UM MÉTODO EFETIVO PARA DETECÇÃO DE ANOMALIAS EM SISTEMAS DE SAÚDE PÚBLICA

LUIZ FERNANDO MAGALHÃES CARVALHO

UM MÉTODO EFETIVO PARA DETECÇÃO DE ANOMALIAS EM SISTEMAS DE SAÚDE PÚBLICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Wagner Meira Júnior

Belo Horizonte

Dezembro de 2015

LUIZ FERNANDO MAGALHÃES CARVALHO

AN EFFECTIVE METHOD FOR ANOMALY DETECTION IN PUBLIC HEALTHCARE SYSTEMS

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: WAGNER MEIRA JÚNIOR

Belo Horizonte

December 2015

Carvalho, Luiz Fernando Magalhães

C331e An effective method for anomaly detection in public healthcare systems / Luiz Fernando Magalhães Carvalho. — Belo Horizonte, 2015
xxvii, 119 f. : il. ; 29cm

Dissertação (mestrado) — Federal University of Minas Gerais
Orientador: Wagner Meira Júnior

1. Anomaly detection. 2. Data mining. 3. Healthcare. I. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

An effective method for anomaly detection in public healthcare systems

LUIZ FERNANDO MAGALHAES CARVALHO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO CÉSAR MACHADO PEREIRA
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 22 de dezembro de 2015.

I dedicate this work to my family and friends.

Acknowledgments

First of all, I thank my Family. My mother Regina and my father Luiz for the love, the example and all the support during my life. My sisters Gabriela and Raquel for the affection and fellowship. My grandparents, uncles and cousins for being lovely present in my life. Abby and Karina, for the true love. I am deeply thankful for all my friends, especially Thome, Vitor, Bernardo, Saulo, Neto and Daniel.

I could not have done this work without all the help that I received from my colleagues from UFMG, especially my friends Carlos Carvalho and Douglas Eduardo. Carlos has been helping me a lot since the end of my graduation. Thanks for being a good friend and a patient partner during all this time. Douglas was my first advisor and a key responsible for the way my life has followed so far. Thank you for the invite to research with you.

I am also thankful for the fellowship from everyone who has been with me at UFMG since 2008: colleagues from the graduation, Speed Lab, LBS, Infosas project and soccer games, especially my friends Felipe Ferre, Luis Gustavo, Pedro Calais, Roberto Carlos, Elverton, Osvaldo and Douglas Teixeira. It was a honor to work and meet you almost every day. I also appreciate the staff of the Computer Science Department, especially Sonia for the attention and competence.

I would like to thank those who have been great leaders for me helping me a lot in my research. First, I appreciate the patience and wisdom of professor Alberto Laender who were a great example for me during the time that we worked together. Professor Osvaldo Carvalho for being a nice and competent coordinator during the Infosas project. My godfather George Jamil for the fellowship, example and inspiration. Professor Martin Ester for gently receiving me in Simon Fraser University and giving me all the attention and patience that I could receive for developing my master in Canada. For sure it was one of the best periods of my life, I have learned a lot with you. I will never forget the Hiking day and all the funny moments that I had there. And finally, my advisor Wagner Meira, who has been a huge inspiration and example for me in the past 5 years. Thank you for all the help, support, advices and wisdom.

"Everything that a man ignores does not exist for him. Therefore, the universe of each one is summed up to the size of his knowledge."
(Albert Einstein)

Resumo

Detecção de anomalias é uma tarefa relevante e amplamente aplicada em diferentes cenários. Com os avanços em gerenciamento de saúde e tecnologia da informação, a detecção de anomalias em saúde se consolidou como um importante tópico na comunidade científica. Porém, o funcionamento dos métodos tradicionais se baseia na estrutura dos hospitais e em regras médicas, informações que, além de serem escassas, podem ser modificadas com o objetivo de se esconder evidências de fraudes.

Neste trabalho, propomos um método para detecção de anomalias em saúde que se baseia na demanda das cidades para detectar hospitais anômalos. Para isso, usamos informações que geralmente são abertas a consultas. Nosso método é composto de duas etapas: análise de anomalias e transferência de escore. Na etapa de análise de anomalias é realizada uma análise contextual das cidades com o objetivo de atribuir um escore para cada uma. Na etapa de transferência de escore, cada hospital recebe um escore considerando sua relação com as cidades.

Nós aplicamos o método em uma base de dados real do Sistema Único de Saúde do Brasil - SUS - considerando dez tipos de procedimentos que custaram mais de 8 bilhões e meio de dólares entre 2008 e 2012. Os resultados mostram que o método foi capaz de identificar casos de anomalias que não seriam encontrados sem as informações sobre as cidades e que a análise contextual de anomalias melhora os resultados em comparação com a análise pontual. Além disso, apresentamos exemplos de hospitais anômalos, ressaltando como o método foi capaz de identificá-los.

As principais contribuições deste trabalho são: I) um método simples e efetivo para detecção de anomalias em saúde pública. II) Nosso método não requer informações sobre os provedores de saúde e nem regras médicas. III) A análise sob a perspectiva dos consumidores possibilitou a identificação de anomalias que não seriam encontradas pelos métodos tradicionais. IV) Aplicamos o método em uma base de dados real e apresentamos um estudo de caso detalhado.

Palavras-chave: Detecção de anomalias, Mineração de dados, Saúde pública.

Abstract

Anomaly detection is an important task that has been largely applied to different scenarios. The improvements of technology for healthcare management and information storage has been enabling anomaly detection in healthcare. Traditional methods are based on the capacity of the hospitals and on medical rules. However, these information are rarely available and sometimes they are modified in order to hide evidences of fraudulent activities.

In this work we propose a simple method for anomaly detection in healthcare which is based on the analysis of the cities demand in order to detect anomalous and potentially fraudulent hospitals. We require only information that is usually available. The method consists of two steps: anomaly analysis and score transfer. In the anomaly analysis we perform a contextual analysis of the cities in order to assign an anomaly score to each one. In the score transfer, each hospital receives a score considering its relation with the cities.

We applied the method to a real database from the Brazilian public healthcare considering medical procedures that cost more than 8.5 billion dollars to the Brazilian government from 2008 to 2012. The results show that the method is able to find anomalous cases that may not be found if the features about the cities were not considered. Comparing the current method with our work with punctual anomalies, we verified an improvement caused by the analysis of contextual anomalies. We also performed a case study in which we show some evident examples of potential fraudulent hospitals, highlighting how our method was able to detected them.

Our main contributions are I) a simple and effective method for anomaly detection in healthcare. II) Our method does not require information about the providers nor medical rules. III) The analysis from the consumer perspective allows the detection of anomalies that could not be detected with traditional methods. IV) We applied the method to a real database and performed a detailed case study.

Keywords: Anomaly detection, data mining, healthcare.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Anomaly detection is an important task that has been applied to several scenarios and largely studied in many fields related to Data Analysis and Data Mining. The term anomaly refers to observations or patterns that are so unusual that it is important or interesting to understand its origin, especially if it presents a real life relevance.

Among many existing definitions, the definition of anomaly by Hawkins [1980] is probably the most adopted since its publication: *An anomaly is an observation which deviates so much from the other observations as to arouse suspicious that it was generated by a different mechanism.*

Another similar term and also important in Data Analysis and Data Mining is the concept of noise. However, despite the fact that both anomaly and noise are outliers objects, it is important to highlight that anomalies are observations more unusual than noise according to the subjective judgment of the analyst. Figure 1.1 represents these concepts according to the deviation degree.

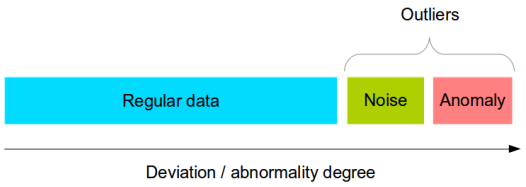

Figure 1.1. Representation of the concepts of noise, anomalies and outliers according to the deviation degree.

Figure 1.2 shows some examples of anomalies that can be easily identified given their patterns that strongly deviate from the other observations.

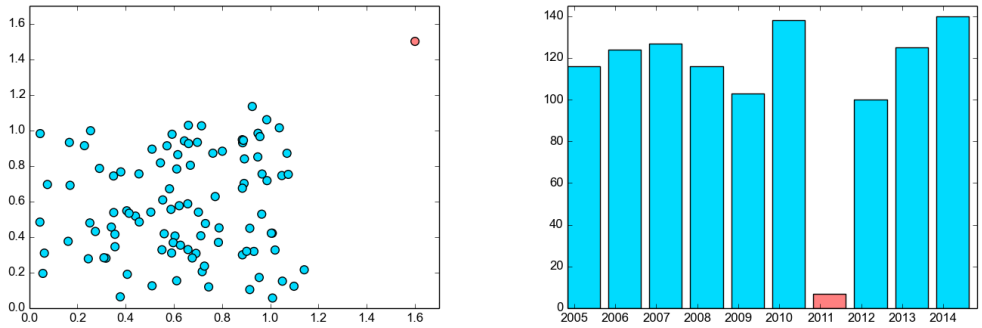

Figure 1.2. Examples of anomalies.

In the next section, we briefly describe the current scenario of anomaly detection and discuss its application in healthcare.

1.1 Context

Although the problem of anomaly detection is not new, there has been some changes in the last years. Recently, it was observed a huge increase of the contribution by the Computer Science community to the problem, especially by the communities of data mining, machine learning, data visualization and databases. According to Aggarwal [2013], most of the first works were performed by the statistics communities. While the first statistical methods are mathematically precise and formal, they lack some important aspects to enable the analysis in the current scenario, though.

In the recent years, it was observed a huge increase in the amount of data produced in all sectors of the society, especially due to the popularization of mobile devices, sensors for multiple activities (such as described in Surdak [2014]) and improvements in our capacity of storing and dealing with huge datasets. This phenomenon, known as Big Data, can be described by the huge amounts of data generated randomly and spontaneously throughout the world, in numberless ways and supported by the fast introduction of technology and the continuous reduction of IT infrastructural costs such as storage, transmission, and many other components, as stated by Jamil and Carvalho [2015]. Given this scenario, the problem of anomaly detection in real datasets is crucial and must follow the pace of the technological development of data generation and storage.

Among many others, we present some examples of important applications. The first one is the task of detecting frauds, that represents a specific type of anomaly. Nowadays, not only almost all financial transactions are recorded and stored, but they

can also be easily performed by anyone with a mobile device. Thus, quick detection of patterns related to fraudulent activities is a key requirement in such applications. Another important task is the detection of failures and defects. With the improvements of sensors and computer vision devices, the factories can detect instantly defective product, avoiding losses and further errors. This problem is known as industrial damage detection.

In spite of its importance and all previous efforts, anomaly detection is still a challenging problem. Besides the challenge involving the huge amount of data, some other core challenges are:

- Anomaly is a subjective judgment: it is not easy to define the boundary between anomalies and regular observations. In addition, it is also hard to distinguish anomalies from noise.

- Anomalies are rare and correspond to a very small part of the occurrences. Dealing with this unbalance is not trivial.

- When the anomaly is a consequence of designed malicious activities, such as frauds, the authors are usually concerned about hiding or modifying all the clues that would help the detection.

- In some scenarios, the definition of regular and non-regular behaviour is not static. It may change for different periods, for example.

- In most of the applications, there is neither labeled datasets nor ground truth for describing or modeling anomalies.

1.2 Anomaly detection in healthcare

Anomaly detection in records of healthcare is an important task that may reveal logistic problems, overloads, regional lack of professionals or services, disease outbreaks, errors in the data and suspicious activities. Hence, it is a key task to support cost reduction, improve records quality, support investments planning and especially to reduce fraud occurrence. This last task is crucial and can avoid loss of huge amounts of money. For example, according to the *FBI*, although 17% of the GDP of the United States was invested in healthcare in 2013, from 3% to 10% of the activities were fraudulent[1],

[1]U.S. Federal Bureau of Investigation. Financial crime report 2010-2011. www.fbi.gov/stats-services/publications/financial-crimes-report-2, 2012.

resulting in a waste of 125 to 175 billion dollars[2].

However, despite its importance, detecting anomalies in healthcare systems is challenging due to many reasons, such as the poor quality of data, lack of data, complexity and dynamism of the field, restricted access to data due to privacy issues and the lack of a global standardization for healthcare organizations. In addition, even if a complete and reliable database is available, it is not feasible to manually analyze all values and records declared by the healthcare providers once that auditing processes are usually expensive and complex. Thus, it is necessary to automatically select (or rank) those entities to be audited in order to reduce the rate of false positives. Predictive models are popular solutions for this purpose, but for the best of our knowledge there is no labelled dataset for anomaly detection in healthcare. Building one is not trivial due to the complexity and dynamism of the problem. As it is also hard to define the expected pattern for all entities, the most appropriate solution is the use of unsupervised learning.

Popular solutions for anomaly detection in healthcare are based on either supervised learning [He et al., 1997] or predefined medical rules [Major and Riedinger, 2002; Li et al., 2008]. Supervised learning demands a labelled dataset, which is rarely available, whereas medical rules may overfit to some specific scenario, such as one type of medical procedure. Furthermore, both approaches are expensive as they require manual work: a labelled dataset and medical rules are manually built by healthcare experts. The dynamic nature of the problem is also leverages the accuracy of these methods, since covering all types of anomalies is complex and new types appear frequently.

Another drawback of existing works, such as Ortega et al. [2006], is their dependency on specific data about healthcare providers. In many scenarios there is not enough information to support such analysis and, even when it is available, it is usually not reliable, since the providers, themselves, generate it. Further, some anomalous patterns can only be determined through the analysis of the entities associated with their activities. In these cases, although the providers are the targets, a strategy for identifying the anomalous ones is to find unexpected patterns in their consumers. In this work we investigate this latter approach.

1.3 Goal and contribution

The goal of this work is a method for anomalous providers detection in healthcare and its application to a real database. The method deals with two types of entities:

[2]R. Kelley. Where can $700 billion dollar in waste be cut annually from the us healthcare system? http://www.larson.house.gov/images/pdf/700billioninwaste.pdf, 2013.

hospitals (providers) and cities (consumers). From the information about the cities, the method detects anomalous hospitals considering the amount of procedures that each hospital performed in the population of each city.

The anomaly projection from cities to hospitals is justified by the lack of information about the hospitals structure and by the fact that the cities patterns enable the anomaly analysis from a new perspective, which is usually not possible if only information about the hospitals is employed.

The main contributions of this work are:

- A simple and effective method for anomaly detection in healthcare. In our model, the goal is to find anomalous hospitals through score transfer from cities. For the best of our knowledge, this is the first method for anomaly detection in healthcare that is based on score transfer concerning different entity types.

- Unlike traditional methods, our method requires neither medical rules nor features about the healthcare providers. In addition, it allows anomaly detection in scenarios where traditional methods usually do not work: when no features about the hospitals are available and when their anomalous behaviour can only be identified through anomaly analysis of the consumers.

- We applied the method to a real database of the Brazilian public healthcare system. We investigated ten types of procedures that cost more than 8.5 billion dollars between Jan/2008 and Dec/2012. We present some evident cases of anomaly and also an analysis over the amount of money that could have been saved if no anomaly occurred.

- The method is divided into two steps: anomaly analysis and score transfer. As our method can be implemented with many combinations of algorithms, we discuss approaches and algorithms that could be applied on each step considering different aspects of the application. We performed a comprehensive number of experiments with multiple algorithms for anomaly analysis, score transfer and score normalization. We also compare the results produced by contextual and punctual anomaly detection.

1.4 Document organization

The rest of this thesis is organized as follows. In Chapter 2 we present the main concepts related to the problem and review some related works. Chapter 3 presents a general

view of the method, the modeling and the implementation proposed. Chapter 4 shows the database, the experiments and the results. In Chapter 5 we present a case study with some evident cases of anomaly and a financial analysis of the results. Finally, Chapter 6 concludes the work and presents the future work.

Chapter 2

Background

In the recent years, many relevant works were produced to tackle the problem of anomaly detection in many fields.

For a complete view of the most important existing techniques and applications, we recommend the surveys by Chandola et al. [2009]; Hodge and Austin [2004], which also organize and compare the existing works.

Next we list the most relevant books that cover all the core issues and works related to anomaly detection. Hawkins [1980] was the first book to define the problem and cover the existing work, which is basically related to statistical methods, such as Bayesian and Distribution-based approaches. In Barnett and Lewis [1994] and Rousseeuw and Leroy [2005], which can be considered so far the most popular books about the topic, most of the content also rely on the relation between regression and outlier analysis. Recently, Aggarwal [2013] was published covering both the statistical methods and all the relevant works proposed by the Computer Science community.

We believe that in the next years a significant effort will be performed to adapt the existing solutions to the current scenario of Big Data following the new models and paradigms, such as the Map Reduce (Dean and Ghemawat [2008]) for massive volume of data.

Next we present a background about the problem of anomaly detection: the types of anomalies, the main types of algorithms, the main applications and a discussion about anomaly detection in healthcare.

2.1 Types of anomalies

According to the nature of the unusual observations, the anomalies may be classified as punctual, contextual or collective, as stated in Chandola et al. [2009].

Punctual anomalies: punctual anomalies are those instances that present extreme or rare values when compared to the other instances. In order to detect punctual anomalies, it is not necessary to analyze the context or the situation in which they occur. For example, if the age of students of a school is measured, the very extreme values are punctual anomalies as the only information used is their ages, without grouping or selecting the students to which each student is compared. This example is shown in the left most image of Figure 2.1: the boxplot indicates the median, the first quartile, the third quartile and the extreme values. The anomaly is far above the second greatest value. Punctual anomalies can also occur when multiple features are considered. For example, as illustrated in the right most image of Figure 2.1, if we measure the students height and weight, the extreme instances are also punctual anomalies, once that no context is established. Instead, all students are compared considering only these measures.

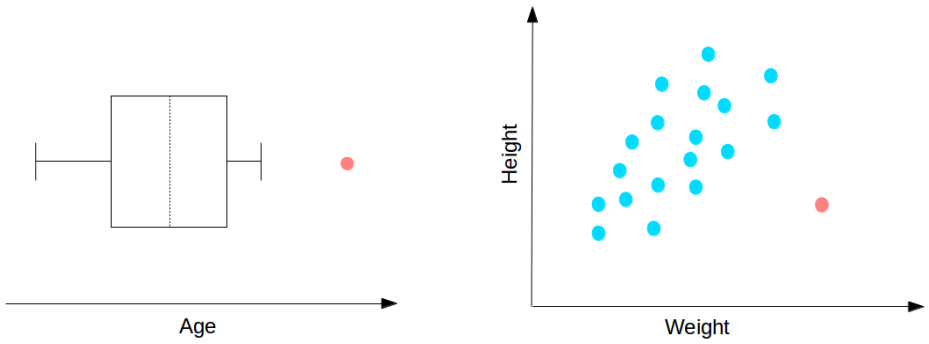

Figure 2.1. Example of punctual anomalies.

Contextual anomalies: contextual anomalies are those unusual observations when the context is considered. A contextual anomaly within a context can be a regular instances in others. In order to detect contextual anomalies two sets of features should be used: one to define the context and one to evaluate the behaviour.

For example, lets suppose that a study aims to find anomalous regions according to pressure and temperature. If the values of all regions are compared, the anomalies found could not be relevant to the study, as they are punctual anomalies. However, if the latitude and longitude are considered in order to establish the context, the anomalies would be relevant cases of unexpected occurrences. Figure 2.2 shows an example. In the left most image, the context of the red instance is defined according to the latitude and longitude: the green instances compose its neighbourhood. The middle image shows an example of regular instance: the red point is located close to its contextual neighbours in the behavioural space. However, the red instance in the right most ex-

ample is anomalous: although it is not isolated in the behavioral space, its behaviour is not in accordance with its contextual neighbourhood.

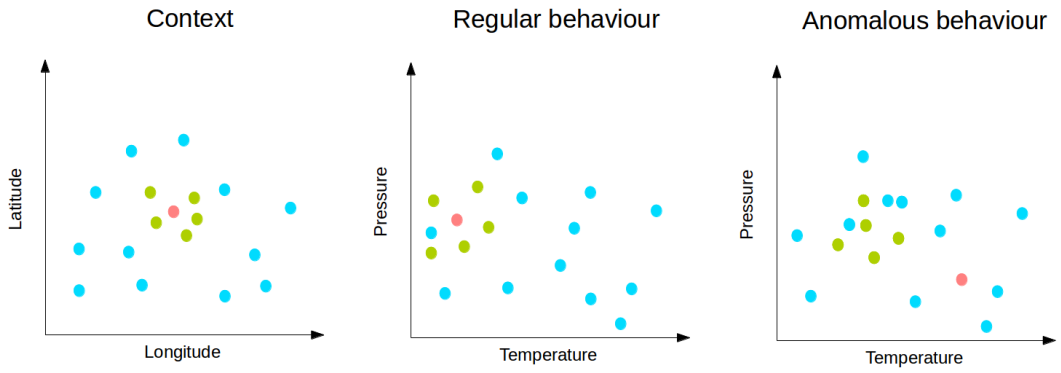

Figure 2.2. Example of contextual anomalies.

The work by Song et al. [2007] proposes a method for contextual anomaly detection and compares the observation of punctual anomalies, referred as the clearest outliers, with contextual anomalies, referred as conditional anomalies. In Kou et al. [2006] two methods are proposed for contextual anomaly detection considering spatial properties and their impact. In Schubert et al. [2014] the concept of spatial neighborhood is applied on a contextual anomaly detection for multiple applications: spatial, video, and networks.

Collective anomalies: collective anomalies are those instances that are anomalous when occur together, although each one of these instances is not an anomaly by itself. In order to identify collective anomalies, it is required a relationship among data instances which can be expressed as graphs or sequential data, for example.

In Vatanen et al. [2012], it is proposed a framework for semi-supervised anomaly detection of collective anomalies. The method assumes a fixed mixture model to describe the background in order to detect collective anomalies. The work in Jiang et al. [2014] presents a scalable framework for real time detection of collective anomaly over a collection of data streams.

2.2 Types of algorithms

The problem of anomaly detection can be solved by unsupervised or supervised algorithms. Unsupervised algorithms do not require labeled instances while supervised algorithms build classification models based on known instances.

2.2.1 Unsupervised algorithms

The main assumption of unsupervised algorithms is that the regular behaviour is more frequent than the anomalous behaviour. Thus, according to one or more criteria, the instances are analyzed in order to detect patterns that are isolated from the majority. Next, we present the main classes of algorithms for unsupervised anomaly detection: probabilistic, linear and proximity-based.

2.2.1.1 Probabilistic models

The key assumption of probabilistic models is that normal instances occur in high probability regions of the data distribution whereas anomalies occur in low probability regions. Next we present some popular approaches that show the principles of these methods.

The most simple approach consists of choosing a probabilistic distribution and trying to fit the data to it. The key idea is to assign anomaly probability inversely proportional to the probability of the observation of the instances in the distribution. The most common model is the Gaussian model. The parameters are chosen using techniques for finding the Maximum Likelihood Estimator, such as the iterative method of Expectation Maximization [Dempster et al., 1977]. Then, the data is fit to the model with the most likely parameters and the instances with less likely are considered anomalous. A pioneer study in Shewhart [1931] proposes that instances that differ from the mean with a magnitude of three times the standard deviation should be considered anomalous.

A basic approach that does not require parameters estimation is based on histogram analysis, such as the method in Goldstein and Dengel [2012]. It can be performed in multiple ways but one of the most common strategy is to assign anomaly degree inversely proportional to the frequency of the observation. The challenge in this case is to control granularity associated with the size of the bins. If they are too small, some regular instances may be considered rare. If they are too large, there may be a precision problem.

Next we discuss the main advantages of probabilistic and statistical models for anomaly detection. First of all, the score values produced by such methods are usually associated with a confidence interval that provide reliability and additional information for the decision in the results analysis. In addition, if the distribution estimated for the data is good, it is not necessary to have labelled instances and it is possible to justify statistically the results.

The main disadvantage of these methods is the assumption of one particular distribution. Besides the fact that it is not trivial to find such distribution, it is usually incorrect and approximated. It also raises the challenge of calibrating the model. If the model is too general, many parameters should be learned and set, resulting in an overfitted model. On the contrary, if the model is too restrictive, it is likely that the data would not fit the model, producing bad results.

2.2.1.2 Linear models

The main assumption of linear models is that there is correlation between the dimensions. The goal is to detect those lower dimensional subspaces in which the anomalous patterns are more different from regular data.

A trivial linear approach consists of fitting the instances in a linear regression model and assigning degrees of anomaly proportionally to their residual value. The key aspect to be observed in this approach is the trade-off between overfitting and generality of the model. If the model is too general, it would not represent correctly the data and then the anomalies found might not be real anomalies. If the model is too fit to the data, it would incorporate the anomalies so that they could not be detected.

This class of methods is largely used and usually present good results for a vast range of applications. However, they present poor quality when its main assumption, the existence of linear correlation among the dimensions, is not true. In some cases, such correlations exist only in some specific regions of the data. In addition, it is usually hard to justify the anomalies based on empirical evidences.

2.2.1.3 Proximity-based methods

Proximity-based methods assume that anomalies occur in sparse and isolated areas whereas typical instances are located in dense regions.

These methods require a notion of distance and similarity in order to compare the instances against each other. Although there are multiple metrics for computing these values, as shown in Pang-Ning et al. [2006], the most popular are the Euclidean distance, Jaccard index and Cosine similarity.

This class of methods can be divided into three sub-classes: cluster-based, distance-based and density-based.

The main assumption of cluster-based algorithms is that regular instances belong to a cluster whereas anomalies do not belong to any cluster. Some methods can be based on further aspects to define anomalies, such as the size of the clusters. The most popular method for cluster-based anomaly detection is *DBSCAN*, published by Ester

et al. [1996] which aims to cluster noisy data. Other clustering algorithms that are able to identify anomalies, as they do not assign all instances to a cluster, are *ROCK* proposed by Guha et al. [1999] and *SNN* proposed by ErtÃ¶z et al. [2004].

Distance-based algorithms exploit the distance of the instances to their neighbours. The main assumption is that anomalies are far from their neighbours whereas regular points are close to them. In Ramaswamy et al. [2000], it is proposed a formulation for distance-based anomaly detection based on the classic algorithm KNN: the anomaly degree of an instance I is proportional to the distance between I and its k^{th} closest neighbour. The method of reverse nearest neighbour, proposed in Hautamaki et al. [2004], assumes that instances that are not in the neighbourhood of their neighbours are likely to be anomalous.

Density-based methods define that anomalies occur in regions with low instance density. The main difference between cluster-based and density-based algorithms is that in the former the instances are partitioned into groups whereas in the latter the partition is based on the data space. The *Local Outlier Factor - LOF*, proposed by Breunig et al. [2000], detects anomalies that present density distributions significantly different from their neighborhood, even when the neighborhood of instances are located in areas of different densities. The method *Connectivity-based outlier factor - COF* Tang et al. [2002] treats low density and isolation differently. There is an improvement compared to the *LOF* effectiveness when dealing with instances with similar neighbourhood density as an outlier. In Jin et al. [2006], the method *Influenced Outlierness* is proposed. It applies the concept of symmetric neighborhood relationship in order to improve the local outlier approach for cases in which clusters of different densities are not clearly separated.

The main advantage of proximity-based methods is the fact that they are data driven and do not depend on any assumption about the data distribution. In addition, it is relatively easy to apply such algorithms in most of the databases.

According to Kriegel et al. [2010], the computational complexity of these algorithms limits their scalability: they can present quadratic computational complexity due to the nested loop to compute the distance between all pairs of instances. However, some works have been proposed in order to enhance scalability in large datasets. The method *ORCA* proposed in Bay and Schwabacher [2003] applies pruning and randomization for avoiding the quadratic cost and making possible near linear time performance. The method *RBRP* proposed in Ghoting et al. [2008] applies a powerful pruning approach based on micro cluster partitions. The *RBRP* scales log-linearly as a function of the number of data points and linearly as a function of the number of dimensions.

However, the assumptions of this class of algorithms may lead to failures in some cases. For example, it is possible that groups of anomalies exist close to each other, resulting in a cluster (or a close neighbourhood or a dense region) that cannot be detected by the algorithms. The method can also detect wrong anomalies or miss true anomalies if the parameters are not correctly chosen, such as the number of neighbours or the neighbourhood radius.

2.2.2 Supervised algorithms

The general goal of the methods based on classification is to learn how to distinguish instances of each class (regular and anomalies) given the features and a set of labeled instances. The process is divided into two phases: training, when the model for classification is created using the labeled instances and test, when the model is applied to classify the unknown instances.

There are many algorithms for classification that can be divided into many categories: Neural Networks-based (such as Zhang [2000], Freund and Schapire [1999]), Bayesian-based (John and Langley [1995], Webb et al. [2005]), Rule-based (Kohavi [1995], Veloso et al. [2006]) and those based on decision trees (Breiman [2001], Quinlan [1993]). We believe that the *One Class SVM* [Schölkopf et al., 1999], is one of the best algorithms for dealing with anomaly detection. Its basic idea is to map the training data into the kernel space and to separate them from the origin with maximum margin.

If accurate labels are available, this class of algorithms usually provides good and fast solutions. However, the availability of labels is a core issue in many data mining problems. When dealing with the problem of anomaly detection, the issue becomes more complex due to the reasons discussed next. First of all, the dynamic nature of the problem might change the labels over different places, time, legislation or other aspects. For example, the behaviour of a tracked bird can be anomalous in the hot season but regular in the cold season. In these situations, the labels should be carefully analyzed to make sure that they are consistent with the reality. In addition, it is usually very hard and expensive to obtain correct labels for anomaly detection, once that in most situations not only the process is performed manually, but also it is complex to cover all the possible unusual cases.

Even if reliable labels are available, applying supervised techniques for dividing the instances into regular and anomalous is challenging, though. The main reason is the problem of class imbalance, as the anomalous class usually represents a very small portion of the instances.

According to the perspective of labels availability, another possible approach is the

semi-supervised anomaly detection. Semi-supervised algorithms usually require only a small and strategic amount of labeled records for training. Given all the challenges related to the labels, semi-supervised methods are more feasible solutions than the supervised learning, however it is also not possible in many cases due to the problem of covering all the possible type of anomalies.

2.3 Output format

The solution provided by an algorithm for anomaly detection can be output in two formats: labels or score values. In the first case, each instance is assigned as either regular or anomaly. This format does not allow any distinction concerning the degree of exceptionally of the instances. We observe that although it is also possible to classify the instances to subclasses informing the type of anomaly, it does not quantify the level of abnormality.

On the other hand, the assignment of a score value for each instance allows the comparison of anomaly degrees, to rank the instances and also to label the instances through the application of a threshold. In addition, the scores may present different meaning depending on the method applied to produce them. If probabilistic models are applied, the score might represent how likely an instance is an anomaly. If the method is based on neighbourhood comparison, the scores could be computed as the distance to them. In the case of linear methods, it could be the deviation. Thus, this output format represents a more challenging but more meaningful and precise solution.

2.4 Applications

The problem of anomaly detection is important in an uncountable number of applications and scenarios. However, there are some core applications of anomaly detection that have been attracting efforts of several works. Among others, we present here four important applications: fraud detection, detection of intrusion, medical anomalies and textual anomalies.

Fraud detection is one of the most popular type of anomaly in the research community and presents increasing importance with the popularization of web/mobile transactions and e-commerce. Next we discuss some relevant works concerning the problem. The work in Fawcett and Provost [1997] proposes an adaptive system for fraud detection based on user profiling. The target application is the identification of fraud in mobile phone, which is one of the most popular applications of the problem. In

Gaber et al. [2013] it is presented a synthetic log generator for this application of mobile phone frauds. Recently, Tseng et al. proposed the Framework *FrauDetector* in Tseng et al. [2015] for fraud detection in phone calls through information propagation in the graph of users and phone numbers. With the popularization of the e-commerce and web applications for reviews, some new types of frauds have emerged, such as frauds in reviews, approached by Hu et al. [2011] and frauds in clicks on the Web [Pearce et al., 2014].

Intrusion detection is usually related to network and systems security. The goal is to identify malicious activities, such as malicious programs, hackers invasion, unauthorized behaviour and policy violations. The problem is not trivial if we observe that, in most cases the volume of data to be analyzed is huge and the detection must occur in real time to avoid damages. Next we list some relevant work concerning the problem of intrusion detection. The work Denning [1987] describes a pioneer, popular and generic model for system intrusion detection which is based on statistical models and rules for detecting abnormal patterns. In Hofmeyr et al. [1998], it is shown that the sequences of system calls represent a good discriminator between regular and invasive activities and systems. The work in Portnoy et al. [2001] apply clustering techniques for intrusion detection in networks environments. Recently, Tamersoy et al. [2014] presented a solution for malware detection that is based on information propagation on a huge graph built with real information of users files.

