Um Método para Detecção e Diagnóstico de Outliers em Dados Urbanos via Análise Multidimensional

Thiago I. A. Souza[1], **Deborah Magalhaes**[1], **Andre L. L. Aquino**[2], **Danielo G. Gomes**[1]

[1]Universidade Federal do Ceará (UFC)
Grupo de Redes de Computadores, Engenharia de Software e Sistemas (GREat)
Av. Mister Hull, s/n – Campus do Pici – Bloco 942-A
60455-760 – Fortaleza – CE – Brasil

[2]Instituto de Computação - Universidade Federal de Alagoas (UFAL)
Caixa Postal 57.072-970 - Maceió - AL - Brasil

[thiagoiachiley, deborah, dgomes]@great.ufc.br, alla@laccan.ufal.br

Abstract. *Since 2007, for the first time in history, more people live in cities than in the countryside and this number only tends to grow. More people in the cities mean more stress on urban infrastructures, greater demand for public services, and an ever-increasing rate of heterogeneous (multidimensional) data generation. Data are essential for the implementation of evidence-based public policies. In this paper, we propose a method for detecting and diagnosing multi-dimensional urban data outliers in 4 sequential stages: (i) modeling the matrix data in a 3D tensor; (ii) Tucker3 decomposition to extract latent factors; (iii) outliers detection statistics, and (iv) diagnostic techniques in the inspection of outliers causes. Using real data from the Smart Citizen platform, our method allows us to identify the environmental variables that most impact the outliers. Moreover, ROC curves indicated an accuracy gain of 20% over the multivariate approach.*

Resumo. *Desde 2007, pela primeira vez na História, mais pessoas vivem nas cidades do que no campo e este número só tende a crescer. Mais pessoas nas cidades significa maior estresse nas infraestruturas urbanas, maior demanda por serviços públicos e também uma taxa de geração de dados heterogêneos cada vez maior. Dados são essenciais para implementação de políticas públicas baseadas em evidências. Neste artigo, propomos um método para detecção e diagnóstico de outliers em dados urbanos via análise multidimensional em 4 passos sequenciais: (i) modelagem dos dados matriciais em um tensor 3D; (ii) decomposição Tucker3 para extração dos fatores latentes; (iii) estatísticas de detecção de outliers, e (iv) técnicas diagnósticas na inspeção das causas dos outliers. A partir de dados reais da plataforma Smart Citizen, nosso método permite identificar as variáveis ambientais que mais impactam os outliers. Além disso, as curvas ROC indicaram um ganho de acurácia de 20% com relação à abordagem multivariada.*

1. Introdução

Vivemos em um século urbano. Segundo a ONU (Organização das Nações Unidas), 54% da população mundial vive em zonas urbanas e estima-se que em 2030

este número atinja 66% [Programme 2016, United Nations and Social Affairs 2015]. Aumento na população urbana implica em maior estresse na infraestrutura das cidades e consequente aumento de seus problemas correlatos (e.g. mobilidade urbana, segurança, saúde pública). Por outro lado, graças às possibilidades das Tecnologias de Informação e Comunicação (TIC) e do Big Data notamos, ainda que em pequena escala, iniciativas concretas de cidades inteligentes, inclusive no Brasil[1].

Dados são essenciais para tomadas de decisões baseadas em evidências, boas práticas em investimento e gestão da infraestrutura de uma cidade. Os dados abertos, em particular, estão transformando significativamente o modo com que os governos locais compartilham informações com os cidadãos e entregam serviços[2]. De forma complementar, recentemente a Internet das Coisas (*Internet of Things*, IoT) tem se tornado uma facilitadora-chave para cidades inteligentes, através da qual dispositivos como sensores e atuadores são componentes fundamentais para a detecção e monitoramento de eventos relacionados ao meio ambiente, clima, energia, dentre outros [Zhang et al. 2017].

Na perspectiva das cidades inteligentes, o monitoramento ambiental se destaca, uma vez que a observação do comportamento de variáveis como gases poluentes, temperatura, ruído e luminosidade são informações vitais para a saúde das pessoas [Rathore et al. 2016]. Entretanto, com o aumento do volume de dados, as técnicas tradicionais de processamento e procedimentos analíticos apresentam desempenho muito limitado [Babar and Arif 2017] [Steed et al. 2013]. Esse problema se torna mais crítico à medida que mais dados são coletados e surgem *outliers* (valores discrepantes).

Outliers são observações que parecem ser inconsistentes com o restante do conjunto de dados, sendo importante identificá-los para explorar seus possíveis padrões de anormalidade [Camacho et al. 2016]. Apesar do fato de que os *outliers*, em geral, são causados por erros de medição, eles podem às vezes indicar eventos de interesse (e.g. altos níveis de poluição do ar, ruído ambiental, ilhas de calor [Ibrahim et al. 2016]). Além disso, tais observações podem permanecer invisíveis com a aplicação de métodos bidimensionais tradicionais, como a Análise de Componentes Principais (*Principal Component Analysis*, PCA) [Khatib et al. 2016], que em geral é utilizado nas soluções de detecção de *outliers* em abordagens de natureza multivariada [Guardiola et al. 2014], [Camacho et al. 2016]. Embora PCA seja um método multivariado popular para a detecção de *outliers* em uma variedade de domínios, sua aplicação mapeia a estrutura 3D natural dos dados em uma forma 2D, capturando apenas as variações bidimensionais.

Diante do exposto, neste artigo propomos um método para o monitoramento ambiental urbano que considera a natureza multidimensional dos dados através da aplicação do método multidimensional Tucker3 [Slavakis et al. 2014] cujo objetivo é explorar todas as dimensões das medidas sensoriadas que são úteis à compreensão do monitoramento do ambiente urbano e apontar melhores resultados de detecção e diagnóstico de *outliers*. Para tal, nosso método apresenta as seguintes etapas: (i) modelagem dos dados matriciais em um tensor 3D (subseção 4.1); (ii) decomposição Tucker3 para extração dos fatores latentes (subseção 4.2); (iii) estatísticas de detecção de outliers (subseção 4.3), e (iv) técnicas diagnósticas na inspeção das causas dos *outliers* (subseção 4.4).

