# DETECÇÃO DE OUTLIERS EM DESPESAS GOVERNAMENTAIS COMO MECANISMO DE SUPORTE A AUDITORIA E COMBATE A CORRUPÇÃO

***ALCANTARA, FRANK***

***TSUNODA, DENISE***

**Resumo**
O direito a prestação de contas passa, obrigatoriamente, pela clareza, precisão e disponibilidade das informações relativas às contas públicas. Nosso país encontra-se em um momento de mudanças, transparência e combate a corrupção. Este trabalho mostra que é possível criar um mecanismo de auditoria fiscal e suporte ao combate a corrupção utilizando técnicas de mineração de dados para destacar anomalias (outliers) nas contas públicas. Utilizou-se dados reais oriundos do Portal da Transparência do Governo Federal, relativos as compras, ainda na fase de empenho, selecionados aleatoriamente nos lançamentos efetuados no primeiro semestre de 2011. Estes dados foram submetidos a um processo de mineração para a seleção e identificação de despesas anômalas através do uso de um algoritmo de agrupamento baseado em distância. Foram destacados os dez registros com maior índice de anomalia.
**Palavras Chaves:** *Outliers, corrupção, mineração de dados, contas públicas, detecção de anomalias.*

## 1. CONTEÚDO

## 2. INTRODUÇÃO

O tema corrupção parece estar consolidado na agenda de preocupações brasileiras. Apesar de jamais sair de pauta, existem evidências de que o problema não é enfrentado de maneira firme e resolutiva (PEREIRA, 2005). Não há indício mais claro desta falta de combate que o não aparecimento da corrupção entre as *preocupações capazes de tirar o sono dos brasileiros* (G1, 2011). Mesmo sem tirar o sono, a corrupção foi responsabilizada pelo desperdício de R$ 41,5 bilhões por ano, valores de 2008 (DECOMTEC - FUNDAÇÃO DAS INDÚSTRIAS DO ESTADO DE SÃO PAULO, 2010), valor mais que suficiente para justificar a criação de toda e qualquer ferramenta, ou processo, que ajude a minimizar este problema.

Frequentemente quantização da corrupção, quantitativa ou qualitativamente, está limitado aos resultados de pesquisas de organizações não governamentais, nacionais ou internacionais. Anualmente a *Transparency International* publica o relatório: *Corruption Perception Index*. Um relatório que classifica os países de acordo com a percepção da corrupção do setor público. Em 2010 o Brasil apareceu na 69º posição. Atrás de países como Ruanda, Gana, Namíbia, Arábia Saudita, Costa Rica, Coreia do Sul, Uruguai e Chile, entre outros (TRANSPARENCY INTERNATIONAL, 2010). Uma lista diversa que não apresenta nenhum fator exclusivo que indique uma característica (PIB, etinia, colonização, religião) específica que justifique esta colocação.

Desde a promulgação da Lei de Responsabilidade Fiscal (Lei Complementar nº 101, de 4.5.2000) que o Brasil está obrigado a diminuir os efeitos da corrupção através da criação de novos mecanismos de controle e a imposição de restrições à conduta dos administradores públicos (CALAU e FORTIS, 2006). Guiados por esta lei os governos estaduais, municipais e federal patrocinam a criação de portais para divulgação de informações, financeiras e fiscais, visando tornar transparente a administração dos órgãos de governos. Transparência aqui entendida como sendo a produção e divulgação sistemática de informações ao público, notadamente via internet. A transparência é um dos pilares em que se assenta a Lei de Responsabilidade Fiscal (CALAU e FORTIS, 2006) e a principal razão do Brasil aparecer entre os melhores países do mundo em outro relatório internacional. No Open Budget Survey da International Budget Partnership, que classifica os países de acordo com a abertura dos orçamentos governamentais. Resultado da transparência nos dados, acesso à informação orçamentária, participação pública e facilidade de aferição. Em 2010, no Open Budget Survey, o Brasil ficou na nona colocação atrás apenas do Chile, EUA, Suécia, Noruega, França, Reino Unido, Nova Zelândia e África do Sul.

A transparência, por si e em si, é apenas um dos fatores que podem ser utilizados para minimizar a percepção de corrupção do governo. O controle das ações governamentais é outro fator importante para este mesmo objetiv006F. Controle exercido através da fiscalização da conduta financeira e jurídica dos funcionários e de mecanismos de controle externo, como tribunais de contas, auditorias, comissões de inquérito; controle social, exercido tanto pela mídia como por grupos organizados da sociedade civil (PEREIRA, 2005).

