Bruno Barbarioli

Anomaly Detection Under Cost Constraint

Belo Horizonte

2017, Fevereiro

Bruno Barbarioli

Anomaly Detection Under Cost Constraint

Universidade Federal de Minas Gerais – UFMG

Instituto de Ciências Exatas

Programa de Pós-Graduação

Supervisor: Renato Martins Assunção

Belo Horizonte
2017, Fevereiro

Este trabalho é dedicado à minha família pelo apoio durante essa longa trajetória.

Acknowledgements

Agradeço em primeiro lugar aos meus pais pelo apoio durante todos esse anos em que me dediquei à academia. Sem o seu suporte em todos os sentidos, esse caminho teria sido impossível.

Agradeço ao professor Renato Assunção por ter sido um excelente orientador, por dedicar seu tempo para me transmitir um pouco do seu conhecimento nos nossos encontros semanais. Agradeço também ao professor Marcos pelas disciplinas que cursei sob sua docência, pelos conselhos e as cervejas eventuais com os colegas de pós-graduação.

Agradeço aos colegas de mestrado (Jussiane, Arthur, Leonardo e Frederico) que me ajudaram durante as inúmeras dificuldades durante as disciplinas do curso. Muito grato também sou aos meus amigos do Leste (Larissa, Douglas, Luis, Milton) que me ajudaram nos meus primeiros passos no mundo da programação.

Por fim, agradeço à todos que de alguma forma contribuíram para a minha trajetória. O meu eventual sucesso também é de todos vocês.

Abstract

A detecção de anomalias é usualmente utilizada em análise de fraudes. Entretanto, restrições de orçamento podem tornar o processo impraticável quando um número grande anomalias é identificada. O presente trabalho propõe um método para selecionar casos probabilisticamente baseado no seus impactos, mas garantindo que a discrepância relativa entre os valores observados e os valores esperados seja levada em consideração. Ele usa uma modificação do *False Discovery Rate step-up procedure* para melhorar a precisão e garantir a escalabilidade. Aplica-se então o método proposto à um projeto destinado a monitorar o sistema de pagamentos do serviço de saúde público brasileiro a fim de encontrar comportamentos fraudulentos.

Palavras-chaves: Detecção de anomalias, Detecção de fraudes, Mineração de Dados, Aprendizado de Máquina, Classificação.

Abstract

Anomaly detection is commonly used in fraud analysis. However, budget constraints can turn the audit process infeasible when a sizable number of anomalies are identified. We propose a method to select cases probabilistically based on their impact, but guaranteeing that the relative discrepancy between the observed values and the expected behavior is also taken into account. It uses a modification of the False Discovery Rate step-up Benjamini Hochberg procedure to improve accuracy and insure scalability. We apply the proposed method to a project designed to monitor the Brazilian public health care payment system in search for fraudulent behavior.

Key-words: Anomaly Detection, Fraud Detection, Data Mining, Machine Learning, Classification.

Extended Abstract

Uma das principais aplicações das técnicas de mineração de dados é a detecção de fraudes e abusos. Existem diversos exemplos bem sucedidos do uso de tal técnica em instituições bancárias e de crediário, seguradoras, empresas de comunicação e de serviços de saúde. Recentemente técnicas estatísticas e de mineração de dados começaram a ser utilizadas para detectar casos suspeitos. Essas técnicas calculam um escore para cada caso dependendo dos aspectos que indicam um comportamento fraudulento, sendo que os casos que possuem um escore alto são classificados como suspeitos.

A auditoria não é somente uma ferramenta para encontrar e punir contraventores, mas ela também possui efeitos indiretos. Ela pode persuadir indivíduos a obedecer às regras se eles souberam que existe uma probabilidade positiva de serem auditados. Esse efeito preventivo é indireto, mas extremamente importante de qualquer forma. Fazer com que a probabilidade de ser auditado seja positiva é a razão pela qual os auditores sempre amostram pequenas empresas e indivíduos de baixa renda ao invés de se concentrar nos grandes pagadores.

No presente estudo, foi adaptado uma proposta inicial por Assunção e Pinheiro (1988) para amostrar unidades, desenvolvendo assim um método para detectar anomalias relacionadas à atividades fraudulentas sob a restrição de custos. A solução ingênua para esse problema seria focar nas possíveis fraudes de maior impacto financeiro. Essa solução, todavia, pode criar uma sensação de impunidade entre os menores fraudadores. Desenvolve-se assim um método que envolve todas as unidades sob avaliação, mas direciona de forma prioritária os recursos para as maiores entre elas. Isto é feito de tal forma que o número total de unidades avaliadas é limitado pelo total de recursos disponíveis para a auditoria. É criada assim uma curva de rejeição que classifica as unidades anômalas de forma probabilística, atribuindo assim uma probabilidade maior de classificação para aquelas unidades que possuem um tamanho maior. Utiliza-se em seguida um método completamente novo de classificação baseado na modificação do *False Discovery Rate step-up procedure* que não somente melhora a precisão mas também assegura a escalabilidade necessária para se avaliar um grande número de unidades.

A motivação por trás do presente estudo vem de um sistema de detecção de fraudes no pagamentos do serviço de saúde público brasileiro chamado InfoSAS. Tal sistema foi criado para identificar anomalias que poderiam ser fraudes ao erário. Ele processa todos os pagamentos feitos pelo Sistema Único de Saúde (SUS) para os mais de duzentos mil estabelecimentos provedores de serviços de saúde. A metodologia foi aplicada à um dos inúmeros procedimentos realizados pelos estabelecimento e assim pode-se verificar sua eficácia em classificar um determinado número de estabelecimentos como anômalos. Os resultados estão de acordo com o esperado, demonstrando assim a viabilidade do uso desse novo método para a classificação quando diante de limitações de recursos.

