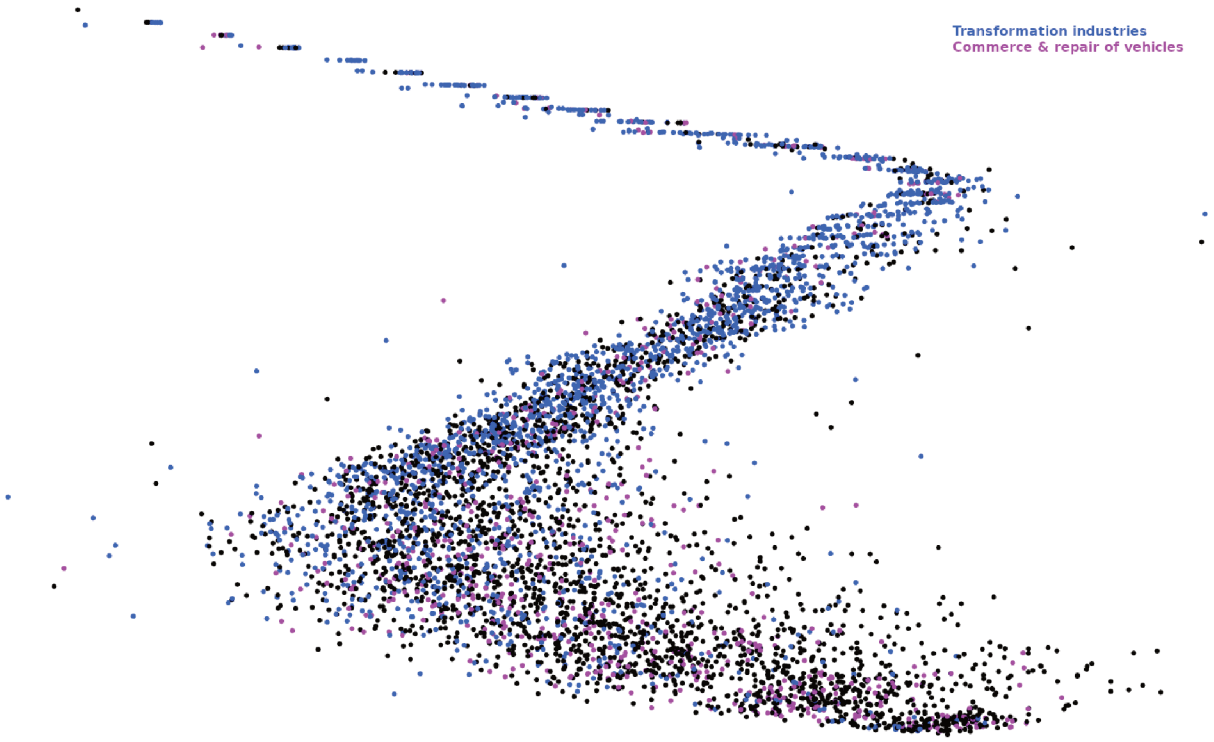


Figure 30 – Scatterplot showing the two major economic activities causing LAs. "transformation industries" are the majority in most cities, but "commerce and repair of vehicles" also represents the majority of LAs in a significant number of cities.

of which region they belong to, fulfilling requirement *r3*.

Figure 32 shows how each visualization can complement each other to provide a deeper analysis. When coloring the scatterplot according to causer agent, it can be noticed that "machines and equipment" is the causer agent with most LACs for the majority of the cities(red points). The parallel sets help to shed more light on this observation by showing that 4 of the 5 regions have "machines and equipment" as causer agent with the largest number of LACs registered. In this figure the categories(ribbons) are sorted from largest to smallest and one can see the greater width of the first ribbon, which represents accidents with "machines and equipment". The exception is the Norte Region, even though this causer agent is also top listed there.

Once again using the complementary capacities of both visualizations, we found out that upper members are the more common body part affected by LAs, followed by lower members. The scatterplot in Figure 33 suggests that upper member lesions are prevalent for the majority of localities, the exception being rural areas. This can be noticed by analyzing the distribution of the points, the purple points are well distributed over the entire layout, and clearly dominate the majority of the layout, while the pink points are largely constrained to the bottom of the layout, where rural locations are mostly located. Figure 34 presents a portion of a parallel sets that shows the relationship between place

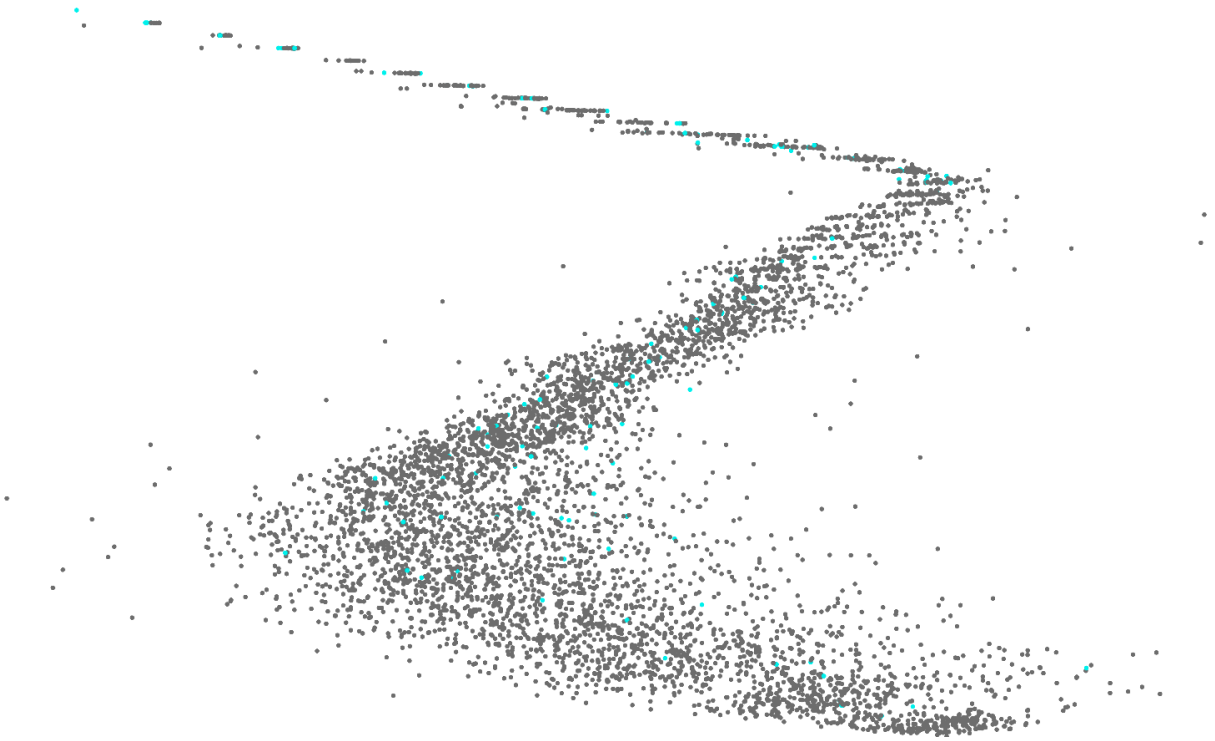
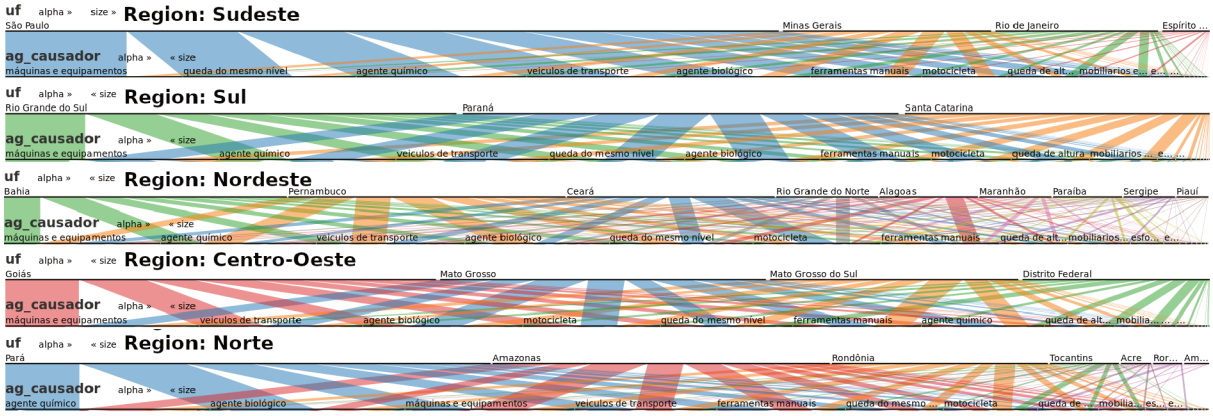