Medical anomaly detection is an important problem that helps the improvement of disease diagnosis and treatment. Some of the most popular related approaches are based on time series and image analysis. The main challenge is related to the damage in case of errors: in this case, the occurrence of false negative must be zero. The work in Lin et al. [2005] proposes the detection of time series discords aiming at a fast alert of unusual medical conditions in health monitor techniques, such as electrocardiograms. In Wong et al. [2003] a Bayesian network is applied to quick detect disease outbreaks. In Laurikkala et al. [2000] it is presented an analysis of outlier detection on medical data through the use of box plot, referred by the authors as informal anomaly detection.

The detection of textual anomalies is not a new task, however its importance has been growing with the popularization of social networks and with the increase of the amount of new documents in the Web, which demands automatic solutions for opinion and topic mining. The core challenges of the problem are the volume and sparsity of the data and temporal issues related to the information. In Baker et al. [1999] the anomaly detection is performed aiming the detection of new class of text through a framework that combines multiple models. The work in Srivastava [2006] shows how the NASA solved the specific problem of detecting anomalies in aerospace problem

reports through clustering techniques.

Other important and popular applications of anomaly detection are: industrial damage and image processing. In the next section, we present a general view and some relevant works related to our main topic: anomaly detection in healthcare.

2.5 Anomaly and fraud detection in healthcare

Anomaly detection is a crucial task in healthcare management that can improve the conditions and avoid loss of huge amounts of money, especially if the fraud occurrence is reduced.

A complete description of the popular types of healthcare frauds is presented in Fabrikant et al. [2014]. In spite of the existence of aspects that contrast healthcare systems of different countries and states, there are some recurrent types of frauds in most of them:

- **Billing for services not rendered:** this is the most common fraud in healthcare and occurs when the provider charges the government for medical procedures that were not performed.
- **Up-coding:** the up-coding fraud consists of charging for a medical procedure that is more expensive and complex than the procedure that was truly performed.
- **Duplicate billing:** occurs when the provider charge two or more times for the same medical procedure that was performed once.
- **Un-bundling of claims:** in this type of fraud, the provider charges individually for a group of procedures that would cost less if they were paid together.
- **Medically unnecessary services or excessive services:** these frauds are hard to detect as they involve charges of procedures that were truly performed but should not have been performed from a medical point of view.

In addition, these frauds can occur in different degrees of intensity that require different approaches for detection. If the frauds are committed in a slow pace, it is usually harder to detect them, but it is easier to identify the authors when the frauds are detected. According to Capelleveen [2013], this approach is known as *Steal a little all the time*. On the other hand, the approach of *Hit and run* consists of an intense fraudulent activity. Although it is trivial to identify the frauds, it is hard to identify the authors as they forge documents and go out of business in order to not be punished.

In the research community, the most popular approaches for fraud detection in healthcare are: *Peer Group Analysis*, *Clustering Analysis*, *Break Point Analysis* and *Single Anomalies*. The *Peer Group Analysis* (Bolton and Hand [2002]) approach aims to identify entities that started to present different behaviour from other entities that used to be similar. As previously discussed, the *Clustering Analysis* is based on identification of isolated entities or groups of entities. The *Break Point Analysis* is similar to *Peer Group Analysis*, however, in this approach, the entities are compared to its own past behaviour in order to identify points of change. As defined by Chandola et al. [2009], *Single Anomalies* are entities whose behaviour do not conform with an expected behaviour defined as regular. This definition is usually based on medical rules or providers capacity.

With the improvements of management systems for healthcare observed in the past years, many works have been developed to deal with the produced data and extract knowledge from it. The pioneer work in He et al. [1997] applies records labelled by experts to create a neural network to identify medical frauds. In Yang and Hwang [2006] the generic framework *MCI HCFAD* is introduced for healthcare fraud and abuse detection. In Aral et al. [2012], the problem of unnecessary services is investigated through the identification of frauds on medical prescription.

2.6 Discussion

As presented in section 1.2, existing works deal with the profile of the patients and with the analysis of the providers capacity. Although these information can provide good results, they usually are not available or not correct.

In addition, the existing works about fraud detection in healthcare deal with the two public healthcare systems of the *U.S*: the *Medicare* and the *Medicaid*. Most of these works, such as Agrawal et al. [2012] and Becker et al. [2005], consider all the details and peculiarities of the American systems in their definition. However, the legislation and rules of the countries and states have great impact in healthcare. Thus, the differences between the *Medicare*/*Medicaid* and the healthcare systems of other countries make it very hard to apply and adapt the existing methods to other countries.

For the best of our knowledge, our method for anomaly detection in healthcare is the first one to:

- allow the detection of anomalous healthcare providers from the consumers analysis. This approach allows the discovery of anomalies (and potential frauds) that

could not be found by traditional methods that consider only information about the providers.

- discover anomalies in the Brazilian public healthcare system. Although the database with records of all transactions is available on the Web, for the best of our knowledge, there is no work able to identify the anomalous providers and justify their anomaly through evidences based on such database. Although our case study was focused only on the Brazilian healthcare system, we believe that the method can be applied to most of the healthcare systems in the world.

As we show in the next chapter, we apply the method to identify anomalous amounts of a procedure type. Thus, among the popular types of frauds in healthcare, we believe that we are able to identify anomalies caused by I) billing for services not rendered, II) up-coding, III) duplicate billing, IV) medically unnecessary services and V) excessive services.

Our method can be applied for detecting punctual and contextual anomalies. The detection of collective anomalies is beyond the scope of this work. In our experiments, we compare the results produced by punctual and contextual anomaly detection.

We believe that it is crucial to output the degree of abnormality of the instances. Thus, as detailed in the next chapter, the method outputs a score value for each instance. In addition, our method is generic. As we show in the next chapter, it can be implemented with different types of unsupervised algorithms for anomaly detection. In our experiments, we show that we were able to achieve good results using simple and intuitive algorithms, such as the *KNN*, as we employ transfer learning.

Chapter 3

Method

In this chapter we describe our method, detail its modeling for public healthcare, show the steps and their implementation and discuss other aspects: temporal granularity and normalization. We conclude the chapter with a discussion about its effectiveness, cost and limitations.

3.1 Overview

In this section we describe our method: the provider/consumer model and our methodology.

3.1.1 Provider/consumer model

Providers and consumers are actors in most, if not all, services. It is not different for healthcare. Thus, although we apply the method for anomaly detection in healthcare, it can be described as method for anomaly detection in services. The providers are those entities that perform and sell the services whereas the consumers use the services and pay for them. The relation between these two types of entities can be represented through a bipartite graph as shown in Figure 3.1. Each provider can be linked with multiple consumers and vice versa. The weight of each edge measures the number (or amount of money) of services performed between each pair.

The anomaly detection is performed through capacity analysis for providers and demand analysis for consumers. Anomalous providers are those performing anomalous amounts of services compared to the other providers considering their capacities. Likewise, if the demand is analyzed, the anomalous consumers are those instances that consume abnormal amounts of services.

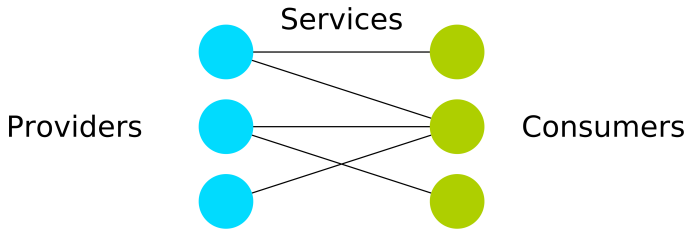

Figure 3.1. Bipartite graph representing the relation between providers and consumers.

Detecting anomalous providers is a trivial task if features about their capacities are available. The same occurs if the target entity are the consumers and features about their demand are available. However, in both cases, it is not possible to evaluate directly the abnormality of the entities without the features about them. Our method addresses this problem: instead of estimating directly the anomaly in the target entity type, we determine anomalies in the other entity type and then we estimate the abnormality of the target type considering the relation between pairs of entities.

Thus, the method can be applied to scenarios in which:

- the goal is to detect anomalous entities of one type (providers or consumers) without having information about it (or the information is not reliable);
- it is available features to estimate anomalies of the other type;
- it is known the amount of services concerning each pair of provider and consumer.

3.1.2 Methodology

As shown in Figure 3.2, after modeling, the method can be divided in two steps: *anomaly analysis* and *score transfer*. In the anomaly analysis step, an anomaly score is assigned to the entities of the type that we have features about. In the score transfer, it is assigned a score for the instances of the other type.

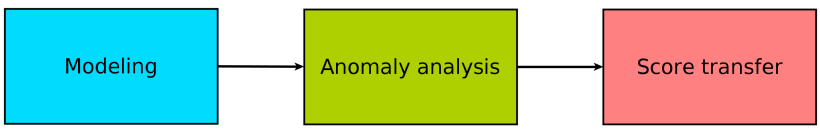

Figure 3.2. Steps of the method for anomaly detection.

If the goal is to detect anomalous providers and there are features only about the consumers, the step of anomaly analysis would consist of assigning anomaly scores

to consumer instances through the consumers demand analysis. Then, in the step of score transfer, the anomalous providers are estimated considering the consumer scores and the weight of the edges. Otherwise, if we want to detect anomalous consumers and there are features only about the providers capacity, we estimate the anomaly degree of the providers and then we transfer the score from providers to consumers.

Next, we detail how we modeled the problem and how we implemented these two steps of anomaly analysis and score transfer.

3.2 Modeling anomalies in public healthcare

In this section we show how we design a method for applying to public healthcare systems. Our assumption is that the public healthcare operates as: each hospital performs medical procedures in the population and these procedures are paid by the government upon request.

From the assumption that fraudulent entities present anomalous patterns, our goal is to detect anomalous hospitals that declared and charged for unexpected amounts of procedures. Although it does not imply that the anomalies found are cases of frauds, they are more likely to be fraudulent and their investigation should be priority.

We assume that there is no information about the hospitals capacity to evaluate whether the amount of procedures is unusual given their capacity. In addition, even if we had a database on their capacity, we should not trust it because the real values could be modified without affecting the amount of money received by each hospital. On the other hand, if we analyze the number of procedures that each hospital declared to have done, we are dealing directly with the key information that affects the amount of money received by them.

Usually, each procedure that a hospital performs has to be declared and some information about the patient are demanded by the government in order to pay for it. A basic information that is demanded is the city where the patient lives. If a hospital declares more procedures than the usual, the aggregated amount of procedures in the population of one or more cities is going to be greater than the actual amount.

As different cities have different sizes, we cannot compare them considering the absolute number of procedures. Instead, for evaluating the cities behaviour, we consider the rate of procedures computed as the number of procedures of each city divided by its population size.

Therefore, we model the method as a bipartite graph between hospitals and cities, as illustrated in Figure 3.3. An edge represents the number of procedures performed

by the hospital in the population of the city. In addition, it is required some additional features about the cities in order to compute their rates of procedures.

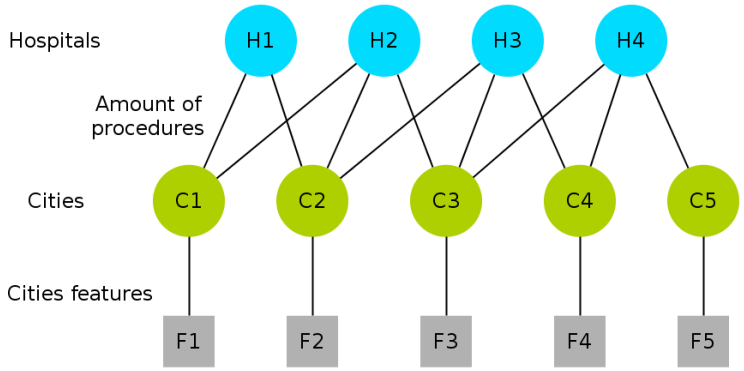

Figure 3.3. Modeling of the method considering hospitals and cities.

Based on this modeling, our problem can be defined as: *How to estimate the anomaly degree of all hospitals given: (I) a dataset with features about the cities and (II) the number of procedures performed by each hospital in the population of each city?*

Toy example: suppose three cities: A, B and C. Their population are served by four hospitals: 1, 2, 3 and 4. Given the number of occurrences of a procedure in each pair of hospital and city, we want to detect the anomalous hospitals. Suppose that the score assignment was performed based on the cities demand. Table 3.1 shows the score of each city.

Table 3.1. Anomaly scores for cities A, B and C.

City A	City B	City C
0.21	1	0.82

As city A presents low anomaly score, we consider that it is not anomalous whereas cities B and C present high anomaly degree. The fraction of people from each city treated in each hospital is shown in Figure 3.4.

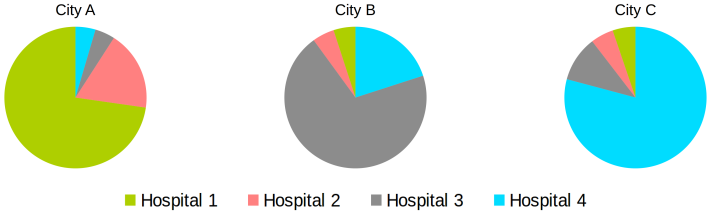

Figure 3.4. Amount of the population of each city treated in each hospital.

It is possible to conclude that hospitals 3 and 4 are more likely to be anomalous because they are the healthcare providers more related to the more anomalous cities.

3.3 Anomaly analysis

The anomaly analysis is the first step of the method. The goal is to assign, for each city C, an anomaly score $S(C)$ considering its behaviour, represented by the rate of procedures performed on its population.

This step can be implemented through several algorithms for anomaly detection. The key aspect of this step is the definition of the type of anomaly: as shown in the previous chapter, our method is able to identify punctual or contextual anomalies.

3.3.1 Punctual analysis implementation

If we look for punctual anomalies, almost all algorithms for anomaly detection may be applied. Next we briefly describe two simple algorithms that we implemented for a punctual analysis in Carvalho et al. [2015]: the *Reverse nearest neighbour* and the *LOF*, described in Section 2.2.1.3. These algorithms can be applied for detecting punctual anomalies, but not for contextual anomalies, as they use the same feature to group the instances and measure their behaviour.

Reverse nearest neighbour: the assumption of this algorithm is that if an object is not among the closest neighbours of its K closest neighbours, it is isolated and then it is an anomaly.

The main step of the algorithm is the construction of a directed graph in which the nodes represent the instances and the edges link an instance to its K closest neighbours. Thus, each node has outdegree equal to K. The anomaly score $S(C)$ of each instance C is assigned as $S(C) = \frac{1}{IN(C)}$, being $IN(C)$ the indegree of its node.

Local Outlier Factor: the *LOF* is a density-based algorithm described in Section 2.2.1.3. According to this algorithm, anomalies are those instances located in regions with density of instances significantly different from their neighbors.

The neighbourhood of an instance C is composed of those instances located in a distance smaller than or equal to the distance $D_K(C)$ from C to its K nearest neighbour. The size $|N|$ of the neighbourhood $N(C)$ can be greater than K in the case of ties.

The algorithm is based on three concepts: *Reachability*, *Average Reachability* and *LOF Value*.

For each instance C, the $Reachability(O, n_i)$ between C and each one of its neighbours n_i is defined as:

$$Reachability(C, n_i) = max(dist(C, n_i), D_K(n_i))$$

The $AverageReachability(C)$ is the average *Reachability* of C to all of its neighbours:

$$AR(C) = \frac{\sum_{n_i \in N(C)} Reachability(C, n_i)}{|N(C)|}$$

The value of $LOF(C)$ is the anomaly degree $S(C)$ of C and is defined as:

$$S(C) = LOF(C) = \frac{\sum_{n_i \in N(C)} \frac{AR(C)}{AR(n_i)}}{|N(C)|}$$

If an instance is not isolated, we expect that its average reachability is similar to the average reachability of its neighbours, so the expected value for the LOF in this case is 1. If the LOF of an instance is much larger than 1, it is likely to be anomalous.

3.3.2 Contextual analysis implementation

If the goal is to detect contextual anomalies, the options are restricted to those algorithms that can be modeled with two types of features: contextual and behavioural. The contextual feature defines the neighbourhood of the instances whereas the behavioural feature is applied for computing their anomaly degree within their context. As previously shown, we apply the rate of procedures as the behavioural feature. Thus, after grouping the cities according to their context, we compare them by considering their distance in terms of rates of procedures.

In the current work, we deal with contextual anomalies as detailed in the next chapters. We implemented the anomaly analysis step with two algorithms: one is a simple statistical solution referred to as *distribution-based solution* and the other is the *KNN* algorithm. The choice of the algorithm is based on simplicity and effectiveness. In addition, they employ different approaches for anomaly detection: the first is probabilistic while the second is based on proximity.

Distribution-based solution: this solution applies a t-test in order to assign

the score to each city. For each city, two distributions of distances are analyzed. The first is a random distribution composed of behavioural distances between random pairs of cities. The second distribution is composed of the behavioural distances between the city and its contextual neighbours.

For each city C, its score $S(C)$ is given by the p-value produced by the t-test with these two distributions. The p-value indicates the probability of observing the same distributions if no correlation exists between them.

If the behavioural distances from a city to its contextual neighbours are similar to the random distribution, the p-value is high and the city is likely to be anomalous as it does not behaviour according to its contextual neighbours. On the other hand, if the behavioural distances are smaller than the random case, the p-value is small, indicating that the city presents similar behaviour to its contextual neighbours.

The only parameter of this algorithm is the number K of cities in the neighbourhood of each city. The size of the random distribution was fixed at $5,000$ in order to provide a reliable sampling between random pairs of cities. The distribution-based algorithm is presented in Algorithm 1.

Algorithm 1 Distribution-based solution for anomaly analysis in the cities.

```
random_distribution ← new_array(5000)

{Generate the random distribution.}
while (iteration ≠ 5000) do
   random_position_1 ← random_int(0, number_cities)
   c1 ← cities[random_position_1]
   random_position_2 ← random_int(0, number_cities)
   while (random_position_2 = random_position_1) do
      random_position_2 ← random_int(0, number_cities)
   end while
   c2 ← cities[random_position_2]
   random_distribution.append(distance(c1.rate, c2.rate))
   iteration += 1
end while
iteration ← 0

{For each city, perform the t-test and assign the score.}
for all (current_city in cities) do
   neighbourhood_distribution ← new_array(K)
   for all (neighbour in current_city.neighbourhood) do
      neighbourhood_distribution.append(distance(current_city.rate, neighbour.rate))
   end for
   current_city.score ← ttest.pvalue (random_distribution, neighbourhood_distribution)
end for
```

KNN: The KNN algorithm is a popular algorithm in data analysis and consists,

basically, of evaluating the behaviour of the K closest neighbours of an instance for voting. Applying the KNN to the contextual analysis, the score of a instance is given by its behavioral distance to its K contextual neighbours.

The KNN also takes only the parameter K of the contextual neighbourhood size. Its implementation is shown in Algorithm 2.

Algorithm 2 KNN algorithm for anomaly analysis in the cities.

```
{For each city, assign as score the behavioural distance to the neighbours.}
for all (current_city in cities) do
   for all (neighbour in current_city.neighbourhood) do
      current_city.score += distance(current_city.rate, neighbour.rate)
   end for
end for
```

In Section 4.2 we present an experimental comparison between the distribution-based solution and the KNN.

3.4 Score transfer

The score transfer step consists of assigning an anomaly score $S(H)$ to each hospital H. The inputs are the anomaly degree $S(C)$ of each city C and the number of procedures $W(H, C)$ between each pair city and hospital.

3.4.1 Transfer approach

The main aspect to be considered is the transfer approach. Here we present three potential options: linear transfer, propagation or optimization. The meaning of the score transfer in the real application should be observed for choosing the transfer approach, so we also discuss their applicability in healthcare.

Without loss of generality, we assume here that we want to transfer the score from consumers to providers.

Linear transfer: this is a simple and intuitive solution for the score transfer that can be applied on almost all scenarios. It consists of applying a function f to perform a linear combination between the consumer scores and edges weight:

$$S(P) = \sum_{C_i \in Consumers} f(S(C_i), W(C_i, P))$$

Propagation-based: the score propagation can be seen as a loop of score transfer between providers and consumers. This approach should be implemented in scenarios where it makes sense to penalize a provider P_1 if it is connected to the same consumers of an anomalous provider P_2. Although there are multiple algorithms to implement this approach, all of them are based on the same mechanism. Initially, the score of all providers are the same whereas the score of each consumer is the value received in the step of anomaly analysis of the method. The link between each pair (P, C) of provider and consumer defines the propagation intensity. The propagation loop is performed alternately from consumers to providers and from providers to consumers until the stop condition is reached. The propagation approach can be implemented with several algorithms such as *Page Rank* (Page et al. [1999]), *Hits* (Kleinberg [1999]) and *Salsa* (Lempel and Moran [2001]).

Optimization-based: the key idea of the optimization solution is the opposite of the propagation. The goal is to concentrate the high score only in those providers that cover better the anomalous consumers. The main challenge of this approach is the high cost: despite the quality of the results, solving the transfer problem as an optimization problem is usually NP-hard. One of the most popular and generic method for optimization problems is the *Simplex* method as described by Dasgupta et al. [2006]. In order to achieve good and fast results to complex optimization problems, a popular solution is the use of Genetic Algorithms, as described in Mitchell [1998], or other bio-inspired algorithms, such as Ant colony optimization, described by Dorigo et al. [2006].

Score transfer in healthcare: we implemented two of these approaches to solve the score transfer problem: the linear transfer due to its simplicity and intuitiveness, and a genetic algorithm in order to experiment a optimization-based solution without dealing with the exponential cost. Next we detail the implementations.

Although the propagation-based solution could be an useful approach for score transfer in many scenarios, we do not believe that this solution conforms to the health-care scenario due to the repeated propagation. The anomalous behaviour are caused by few hospitals that infect the cities related to them. Thus, there is no reason for keeping propagating high scores beyond the anomalous hospitals and the infected cities. For instance, if a regular hospital $H1$ is linked to a city that has been affected by an anomalous hospital $H2$, $H1$ should not receive a high score, but it would happen if a solution based on propagation were applied.

3.4.2 Linear transfer implementation

For the linear transfer we implemented two different functions to control the score transfer: *simple linear transfer* and *proportional linear transfer*.

The first function is the *simple linear transfer*. For each pair of hospital H and city C, the score $S(C_i)$ of C is multiplied by the amount of procedures $W(H, C)$ performed by H on population of C:

$$S(H) = \sum_{C_i \in Cities} S(C_i) * W(H, C_i)$$

The practical meaning of this solution is that each hospital receives the score of each city weighted by the absolute amount of procedures performed by the hospital in the population of the city. Thus, if a city is anomalous and the hospital is strongly related to it, it is likely that the hospital is the responsible for the city behaviour, then the hospital receives significant score from the city.

The implementation of the simple linear transfer is shown in Algorithm 3.

Algorithm 3 Simple linear transfer algorithm for score transfer.

```
for all (hospital in hospitals) do
    hospital.score ← 0
    for all (city in hospital.related_cities) do
        hospital.score += city.score * amount(hospital,city)
    end for
end for
```

The second function is the *proportional linear transfer*: the score of each city C_i is weighted by the fraction of procedures performed by H on its population. $W(C_i)$ represents the whole amount of procedures in C_i and $W(H, C_i)$ represents only the amount performed by H:

$$S(H) = \sum_{C_i \in Cities} S(C_i) * \frac{W(H, C_i)}{W(C_i)}$$

The practical meaning of this solution is that the score of each city is proportionally divided among the hospitals. The score transferred from a city to a hospital depends only on the fraction that the hospital represents for the city. As the absolute amount of procedures is not considered, a big city C_{big} and a small city C_{small} may transfer the same portion of their scores to a hospital H if the percentage of procedures in C_{big} and C_{small} performed by H is the same.

The proportional linear transfer implementation is shown in Algorithm 4.

Algorithm 4 Proportional linear transfer algorithm for score transfer.

```
for all (hospital in hospitals) do
  hospital.score ← 0
  for all (city in hospital.related_cities) do
    hospital.score += city.score * (amount(hospital,city) / amount(city))
  end for
end for
```

3.4.3 Genetic algorithm implementation

In order to apply the optimization-based solution, we also implemented the score transfer with a genetic algorithm. Genetic algorithm is a type of evolutionary algorithm inspired by the natural selection process. It implements heuristics in order to find good solutions for complex problems, specially optimization and search problems.

The algorithm takes as input the amount of procedures between each pair of hospital and city ($W(H, C)$) and the score of each city ($S(C)$) and outputs the best score assignment for each hospital $S(H)$.

In this section we present the basic concepts related to genetic algorithms and show how they were implemented.

Generation: genetic algorithms simulate the evolution of a population over the generations. The core of such algorithms is a loop in which each iteration simulates a generation with a different population. In our implementation, the parameter G defines the number of generations.

Individual and population: each individual in a genetic algorithm represents a candidate solution for the problem. The individuals are defined by a genotype and a fitness value. In each generation, a set of individuals, called population, is created and evaluated. The parameter P defines the number of individuals in each population.

Genotype and Phenotype: The genotype of an individual codes its genetics and usually is a binary string. The map of this genetic code generates its phenotype, that is the solution that the individual represents to the problem.

In our implementation there is no difference between the genotype and the phenotype of the individuals. The solution to the problem represented by each individual is an array with a score value for each hospital. For example, if only 4 hospitals H_A, H_B, H_C and H_D were active, an individual would be [0.23,0.45,0.87,0.12] indicating the respective score for each hospital.

Fitness function: the fitness function evaluates the quality of the results produced by each individual according to an objective function. According to the meaning of anomalies in our application, our fitness function should assign low fitness values to good individuals, in which

- anomalous hospitals have high scores,
- regular hospitals have low scores.

Otherwise, the fitness value should be high. Next we detail the fitness function implemented.

The fitness function of each individual I considers its contextual neighbourhood of size K and it is based on two metrics: the ideal amount of procedures of the cities and the expected amount of procedures of the cities based on the individual scores.

The ideal amount of procedures of each city C_i estimates the whole amount of procedures in the population of C_i if no anomalies existed. We estimate the ideal amount of procedures in each city C_i as the number of procedures performed in its population according to the average behaviour of its neighbours. The algorithm to compute the ideal amount of procedures in each city is shown in Algorithm 5.

Algorithm 5 Algorithm to compute the ideal amount of procedures in each city.

```
for all (city in cities) do
    average_rate ← 0
    for all (neighbour in city.neighbourhood) do
        average_rate += neighbour.rate / size(city.neighbourhood)
    end for
    city.ideal_amount ← city.population * average_rate
end for
```

In order to compute the expected amount of procedures of the cities, we consider that the score assigned by each individual I to each hospital H estimates the amount of procedures that should not have been done by H. For each hospital H, the complement of its score $(1 - S(H))$ estimates the regular amount of procedures associated with it. For instance, if the score of H is 0.9 according to I, we estimate that 90% of the procedures are anomalous and only 10% of its real amount should have been done if H were not anomalous. The application of this operation in all hospitals gives us the expected amount in each city.

For each individual I, the fitness is computed as the sum of the difference between the ideal and expected amount in each city. It means that we evaluate the quality of each individual as the difference, expressed in number of procedures, between the ideal

scenario (in which no anomaly exists) and the expected scenario if the anomalous procedures according to I did not exist.

The fitness function is represented in Algorithm 6.

Algorithm 6 Fitness computation to measure the quality of an individual.

```
for all (hospital in hospitals) do
   expected_rate ← 1 - individual.score(hospital)
   for all (city in cities) do
      expected_amount(city) += amount(hospital,city) * expected_rate
   end for
end for

for all (city in cities) do
   fitness += absolute(ideal_amount(city) - expected_amount(city))
end for
```

Initial population: in genetic algorithms, the initial population is usually randomly generated. However, we have to consider our scenario before generating a score distribution for the hospitals: our assumption is that most of hospitals are regular whereas just rare cases represents anomalies. As a solution, the initial population follows an exponential distribution.

Figure 3.5 presents the score distribution of four examples of individuals generated with the exponential distribution.

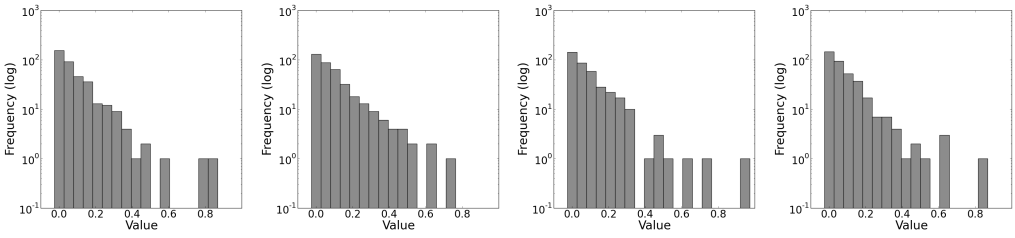

Figure 3.5. Examples of individuals generated following an exponential distribution.

After the first generation, the populations are generated through reproduction operations, which are not based on the score distribution.

Reproduction: the reproduction consists of operations performed over the population of a generation in order to create the population of the next generation. We applied two types of reproduction operations: crossover and mutation.

The first step of the reproduction consists of selecting good individuals for the operation. The strategy applied is the tournament. In this strategy, some individuals are randomly chosen and the one with best fitness is selected. The parameter TS (tournament size) indicates the number of individuals in each tournament.

The crossover simulates the reproduction between two individuals of a generation G_i resulting in two new individuals for generation G_{i+1}. During the crossover, the genetic of the parents' individuals are combined. In our implementation, we choose at random a position p of the genotype array. The range 0 to p of the genotype of the first child comes from the first parent whereas the remaining come from the second parent. The second child is generated in the opposite way: genes 0 to p come from the second parent and the remaining genes come from the first parent, as illustrated in Figure 3.6.

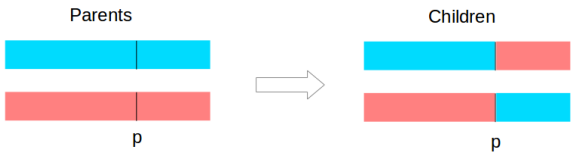

Figure 3.6. Crossover operation.

The mutation consists of a random change in the individual's genotype. Our mutation approach consists of changing the score of one hospital in one individual by a random number between 0 and 1. This new value is generated according to an uniform distribution.

The parameters CP and MP defines the probability of occurrence of crossover and mutation, respectively.

Elitism: the elitism is an operation which consists of reproducing the individual with best fitness of generation G_i in the generation G_{i+1}. The elitism occurrence is defined by a binary parameter E in our implementation.

Algorithm: Algorithm 7 presents the genetic algorithm that we implemented for the score transfer step.

The records can be aggregared by different temporal units, such as days, weeks or months. The next section shows how the method can be adjusted in order to deal with different temporal granularity.

Algorithm 7 Genetic algorithm implemented to find the best score assignment for the hospitals.

```
generation ← 0
Population_0 ← random_population(P)

while (generation ≠ G) do
   for all (individual ∈ population_generation) do
      compute_fitness(individuals)
   end for
   population_generation+1 ← empty_population(P)
   if (E) then
      population_generation+1 ← elitism(population_generation)
   end if
   while (size(population_generation+1) ≠ P) do
      if (random_float(0,1) < crossover_probability) then
         I_a ← tournament(population_generation)
         I_b ← tournament(population_generation)
         population_generation+1 ← crossover(I_a, I_b)
      end if
      if (random_float(0,1) < mutation_probability) then
         I_c ← tournament(population_generation)
         population_generation+1 ← mutation(I_c)
      end if
   end while
   population_generation ← population_generation+1
   generation ← generation + 1
end while
```

3.5 Dealing with variable temporal granularity

When the data is aggregated in one period, such as week or month, the analysis can be applied as described above. However, the records are usually represented as time series concerning multiple blocks of period U.

In order to deal with different temporal granularity, the analysis is performed in periods called windows. The parameter W defines the number of time blocks in each window and S defines the sliding distance between two adjacent windows. For instance, if the value of W and S are respectively 6 and 3, each window has size $6U$ and there are 3 blocks ($3U$) separating two adjacent windows, as illustrated in Figure 3.5.

