

Identificação de fraudes no sistema de abastecimento de água por meio da utilização da técnica de detecção de *outliers*

Kelly Cerqueira Guerra (UESC) kelly.guerra@hotmail.com
Diego de Castro Fettermann (UFSC) d.fettermann@ufsc.br
Glaucio Garcia Martins Pereira da Silva (UFSC) glaucio.silva@ufsc.br
Carlos Ernani Fries (UFSC) carlos.fries@ufsc.br

Resumo:

A presença de fraudes é uma fonte de prejuízos em diversos setores, como no sistema bancário, distribuição de energia e abastecimento de água. Este artigo tem como objetivo propor uma alternativa para identificação de fraudes em sistemas de abastecimento de água baseado na detecção de *outliers*. A proposta foi aplicada em uma amostra de 55 economias de uma empresa de abastecimento de água. Os resultados encontrados permitiram reduzir a aleatoriedade da busca por fraudes no sistema. Foi possível detectar todos os casos de fraudes no consumo de água apresentados na amostra analisada, se mostrando uma alternativa capaz de direcionar os locais a serem inspecionados quanto à existência de irregularidades.

Palavras chave: Fraude, Abastecimento de água, *outliers*.

Fraud identification in the water supply system through the use of outliers's detection

Abstract

The fraud is a source of waste in several sectors such as the banking system, power distribution and water supply. This article aims to identify fraud in water supply systems based on outliers detection. The proposal was applied to a sample of 55 households from a water supply company. The findings allowed to reduce the randomness of the search for fraud in the supply system. It was possible to detect all cases of fraud in the sample, showing an alternative for directing the households to be inspected for irregularities.

Key-words: Fraud, Water supply, outliers

1. Introdução

As perdas durante o abastecimento de água acontecem devido à ocorrência de vazamentos, erros de medição, ligações clandestinas, fraudes e falhas no cadastro comercial (TSUTIYA, 2006). A detecção de vazamentos e fraudes constitui-se em um problema para as empresas de abastecimento de água, em razão da dificuldade de detecção e da falta de técnicas eficientes, resultando em prejuízos financeiros para a concessionária em razão do volume de água não faturado. As empresas de abastecimento de água enfrentam dificuldade para localizar os pontos onde ocorrem fraudes e vazamentos. O principal método de detecção utilizado é a inspeção com a utilização de equipamentos acústicos, um método vagaroso e dispendioso, que exige percorrer toda extensão das tubulações em busca do ponto de vazamento (GUMIER; LUVIZOTTO JUNIOR, 2007). Na maioria dos casos, a seleção dos locais para inspeção é

realizada de forma aleatória. Mesmo quando existe a tentativa de direcionar a inspeção, a seleção das áreas é realizada empiricamente, dificultando a localização das fraudes (GUMIER; LUVIZOTTO JUNIOR, 2007).

A literatura apresenta alguns métodos que poderiam ser utilizados para a detecção de fraudes no sistema de abastecimento de água. Esses métodos pretendem aumentar o foco de investigação, direcionando a aplicação dos instrumentos de inspeção em locais com maior probabilidade de existência de fraude. Um dos métodos mencionados na literatura para a detecção de fraudes consiste na utilização de metodologias para a identificação de dados atípicos (*outliers*) (KNORR et al., 1999; BEN-GAL, 2005; ARAÚJO, 2010; PAIS, 2011; SILVA, 2011). Um *outliers* consiste em uma observação que se desvia muito das demais observações, a ponto de suspeitar-se que tenha sido gerada por um mecanismo diferenciado (Hawking, 1980), neste caso a fraude. O objetivo principal desse trabalho é propor uma alternativa para detecção dos locais das fraudes no sistema de abastecimento de água. A partir da aplicação dessa proposta pretende-se identificar com maior precisão casos de fraudes (violações nos hidrômetros). Esta sistemática será aplicada a uma amostra de dados de consumo mensal de água de economias de uma empresa de abastecimento de água. A aplicabilidade dessa alternativa se restringe a economias residenciais. Os valores mensais de consumo dessas economias podem sofrer grande variação devido a outras causas além de fraudes e vazamentos, fator este que dificulta a detecção do problema. Outro fator limitante é o uso de dados advindos de hidrômetros antigos, os quais podem conferir problemas nas medidas de consumo.