[1]http://revistaplanet.com/

[2]http://www.dataforcities.org/

Este artigo é a sequência de [Souza et al. 2017]. Como principais diferenças apontamos: (i) a aplicação de um método de decomposição tensorial para a extração de fatores latentes; (ii) a combinação dos resultados do método multidimensional com estatísticas multivariadas de monitoramento de processos; (iii) uma proposta de aplicação de um método diagnóstico de *outlier*, comparando-o com o aplicado em [Souza et al. 2017]; (iv) e tratamento dos dados ambientais sensoriados nas cidades de Elda e Rois (Espanha), Nuremberg (Alemanha) e Tallinn (Estônia) como um arranjo multidimensional.

2. Trabalhos Relacionados

A detecção de *outliers* é um campo ativamente pesquisado em muitos domínios, tais como, detecção de intrusão e fraude, saúde pública, processamento de imagem, dentre outros [Chandola et al. 2009]. Recentemente, a temática vem sendo abordada no contexto de ambientes urbanos, por exemplo [Souza et al. 2017], em que ferramentas analíticas bidimensionais foram aplicadas na detecção de *outliers* considerando a natureza multivarada dos dados. Entretanto, nesse trabalho a natureza multidimensional dos dados não foi considerada no estudo. Segundo [Osanaiye et al. 2016] a eficiência da detecção de *outliers* depende da natureza dos dados de entrada. Ainda segundo [Osanaiye et al. 2016], a entrada é uma coleção de instâncias de dados sob a forma de padrões, amostras e observações descritas por um conjunto de atributos representados em forma binária, categórica ou numérica. Cada instância pode consistir em atributos únicos (univariados), múltiplos (multivariados) [Chandola et al. 2009] ou múltiplos em distintas dimensões (multidimensionais) [Fanaee-T and Gama 2016].

Com os recentes avanços das tecnologias de sensores, é possível analisar dezenas de variáveis sensoriadas em um ambiente através de diferentes locais e horários. Dada à variação espaço-temporal existente nos dados coletados de ambientes urbanos, diversos trabalhos na literatura surgem considerando a natureza multidmensional desses dados com o objetivo de identificar os locais ou períodos de tempo relacionados às medidas anormais [Dong et al. 2010], [Li et al. 2011], [Singh et al. 2006].

O presente trabalho diferencia-se dos anteriores por combinar técnicas de detecção e diagnóstico de *outliers* com a ferramenta analítica multidimensional Tucker3 para explorar a interação entre as dimensões espaço-temporal com as medidas ambientais heterogêneas sensoriadas.

3. Fundamentos Analíticos

Com a enorme quantidade de dados gerados dos ambientes urbanos nos últimos anos, o processamento e análise desses dados podem ser desempenhados através de ferramentas analíticas, como [Suzhi et al. 2015]: modelagem estocástica, mineração de dados, aprendizado de máquinas e análise de dados em larga escala. Para tanto, é importante conhecer a natureza desse conjunto de dados para escolher quais ferramentas analíticas serão aplicadas. Os dados coletados podem apresentar natureza univariada, multivariada ou multidimensional. Enquanto os dados univariados representam amostras relacionadas ao mesmo fenômeno escalar (por exemplo, monitoramento apenas de temperatura), dados multivariados representam amostras relacionadas à diferentes fenômenos (por exemplo, monitoramento simultâneo de temperatura, pressão, umidade, etc) [Aquino et al. 2014]. Por outro lado, dados multidimensionais não apenas representam amostras relacionadas

à fenômenos heterogêneos, como também situa-os, por exemplo no espaço, ou seja, podem representar mais de duas dimensões usuais. Tais dados podem ser representados por arranjos multidimensionais, que são chamadas de tensores [Kolda and Bader 2009]. Portanto, um escalar, representado por x, é um tensor de ordem zero; um vetor, denotado por $\mathbf{x}$, é um tensor de primeira ordem; uma matriz, denotada por $\mathbf{X}$, é um tensor de segunda ordem; e um arranjo tridimensional, denotado por $\underline{\mathbf{X}}$, também chamado apenas de tensor, é um tensor de terceira ordem [Kolda and Bader 2009], como ilustrado na Figura 1(a). Além disso, a organização de um tensor pode ser alterada transformando-o em matrizes. Este processo é chamado de matriciação. Para um tensor de terceira ordem $\underline{\mathbf{X}}_{[i\times j\times k]}$, as diferentes formas de matriciação são apresentadas na Figura 1(b). É importante destacar que o processo inverso da matriciação é a desmatriciação, em que a matriz é conduzida seguindo o processo contrário da matriciação para formar o tensor.

Assim, em uma abordagem tensorial, quando fixamos um dos três índices (i, j, k) de um tensor de terceira ordem $\underline{\mathbf{X}}_{[i\times j\times k]}$, formamos as chamadas fatias, isto é, seções bidimensionais de um tensor, a saber: Fatia Horizontal $\mathbf{X}_i$, representando a i-ésima fatia horizontal para o tensor matriciado no primeiro modo $\mathbf{X}_1$, gerando uma matriz de dimensão (j, k); Fatia Vertical $\mathbf{X}_j$, representando a j-ésima fatia vertical para o tensor matriciado no segundo modo $\mathbf{X}_2$, gerando uma matriz de dimensão (k, i); e Fatia Frontal $\mathbf{X}_k$, representando a k-ésima fatia frontal para o tensor matriciado no terceiro modo $\mathbf{X}_3$, gerando uma matriz de dimensão (i, j).

3.1. Técnicas de Análise Multivariada

A Análise de Componentes Principais, uma das principais técnicas de análise multivariada, tem como objetivo identificar a relação entre as características extraídas dos dados, a fim de reduzi-las, eliminando sobreposições insignificantes e escolhendo as formas mais relevantes a partir de combinações lineares das variáveis originais. Os componentes principais com maiores autovalores (representando a variância explicada do modelo) contêm as informações mais relevantes da matriz de dados original e com a análise desses componentes, os padrões dominantes podem ser revelados [Camacho et al. 2016]. Esta abordagem multivariada, utilizando a técnica PCA, será comparada com os resultados obtidos a partir da abordagem multidimensional através da técnica de decomposição tensorial Tucker3.