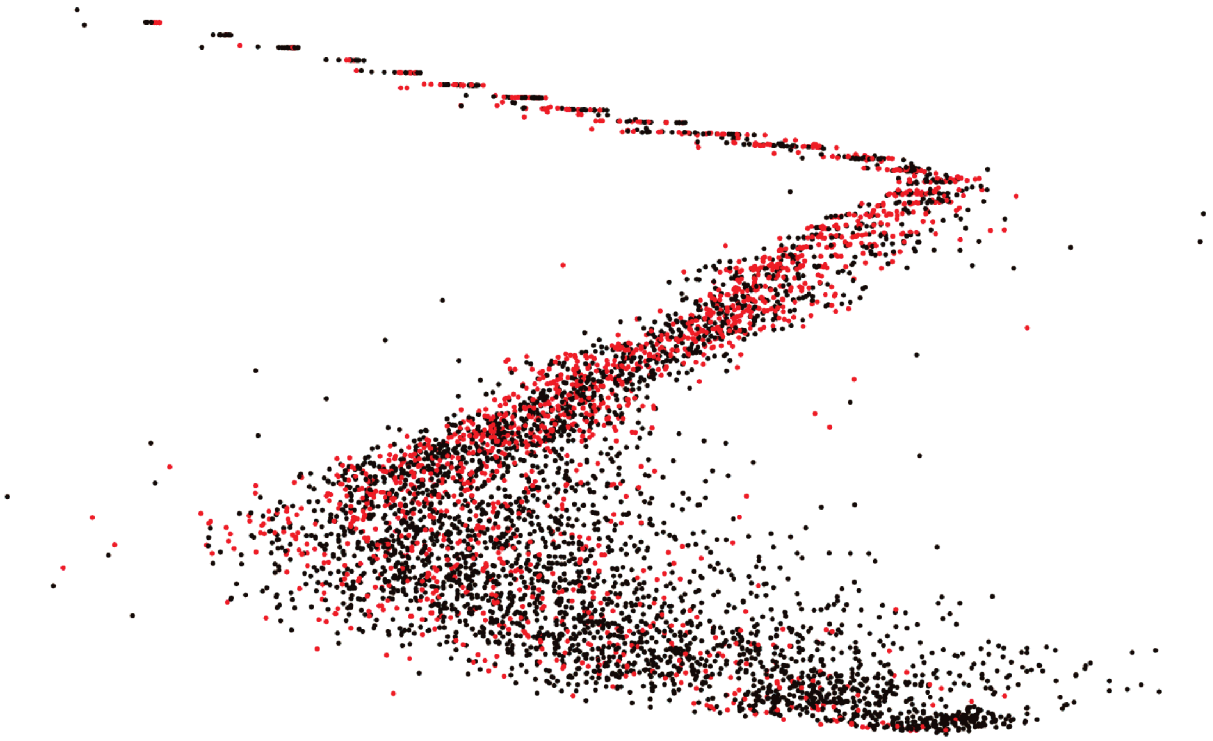
Figure 31 – Scatterplot showing the distribution of cities with the majority of LAs in "human health and social services", which represents few but important cities.

of occurrence of a LA and the resulting lesions. Analyzing this parallel sets, one can notice that the majority of LAs occur in the employment location (blue ribbons) and these LAs result in a greater number of lesions to upper members than to lower members. However, for rural areas (green ribbons) this distribution is more balanced, supporting the findings in the scatterplot. By hovering over these ribbons a tooltip shows the percentage these LAs represent from the total. LAs occurring in the employment location resulting in lesions to upper members are 33.74% and resulting in lesions to lower members are 11.69%. For LAs happening in rural areas, lesions to upper members are 1.6% and lesions to lower members are 1.19%. This example also shows how the interactive tools help the analysis with the parallel sets, in this case the tooltip. The parallelograms formed by the ribbons are sometimes misleading when one wants to compare size, but the tooltip shows clearly the amount each ribbon represents. This parallel sets depicted also helped us to identify that accidents happening on public ways (red ribbons), also have a more balanced distribution of LAs resulting in lesions to upper members (8.57%) and lesions to lower members (10.56%), which we found difficult to identify by looking only at the scatterplot. This type of finding can help the government to prevent specific types of accidents by showing in which place a type of accident is more likely to occur. All the analysis up to this point fulfill requirement *r1*.

Identifying areas that may be lacking attention can be readily done when one combines



(A) Causer agents by region.



(B) Machines and equipment causer agent highlight.

Figure 32 – Analysis of causer agents, using parallel sets (A) and the scatterplot (B).The Parallel sets show that all regions except one report "machines and equipment" as main causer of accidents, while the scatterplot shows that it is the main LA causer in the majority of cities.

the findings of both visualizations. For example, in Figure 35A the scatterplot shows that a few cities have the majority of LAs reported by syndicates, while the law mandates that LAs must be registered by the employer. Those cities are all located in the Nordeste region, shown in Figure 35B. Analyzing this category in the parallel sets one can confirm that the Nordeste region is the only one with a noticeable number of accidents reported by syndicates, around 1.6%. This analysis fulfills requirement *r2*.