2. Revisão de Literatura

Nesta seção é apresentado o referencial teórico para a condução do trabalho, o qual contempla o sistema de abastecimento de água e as técnicas de detecção de *outliers* recomendadas pela literatura.

2.1 Sistema de abastecimento de água

O sistema de abastecimento de água pode ser caracterizado como o conjunto de obras, equipamentos e serviços destinados ao fornecimento de água potável e em quantidade suficiente para uma comunidade com a finalidade de consumo doméstico, industrial, em serviços públicos entre outros usos (AZEVEDO NETTO et al., 1998). A Figura 1 ilustra os elementos constituintes de um sistema de abastecimento de água.

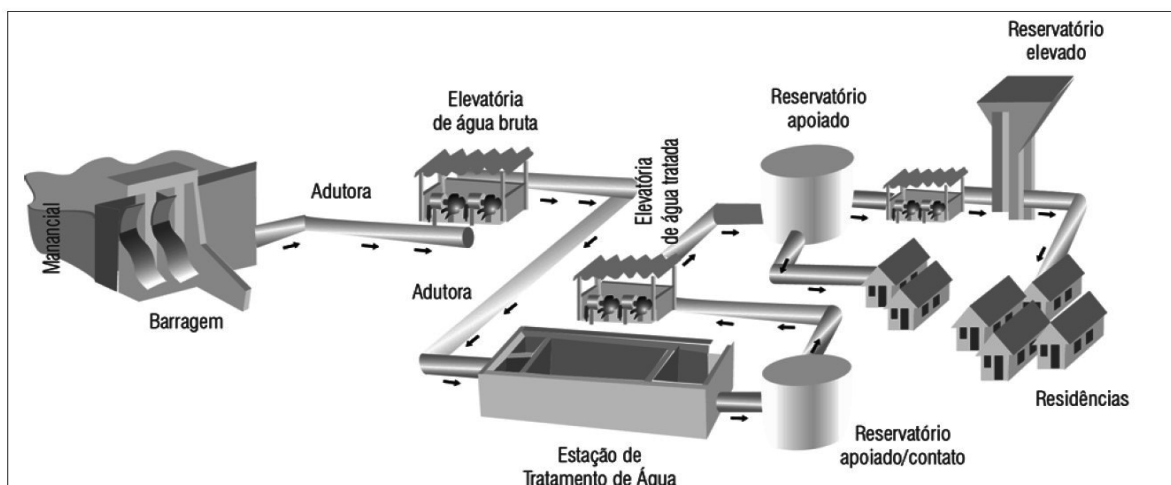


Figura 1 - Sistema de Abastecimento de Água. Fonte: EMBASA - Relatório Anual para Informação ao Consumidor (2011)

As perdas podem ocorrer em todos os pontos de um sistema de abastecimento, desde a captação até as ligações prediais (GOMES, 2009). No Brasil, o índice de perdas no sistema de abastecimento de água é de cerca de 40% (SNIS, 2010). Esta perda de água não se resume apenas ao volume de água produzida que se perde no trajeto até o uso final pelos clientes, mas também a água entregue e por alguma ineficiência, não for faturada. (TSUTIYA, 2006).

Estas perdas ainda podem ser caracterizadas como perdas aparentes e perdas reais (GUMIER; LUVIZOTTO JUNIOR, 2007). As perdas aparentes (ou não físicas) estão relacionadas ao volume de água produzido e não faturado pela empresa devido a fraudes, erros de medição, ligações clandestinas, entre outros. As perdas reais (ou físicas) são ocasionadas por vazamentos, que ocorrem em todo o sistema de abastecimento (VILANOVA et al., 2010).

As fraudes são caracterizadas por intervenções improvisadas no hidrômetro (aparelho que mede a quantidade de água consumida) afim de que seja medida apenas uma parcela da água consumida pelo imóvel. Outra situação bastante comum são as ligações inativas, que não fazem parte do cadastro comercial, mas por ação do cliente se tornaram ativas novamente sem o conhecimento da empresa (TSUTIYA, 2006). As ligações clandestinas podem ser feitas de tal forma que fica difícil sua detecção pela companhia de saneamento, exigindo que sejam realizadas análises de consumo ou pesquisas de campo para efetivamente identificar a ligação (TSUTIYA, 2006).