Dentre os esforços de transparência destaca-se o Portal da Transparência do Governo Federal do Brasil (GOVERNO FEDERAL, 2010). O Portal da Transparência sobressai tanto graças a qualidade da informação disponível quanto pela quantidade destas informações. Considerando-se o tamanho do estado brasileiro, a quantidade de dados disponíveis e a organização destes dados acabam dificultando o acompanhamento detalhado das ações fiscais e financeiras do governo. Na página Detalhamento Diário de Despesas (GOVERNO FEDERAL, 2004), por exemplo, as despesas federais podem ser consultadas de acordo com um sistema de filtros de seleção, com três opções: Empenho, liquidação e pagamento, Atendendo o que foi definido pela lei LEI No 4.320, DE 17 DE MARÇO DE 1964 (PRESIDÊNCIA DA REPÚBLICA, 1964). O empenho representa o primeiro estágio da despesa orçamentária. É registrado no momento fiscal da contratação do serviço, aquisição do

material ou bem, obra ou amortização da dívida (GOVERNO FEDERAL, 2010). Do ponto de vista do controle e transparência o empenho distingue-se devido ao imediatismo e primazia. Segundo a Lei nº 4.320-64, não pode existir pagamento sem liquidação e esta, por sua vez, não poderá ocorrer sem a criação do empenho respectivo (PRESIDÊNCIA DA REPÚBLICA, 1964).

No caso do governo federal, visto pelo Portal da Transparência, no período observado, são emitidos, em média, 7500 itens de empenho por dia. Cada um destes itens está sob o controle de uma Unidade Gestora, de um Órgão Superior, de uma Entidade Vinculada e, por último do Tribunal de Contas da União. Responsável final pelo julgamento de todas as contas de administradores públicos e demais responsáveis por dinheiros, bens e valores públicos federais, bem como as contas de qualquer pessoa que der causa a perda, extravio ou outra irregularidade de que resulte prejuízo ao erário. Tal competência administrativa-judicante, entre outras, está prevista no art. 71 da Constituição brasileira (TRIBUNAL DE CONTAS DA UNIÃO, 2010).

Este artigo apresenta uma técnica de mineração de dados para a detecção de empenhos anômalos, com o objetivo de destacar procedimentos suspeitos e facilitar a detecção de erros. Sejam eles causados por erros inocentes ou por má fé. Para isto serão utilizadas técnicas de detecção de outliers.

Um dos problemas básicos da mineração de dados (além dos problemas de classificação, agrupamento, predição e associação) é a detecção de outliers, ou anomalias (BEN-GAL, 2005). A detecção de outliers constituí-se na busca por objetos em um conjunto de dados que não obedecem às leis que são válidas para a maior parte dos elementos contidos neste conjunto (PETROVSKIY, 2003). Um outlier é um dado com comportamento tão diferente dos seus semelhantes que desperta a necessidade da investigação. Os resultados da detecção são apresentados para análise humana, de forma que, o resultado da análise depende da qualidade da mineração de dados (PETROVSKIY, 2003).

As técnicas de detecção de outliers podem ser utilizadas para localizar itens de empenho, nos dados do Portal da Transparência, que apresentem qualquer comportamento anômalo e destacar estes empenhos para análise posterior. Do ponto de vista deste artigo, não há diferença entre os comportamentos anômalos causados por corrupção e má fé, por erros inocentes ou de digitação, ou por procedimentos administrativos perfeitamente corretos e legais, mas pouco usuais.

## 3. REFERENCIAL TEÓRICO

Em síntese todas as técnicas de detecção de outliers se referem à seleção de elementos amostrais cujo comportamento é, de alguma forma, discrepante quando comparado ao comportamento dos elementos ao seu redor. Cada autor, ou linha de pesquisa, se refere às técnicas de descoberta e separação destes elementos de uma forma diferente adequada as características dos dados utilizados ou a ciência predominante no estudo. Detecção de outliers, detecção de anomalias, detecção de ruído, detecção de desvio e mineração de exceção (HODGE e AUSTIN, 2004) são alguns dos termos usados frequentemente por autores das áreas de mineração de dados, e estatística, para identificar estas anomalias. Estas exceções, estes outliers, estes elementos dissonantes, ocorrem devido a falhas mecânicas, erros inocentes, fraudes ou condições naturais que provocam alterações sensíveis em suas características matemáticas, ou físicas, criando um comportamento distinto entre seus pares. Na literatura de mineração de dados, pura e específica, os outliers são frequentemente citados nos capítulos referentes ao tratamento prévio, ou adequação dos dados (CIOS, PEDRYCZ, et al., 2007) (HAND, MANNILA e SMYTH, 2001) (HAN e KAMBER, 2006). Nestas aplicações clássicas e exclusivas o objetivo mais comum é descobrir um comportamento padrão, um modelo, uma regra, uma associação que permita inferir novos conhecimentos. Neste cenário os outliers são indesejados. A ocorrência de um ou mais elementos que não obedecem à regra geral pode ser o suficiente para distorcer ou invalidar o modelo.