List of Figures

Figure 1 – Calculating the p-value using the *rejection curve* 18
Figure 2 – Size Variable Histogram. Size is the basis 10 logarithm of the annual expenditure of the health provider. 22
Figure 3 – Nonparametric Conditional Probability Curve $\mathbb{P}_Y(S)$ 22
Figure 4 – Rejection Curve $\mathbb{C}_Y(S)$ 23
Figure 5 – Rejection Curve $\mathbb{C}_Y(S)$ for different δ_0's. Dotted line: $\delta_0 = 0.001$; Continuous line: $\delta_0 = 0.01$; Dashed line: $\delta_0 = 0.025$. 24
Figure 6 – Rejection Curve $\mathbb{C}_Y(S)$ for different α's. Dotted line: $\alpha = 0.01$; Continuous line: $\alpha = 0.05$; Dashed line: $\alpha = 0.10$. 24

List of Tables

Table 1 – Number of errors committed when testing m hypothesis 17
Table 2 – Parametric Classification of Individual Scores 25
Table 3 – Nonparametric Classification of Individual Scores 25

Contents

1 INTRODUCTION 10

2 METHODOLOGY 12

2.1 Determining the Theoretical Rejection Curve 13

2.2 Estimation Procedure 14

2.2.1 Nonparametric Estimation 15

2.2.2 Parametric Estimation 15

2.2.3 Estimating the Rejection Curve 16

2.3 Multiple Hypothesis Testing 16

2.3.1 Bonferroni's Correction 16

2.3.2 False Discovery Rate (FDR) 16

3 HEALTHCARE PAYMENT SYSTEM OUTLIER DETECTION (HEPSOD) 19

4 USING HEPSOD IN THE INFOSAS 21

4.1 Parameter Estimation 21

4.2 Estimation of Individual Scores 25

4.3 The FDR Adaptive Step-up procedure 26

5 CONCLUSIONS 27

BIBLIOGRAPHY 29

1 Introduction

One of the main applications of data mining anomaly detection techniques is fraud and abuse identification. There are many documented successful examples of such techniques use in banking and credit card institutions (SUDJIANTO et al., 2010; CHAUDHARY; YADAV; MALLICK, 2012; CHAN et al., 1999; NGAI et al., 2011), insurance companies (VIAENE et al., 2002; TENNYSON; SALSAS-FORN, 2002), telecommunications (BECKER; VOLINSKY; WILKS, 2010; HILAS, 2009; ESTÉVEZ; HELD; PEREZ, 2006), and health care claim payments (BERWICK; HACKBARTH, 2012; WAN; SHASKY, 2012; LIOU; TANG; CHEN, 2008; ORTEGA; FIGUEROA; RUZ, 2006). Systems for processing data in these domains have been implemented and allow for automatic case selection for auditing. Recently, data mining and statistical techniques started to be used as tools to detect suspect cases (BOTS; LOHMAN, 2003; BHOWMIK, 2011). The data mining techniques calculate a score for each case depending on specific aspects that indicate a fraudulent behavior. Those cases with high scores are deemed suspicious. Audit is not only the tool to find and punish those committing frauds but it also has indirect effects. It can persuade individuals to comply with the rules if they know that there is a non-null probability of being audited. This preventive and deterrent effect is an indirect one, but extremely important nonetheless. Making this audit probability non-null for everyone is the reason why tax auditors always sample small business and taxpayers rather than concentrating only on the large ones (CLEARY; TAX, 2011).

The domain we are mainly interested is the detection of health care payment frauds. The health care economic sector is an attractive fraud target due to its large size and the volume of money involved. The National Health Care Anti-Fraud Association estimates that over $60 billion is lost to health care fraud in United States each year (see <www.nhcaa.org>). In this sector, after a preliminary spotting of suspicious cases, a time-consuming and expensive human audit process is needed to ascertain the validity of the evidence found, as well as posterior in-field to obtain confirmatory fraud demonstration (CHALKLEY; MALCOMSON, 2000; FU et al., 2004). Therefore, a false positive has a very high cost for the system: the in-field auditing of a false positive plus the opportunity cost of not visiting a truly positive case instead. In the case of health care payments frauds, the institutions face limited human resources to carry out the time-consuming and expensive auditing. At the same time, the relatively large number of high score suspicious cases (see section 3) represent far more than the resources allow to audit. This resource constraint problem creates a need for a systematic approach to identify which cases to audit.

In the present study, we adapt an initial proposal by (ASSUNçãO; PINHEIRO,

1988) for sampling units to develop a method to detect outliers related to fraudulent activities under a cost constraint. The naive solution would be to focus on the frauds that could have the largest financial impact. This solution has the unintended result of creating a sense of impunity among the smaller perpetrators. We designed a method that covers all units under scrutiny while directing more resources to the largest ones. This is done in such a way that the total number of cases to audit is constrained within an upper bound determined by the scarce resources available. It also introduces a completely new approach based on a modification of the False Discovery Rate (BENJAMINI, 2010) step-up Benjamini Hochberg (HB) procedure that not only improves the accuracy of the previous proposal, but it further insures the scalability necessary to evaluate a large number of establishments and medical procedures simultaneously.

The motivation behind the current study comes from a fraud detection system for health care services payments called InfoSAS. It is a system designed to detected anomalies that could possibility be frauds to the government. It processes all payments carried out by the public health system in Brazil to the more than 200 thousand health service providers. Presently, in each year, these services comprise more than 3.5 billion ambulatory and 11 million hospitalization procedures, classified in more than 5000 different types. It is currently one the largest data mining systems in Brazil and it identifies anomalous medical providers that should be visited in-field. We show how we combine the anomaly scores and achieve a balance between value involved and the need to not focus exclusively on large providers.

2 Methodology

Usually, anomaly scores are based on ratios between an observed value and a prescriptive or expected behavior under normal conditions, allowing for natural variation. Hence, the score can evaluate the *relative* amount of upper airway and neck surgeries in a given municipality administered by a single health service provider, and then it can be compared with the amount one expect based on the municipality population size and composition. This ratio misses the values involved in the case of equivalent fractions as, for example, when we have $110/100 = 110000/100000$. Therefore, besides the ratio itself, one must also take into account the values involved. However, if we focus on the difference we can miss gross relative discrepancies, stimulating the occurrence of frauds among small size establishments. We describe next a principled way to take both aspects into account.

The method can be summarized as follows. Let Y be a variable obtained as a ratio between observed and expected quantities. Therefore, Y is a measure of *relative* discrepancy. The larger the value of Y, the more evidence we have for its anomalous behavior. Let S be a size variable correlated with Y and α be the general proportion of cases that can be audited, a value determined by the available resources. Let $\mathbb{P}_Y(S)$ be the conditional probability of finding an anomalous case given that the size is S. We determine a function $\mathbb{C}_Y(S)$, called *rejection curve*, giving a Y-threshold for each size S. This curve is computed to satisfy two conditions: the global probability of some case being anomalous is α; given its size S, a case will be anomalous with the desired probability $\mathbb{P}_Y(S)$.