2.2 Detecção de outliers

Uma das principais etapas da análise estatística de dados é o estudo da qualidade dos dados observados. Em alguns conjuntos de dados é possível perceber que existem elementos que se distanciam significativamente dos demais, sendo comumente designados de *outliers* (MUÑOZ-GARCIA et al., 1990). Não existe uma definição única para *outliers*, podendo variar desde uma observação que se desvia muito das demais levantando suspeitas de seja gerada por um mecanismo diferenciado (HAWKINS, 1980), ou mesmo uma observação desviada significativamente das demais (LIU, 2004). Verifica-se de comum entre os autores que os *outliers* são dados diferenciados do conjunto geral de dados, podendo inclusive ser gerados por mecanismos diferentes (HAWKINS, 1980), no caso desse estudo este mecanismo seria a fraude.

Os *outliers* se relacionam com os outros elementos da amostra e aparecem em uma quantidade pequena, pois se existissem muitos *outliers* seria impossível estabelecer um padrão (SILVA, 2004). Em algumas aplicações, os eventos incomuns são mais interessantes do que os normais, pois podem conduzir a novas descobertas (KNORR et al., 1999). As causas do surgimento dos *outliers* podem ajudar a definir a forma como devem ser tratados. Independente de suas causas, a fase inicial do estudo de *outliers* é a detecção das observações discrepantes. Em alguns casos é possível identificá-las por meio da análise dos valores amostrais ou por meio da interpretação gráfica, em outros casos é necessário utilizar métodos mais sofisticados (FIGUEIRA, 1998). A escolha do método utilizado para a detecção de *outliers* depende de diversos fatores como a distribuição dos dados, se os parâmetros de distribuição (por exemplo, média e variância) são conhecidos, o número esperado de dados discrepantes e o tipo de *outliers* esperados (por exemplo, superior ou inferior) (KNORR et al., 1999).

Os métodos de detecção de *outliers* podem ser divididos em métodos univariados e métodos multivariados. Entre os métodos univariados mais utilizados está o Z-Score com ou sem ajuste de Bonferroni (COUSINEAU; CHARTIER, 2010). Além desse, também são recomendados o teste *box plot*, teste do z score modificado, teste de Grubbs, teste de Dixon, teste de Cochran. Os métodos de detecção multivariados são divididos em métodos estatísticos, que são baseados em parâmetros da distribuição estimados, e métodos de

mineração de dados (*data mining*) (BEN-GAL, 2005).

Durante muito tempo, os *outliers* foram considerados como erros, que serviam apenas para atrapalhar e confundir as investigações. Porém, atualmente a detecção de *outliers* é amplamente utilizada e pesquisada por diversas disciplinas (ARAÚJO, 2010). Dentre as diversas aplicações da detecção de *outliers* pode-se citar: detecção de fraude, processamento de pedido de empréstimo, detecção de invasão de rede, diagnósticos de falhas, monitoramento de condições médicas, análise de irregularidade em votações, ensaios clínicos, previsão de mau tempo, análise de desempenho de atletas (ARAÚJO, 2010; BEN-GAL, 2005).

3. Metodologia

Os procedimentos do método foram desenvolvidos em seis etapas. Primeiramente é apresentado o caso da empresa de abastecimento de água em estudo. Posteriormente é realizada uma análise descritiva dos dados de consumo das economias analisadas. Na etapa seguinte é realizado um tratamento dos dados de consumo. A etapa quatro apresenta a aplicação das técnicas de detecção de *outliers* para identificação de fraudes no sistema. A próxima etapa analisa os resultados obtidos e sua capacidade em identificar casos de fraudes no sistema e por fim a etapa seis apresenta as recomendações para a ampliação da aplicação das técnicas propostas em um conjunto maior de dados. O detalhamento de cada uma das etapas da proposta é apresentado juntamente aos resultados.

4. Resultados

4.1 Descrição do caso

A aplicação da proposta para detecção de fraudes na distribuição de água foi realizada em uma empresa localizada na região Nordeste do. Essa unidade atende um volume de 53 mil economias com um volume de água tratada e distribuída de 22.100 m³ ao dia. A detecção de fraudes no sistema de abastecimento é um problema na empresa, sendo que o procedimento de fiscalização realiza a seleção de economias de forma aleatória, tornando a identificação de fraudes uma tarefa pouco eficiente e onerosa à empresa.