Applying such concept in our modeling, in each window w, each city C receives an anomaly score $S(C)_w$ considering its behaviour within w. Then, in the next step of the method, each hospital H receives an anomaly score $S(H)_w$ considering both the score $S(C)_w$ of each city and the amount of procedures between each pair $W(H, C)_w$.

The values of W and S have great impact on the results. If the value of W is too small, the result becomes too sensitive to small variations and loses its reliability. On

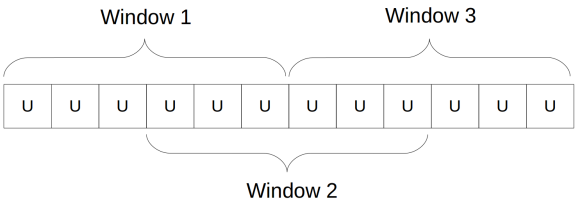

Figure 3.7. Example of time division with window $6U$ and sliding $3U$.

the contrary, if W is too large, the score tends to lose its significance as the anomalous behaviours would be smoothed. A good value for W would enable us to distinguish the anomalous from regular cities in a reliable way.

The value of S defines the number of windows in the analysis and the overlapping between them. If S is low, we produce many windows with great overlapping. In this case, each period U is analyzed in multiple windows. If S is close or equal to the window size, we reduce the number of windows and the overlapping, and each window tend to be independent from the others. The value of S cannot be greater than the window size, otherwise one or more months are skipped in the analysis.

3.6 Score normalization

Different algorithms analyzing different types of procedure produce different score ranges. Thus, after assigning a score for each city in each window, we perform a normalization step to set the score range from 0 to 1. There are two issues related to this normalization process: the normalization moment and the normalization method.

The normalization moment refers to the decision of normalizing separately the scores values of each window or normalizing all the values of all windows. In the first case, in each window, at least one city would present maximum score. In the second, the maximum score refers to the most anomalous city in the whole period and it is possible that no city receives high score in one or more windows.

The disadvantage of normalizing the scores for the whole period is that the analysis can be affected by isolated events and occurrences: one occurrence in one window would affect the scores of all windows. However, despite of this disadvantage, we adopt this solution for three reasons. First, although it is possible that all the high scores occur in the same window, we can still rank the cities of the other windows. Second, normalizing all the scores of all windows together makes it possible to rank and compare all the cases of anomalies, even in different windows. And, most important, if we normalize each window separately, we cannot measure which entities are the most

anomalous for the whole period: all the most anomalous cities of each window would be equally anomalous. It would affect the score transfer and reduce the precision of the results.

For the normalization method we implemented two algorithms: the trivial normalization approach in which all the scores are divided by the greatest score value and the *Unified method* described in Erich and Zimek [2011]. The last one consists of a method that, given a set of instances and their values, assigns scores between 0 and 1 indicating the probability of the instances being anomalies.

3.7 Discussion and limitations

In this section we discuss the effectiveness, the cost and the limitations of the method and our modeling.

Effectiveness: our goal is to detect anomalous hospitals that claimed for more procedures than expected. Our approach consists of detecting cities with anomalous rate of procedures in order to infer which hospitals caused such anomalies. We believe that it is an effective solution due to the following assumptions:

- If a hospital claims the payment of a procedure, it has to be informed the city where the patient lives.
- If the amount of procedures claimed by a hospital is greater than the expected, it will eventually affect the amount of procedures in one or more cities.
- If we have some features about the cities, such as the population size, it is possible to estimate the rate of procedures of each city in order to analyze them and identify anomalies. It is performed in the step of anomaly analysis.
- Once that we know the anomaly degree of the cities and the amount of procedures performed by each hospital in the population of each city, it is possible to infer the anomaly degree of the hospitals in the score transfer step.

In addition, we believe that the amount of people from each city attended by each hospital is a trivial information that is included in most of healthcare databases. Moreover, as most of the countries or cities perform demographic surveys, it is usually easy to find reliable features about the cities. Thus, we conclude that this analysis can be performed in a vast range of healthcare databases.

Cost: we implemented two algorithms for each step of the method: distribution-based and KNN for the anomaly analysis and the linear transfer and genetic algorithm for the score transfer. The step of anomaly analysis is the most expensive step as the computation of the nearest neighbours of each city is the most expensive operation of the method.

The two algorithms implemented for the anomaly analysis require information about the neighbourhood of each city. We precompute the contextual neighbourhood of each city through a nested loop with quadratic cost in order to find the distance between all pairs of cities. Although this solution is not scalable to very large datasets, we do not believe that scalability is an important requirement for the application as we deal with pairs of cities: the size of the databases would not scale to very large amounts.

If the method would be applied for large dataset, it could be applied some strategies for cost reduction in the K-nearest neighbour computation, as described in Section 2.2.1.3.

Limitations: there are two scenarios in which the modeling may not be effective: when we deal with very small or big cities and when the additional procedures is distributed in many cities.

In small cities few procedures might represent big variations in the rate: the confidence in these cities is small. However, we could solve this problem applying the *Empirical Bayes Estimator* on the cities rates as described in Section 4.2.1.1. There is also a problem with big cities: if the population is too big, the rate of procedures performed in the city is hardly affected by the anomalous hospitals. Hence, even if a hospital is fraudulent, the results of its activities is smoothed by the big population and the rate of procedures of the city does not indicate occurrence of anomaly. However, the solution for this problem is not in the scope of this work and as a future work, we plan to repeat this experiment dividing big cities in districts according to the census division to avoid the analysis over huge populations. Therefore, it is very unlikely that big cities receive high scores in our analysis.

Although the method is able to find anomalous providers that causes relevant changes in the cities rate, detecting providers related to many low-score cities is unlikely. For example, if a hospital keeps small fraudulent activities in many cities, the hospital is going to receive a low score from each city and its final score it is also going to be low. In order to detect hospitals with such behaviour, an effective approach is verifying the number of cities related to the hospital or their geographical distances from it. If the hospital is related to many cities or they are too far, it is possible that

the hospital is performing many smalls frauds. Evaluating the anomaly detection with this perspective is also a future work.

Chapter 4

Experiments

In this chapter we show the experiments performed in order to evaluate our method in the real database of the Brazilian public healthcare system. Besides describing the database, we divide our experiments into the two steps of the method: anomaly analysis and score transfer. The goal is to find the best algorithm/setup for each step and to evaluate the results.

Although all cities and hospitals have a unique identifier number in the database, we modified these identifiers in all results and examples shown in this document due to ethical and privacy issues.

4.1 Dataset

We apply our analysis to a real database from the Brazilian public healthcare system composed of five years of data, from January of 2008 until December of 2012. For each medical procedure type, we know the monthly amount of procedures performed by each hospital in the population of each city. These values represent the amount of procedures paid by the government. For the best of our knowledge, this is one of the most complete databases from public healthcare in the world. The dataset is available in the *Datasus* web page[1].

In addition to the database quality, we believe that improving the efficiency of the Brazilian Public Healthcare is a core task given the current situation. Despite being one of the countries with largest percentage of GDP spent in healthcare, its life expectancy is still low compared with other countries. According to the Bloomberg

[1] http://www2.datasus.gov.br/DATASUS/index.php

ranking, Brazil presents a bad score for healthcare efficiency[2]. The problem gets even more severe if we consider that the cost with healthcare in Brazil will increase a lot as the average age of the Brazilian population will increase fast in the next years, according to the United Nations[3].

This section presents information about the procedures selected for the analysis, about the entities (hospitals and cities) existing in the dataset, and about the features of the cities.

4.1.1 Procedures

In order to perform our experiments, we selected ten types of procedures with help of two public healthcare experts according to two criteria: seasonality and volume. Among the types of procedures for which it is not expected significant variation in the frequency nor sensibility to outbreaks, we selected ten types with great frequency in order to maximize the results' reliability. Table 4.1 presents the ten selected procedures with their frequency and the whole cost in Dollars for the period of analysis (Jan/2008-Dec/2012).

Table 4.1. List of selected procedures with the amount and the cost during the five-years period.

Procedure	Amount	Price ($)
Arteriography	12,153,229	39,538,117
Cardiovascular Surgery	12,235,183	2,234,856,750
Glaucoma Surgery	14,711,471	123,305,700
Highly Complex Orthopedic	12,072,615	253,134,000
Neurosurgery	12,082,015	265,739,775
Obstetrics	19,671,176	4,096,538,250
Oncology	12,165,030	428,805,300
Scintigraphy	12,945,508	226,140,600
Transplant	12,093,000	495,531,225
Ultrasonography	30,833,909	401,291,100
Total	**150,963,136**	**8,564,880,817**

Next, we give a brief description of each type of procedures selected for analysis:

[2]Bloomberg. Most efficient health care 2014. http://www.bloomberg.com/visual-data/best-and-worst/most-efficient-health-care-2014-countries

[3]United Nations. The consequences of the fast olding of Brazilian population. https://nacoesunidas.org/rapido-envelhecimento-da-populacao-levara-brasil-a-sofrer-pressoes-fiscais-a-partir-de-2040-diz-onu/

1. **Arteriography:** it is a class of medical imaging procedure of high complexity. The goal is to visualize blood vessels such as arteries, veins, and the heart chambers in order to locate anomalies and diseases.

2. **Cardiovascular Surgery:** involves all surgical procedures performed in the heart.

3. **Glaucoma Surgery:** involves all types of surgical procedures to restore the eyes conditions against the glaucoma disease.

4. **Highly Complex Orthopedic:** orthopedic procedures of high complexity. Most of them are related to treatments in the spine.

5. **Neurosurgery:** involves all types of surgical procedures in the nervous system.

6. **Obstetrics:** involves procedures related to pregnancy, childbirth, and the postpartum period.

7. **Oncology:** procedures related to the treatment of tumors and cancer.

8. **Scintigraphy:** form of diagnostic and medical imaging procedure that aims at the identification of tumors and diseases through radiation. In this class of procedure, radioisotopes are ingested to produce internal radiation that can be analyzed with gamma cameras.

9. **Transplant:** procedures that aims the replacement of an organ with a disease by a healthy one. It is usually performed to replace the heart, kidneys, liver, lungs, pancreas, intestine or thymus.

10. **Ultrasonography:** consists of a class of medical imaging procedures that apply high frequency waves to diagnose diseases, injuries and to monitor pregnancy conditions.

4.1.2 Entities

There are 5,566 different cities and 8,502 different hospitals in the dataset related to these ten procedures. Table 4.2 presents, for each type of procedure, the number of hospitals that performed the procedure at least once during the five years period. It also shows the number of cities for which the procedure was performed at least once in their population. The last column presents the number of different pairs of hospital and city (H, C) such that H performed at least one procedure in the population of C.

Table 4.2. Number of existing entities for each procedure: cities, hospitals and different pairs of city and hospital.

Procedure	# Cities	# Hospitals	# Pairs
Arteriography	5,007	786	18,651
Cardiovascular Surgery	5,479	374	31,742
Glaucoma Surgery	3,775	299	8,198
Highly Complex Orthopedic	5,089	1,211	20,850
Neurosurgery	5,335	680	22,894
Obstetrics	5,555	4,737	78,631
Oncology	5,474	292	22,400
Scintigraphy	5,406	517	25,758
Transplant	4,561	539	19,670
Ultrasonography	5,566	6,405	150,796

4.1.3 Cities features

In addition to the database about the procedures, we also employed a dataset with some features about the cities provided by the *Brazilian Institute of Geography and Statistics - IBGE*[4]. For each city, we have:

- the population size per year from 2008 to 2012,
- the geographic coordinates of the city center in 2010,
- the Human Development Index - *HDI* in 2010[5].

4.2 Anomaly analysis

This section describes the step of anomaly analysis of our experiments which consists of assigning an anomaly score for each city. We present the experimental setup and the results. In addition, we compare the results of the current contextual analysis with our punctual analysis described in Carvalho et al. [2015].

4.2.1 Experimental setup

The goal of our experimental setup is to investigate the best configuration for the anomaly analysis considering the algorithms, the parameter K of neighbourhood size and the normalization algorithm. As shown in Chapter 3, the contextual algorithms implemented are the *distribution-based solution* and the *KNN*. For the normalization

[4]http://www.ibge.gov.br

[5]HDI is a metric that indicates the quality of the city according to three information: income per person, education level and healthcare quality.

method we tested the trivial normalization and the *Unified method*. For the parameter K we applied four values: 4, 8, 12 and 16.

4.2.1.1 Parameter calibration

In this section we present some directives for calibrating the parameters of experiments.

Temporal granularity: as shown in Section 3.5, it is possible to control the temporal granularity with two parameters: the window size and windows sliding. As we have data about five complete years, we created one window for each year. Hence, our value for both W and S are equal to 12: the first window considers the 12 months of 2008, then we perform a sliding of 12 months to start the next window in January of 2009 and so on.

This approach comprises a complete year in each window, avoiding casual effects of seasonality. In addition, we believe that the information produced (the anomaly degree of each city in each year) becomes robust to small variations and easier to analyze for auditing processes, since that the accounting is usually organized per year.

Thus, in the step of anomaly analysis the information produced by the algorithms is the anomaly score of each city in each year from 2008 to 2012.

Cities unbalancing issue: as we compare rates of procedures in population of all cities, it is a problem dealing with unbalancing cities size. As described in Section 3.7, solving the issue of big cities rates is not in the scope of this work.

In small cities few procedures might represent big variations in the rate: the confidence in these cities is small. To tackle this problem, for each month, after computing the rate of procedure, we applied an *Empirical Bayes Estimator* as described in Marshall [1991] to smooth their variations towards the average. The lower the confidence in the cities rate, the greater the approximation to the average. Therefore, if a population of a city is too small, the method approximates the rate to the average in order to reduce the variation impact.

Contextual and behavioural features: As we deal with contextual anomalies, we apply two different types of features in the analysis: the contextual feature and the behavioural feature. The contextual feature adopted is the *HDI* weighted by the geographical distance, according to the experiments described in Appendix A. As shown in Section 3.2, the behavioural feature is the rate of procedures.

As we have the *HDI* only for the year 2010 and the cities location are static

information, the contextual neighbourhood of each city do not change from 2008 to 2012.

On the other hand, we have the monthly number of procedures and the population size of the cities during this period. Thus, the behaviour of each city, computed as rate of procedures, changes every month. In order to compute one behaviour value for each city in each window (each year), we compute the Euclidean Distance of the rate of procedures in all months within w.

4.2.1.2 Choosing the algorithms

In order to choose one algorithm for anomaly analysis and one for score normalization, we run the experiments with all the algorithms combinations fixing the value of K as 8. Then, we perform a ranking comparison and an analysis of the scores distribution.

For the ranking analysis, it is not necessary to perform score normalization as we do not consider the score values, only the positions in the rankings. We are interested in the top positions of the rankings: the practical goal of the method is to identify the most anomalous cities. Thus, if most anomalous cities in both rankings are the same, we consider that the results are similar. In addition, only the cities with great score values present relevant impact in the score transfer step.

Figure 4.1 presents the number of cities in common in the top positions of the rankings of both algorithms. The results show that for all procedures and all windows, the top 100 cities are the same in both rankings and they are also disposed in the same order. As both algorithms produce the same practical results, the algorithm to be applied on our experiments should be the one presenting the best score distribution considering that anomalies are rare.

In the score distribution analysis we verified that the distribution produced by the four combinations between the two algorithms for anomaly analysis and the two algorithms for score normalization. Figure 4.2 presents the distribution of the scores produced by the distribution-based solution and normalized by the trivial normalization method. In Figure 4.3 we show the score distribution produced by the distribution-based solution normalized by the unified method. Figures 4.4 and 4.5 present the score distribution for the *KNN* results normalized by the trivial method and unified method, respectively. In all the figures, the y-axis is in the log scale and the values are the scores of all cities in all windows.

As we apply an unsupervised approach, our assumption is that anomalies are rare instances presenting isolated behaviour. According to this assumption, we consider that the results provided by the distribution-based solution are not good. In the execution

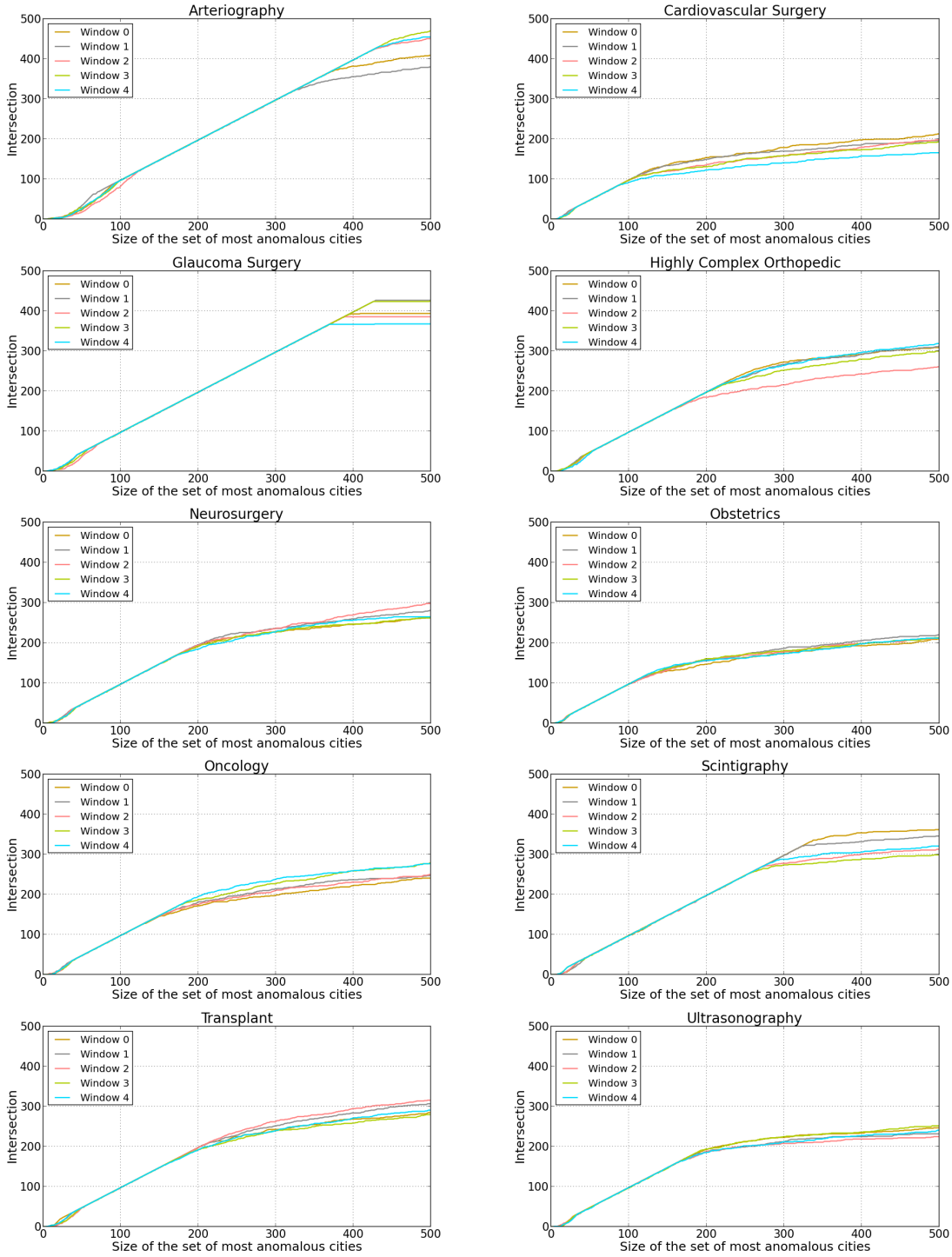

Figure 4.1. Number of cities that occur in the top positions of both rankings.

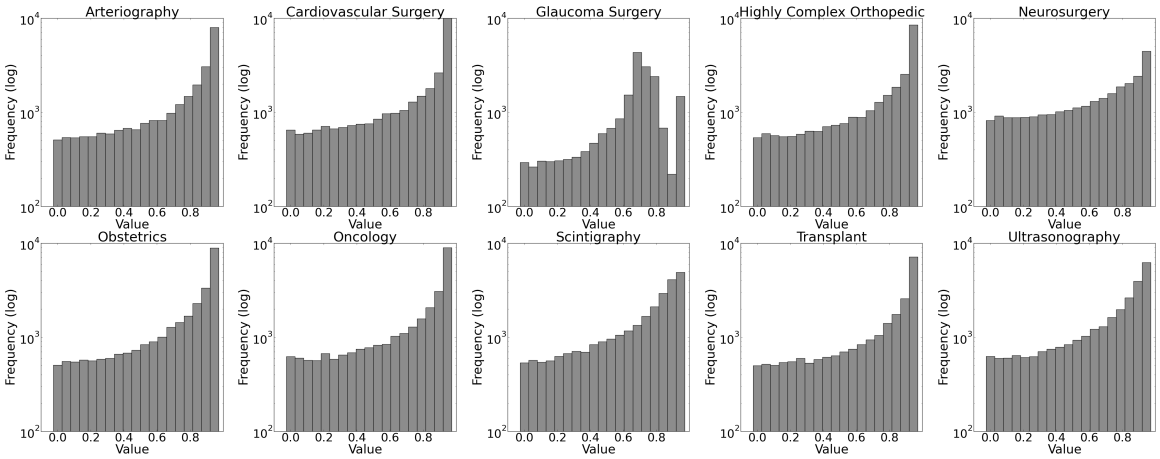

Figure 4.2. Score distribution produced with the distribution-based solution and the trivial normalization.

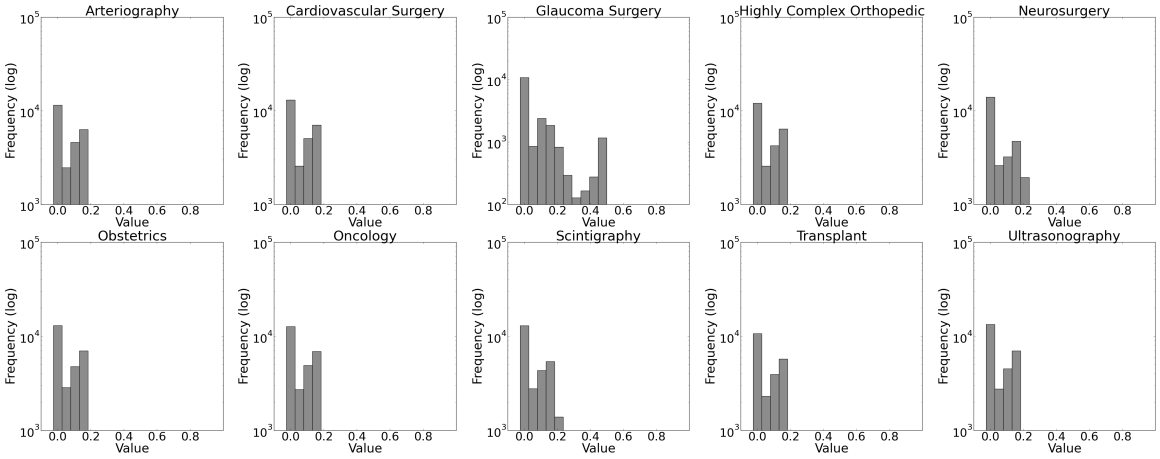

Figure 4.3. Score distribution produced with the distribution-based solution and normalized with the Unified method.

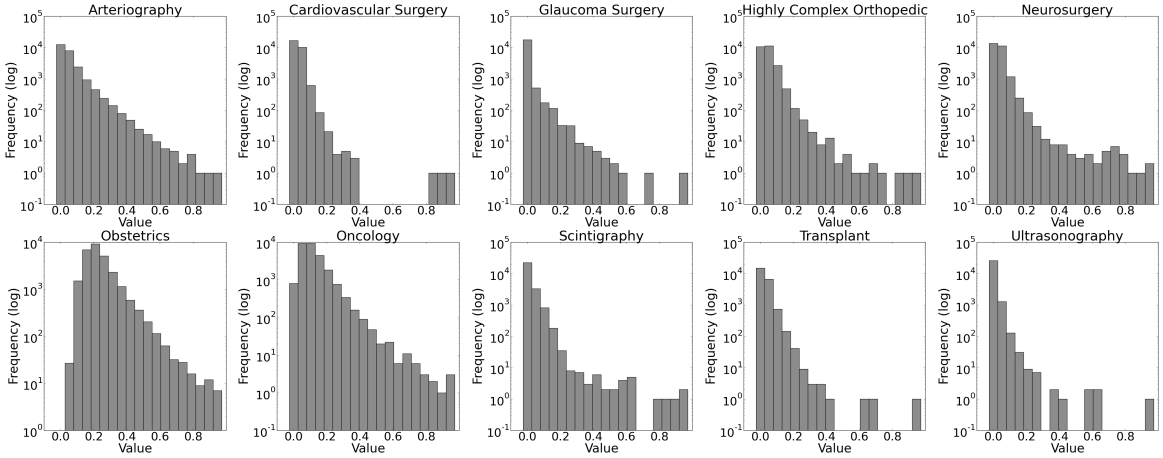

Figure 4.4. Score distribution produced with the KNN and the trivial normalization.

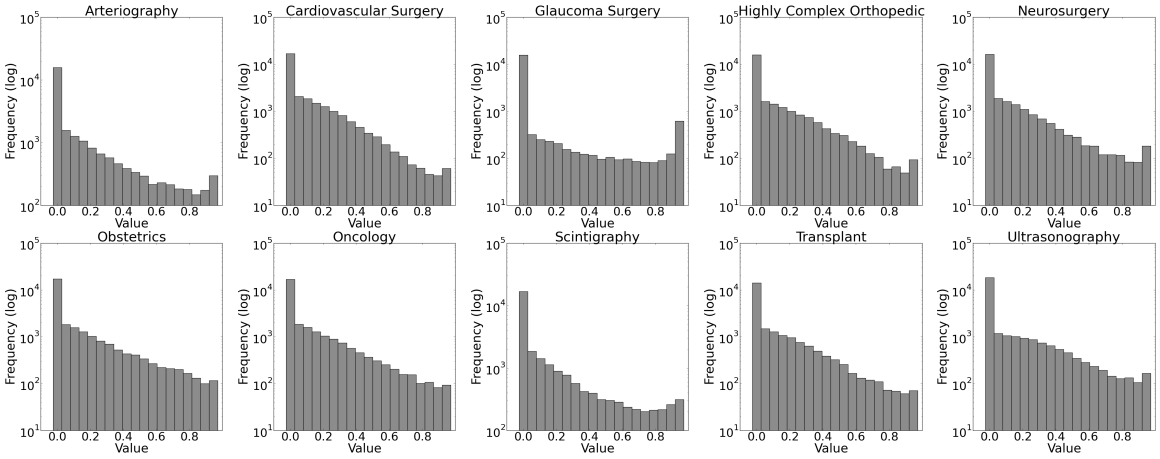

Figure 4.5. Score distribution produced with the KNN and normalized with the Unified method.

of the distribution-based solution normalized with the trivial normalization, most of the cities were assigned the maximum score. Besides the problem of precision that it would cause to find the anomalous hospitals, it does not conform with the definition of anomaly. The problem with the distribution-based solution normalized with the unified method is that no city was considered anomalous.

On the other hand, the score distribution provided by the KNN is more suitable to the goal of identifying few isolated instances. The difference between the KNN results normalized with the trivial solution and the unified method is that the trivial solution produced less anomalies then the unified method.

Figure 4.6 compares the cumulative distribution function - CDF of the score produced by the KNN and normalized with both algorithms. We observe that the curves produced with the trivial solution present a fast increasing and a long tail, indicating that just a few instances receive high scores. The curves produced with the unified method increase slowly and constantly as the scores values are more equally distributed.

Adopting a conservative decision, we consider that the normalization with the trivial solution is better: just a few anomalies while almost all cities are regular. In addition, if many cities receive high scores, it causes a loss of precision in the score transfer step, as many hospitals would also receive high scores.

Therefore, although the ranking produced with both algorithms are the same in the top positions, we adopt the KNN algorithm and the trivial normalization for our experiments due to the score distribution.

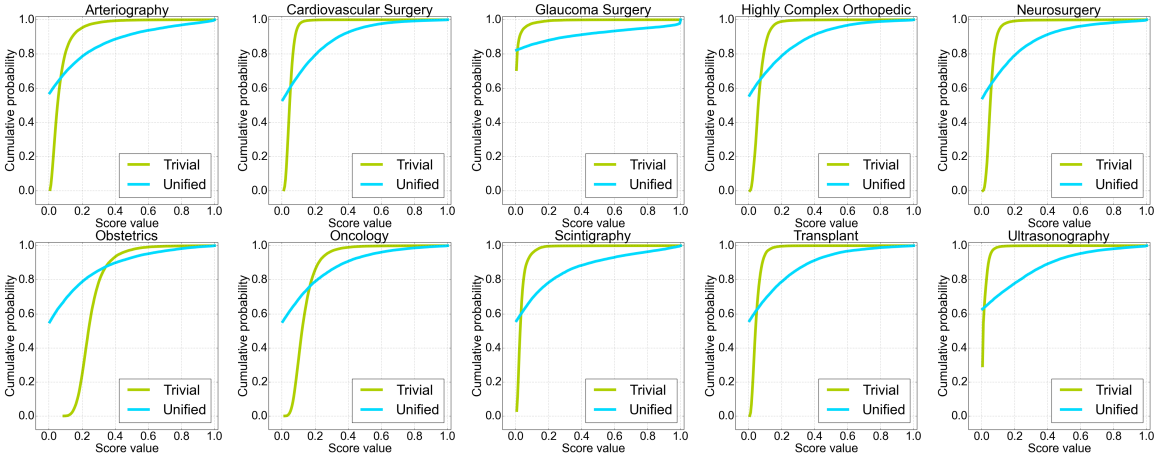

Figure 4.6. Cumulative probability of the scores produced with the KNN and normalized with both algorithms.

4.2.1.3 Choosing K

The parameter K indicates the number of contextual neighbours to be considered in the analysis. If K is too small, the algorithm becomes vulnerable to small variations. If K is too large, it may suffer from undesirable smoothing effects. We present the experiments to choose the value of K among four values: 4, 8, 12 and 16.

The experiment consists of executing the anomaly analysis for all procedures with these four values of K. Then we compare the produced rankings with the *Kendall Tau* metric Kendall [1938]. This metric computes the rate of pairs of items with different relative position in two rankings. Thus, the higher the distance, the more pairs of elements (A, B) exists such that A is more anomalous than B in one ranking while B is more anomalous than A in the other ranking.

Table 4.3 shows the *Kendall Tau* distances for the ten procedures. We note that, although each pair of ranking is compared in each one of the five windows, we show only the average distance. We verified statistically that there is no significant difference among the values of the five windows: the average variance is equal to 0.02. The greatest variance of 0.038 occurs when we compared the values of $K = 4$ and $K = 16$ for the Obstetrics procedure.

According to the results, there is little difference between the rankings produced by the four values of K. As we performed the previous experiments with $K = 8$, we verified that this value produced good results for the cities analysis. Thus, we believe that it is a good value to be adopted in the experiments.

Table 4.3. Kendall Tau distances between the rankings generated with different values for K.

K values	Art.	C.S.	Glauc.	H.C.O	Neuro.	Obst.	Onco.	Scint.	Trans.	Ultra.
4 x 8	0.091	0.076	0.109	0.079	0.091	0.122	0.091	0.079	0.099	0.115
4 x 12	0.109	0.090	0.143	0.093	0.108	0.143	0.107	0.094	0.117	0.137
4 x 16	0.118	0.096	0.160	0.099	0.117	0.153	0.115	0.102	0.125	0.150
8 x 12	0.060	0.048	0.076	0.050	0.059	0.076	0.057	0.050	0.063	0.075
8 x 16	0.077	0.059	0.106	0.062	0.075	0.094	0.072	0.066	0.079	0.099
12 x 16	0.046	0.035	0.058	0.036	0.046	0.055	0.043	0.039	0.047	0.059

4.2.2 Results and evaluation

So far, we defined that:

- the algorithm to be applied is the *KNN* with K equal to 8;
- the window size W and the sliding S are both set to 12;
- the normalization is performed for the whole period with the trivial approach of dividing the scores by the largest value;
- and that the Bayesian rates are applied as behavioural feature to reduce the impact of variations in small cities.

Next, we present the results produced with this setup through visual analysis and manual labelling.

Visual analysis: as the normalization is performed considering all five windows, the maximum score of each procedure occurs in one city in one window. Table 4.4 presents the city with largest score for each type or procedure and also shows in which window the maximum score occurred.