As a second step we use a modified version of the False Discovery Rate (FDR) procedure to improve the accuracy of the classification and insure its scalability. Let m be the number of hypotheses tests being performed and m_0 the number of true null hypotheses among them. In this case, one hypotheses test is done for each observation, i.e. for each of the medical procedure performed by each establishment. The null hypotheses represents that the establishment has a normal score compared to its peers, whereas the alternative hypotheses classifies it as anomalous. After the *rejection curve* has been determined, we use it to calculate a p-value for each observation under analysis. Then we use the BH step-up procedure to classify them and control for the false positive results among all rejected null hypotheses. We adapted the method initially proposed by (BENJAMINI; HOCHBERG, 2000) and further extended by (STOREY; TAYLOR; SIEGMUND, 2004) (BENJAMINI; KRIEGER; YEKUTIELI, 2006) to restrict to α the proportion of observations classified as anomalous. The adaptive procedure in its standard format requires us to estimate m_0, but instead of doing so we use $(1 - \alpha) * m$ as its value, since we already know the number of observations that we want to classify as abnormal.

2.1 Determining the Theoretical Rejection Curve

There is a duality between $\mathbb{C}_Y(S)$ and $\mathbb{P}_Y(S)$ in the sense that, once the distributions of Y and S are known, one of the functions can be obtained from the other. In the present approach, we determine the function $\mathbb{P}_Y(S)$ from a probability model and a set of constraints. After that, the $\mathbb{C}_Y(S)$ curve is determined.

We impose a monotonicity constrain: $\mathbb{P}_Y(S)$ increases with S in order to direct the resources to the larger units, with potentially greater financial consequences. Besides that, two other restrictions are applied to the function:

$$0 \leq \mathbb{P}_Y(S) \leq 1\,,$$

as it must if it is a probability, and the expected value of this random probability must satisfy

$$\mathbb{E}(\mathbb{P}_Y(S)) = \alpha \tag{2.1}$$

It is important to notice that $\mathbb{P}_Y(S)$ will be *greater* than α for large sized establishments and smaller that α for the small ones, reaching the overall unconditional level α.

Taking these restrictions into consideration, the following model was chosen due to its simplicity for the conditional probability function:

$$\mathbb{P}_Y(S) = Ae^{BS} + C \tag{2.2}$$

where A, B and C are constants to be determined. Due to the previous constraints the following inequalities must be satisfied:

$$A, B < 0$$

$$1 > C > 0$$

$$|A| < C$$

We select two size values $s_0 < s_1$ and apportion the percentage α of items that can be audited:

$$\begin{aligned} \mathbb{P}_Y(s_0) &= Ae^{Bs_0} + C = \delta_0\alpha \qquad (2.3)\\ \mathbb{P}_Y(s_1) &= Ae^{Bs_1} + C = \delta_1\alpha \qquad (2.4) \end{aligned}$$

where δ_0 and δ_1 are between 0 and 1. Substituting the chosen function (2.2) in (2.1) we obtain:

$$E(\mathbb{P}_Y(S)) = A\psi_S(B) + C = \alpha \tag{2.5}$$

where $\psi_S(B) = E(e^{BS})$ is the moment generating function of S evaluated at B.

The system of equations has a trivial solution when we set $\delta_0 = \delta_1 = 1$ which is given by $B = 0$ and $A + C = \alpha$. This solution means that the rejection curve assumes a constant value which is equal to the $1 - \alpha$ percentile of the Y distribution. To ensure that the $\mathbb{P}_Y(S)$ is strictly increasing in S we impose the restrictions $\delta_0 \neq \delta_1$ and $B < 0$.

The equations (2.3), (2.4) and (2.5) can be rewritten as follows:

$$\begin{aligned} A &= \frac{(\delta_1 - \delta_0)\alpha}{e^{Bs_1} - e^{Bs_0}} \\ C &= [\delta_1 - \frac{(\delta_1 - \delta_0)e^{Bs_1}}{e^{Bs_1} - e^{Bs_0}}]\alpha \\ \frac{\psi_S(B) - e^{Bs_1}}{e^{Bs_1} - e^{Bs_0}} &= \frac{1 - \delta_1}{\delta_1 - \delta_0} \end{aligned}$$

The last three equations can be simplified if we consider $\delta_1 = 1$ implying that the probability of the analyzed variable being classified as an anomaly is α when its corresponding size is s_1. Thus, the equations become:

$$A = \frac{(1 - \delta_0)\alpha}{e^{Bs_1} - e^{Bs_0}} \tag{2.6}$$

$$C = \frac{(\delta_0 e^{Bs_1} - e^{Bs_0})\alpha}{e^{Bs_1} - e^{Bs_0}} \tag{2.7}$$

$$\psi_S(B) = e^{Bs_1} \tag{2.8}$$

Once we have obtained A, B, and C, we can determine the rejection curve $\mathbb{C}_Y(S)$.

The rejection curve determination is based on the idea that values of the variable Y that are greater than a threshold should be classified as outliers. This threshold is given by the rejection curve and thus it translates into the following:

$$\mathbb{P}_Y(S) = P(Y > \mathbb{C}_Y(S)|S) = 1 - F_Y(\mathbb{C}_Y(S)|S) \tag{2.9}$$

where $F_Y(\mathbb{C}_Y(S)|S)$ is the cumulative distribution function of Y conditioned on S. If we assume that the variable Y and the size variable S are independent (2.9) becomes

$$\mathbb{P}_Y(S) = 1 - \mathbb{F}_Y(\mathbb{C}_Y(S)) \tag{2.10}$$

where $\mathbb{F}_Y(.)$ is the cumulative distribution function of Y. Hence,

$$\mathbb{C}_Y(S) = \mathbb{F}_Y^{-1}(1 - \mathbb{P}_Y(T)) \tag{2.11}$$

and the threshold is given by the $1 - \mathbb{P}_Y(S)$ quantile of the Y distribution.

2.2 Estimation Procedure

The Y and S distributions are unknown, and thus we use estimation procedures to determine the rejection curve. The S distribution is used to determine $\mathbb{P}_Y(S)$ through its constants A,B and C and the Y distribution is used to determine $\mathbb{C}_Y(S)$. Equation (2.8) provides an estimate for B. Two different procedures are used to solve this equation, a parametric and a nonparametric.

2.2.1 Nonparametric Estimation

The nonparametric procedure uses the well known result that the moment generating function uniquely determines the probability distribution of a random variable (see (BILLINGSLEY, 1995, p. 342–345)). No assumption is made about the distribution of the size variable S except that it has all its moments $\mathbb{E}(S^k) < \infty$ and hence its moment generating function $\psi_S(t)$ is well defined. This is an easily satisfied requirement in virtually all practical cases. So, for large enough moment order p, we have:

$$\psi_S(B) \approx 1 + \sum_{i=1}^{p} \frac{m(i)B^i}{i!}$$

where $m(i)$ represents the i moment of the variable S. Replacing the parameters $m(i)$ with the sampled ones, the estimation of B is given by the solution of the following equation:

$$1 + \sum_{i=1}^{p} \frac{\hat{m}(i)B^i}{i!} = e^{Bs_1} \tag{2.12}$$

The value of p is determined simultaneously with B. First, an initial value p_0 is used to compute the first estimate B_0 of B. Second, a new value $p_1 = p_0 + 1$ is chosen with its estimate B_1. This procedure is carried out until convergence of B is reached. The number of iterations is used as the value of p.