4.2 Análise descritiva dos dados

A empresa disponibilizou uma série histórica de consumo de água de 13 períodos (meses), compreendidos entre julho de 2012 e julho de 2013, de 55 economias residenciais. Nesta amostra, foram disponibilizados 50 casos aleatórios de economias residenciais e cinco casos em que foi detectada fraude no sistema de abastecimento de água, totalizando o conjunto de dados de consumo de 55 economias. A Tabela 1 apresenta a análise descritiva do consumo de água mensal da amostra das economias residenciais e das economias em que foi identificada fraude no sistema.

	Economias sem fraude (n=50)			Economias com fraude (n=5)		
	Consumo Médio (m³)	Desvio Padrão	Coefficiente de Variação	Consumo Médio (m³)	Desvio Padrão	Coefficiente de Variação
Jul/2012	28,68	19,18	0,67	9,0	8,22	0,91
Ago/2012	28,44	17,75	0,62	7,4	7,40	1,00
Set/2012	28,60	15,43	0,54	2,2	2,17	0,99
Out/2012	24,50	14,38	0,59	1,8	1,64	0,91
Nov/2012	27,16	13,03	0,48	2,0	1,22	0,61
Dez/2012	27,60	14,11	0,51	3,4	3,21	0,94
Jan/2013	29,68	13,73	0,46	0,6	1,34	2,23
Fev/2013	34,12	19,33	0,57	1,4	1,34	0,96
Mar/2013	26,28	18,03	0,69	5,4	7,40	1,37
Abr/2013	25,92	14,93	0,58	6,2	5,72	0,92
Mai/2013	27,02	12,63	0,47	8,2	6,72	0,82
Jun/2013	25,90	11,56	0,45	11,6	5,68	0,49
Jul/2013	31,74	13,58	0,43	14,4	7,80	0,54
Média	28,13	15,21	0,54	5,66	4,60	0,98

Tabela 1 - Análise descritiva do consumo de água mensal

Nessa análise se verifica uma grande variabilidade nos dados, tanto entre as economias com e sem a identificação de fraude, muito em razão da heterogeneidade das economias que compõem o sistema. O coeficiente de variação nos dados é sempre alto, nunca inferior a 0,43 (JULHO, 2013). A Figura 2 apresenta a série histórica de consumo de água das economias analisadas.

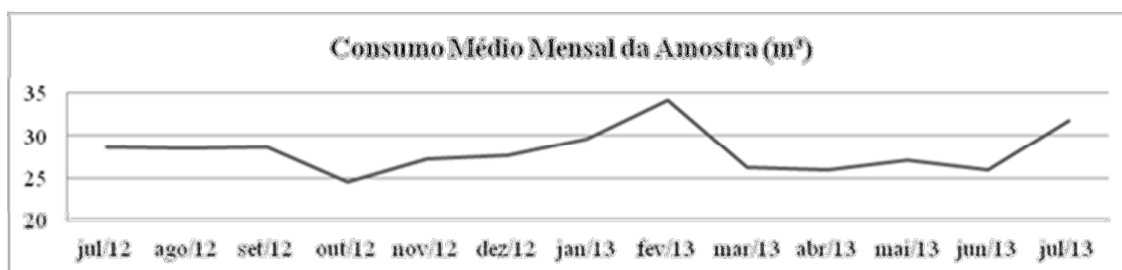


Figura 2II - Consumo médio mensal da amostra

Existem diversos métodos para a detecção de *outliers* univariados, como no caso deste estudo, tais como *box plot*, teste do z-score, teste do z score modificado, teste de Grubbs, teste de Dixon, teste de Cochran, correlação linear e regressão linear (ANDREOTTI, 2005). Entre diversas dessas opções, a utilização da técnica de z-score modificado obteve melhor desempenho em um caso para detecção de fraudes em energia elétrica (NASCIMENTO et al., 2012). Cousineau e Chartier (2010) também mencionam a utilização do z-score como um dos mais utilizados. Os métodos escolhidos foram o Z-Score e o Z-Score modificado, em razão de sua eficácia e ampla aceitação (SHIFFLER, 1988; CROSBY, 1994; SEO, 2006), sendo que seus resultados para a detecção de fraudes serão comparados.