Table 4.4. Cities with largest score for each procedure type.

Procedure	City	Window	Year
Arteriography	2536	3	2010
Cardiovascular Surgery	2536	3	2010
Glaucoma Surgery	4240	4	2011
Highly Complex Orthopedic	3515	3	2010
Neurosurgery	4131	5	2012
Obstetrics	2193	3	2010
Oncology	2536	3	2010
Scintigraphy	1198	5	2012
Transplant	5030	2	2009
Ultrasonography	4296	5	2012

Next we show the rate and the amount of procedures in the city that received the maximum score for each procedure type. In each of the ten plots, the x-axis indicates the month while the y-axis indicates the rate of procedures. The red line shows the

rate in the target city and the numbers close to the red line indicate the amount of procedures in the city. The gray line represents the average rate in the country considering only the cities in which the procedure was performed at least once.

Figure 4.7 shows the time series of rate in City 2536, which received the highest score for the procedure of Arteriography. As shown in Table 4.4, the highest score was assigned in window 3 (year 2010). For example, in the month 07/2010 the number of procedures in the city was 213, which represents a rate of 115 procedures per 100,000 people. The country average rate in the same month was about 4 procedures per 100,000 people. In windows 1, 2, 4 and 5 this city received scores 0.45, 0.40, 0.72 and 0.55, respectively, being in the ranks 2, 3, 5 and 20 of these windows.

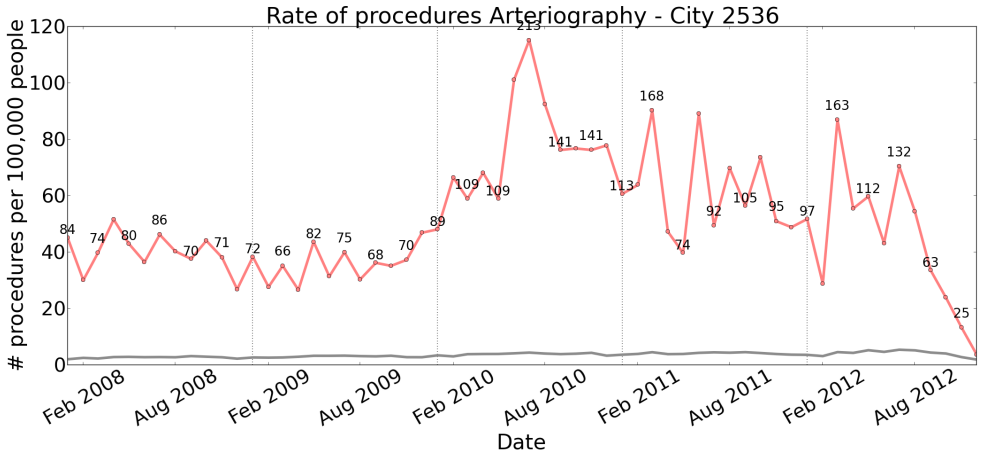

Figure 4.7. Rate and number of procedures in the most anomalous city according to Arteriography procedure.

City 2536 also received the maximum score for the Cardiovascular Surgery procedure. Figure 4.8 presents its behaviour. Although the maximum score of City 2536 occurs in window 3, this city is also in the first rank of all windows.

For the procedure of Glaucoma Surgery, the maximum score occurs in City 4240 in window 4. As shown in Figure 4.9, although the number of procedures in this city is null in most of months, in November of 2011, the rate of procedures in this city is 53,816 per 100,000 people, meaning that according to the records, more than 50% of its population performed Glaucoma Surgery. In this same month, the country average was about 200 procedures per 100,000 people.

City 3515 received high scores in all windows for the procedure type of Highly Complex Orthopedic. The score 1.0 occurs in the year 2010. Figure 4.10 shows that, during this year, the country average was about 1.7 procedures per 100,000 people whereas the rate in the city was more than 21 procedures per 100,000 people.

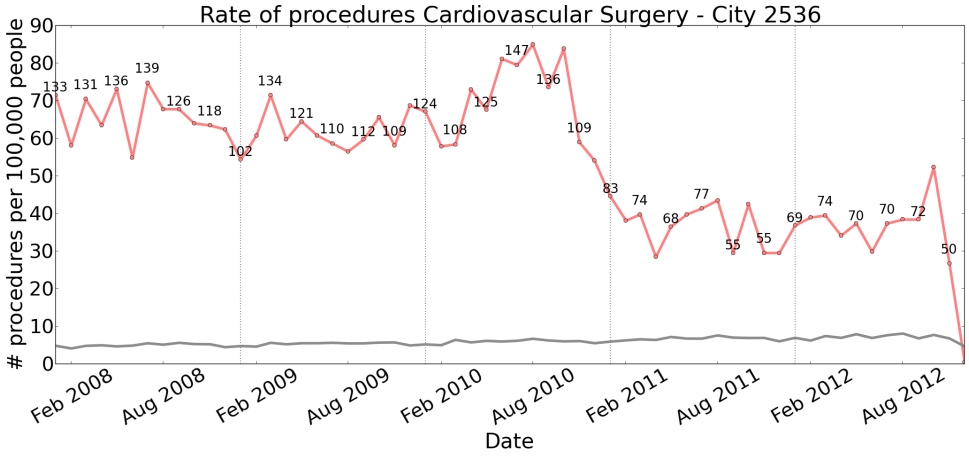

Figure 4.8. Rate and number of procedures in the most anomalous city according to Cardiovascular Surgery procedure.

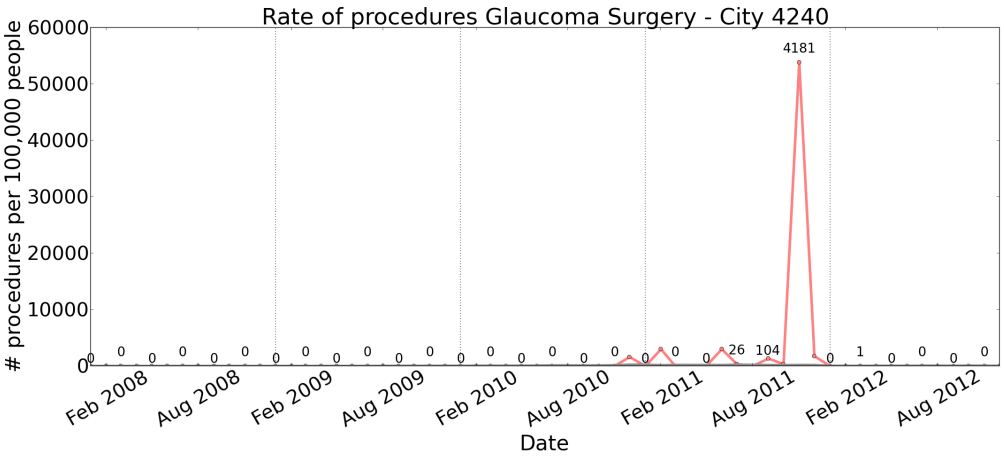

Figure 4.9. Rate and number of procedures in the most anomalous city according to Glaucoma Surgery procedure.

For the Neurosurgery procedure, the most anomalous case was in 2012 in city 4131, when the rate in the city was about 12 times larger than the average, as shown in Figure 4.11. This same city was also anomalous in the other years.

Figure 4.12 shows the time series of city 2193, which was the most anomalous for Obstetrics procedures in 2010. This city is also anomalous in the other windows. The number of procedures is significant, i.e., the rate is not susceptible to small variations. Therefore, as the rate is above the country average during almost the whole period, city 2193 presents anomalous behaviour according to Obstetrics claims.

For the Oncology procedure, city 2536, which also presented the highest scores for Arteriography and Cardiovascular Surgery, was the most anomalous. The maximum

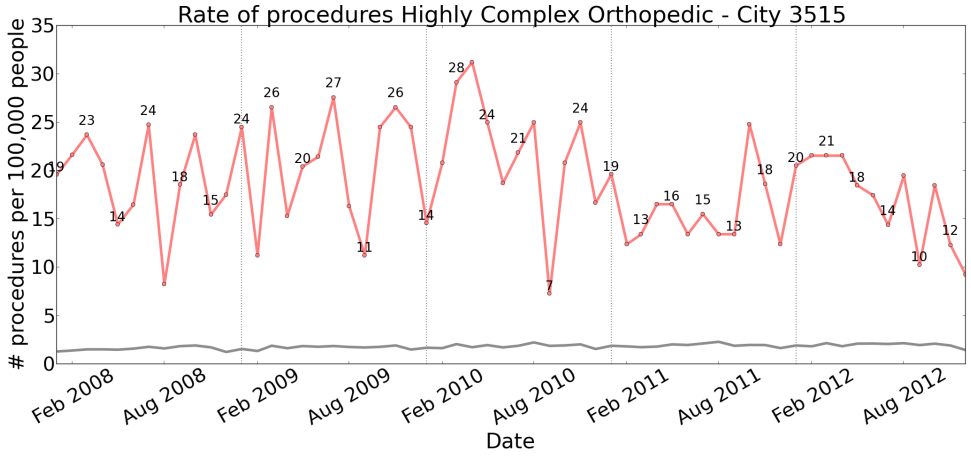

Figure 4.10. Rate and number of procedures in the most anomalous city according to Highly Complex Orthopedic procedure.

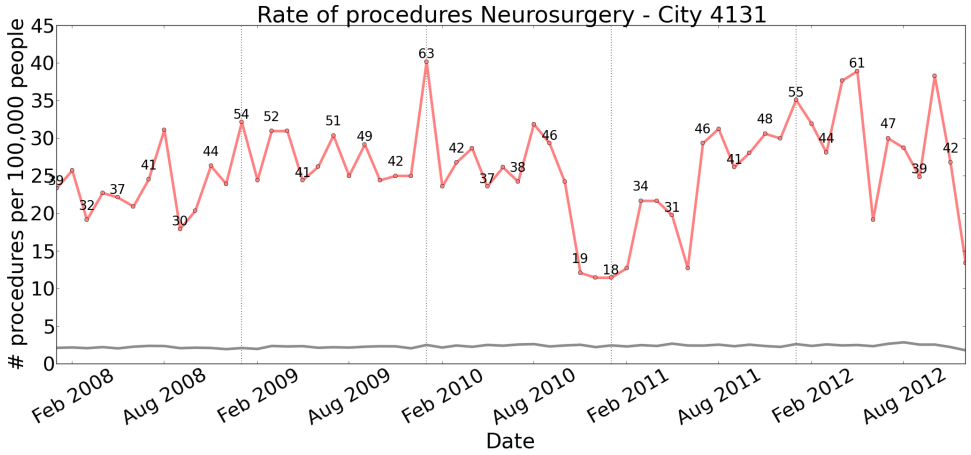

Figure 4.11. Rate and number of procedures in the most anomalous city according to Neurosurgery procedure.

score occurs in 2010, when its average rate was about 5 times the country average rate. Figure 4.13 presents the behaviour of city 2536 according to Oncology procedure.

For the Scintigraphy procedure, city 1198, shown in Figure 4.14, received score equal to 1.0 in 2012. The month with the highest rate was September/2011, when the rate in the city was 725 procedures per 100,000 people, whereas the country average was 14 procedures per 100,000 people.

For the Transplant procedure, the highest score occurred in 2009 in City 5030, shown in Figure 4.15. In June/2009 the rate in this city was almost 25 times the country average rate.

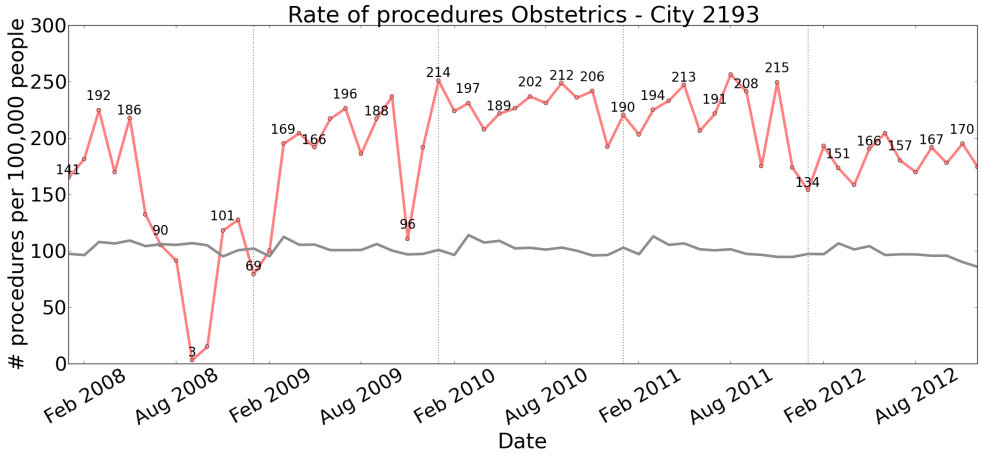

Figure 4.12. Rate and number of procedures in the most anomalous city according to Obstetrics procedure.

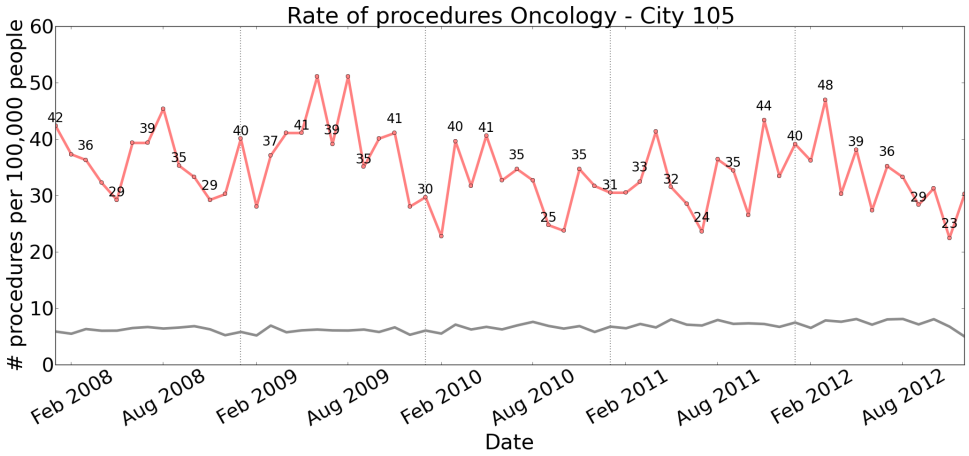

Figure 4.13. Rate and number of procedures in the most anomalous city according to Oncology procedure.

Finally, we show the city with highest score for Ultrasonography, which occurred in 2012 in city 4296. According to Figure 4.16 it is possible to see that the monthly number of procedures was smaller than 15 until the end of 2011. However, it presented anomalous behaviour in the end of 2011 and beginning of 2012: the number of procedures increased to about 2,000. During this period, the rate of procedures was about 20,000 per 100,000 people.

Manual evaluation: after showing the most anomalous city for each procedure, we present now a manual evaluation of the results performed by an expert in healthcare management. As we are dealing with a non-labeled dataset, the evaluation is a

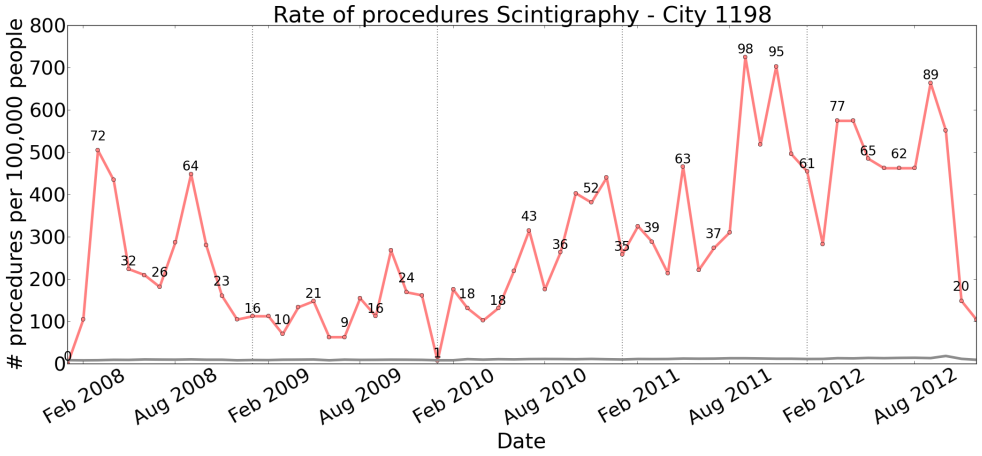

Figure 4.14. Rate and number of procedures in the most anomalous city according to Scintigraphy procedure.

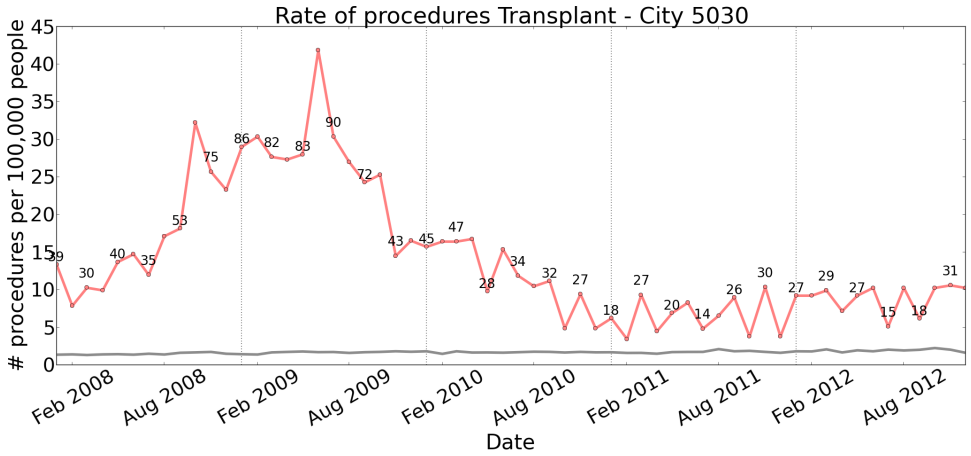

Figure 4.15. Rate and number of procedures in the most anomalous city according to Transplant procedure.

challenging task. Although it is not the ideal strategy for the evaluation, the manual evaluation gives us a notion of the results quality with a feasible cost. Our methodology for the manual evaluation is described next.

For each procedure type, we choose one window and evaluate the top 10 cities of the ranking. We assume that the solution quality would not vary so much among the windows of the same procedure type. The window evaluated of each procedure followed the alphabetical and numerical order as shown next. Arteriography: window 1 (year 2008); Cardiovascular Surgery: window 2; Glaucoma Surgery: window 3; Highly Complex Orthopedic: window 4; Neurosurgery: window 5 (year 2012); Obstetrics: window 1; Oncology: window 2; Scintigraphy: window 3; Transplant: window 4;

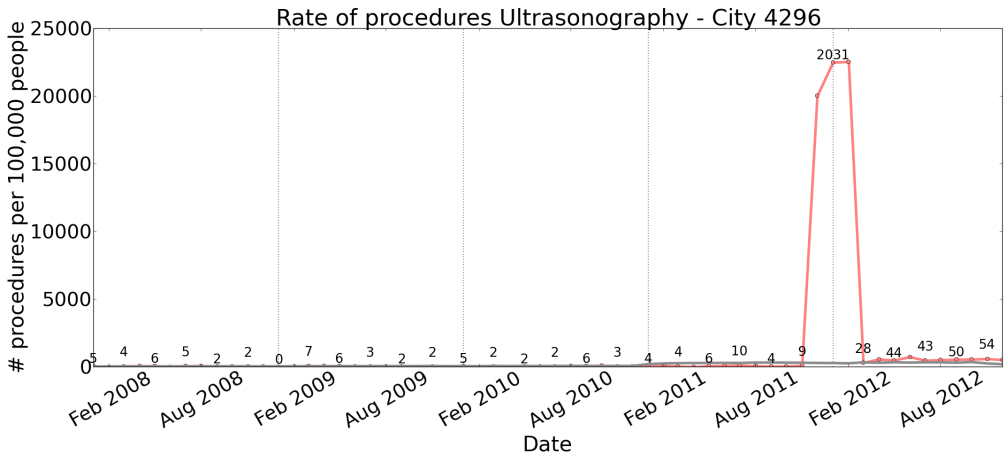

Figure 4.16. Rate and number of procedures in the most anomalous city according to Ultrasonography procedure.

Ultrasonography: window 5.

Although this labeling depends on subjective interpretation, we usually consider that an anomalous city presents both significant difference from the country average and a representative number of procedures. According to our judgment, if the amount of procedures is smaller than 20, it is likely that the city is small and fluctuations are acceptable.

Table 4.5 shows the results: for each position and each procedure, we indicate if the city is anomalous (letter A and red color) or regular (letter R and the blue color). We conclude that the majority of the top 10 ranked cities are occupied by true anomalies: about 80%. For all procedures, the top 4 positions are anomalous cities.

There are two types of ranking configurations that we consider correct for the top 10 cities: when all of them are anomalies and when there is no other anomaly from the occurrence of the first case of non-anomalous city. The only procedure that do not conform to this expected behaviour is the Obstetric Surgery, in which the fifth and sixth positions are occupied by non-anomalous cities whereas the remaining positions are anomalous cities. Therefore, for the top 10 positions of the ranking, we consider that we were able to achieve good rates of true positives.

As it is impractical to analyze manually all the cities, in order to analyze the occurrence of false negatives, we select at random 30 cities among the remaining positions of the ranking (from the eleventh position) of each procedure and manually label them. In this analysis we did not found any case of anomaly. Although it is not a reliable strategy to measure the false negative rate, this analysis indicates that it is not easy to find an anomalous city at random: from 300 observations (30 from each of the

Table 4.5. Results of the manual evaluation of the top ten cities in the ranking of each procedure type.

Position	Art.	C.S.	Glau.	H.C.O	Neur.	Obs.	Onc.	Sci.	Tra.	Ult.
1st	A	A	A	A	A	A	A	A	A	A
2nd	A	A	A	A	A	A	A	A	A	A
3rd	A	A	A	A	A	A	A	A	A	A
4th	A	A	A	A	A	A	A	A	A	A
5th	A	A	A	A	A	R	A	A	A	A
6th	R	A	A	A	A	R	R	A	A	A
7th	R	R	A	A	A	A	R	A	A	A
8th	R	R	A	A	A	A	R	A	A	A
9th	R	R	A	A	A	A	R	A	R	A
10th	R	R	A	R	A	A	R	A	R	A

ten procedures), none of them was anomalous.

In order to show that the rate of procedures tends to be anomalous in the top ranked cities whereas it is regular in the remaining positions, we show in Figure 4.17 the time series of average rate for three ranges of the ranking:

1. The top 10 positions.
2. From 11^{th} to 100^{th} position.
3. From 100^{th} to the end of the ranking.

According to the result, cities in the first position tend to present high rates of procedures occurrence. Cities from positions 11^{th} to 100^{th} also present high rates but lower than the top ranked cities. Finally, the remaining cities present low rates of procedures occurrence.

4.2.3 Comparison between punctual and contextual anomalies

Recently, we published the work in Carvalho et al. [2015] describing the progress of the method and some results. The core differences between the current work and this previous work are shown in Table 4.6.

We consider that the main improvement of this current work compared to the published work is the change from punctual anomalies to contextual anomalies. We believe that this change represents a significant improvement in the detection of anomalous cities, which eventually also improves the detection of anomalous hospitals. The main reason for this change is the fact that it is much more consistent to compare rates of procedures of cities with similar profile, especially if we deal with cities of a huge country, such as Brazil.

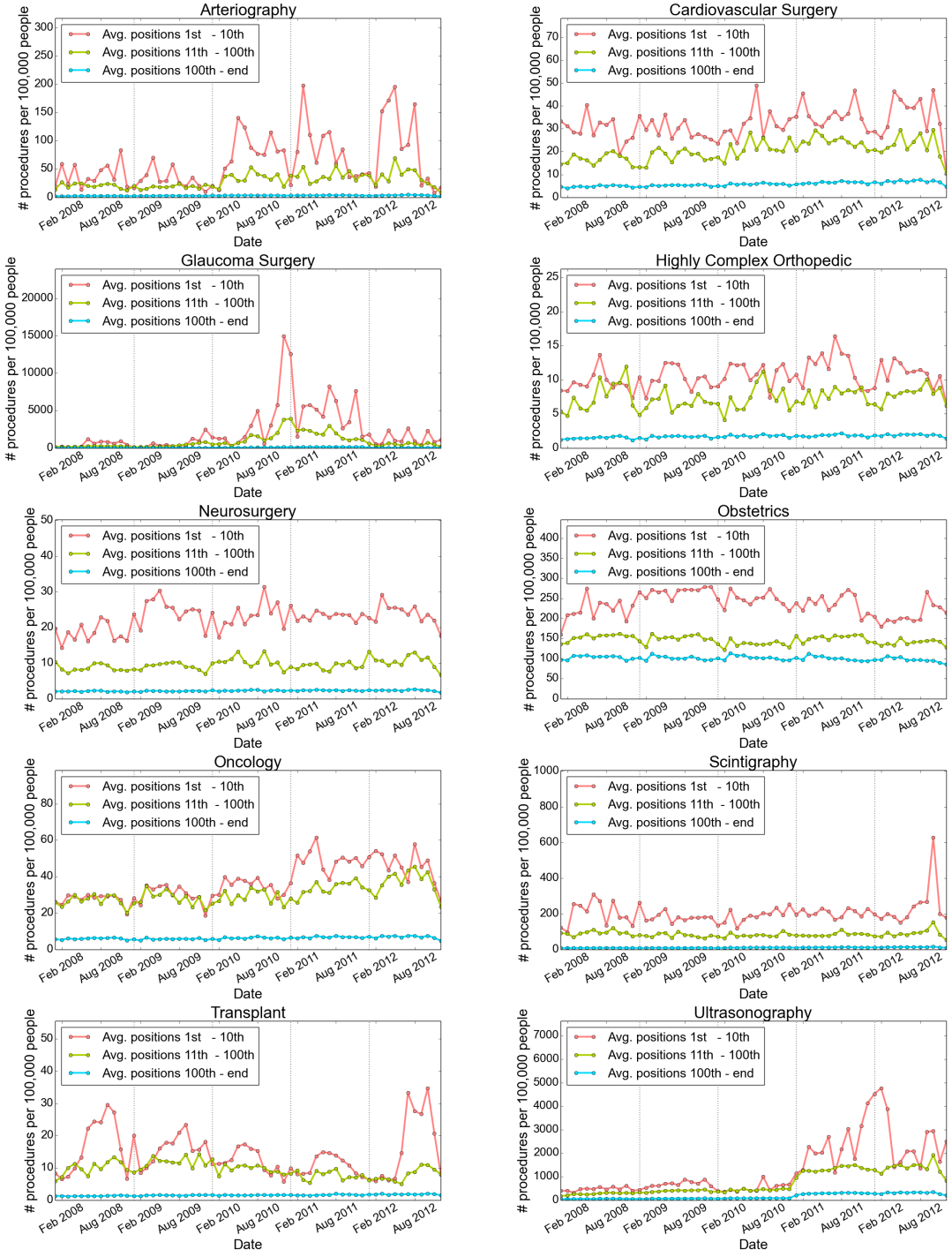

Figure 4.17. Average rate of procedures for three ranges of the ranking of each procedure.

As the work in Carvalho et al. [2015] deals with punctual anomalies, the rate of procedures in the cities is both the contextual and behavioural feature. It means that the cities are compared with cities most similar rates of procedures. We perform a

Table 4.6. Core differences between the current work and the published version.

Aspect	Carvalho et al. [2015]	Current work
Anomaly analysis	Punctual anomalies	Contextual anomalies
Database	Three procedures types (total cost of $3.5 billion)	Ten procedures types (total cost of $8.5 billion)

comparison between the results produced by both works. Although in Carvalho et al. [2015] we only use 3 procedure types, we repeated the analysis with punctual anomalies for the other seven procedures in order to perform this comparison.

4.2.3.1 Neighbourhood intersection

First of all, we want to verify if the neighbourhood of the cities are similar in the contextual and punctual analysis. For each procedure and for each window, we measure the number of common cities in the neighbourhood of each city. As we use $K = 8$, the intersection can range from 0 to 8.

Figure 4.18 shows the distribution of the intersection size in the neighbourhood of the punctual and contextual anomalies. The results indicate that, for the vast majority of the cities, the intersection between the punctual and contextual neighbourhood are empty.

Therefore, punctual and contextual anomalies produce very different neighbourhood for the cities. It means that, for each city C, the 8 cities with most similar rates of procedures of C are not the same cities with similar HDI and location.

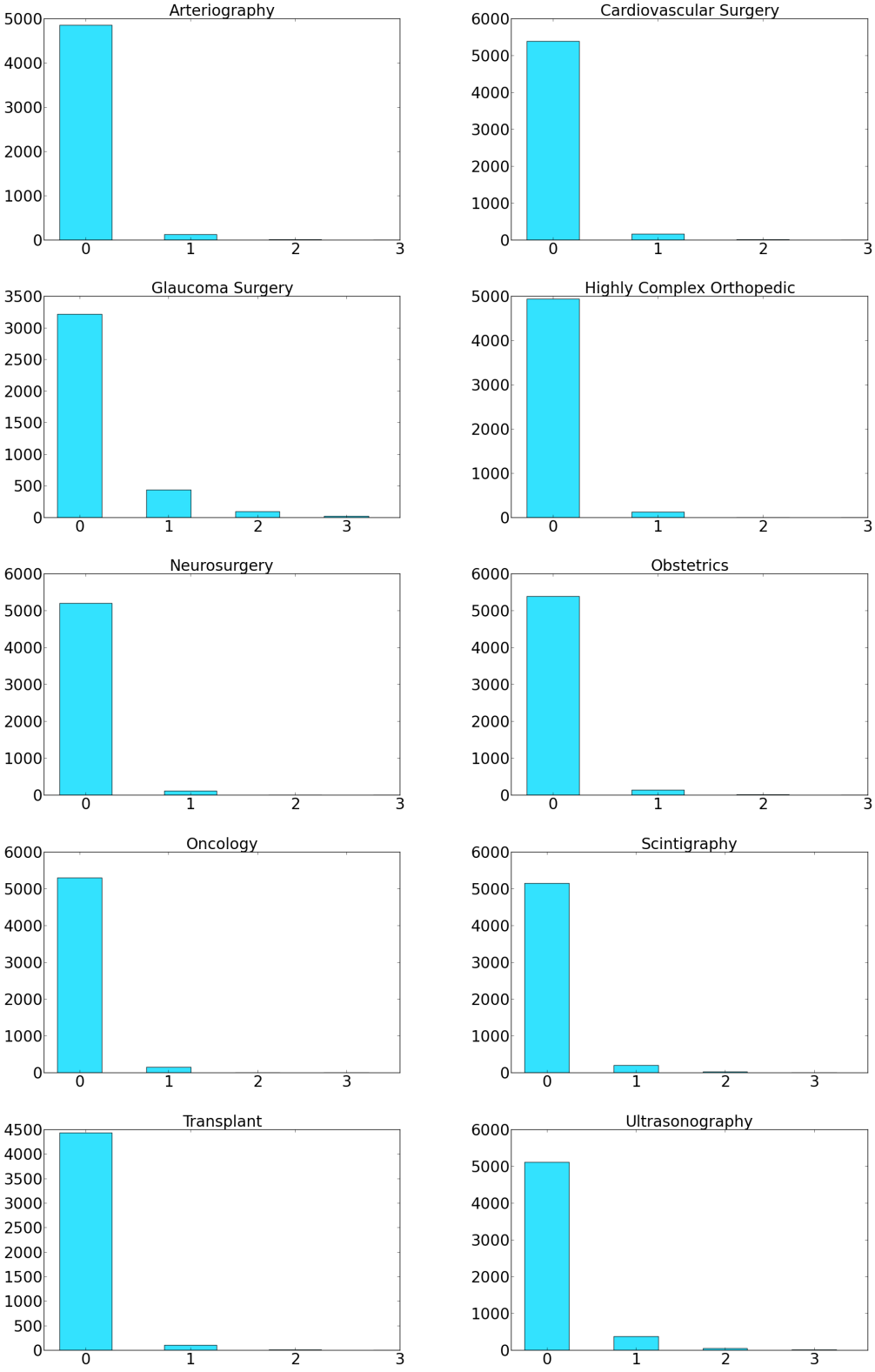

Figure 4.18. Distribution of the intersection of the cities neighbourhood considering punctual and contextual anomalies for all procedures and windows.