Em ambientes financeiros (WESTPHAL, 2009), na medicina (KUMAR, KUMAR e SINGH, 2008), em segurança de redes (MOHAMED e KAVITHA, 2011), redes sociais (TAYLOR & FRANCIS GROUP, 2009), física quântica (WEINSTEIN, 2009), astronomia (ZHANG, LUO e ZHAO, 2005) e mesmo em pesquisas geoquímicas (FILZMOSERA, GARRETTB e REIMANN, 2004), frequentemente são exatamente os outliers que contém o conhecimento mais valioso e merecem análises detalhadas e cuidadosas.

De acordo com a participação humana, as técnicas de detecção de outliers podem ser classificadas em supervisionadas e não supervisionadas. Uma técnica de detecção supervisionada considera um conjunto de dados para treinamento que contém elementos pré-classificados tanto na classe normais quanto na classe outliers. As técnicas de detecção não supervisionadas não fazem nenhuma consideração inicial e não conhecem classes de treinamento (GOGOI, BHATTACHARYYA, et al., 2011).

De acordo com o número de dimensões contidas nos dados as técnicas de detecção podem ser classificadas em univariadas ou multivariadas. A técnica de análise univariada procura fazer inferências trabalhando com cada variável contida nos dados de forma isolada. É apropriada para situações onde as variáveis são independentes. Esta limitação restringe o uso destas técnicas de detecção em bancos de dados reais onde, frequentemente, existe um grande número de variáveis interdependentes e inter-relacionadas. Alguns autores que utilizam as técnicas de detecção multivariada dividem os outliers em brutos e estruturais. Brutos são os outliers cuja discrepância é causada por um ou mais atributos individuais. Um outlier bruto é aquele cuja discrepância é causada em uma dimensão por, no mínimo, uma variável. Estruturais são os outliers que não podem ser classificados como brutos (ZHANG, LUO e ZHAO, 2005).

As técnicas de detecção de outliers ainda podem ser divididas em gráficas e estatísticas. Técnicas gráficas, tais como o box, scatter e spin plots, podem ser utilizadas para a localização visual de outliers (HAND, MANNILA e SMYTH, 2001) em problemas com até três variáveis interdependentes. Acima disso, e sem supervisão, recorre-se às técnicas estatísticas. Técnicas estatísticas, ou métodos, estatísticos, utilizados para a detecção de outliers são chamados de robustos.

Quanto aos tipos de algoritmos (ALI e XIANG, 2010) classificam-se as técnicas de detecção de outliers em: Baseados em distribuição estatística, identificam os outliers considerando o

modelo de distribuição ou de probabilidade; Baseados em distância, considera os elementos da amostra que não possuem vizinhos próximos como outliers (KNORR, NG e TUCAKOV, 2000) (RAMASWAMY, RASTOGI e SHIM, 2000); Baseados em densidade local, identifica os outliers considerando a distância entre todos os elementos em um espaço dimensional (BREUNIG, KRIEGEL, et al., 2000); Baseados em desvio, um algoritmo linear que funciona localizando uma série de dados familiar e considera qualquer distúrbio na série como sendo um outlier (ARNING, AGRAVAL e RAGHAVAN, 1996); Baseados em frequência de padrão, define o grau de anomalia de um elemento considerando a quantidade de padrões frequentes ele que contém (HE, XU, et al., 2005).

Como os métodos de detecção de outliers são baseados em conjuntos dispares de pressupostos, e conceitos, a comparação entre estes métodos é, frequentemente, infrutífera. Contudo, a eficiência de métodos baseados nos mesmos conceitos e pressupostos é frequentemente avaliada (OTEY, PARTHASARATHY e GHOTING, 2005) em busca de algoritmos melhores e mais rápidos, ou simplesmente para validação de eficiência em estruturas de dados diferentes. Por sua própria natureza, a estrutura de dados que gera o outlier determina que método será mais eficiente para sua detecção. Sendo assim, não existe um método único, universalmente aplicável, ou genérico, para a detecção de outliers (HODGE e AUSTIN, 2004).

## 4. METODOLOGIA

No Portal da Transparência os dados de Despesas Diárias podem ser consultados por período, Fase da despesa, órgão Superior e Favorecido, como apresentado na Figura 1:

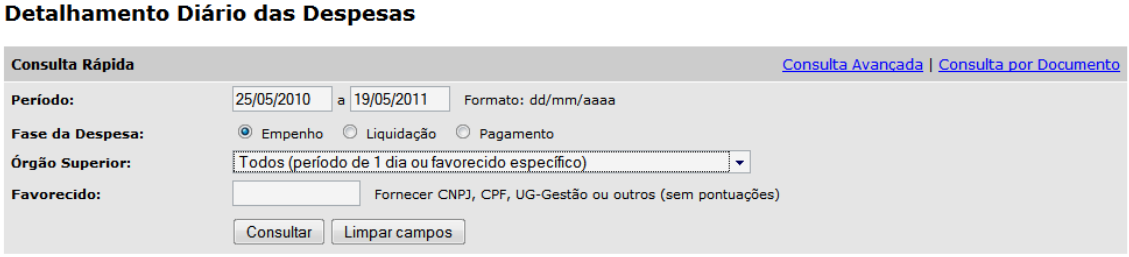


**Figura 1 - Caixa de Seleção de Despesas**

Foi necessária a criação de um banco de dados para seleção e mineração, com os dados diários das despesas do governo Estes dados estão disponíveis online na forma de tabelas em páginas HTML. Desta forma, foi necessária a criação um web crawler1 especificamente para recuperar estes dados. O crawler foi desenvolvido em PHP 5.3.5 e o banco de dados escolhido foi o MySql 5.0.7.

Este estudo utiliza a primeira fase do processo de despesa orçamentária: O empenho. Os dados mais antigos disponíveis no portal datam do dia 25 de junho de 2010. Aparentemente existe uma janela de dados flutuante com duração de aproximadamente um ano, limitando e determinando os dados que podem ser utilizados. Cada consulta diária retorna uma tabela com 15 linhas de documentos de empenho. Cada linha corresponde a um documento e contém seu código, sua data, o Órgão Superior Órgão / Entidade Vinculada, a Unidade Gestora, o Elemento de Despesa, o Favorecido e o Valor.

Cada documento de empenho contém uma tabela com os dados referentes aos materiais, serviços, bens ou obras empenhados além do detalhamento dos responsáveis pelo empenho, liquidação e pagamento da despesa. Em média, cada dia retorna aproximadamente 450 páginas de tabelas de 15 linhas ou 7500 documentos de empenho, por sua vez cada um contendo um número médio de dois itens. Na nossa amostra consiste de dez dias com um total de 4383 páginas, 64.873 documentos de empenho e 123.818 itens de empenho.

O crawler desenvolvido sorteia, de forma randômica, um dia de semana no primeiro semestre de 2011 e, varrendo as páginas de listagem de documentos de empenho e armazena estas páginas em uma tabela banco de dados (rawdata). Paralelamente, um segundo processo, varre a tabela rawdata, identifica os links referentes aos documentos de empenho e carrega cada um dos itens destes documentos em outra tabela (detalha_gastos). Estes dois processos, independentes reduzem o tempo de amostragem de cada item enquanto permitem a verificação da validade dos dados comparando a soma do valor total dos itens do empenho com o valor total apresentado na listagem diária. Um fragmento do modelo de entidades e relacionamentos (MER) deste banco de dados está apresentado na Figura 2.

[1] A expressão em inglês crawler ou web crawler representa uma classe de programas capazes de recolher informações em sites web e armazenar estas informações para uso posterior.