3.2. Técnicas de Análise Multidimensional

Diversos problemas do mundo real apresentam natureza multidimensional, exigindo uma representação tensorial que tornou-se comum ao longo dos últimos anos em uma variedade de domínios de aplicações, incluindo análise de redes sociais [Xu et al. 2015], neurociências [Chen et al. 2015], sensoriamento remoto [Zhang et al. 2011], etc. A análise de dados multidimensional revela-se mais oportuna do que os métodos bidimensionais, uma vez que estende os métodos tradicionais de decomposição de matrizes para capturar estruturas multilineares e correlações subjacentes em conjuntos de dados de ordem superior [Kolda and Bader 2009]. Os métodos de decomposição tensorial visam representar um tensor como uma combinação linear de produtos externos de tensores de baixa ordem [Suzhi et al. 2015]. Tais métodos são extremamente úteis para processamento e análise de grandes quantidades de dados uma vez que facilitam à interpretação inteligível dos dados. Apresentamos a seguir, o método de decomposição tensorial utilizado nesta pesquisa, a saber, Tucker3.

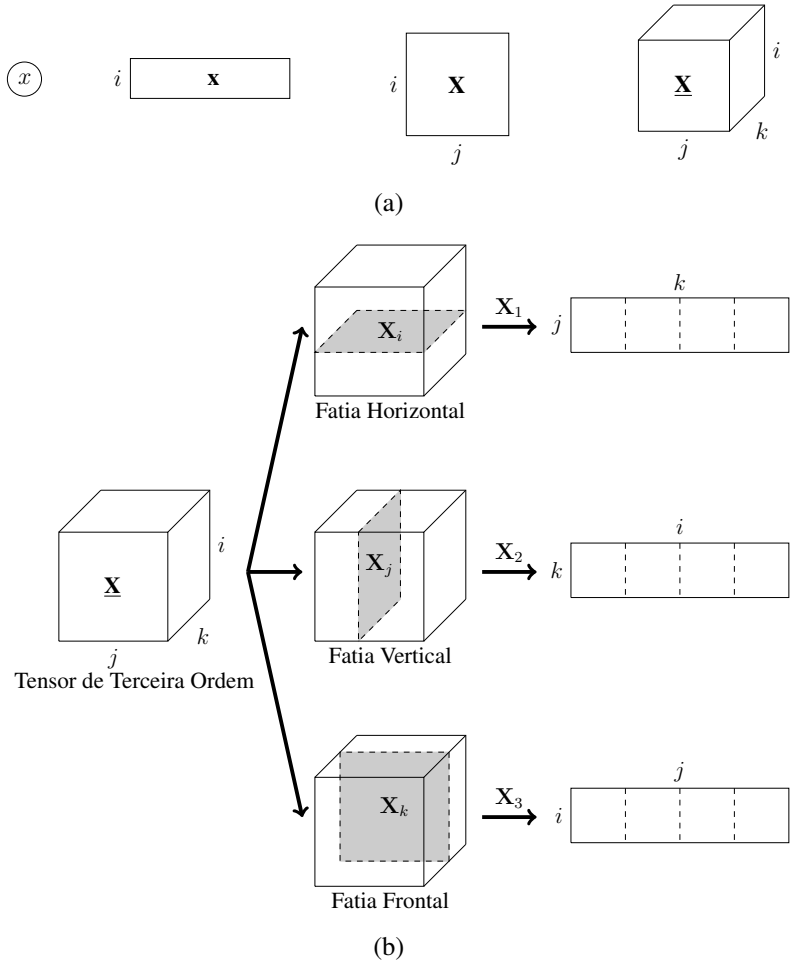


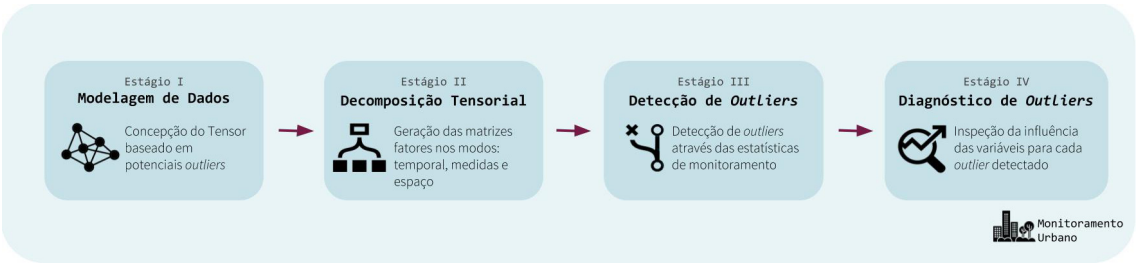
Figura 1. Representação Tensorial: (a) Ilustração da ordem tensorial: x tensor de ordem zero; x tensor de primeira ordem; X tensor de segunda ordem; $\underline{\mathbf{X}}$ tensor de terceira ordem; e (b) Ilustração dos três modos da matriciação de um tensor de terceira ordem.