4.2.3.2 Contextual difference in the neighbourhood

Our goal in this analysis is to compare the punctual and contextual results according to the HDI and location distances between each city and its neighbours. As these features compose the set of contextual features in the contextual analysis, we expect that the difference between cities and neighbours are larger in the punctual analysis.

HDI: in Figure 4.19 we show the distribution of HDI difference between the cities and their neighbours in all windows of each procedure type. As expected, the HDI between punctual neighbours is less similar than between contextual neighbours. Thus, in the contextual analysis, we compare the cities with neighbours that are more similar according to education, healthcare and economics aspects.

Location: Figure 4.20 shows the distribution of the differences between cities and neighbours according to the geographic distance. The results show a huge difference between the geographic distance in the cities neighbourhood of both analysis. Therefore, the punctual neighbourhood is composed of cities that can be physically distant. From the healthcare perspective, it is not consistent to compare cities that are so far from each other.

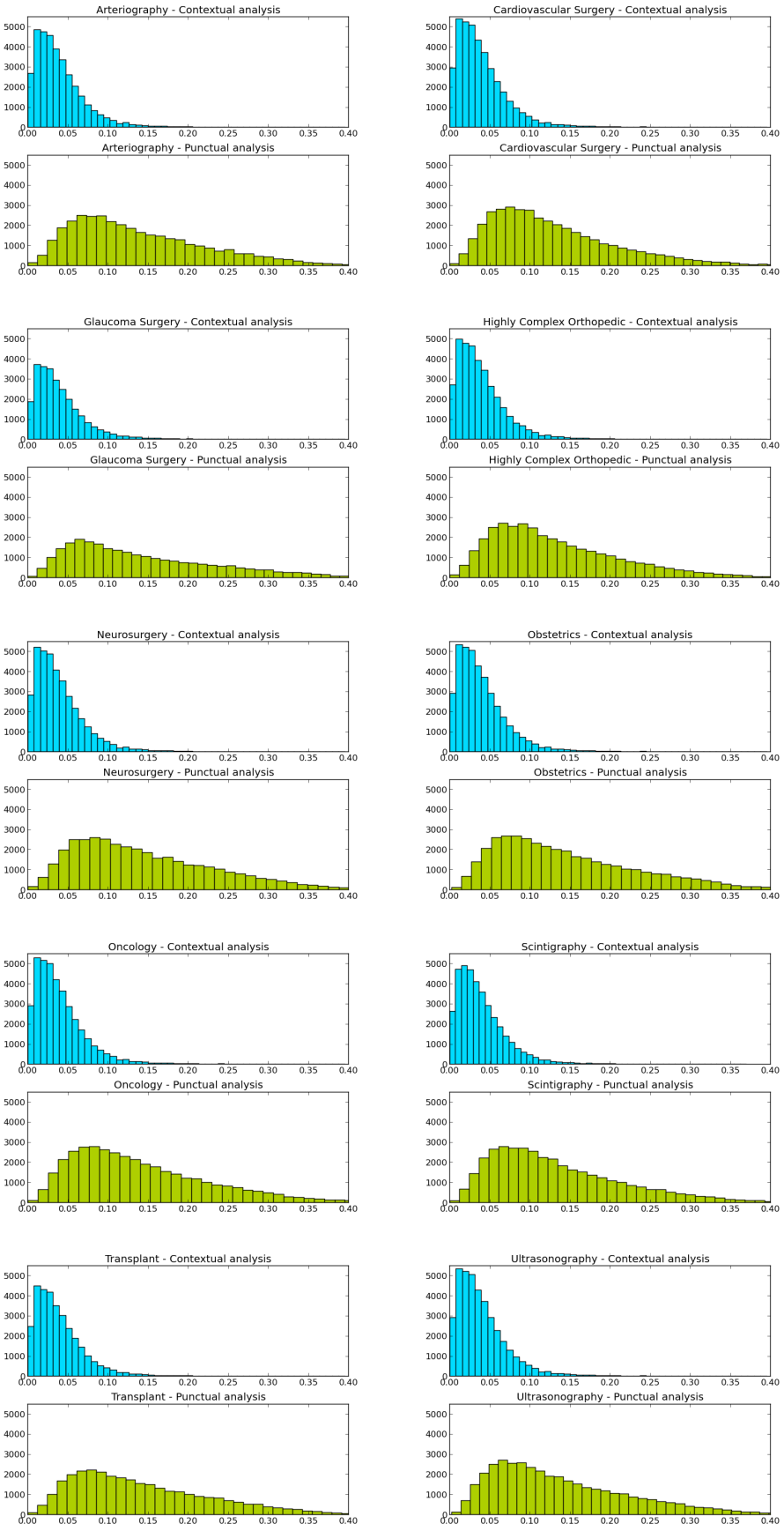

Figure 4.19. Distribution of the *HDI* distance between cities and their neighbourhood of the contextual and punctual analysis..

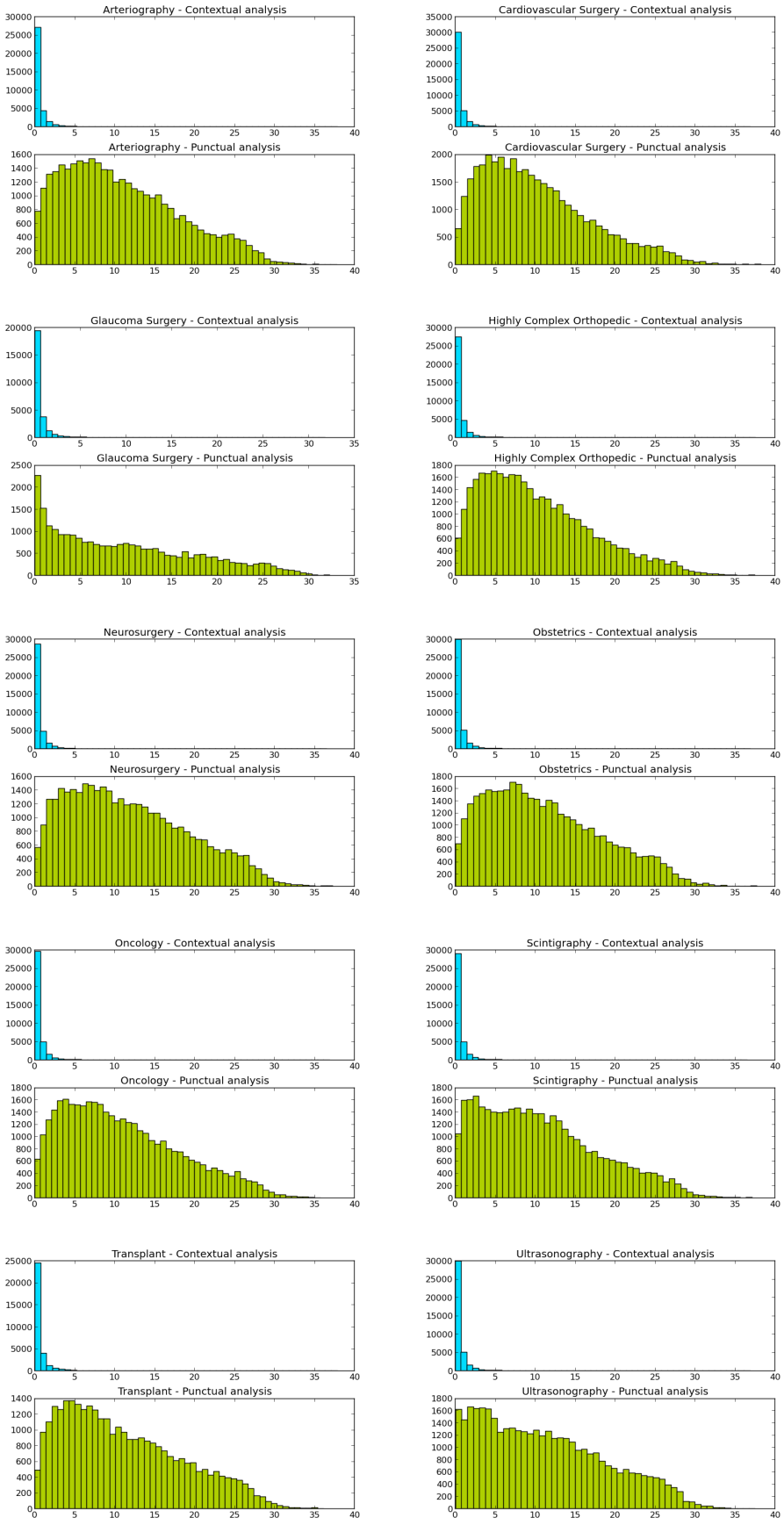

Figure 4.20. Distribution of the geographic distance between cities and their neighbourhood of the contextual and punctual analysis..

4.2.3.3 Behavioral difference in the neighbourhood

We also compared the behavioural distances between the cities and their neighbours, which is expressed by the Euclidean distance in the rate of procedures. The behavioral distances in the punctual analysis are minimal, once that the neighbours of the cities are those cities with smaller behavioural distances. In this comparison, we want to check whether the behavioural distances of the contextual analysis are significantly greater than the behavioural distances of the punctual analysis.

Figure 4.21 shows the distribution of the behavioural distance between the cities and their 8 neighbours in all windows of each procedure. It is interesting to observe that the X scale of the plots depends on the rate of the procedures: frequent procedures present large values whereas the opposite situation is observed in rare procedures.

Although the distances are smaller in the punctual analysis, we consider that the contextual analysis also produces small behavioural distances.

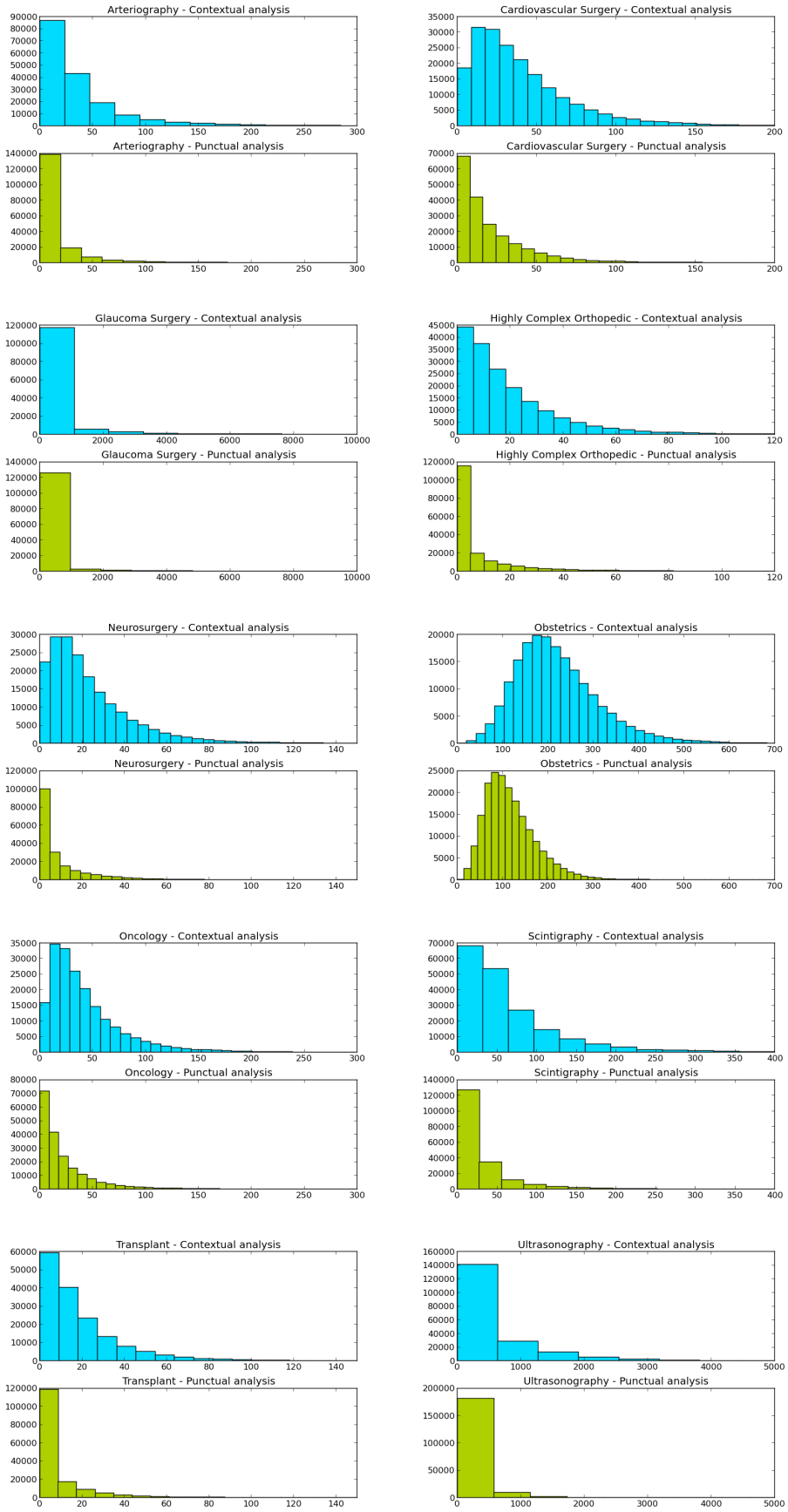

Figure 4.21. Distribution of the behavioural distance between cities and their neighbourhood of the contextual and punctual analysis.

4.2.3.4 Ranking difference

We have shown so far that punctual and contextual neighbourhood are very different: besides the rare intersection, the profile of the neighbourhood is very different according to contextual and behavioural features. Next we discuss the impact of these difference in the ranking of the cities.

Figure 4.22 shows the number of cities that appear in the top positions of both rankings. The set of top positions is ranged from 1 to 500. The dotted line show the expected behaviour if the two rankings were identical. The results show that for all procedures there is a significant difference between the rankings, as expected.

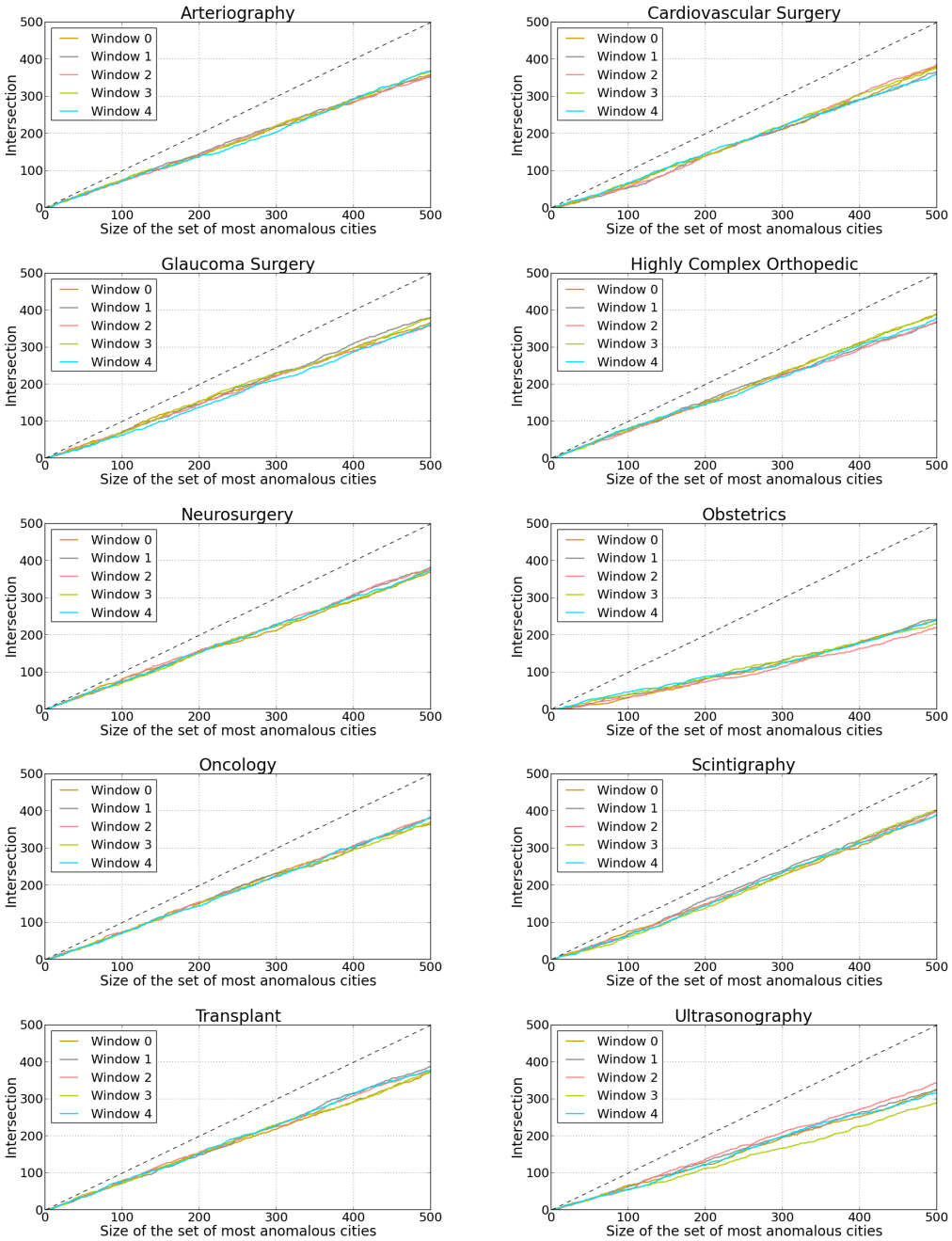

Figure 4.22. Number of cities that occur in the top positions of the punctual and contextual analysis.

4.2.3.5 Examples

Next we present some examples of cities that were ranked better with the contextual analysis than with the punctual analysis.

In the year 2009, city 3666 occupied the second position of the punctual ranking.

However, as shown in Figure 4.23, the city is not anomalous in that year (shaded area), when it is in position 68 in the contextual analysis.

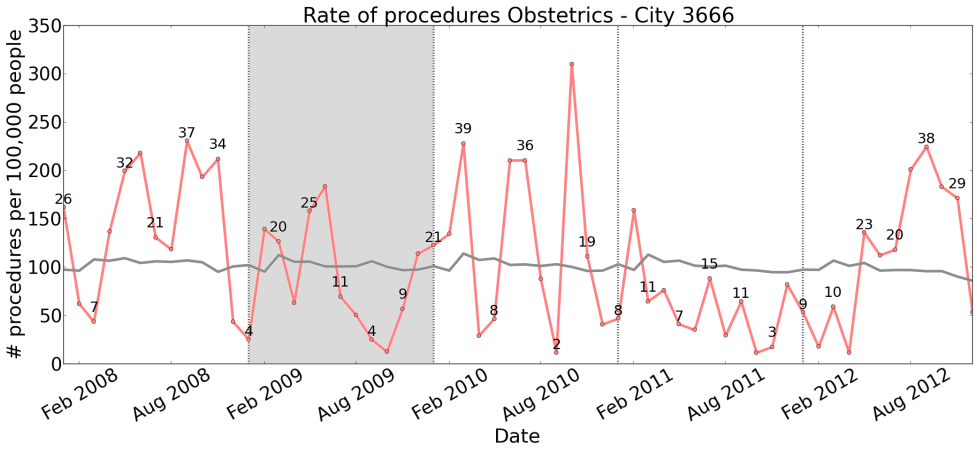

Figure 4.23. Rate and number of procedures in city 3666.

City 4599 is the city in the second position of the ranking produced with the contextual analysis in 2009. Figure 4.24 shows that this city is anomalous since its rate is high and the number of procedures is significant. However, according to the punctual analysis, this city is not anomalous occupying position 1228 of the ranking of the same year.

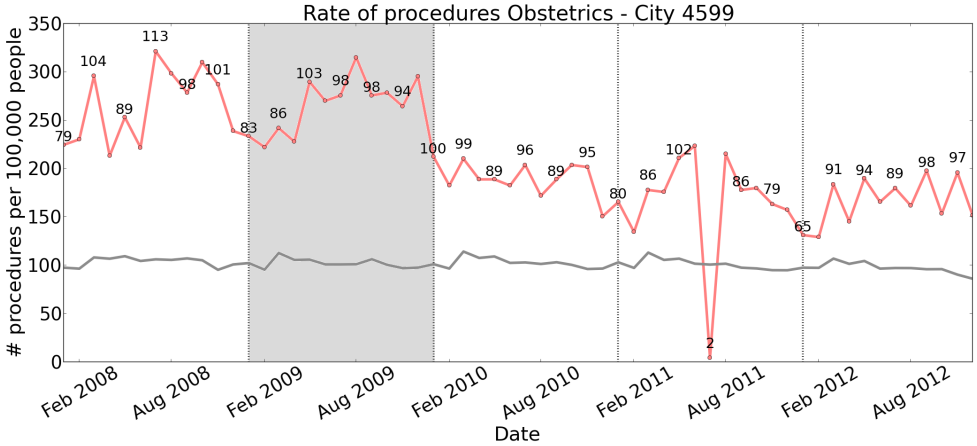

Figure 4.24. Rate and number of procedures in city 4599.

Thus, there is a significant difference between the neighbourhood produced with punctual and contextual analysis.

We have shown that the neighbourhood produced with the contextual analysis is much more consistent than the punctual neighbourhood. According to the behavioural

distance, we have shown that the difference between contextual pairs and punctual pairs of neighbours is not so significant. However, pairs of contextual neighbours are significantly more closer than punctual neighbours according to the geographic location, economics, educational and healthcare aspects. In addition, we presented some examples of cities that were properly ranked by the contextual analysis.

As the next step of score transfer depends on the anomaly analysis, the employment of contextual analysis also improves the evaluation of the hospitals.

4.3 Score transfer

The step of score transfer consists of assigning a score for each hospital in each window considering

- the score received by each city in each window, and
- the amount of procedures performed by each hospital in the population of each city in each window.

In this section we detail the experimental setup to choose the algorithms parameters for the score transfer and present the results.

4.3.1 Experimental setup

As described in Section 3.4, we implemented two algorithms for score transfer: the linear combination and the genetic algorithm. In this section we detail how we calibrate and compare the algorithms in order to produce the better ranking and score assignment to the hospitals.

4.3.1.1 Linear transfer

In Section 3.4.2 we show two implementations of the linear transfer algorithm: *simple linear transfer* and *proportional linear transfer*. We compared both options considering the aspects involved in healthcare (qualitative analysis) and considering the results (empirical analysis).

Qualitative analysis: in the simple linear transfer, for each pair (H, C) of hospital and city, the score is weighted by the absolute amount of procedures performed by H in the population of C. For instance, even if C is a very anomalous city and H is the only hospital related to it, C might transfer low score to H if C is a small city

with few procedures. Hence, according to this strategy, the key aspect is the absolute number of procedures: anomalies in pairs linked by many procedures are more relevant than in pairs linked by few procedures.

In the proportional linear transfer, the score is weighted by the fraction of procedures in C performed by H. It focus on the responsibility of the hospitals: if H is the only hospital related to an anomalous city C, the score transferred from C to H is going to be high, regardless of the absolute amount of procedures.

According to an analysis with help of two healthcare experts, we concluded that the proportional linear transfer is more appropriate because our method would hardly detect an anomalous big city. As we are able to detect small and medium cities, it is better to point as anomalous the hospitals that are the top responsible for extreme anomalous cities than hospitals related to anomalous big cities.

Empirical analysis: in our empirical analysis we compared the rankings and the score distribution produced by both approaches of linear transfer.

First, we compared the rankings produced by both solutions. Figure 4.25 shows the number of entities in common within the top positions of both rankings. It is possible to see that the rankings are very different, so the function of the linear transfer has great impact on the results.

Then, we verified the scores distribution for the hospitals after the score transfer. Figures 4.26 and 4.27 show the distribution produced by the simple linear transfer and the proportional linear transfer, respectively. The y-axis is in the log scale. Although both solutions produced similar score distribution, we believe that the distribution produced by the proportional linear transfer is more appropriate for anomaly detection: few hospitals received high score whereas many hospitals received low score.

Hence, we adopt the proportional linear transfer due to to qualitative and empirical reasons.

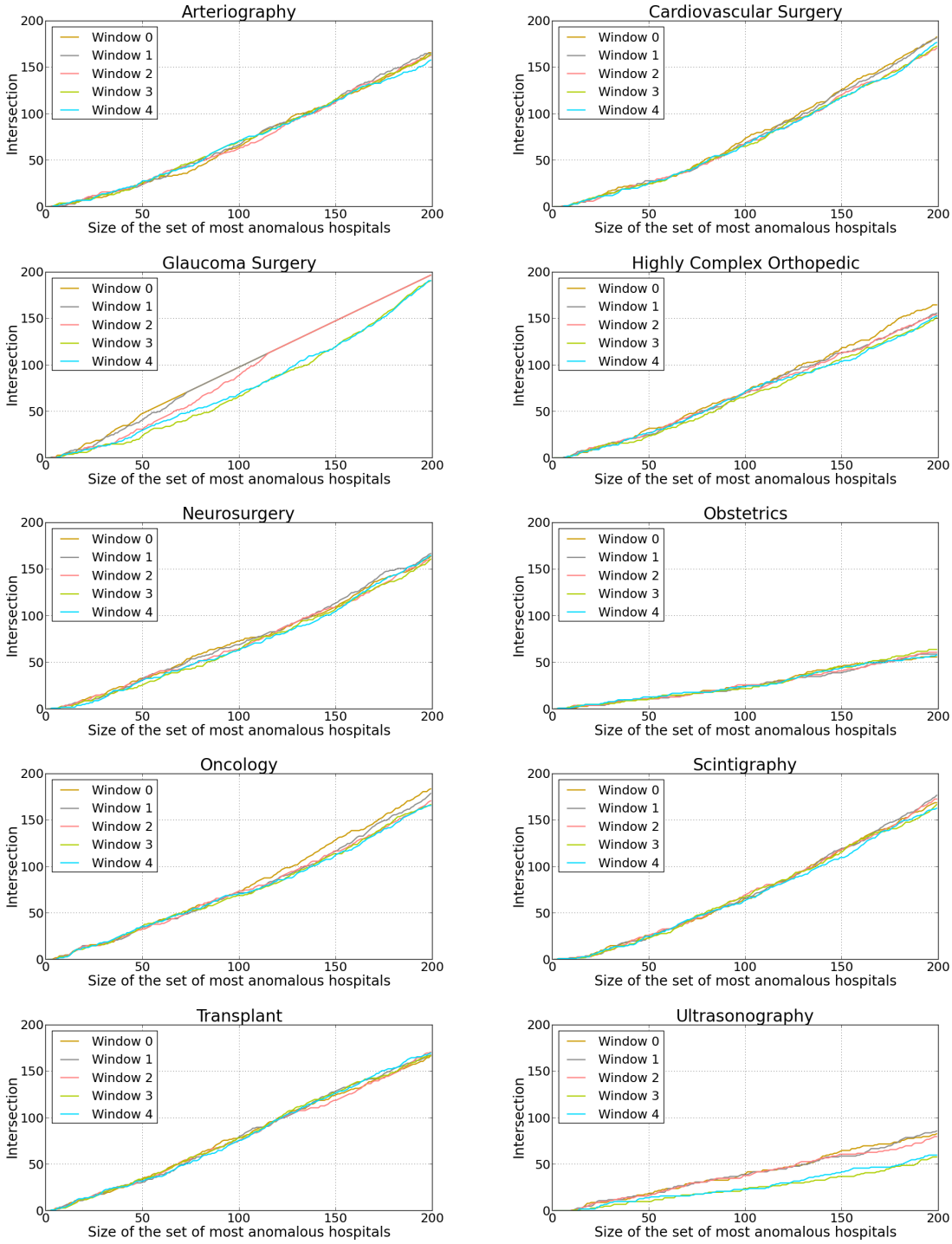

Figure 4.25. Comparison between the rankings produced by the two linear transfer approaches.

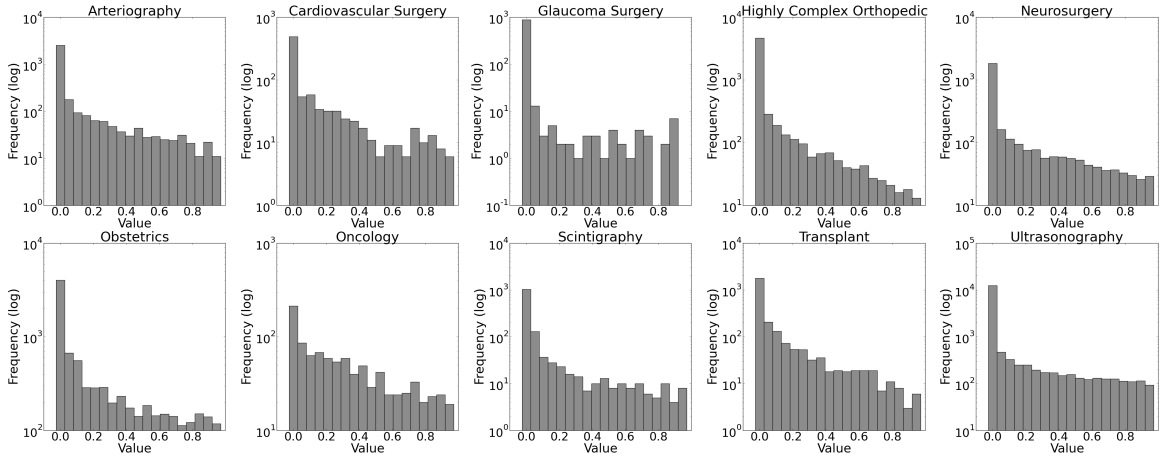

Figure 4.26. Hospitals score distribution produced with the simple linear transfer.

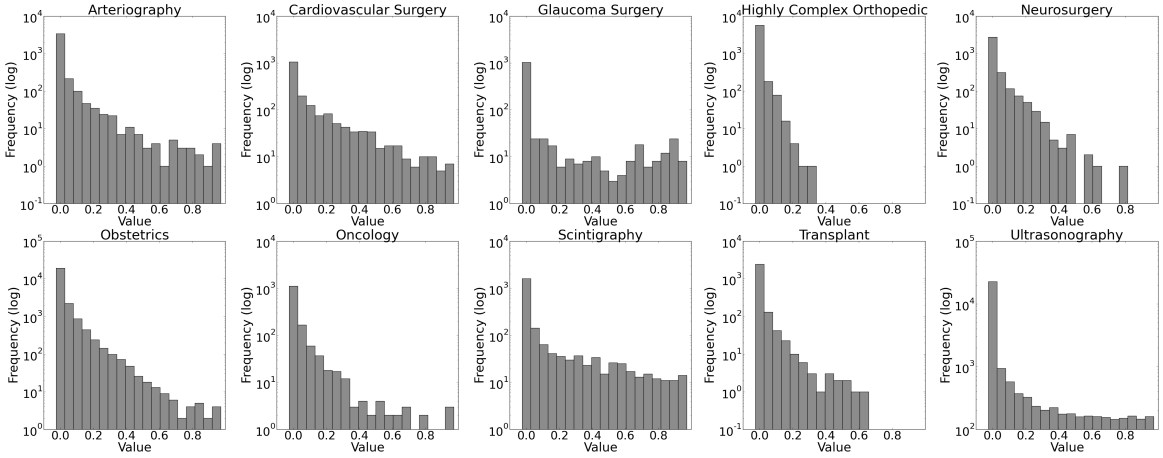

Figure 4.27. Hospitals score distribution produced with the proportional linear transfer.

4.3.1.2 Genetic algorithm

As shown in Section 3.4.3, we implemented a genetic algorithm in order to apply an optimization-based solution for the score transfer, without dealing with the exponential computational cost. For each window of each procedure, we run the algorithm in order to produce a ranking. Next we show how we defined the values for the parameter of population size P, number of generations G, crossover probability CP, mutation probability MP, tournament size TS and elitism E.

The core aspect to be observed in genetic algorithms is the selection pressure which determines the compromise between the speed of convergence of the solution and the search space exploitation. If the selection pressure is too strong, the best fitness is reached too fast before the required search space exploitation, so it is likely

that a local optimal solution will be output. On the other hand, if the selection pressure is too weak, it might not converge to a good solution or take a long time for it.

Each parameter affects the selection pressure. Next, we show our experiments to find the best values for each parameter. We believe that the parameters of the genetic algorithm have the same effect in all procedure types and in all windows. Thus, in our analysis to define the parameters we present only the results for Cardiovascular Surgery in the third window (year of 2010).