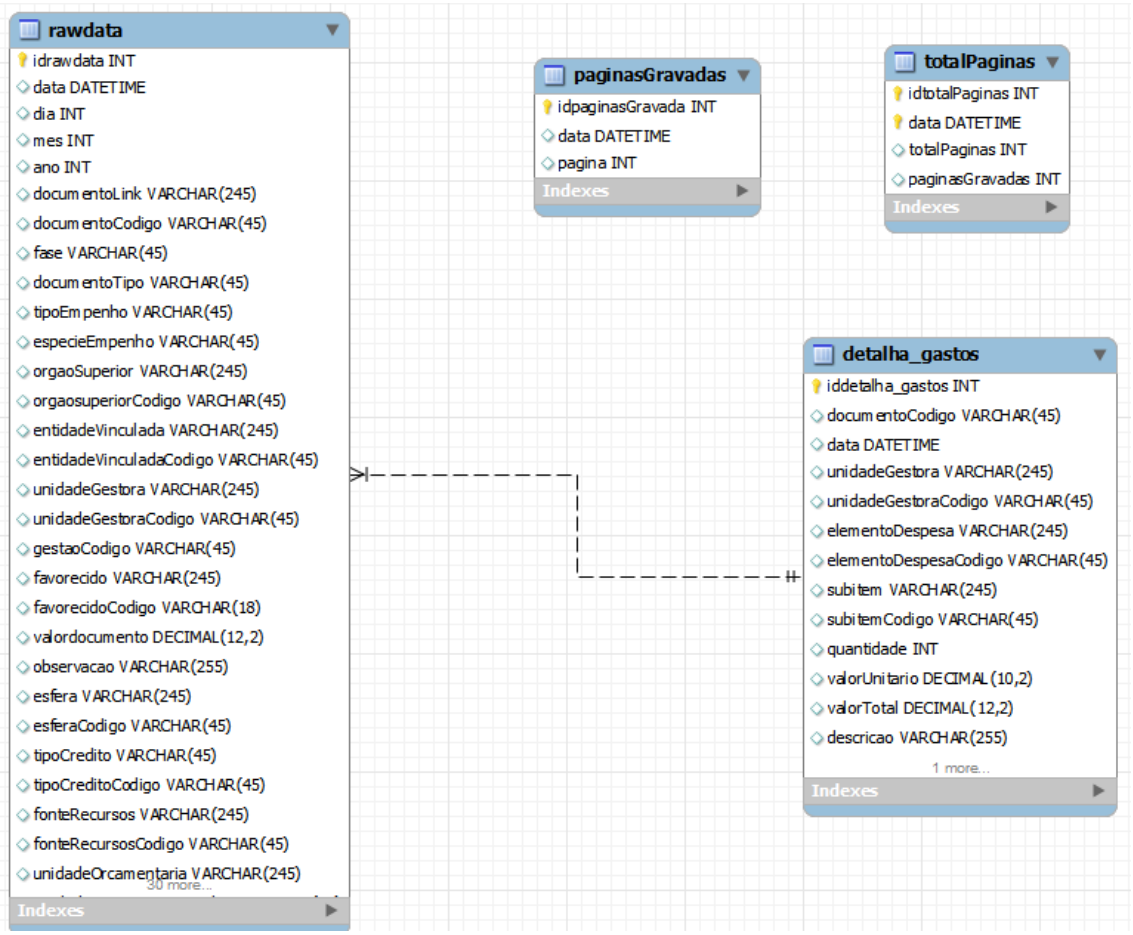


**Figura 2 - MER para recolhimento de dados**

A amostragem probabilística foi realizada com um sorteio randômico de dez dias através de uma chamada a API do site Random.org. O site usa ruído atmosférico para gerar números randômicos verdadeiros (RANDOM.ORG, 2010). A amostragem escolhida atende as características de validade, confiança e precisão para uma amostragem probabilística (LEVY e LEMESHOW, 1999). Para realizar a amostragem foi criada uma chamada à API do Random.org para recolher 1000 números randômicos de quatro dígitos. Destes foram selecionados os 10 primeiros números adequados ao formato dd-mm, com algumas restrições: O dia, dd-mm, escolhido estaria, forçosamente, entre o dia primeiro de janeiro de 2011 e o dia 27 de junho de 2011; O dia deveria estar entre segunda e sexta-feira. Foram selecionados os dias 24-05, 14-02, 17-06, 11-04, 25-03, 20-06, 09-06, 25-01, 19-05, 02-03, nesta ordem, que produziram um total de 123.818 itens de empenho.

Durante o processo de amostragem foram perdidas 22 páginas devido a erros na formatação da página, o que inviabilizou o processo de armazenamento, ou erros provocados por problemas no próprio Portal da Transparência, que inviabilizaram a recuperação das páginas.

Uma vez que os empenhos necessários foram recuperados do Portal da Transparência, os dados da tabela *detalha_gastos* foram convertidos em um arquivo de textos no formato adequado a importação de dados para uso no aplicativo Rapidminer da Rapid-i (RAPID-I, 2010).

O algoritmo de detecção de outliers escolhido é baseado em distância, como proposto por Ramaswamy, Rastogi e Shim (2000). Tal algoritmo utiliza a distância $D^k(p)$ para representar a distância entre o ponto $p$ e os seu elemento $k^{th}$ (k-ésimo) vizinho. Classificando os pontos de acordo com sua distância $D^k(p)$, os n pontos quaisquer com maior distância serão os outliers desejados (RAMASWAMY, RASTOGI e SHIM, 2000). É possível a utilização de

qualquer métrica para medir a distância entre estes pontos e, em alguns casos, como por exemplo, em mineração de textos, medidas de distâncias não métricas podem ser utilizadas, permitindo um alto grau de generalização para este método (RAMASWAMY, RASTOGI e SHIM, 2000). A seguir está o algoritmo desenvolvido por Ramaswamy, Ratogi e Shim para o cálculo da distância $D^k(p)$:

```
Procedure getKthNeighborDist(Root,p,k,minDkDist)
Begin
    nodeList := {Root}
    p.Dkdist := ∞
    nearHeap := ∅
    While nodeList is not empty do{
        delete the frist element, Node from nodeList
        if (Node is a leaf){
            for each point q in Node do
                if (dist(Ip,q) < p.DkDist){
                    nearHeap.insert(q)
                    if nearHeap.numPoings() > k) nearHeap.deleteTop()
                    if nearHeap.numPoings() = k)
                        p.DkDist = dist(p, nearHeap.top())
                    if (p.Dkdist ≤ minDkDist) return
                }
        }
        else {
            append Node's children to nodeList
            sort modeList by MINDIST
        }
        for each Node in nodeList do
            If (p.DkDist ≤ MINDIST(p.Node))
                Delete Node from nodeList
    }
End
```