3.2.1. Decomposição Tensorial baseada em Tucker3

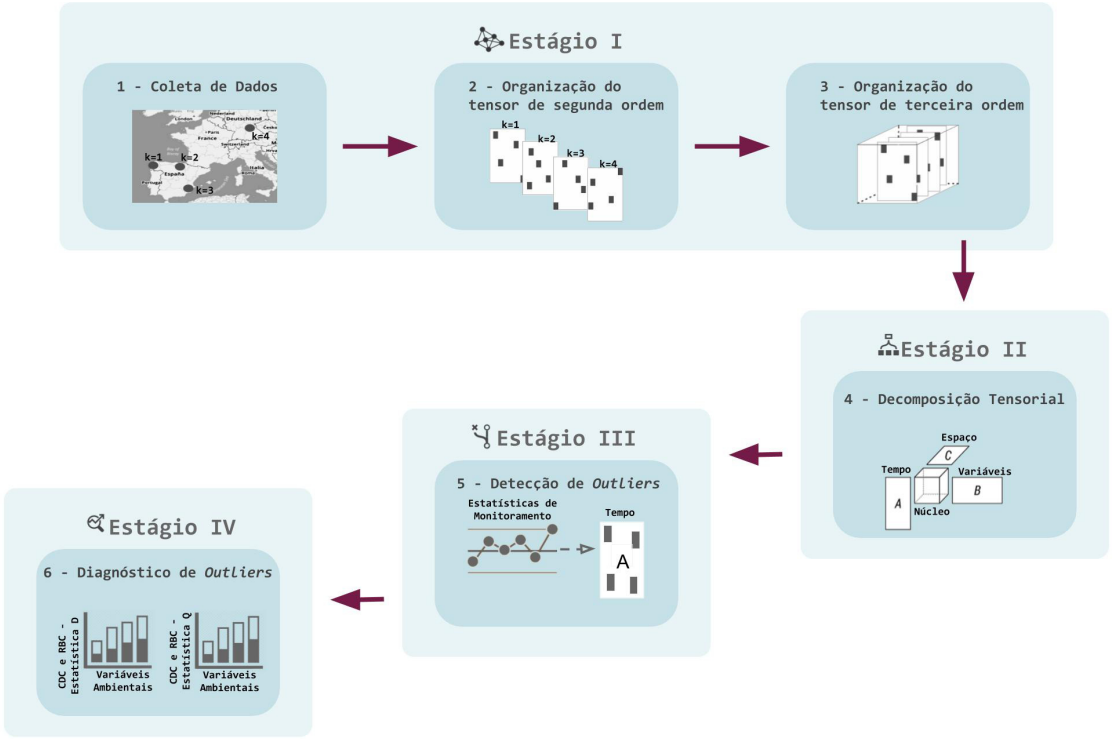
A decomposição tensorial baseada em Tucker3 é uma técnica multidimensional que generaliza os métodos multivariados, Análise de Componentes Principais (PCA) e de Decomposição de Valor Singular (SVD), para arranjos de ordem superior [Kolda and Bader 2009]. Em geral, as aplicações criam um tensor de terceira ordem de dimensões i (modo temporal) $\times$ j (modo medição) $\times$ k (modo espacial). A técnica reduz cada dimensão (ou modo) de um tensor de terceira ordem em um conjunto de componentes/fatores e descreve suas relações. Portanto, o método multidimensional Tucker3 decompõe o tensor $\underline{\mathbf{X}}_{[i\times j\times k]}$ da seguinte maneira

$$\underline{\mathbf{X}} = \underline{\mathbf{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}, \tag{1}$$

em que $\mathbf{A}_{[i\times p]}$, $\mathbf{B}_{[j\times q]}$ e $\mathbf{C}_{[k\times r]}$ são as chamadas matrizes fatores do modelo, multiplicadas nos respectivos modos (indicados no índice de cada matriz fator da Equação 1) pelo tensor núcleo $\underline{\mathbf{G}}_{[p\times q\times r]}$, onde $p < i$, $q < j$ e $r < k$.



(a)



(b)

Figura 2. Monitoramento Urbano - Proposta de Detecção e Diagnóstico de *Outliers*: (a) estágios do método proposto; (b) etapas de cada estágio.

4. Monitoramento Urbano - Proposta de Detecção e Diagnóstico de Outliers

A Figura 2 ilustra a metodologia proposta para o monitoramento urbano na detecção e diagnóstico de outliers. O procedimento consiste em quatro etapas principais: Estágio I - *Modelagem de Dados*, Estágio II - *Decomposição Tensorial*, Estágio III - *Detecção de Outliers* e Estágio IV - *Diagnóstico de Outliers*. Enquanto a Figura 2a apresenta os quatro estágios da nossa metodologia, a Figura 2b detalha as etapas de cada estágio. Por exemplo, o estágio *Modelagem de Dados* (Figura 2a) é representado através das etapas 1-3 apresentadas na Figura 2b, a saber, as etapas: 1 coleta de dados, 2 organização do tensor de segunda ordem e 3 organização do tensor de terceira ordem. Da mesma forma, o estágio *Decomposição Tensorial* é apresentado pela etapa 4, que é a aplicação da decomposição tensorial Tucker3 ao conjunto de dados organizados em um tensor de terceira ordem. O estágio *Detecção de Outliers* é representado pela etapa 5, que é a aplicação das estatísticas D e Q [Camacho et al. 2016] e, final-

mente, a etapa *Diagnóstico de Outliers* é ilustrada pelo passo 6, que é a aplicação dos métodos Contribuição de Decomposição Completa (*Complete Decomposition Contribution*, CDC) e Contribuição Baseada na Reconstrução (*Reconstruction-Based Contribution*, RBC) [Alcala and Qin 2011] no diagnóstico dos *outliers*.

4.1. Modelagem de Dados

Neste estudo, um conjunto de dados reais foi obtido da plataforma Smart Citizen [3], um projeto que conecta as pessoas com seus ambientes através da localização geográfica para coleta e compartilhamento de dados. O projeto Smart Citizen fornece dados ambientais em tempo real, compostos pelas seguintes variáveis ambientais: temperatura, umidade, ruído, monóxido de carbono (CO), dióxido de nitrogênio (NO_2) e luminosidade. Neste trabalho, ao longo de 2.130 horas, as 6 variáveis ambientais foram monitoradas em 4 cidades diferentes (Elda - Espanha, Rois - Espanha, Numberg - Alemanha e Tallinn - Estônia). Esses locais foram selecionados porque oferecem nós sensores online onde as medições podem ser realizadas em tempo real sem dados faltantes. Consequentemente, o conjunto de dados multidimensional é modelado como um tensor de terceira ordem com as seguintes dimensões $\underline{\mathbf{X}}_{[2130x6x4]}$.