The basic assumption of the algorithm is that, during the generations, it is expected a gradual improvement in the fitness of the best individual. In order to verify the existence of convergence, we empirically fixed the values of the parameters as:

$$P = 70,\ G = 50,\ CP = 0.6,\ MP = 0.1,\ TS = 3,\ E = TRUE$$

Figure 4.28 shows the value of the best fitness of each generation. As expected, there is convergence: the best solution is about $44,000$ in the first generation and the fitness of the last generation is about $40,500$. As we employ elitism, the solution of one generation is never worse than the previous generation.

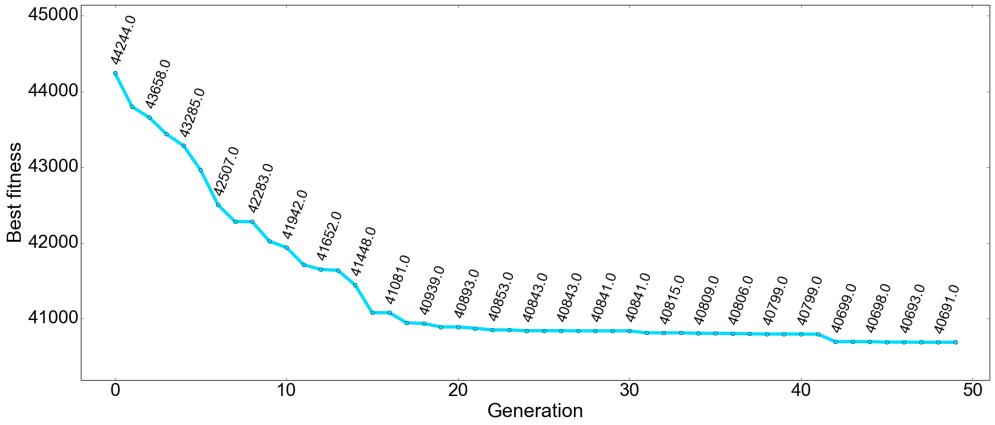

Figure 4.28. Convergence of the best solution in a first execution.

Next we present the experiments to find the best values for the parameters. For each parameter, following the decreasing importance order: P, G, CP and MP, TS and E, we vary the values and measure the best solution while the others parameters are kept fixed. Once that we find the best value for one parameter, we apply this value for the remaining experiments. All experiments were repeated 7 times and only the averages are shown in the results in order to ensure reliability.

Population size: first, we verified the best value for the population size. We change the population size and keep fixed the other parameters. We expect that the greater the population, the better the solution because bigger populations allow more diversity and vast exploitation of the search space. As each hospital may assume many score values, the search space is huge, then we include big populations in our experiments: P is ranged from 50 to $P = 800$ with step equal to 30.

Figure 4.29 shows the best fitness value of the executions with different population sizes. As expected, the best solutions were provided with great values for P. From $P = 300$, the fitness improvement is slower than the improvement for small values of P. Although it was verified that the solution quality is proportional to the size of P, the execution cost is too high when P is big. Thus, we believe that 320 is a good value for the remaining experiments considering the trade off between the execution cost and the results quality.

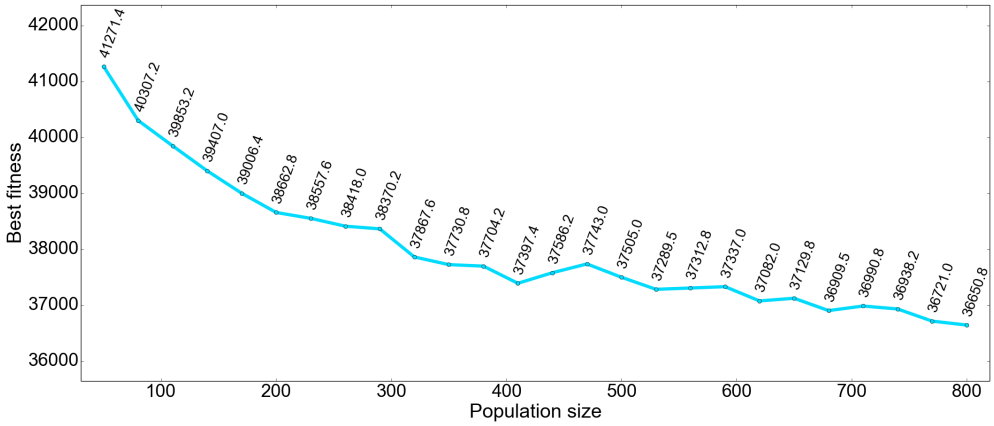

Figure 4.29. Best fitness values for different values of population size.

Number of generations: using the value of $P = 320$ and fixing the other parameters, we varied the number of generations to find the best value. As the search space is huge, we also tried large values for G: from 50 to 440 with step equal to 30. Again, we expect that the greater the value, the better the solution.

The results shown in Figure 4.30 indicate that the quality of the results is proportional to the number of generations. The best solutions found are those associated with large values of G. Thus, we adopt the value of $G = 400$ for the remaining experiments.

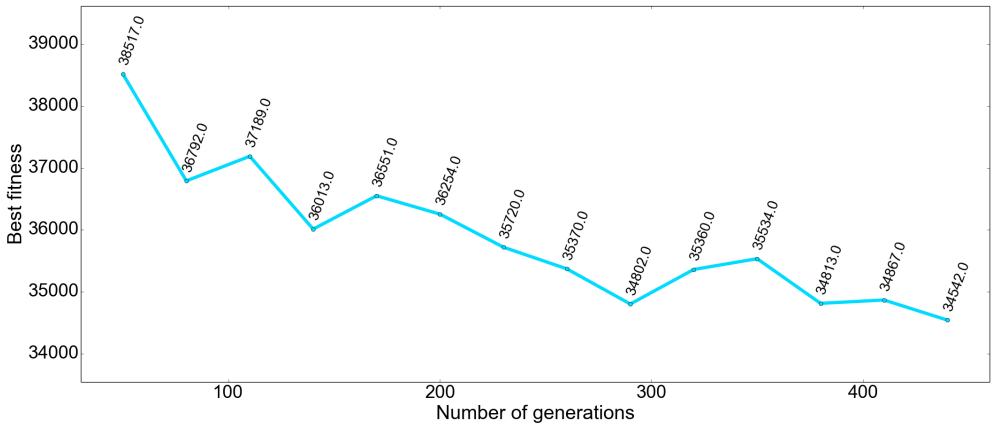

Figure 4.30. Best fitness values for different number of generations.

Reproduction probabilities: Adopting the values of $P = 320$ and $G = 400$, we verified the solution provided by different combinations of values for the crossover probability CP and mutation probability MP. The crossover is the most important operation in genetic algorithms and usually receives high probability. The mutation is an important operation to expand the search space of genetic algorithms, however it has to be performed with lower probability in order not to disturb the evolution convergence. The values applied for CP are 0.6, 0.7 and 0.8 and the values for MP are 0.05, 0.1 and 0.15. Table 4.7 presents the results produced by the combination of these probabilities values for the reproduction operations.

Table 4.7. Best fitness for multiple combinations of crossover and mutation probabilities.

CP/MP	0.05	0.1	0.15
0.6	34,702	34,374	34,412
0.7	35,066	34,588	34,152
0.8	35,219	35,198	34,106

We adopt the values of $CP = 0.8$ and $MP = 0.15$, which generated the best fitness, although the results were very similar. According to the usual mutation probability applied on genetic algorithms, our MP is high, however it depends on the application. In our application in healthcare, this value of 0.15 is not causing strong change in the convergence while helps the search space exploitation.

Tournament size: The tournament size has also great impact in the selection pressure of the algorithm. If the tournament is small, the pressure is reduced as we select the best individual in a small group. Although it is good for the search space

exploration, it reduces the convergence of the solution, as the individuals selected for reproduction might be not so good. On the other extreme, if the value of TS is too high, we tend to repeat more the same individuals selected for reproduction and it might cause a premature convergence of the solution. Figure 4.31 shows the solution produced with four values for TS: 2, 3, 4 and 5.

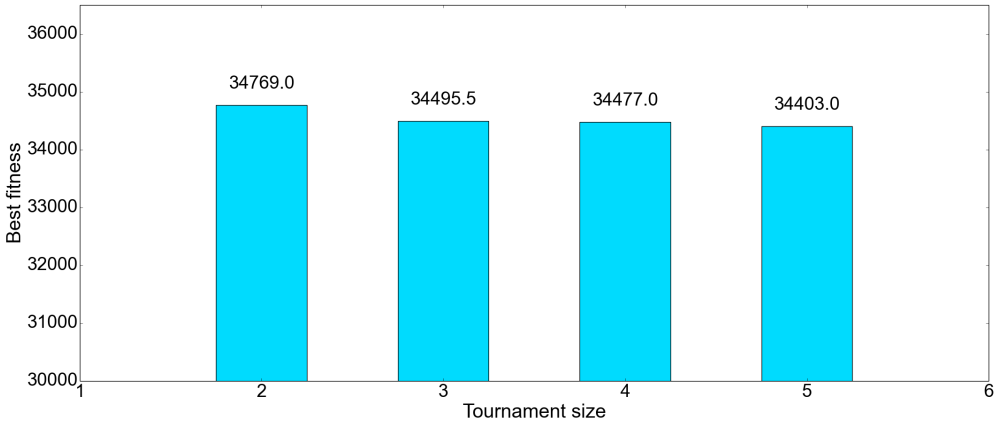

Figure 4.31. Best fitness values for different tournament sizes.

Again, the results were very similar. We assign the value of $TS = 5$ for the next experiments. Besides producing the best solution, we believe that this large value for TS is suitable if we observe that the population of 320 can be considered large as well.

Elitism: Finally, we evaluate the quality of the results with and without elitism. Although the elitism avoid worsening the results between two adjacent generations, it increases the selection pressure and sometimes it may result in worse results. The effect of the elitism cannot be predicted and varies according to the application.

As shown in Figure 4.31 the execution with $TS = 5$ and with elitism produced solutions with average fitness $34,403$. Keeping the parameters, we executed the genetic algorithm without elitism and obtained an average solution equal to $34,492$. Although the results are very similar, we adopt the elitism in order not to loose the best solutions generated.

Figure 4.32 shows the evolution of the best fitness with $E = TRUE$ and $E = FALSE$ in one of the executions. It is possible to see that in the execution without elitism is sometimes worse than the previous generation whereas it never happens when it is employed .

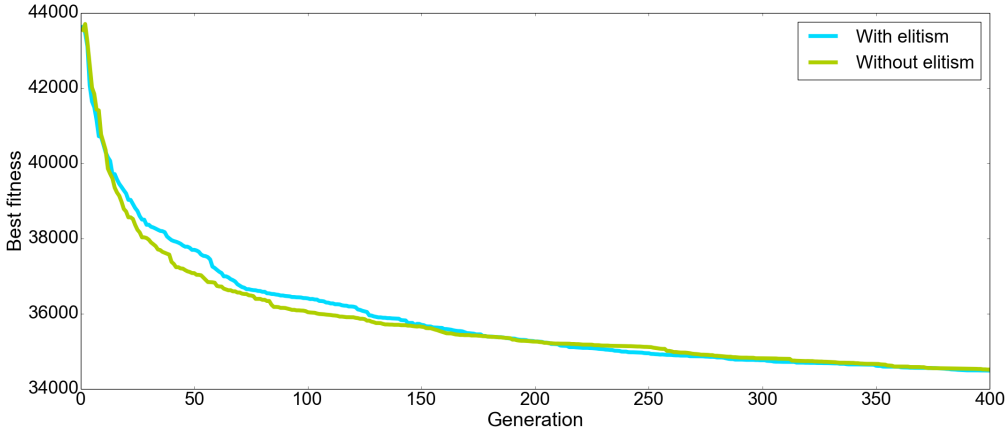

Figure 4.32. Evolution of the best fitness with and without elitism.

Hence, for our experiments we assign the following values for the parameters.

$$P = 320,\ G = 400,\ CP = 0.8,\ MP = 0.15,\ TS = 5,\ E = TRUE$$

Using the values found in the parameter calibration step, we executed the genetic algorithm in all windows of all procedures types. Figure 4.33 shows the score distribution for each procedures types considering all windows in each one.

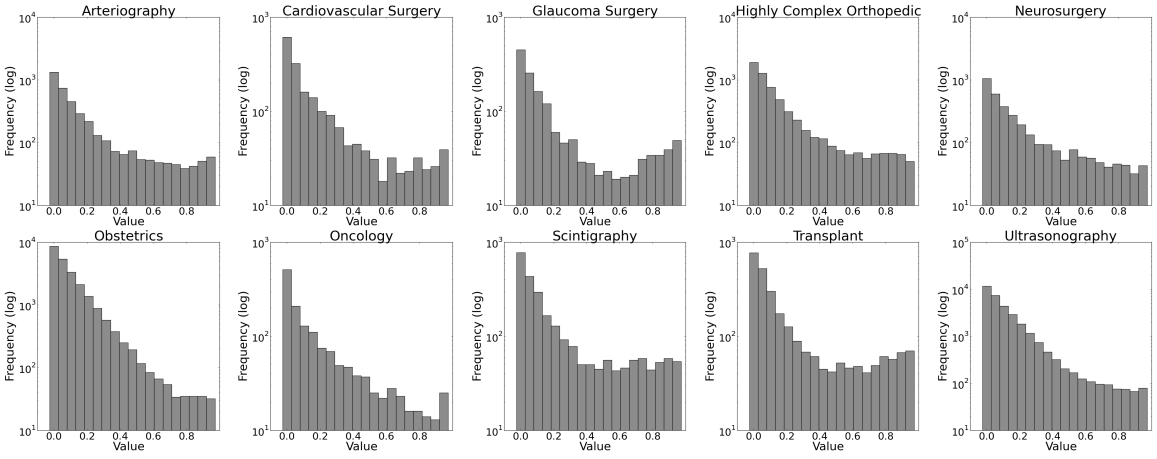

Figure 4.33. Score distribution generated with the genetic algorithm.

Next, we compare the results produced by the linear transfer and the genetic algorithm.

4.3.2 Results and evaluation

In this section we compare and evaluate the results produced by the two algorithms for score transfer. First, we compare the rankings of the linear transfer and genetic results. Next, we present a manual evaluation of the top positions of each ranking to define which algorithm produced the best results.

The first analysis is the ranking comparison in the top positions. Figure 4.34 shows, for each procedure and window, the number of hospitals in common in the top positions of the linear transfer and genetic rankings. The results show that the two algorithms produce very different rankings.

We also evaluate manually the quality of the results produced by the linear transfer and genetic algorithms for score transfer. The manual evaluation was performed with help of a public healthcare expert and was designed as following.

- For each procedure type, we manually evaluated a specific window. We believe that the solution quality would not vary so much among the windows of the same procedure type. In the future we want to evaluate all windows of all procedure types when more healthcare experts are available for helping us.

- The window evaluated of each procedure followed the alphabetical and numerical order as described in Section 4.2.2.

- For each procedure we evaluate the top 7 hospitals in the rankings of both algorithms.

- The manual labelling is based on visual analysis: for each hospital H, we analyze the time series of number of procedures performed by H, the rate of procedures in each city C_H with strong relation to H, the time series of amount of procedures in C_H and the amount of score transferred from C_H to H. Basically, we use images and plots to judge if the hospital is anomalous based on the cities affected by them. All types of visual information used in the manual labeling are shown in the case study of the next chapter.

We do not define a score threshold nor a position in the ranking for separating anomalies hospitals from regular ones as it is not possible to predict the number of real anomalies. Therefore, our evaluation of true positives is merely qualitative. Although it is not the best approach for evaluation, it is the only feasible approach. In addition, as the anomalies are very rare, having anomalies in the first positions shows that the method could meet the goals.

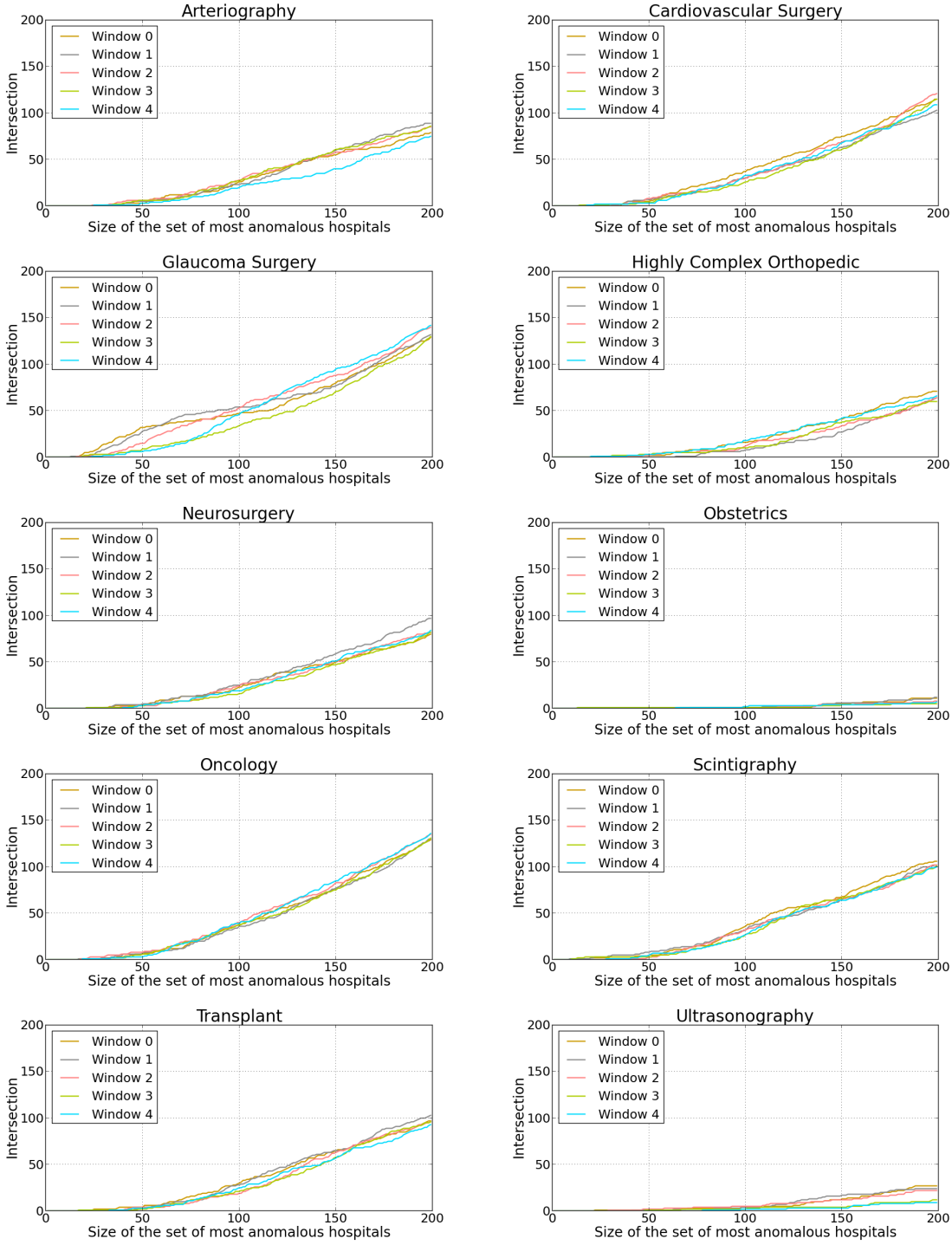

Figure 4.34. Number of hospitals in common in the top positions of the linear transfer and genetic rankings.

Again, as the anomalies are very rare, we have not estimated the true negative as well. The experiment to measure the true negative rate would consist of selecting hospitals with low scores and verifying whether they are indeed non-anomalous. However,

with some initial experiments we manually verified that all randomly selected hospitals with low scores are non-anomalous. Thus, unless we manually label all the hospitals, we do not believe that it is possible to estimate the true negative rate in a reliable way.

Table 4.8 shows the result of the manual evaluation for the top 7 hospitals in the ranking produced by the algorithms in the window associated with each procedure type, as shown before. For each hospital, we labeled it as either anomalous (red color) or regular (blue color).

Table 4.8. Results of the manual evaluation of the top seven hospitals in the rankings produced by the linear transfer and genetic algorithms.

Algorithm/Position	Art.	C.S.	Glau.	H.C.O	Neur.	Obs.	Onc.	Sci.	Tra.	Ult.
Linear 1st	A	A	A	A	A	A	A	A	A	A
Linear 2nd	A	R	A	A	R	A	A	A	A	A
Linear 3rd	A	A	A	R	R	A	R	R	A	A
Linear 4th	R	R	A	R	A	A	R	A	R	A
Linear 5th	A	A	A	R	R	A	A	A	R	A
Linear 6th	A	R	A	R	A	R	R	A	R	A
Linear 7th	A	R	A	R	A	R	R	R	R	A
Genetic 1st	R	R	R	R	R	A	R	R	R	R
Genetic 2nd	R	R	R	R	R	R	R	R	R	R
Genetic 3rd	R	R	R	R	R	R	R	R	R	R
Genetic 4th	R	A	R	R	R	R	R	R	R	R
Genetic 5th	R	R	R	R	R	R	R	R	R	R
Genetic 6th	R	R	R	R	R	R	R	R	R	R
Genetic 7th	R	R	R	R	R	R	A	R	R	R

The results show that the linear transfer algorithm produced better results: most of the top positions of the ranking produced by it are anomalous hospitals whereas in the top positions of the genetic most of the hospitals are regular.

The best results occur in the procedures of High Complex Orthopedic, Obstetric and Transplant. For these procedures, the first positions of the ranking are anomalous hospitals and starting at the first position with a regular hospital, no anomalies occur. A more detailed analysis is necessary to confirm if this pattern occurs in the complete ranking.

For the procedures of Glaucoma Surgery and Ultrasonography, the 7 top positions are anomalous hospitals. For these procedures, we also analyzed further positions in the ranking. For Glaucoma Surgery the hospitals in positions 8, 9, 10, 11, 12, 13 and 15 are also anomalous. For Ultrasonography there is no anomalous hospital between positions 8 and 15. We believe that these are good results as well.

For the five remaining procedures, the first positions are occupied by both anomalous and regular hospitals. Although it is not the ideal scenario, we verified that in

the first positions of these rankings there are very anomalous hospitals. Some of these cases are shown in the case study of Chapther 5.

Finally, we want to understand why the genetic algorithm was not able to produce good results. One possible reason is that, although we applied big populations and many generations, the algorithm was not able to converge to a good solution. The other possible reason is that our fitness function might not be good for the optimization.

As we consider that the ranking produced by linear transfer is good, we expect that its fitness function is good as well. Otherwise, we can conclude that out fitness function is not appropriate for the goal in the application. In order to verify this issue, we check the value of the fitness function of the solution provided by the linear transfer and by the genetic algorithms. For each procedure, we analyze the same windows analyzed in the manual labeling.

Table 4.9. Comparison of the fitness of the linear transfer and the genetic solution.

Procedure	Window	Fitness (genetic)	Fitness (linear)
Arteriography	1	23,451	40,833
Cardiovascular Surgery	2	32,217	43,012
Glaucoma Surgery	3	812,178	1,169,837
Highly Complex Orthopedic	4	13,324	179,13
Neurosurgery	5	18,572	26,549
Obstetrics	1	395,366	514,402
Oncology	2	35,274	44,014
Scintigraphy	3	88,624	160,112
Transplant	4	13,735	19,856
Ultrasonography	5	3,688,832	3,920,244

Table 4.9 presents the fitness values. As the fitness values of the linear transfer are not good compared to the genetic solution, we conclude that the fitness function is not appropriate for producing good rankings and score assignments.

This conclusion does not discard the other possibility of insufficient population size and generations for the convergence. However, it is necessary to apply a good fitness function to perform correctly such analysis.

Hence, the linear transfer provided the best results. As a future work, we want to develop a good fitness function able to evaluate correctly the quality of a solution according to our notion of anomaly.

In the next chapter we present some likely anomalies that we were able to found by executing the score transfer with the linear transfer algorithm.

Chapter 5

Case study

We executed the method with the KNN for the anomaly analysis and with the linear transfer for the transfer score step as described in Chapter 4. In this chapter we present a detailed analysis over the most likely anomalous hospitals found focusing on how and why our method was able to identify them. We also present some analysis concerning the financial cost of the procedures. Hospitals 8199, 7857, and 5213 are the three anomalous hospitals selected for the case study. We observe that auditing steps are necessary to either conclude that they have committed fraud or the anomalies were caused by genuine reasons that may be justified by other aspects.

5.1 Financial analysis

Besides the number of procedures, information about the cost is also available in the *DATASUS* database. Considering these information, we have performed some analysis in order to estimate how much money was involved in anomalous activities. Next we describe our methodology to estimate the financial information and show the results.

5.1.1 Methodology

The goal of the financial analysis is to compute the difference between the actual cost of the procedures and the the ideal cost of the procedures if no anomalies has occurred. We refer to this difference as the *residual cost*. The actual cost of the procedures is available in the dataset and expressed as the cost of the procedures performed by each hospital in the population of each city per month. The ideal cost of the procedures is estimated considering the average price of the procedures and the ideal number of procedures in the cities, as detailed next.

Given a procedure type, we estimate the ideal number of procedures $ideal(C, W)$ in each city C during window W as the average rate of procedures of its contextual neighbourhood $N(C)$ applied to its population $pop(C, W)$. The neighbourhood size is 8.

$$ideal(C, W) = \frac{\sum_{n \in N(C)} rate(n, W}{|N(C)|} * pop(C, W)$$

After we compute the ideal amount $ideal(C, W)$ of procedures in each city C, we also estimate the ideal number of procedures $ideal(H, C, W)$ that each hospital H should have performed in the population of C during window W. The value of $ideal(H, C, W)$ is estimated considering the fraction that H represents in the total number of procedures in C.

$$ideal(H, C, W) = ideal(C, W) * \frac{actual(H, C, W)}{actual(C, W)}$$

Once that it is known the actual amount $actual(H, C, W)$ and the ideal amount $ideal(H, C, W)$ of procedures between each pair (H, C), we estimate the residual number of procedures $residual(H, C, W)$ that indicates the number of extra procedures between the ideal and the actual number.

$$residual(H, C, W) = actual(H, C, W) - ideal(H, C, W)$$

As the cost vary among hospitals, in order to estimate the financial cost, we estimate the average price $avg\_cost(H, W)$ per procedure in H as the total cost in H divided by the total number of procedures performed by H during W.

Finally, we compute the residual cost $residual\_cost(H, C, W)$ that indicates the amount of money that could have been saved if no anomalies existed between each pair H, C during W.

$$residual\_cost(H, C, W) = residual(H, C, W) * avg\_cost(H, W)$$

5.1.2 Results

We performed the financial analysis over all ten types of procedures from 2008 to 2012.

The results are shown in Figure 5.1 in which the bars show the real and the ideal cost of the procedures and the numbers above the bars indicate the residual cost [1].

[1] The original values were expressed in the Brazilian currency (Real). The Real value ranged from 0.40 to 0.65 US dollars approximately from 2008 to 2012. In this paper we adopt an exchange rate of 0.525.

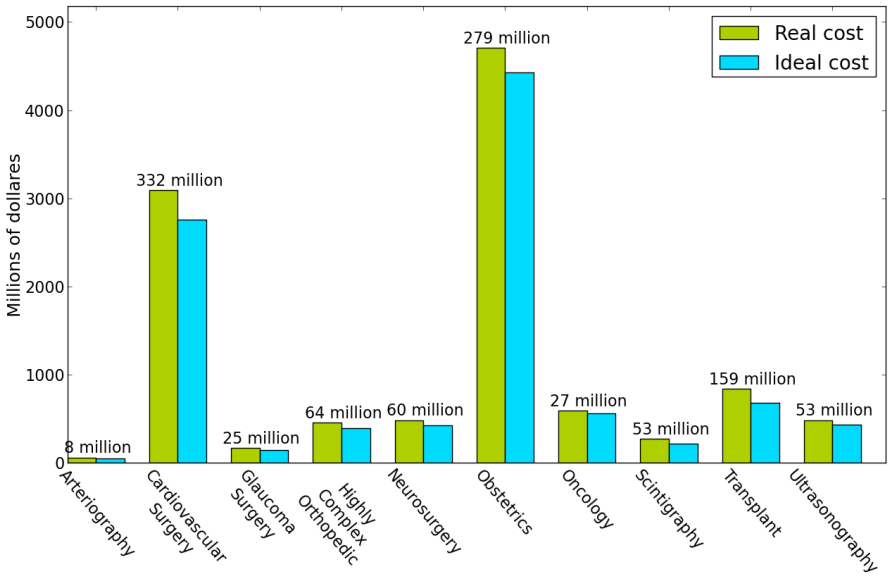

Figure 5.1. Real and ideal cost of the procedures. The numbers above the bars indicate the residual cost.

The total residual cost considering all procedures is about one billion dollars.

Although the Obstetric procedure presents the largest cost, we estimate that the procedure of Cardiovascular Surgery presents the largest residual cost. As the average cost of each Obstetric procedure (208 dollars in the average) is also greater than each Cardiovascular Surgery (183 dollars in the average) procedure, we believe that the only reason for these results of residual analysis is that there are more anomalies concerning the Cardiovascular Surgery than the Obstetric procedure.

Table 5.1. Hospitals with largest residual cost for each procedure.

Procedure	Hospital ID	Residual cost ($)
Arteriography	*2462*	1,695,112
Cardiovascular Surgery	*6824*	57,086,700
Glaucoma Surgery	*5463*	7,731,363
Highly Complex Orthopedic	*4173*	8,956,014
Neurosurgery	*8286*	6,665,048
Obstetrics	*6211*	26,765,865
Oncology	*1276*	4,823,486
Scintigraphy	*7266*	4,412,882
Transplant	*3600*	19,434,974
Ultrasonography	*3110*	1,822,442
TOTAL	-	139,393,886

Table 5.1 shows, for each procedure type, the hospital with largest residual costs.

This table also shows the total amount of money that could have been be saved if these ten hospitals performed the ideal number of procedures.

5.2 Detailed analysis

In this section we present three hospitals likely to be anomalous and detail how the method was able to detect them.

5.2.1 Hospital 8199

Hospital 8199 is in the top positions of the anomaly ranking for the procedures of Arteriography, Cardiovascular Surgery, Neurosurgery, Obstetrics, Scintigraphy, Transplant and Ultrasonography. In this analysis we present the hospital behaviour considering the procedure of Cardiovascular Surgery, in which its anomaly is the most likely. For this procedure, hospital 8199 was the first in the anomaly ranking of all windows.

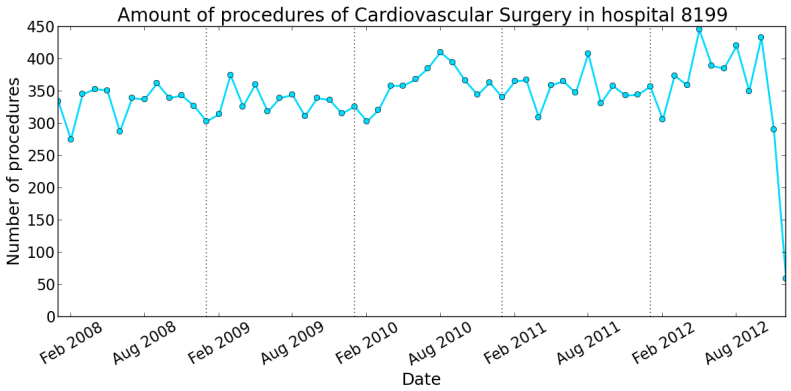

Figure 5.2. Monthly number of procedures of Cardiovascular Surgery in hospital 8199.

First of all, we show in Figure 5.2 the monthly amount of Cardiovascular Surgery procedures performed by hospital 8199. Considering only the number of procedures, hospital 8199 does not look anomalous as the amount does not vary much. However, if we look to the cities related to the hospital, we can verify its anomaly.

Figure 5.3 presents the graphical view of hospital 8199 and the cities related to it. The blue node is the hospital. The nodes connected to the hospital are the cities most related to it between 2008 and 2012. The color of each city represents its average anomaly degree during the five years of analysis. The edges indicate the number of procedures performed by 8199 in the population of each city in the whole period.

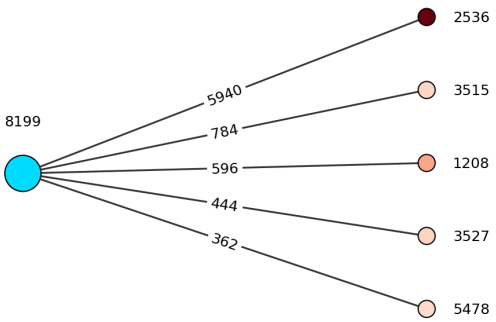

Figure 5.3. Hospital 8199 and its connection to the most served cities from 2008 to 2012.