Como alternativa a criação das funções necessárias em Perl, como fizeram os autores (RAMASWAMY, RASTOGI e SHIM, 2000), ou outra linguagem de programação, este estudo utilizou um software livre, licenciado sob GPL para criar uma estrutura de processamento que pudesse ser facilmente reproduzida. Sem custos de aquisição de software e sem a necessidade do domínio de linguagens de programação. Neste software, o Rapidminer, foi criado um processo padrão contendo: Um módulo de leitura de arquivos csv (read csv), um módulo de amostragem (sample) e um módulo para executar a detecção de outliers (detect outlier) por distância segundo o algoritmo descrito por Ramaswamy, Rastogi e Shim (2000). A configuração dos módulos utilizados foi realizada com a manutenção dos valores padrão apresentados na inserção dos módulos. Sendo necessário realizar apenas as seguintes alterações: Incluir o path do arquivo csv que será importando pelo módulo de leitura de csv (Read CSV); limitar em 2000 elementos o valor máximo de amostras permitidas pelo módulo de amostragem (sample); Definir, tanto do número $k^{th}$ de vizinhos, quanto do número n de outliers desejados, na configuração do módulo de detecção de outliers (Detect Outlier), para este estudo ambos devem ser configurados para o valor 10; Escolher o procedimento para o cálculo das distâncias $D^k(p)$ na configuração do módulo de detecção de outliers (Detect Outlier). Para este estudo escolheu-se a distância euclidiana.

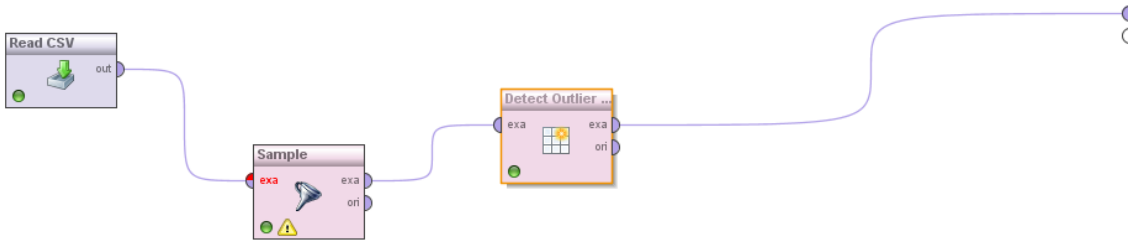


**Figura 3 - Processo de Detecção no Rapidminer**

Ainda considerando a simplicidade do processamento, o processo desenhado ficou contido em si mesmo. Desta forma, o resultado do processamento contendo os outliers é enviado apenas para a própria interface de operação do Rapidminer sendo visível em tela.
Para a execução do processo no Rapidminer foi necessária à instalação do módulo parallel processing da própria Rapid-I e o aumento na capacidade de memória RAM do computador de testes de 1Gbytes para 4Gbytes.

## 5. RESULTADOS E CONSIDERAÇÕES FINAIS

Como esperado, o processo de detecção classificou 10 elementos como outliers. Este número desejado foi determinado na configuração do módulo de detecção de outliers no processo construído no Rapidminer com o objetivo de destacar os outliers de maior distância em uma quantidade que permitisse verificação manual dos resultados. Estes elementos estão listados na Tabela 1. Observe que entre os dez elementos selecionados destaca-se um documento de empenho contendo uma linha com valor unitário e valor total de R$235.789.672,00. Outros elementos possuem o valor total menor que o valor unitário. O processo de coleta e mineração de dados, utilizando o algoritmo escolhido, foi capaz de destacar alguns lançamentos que atendem as especificações de um outlier como descrito anteriormente. A avaliação do motivo desta classificação, do ponto de vista operacional, fiscal e legal dependerá dos órgãos e pessoas responsáveis e competentes para esta avaliação.