2.2.2 Parametric Estimation

In this case the size variable S is assumed to have a distribution $\mathbb{P}_{\hat{\theta}}$ with $\hat{\theta}$ parameters. After the parameters are estimated we can then proceed to the estimation of B from the following equation:

$$\psi_{S,\hat{\theta}}(B) = e^{Bs_1}$$

For example, if we assume that $S \sim N(\mu, \sigma^2)$ and represent the parameters estimates by $\hat{\mu}$ and $\hat{\sigma}^2$, we can obtain $\hat{B}$ as the solution of the following equation:

$$e^{\hat{\mu}B + \frac{\hat{\sigma}^2 B^2}{2}} = e^{Bs_1}$$

which is:

$$\hat{B} = \frac{2(s_1 - \hat{\mu})}{\hat{\sigma}^2}$$

Sometimes it will not be possible to determine a closed-form solution to (8) and numeric methods need to be used to find a solution. As an example, if $S \sim \Gamma(\alpha, \beta)$, and $\hat{\alpha}$ and $\hat{\beta}$ are the associated estimators, equation (8) turns out to be:

$$(\frac{\hat{\beta}}{\hat{\beta} - B})^{\hat{\alpha}} = e^{Bs_1} \ ,$$

which has no closed-form solution. Once $\hat{B}$ is estimated, the constants A and C can be obtained from (6) and (7).

2.2.3 Estimating the Rejection Curve

The methodology used to estimate $\mathbb{C}_Y(S)$ is nonparametric and it is based on the function $\mathbb{P}_Y(S)$ and on the distribution of the variable being evaluated Y. We can write (11) as

$$\begin{aligned} F_Y(\mathbb{C}_Y(S)) &= \theta \qquad (2.13) \\ 1 - \mathbb{P}_Y(S) &= \theta \qquad (2.14) \end{aligned}$$

The idea is to vary θ and solve these equations. As the size variable S is limited to an interval (s_{min}, s_{max}), the $\mathbb{C}_Y(S)$ curve only needs to be determined on this interval. This means using only values of θ in the interval $[1 - \mathbb{P}_Y(s_{max}), 1 - \mathbb{P}_Y(s_{min})]$. Once the $\mathbb{C}_Y(S)$ points are determined from equations (2.13) and (2.14), we can use a nonparametric procedure to fit the rejection curve.

2.3 Multiple Hypothesis Testing

In our application, we have more than one ratio variable Y. We wish to obtain a single set of anomalous observations, regardless of the number of ratio variables being evaluated. In order to do so we use two different procedures to deal with the multiple hypothesis testing problem: the Bonferroni's Correction and the False Discovery Rate (FDR).

2.3.1 Bonferroni's Correction

We use the Bonferroni procedure to split the global α probability value. Each individual rejection curve will select a percentage α/n of its observations as outliers, where n is the number of ratio variables being used for audit.

2.3.2 False Discovery Rate (FDR)

The FDR procedure can be used either to enable multiple hypothesis testing or to refine the result obtained when using a single ratio variable Y. We briefly introduce the method consecrated in (BENJAMINI; HOCHBERG, 1995), however focusing on the adapted procedures that led to a extension of our method.

Let R be the number of hypothesis rejected, and let m be the number of hypothesis being tested of which m_0 are true, as defined previously. Table 1 describes the possible outcomes.

The number of hypothesis being tested are the observations being classified, so it is known in advance. R is traditionally an observable random variable, whereas U, V, S and T are unobserved random variables. The proportion of null hypothesis being erroneously

Table 1 – Number of errors committed when testing m hypothesis

	Declared non-significant	Declared significant	Total
True null hypothesis	U	V	m_0
Non-true null hypothesis	T	S	$m - m_0$
	m-R	R	m

rejected among all rejected ones is $Q = V/(V+S)$, which is a random variable as well. We define the FDR Q_e to be the expectation of Q,

$$Q_e = E(Q) = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)$$

When testing the hypothesis $H_1, H_2, ..., H_m$ let $P_1, P_2, ..., P_m$ be the corresponding p-values. Let $P_{(1)} \leqslant P_{(2)} \leqslant, ..., \leqslant P_{(m)}$ be the ordered p-values where H_i denotes the null hypothesis corresponding to P_i. Define the multiple hypothesis test as the following: let k be the largest i for which $P_i \leqslant \frac{i}{m}q$, where q is the level at which we wish to control Q similarly to the level os significance in the traditional hypothesis test. Then reject all $H_i, i = 1, 2, ..., k$. This procedure came to be known as Benjamini – Hochberg Step-up procedure (BH Step-up).

For independent test statistics the above procedure controls the FDR at q, i.e:

$$E(Q) \leqslant \frac{m_0}{m} q \leqslant q \tag{2.15}$$

An extension of the original FDR procedure was introduced by (BENJAMINI; KRIEGER; YEKUTIELI, 2006), which is called Adaptive Linear Step-up procedure. As we can see from equation 15 the term m_0/m is a factor of conservativeness, since it is always less or equal to 1 due to the fact that $m_0 \leqslant m$. If m_0 was known, we could alter the the BH Step-up and instead of using $\frac{i}{m}q$ we could use $\frac{i}{m_0}q$ which would in turn control the FDR at exactly q, i.e:

$$E(Q) = q \tag{2.16}$$

As m_0 is traditionally unknown, different estimation procedures have been created to estimate its value based on the entire sample of p-values. A few examples can be found in (BENJAMINI; HOCHBERG, 2000), (STOREY; TAYLOR; SIEGMUND, 2004), (YEKUTIELI; BENJAMINI, 1999).

When applying the FDR procedure to anomaly detection using the *rejection curve*, the first step is to calculate the p-value of each observation under analysis. According to

(HOGG; CRAIG, 1995)[p. 255–256] a p-value is the observed probability of a statistics being at least as extreme as the particular observed value, when the null hypothesis is true. Therefore, we devise a method to calculate the p-value using the *rejection curve* as displayed on Figure 1.