From the bipartite graph we conclude that, among the cities that are strongly served by hospital 8199, cities 2536 and 1208 are the most anomalous. Next we analyze both cities and their relation to hospital 8199.

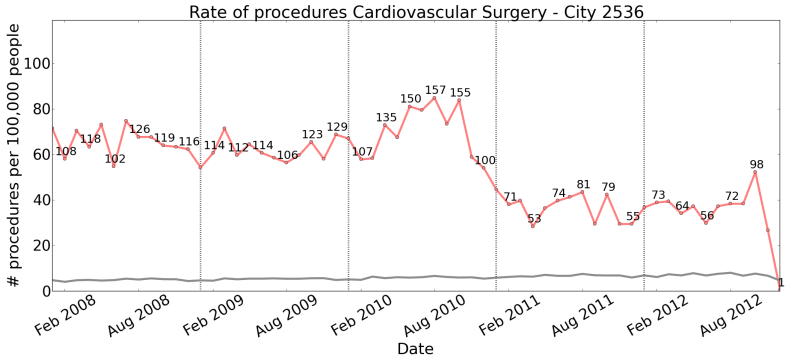

Figure 5.4. Rate and number of procedures of Cardiovascular Surgery in city 2536.

Figure 5.4 presents the rate of Cardiovascular Surgery procedures in city 2536 (red line) compared to the country average rate (gray line). For each month, we also show the number of procedures. From this analysis, it is possible to conclude that city 2536 is anomalous.

So far, we have shown that city 2536 is the main city served by hospital 8199 and that city 2536 is anomalous. In order to show that the anomaly in city 2536 is caused by hospital 8199, we show in Figure 5.5 that this hospital is the main healthcare provider for Cardiovascular Surgery in this city: in all windows, the hospital performed almost all the procedures in its population.

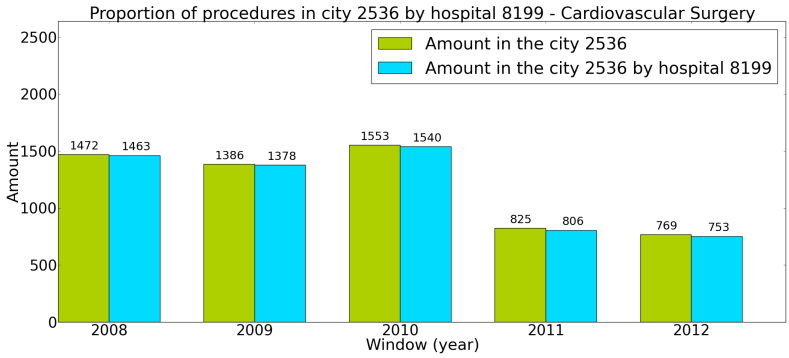

Figure 5.5. Whole number of procedures in city 2536 and amount performed by hospital 8199.

Analyzing the relation between city 1208 and hospital 8199, we verified that the number of procedures relating these entities reduced from the year of 2011. As city 1208 is the most related city to hospital 8199, we would expect that the amount in 8199 were also reduced. However, the overall number of procedures in 8199 does not vary much (Figure 5.2). This gap can be explained if we analyze city 1208, that was also target of anomaly activities of hospital 8199.

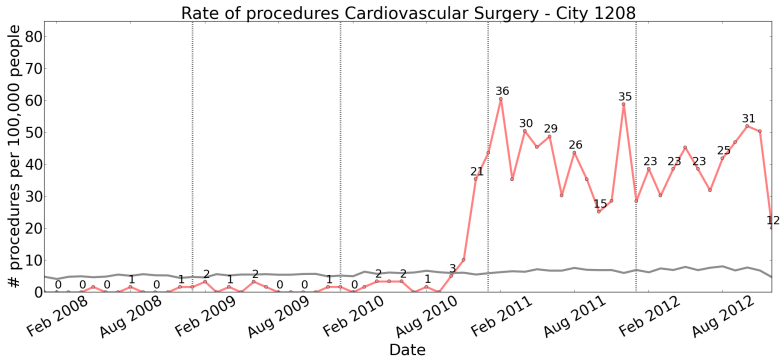

Figure 5.6. Rate and number of procedures of Cardiovascular Surgery in city 1208.

Figure 5.6 shows the rate and number of procedures in city 1208. The city is anomalous as the number of procedures grew dramatically from the end of 2010.

According to Figure 5.7, which shows the number of procedures in city 1208 performed by hospital 8199, we conclude that from the moment that 8199 started to be a healthcare provider of 1208 population, the city started to present anomalous rate of procedures. Therefore, city 1208 is anomalous and hospital 8199 is the responsible for the anomaly.

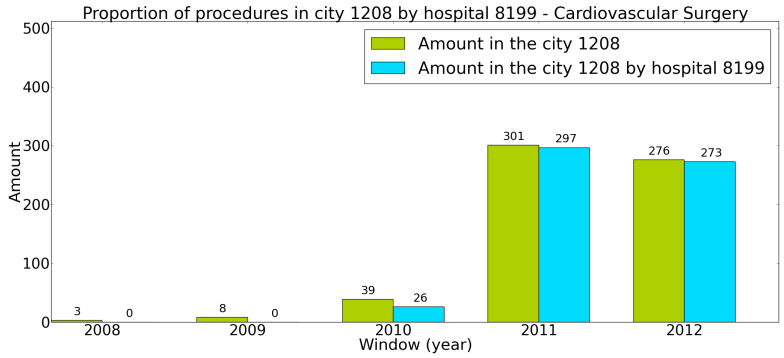

Figure 5.7. Whole number of procedures in city 1208 and amount performed by hospital 8199.

From 2008 to 2012, hospital 8199 performed procedures of Cardiovascular Surgery in the population of 324 cities. As shown, the amount in the hospital is constant so that we would not discover its anomaly without applying a method for anomaly detection which considers the cities related to it. We have found two cities that illustrate such behaviour: from the moment that hospital 8199 reduced its activities in the population of city 2536, it started to perform an anomalous number of procedures in the population of city 1208.

Table 5.2. Real, ideal and residual cost (in Dollars) in hospital 8199 for all procedures.

Procedure	Real cost ($)	Ideal cost ($)	Residual ($)
Arteriography	1,690,235	1,318,349	371,886
Cardiovascular Surgery	66,954,175	53,566,887	13,387,288
Glaucoma Surgery	214	179	35
Highly Complex Orthopedic	3,582,055	3,509,521	72,535
Neurosurgery	4,350,941	3,001,968	1,348,973
Obstetrics	9,419,764	5,077,617	4,342,147
Oncology	3,817,857	3,130,901	686,956
Scintigraphy	1,463,997	1,334,323	129,674
Transplant	3,554,757	2,371,259	1,183,498
Ultrasonography	1,440,512	1,043,252	397,261
TOTAL	**96,274,507**	**74,354,255**	**21,920,253**

Table 5.2 shows the money received by hospital 8199 and the money that it should have received if the number of procedures performed were regular according to our methodology for financial analysis. The procedure of Cardiovascular Surgery presents the largest difference between the real and the ideal money produced in the period between 2008 and 2012. Considering all the procedures, the hospital could

have saved almost 22 million dollars from the government if the regular number of procedures were performed.

5.2.2 Hospital 7857

Hospital 7857 is one of the most anomalous hospitals concerning the procedure of Glaucoma Surgery. In the years of 2010 and 2011, it is in the first position of the ranking.

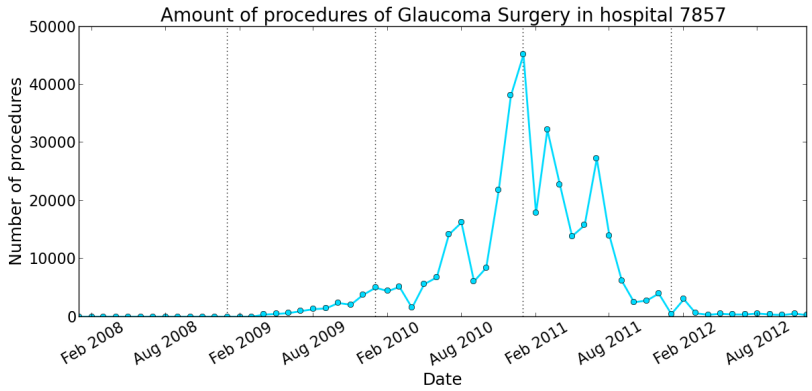

Figure 5.8. Number of procedures of Glaucoma Surgery in hospital 7857.

Figure 5.8 shows the monthly number of procedures in the hospital. On the contrary of the previous case, the time series of hospital 7857 indicates the anomalous behaviour since the amount suffer drastic changes in 2010 and 2011.

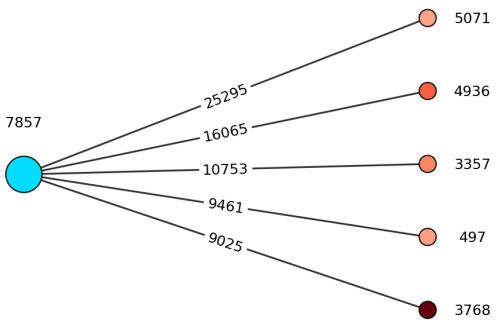

Figure 5.9. Hospital 7857 and its connection to the most related cities from 2008 to 2012.

The cities that are most related to hospital 7857, their anomaly degree and the number of procedures concerning the pairs are indicated in the graph of Figure 5.9. Next we present a detailed analysis over the cities 5071, 4936, 3357, 497 and 3768.

Figures 5.10, 5.11, 5.12, 5.13 and 5.14 show the rate, the amount and the contribution of hospital 7857 of procedures in termos of Glaucoma Surgery in cities 5071, 4936, 3357, 497 and 3768, respectively. Not only it is possible to conclude that all these five cities are anomalous in the years of 2010 and 2011, but also that the anomaly is caused by hospital 7857. Thus, hospital 7857 is a case of anomaly detected through the analysis of the rate of procedures in the cities as modeled with our method.

Among the ten analyzed procedures, hospital 7857 performs only the procedure of Glaucoma Surgery and the real cost of the procedures between 2008 and 2012 was 9,946,218 dollars, which is 121,329 dollars smaller than the ideal amount of 10,067,567 dollars. Although the hospital presented anomalous behaviour in some cities, in the general, the amount of money received was not anomalous. Thus, if a financial analysis were performed, hospital 7857 would not be detected as anomaly.

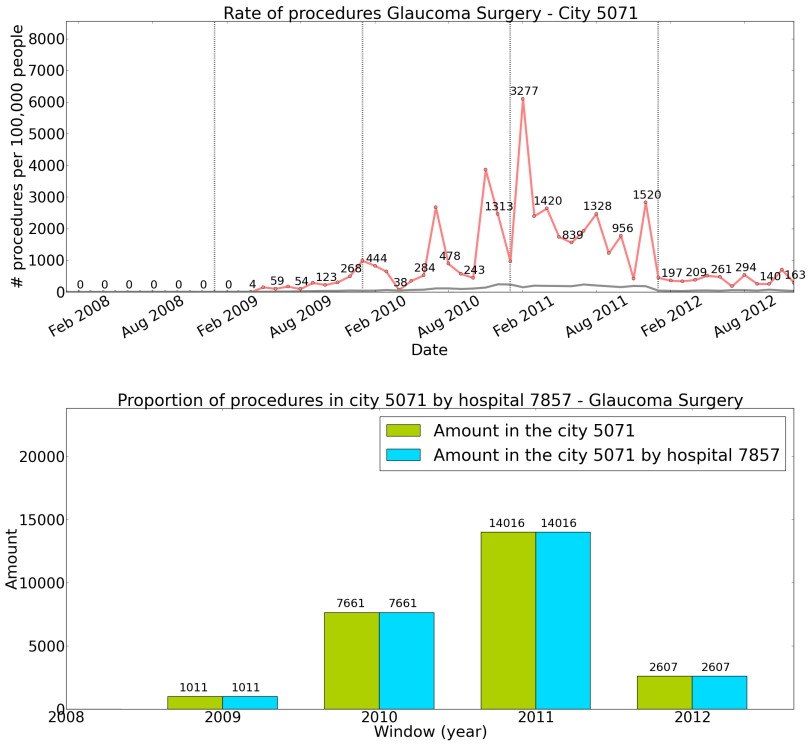

Figure 5.10. Rate and amount of Glaucoma Surgery procedures in city 5071 and the amount performed by hospital 7857.

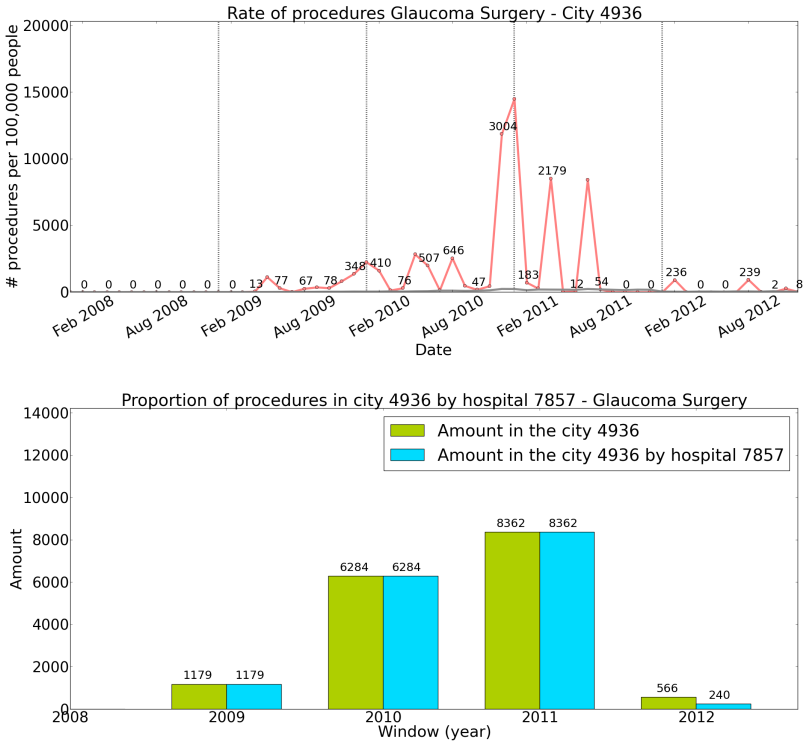

Figure 5.11. Rate and amount of Glaucoma Surgery procedures in city 4936 and the amount performed by hospital 7857.

5.2.3 Hospital 5213

Hospital 5213 is an evident anomaly concerning the procedure of Scintigraphy being in the first position of the ranking of all five windows.

Figure 5.15 shows the amount of Scintigraphy procedures in hospital 5213 in each month. The variations are not frequent and not significant. Thus, we can conclude again that the proposed method helped finding anomalies that would not be found if only data about the hospitals were applied.

The five cities most related to hospital 5213 and their anomaly degrees are indicated in Figure 5.16.

Figures 5.17, 5.18, 5.19, 5.20 and 5.21 show that cities 1695, 3506, 3356, 744 and 4279 are anomalous and that hospital 5213 is responsible for the anomalous behaviour identified in the cities.

Table 5.3 shows the real, ideal and residual cost of the procedures in hospital 5213. Although the difference is positive for some procedures types and negatives for others, in the total, it received more than 1.5 million dollars above the expected value if the number of procedures were regular. Again, we believe that hospital 5213 would not be so anomalous if a financial analysis were performed. However, considering the

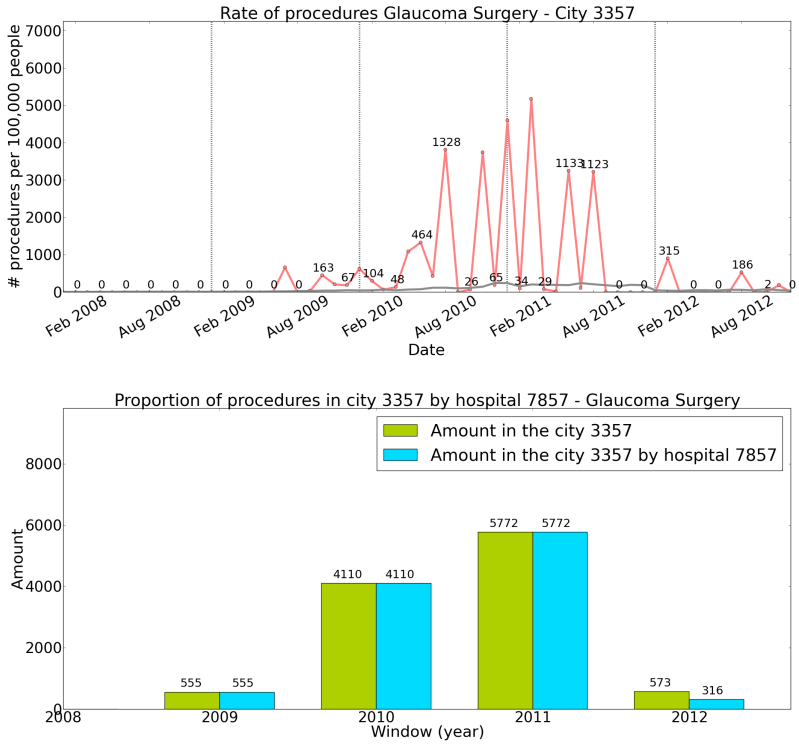

Figure 5.12. Rate and amount of Glaucoma Surgery procedures in city 3357 and the amount performed by hospital 7857.

Table 5.3. Real, ideal and residual cost in hospital 5213 for all procedures.

Procedure	Real cost ($)	Ideal cost ($)	Residual ($)
Arteriography	271	304	-33
Cardiovascular Surgery	153,154	178,997	-25,844
Glaucoma Surgery	217,596	239,831	-22,235
Highly Complex Orthopedic	2,872,410	3,223,415	-351,005
Neurosurgery	19,053,345	21,843,198	-2,789,853
Obstetrics	13,660,514	12,963,456	697,058
Oncology	7,135,795	4,445,261	2,690,535
Scintigraphy	1,239,647	1,416,557	-176,910
Transplant	3,554,757	2,371,259	1,183,498
Ultrasonography	1,440,512	1,043,252	397,261
TOTAL	**49,328,002**	**47,725,529**	**1,602,473**

number of procedures, especially Scintigraphy procedures, the hospital is anomalous.

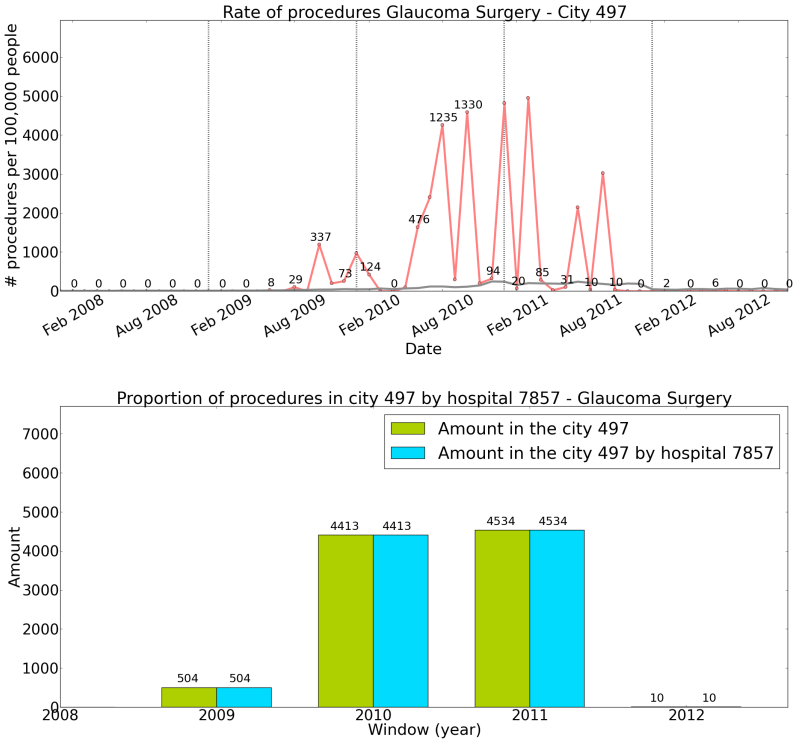

Figure 5.13. Rate and amount of Glaucoma Surgery procedures in city 497 and the amount performed by hospital 7857.

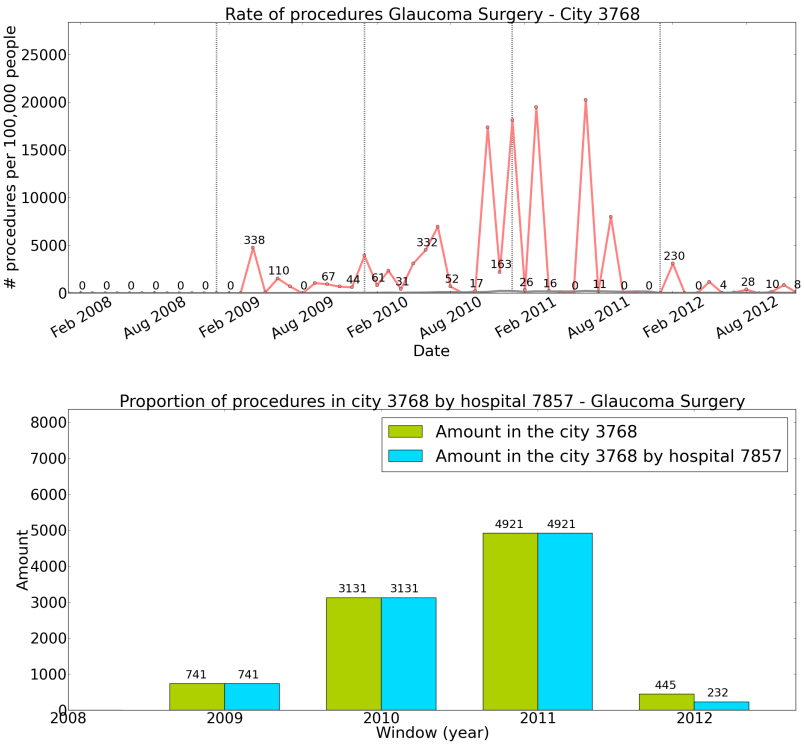

Figure 5.14. Rate and amount of Glaucoma Surgery procedures in city 3768 and the amount performed by hospital 7857.

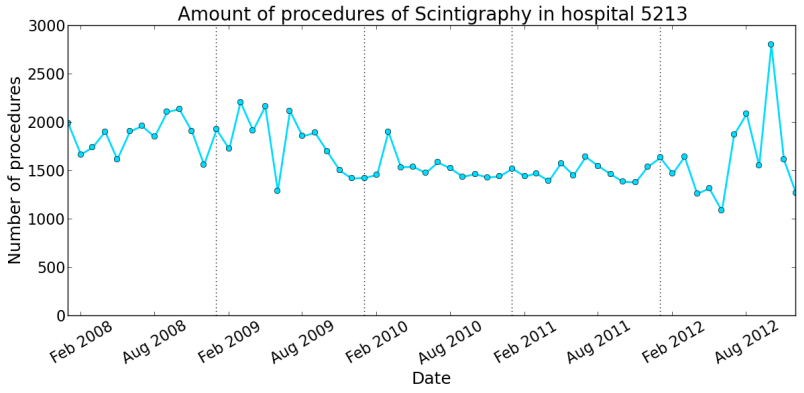

Figure 5.15. Monthly number of procedures of Scintigraphy in hospital 5213.

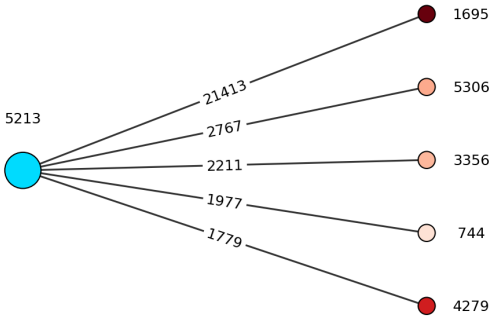

Figure 5.16. Hospital 5213 and its connection to the most related cities from 2008 to 2012.

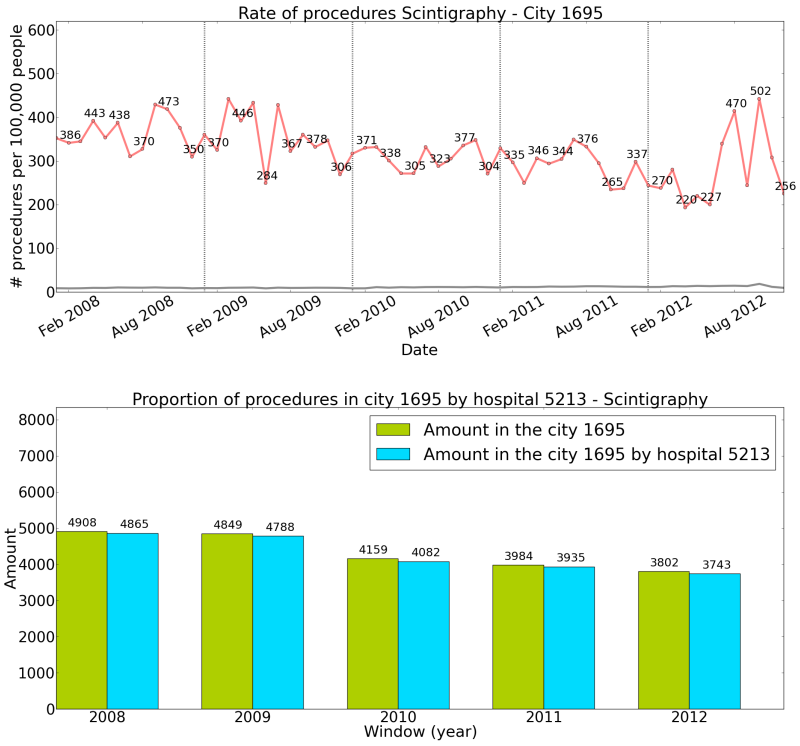

Figure 5.17. Rate and amount of Scintigraphy procedures in city 1695 and the amount performed by hospital 5213.

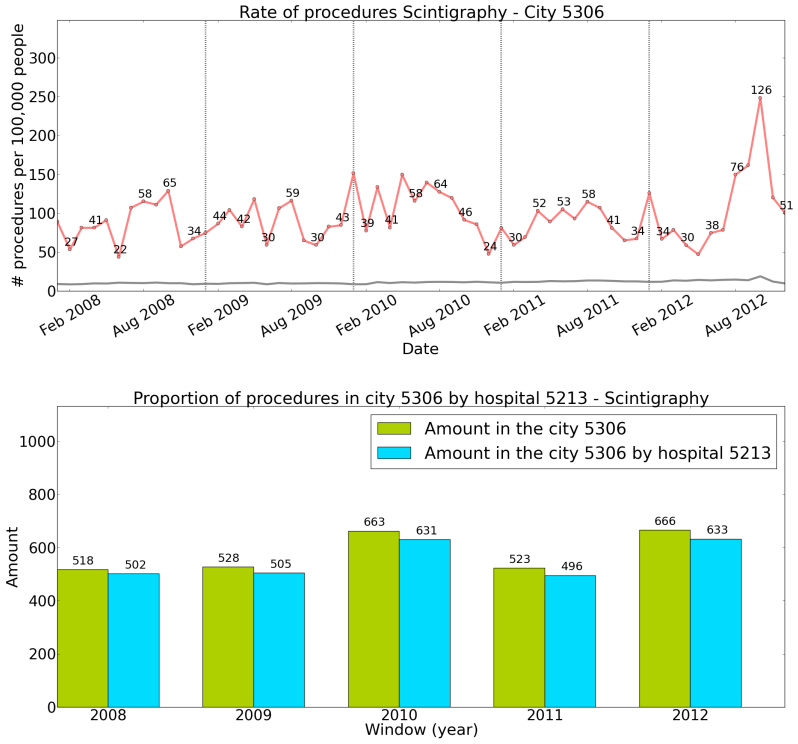

Figure 5.18. Rate and amount of Scintigraphy procedures in city 3506 and the amount performed by hospital 5213.

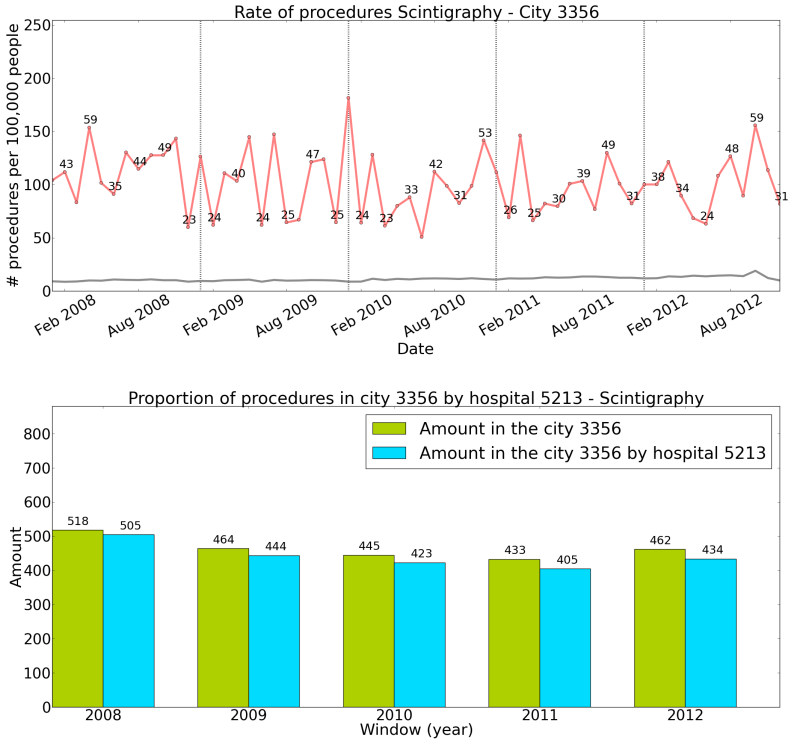

Figure 5.19. Rate and amount of Scintigraphy procedures in city 3356 and the amount performed by hospital 5213.

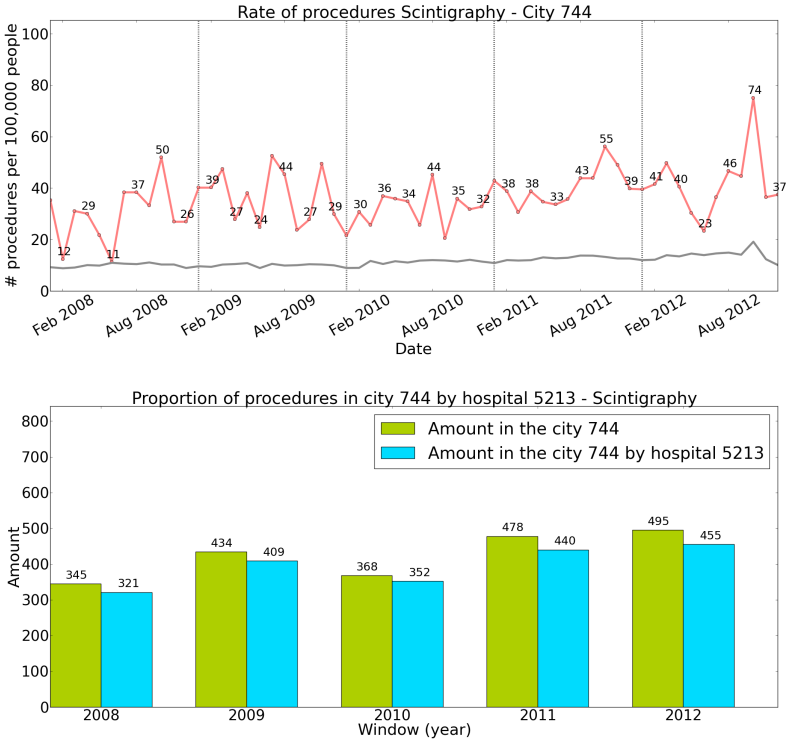

Figure 5.20. Rate and amount of Scintigraphy procedures in city 744 and the amount performed by hospital 5213.