In the scatterplot it is possible to notice some isolated cities. The majority of these

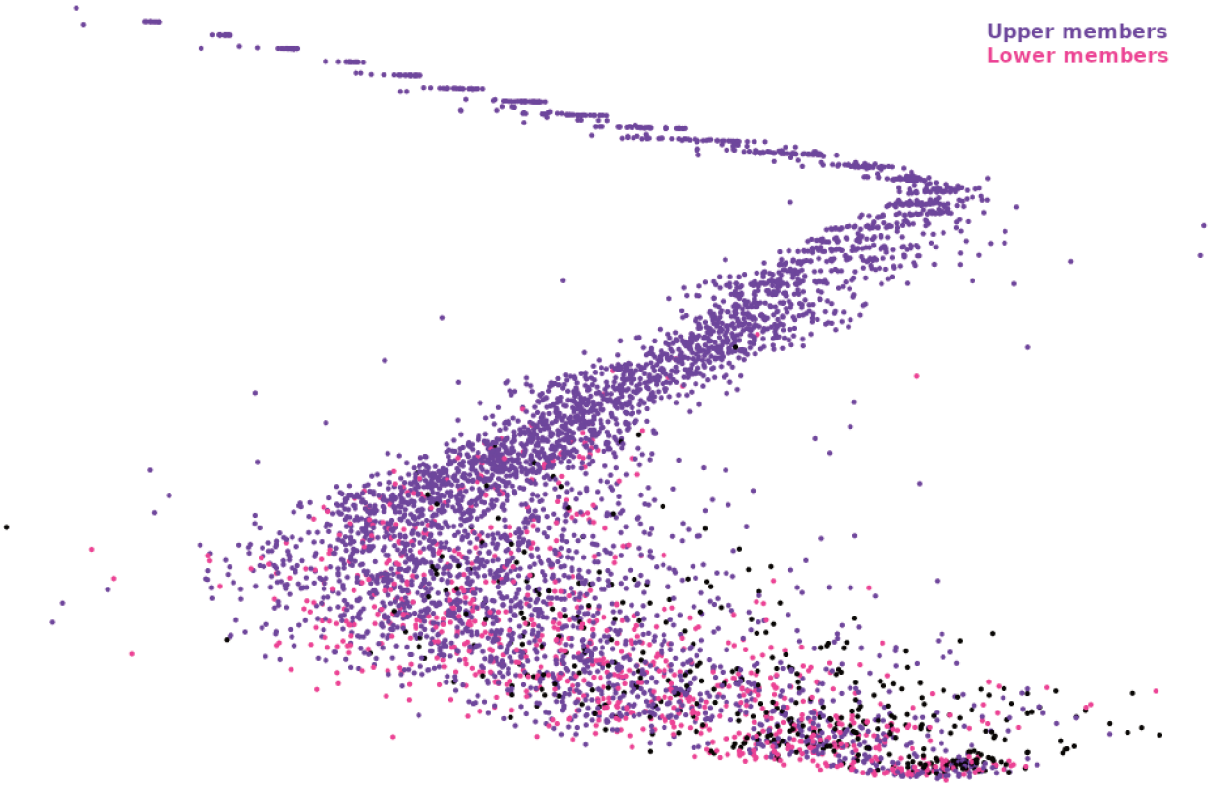
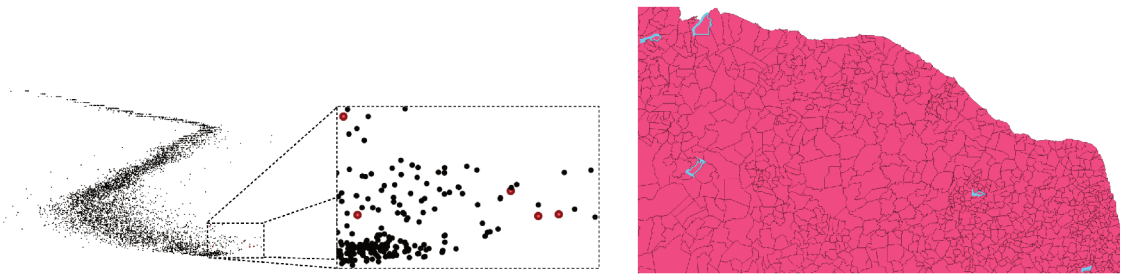


Figure 33 – Scatterplot presenting the most commonly injured body parts. Lesions to upper members are more prevalent in the majority of cities, but lesions to lower members are more common in rural areas.

Figure 34 – Distribution of lesion location according to LA's place of occurrence.

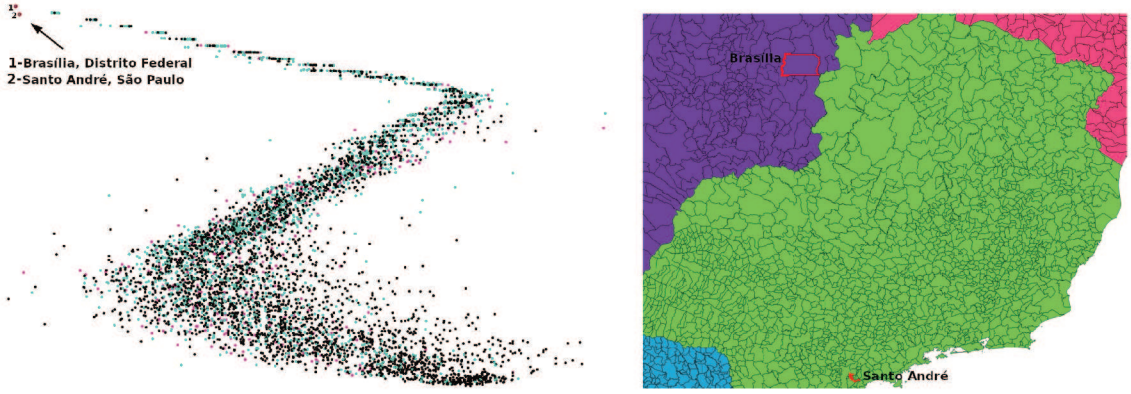


(A) Scatterplot zoom. (B) Geographical location.

Figure 35 – Analysis of cities reporting syndicates as main LAC issuer, showing that they are all located in the Nordeste region.

cities are small, with few accidents registered, 1 or 2 in most cases. Although we were not able to draw any conclusions about the positioning of these cities, it might be interesting for domain experts to investigate them, in the sense of verifying the reasons for the few

accidents registered. One interesting situation involving two isolated cities however is shown in Figure 36. These cities are Brasília, the capital of Brazil, and Santo André, a city in the metropolitan region of São Paulo. Figure 36A shows the positioning of both cities in the scatterplot, and Figure 36B shows their geographical positions in the political map. It can be seen how those cities are far apart geographically, yet have similar behavior, as suggested by their LAMP placement.



(A) Brasília and Santo André's location in the scatterplot.

(B) Brasília and Santo André's geographical location.