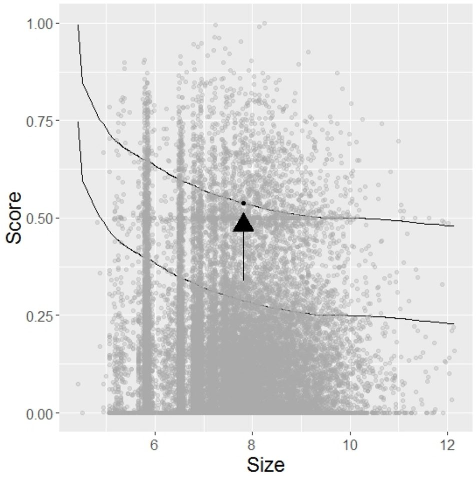

Figure 1 – Calculating the p-value using the *rejection curve*

The method consists on dislocating the *rejection curve* to each observation, and then calculating the p-value as the number of establishments above the curve divided by the total number of establishments. Since it uses the *rejection curve* as the threshold for the "at least as extreme as the particular observed value", it ascertains that the size of each observation is taking into account when calculating the p-value.

After the p-values are calculated and sorted we use a modification of the Adaptive Linear Step-up procedure described earlier to classify the observations. Normally, in order to use this method we would have to estimate m_0. However, in our particular case this is not necessary. Since we know how many observations we want to classify as anomalous, we know the true number of null hypotheses being tested, i.e, we know the exact value of m_0. Since α is the number of observations being classified as anomalous culminating on having the null hypothesis rejected, m_0 can be defined as:

$$m_0 = (1 - \alpha) \times m$$

Finally, we can use the Adaptive Linear Step-up procedure to classify the observations. However, in lieu of defining q as we usually do in a hypothesis test, we are going to use it as a tuning parameter. The idea is that allowing for a greater number of false discoveries among the discoveries, we can improve the classifier's accuracy regarding α . Therefore, there is a *trade-off* between accuracy of the classifier and possible false classifications.

3 Healthcare Payment System Outlier Detection (HEPSOD)

The Brazilian health care system can be described as a hybrid between a public system, called Sistema Único de Saúde (SUS) and a private one, the Sistema de Saúde Suplementar (SSAM). The SUS is one of the largest public health care systems in the world, providing services ranging from simple ambulatory procedures to organ transplants. The Brazilian Constitution guarantees free and universal access to every Brazilian and 80% of the its population depends solely on SUS health care services. The numbers are staggering. In the year of 2014, 4.1 billion ambulatory procedures, 1.4 billion medical appointments, 11.5 million hospitalizations, 19 million oncology procedures and 3.1 million chemotherapy procedures were performed through SUS. The Brazilian Health Minister had a R$91.5 billion budget for 2015, which happened to be the largest among all federal government sectors.

This massive infrastructure created the necessity of a monitoring system to evaluate and oversee the expenditures in order to detect any illegal activities within the health care providers. The federal health care department in conjunction with the Universidade Federal de Minas Gerais created the InfoSAS project with such intent. This is one of the largest data mining projects ever implemented in Brazil and it is designed to monitor the health care procedures using 4 large public data systems: the hospital information system (SIH), the ambulatory system (SIA), the healthcare establishment database (CNES) and the population data from the Instituto Brasileiro de Geografia e Estatística (IBGE).

The InfoSAS project has two assumptions: the fraudulent behavior occurs either when the service provider charges more for a procedure than its real value, or when it files for reimbursement of a number of procedures greater than what it has actually performed. For each case, the team created different algorithms to evaluate and detect anomalies within the database so as to guide the audit department to work in a more efficient way. On a monthly basis, the production and population data are inputed on a basic fact sheet that feeds the data mining algorithms. These algorithms generate a more comprehensive fact sheet containing tables and graphics providing detailed information about a procedure within a establishment during a period of time.

In the present study the algorithms being used deal with the second case, regarding the number of procedures of a given type carried out by each health care provider. These algorithms use time series analysis to attribute a discrepancy score to a procedure type within a single health care provider during a fixed period of time. Even though procedures are computed within a single health care provider, they are compared with their peers all around the country. There are five such algorithms that are then combined into a single one, making it easier to evaluate. This merger takes into consideration the weight that each of those algorithms have on the anomaly detection efficacy, according to health experts that worked in the project. The score created by this technique range from 0 to 1, where 0 represents that the establishment being assessed conforms with the expected level of production while 1 represents a total inconformity with the expected level. Therefore, we need to fix a threshold between 0 and 1 to classify the different establishments between normal and anomalous.

4 Using HEPSOD in the INFOSAS

The time period analyzed is from first of January/2014 to 31st of December/2014. The audit case selection procedure starts by choosing the variable that measures the size of the health care establishments under assessment. We choose the annual expenditure of the health provider. Since this variable has a highly asymmetric distribution, we took its logarithm in basis 10 as the size variable S. Second, we have to fix the number of anomalous observations that are going to be audited in the field. This input is set by the resources available for the auditing task. In this paper, only for illustrative purposes, we selected $\alpha = 0.05$. We included an analysis of how changes in the value of this parameter affect the conditional probability and the rejection curve. We have set $s_0 = 4.42 = \min\{S_i\}$, the minimum of the observed sizes, and its relative weight of rejection δ_0 as 0.01. The median size was the value chosen as $s_1 = 7.29$. We also included an analysis of how δ_0 influences the rejection curve.

The analysis goes as follows: first we apply the classification procedure to the unified score analyzing all of its instances. Then we apply it to the individual scores that compose the unified one and compare the combined results. Finally, we implement the FDR Adaptive Step-up Procedure to the unified score in order to improve its classification accuracy.

4.1 Parameter Estimation

For the parametric procedure, the first step is to analyze the logarithm of the annual expenditure distribution and determine if a theoretical and known distribution fits closely to the empirical data. As we can see from Figure 2, the histogram of the logarithm transformation of annual expenditure does not follow any easily recognizable distribution. A Gaussian distribution seems to fit well the half upper part of the distribution but the peaks on its lower tail prevents a Gaussian as a good model for these data.