O estágio *Modelagem de Dados* consiste em três etapas. Na etapa 1, a Figura 2b mostra os nós sensores selecionados, cada um representando a dimensão espacial (localização) k, onde k = 1, ..., 4, totalizando as quatro cidades selecionadas. O passo 2 mostra que os dados coletados de cada cidade estão organizados em uma matriz composta por potenciais *outliers*, destacados por blocos cinzentos. Finalmente, no passo 3, as quatro matrizes são concatenadas uma atrás da outra, caracterizando um tensor de terceira ordem. Assim, as três dimensões (tempo $\times$ variáveis $\times$ espaço) podem ser exploradas simultaneamente através da decomposição tensorial Tucker3.

4.2. Decomposição Tensorial

No estágio *Decomposição Tensorial*, o método multidimensional Tucker3 gera as matrizes fatores **A**, **B** e **C**, representando as dimensões tempo (modo temporal), as variáveis ambientais (modo medições) e as cidades (modo espacial), respectivamente, como pode ser observado na etapa 4 da Figura 2b. Uma vez que o objetivo é a detecção de *outliers*, nos concentramos apenas na análise do modo temporal (matriz fator **A**). O número de componentes principais de cada matriz fator é escolhido com base na porcentagem acumulada da variância explicada [Kroonenberg 2008]. Portanto, caso a porcentagem cumulativa dos primeiros componentes esteja acima de um limite (por exemplo, 75% [Fanaee-T and Gama 2016]), o número adequado de componentes é selecionado como sendo os componentes que ultrapassam esse limite.

4.3. Detecção de *Outliers*

No estágio *Detecção de Outliers*, as estatísticas D e Q são aplicadas aos componentes das séries temporais multivariadas extraídas da matriz fator **A**. A estatística D, também chamada de estatística T^2 de Hotelling, é uma métrica (distância de Mahalanobis no subespaço de projeção) para o monitoramento de séries temporais multivariadas [Hotelling 1947], como é o caso da matriz **A**. Por outro lado, a estatística Q, também chamada Erro de Predição Quadrada (SPE, do inglês *Squared Prediction Error*), captura os

[3] `https://smartcitizen.me/`

componentes com menor variância explicada e indica o quão bom/deficiente são os dados através da soma das variações inexplicadas de tais componentes [Hotelling 1947]. Ambas as estatísticas são complementares e geram gráficos de controle multivariados para o monitoramento de processos, permitindo a detecção de valores discrepantes conforme empregado em [Camacho et al. 2016]. Destacamos também que, embora apenas o modo temporal tenha sido considerado, essa dimensão é derivada do processo de decomposição tensorial. Portanto, ele traz informações sobre as variáveis ambientais ao longo do tempo, como também informações sobre as cidades. Isso torna este modo muito mais rico quando comparado a um único modo derivado do método multivariado PCA. Além disso, abordar o problema bidimensional reduz sua complexidade. O problema de monitoramento da série temporal é reformulado e um *outlier* é definido como o instante em que a série cai fora do padrão de normalidade presente nos gráficos de controle das estatísticas D e Q. Os potenciais *outliers* detectados pelas estatísticas são destacados na Figura 2b (etapa 5) por blocos cinza. É importante destacar que, tradicionalmente, as estatísticas D e Q são aplicadas nas saídas do PCA. Entretanto, nesta pesquisa, aplicamos essas estatísticas nas saídas de decomposição Tucker3. Portanto, comparamos ambas as abordagens para avaliar seu desempenho e oferecer uma diretriz para pesquisadores que buscam técnicas adequadas de análise de dados no monitoramento ambiental.

4.4. Diagnóstico de *Outliers*

Finalmente, no estágio IV (Figura 2a), a influência das variáveis ambientais em cada *outlier* detectado pelas estatísticas D e Q, é identificada através dos métodos de diagnóstico CDC e RBC [Alcala and Qin 2011]. Nesta perspectiva, o diagnóstico é realizado na etapa 6 (ver Figura 2b) avaliando a contribuição de cada variável para as diferentes estatísticas de monitoramento (estatísticas D e Q). Ou seja, os métodos designam as variáveis relacionadas aos respectivos *outliers* e indica a contribuição de cada variável bem como o seu grau de influência através de gráficos de barras.

5. Resultados

Os resultados revelam o impacto da detecção de *outliers* e da eficácia de seu diagnóstico no monitoramento de ambientes urbanos, através da combinação das estatísticas D e Q com os resultados extraídos do método de decomposição tensorial Tucker3. Todas as simulações realizadas nesta pesquisa foram implementadas no *software* MATLAB, utilizando um processador *Intel Core I5* com velocidade de 2.66GHz e 4GB de memória.

5.1. Detecção de *Outliers*

A Figura 3 mostra os resultados da detecção de *outliers* através das estatísticas D e Q. Essas duas estatísticas foram aplicadas sobre os dois componentes selecionados a partir da matriz fator **A** gerada pelo método tensorial Tucker3. Os componentes da matriz fator **A**, selecionados pelo critério da variância explicada, retornaram 95,8 % de variância explicada pelo modelo, gerando um espaço bidimensional representativo responsável pela variabilidade máxima dos componentes. Para o primeiro componente (CP-1), a estatística D identificou um *outlier* na amostra #1208 (Figura 3a). Além disso, a estatística D detecta no segundo componente (CP-2), um evento anormal na amostra #1523 (Figura 3c). Em relação aos resultados retornados pela estatística Q, a amostra #1523 se destaca no primeiro componente (Figura 3b). No entanto, para o CP-2, a estatística Q não mostrou

nenhuma variação significativa, o que nos permite inferir que não houve comportamento anormal nos dados (Figura 3d). Assim, nota-se que, para ambos os componentes selecionados, as estatísticas D e Q apontam que a observação #1208 e #1523 se desviam das demais, caracterizando-se portanto como *outliers*.