4.3 Tratamento dos dados

A grande variabilidade dos dados de consumo de água das economias da amostra é explicada, em parte, pela variação entre essas economias. Dentro da amostra aleatória de 50 economias residenciais, existem casos de residências grandes, com diversos pontos de água e residências pequenas com menor quantidade de pontos e consumo. Uma alternativa para reduzir a variabilidade de dados de consumo consiste na identificação de padrões de consumo de água

entre as economias, em que cada agrupamento de economias possua maior semelhança no consumo. Com esse objetivo propõe-se utilizar a técnica multivariada análise de cluster (TINSLEY; BROWN, 2000; HAIR et al., 2005; NORUSIS, 2011). Foi analisada a série de 13 meses de consumo de água do conjunto de 50 economias da amostra por meio da análise de cluster hierárquico, seguindo o método de Ward e o algoritmo do quadrado da distância euclidiana (FIGURA 3).

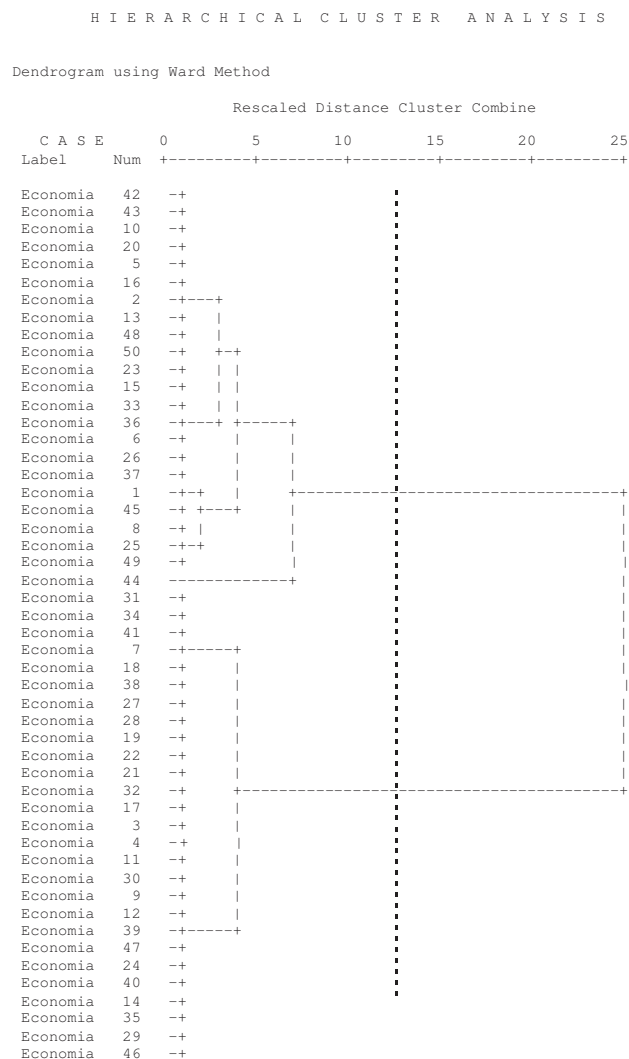


Figura 3 -III Dendrograma da análise de cluster hierárquico. Fonte: SPSS® v 17

O resultado obtido indica a presença de dois grandes agrupamentos, o primeiro agrupamento inclui economias com padrão maior de consumo de água na amostra e o segundo com economias de menor consumo. Os resultados indicam uma redução da variabilidade do consumo de água entre os grupos identificados. Mesmo assim, o coeficiente de variação apesar de reduzido ainda atinge valores altos, como de 0,53 (08/2013) entre as economias de menor consumo, no grupo 2 (TABELA 2).

GRUPO 1 (n=23)

GRUPO 2 (n=27)

	Consumo Médio (m³)	Desvio Padrão	Coefficiente de Variação	Consumo Médio (m³)	Desvio Padrão	Coefficiente de Variação
Jul/2012	43,52	18,27	0,42	16,03	7,07	0,44
Ago/2012	41,48	16,48	0,40	17,33	9,24	0,53
Set/2012	39,65	13,44	0,34	19,18	9,78	0,51
Out/2012	35,48	11,11	0,31	15,15	9,39	0,62
Nov/2012	37,56	9,09	0,24	18,30	8,51	0,47
Dez/2012	39,04	10,31	0,26	17,85	8,40	0,47
Jan/2013	40,96	8,78	0,21	20,07	9,05	0,45
Fev/2013	49,43	15,56	0,31	21,07	10,74	0,51
Mar/2013	37,78	19,96	0,53	16,48	7,68	0,47
Abr/2013	37,09	14,60	0,39	16,41	5,94	0,36
Mai/2013	36,35	11,83	0,33	19,07	6,32	0,33
Jun/2013	33,78	11,21	0,33	19,18	6,63	0,35
Jul/2013	36,43	11,64	0,32	27,74	14,03	0,51
Média	39,12	13,25	0,34	18,76	8,68	0,46