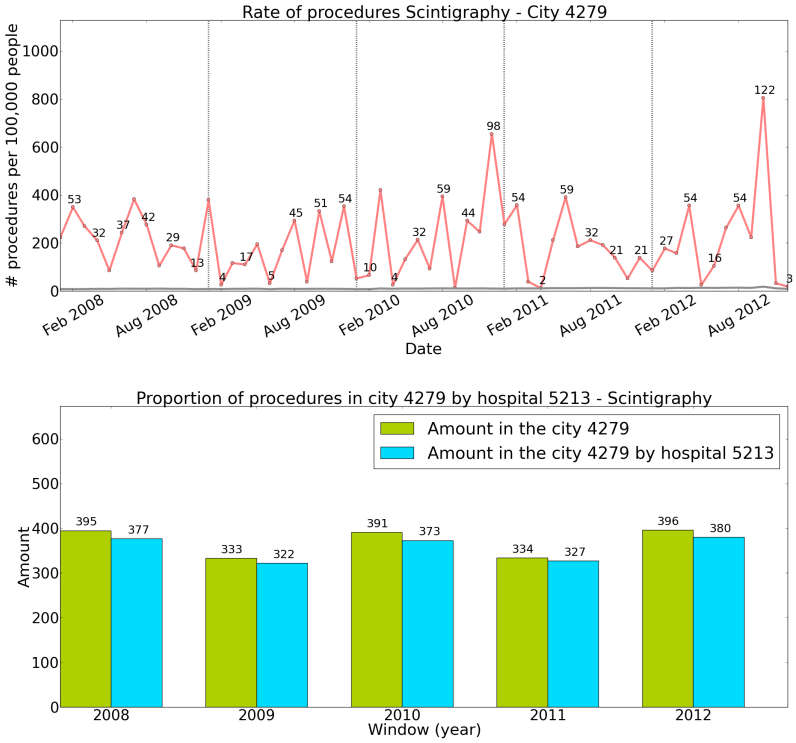

Figure 5.21. Rate and amount of Scintigraphy procedures in city 4279 and the amount performed by hospital 5213.

Chapter 6

Conclusion

In this work we propose an effective solution for the problem of detecting anomalous hospitals in healthcare systems. According to this goal, the main challenge is the lack of information about the hospitals in order to perform the capacity analysis. In addition, this information could be modified in order to disguise the anomalies. Instead, we deal with the key information that impacts in the money paid by the government for the hospitals: the number of procedures that they claim as done. Our method models the two entities, hospitals and cities, as a bipartite graph in which the edges represent the number of medical procedures that each hospital performed in the population of each city. In the first step of the method each city receives an anomaly score considering its demand analysis. In the second step of score transfer, each hospital receives an anomaly score based on the consumers score and in the number of procedures between each pair of consumer and provider.

The goal in the step of anomaly analysis is to detect contextual anomalies. The contextual feature applied is the *HDI* weighted by the geographic location of each city whereas the behavioural distance is the rate of procedures of the cities. In order to assign an anomaly score for each city we applied the KNN algorithm, although we also tried a probabilistic solution. For the score transfer, we tried a solution based on linear combination and a genetic algorithm. The best results were provided by the version of the linear combination in which each hospital H receives the score of each city C weighted by the fraction that H represents in the whole amount in C.

We perform the experiments considering 10 different types of medical procedures from the Brazilian public healthcare system, which is considered a reliable and complete database. The total cost of the procedures analyzed from 2008 to 2012 is over 8 billion dollars. Evaluating the results is not a trivial task as there is no labeled dataset nor any type of ground truth. Based on our visual analysis with help of experts in public

healthcare management, we concluded that our results were good and that we were able to detect anomalous hospitals. Our case study shows some evident cases of anomalies. We also estimated that about one billion dollars could have been saved if all hospitals performed the regular amount of procedures.

We believe that our method can be applied to a large range of public healthcare systems in the world since it depends on two trivial information: the population of each city (or region) and the number of people from each city (or region) cared in each hospital.

6.1 Future work

In this section we present future works for anomaly detection in healthcare systems and discuss some other potential application of the method.

6.1.1 Future work in healthcare

Considering the application in the healthcare, we want to perform experiments in further procedures types and develop solutions for dealing with procedures vulnerable to seasonality. In addition, we want to design, implement and evaluate solutions for the scenarios in which the method presents limitations:

- Big cities issue: we want to divide big cities into smaller regions using geocoding and perform the experiments again so that we may be able to detect anomaly in big cities. This is a challenging task because instead of dealing with the cities of the patients, we have to deal with their full address in order to map the regions. The access to this database is also a challenge as it is not available and it involves privacy issues.

- Distributed anomalies: we intent to develop a method to detect anomalies in cases in which a hospital performs small anomalous activities in many cities. The most trivial solution is to evaluate how likely it is the relation between the hospitals and cities. This analysis can be based on the amount of cities served by the hospitals and on their geographical distances.

We also want to apply the method for detecting anomalies in healthcare of other countries and to implement and evaluate it with further unsupervised algorithms and ensemble techniques.

The results evaluation is also a challenging task of the work. In addition to the visual analysis and manual labeling, we want to design a model to simulate synthetic

anomalies in the database. The evaluation with synthetic anomalies would give us the opportunity of measuring precisely the precision and the recall. However, this is not an easy task. Next we discuss some of the aspects that should be considered in such model.

Duration: the longer the duration of the anomaly, the easier is to identify it. For example, consider that each window is related to one year. If the anomaly exists for one or two months, the behavioural difference does not change significantly, as it is computed through Euclidean Distance among all months. On the other hand, if the anomaly persists for many months, it has great impact on the behavioural distance.

Anomaly intensity: the greater the anomaly intensity, the easier is to detect it. In order to define the amount of additional procedures, it is important to consider the likelihood of the amount based on the entities sizes. In addition, it is also important to not bias the number of additional procedures on the method operation. For instance, we cannot compute the intensity based on the cities rate.

Anomaly concentration in the cities: a hospital may perform anomalous activities either in few cities or divide the additional procedures into many cities. When the additional procedures are distributed among many cities, it is more difficult to detect them, as the city rates are not so affected.

Size of the anomalous cities: the size of the cities where the synthetic anomalies are inserted has great impact on the results. If the cities are small, it easy to detect them. On the other hand, if the cities are big, the rate of procedures in the city is not so affected.

We present above a suggestion of algorithm to create the synthetic anomalies considering these aspects.

1. Select at random one hospital and the duration of the anomalies.
2. Estimate the amount of additional procedures in of the hospital.
3. Distribute the synthetic anomalies among the cities related to the hospital according to the concentration level and the cities sizes.

6.1.2 Further scenarios

We aim to investigate the effectiveness of the method in other scenarios. For this purpose, the main challenge is to obtain a reliable and complete database. Next we

describe some potential applications of our method for anomaly detection in services.

Engineering: the method could be used to detect anomalies concerning partnership between government and private companies, such as engineering companies.

For example, suppose that a government hired private companies to build and fix roads. The goal is to detect potential incidents of corruption in the government. The providers are the engineering companies, the consumers are the government entities and the services are amount of roads, represented in square meter.

The obvious way to detect abuse or other anomalous activities by the government would be estimating whether the amount of services hired are really necessary or if it was really performed. However, it is not trivial to estimate these information. On the other hand, it is trivial to evaluate the capacity of the companies using features such as number of workers and number of machines. Thus, the method could be applied to detect anomalies considering the capacity of the companies and then the anomalous government entities could be also identified.

Food supply: another application for the method is to detect anomalies in companies which supply food for public schools. The food providers might change information about their capacity in order to charge for more food than necessary. However, if the demand of the schools is analyzed considering the number of students, it is possible to transfer the scores from schools to the companies in order to detect anomalous companies of food supply.

Healthcare through trajectories analysis: The method can be also applied to healthcare considering the trajectory of the patients (consumers). The goal is to detect anomalous doctors, hospitals or other providers through the identification of anomalous trajectories of patients.

The analysis of patients trajectory might reveal unreal or unnecessary procedures even if the provider patterns are regular. For example, suppose that a patient performed multiple hemodialysis procedures due to a disease in the blood. From the moment that this patient performed a kidney transplant, it is not expected further occurrences of hemodialysis procedures. Otherwise, it may represent an anomalous activity. One possible solution for discovering these potential cases of frauds in patients trajectory is applying rule-based algorithms in the anomaly analysis step of the method.

Bibliography

Aggarwal, C. C. (2013). *Outlier analysis*. Springer Science & Business Media.

Agrawal, R., El-Bathy, N., and Seay, C. (2012). Medicaid fraud detection using data broker services. *SIGHIT Rec.*, 2(1):25--25. ISSN 2158-8813.

Aral, K. D., GÃ$\frac{1}{4}$venir, H. A., Sabuncuoglu, I., and Akar, A. R. (2012). A prescription fraud detection model. *Computer Methods and Programs in Biomedicine*, 106(1):37 – 46. ISSN 0169-2607.

Baker, L. D., Hofmann, T., McCallum, A., and Yang, Y. (1999). A hierarchical probabilistic model for novelty detection in text. In *Proceedings of International Conference on Machine Learning*. Citeseer.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, volume 3. Wiley New York.

Bay, S. D. and Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 29--38, New York, NY, USA. ACM.

Becker, D., Kessler, D., and McClellan, M. (2005). Detecting medicare abuse. *Journal of Health Economics*, 24(1):189--210.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, pages 235--249.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5--32. ISSN 0885-6125.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93--104. ISSN 0163-5808.

Capelleveen, G. C. (2013). Outlier based predictors for health insurance fraud detection within us medicaid.

Carvalho, L. F., Teixeira, C. H., Dias, E. C., Meira Jr, W., and Carvalho, O. F. (2015). A simple and effective method for anomaly detection in healthcare. In *4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining, Vancouver, Canada*.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1--15:58. ISSN 0360-0300.

Dasgupta, S., Papadimitriou, C. H., and Vazirani, U. (2006). *Algorithms*. McGraw-Hill, Inc.

Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107--113. ISSN 0001-0782.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38. ISSN 00359246.

Denning, D. (1987). An intrusion-detection model. *Software Engineering, IEEE Transactions on*, SE-13(2):222–232. ISSN 0098-5589.

Dorigo, M., Birattari, M., and Stützle, T. (2006). Ant colony optimization. *Computational Intelligence Magazine, IEEE*, 1(4):28--39.

Erich, H.-P. K. P. K. and Zimek, S. A. (2011). Interpreting and unifying outlier scores. In *11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, volume 42. SIAM.

ErtÃ¶z, L., Steinbach, M., and Kumar, V. (2004). Finding topics in collections of documents: A shared nearest neighbor approach. In *Clustering and Information Retrieval*, volume 11 of *Network Theory and Applications*, pages 83–103. Springer US.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226--231.

Fabrikant, R., Kalb, P. E., Bucy, P. H., and Hopson, M. D. (2014). *Health care fraud: enforcement and compliance*. Law Journal Press.

Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316. ISSN 1384-5810.

Freund, Y. and Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3):277--296. ISSN 0885-6125.

Gaber, C., Hemery, B., Achemlal, M., Pasquet, M., and Urien, P. (2013). Synthetic logs generator for fraud detection in mobile transfer services. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 174–179.

Ghoting, A., Parthasarathy, S., and Otey, M. (2008). Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery*, 16(3):349–364. ISSN 1384-5810.

Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59--63.

Guha, S., Rastogi, R., and Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 512--521. IEEE.

Hautamaki, V., Karkkainen, I., and Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 430--433, Washington, DC, USA. IEEE Computer Society.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

He, H., Wang, J., Graco, W., and Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, 13(4):329 – 336. ISSN 0957-4174. Selected Papers from the PACES/SPICIS'97 Conference.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126. ISSN 0269-2821.

Hofmeyr, S. A., Forrest, S., and Somayaji, A. (1998). Intrusion detection using sequences of system calls. *Journal of computer security*, 6(3):151--180.

Hu, N., Liu, L., and Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3):614 – 626. ISSN 0167-9236. On quantitative methods for detection of financial fraud.

Jamil, G. L. and Carvalho, L. F. M. (2015). Perspectives of big data analysis for knowledge generation in project management contexts. *Handbook of Research on Effective Project Management through the Integration of Knowledge and Innovation*, page 1.

Jiang, Y., Zeng, C., Xu, J., and Li, T. (2014). Real time contextual collective anomaly detection over multiple data streams. *Proceedings of the ODD*, pages 23--30.

Jin, W., Tung, A., Han, J., and Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship. In Ng, W.-K., Kitsuregawa, M., Li, J., and Chang, K., editors, *Advances in Knowledge Discovery and Data Mining*, volume 3918 of *Lecture Notes in Computer Science*, pages 577–593. Springer Berlin Heidelberg.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338--345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, pages 81--93.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604--632. ISSN 0004-5411.

Kohavi, R. (1995). The power of decision tables. In *Proceedings of the 8th European Conference on Machine Learning*, ECML '95, pages 174--189, London, UK, UK. Springer-Verlag.

Kou, Y., Lu, C.-T., and Chen, D. (2006). Spatial weighted outlier detection. In *SDM*, pages 614--618. SIAM.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2010). Outlier detection techniques. In *Tutorial at the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC.*

Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., and Kavsek, B. (2000). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pages 20--24. Citeseer.

Lempel, R. and Moran, S. (2001). Salsa: The stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131--160. ISSN 1046-8188.

Li, J., Huang, K.-Y., Jin, J., and Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science*, 11(3):275–287. ISSN 1386-9620.

Lin, J., Keogh, E., Fu, A., and Van Herle, H. (2005). Approximations to magic: finding unusual medical time series. In *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, pages 329–334. ISSN 1063-7125.

Major, J. A. and Riedinger, D. R. (2002). Efd: A hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance*, 69(3):309--324. ISSN 1539-6975.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, pages 283--294.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Ortega, P. A., Figueroa, C. J., and Ruz, G. A. (2006). A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN*, 6:26--29.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: bringing order to the web.

Pang-Ning, T., Steinbach, M., Kumar, V., et al. (2006). Introduction to data mining. In *Library of Congress*, page 74.

Pearce, P., Dave, V., Grier, C., Levchenko, K., Guha, S., McCoy, D., Paxson, V., Savage, S., and Voelker, G. M. (2014). Characterizing large-scale click fraud in zeroaccess. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 141--152, New York, NY, USA. ACM.

Portnoy, L., Eskin, E., and Stolfo, S. (2001). Intrusion detection with unlabeled data using clustering. In *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001*, pages 5--8.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-238-0.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427--438. ISSN 0163-5808.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. (1999). Support vector method for novelty detection. In *NIPS*, volume 12, pages 582--588.

Schubert, E., Zimek, A., and Kriegel, H.-P. (2014). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min. Knowl. Discov.*, 28(1):190--237. ISSN 1384-5810.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. ASQ Quality Press.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional anomaly detection. *Knowledge and Data Engineering, IEEE Transactions on*, 19(5):631–645. ISSN 1041-4347.

Srivastava, A. (2006). Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques. In *Aerospace Conference, 2006 IEEE*, pages 17 pp.–.

Surdak, C. (2014). *Data Crush: How the Information Tidal Wave is Driving New Business Opportunities*. AMACOM Div American Mgmt Assn.

Tamersoy, A., Roundy, K., and Chau, D. H. (2014). Guilt by association: Large scale malware detection by mining file-relation graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1524--1533, New York, NY, USA. ACM.

Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D. W.-L. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, pages 535--548, London, UK, UK. Springer-Verlag.

Tseng, V. S., Ying, J.-C., Huang, C.-W., Kao, Y., and Chen, K.-T. (2015). Fraudetector: A graph-mining-based framework for fraudulent phone call detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 2157--2166, New York, NY, USA. ACM.

Vatanen, T., Kuusela, M., Malmi, E., Raiko, T., Aaltonen, T., and Nagai, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. ISSN 2161-4393.

Veloso, A., Meira, W., and Zaki, M. (2006). Lazy associative classification. In *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 645–654. ISSN 1550-4786.

Webb, G. I., Boughton, J. R., and Wang, Z. (2005). Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24. ISSN 0885-6125.

Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. (2003). Bayesian network anomaly pattern detection for disease outbreaks. In *ICML*, pages 808--815.

Yang, W.-S. and Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31(1):56 – 68. ISSN 0957-4174.

Zhang, G. (2000). Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30(4):451–462. ISSN 1094-6977.

Appendix A

Feature analysis

In this chapter, we present our experiments to choose the contextual features to be applied on the anomaly analysis step.

As discussed in Section 2.1, there are two types of anomalies: punctual and contextual. The punctual anomalies are those that present extreme or isolated patterns when compared to all instances. On the contrary, contextual anomalies are those with isolated behaviour within a context. Hence, two set of features are necessary: one to define the context and other to evaluate the behaviour. The context of an instance is defined by similar instances, called contextual neighbours. The behaviour is estimated through the comparison of the behavioural features of the instances with their contextual neighbours. The assumption is that the instances should present similar behaviour to their contextual neighbours in the behavioural space.

We want to analyze the healthcare demand of the Brazilian cities. As the country is huge and the cities are very different in many aspects, we believe that the context is very important. Thus, in this current work we perform contextual analysis.

As shown in Chapter 3, the behavioural feature is the rate of procedures due to its practical meaning: according to our goal in the anomaly analysis, a city is anomalous if the number of procedures performed in its population is unusual. As we have the population size of each city, we divide the number of procedures by the population so that we can compare cities of different sizes.

Once that we defined the behavioural feature, we need to choose the contextual feature. The chosen feature must satisfy the basic assumption that contextual neighbours should have small behavioural distance. Hence, the contextual analysis must create neighbourhoods with similar rates of procedures. As shown in Section 4.1, the information available about the cities are: location, population size, and *HDI*. These three information are the potential contextual features for our analysis.

Next we show how we use the three information for estimating the contextual distances between the cities.

Location: we have the location of each city expressed as $(latitude, longitude)$. From this information, we computed the Euclidean Distance of all pairs of cities. We refer to this distance as the geographical distance. For example, if two cities are located in coordinates $(-19, -43)$ and $(-18, -44)$, their geographical distance is 2.

Population: As we have the population of each city in each year, the population distance of each pair of cities is computed as the Euclidean Distance between their population considering each year between 2008 and 2012.

HDI: as we only have the cities *HDI* of the year of 2010, the *HDI* distance between two cities is estimated by the simple absolute difference in their *HDI* values.

We performed a statistical experiment to verify if the behavioural distance of the neighbourhoods produced with these contextual distances are significantly smaller than random neighbourhoods. The experiment consist of generating two distributions of behavioural distances: a random distribution and a contextual distribution. Both distribution are composed of 50,000 values of behavioural distances. The experiment implementation is shown in Algorithm 8.

The random distribution is composed of values of behavioural distance between random pairs of cities. The neighbourhood distribution is generated as described next. In each iteration, one city $c1$ is chosen and its K contextual neighbours are computed according to the candidate contextual feature. Then, one of these K contextual neighbours, $c2$, is chosen at random and the behavioural distance of the pair ($c1$,$c2$) is appended to the distribution.

Therefore, K represents the size of the neighbourhood from which we choose at random one neighbour. If K is small, it is likely that a close contextual neighbour is chosen. If K is large, a distant neighbour can be chosen and the contextual similarity might not be so small.

After generating the two distributions, a t-test is performed to evaluate their similarity: the p-value produced by the test is inversely proportional to the similarity of the two distributions. The p-value indicates the probability of observing the same results if no correlation exists. If the p-value is smaller than 0.05 and the average value of the neighbourhood distribution is smaller, we conclude that the behavioural distance between contextual neighbours is significantly smaller than random neighbours. If the p-value is greater than 0.05, we conclude that the contextual neighbours produced do not present similar behaviour.

When we increase K, we get closer to the random case. A good candidate feature

Algorithm 8 Experiment to check if the candidates of contextual feature produce contextual neighbourhood with small behavioural distance.

```
number_cities ← 5566
random_distribution ← new_array(50,000)
contextual_distribution ← new_array(50,000)

{Generate the random distribution.}
while (iteration ≠ 50,000) do
   {Choose at random the first city.}
   random_position_1 ← random_int(0, number_cities)
   c1 ← cities[random_position_1]
   {Choose at random the second city which must be different from the first city.}
   random_position_2 ← random_int(0, number_cities)
   while (random_position_2 = random_position_1) do
      random_position_2 ← random_int(0, number_cities)
   end while
   c2 ← cities[random_position_2]
   {Append their behavioural distance to the distribution.}
   random_distribution.append(Euclidean_distance(c1.rates, c2.rates))
   iteration += 1
end while
iteration ← 0

{Generate the contextual distribution.}
while (iteration ≠ 50,000) do
   {Choose at random the first city.}
   random_position_1 ← random_int(0, number_cities)
   c1 ← cities[random_position]
   contextual_distances ← new_array(K)
   {Iterate over all cities to look for the k closer cities.}
   for all (candidate in cities) do
      distance ← distance(c1.contextual_feature, candidate.contextual_feature)
      if (distance < max(neighbourhood_distances)) then
         contextual_distances.remove(max(neighbourhood_distances))
         contextual_distances.insert(distance)
      end if
   end for
   {Choose at random one of the k distance to append to the contextual distribution.}
   random_position_2 ← random_int(0,K)
   contextual_distribution.append(contextual_distances[random_position_2])
   iteration += 1
end while
```

for the contextual analysis should present ascending p-value as we increase the value of K: close neighbours should present smaller behavioural distances compared to random cities.

Tables A.1, A.2 and A.3 present the p-values produced in the experiment for the ten procedures with three different contextual features: population size, geographical distance and *HDI*, respectively. The neighbourhood size K was set to 5, 50, 500, 3000

Table A.1. P-value considering the population size as the contextual feature.

Procedure / K	5	50	500	3000	All cities
Arteriography	0.003535	0.002757	1.186e-05	0.03763	0.8436
Cardiovascular Surgery	3.852e-11	6.431e-11	< 2.2e-16	0.002195	0.7373
Glaucoma Surgery	0.5281	0.3004	0.2658	0.2621	0.9694
Highly Complex Orthop.	6.191e-15	1.021e-15	< 2.2e-16	0.06554	0.3534
Neurosurgery	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.0005257	0.8096
Obstetrics	< 2.2e-16	< 2.2e-16	< 2.2e-16	9.124e-08	0.5498
Oncology	2.577e-10	4.99e-13	< 2.2e-16	0.07692	0.1384
Scintigraphy	0.002151	0.001949	0.01419	0.9136	0.4142
Transplant	1.09e-07	1.031e-12	5.014e-13	0.6769	0.6507
Ultrasonography	2.279e-09	2.102e-07	1.718e-09	4.111e-05	0.8198

Table A.2. P-value considering the geographical location as the contextual feature.

Procedure / K	5	50	500	3000	All cities
Arteriography	< 2.2e-16	< 2.2e-16	0.1027	0.09721	0.928
Cardiovascular Surgery	< 2.2e-16	< 2.2e-16	< 2.2e-16	2.09e-15	0.6609
Glaucoma Surgery	< 2.2e-16	< 2.2e-16	5.287e-08	6.597e-16	0.01972
Highly Complex Orthop.	< 2.2e-16	< 2.2e-16	0.0003484	0.0022	0.2356
Neurosurgery	< 2.2e-16	< 2.2e-16	0.06736	0.5762	0.6601
Obstetrics	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.9025
Oncology	< 2.2e-16	< 2.2e-16	3.058e-12	0.0001262	0.4049
Scintigraphy	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.2673	0.1265
Transplant	< 2.2e-16	< 2.2e-16	0.434	0.06133	0.5515
Ultrasonography	< 2.2e-16	< 2.2e-16	4.523e-16	0.8387	0.05987

Table A.3. P-value considering the *HDI* as the contextual feature.

Procedure / K	5	50	500	3000	All cities
Arteriography	0.02297	0.1501	0.01645	0.8827	0.9752
Cardiovascular Surgery	< 2.2e-16	< 2.2e-16	< 2.2e-16	3.652e-10	0.1006
Glaucoma Surgery	3.177e-05	0.003674	0.0225	0.1856	0.5353
Highly Complex Orthop.	4.505e-11	3.481e-15	2.963e-15	0.3465	0.5083
Neurosurgery	0.0102	0.0003325	7.586e-05	0.4703	0.4437
Obstetrics	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.8728
Oncology	< 2.2e-16	< 2.2e-16	3.32e-14	0.3674	0.01136
Scintigraphy	< 2.2e-16	< 2.2e-16	< 2.2e-16	4.112e-11	0.3939
Transplant	9.278e-07	1.703e-06	3.628e-07	0.2257	0.1463
Ultrasonography	< 2.2e-16	< 2.2e-16	1.311e-15	1.976e-07	0.6152

and to all the cities according to the number of cities that performed each procedure as shown in Table 4.2. In all executions, the average value of the random distribution is greater than the average in the neighbourhood distribution. Thus, we can interpret the results as how significantly smaller is the behavioural distance between neighbours compared to random pairs.

From the results, we conclude that all the three features are good candidates: with just few exceptions, the p-value is small when K is small and it increases when the value K becomes large. The meaning of these results is that cities with similar distance according to population size, geographical location or *HDI* have also similar rates of procedures.

As we could not choose the contextual feature with this experiment, we also created two new features to repeat the experiment, as shown next. With these new features, we avoid that geographically distant cities are considered contextual neighbours due to their similar *HDI* or population.

1. *HDI* distance weighted by the geographical distance: the contextual neighbours are those cities that are both geographic close and with similar development index.

2. population distance weighted by the geographical distance: contextual neighbours are close cities with similar population size.

Table A.4. P-value considering the *HDI* weighted by the geographical distance as the contextual feature.

Procedure / K	5	50	500	3000	All
Arteriography	< 2.2e-16	1.457e-11	0.395	0.03827	0.5243
Cardiovascular Surgery	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.095e-13	0.641
Glaucoma Surgery	< 2.2e-16	< 2.2e-16	2.722e-05	3.802e-06	0.8633
Highly Complex Orthop.	< 2.2e-16	< 2.2e-16	7.078e-06	0.03083	0.8115
Neurosurgery	< 2.2e-16	< 2.2e-16	0.01423	0.1478	0.02954
Obstetrics	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.02294
Oncology	< 2.2e-16	< 2.2e-16	1.415e-12	0.01213	0.9162
Scintigraphy	< 2.2e-16	< 2.2e-16	< 2.2e-16	5.108e-06	0.2444
Transplant	< 2.2e-16	8.111e-12	0.005658	0.7929	0.6275
Ultrasonography	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.066e-05	0.4717

The results of the experiment with these new features are shown in Tables A.4 (*HDI* weighted by the geographical distance) and A.5 (population size weighted by the geographical distance). Again, in all experiments the average distance of the neighbourhood distribution was smaller than in the random distribution. Thus, they show

Table A.5. P-value considering the population size weighted by the geographical distance as the contextual feature.

Procedure / K	5	50	500	3000	All
Arteriography	< 2.2e-16	0.001629	0.66	0.009262	0.3301
Cardiovascular Surgery	< 2.2e-16	< 2.2e-16	2.775e-16	0.3044	0.9485
Glaucoma Surgery	< 2.2e-16	7.858e-14	2.481e-05	0.2144	0.9127
Highly Complex Orthop.	< 2.2e-16	< 2.2e-16	9.447e-07	0.001916	0.1976
Neurosurgery	< 2.2e-16	< 2.2e-16	0.05109	3.593e-10	0.7829
Obstetrics	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.9713
Oncology	< 2.2e-16	< 2.2e-16	5.656e-06	0.00408	0.2188
Scintigraphy	< 2.2e-16	< 2.2e-16	8.87e-05	0.0003631	0.5979
Transplant	< 2.2e-16	7.286e-11	0.006763	1.966e-10	0.4098
Ultrasonography	< 2.2e-16	< 2.2e-16	7.207e-10	0.3893	0.6106

how significantly smaller is the behavioural distance between neighbours compared to random pairs.

Again, all the candidates for the contextual features are good and have great correlation with the behavioural distance. As all of them are similar in the quantitative analysis, our decision is based on the qualitative analysis: we believe that the geographic distance and the *HDI* are the most relevant features to set the cities context.

The *HDI* is generic and measure the overall situation of the cities according to the education, healthcare and economics aspects. The location is also important because cities located in the same region tends to present similar culture. Hence, the contextual distance between each pair of cities is given by the *HDI* distance weighted by their geographic distance.

After choosing the two features applied on the contextual analysis, we verified whether or not these two information are correlated. If the location and *HDI* present high correlation, it would not be necessary to combine them to generate the contextual feature. Otherwise, they complement each other and both are important to be considered.

Figure A.1 shows the relation between these two information: each point is a Brazilian city and its color indicates the *HDI* level. Although in the North the *HDI* tends to be smaller than in the south, it is possible to see that for most of the cities the geographic neighbours present varied *HDI* levels. Thus, we conclude that the two information complement each other.

Therefore, we shown that the contextual neighbours of a city should be the geographical close cities with similar development index.

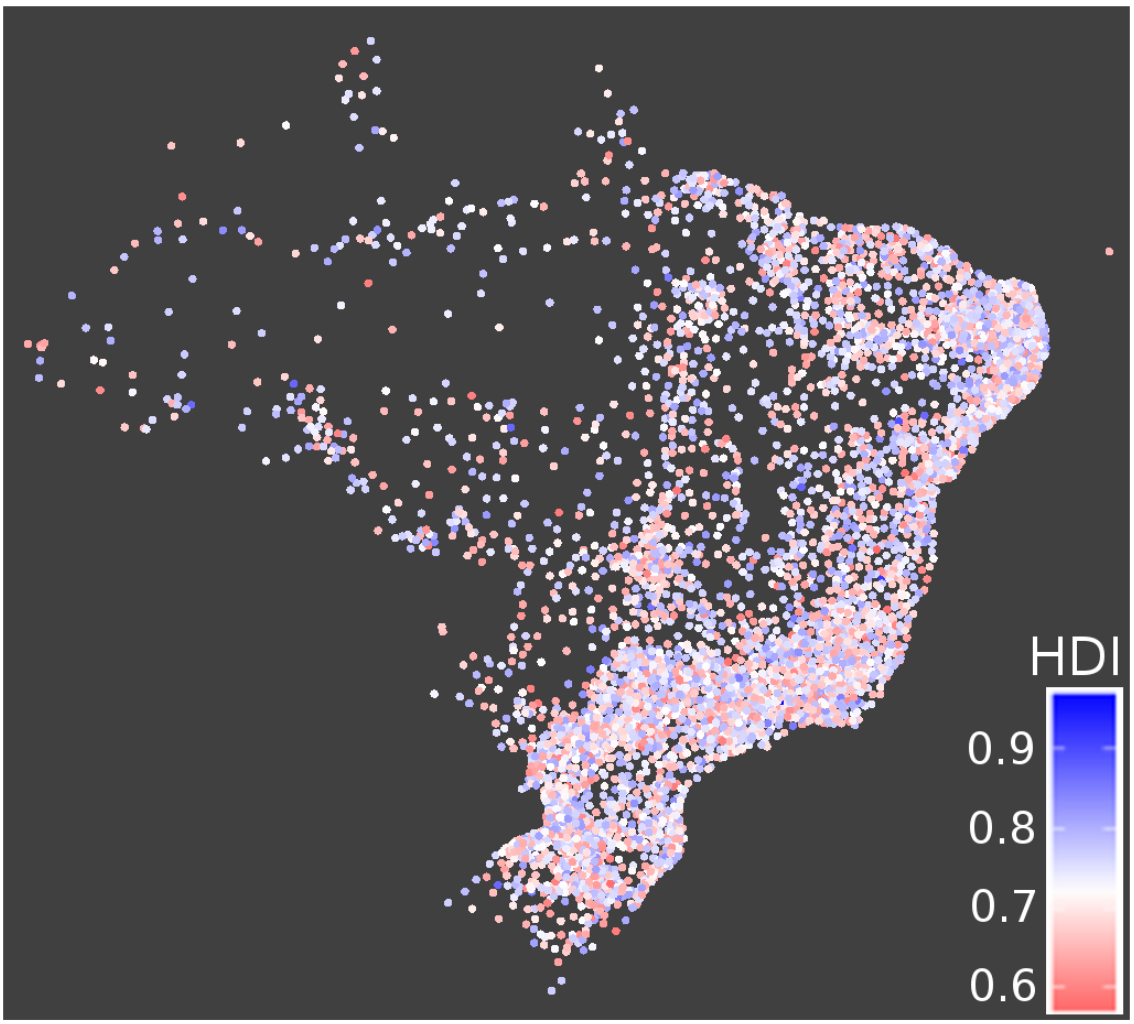

Figure A.1. *HDI* of the Brazilian cities.