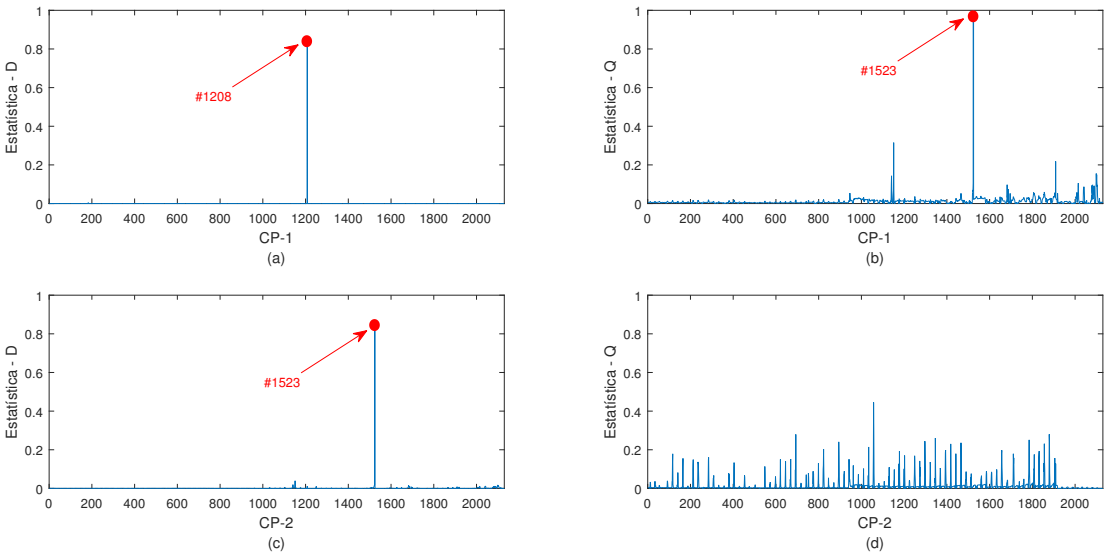


Figura 3. Detecção de *Outliers* - Abordagem Multidimensional: (a) Estatística D (CP-1); (b) Estatística Q (CP-1); (c) Estatística D (CP-2); (d) Estatística Q (CP-2).

.

Tabela 1. Taxa percentual do diagnóstico correto das variáveis para *outlier* #1208.

Método Diagnóstico	Estatística de Detecção	
	D	Q
CDC	82.52%	71.54%
RBC	95.48%	92.40%

Tabela 2. Taxa percentual do diagnóstico correto das variáveis para *outlier* #1523.

Método Diagnóstico	Estatística de Detecção	
	D	Q
CDC	77.80%	61.33%
RBC	85.90%	81.30%

Consultando a plataforma Smart Citizen a partir da qual os dados foram coletados para esta pesquisa, confrontamos os resultados obtidos e observamos ocorrências de eventos anormais nas quatro matrizes correspondentes às cidades selecionadas. O comportamento discrepante foi identificado nas amostras #1208 (correspondente ao instante 22 horas de 01/20/2017, veja a Figura 3a) e #1523 (correspondente ao instante 24 horas do dia 03/02/2017, veja Figuras 3b e 3c). Portanto, as estatísticas D e Q monitoram as tendências e os padrões das séries temporais analisadas. Entre as amostras monitoradas, identificamos as mais discrepantes, permitindo estabelecer inferências sobre os padrões de comportamento do conjunto de dados urbanos.

5.2. Diagnóstico de *Outliers*

Além de detectar eventos anormais, identificamos as variáveis que contribuíram para os desvios detectados. Assim, uma vez que os *outliers* foram detectados, a

contribuição de cada variável é calculada no subespaço dos dois componentes selecionados. Portanto, aplicamos os métodos CDC e RBC sobre cada *outlier* detectado pelas estatísticas D e Q (Figura 4).

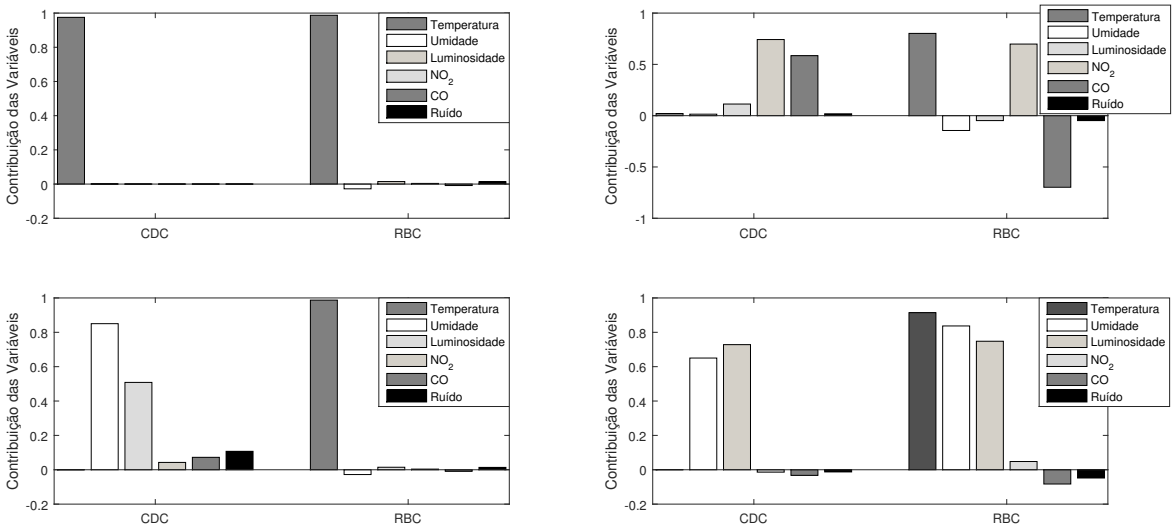


Figura 4. Diagnóstico de *Outliers*: (a) para Estatística D - Amostra #1208; (b) para Estatística Q - Amostra #1208; (c) para Estatística D - Amostra #1523; (d) para Estatística Q - Amostra #1523.