Tabela 2 - Análise descritiva do consumo de água mensal dos grupos 1 e 2

A partir disso, foram realizados testes para verificar a aderência dos dados à distribuição normal. Com este objetivo foi utilizado o teste Shapiro-Wilk. Segundo a estatística do teste, não foi possível rejeitar a hipótese de não aderência à distribuição normal em todos os períodos analisados (p valor < 0,05). Em razão disso, se aceita a hipótese de aderência dos dados à distribuição normal (TABELA 3).

Shapiro-Wilk			
	Estatística	df	Valor p
Julho/2012	0,940	27	0,123
Agosto/2012	0,962	27	0,418
Setembro/2012	0,946	27	0,175
Outubro/2012	0,954	27	0,274
Novembro/2012	0,988	27	0,986
Dezembro/2012	0,982	27	0,896
Janeiro/2013	0,945	27	0,165
Fevereiro/2013	0,974	27	0,698
Março/2013	0,933	27	0,083
Abril/2013	0,977	27	0,794
Maio/2013	0,951	27	0,221
Junho/2013	0,962	27	0,407
Julho/2013	0,949	27	0,204

Tabela I3 - Testes de aderência dos dados à distribuição normal. Fonte: SPSS® v 17

4.4 Aplicação do método de detecção de outliers

O Z-Score é um método simples utilizado em dados com distribuição normal que identifica outliers utilizando a média e o desvio padrão da amostra (SHIFFLER, 1988; SEO, 2006). O Z-Score é calculado através da Equação 1 (SHIFFLER, 1988):

$$Z_{score} = \frac{x_i - \bar{x}}{s} \quad (\text{Equação 1})$$

Onde

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{Equação 2})$$

É sugerido que um dado seja considerado *outliers* quando o valor absoluto do Z-Score for maior que 3, no entanto, a depender das características da amostra ou dos critérios de decisão especificados, este valor pode ser alterado (NASCIMENTO et al., 2012). Uma limitação para o uso do Z-Score é o desvio padrão elevado em razão dos dados com valores extremos. Dessa forma, os *Outliers* menos extremos passam despercebidos por causa dos *outliers* mais extremos (SEO, 2006). Para evitar esse problema, foi desenvolvido o método do Z-Score modificado (SEO, 2006; NASCIMENTO et al., 2012). A média e o desvio padrão são substituídos por indicadores mais robustos: a mediana e a média do desvio absoluto da mediana (MAD) (EQUAÇÃO 3).

$$MAD = \text{median}\{|x_i - \bar{x}_1|\} \quad (\text{Equação 3})$$

A utilização desses indicadores resulta no método Z-Score modificado (SEO, 2006; NASCIMENTO et al., 2012). Sendo assim, o Z-Score modificado (M_i) é calculado a partir da mediana e do desvio absoluto da mediana (MAD), além de uma constante (EQUAÇÃO 4) (SHIFFLER, 1988). São considerados *outliers* as observações com $M_i < K$, sendo K um valor selecionado pelo pesquisador.

$$M_i = \frac{0,6745}{MAD} (x_i - \bar{x}_1) \quad (\text{Equação 4})$$

Utilizando o software Excel® foi calculado o Z-Score e o Z-Score modificado para cada residência do Grupo 2 e para as residências com fraude já confirmada durante os treze períodos. As observações pertencentes ao Grupo 1 não foram envolvidas, pois se tratam de valores de consumo superiores e dificilmente possuam fraude. Foram considerados como fraude os casos com identificação de *outliers* por, pelo menos, três períodos (meses) consecutivos.

No cálculo do Z-Score foi considerado $K=-1,5$ e pode-se perceber que, dentre as cinco economias com fraude confirmada o método foi capaz de identificar quatro delas e, no grupo das demais economias, duas foram identificadas como possíveis fraudes. Para o Z-Score modificado foi utilizado $K=-1,3$, sendo possível observar que todas as economias com fraude confirmada foram identificadas como tal e, nas demais economias duas foram apontadas como possível fraude.