Resorting to the non-parametric approach, we calculate the empirical moments up to $p = 6$ and estimate the parameter B in equation (2.12) using the moment generating function. More specifically, using the empirical moments $\hat{m}(i)$, we must find B such that

$$\begin{aligned} e^{7.29B} &= 1 + \frac{7.37B}{1} + \frac{55.79B^2}{2} + \frac{432.73B^3}{6} + \\ &\quad \frac{3438.88B^4}{24} + \frac{27975.82B^5}{120} + \frac{232741.4B^6}{720} \end{aligned}$$

The conditional probability function that results from this nonparametric approach is:

$$\mathbb{P}_Y(S) = 0.1704 - 0.2891e^{-0.1202S}$$

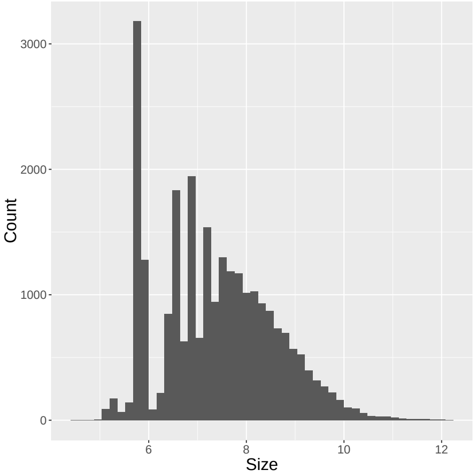

Figure 2 – Size Variable Histogram. Size is the basis 10 logarithm of the annual expenditure of the health provider.

Figure 3 shows the graph of $\mathbb{P}_Y(S)$ versus the size variable S. The conditional probability has the specified functional form. As the establishment size increases, the probability of classifying a score as anomalous raises as well. Observe that $\mathbb{P}_Y(S) > \alpha = 0.05$ for $S > 7.2$, reaching almost $2\alpha = 0.10$ when $S \approx 12$ while it is smaller than α for $S < 7.2$. For the same relative discrepancy on the variable Y, we reject the case more easily if S is large than when S is small. However, this is done in such a way that we still have an overall rejection rate equal to α. The inequality $|A| < C$ has not been satisfied. However, this inequality exists to assure that the probability function only has values between 0 and 1, where it is defined. This is not a problem in this particular case, because the size variable has values in the interval $(4.42, 12.14)$ thus maintaining the probability function within its theoretical limits.

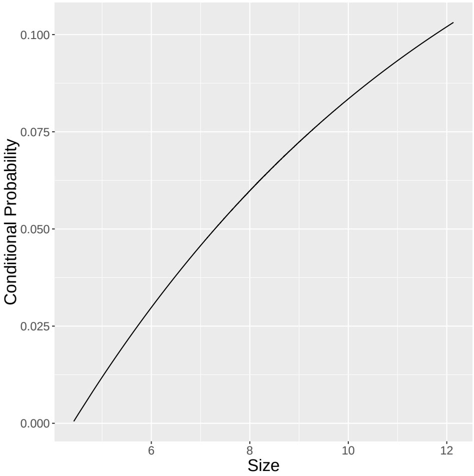

Figure 3 – Nonparametric Conditional Probability Curve $\mathbb{P}_Y(S)$

Remarkably, we obtain very similar results using the parametric approach based on a Gaussian approximation for the empirical distribution of S shown in Figure 2. The conditional probability function that results from the parametric estimation is:

$$\mathbb{P}_Y(S) = 0.1689 - 0.2882e^{-0.1214S}$$

After obtaining $\mathbb{P}_Y(S)$, we determine the rejection curve $\mathbb{C}_Y(S)$. This curve depends on the specific variable Y under monitoring. We illustrate the results for one target score variable, *Paediatric Clinic: Treatment of infectious and parasitic diseases*. This is a discrepancy score unifying five other discrepancy scores, all of them based on the ratio variables between the amount of these procedures carried out by one health provider and what we expect based on the population size and composition it serves. Figure 4 shows that the rejection curve has a functional form such that as the establishment size increases the threshold decreases. Hence, the model works as intended, and the number of scores classified as anomalous is 6.55% of the total. This is approximately what we had determined when we set the theoretical overall level as $\alpha = 5\%$.

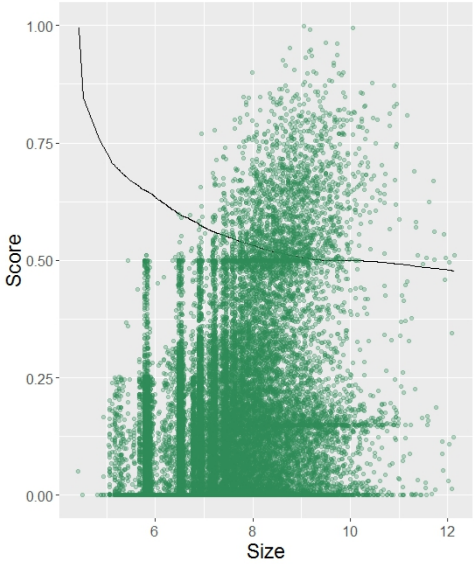

Figure 4 – Rejection Curve $\mathbb{C}_Y(S)$

We included an analysis of how different values of δ_0 affect the rejection curve, as it determines the probability of classifying an observation of size s_0 as anomalous. Since s_0 was set to be $s_0 = 4.42 = \min\{S_i\}$, i.e. the minimum of the observed sizes, we can use it as a tuning parameter that determines the relative slope of the curve. In our example it resulted in almost identical rejection curves as we can see from Figure 5. This outcome is due to the limited number of observations with relative small sizes compared to the entire population. Since the curve has only a few observations to classify at small sizes, the δ_0 does not influence its overall shape and slope, so the choice among $\delta_0 = 0.001$, $\delta_0 = 0.01$ and $\delta_0 = 0.025$ does not affect the classification.

A comparison can be made for different levels of α. In Figure 6 we traced the rejection curve for 3 different levels of $\alpha : 0.01, 0.05, 0.10$. We can see that, as we choose

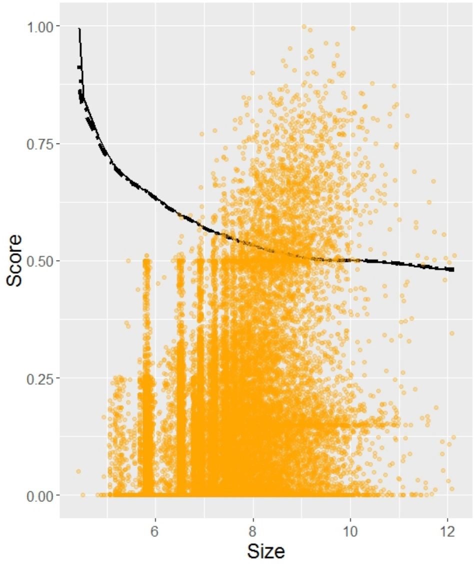

Figure 5 – Rejection Curve $\mathbb{C}_Y(S)$ for different δ_0's. Dotted line: $\delta_0 = 0.001$; Continuous line: $\delta_0 = 0.01$; Dashed line: $\delta_0 = 0.025$.

to select less anomalous cases to audit, the rejection curve moves upwards. It takes an interesting shape for $\alpha = 0.10$ due to the score distribution shape, as it has to contour a densely populated section of the observation points. The percentage of anomalous observations for $\alpha = 0.01, 0.05, 0.10$ is equal to 1.40%, 6.55%, and 11.60%, respectively, close to the α target.