Na Figura 4a, é apresentado o diagnóstico do *outlier* na amostra #1208 para a estatística D, em que tanto o método CDC quanto RBC apontam para variável temperatura como a responsável por causar o desvio no comportamento dos dados no subespaço dos dois componentes principais selecionados. No entanto, quando o diagnóstico é aplicado para a estatística Q (Figura 4b), o método CDC aponta para a influência dos gases poluentes (variáveis NO_2 e CO) na contribuição para o comportamento anormal, e o método RBC aponta não apenas para os gases poluentes como também a variável temperatura. Percebe-se que não houve influência significativa para as demais variáveis. Portanto, as variáveis temperatura, NO_2 e CO são as que têm maior influência no desvio dos dados, gerando um evento anormal na amostra #1208.

Em relação as variáveis que influenciaram a geração do *outlier* #1523 (Figuras 4c e 4d), também no subespaço dos componentes (CP-1 e CP-2), identificamos uma proeminência nas variáveis umidade e luminosidade através do método CDC realizado sobre a estatística D, e o método RBC apresentando a temperatura como a variável de maior intensidade na contribuição do desvio dos dados em relação às demais variáveis (Figura 4c). A Figura 4d complementa a análise, em que novamente as variáveis umidade e luminosidade se destacam através do método CDC, e o método RBC corrobora retornando-as e apontando que a variável temperatura também se destaca sobre as demais.

As Tabela 1 e 2 apresentam as taxas de diagnóstico correto para ambos os métodos CDC e RBC utilizados neste trabalho. A primeira coluna mostra os métodos utilizados para o diagnóstico de *outliers*, a segunda coluna mostra os resultados do diagnóstico quando o *outlier* é detectado pela estatística D e a terceira coluna quando detectado pela

estatística Q. Observa-se que as taxas de diagnóstico são melhores para o método RBC, enquanto que para o método CDC as taxas são ligeiramente menores em termos percentuais. O método RBC se mostrou mais eficiente em relação ao CDC no diagnóstico, revelando a influência de determinadas variáveis cujas contribuições não haviam sido apontadas pelo método CDC. Esta comparação permite-nos inferir que o poder diagnóstico do método RBC (aqui implementado e aplicado), supera do método CDC aplicado em [Souza et al. 2017], revelando-se uma ferramenta útil na discriminação da inflência das variáveis na geração dos *outliers*.

5.3. Comparação - Multivariado × Multidimensional

Comparamos os resultados das abordagens, multivariada e multidimensional, para identificar quais apresentam ganhos em termos de detecção de *outliers* para auxiliar na análise de dados a partir do monitoramento ambiental urbano. Portanto, realizamos uma média das quatro matrizes analisadas neste trabalho, isto é, a média do tensor $\underline{\mathbf{X}}$ ao longo da dimensão espacial k, obtendo uma matriz de médias das variáveis ambientais. Após a obtenção da matriz, aplicamos uma técnica clássica de análise de dados bidimensional, Análise de Componentes Principais e, em seguida, reaplicamos as estatísticas D e Q para explorar a natureza multivariada dos dados e compará-los com os resultados da abordagem multidimensional. A Figura 5 apresenta os resultados das estatísticas D e Q aplicadas nas saídas do método PCA na matriz de médias com dimensões 2130×6.

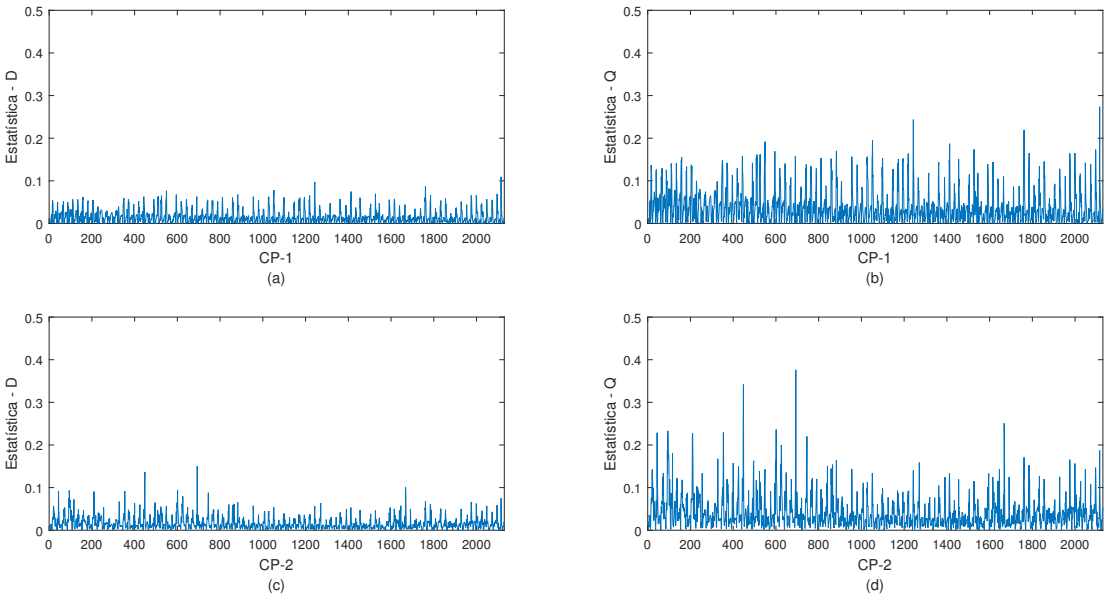


Figura 5. Detecção de *Outliers* - Abordagem Multivariada: (a) Estatística D (CP-1); (b) Estatística Q (CP-1); (c) Estatística D (CP-2); (d) Estatística Q (CP-2).