Figure 36 – Analysis of Brasilia and Santo André using the projection + political map visualization. They present a similar behavior, despite their geographical distance.

The parallel sets for both cities presented in Figure 37 allows one to explore each city individually and shows in details how they have similar behavior, while are different from the remaining cities in the regions they belong to (Centro-Oeste and Sudeste, respectively), in terms of LA occurrence.

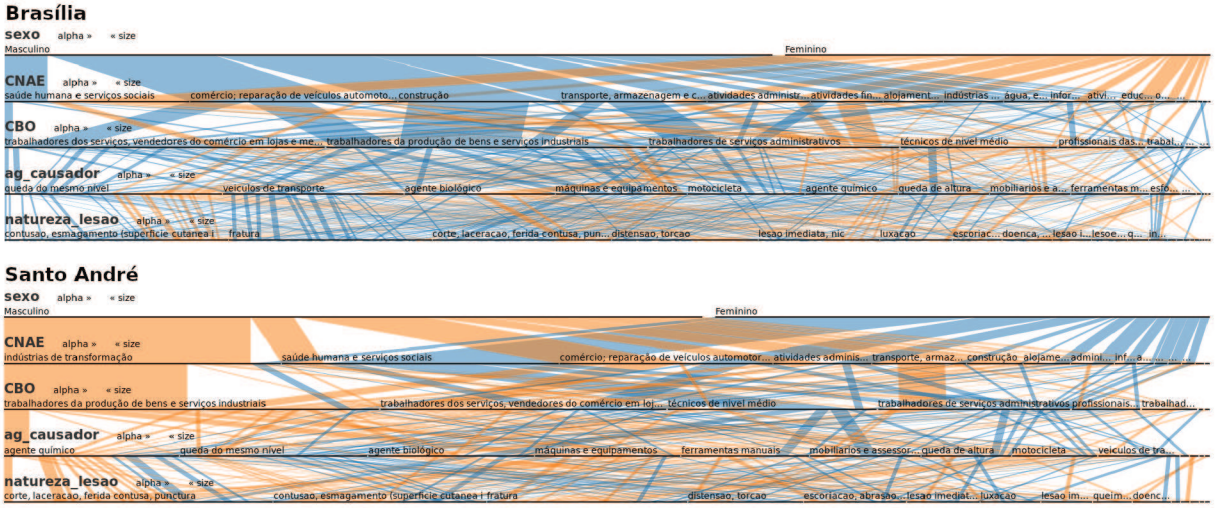


Figure 37 – Parallel sets of Brasília and Santo André showing similar LA profiles.

Brasília has a more balanced distribution of man and women accidents in the economic activities, suggesting a better work force distribution, with 62% of LACs related to male LAs, which is 10% less than the region average(72%). The economic activity with more registered accidents is "human health and social services", as already seen in the scatterplot, but the parallel sets allows us to perceive other economic activities with a meaningful number of LAs registered. Still related to economic activity, "transformation industries" is only responsible for 6% of accidents in Brasília, distant from the rest of the region. Regarding occupations, "Arts and Science", and "Human Health" have more women accidents, and "production of goods and services" is more associated with accidents with man. "Service providers and sellers" is the occupation with the largest number of LACs. Brasília, being the capital of Brazil, is an administrative center, concentrating the highest number of public agencies, and has a particular focus on providing services. These characteristics seem to reflect well in its economic activities and occupations registering most of the LACs. Regarding causer agent, the one with most occurrences is "fall from the same level", followed by "transportation vehicles". In respect to injury types, the most common in Brasília is "contusion", while in the region is "fracture".

As expected from the scatterplot, Santo André displays a similar LA profile to Brasília. The distribution of accidents between men and women is also more balanced, 59% are men, which is 6% less than the region average. "human health and social services" is also an economic acitivity responsible for a significant number of LACs (23%), even though the top listed here is "transformation industries" (27%). The occupation profile is also similar to Brasília. Women LAs are more prevalent in occupations in "Arts and Science" and in "Health". Regarding causer agent, the one with most LACs in Santo André is "chemical agent", closely followed by "fall from the same level", while the other top listed causers are the same from Brasília. This analysis of both cities fulfill requirement *r3*.

The parallel sets + treemap visualization is effective in characterizing entire regions as well, Sul and Sudeste regions have similar behavior, as can be seen in Figure 38. These are the regions with the most balanced proportion of male to female LAs, 65% are males. The economic activities responsible for the greater number of LACs are: "transformation industries", "commerce and repair of vehicles", and "human health and social services". This similar behavior is extended to the other attributes, and the only difference is the distribution of LAs in the states. Figure 39 presents the treemaps of both these regions, showing the total number of LACs for each state in them. It is possible to notice that the Sul region is more homogeneous than Sudeste, with its three states registering similar numbers of LACs, while in the Sudeste region the state of São Paulo is responsible for 60% of all LACs in the region, and Espírito Santo is responsible for only 4%.

Figure 40 shows the parallel sets for the Nordeste and Centro-Oeste regions. A more significant number of accidents in the rural area can be perceived in these regions, around 8%. Additionally the percentage of accidents with men is greater than 70%. Similar

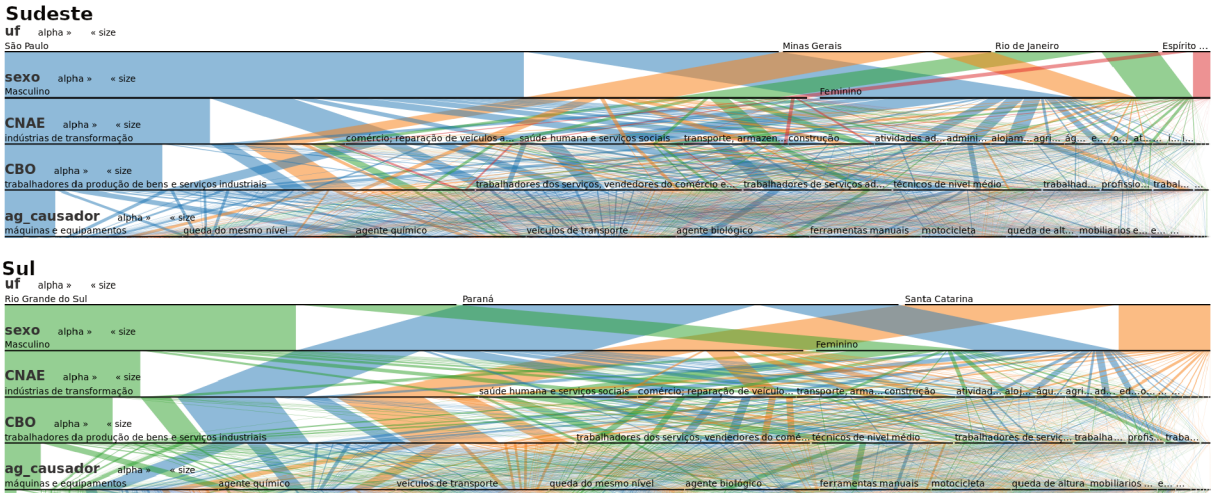


Figure 38 – Parallel sets of the Sudeste and Sul regions showing their similar behavior.