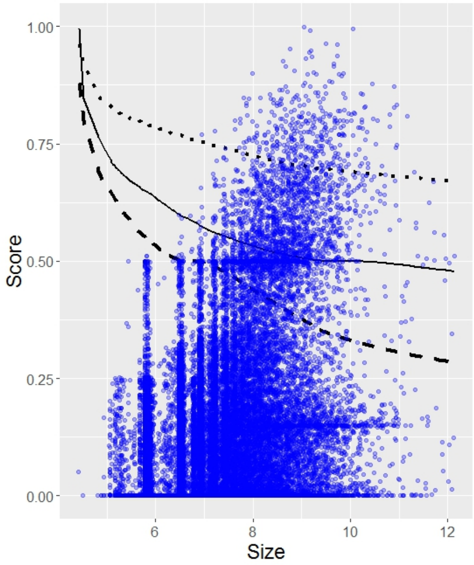

Figure 6 – Rejection Curve $\mathbb{C}_Y(S)$ for different α's. Dotted line: $\alpha = 0.01$; Continuous line: $\alpha = 0.05$; Dashed line: $\alpha = 0.10$.

4.2 Estimation of Individual Scores

In this section, we apply our case selection method to the discrepancy scores that compose the unified score analyzed in the previous section. As explained at the end of section 3, this unified score is composed by five individual scores capturing different discrepancy aspects of the time series of procedures carried out by the health care provider. All scores represent different algorithms applied to the same database of a single health care procedure provided by different establishments. The results for the parametric estimations are displayed on Table 2 and for the nonparametric estimation are displayed on Table 3. The global level of significance is $\alpha = 0.05$ and we used the Bonferroni correction with the same weight given by the unified method to create each individual rejection curve, instead of simply dividing by n.

Table 2 – Parametric Classification of Individual Scores

Score	A	B	C	Weight	Anomalous
KNN	-0.0432	-0.1214	0.0253	0.15	268 (1.05%)
Insta	-0.0288	-0.1214	0.0169	0.10	149 (0.59%)
Med Est	-0.0288	-0.1214	0.0169	0.10	154 (0.60%)
Med Mun	-0.0432	-0.1214	0.0253	0.15	281 (1.10%)
Produc	-0.1441	-0.1214	0.0845	0.50	822 (3.23%)

The expected number of anomalous observations are 0.75%, 0.5%, 0.5%, 0.75% and 2.5% and the respective results from the methodology are 1.05%, 0.59%, 0.60%, 1.10%, 3.23%. We can see that they are close to the expected numbers, and even when applied to scores with different distributions the outcomes are similar, indicating that the method is robust to the variable's distribution.

Table 3 – Nonparametric Classification of Individual Scores

Score	A	B	C	Weight	Anomalous
KNN	-0.0430	-0.1202	0.0256	0.15	267 (1.05%)
Insta	-0.0289	-0.1202	0.0170	0.10	148 (0.59%)
Med Est	-0.0289	-0.1202	0.0170	0.10	154 (0.60%)
Med Mun	-0.0430	-0.1202	0.0256	0.15	282 (1.10%)
Produc	-0.1446	-0.1202	0.0852	0.50	822 (3.23%)

We can see that the results of the parametric and nonparametric estimation are very similar. When combined, the number of observations classified as anomalous is equal to 4.4% of the total number of cases, disregarding the estimation procedure. Some establishments are classified as anomalous by more than one score, and thus generating an overlapping when merged together. This situation implies that the total number of anomalous establishments analyzed by the individual scores sums up to a smaller number than the total one under the unified score. Another point to be taken into consideration is that virtually all of the establishments that were classified as anomalous by the individual scores were also classified as such by the unified one. We are studying ways to take into account the correlation between individual discrepancy scores to better control this joint behavior.

4.3 The FDR Adaptive Step-up procedure

Lastly, we apply the FDR procedure to the unified score in order to improve its accuracy. The result obtained when we applied the classification procedure using the nonparametric estimation was 6.55% when α was set to 5%, i.e. roughly 30% over our intended target.

In order to use the FDR procedure we first calculate the p-value of each establishment under analysis. The algorithm responsible for this operation has a complexity of $\theta(n^2)$ and due to the massive amount of observations, approximately 25000, we had to devise an approach to reduce the computational time. Instead of calculating the p-value of each observation, we used the fact that the *rejection curve* systematically overestimate the number of observations being classified as anomalous, and calculated the p-values only of those that were initially above the original curve which we will call m'. This technique reduces in 2 orders of magnitude the number of observations being evaluated and thus reduces considerably the computational cost.

After calculating the p-values we need to calculate m_0, the number of true null hypothesis. Since the new m' under evaluation is 1668 after we have reduced the scope of the classification procedure, we have:

$$m_0 = m' - \alpha \times m = 1668 - 0.05 \times 25461 = 395$$

We can now use the Adaptive Step-up Procedure to classify the observations. We use the q in the $\frac{i}{m_0}q$ as a tuning parameter to adjust the accuracy of our classification regarding how close the number of anomalous observations is to α. Bearing in mind that q is also the number of false discoveries among the discoveries, we need to be aware of the *trade-off* that such technique imposes. The results were remarkable nonetheless, as the number of observations being classified as anomalous were 5.02%, within less than 1% of the chosen α. The q resulting from the tuning procedure was 0.015, which does not result in a great number of false discoveries among the ones obtained.

5 Conclusions

Anomaly detection has become an important field due to its broad applicability and the impact of big data in commercial, social and scientific environments. As a result, intense research of new techniques and methods to detect and classify their presence are underway, and specific approaches are being crafted to deal with individual problems (CHANDOLA; BANERJEE; KUMAR, 2009).

In the present study, we created a method to detect anomalies related to fraudulent activities. There are two different types of repercussions of such work. The direct effects can be listed as: recover undue payments; public sanction; law enforcement applicability visible within the community and thus encouraging others to comply. The indirect effects can be summed as deterrent and preventive effects on other local level audits.

In this particular case, budget constraints in the institution responsible for the audit process creates a need for limiting the number of observations being classified as anomalous. The naive solution would be to focus on the frauds that could have the biggest impact, financially or else wise. This solution may lead to smaller perpetrators knowing they are free to act. We designed a method that covers all units under scrutiny while directing the resources primarily to the largest ones.