**Tabela 1 - Outliers Detectados**

| Código do Documento | Data | Sub Item | Valor Unitário | Valor Total |
|---|---|---|---|---|
| 170381000012011NE000002 | 19/5/2011 | INDENIZACOES | 235789672.00 | 235789672.00 |
| 393003392522011NE000252 | 11/4/2011 | ACRE | 23050141.67 | 23050141.67 |
| 275068272092011NE800001 | 11/4/2011 | INDENIZACOES | 11000000.00 | 11000000.00 |
| 250057000012011NE801456 | 9/6/2011 | APOIO ADM., TEC E OP. | 12149056.52 | 1012380.88 |
| 773200000012011NE440491 | 25/1/2011 | INDEN. AUXILIO-TRANSPORTE | 88935000.00 | 88935000.00 |
| 170600000012011NE000329 | 25/1/2011 | AMORTIZ. DIVIDA REF./INST.FIN. | 30585349.00 | 30585349.00 |
| 170600000012011NE000331 | 25/1/2011 | JUROS DA DIVIDA INSTIT.FIN. | 4659240.00 | 4659240.00 |
| 250057000012011NE800152 | 25/1/2011 | APOIO ADM. TEC. E OP. | 11769753.37 | 252461.21 |
| 240901000012011NE001831 | 19/5/2011 | INST. ASSIST., CULT. OU EDU. | 6500000.00 | 6500000.00 |
| 170381000012011NE000003 | 19/5/2011 | INDENIZACOES | 6000000.00 | 6000000.00 |

O objetivo original foi atingido e, através do processamento de dados reais, disponíveis publicamente na internet, foi possível mostrar a utilidade de técnicas de mineração de dados como ferramenta de apoio à auditoria fiscal e combate à corrupção.

Este artigo não encerra o tema, apenas indica um caminho que pode ser adotado para melhorar o combate à corrupção no Brasil. A análise até aqui sugere alguns procedimentos que podem ser adotados no futuro:

- A melhoria do crawler visando um aumento na velocidade de captação e a criação de rotinas de pré-processamento que separem os documentos com erros evidentes. Diminuindo os recursos e tempo necessários a detecção.
- A comparação com outros algoritmos de detecção baseados em distância, ou não, quanto à precisão, velocidade e recursos computacionais utilizados;
- A criação de um processo para a visualização destes outliers em gráficos;

## 6. OBRAS CITADAS


ALI, A. B. M. S.; XIANG, Y. **Dynamic and Advanced Data Mining fro Processing Techonological Development:** Innovations and Systematic Aproaches. 1ª Edição. ed. Hershey: IGI Global, 2010.

ARNING, A.; AGRAVAL, R.; RAGHAVAN, P. **A Linear Method for Deviation Detection in Large Databases**. The Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: [s.n.]. 1996. p. 6.

BEN-GAL, I. OUTLIER DETECTION. In: O, M.; L., R. **Data Mining and Knowledge Discovery Handbook:** A Complete Guide for Practitioners and Researchers. Tel-Aviv: Kluwer Academic Publishers, 2005. p. 117 -130.

BRASIL, G. F. D. Sobre o Portal. **Portal da Transparência**, 2004. Disponivel em: <http://www.portaltransparencia.gov.br/sobre/>. Acesso em: 19 junho 2011.

BREUNIG, M. M. et al. **LOF:** Identifying Density-Based Local Outliers. Proc. 29th ACM SIDMOD Int. Conf. on Management of Data (SIGMOD 2000). Dallas: [s.n.]. 2000. p. 12.

CALAU, A. A.; FORTIS, M. F. D. A. **Transparência e Controle social na Adiministração pública brasileira:** avaliação das prinicpais inovações introduzidas pela Lei de Responsabilidade Fiscal. XI Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública. Ciudade de Guatemala - Guatemala: [s.n.]. 2006. p. 16.

CIOS, K. J. et al. **Data Mining A Knowledge Discovery Approach**. 1ª Edição. ed. New York, NY - USA: Springer Science+Business Media, LLC, 2007. ISBN ISBN-13: 978-0-387-33333-5.

DECOMTEC - FUNDAÇÃO DAS INDÚSTRIAS DO ESTADO DE SÃO PAULO. **Corrupção: custos econômicos e propostas de combate**. Fiesp - Fundação das Indústrias do Estado de São Paulo. São Paulo, p. 35. 2010.

FILZMOSERA, P.; GARRETTB, R. G.; REIMANN, C. Multivariate outlier detection in exploration geochemistry. **Computers & Geosciences**, Viena, 16 Novembro 2004.

G1. Brasileiro Teme mais a volda inflação que a violência, diz pesquisa. **G1 Economia**, 2011. Disponivel em: <http://g1.globo.com/economia/noticia/2011/06/brasileiro-teme-mais-volta-da-inflacao-do-que-violencia-diz-pesquisa.html>. Acesso em: 20 junho 2011.

GOGOI, P. et al. A Survey of Outlier Detection Methods in Network Anomaly Identification. **The Computer Journal**, Oxford, v. 54, 22 Setembro 2011.

GOVERNO FEDERAL. Detalhamento Diário de Despesas. **Portal da Transparência - Governo Federal do Brasil**, 2004. Disponivel em: <http://www.portaltransparencia.gov.br/despesasdiarias/>. Acesso em: 15 junho 2011.

GOVERNO FEDERAL. Portal da Transparência - Detalhamento Diário de Despesas - Saiba Mais. **Portal da Transparência**, 2010. Disponivel em: <http://www.portaltransparencia.gov.br/despesasdiarias/saiba-mais>. Acesso em: 20 junho 2011.