4.5 Análise dos resultados

Ao testar qualquer hipótese existem dois tipos de erros que podem surgir. O “Erro tipo I” é a rejeição de uma hipótese nula (H_0) quando ela é, de fato, verdadeira. O “Erro tipo II” consiste em não rejeitar a hipótese nula quando, na verdade, ela é falsa (FONSECA; MARTINS, 2006). Sendo assim, H_0 consiste em determinada economia não possuir fraude. O método testado por essa sistemática pode dar origem aos seguintes erros:

- “**Erro tipo I**”: Rejeita H_0 quando H_0 é verdadeira. No caso, a sistemática acusa a presença de fraude em uma economia em que não existe fraude;
- “**Erro tipo II**”: Não rejeita H_0 quando H_0 é falsa. A sistemática não acusa a presença em uma economia em que existe fraude.

Dentre as 50 economias da amostra que não havia fraude, os métodos utilizados apontaram caso de fraude em duas economias, portanto, a porcentagem de “Erro tipo I” é de 4% para os dois métodos. O método do Z-Score deixou de detectar a presença de fraude em uma economia dentre as cinco em que existia fraude, obtendo 20% de “Erro tipo II”. O método do Z-Score modificado foi capaz de detectar com sucesso todos os casos de fraude confirmada. Dessa forma, a sua porcentagem de “Erro tipo II” é 0%. A Tabela 4 resume as taxas de “Erro tipo I” e “Erro tipo II” e as taxas de erros/acertos para dos dois métodos testados.

Método	Taxa de "Erro tipo I"	Taxa de "Erro tipo II"	Taxa de Erro/Acerto
Z-Score	2/50	0/5	2/53
Z-Score modificado	2/50	1/5	3/52

Tabela 4- Taxas do "Erro tipo I" e do "Erro tipo II" e Taxas de Erro/Acerto

4.6 Análise e recomendações para aplicação da sistemática de detecção de fraudes em grande escala

Um dos fatores que dificultou o desenvolvimento da sistemática foi a quantidade restrita de dados, principalmente em relação ao número limitado de períodos, que impediu identificar com precisão o comportamento sazonal da demanda. Outro fator que restringiu a pesquisa foi a grande variabilidade dos dados de consumo, restringindo a aplicação de outros métodos recomendados para a detecção de fraude. Para aplicação da proposta em grande escala seria recomendado realizar mais testes com os métodos propostos em uma série histórica de dados maior do que a disponibilizada, além de testar em outros casos de fraude confirmada.

Ainda não se pode afirmar que nas economias indicadas com erro do tipo I não existam fraude, visto que nessas economias não foi realizada inspeção. Também não foi possível acessar qualquer tipo de classificação das economias em relação ao número de moradores ou à quantidade de pontos (chuveiros, torneiras e duchas) existentes no imóvel. Algum tipo de classificação pode contribuir para disponibilizar uma menor variabilidade nos dados de consumo contribuindo na aplicação das técnicas para detecção de fraude. Para que sua aplicação se torne parte da rotina de verificação da empresa é imprescindível a operacionalização de um sistema capaz de aplicar as etapas propostas por esta sistemática.

5. Conclusões

Apesar da quantidade limitada de dados, a aplicação da proposta para detecção de fraudes se mostrou eficiente, sendo capaz de proporcionar um direcionamento para a inspeção dos possíveis casos de fraudes no sistema de abastecimento de água a partir da identificação das economias com consumo de água atípico. Dentre os dois métodos testados o Z-Score modificado se mostrou mais eficiente que o Z-Score, pois as taxas de erro foram menores. Vale ressaltar que a etapa de tratamento dos dados se mostrou fundamental para a aplicabilidade da proposta, pois foi capaz de reduzir a variabilidade presente na amostra.

Referências

- ARAÚJO, B.M. *Identificação de outlier sem redes complexas baseada em caminhada aleatória*. Dissertação de Mestrado em Engenharia de Produção, Universidade de São Paulo, São Carlos (USP-EESC), 2010.
- AZEVEDO NETTO, J.M., FERNANDEZ Y FERNANDEZ, M., ARAUJO, R., EIJIITO, A. *Manual de Hidráulica*. (8ª ed.). São Paulo: Edgard Blucher Ltda, 1998.
- BEN-GAL, I. Outliers detection. In: MAIMON, O. & ROCKACH, L. (Eds.). *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers. Norwell, Massachusetts, 2005.
- COUSINEAU, D., CHARTIER, S. Outliers detection and treatment: a review. *International Journal of Psychological Research*, v.3, n.1, p.58-67, 2010.
- FIGUEIRA, M.M.C. *Identificação de Outliers*. *Millenium*, nº12, 1998. Disponível em <<http://www.ipv.pt/millenium/arq12.htm>>.
- FONSECA, J.S., MARTINS, G.A. *Curso de estatística*. (6ª ed.). São Paulo: Atlas, 2006.
- GOMES, H.P. *Eficiência Hidráulica e Energética em Saneamento - Análise Econômica de Projetos*. (2ª ed.). João Pessoa: Editora Universitária da UFPB, 2009.