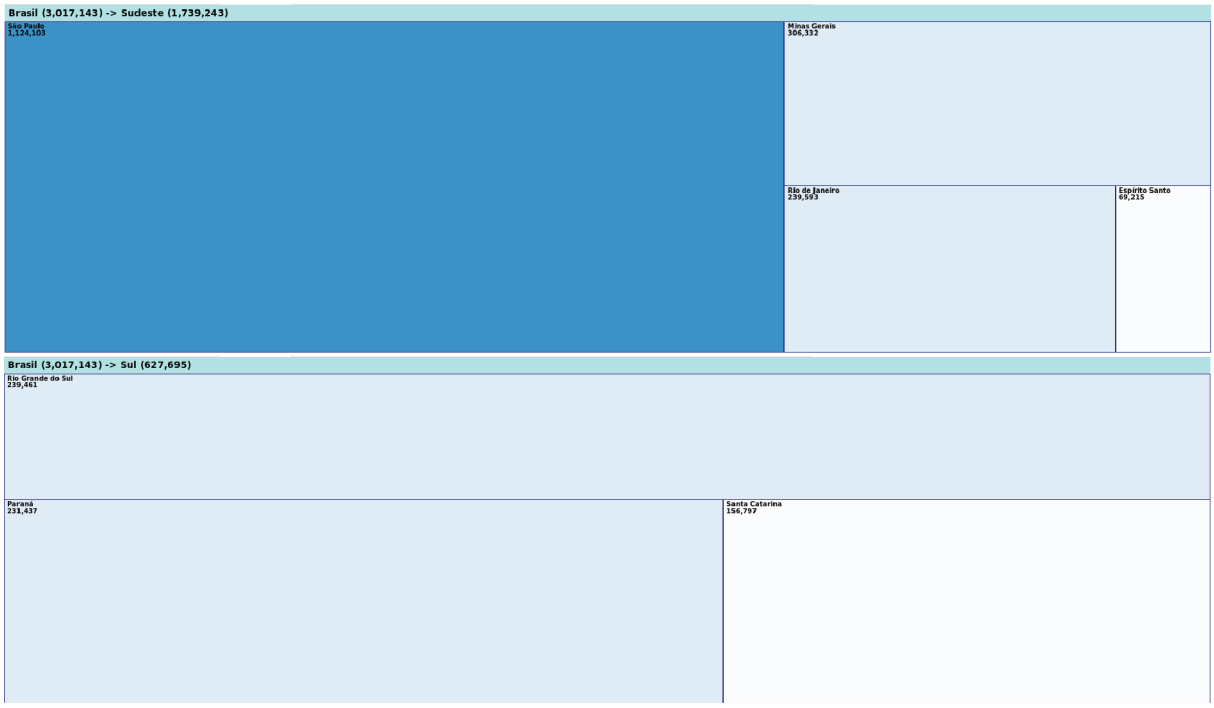


Figure 39 – Treemaps of the Sudeste and Sul regions showing that the distribution of accidents is much more homogeneous in the Sul region than in the Sudeste region.

causer agents are also present in both regions, predominantly “machines and equipment”, “chemical agent”, “biological agent”, and “vehicles”.

Markedly different from the others is the Norte region, as show in Figure 41. Male accidents represent more than 80% of all LACs. The LACs profile related to economic activities is totally different from the others. Even though “transformation industries” is still the activity with most LACs, it is not so prevalent as in other regions, the second

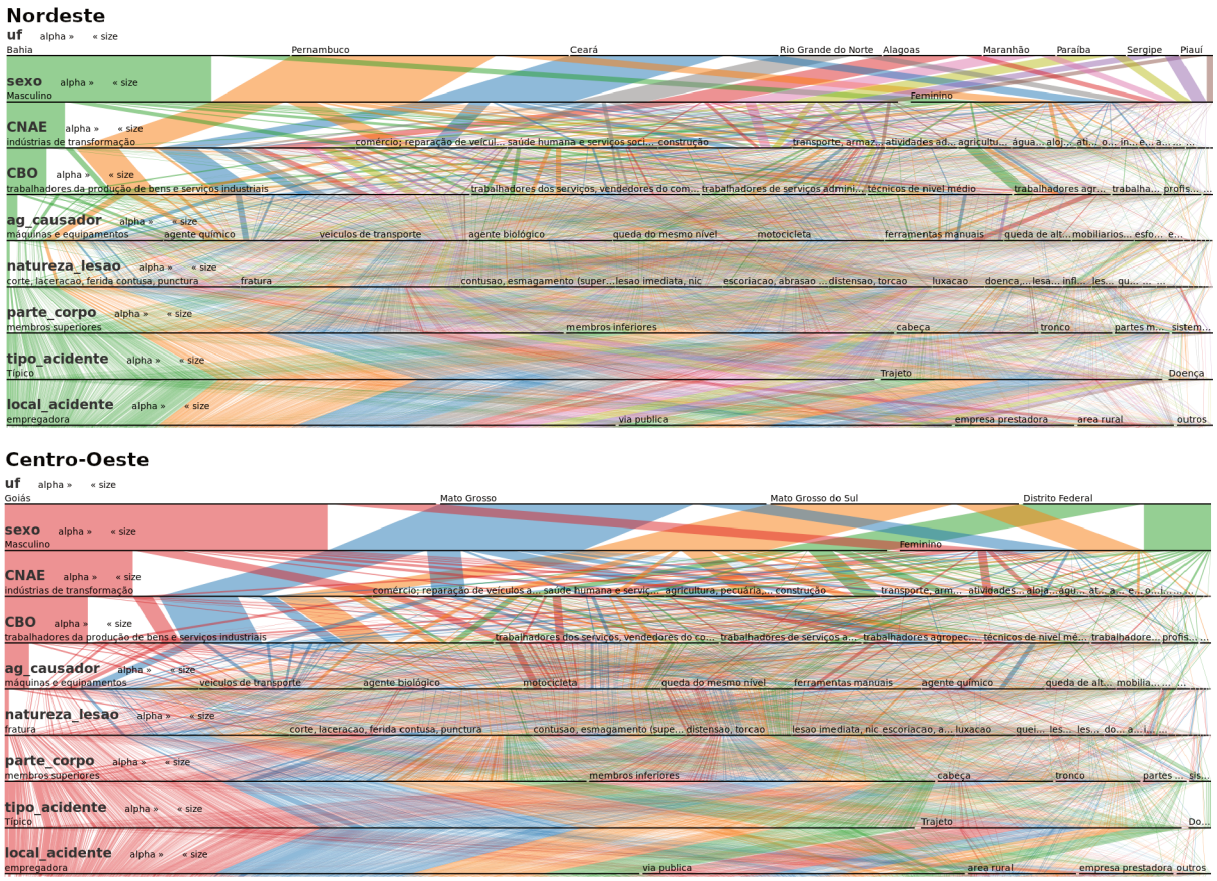


Figure 40 – Parallel sets of the Nordeste and Centro-Oeste regions, in which some similarities can be noticed, such as the number of accidents in the rural area.