The methodology showed promising results when applied to the InfoSAS project. We selected one of several medical procedures being monitored and adopted its scores as the variable being evaluated by the method. The research was conducted using both parametric and nonparametric estimation, and two different approaches, the first using unified scores and the second, using its components. They resulted in different outcomes, whereas the first one classified more observations as anomalous than initially expected, the second one classified less than what was initially intended.

We select a functional specification for $\mathbb{P}_Y(S)$ that suits our aims. However, the methodology we propose is general and can be adapted for any other parametric specification of the probability curve. Although, it does not seem to be an easy way to extend our method for a case where one does not assume a parametric shape for $\mathbb{P}_Y(S)$. It is an interesting research issue to verify what can be done if we assume only general properties for this probability function, such as its continuity and increasing monotonicity.

The modified version of the FDR Adaptive Step-up procedure that we applied to our classification problem resulted in a dramatically increase in accuracy of the *rejection curve*. It also solves the scalability problem with the Bonferroni correction, since the InfoSAS detection system monitors a large number n of ratio variables, and Bonferrori does not scale well when n is large. The rejection level associated with each ratio variable becomes α/n, too small to be useful when dealing with a large number of variables. The FDR does not have this issue since it stacks the observations on a single classification procedure, after their appropriate p-values have been calculated.

Bibliography

ASSUNçãO, R.; PINHEIRO, J. Crítica de razões no censo industrial de 1985. *Revista Brasileira de Estatística*, v. 49, p. 101–118, 1988. Citado 2 vezes nas páginas 6 and 11.

BECKER, R. A.; VOLINSKY, C.; WILKS, A. R. Fraud detection in telecommunications: History and lessons learned. *Technometrics*, Taylor & Francis, v. 52, n. 1, p. 20–33, 2010. Citado na página 10.

BENJAMINI, Y. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 72, n. 4, p. 405–416, 2010. Citado na página 11.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, JSTOR, p. 289–300, 1995. Citado na página 16.

BENJAMINI, Y.; HOCHBERG, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, Sage Publications Sage CA: Los Angeles, CA, v. 25, n. 1, p. 60–83, 2000. Citado 2 vezes nas páginas 12 and 17.

BENJAMINI, Y.; KRIEGER, A. M.; YEKUTIELI, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, JSTOR, p. 491–507, 2006. Citado 2 vezes nas páginas 12 and 17.

BERWICK, D. M.; HACKBARTH, A. D. Eliminating waste in us health care. *Jama*, American Medical Association, v. 307, n. 14, p. 1513–1516, 2012. Citado na página 10.

BHOWMIK, R. Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, Citeseer, v. 2, n. 4, p. 156–162, 2011. Citado na página 10.

BILLINGSLEY, P. *Probability and measure. wiley series in probability and mathematical statistics.* [S.l.]: Wiley New York, 1995. Citado na página 15.

BOTS, P. W.; LOHMAN, F. A. Estimating the added value of data mining: A study for the dutch internal revenue service. *International Journal of Technology, Policy and Management*, Inderscience Publishers, v. 3, n. 3-4, p. 380–395, 2003. Citado na página 10.

CHALKLEY, M.; MALCOMSON, J. M. Government purchasing of health services. *Handbook of health economics*, Elsevier, v. 1, p. 847–890, 2000. Citado na página 10.

CHAN, P. K. et al. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and their Applications*, IEEE, v. 14, n. 6, p. 67–74, 1999. Citado na página 10.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 15, 2009. Citado na página 27.

CHAUDHARY, K.; YADAV, J.; MALLICK, B. A review of fraud detection techniques: Credit card. *International Journal of Computer Applications*, International Journal of Computer Applications, 244 5 th Avenue,# 1526, New York, NY 10001, USA India, v. 45, n. 1, p. 39–44, 2012. Citado na página 10.

CLEARY, D.; TAX, R. I. Predictive analytics in the public sector: Using data mining to assist better target selection for audit. In: *The Proceedings of the 11th European Conference on EGovernment: Faculty of Administration, University of Ljubljana, Ljubljana, Slovenia*. [S.l.: s.n.], 2011. p. 168. Citado na página 10.

ESTÉVEZ, P. A.; HELD, C. M.; PEREZ, C. A. Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications*, Elsevier, v. 31, n. 2, p. 337–344, 2006. Citado na página 10.

FU, H.-H. et al. Application of a single sampling plan for auditing medical-claim payments made by taiwan national health insurance. *Health policy*, Elsevier, v. 70, n. 2, p. 185–195, 2004. Citado na página 10.

HILAS, C. S. Designing an expert system for fraud detection in private telecommunications networks. *Expert Systems with applications*, Elsevier, v. 36, n. 9, p. 11559–11569, 2009. Citado na página 10.

HOGG, R. V.; CRAIG, A. T. *Introduction to mathematical statistics.(5"" edition)*. [S.l.]: Upper Saddle River, New Jersey: Prentice Hall, 1995. Citado na página 18.

LIOU, F.-M.; TANG, Y.-C.; CHEN, J.-Y. Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health care management science*, Springer, v. 11, n. 4, p. 353–358, 2008. Citado na página 10.

NGAI, E. et al. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, Elsevier, v. 50, n. 3, p. 559–569, 2011. Citado na página 10.

ORTEGA, P. A.; FIGUEROA, C. J.; RUZ, G. A. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN*, Citeseer, v. 6, p. 26–29, 2006. Citado na página 10.

STOREY, J. D.; TAYLOR, J. E.; SIEGMUND, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 66, n. 1, p. 187–205, 2004. Citado 2 vezes nas páginas 12 and 17.

SUDJIANTO, A. et al. Statistical methods for fighting financial crimes. *Technometrics*, Taylor & Francis, v. 52, n. 1, p. 5–19, 2010. Citado na página 10.

TENNYSON, S.; SALSAS-FORN, P. Claims auditing in automobile insurance: fraud detection and deterrence objectives. *Journal of Risk and Insurance*, Wiley Online Library, v. 69, n. 3, p. 289–308, 2002. Citado na página 10.

VIAENE, S. et al. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, Wiley Online Library, v. 69, n. 3, p. 373–421, 2002. Citado na página 10.

WAN, T. T.; SHASKY, C. A. Mining medical claims data with exploratory to confirmatory statistical methods. *International Journal of Public Policy*, Inderscience Publishers Ltd, v. 8, n. 1-3, p. 122–135, 2012. Citado na página 10.

YEKUTIELI, D.; BENJAMINI, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, Elsevier, v. 82, n. 1, p. 171–196, 1999. Citado na página 17.