GUMIER, C.C., LUVIZOTTO JUNIOR, E.J. Aplicação de modelo de simulação-otimização na gestão de perda de água em sistemas de abastecimento. *Revista Engenharia Sanitária e Ambiental*, v.12, n.1, p.32-4, 2007. Doi: 10.1590/S1413-41522007000100005.

HAIR, J.F., ANDERSON, R.E., TATHAM, R.L., BLACK, W.C. *Multivariate Data Analysis*. (5ª ed.). New Jersey: Prentice Hall, 2005.

HAWKINS, D. *Identification of outliers*. London: Chapman & Hall, 1980.

CROSBY, T. How to Detect and Handle Outliers. *Technometrics*, v.3, p.315-316, 1994. Doi: 10.1080/00401706.1994.10485810.

KNORR, E.M., NG, R.T., TUCAKOV, V. Distance-based outliers: algorithms and applications. *The VLDB Journal*, v.8, n.3-4, p.237-253, 1999. Doi: 10.1007/s007780050006.

LIU, H., SHAH, S., JIANG, W. On-line outliers detection and data cleaning. *Computers and Chemical Engineering*, v.28, p.1635-1647, 2004. Doi:10.1016/j.compchemeng.2004.01.009.

MUÑOZ-GARCIA, J., MORENO-REBOLLO, J.L., Pascual-Acosta, A. *Outliers: a formal approach*. *International Statistical Review*, v.58, n.3, p. 215-226, 1990. Disponível em <<http://www.jstor.org/stable/1403805>>.

NASCIMENTO, R.M., OENING, A.P., MARCILIO, D.C., AOKI, A.R., ROCHA J.R., E.P., SCHIOCHET, J.M. Algoritmo de detecção e correção de outliers para previsão de carga. IV Simpósio Brasileiro de Sistemas Elétricos (SBSE) (Anais), Goiânia, 2012.

NORUSIS, M.J. *PASW Statistics 18: Advanced Statistical Procedures Companion*. Chicago, IL: SPSS, 2011.

PAIS, R.M.A. *Detection of Outliers and Outliers Clustering on Large Datasets with Distributed Computing*. Dissertação de Mestrado em Informática, Universidade de Lisboa, Lisboa, 2011.

SEO, S. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. 56 f, 2006. Dissertação de Mestrado em Ciência, Escola de Saúde Pública, Universidade de Pittsburgh, Pittsburgh.

SHIFFLER, R.E. Maximum Z Scores and Outliers. *The American Statistics*, v.42, n.1, p.79-80, 1988. Doi: 10.1080/00031305.1988.10475530.

SILVA, F.R. *Uma abordagem para detecção de outliers em dados categóricos*. 2004. 81 f. Dissertação de Mestrado em Ciência da Computação, Instituto de Computação, UNICAMP, São Paulo.

SILVA, M.D. *Um modelo para análise e tratamento de dados de demanda de energia elétrica*. 2011. 111 f. Monografia de Bacharelado em Ciência da Computação, Universidade Federal de Alfenas, Alfenas.

TINSLEY, H.E.A., BROWN, S.D. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. San Diego: Academic Press, 2000.

TSUTIYA, M.T. *Abastecimento de água*. (3ª ed.). São Paulo: Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da Universidade de São Paulo, 2006.

VILANOVA, M.R.N., VIANA, A.N.C., SALVADOR, E., SANTOS, G.C., PASCOA, C. Redução do consumo de energia elétrica em sistemas de abastecimento através da redução das perdas de água: estudo de caso no município de Lavras - MG. In: GOMES, H.P. *Sistemas de Saneamento – Eficiência Energética*. (1ª Ed.). João Pessoa: Editora Universitária da UFPB, 2010.