activity with most LACs, “construction”, is not so high in other regions. LACs related to “human health and social services”, which is top listed in the rest of the country, are only marginal here. The most reported causer agent is “chemical agent”, different from the remaining regions, in which the most reported causer agent is “machines and equipment”. This behavior seems to reflect the poor development of the region, the lowest in Brazil, so we decided to performe a deeper investigation. Analyzing the states of this region its low industrialization is noticiable. The economic activity of “transformation industries” only has a great number of LACs because of the two principal states in the region, Pará and Amazonas, which are responsible for 70% of all accidents in the region. Rondônia, the third state with more LACs, has twice the number of reported accidents in “construction” than in “transformation industries”. In all the other states, “transformation industries” percentages are negligible and LACs are mainly associated with rural activities, such as fishing and agriculture.

A more detailed analysis of Pará and Amazonas, depicted in Figure 42, revealed that the LAC numbers associated with Industry are reported mostly in their capital cities, Belém and Manaus respectively. The rest of the localities in these states present profiles

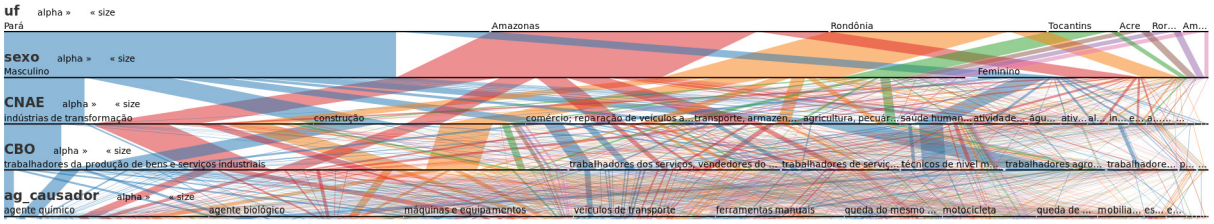


Figure 41 – Parallel sets of the Norte region. A larger proportion of male accidents can be noticed, and also the different LA profile from the other regions.

similar to the rest of the region, with two mesoregions in Pará, Nordeste Paraense and Sudeste Paraense, strongly characterized by accidents related to pecuary and agriculture. This analysis shows that these capitals are outliers in the region, which can be confirmed by their placement in the scatterplot, at the top of the "Z", close to cities such as São Paulo, Rio de Janeiro, and Goiânia. These analysis fullfill requirement r4.

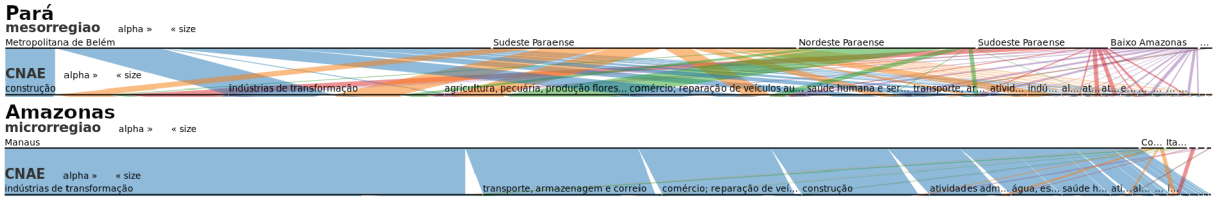


Figure 42 – Parallel sets of the states of Pará and Amazonas. The Capital cities dominate their respective states, and present the only high numbers of LACs in Industry of the region.

5.4 Final Considerations

This chapter presented the results of detailed analysis performed using the system described in Chapter 4. The analysis results demonstrated that the strategy fulfilled all the requirements described in Chapter 3. We were able to identify patterns in local and also higher levels of the Brazilian territorial division, relate cities in different geographical localities, characterize work profiles, identify cities lacking attention, among other findings. Also, the analysis showed how both visualizations efficiently complement each other, shedding more light on the individual findings, but are also effective when used independently.

Chapter 6

Conclusion

In this research work we presented a strategy to perform visual analysis of LAs data, aiming to explore the underlying structural characteristics of the BFLPO dataset, including its heterogeneity and associated hierarchical organization. A system implementing this strategy was developed, as a mean to validate this strategy.

The chosen layouts were able to communicate the underlying data structure, and the interactive tools provided an effective exploration. By using our proposed strategy, we were able to identify profiles associated with individual cities and large geographical areas, as well as behavior patterns associated with the whole country. Combining the observed patterns in both visualizations we were able to find similarities among geographically distant localities, occurrence patterns related to the cities' size and economic developments, the frequency distribution of LA types in Brazil, and to characterize those LAs in terms of occupation type, gender differences, causer agent, among other aspects. Moreover, the visualizations' complementary capacities provided simultaneous analysis of different data aspects. As an example, the scatterplot depicts only cities, and the insights over higher hierarchical levels are gathered by the shape formed by the points' position and the colors, while the treemap and parallel sets visualization allows the characterization of larger geographical areas, in all hierarchical levels. We believe that several of our findings would be difficult to uncover using solely tabular data, or common statistical graphs. Among these finds we identified relationships among cities that would not be easy to identify if each city was analyzed only considering cities located in the same or close localities, and patterns associated with entire regions.

The system can contribute to policy making, because it shows clearly important strategical information, such as major accident causers, localities lacking proper monitoring, odd behaviors, among others. This could potentially aid the government in evaluating if an ongoing strategy is being effective, or even if a new proposed policy would be adequate to the situation. Additionally, the developed system also makes available data more transparent, by easily communicating the LA occurrences situation in the country, enabling a more effective comprehension of their occurrence and providing a simple but effective

communication channel between citizen and government. This could foment a more active participation of the population in matters regarding labor issues, allowing the population to better monitor government policies and demand appropriate solutions.

The strategy and correspondent system, as well as the analysis presented in this thesis focused on LA data provided by the BFLPO. However, we believe this strategy can readily be used to analyze other datasets that present similar characteristics, that is, datasets that present hierarchical and geographical aspects, as well as a massive number of categorical measured attributes. A large number of governmental datasets present these characteristics, and we thus believe it has a large applicability in several strategic analysis scenarios.

6.1 Limitations

During the development of this research work, we were able to identify two types of limitations: data related and strategy related.

Regarding the data, the number of LACs is small, when compared to the number of LAs that really occur, and this is specially noticeable in poorer regions, like Norte. This underrepresentation is a serious problem and potentially impairs a proper analysis, making difficult to explain a specific behavior, as well as to trust in some revealed patterns. Furthermore, some LACs are not properly reported, resulting in loss of information that may be important for the analysis. We partially address this problem, by shedding light on anomalous behavior and by identifying areas that may lack attention. We expect that this will aid the government in identifying deficiencies on the LA reporting process, in order to promote a more strict inspection.