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. 2ª Edição. ed. San Francisco, CA - USA: Elsevier Inc, 2006.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. 1ª Edição. ed. Boston, MS - USA: The MIT Press, 2001. ISBN ISBN: 026208290x.

HE, Z. et al. FP-Outlier: Frequent Pattern Based Outlier. **Computer Science and Information System**, New York, 2005. 113 - 118.

HODGE, V. J.; AUSTIN, J. **A Survey of Outlier Detection Methodologies**. Rotterdan - Holanda: Kluwer Academic Publishers, 2004.

HUBER, P. J. Robust Estimation of a Location Parameter. **Project Euclid**, 1964. Disponivel em: <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms/1177703732>. Acesso em: 10 junho 2011.

INTRODUCTION to R. **R-Project.org**, 2010. Disponivel em: <http://www.r-project.org/about.html>. Acesso em: 02 junho 2011.

KNORR, E. M.; NG, R. T.; TUCAKOV, V. Distance Based outliers: algorithms and Applications. **The VLDB Journal — The International Journal on Very Large Data Bases**, New York, Fevereiro 2000. 17.

KUMAR, V.; KUMAR, D.; SINGH, R. K. Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented In Medical Databases. **IJCSNS International Journal of 272 Computer Science and Network Security**, Seoul, 8, 30 Agosto 2008. 6.

LEVY, P.; LEMESHOW, S. **Sampling of Populations Methods and Applications**. 3ª Edição. ed. New York - USA: JOHN WILEY & SONS, INC., 1999.

MOHAMED, M. S.; KAVITHA, T. Outlier Detection Using Support Vector Machine in Wireless Sensor Network Real Time Data. **International Journal of Soft Computing and Engineering (IJSCE)**, London, 30 Maio 2011. 5.

OTEY, M. E.; PARTHASARATHY, S.; GHOTING, A. **An Empirical Comparison of Outlier Detection Algorithms**. KDD-2005 Workshop - Data Mining Methods for Anomaly Detection. Chicago: [s.n.]. 2005. p. 45-51.

PEREIRA, J. M. Reforma do Estado e controle da corrupção no Brasil. **International Budget Partnership**, São Paulo, abril 2005. 17.

PETROVSKIY, M. I. Outlier Detection Algorithms in Data Mining Systems. **Programmirovanie**, Moscow - Russia, 29, 19 Fevereiro 2003. 10.

PRESIDÊNCIA DA REPÚBLICA. LEI No 4.320, DE 17 DE MARÇO DE 1964. **Presidência da República do Brasil - Casa Civíl**, 1964. Disponivel em: <http://www.planalto.gov.br/ccivil_03/Leis/L4320.htm>. Acesso em: 10 junho 2011.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. **Efficient Algorithms for Mining Outliers from Large Data Sets**. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Texas: ACM. 2000. p. 427-438.

RANDOM.ORG. Introduction to Randomness and Random Numbers. **Random.org**, 2010. Disponivel em: <http://www.random.org/randomness/>. Acesso em: 10 junho 2011.

RAPID-I. Rapidminer. **rapid-i.com**, 2010. Disponivel em: <http://rapid-i.com/content/view/181/196/>. Acesso em: 10 junho 2011.

TAYLOR & FRANCIS GROUP. **Next Generation of Data Mining**. 1ª Edição. ed. Boca Raton: CRC Press, 2009. ISBN ISBN: 13: 978-1-4200-8586-0.

TRANSPARENCY INTERNATIONAL. **Corruption Perceptions Index 2010**. Transparency International. Berlim - Alemanha, p. 12. 2010. (ISBN: 978-3-935711-60-9).

TRIBUNAL DE CONTAS DA UNIÃO. Tibunal de Contas da União - Funcionamento. **Tribunal de Contas da União**, 2010. Disponivel em: <http://portal2.tcu.gov.br/portal/page/portal/TCU/institucional/conheca_tcu/institucional_funcionamento>. Acesso em: 10 junho 2011.

WEINSTEIN, M. Strange Bedfellows: Quantum Mechanics and Data Mining. **Nuclear Physics B-proceedings Supplements**, Stanford, v. 199, p. 74-84, 3 Novembro 2009. ISSN ISSN: 0920-5632.

WESTPHAL, C. **DATA MINING FOR INTELLIGENCE, FRAUD, & CRIMINAL DETECTION**. 1ª Edição. ed. Boca Raton: CRC Press, 2009. ISBN ISBN:13: 978-1-4200-6723-1.

ZHANG, Y.; LUO, A.; ZHAO, Y. **Outlier detection in astronomical data**. Storage and Retrieval for Image and Video Databases. San Jose: [s.n.]. 2005. p. 9.