Regarding our strategy and implemented system, the parallelograms in the parallel sets, when individually evaluated, may sometimes mislead the analysis, because they tend to get distorted when the path from a category to the next one generates sharp angles. Adding some visual cue to the axis, such as ticks, could improve this situation. We currently address this issue by using the tooltip information.

The treemap, although useful to navigate the parallel sets and to add more context to the analysis, does not allow the comparison of localities from the same level, but in different branches, such as mesoregions from different states. This may make such analysis slower and more prone to errors.

The BFLPO dataset is massively categorical, and in order to employ the point-placement techniques, it was necessary to transform this data into numerical values. Thus, our evaluations became strongly dependent on these transformation procedures, which may result in information loss. However, we believe that the layouts produced from the transformed data already produced satisfactory results in terms of revealing useful patterns for analysis.

6.2 Future Work

After we performed the analysis described in Chapter 5, several interesting research directions were found. This research work explored the structural aspect of LAs data, however another interesting aspect of the data is the temporal one. Combining these two aspects in a single analysis strategy could potentially reveal more patterns and improve analysis as a whole. A temporal visual LA analysis tool has already been developed in(BRITO; RODRIGUES; PAIVA, 2019), and we intend to combine the proposed techniques with our approaches, to perform such analysis. Additionally, there are many small improvements that can be made to the developed system, including new interactions, and better communication between the layouts/visualizations. Some examples of those include improving the scatterplot by showing more than one attribute at once, for instance using different shapes for points, transparency, among others. New multidimensional projections could be made based on user selections. The treemap could show different values besides number of accidents. Finally, we intend to perform qualitative experiments with specialists from the BFLPO, which are domain experts, and may significantly benefit from the information provided by our tool. This experiment could lead to further improvements in our strategy and positively impact the decision making process.

Bibliography

ALOWIBDI, J. S.; GHANI, S.; MOKBEL, M. F. Vacationfinder: A tool for collecting, analyzing, and visualizing geotagged twitter data to find top vacation spots. In: **Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. New York, NY, USA: ACM, 2014. (LBSN '14), p. 9–12. ISBN 978-1-4503-3140-1. DOI 10.1145/2755492.2755495.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3: data-driven documents. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2301–2309, Dec 2011. ISSN 1077-2626. DOI 10.1109/TVCG.2011.185.

BRITO, L.; RODRIGUES, M.; PAIVA, J. G. de S. A computational system for temporal visual analysis of labour accident data. In: **2019 23rd International Conference Information Visualisation (IV)**. [S.l.: s.n.], 2019. p. 88–93. ISSN 2375-0138. DOI 10.1109/iv.2019.00024.

CHANG, W. et al. **shiny: Web Application Framework for R**. [S.l.], 2017. R package version 1.0.5.

COX, T.; COX, M. **Multidimensional Scaling**. Boca Raton: Chapman & Hall/CRC, 2000. (Monographs on statistics and applied probability 88). DOI 10.1201/9781420036121. ISBN 1584880945.

DATAVIVA. **DataViva**. 2013. Http://dataviva.info.

DATAWHEEL. **DataUSA**. 2016. Https://datausa.io/.

_____. **DataChile**. 2018. Https://es.datachile.io/.

DíAZ, P.; AEDO, I.; HERRANZ, S. Understanding citizen participation in crisis and disasters: The point of view of governmental agencies. In: **Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces**. New York, NY, USA: ACM, 2014. (AVI '14), p. 395–397. ISBN 978-1-4503-2775-6. DOI 10.1145/2598153.2602227.

FATORE, F. M.; FADEL, S. G. **mp: Multidimensional Projection Techniques**. [S.l.], 2016. R package version 0.4.1.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, n. 2, p. 179–188, 1936. DOI 10.1111/j.1469-1809.1936.tb02137.x.

GRAVES, A.; HENDLER, J. Visualization tools for open government data. In: **Proceedings of the 14th Annual International Conference on Digital Government Research**. New York, NY, USA: ACM, 2013. (dg.o '13), p. 136–145. ISBN 978-1-4503-2057-3. DOI 10.1145/2479724.2479746.

GROEGER, L.; GRABELL, M.; COTTS, C. **Workers' Comp Benefits: How Much is a Limb Worth?** 2015. Https://projects.propublica.org/graphics/workers-compensation-benefits-by-limb.

HEER, J.; BOSTOCK, M.; OGIEVETSKY, V. A tour through the visualization zoo. **Communications of the ACM**, ACM, New York, NY, USA, v. 53, n. 6, p. 59–67, jun. 2010. ISSN 0001-0782. DOI 10.1145/1743546.1743567.

HOXHA, J.; BRAHAJ, A.; VRANDEčIć, D. Open.data.al: Increasing the utilization of government data in albania. In: **Proceedings of the 7th International Conference on Semantic Systems**. New York, NY, USA: ACM, 2011. (I-Semantics '11), p. 237–240. ISBN 978-1-4503-0621-8. DOI 10.1145/2063518.2063558.

ILO. **Safety and Health at work**. 2019. Http://www.ilo.org/global/topics/safety-and-health-at-work/lang–en/index.htm.

INSELBERG, A. The plane with parallel coordinates. **The Visual Computer**, v. 1, n. 2, p. 69–91, Aug 1985. ISSN 1432-2315. DOI 10.1007/BF01898350.

INSS. **Comunicação de Acidente de Trabalho – CAT**. 2018. Https://www.inss.gov.br/servicos-do-inss/comunicacao-de-acidente-de-trabalho-cat/.

International Food Policy Research Institute (IFPRI); DATAWHEEL. **Data Africa**. 2017. Https://DataAfrica.io.

JEONG, D. H. et al. Understanding principal component analysis using a visual analytics tool. **Charlotte visualization center, UNC Charlotte**, v. 19, 2009.

JOIA, P. et al. Local affine multidimensional projection. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, p. 2563–2571, 01 2011. DOI 10.1109/tvcg.2007.70443.

KOSARA, R.; BENDIX, F.; HAUSER, H. Parallel sets: interactive exploration and visual analysis of categorical data. **IEEE Transactions on Visualization and Computer Graphics**, v. 12, n. 4, p. 558–568, July 2006. ISSN 1077-2626. DOI 10.1109/TVCG.2006.76.

LEI, S. T.; ZHANG, K. A visual analytics system for financial time-series data. In: **Proceedings of the 3rd International Symposium on Visual Information Communication**. New York, NY, USA: ACM, 2010. (VINCI '10), p. 20:1–20:9. ISBN 978-1-4503-0436-8. DOI 10.1145/1865841.1865868.

LIMA, C. M.; PAIVA, J. G. Análise de dados do dataviva utilizando técnicas de projeção multidimensional. In: **Proceedings of the 13th Simpósio Brasileiro de Sistemas de Informação**. [S.l.: s.n.], 2017.

LIU, S. et al. A survey on information visualization: recent advances and challenges. **The Visual Computer**, v. 30, n. 12, p. 1373–1393, Dec 2014. ISSN 1432-2315. DOI 10.1007/s00371-013-0892-3.

LUIZ, E. R. **Geographic Information Systems (GIS) - Brasil Coleção de shapefiles, GeoJSON e TopoJSON prontas para uso**. 2014. Http://fititnt.github.io/gis-dataset-brasil/.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, Nov 2008. ISSN 1532-4435.

MENDONçA, P. G. A. de; MACIEL, C.; FILHO, J. V. Visualizing aedes aegypti infestation in urban areas: A case study on open government data mashups. In: **Proceedings of the 15th Annual International Conference on Digital Government Research**. New York, NY, USA: ACM, 2014. (dg.o '14), p. 186–191. ISBN 978-1-4503-2901-9. DOI 10.1145/2612733.2612751.

MUNZNER, T. **Visualization Analysis and Design**. Boca Raton: A K Peters/CRC Press, 2014. (AK Peters Visualization Series). DOI 10.1201/b17511. ISBN 978-1-4665-0893-4.

NIELSEN, M.; GRøNBæK, K. Pivotviz: Interactive visual analysis of multidimensional library transaction data. In: **Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries**. New York, NY, USA: ACM, 2015. (JCDL '15), p. 139–142. ISBN 978-1-4503-3594-2. DOI 10.1145/2756406.2756937.

NONATO, L. G.; AUPETIT, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. **IEEE Transactions on Visualization and Computer Graphics**, p. 1–1, 2018. ISSN 1077-2626. DOI 10.1109/TVCG.2018.2846735.

OECD. **Open Government Data**. 2019. Http://www.oecd.org/gov/digital-government/open-government-data.htm.

PAULOVICH, F. V.; MINGHIM, R. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1229–1236, Nov 2008. DOI 10.1109/TVCG.2008.138.

PAULOVICH, F. V. et al. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, May 2008. ISSN 1077-2626. DOI 10.1109/tvcg.2007.70443.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018.

RADL, W. et al. And data for all: On the validity and usefulness of open government data. In: **Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies**. New York, NY, USA: ACM, 2013. (i-Know '13), p. 29:1–29:4. ISBN 978-1-4503-2300-0. DOI 10.1145/2494188.2494228.

RODRIGUES, N. et al. Visualization of time series data with spatial context: Communicating the energy production of power plants. In: **Proceedings of the 10th International Symposium on Visual Information Communication and Interaction**. New York, NY, USA: ACM, 2017. (VINCI '17), p. 37–44. ISBN 978-1-4503-5292-5. DOI 10.1145/3105971.3105982.

SHARMIN, M. et al. Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In: **Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing**. New York, NY, USA: ACM, 2015. (UbiComp '15), p. 505–516. ISBN 978-1-4503-3574-4. DOI 10.1145/2750858.2807537.

SHNEIDERMAN, B. Tree visualization with tree-maps: 2-d space-filling approach. **ACM Transactions Graphics**, ACM, New York, NY, USA, v. 11, n. 1, p. 92–99, jan. 1992. ISSN 0730-0301. DOI 10.1145/102377.115768.

Smartlab de Trabalho Decente MPT - OIT. **Observatório Digital de Saúde e Segurança no Trabalho**. 2017. Http://observatoriosst.mpt.mp.br.

TUOR, R.; EVéQUOZ, F.; LALANNE, D. Parallel bubbles: Categorical data visualization in parallel coordinates. In: **Actes De La 28Ième Conférence Francophone Sur L'Interaction Homme-Machine**. New York, NY, USA: ACM, 2016. (IHM '16), p. 299–306. ISBN 978-1-4503-4243-8. DOI 10.1145/3004107.3004142.

WICKHAM, H. **tidyverse: Easily Install and Load the 'Tidyverse'**. [S.l.], 2017. R package version 1.2.1.

WITTENBURG, K.; TURCHI, T. Treemaps and the visual comparison of hierarchical multi-attribute data. In: **Proceedings of the International Working Conference on Advanced Visual Interfaces**. New York, NY, USA: ACM, 2016. (AVI '16), p. 64–67. ISBN 978-1-4503-4131-8. DOI 10.1145/2909132.2909286.

YING, Y.; XIALING, T.; WEI, T. Study on governmental cultural resources purchase management based on public information behavior big data. In: **Proceedings of the 8th International Conference on E-Education, E-Business, E-Management and E-Learning**. New York, NY, USA: ACM, 2017. (IC4E '17), p. 72–75. ISBN 978-1-4503-4821-8. DOI 10.1145/3026480.3026491.

ZHIYUAN, H. et al. Application of big data visualization in passenger flow analysis of shanghai metro network. In: **2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)**. [S.l.: s.n.], 2017. p. 184–188. DOI 10.1109/ICITE.2017.8056905.

Appendix

APPENDIX A

Reproducing this work

The source code to generate the visualizations and the needed datasets are available as a git repository on https://notabug.org/mprodrigues/app-mpt Any updated instructions on how to reproduce this research can also be found there.

Following are described the sources of the primary datasets used and how to process them for the visualizations.

A.1 BFLPO dataset

This data was available from https://observatoriosst.mpt.mp.br/ in the tab "Sobre", in the lateral menu "Conjunto de Dados". However, at the time of writing it was inaccessible.

A.2 IBGE data

The data obtained from IBGE is the territorial division of Brazil and the population estimates for each city for the years of 2012 to 2017.

The territorial division data can be downloaded from http://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/divisao_territorial. The file used was “DTB_BRASIL_MUNICIPIO.ods” for 2016 year, however since then updates for the years of 2017 and 2018 are available. From this data set were removed the columns with area codes and a column with the region of each state was added manually. This data was converted to csv and is available in the source repository.

The population estimates can be downloaded from https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de-populacao.html. Only the "Município" e "População Estimada" columns were used.

A.3 GeoJSON files

The GeoJSON files used to generated the maps are available in https://github.com/fititnt/gis-dataset-brasil.

A.4 Category reductions

We used a mix of official/non-official sources to reduce the number of categories in some attributes, and in the case where such did not exist we used our own common sense.

All of the files containing this categories are already available in the repository, following are listed where we got them.

- Causer agent classification was obtained from http://suporte.quarta.com.br/LayOuts/eSocial/Tabelas/Tabela_14.htm and https://docplayer.com.br/132223-Tabela-3-agente-causador-da-doenca-profissional-ou-do-trabalho-descricao-da-situacao-geradora-da-doenca.html.;
- Economic Activity classification was obtained from https://cnae.ibge.gov.br/documentacao/documentacao-cnae-2-0.html;
- Occupation classification was ceded by a BFLPO specialist;
- The parts of the body injured were arbitrarily grouped into head, trunk, upper members, lower members, and systems.

A.5 Regenerating the data used in the visualizations

Simply run the script "summary.R" in the folder "scripts". This will put all necessary data in the folder “data”.