Para ambos os componentes selecionados (CP-1 e CP-2), que representam 91% da variância explicada pelo método multivariado PCA, nenhum evento anormal significativo foi detectado, destacando um padrão de normalidade para os resultados retornados por ambas as estatísticas de detecção de valores discrepantes. Este resultado nos permite inferir que a nossa abordagem multidimensional foi mais eficiente na detecção de eventos anormais em um grande conjunto de dados, uma vez que a abordagem multivariada

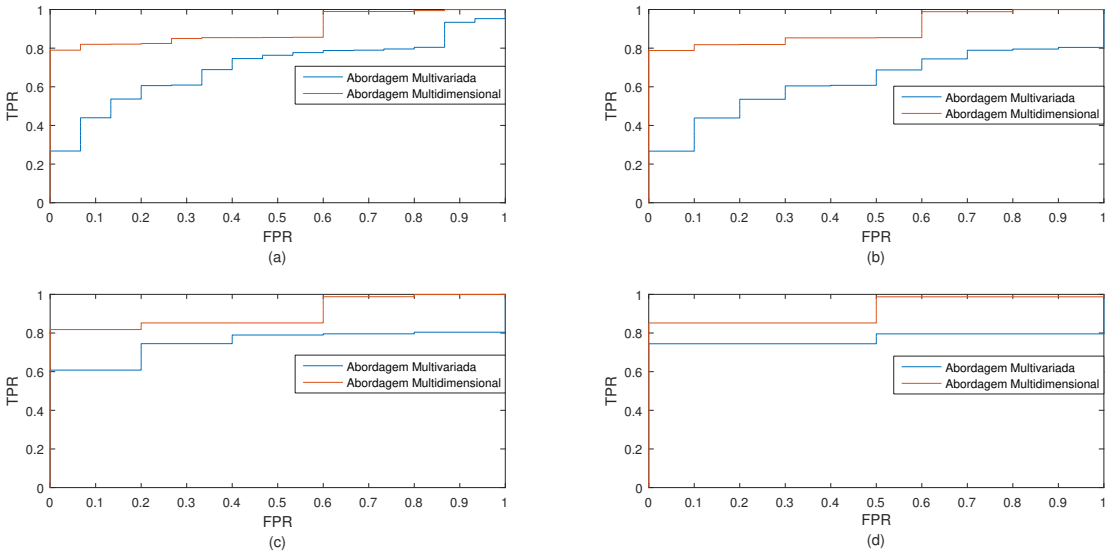


Figura 6. Curvas ROC para comparação entre as abordagens multidimensional e multivariada: (a) Estatística D (CP-1); (b) Estatística Q (CP-1); (c) Estatística D (CP-2); (d) Estatística Q (CP-2).

não identificou padrões relevantes de comportamentos discrepantes no banco de dados analisado.

Avaliamos o desempenho das abordagens, multivariada e multidimensional, em termos das Curvas de Característica Operacional do Receptor (ROC) [Fawcett 2006]. A curva ROC consiste em traçar a taxa positiva verdadeira (*True Positive Rate*, TPR) contra a taxa de falso positivo (*False Positive Rate*, FPR). A ”área sob a curva”(*Area Under Curve*, AUC) reflete o melhor desempenho - maiores valores de AUC significam a melhor abordagem. A Figura 6 apresenta as curvas ROC aplicadas às estatísticas D e Q nos respectivos componentes com variância máxima para cada abordagem. Para o primeiro componente (CP-1), em termos da estatística D, temos um valor de AUC 0,90 para a abordagem multidimensional e temos um valor menor de 0,69 para a abordagem multivariada (Figura 6a). Para a estatística Q, o valor AUC é de 0,89 para a abordagem multidimensional, enquanto que para a abordagem multivariada AUC é 0,62 (Figura 6b). Em relação ao segundo componente (CP-2), a abordagem multidimensional também aparece com maior AUC, com um valor de 0,90 para a estatística D e 0,74 para a abordagem multivariada (Figura 6c). Finalmente, para a estatística Q, a AUC é de 0,91 para a abordagem multidimensional e 0,77 para a abordagem multivariada (Figura 6d). Portanto, as curvas ROC indicam um melhor desempenho para a abordagem multidimensional com os valores mais altos de AUC.

6. Conclusão

Este artigo apresenta um novo método para o monitoramento ambiental urbano. Nossa proposta explora a natureza multidimensional dos dados, a qual combinada com abordagens multivariadas mostrou-se capaz de detectar e diagnosticar dados urbanos discrepantes (*outliers*). A partir dos resultados das curvas ROC, as quais apresentaram valores de AUC para a abordagem proposta em média 20% superiores em relação à aborda-

gem multivariada, podemos afirmar que conseguimos avançar nos métodos de detecção de *outliers* urbanos.

Como estudos futuros, pretendemos realizar uma análise de correlação entre variáveis ambientais, aumentar o número de amostras e de cidades, e aplicar outras técnicas de decomposição tensorial comparando com métodos multivariados clássicos.

Agradecimentos

Thiago Iachiley é bolsista de doutorado da CAPES. André L. L. Aquino agradece ao CNPq, FAPEAL e FAPESP e Danielo G. Gomes agradece ao CNPq (processos 311878/2016-4 432585/2016-8) pelo apoio financeiro.

Referências

Alcala, C. F. and Qin, S. J. (2011). Analysis and generalization of fault diagnosis methods for process monitoring. *Journal of Process Control*, 21:322–330.

Aquino, A. L. L., Junior, O. S., Frery, A. C., Albuquerque, E. L., and Mini, R. A. F. (2014). Musa: Multivariate sampling algorithm for wireless sensor networks. *IEEE Transactions on Computers*, 63:968–978.

Babar, M. and Arif, F. (2017). Smart urban planning using big data analytics to contend with the interoperability in internet of things. *Knowledge-Based Systems*, 77:65–76.

Camacho, J., Villegas, A. P., Teodoro, P. G., and Fernandez, G. M. (2016). Pca-based multivariate statistical network monitoring for anomaly detection. *Computers and Security*, 59:118–137.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58.

Chen, D., Li, X., Wang, L., Khan, S., Wang, J., Zeng, K., and Cai, C. (2015). Fast and scalable multi-way analysis of massive neural data. *IEEE Transactions on Computers*, 64:707–719.

Dong, J.-D., Zhang, Y.-Y., Zhang, S., Wang, Y.-S., Yang, Z.-H., and Wu, M.-L. (2010). Identification of temporal and spatial variations of water quality in sanya bay, china by three-way principal component analysis. *Environmental Earth Sciences*, 60:1673–1682.

Fanaee-T, H. and Gama, J. (2016). Tensor-based anomaly detection: An interdisciplinary survey. *Knowledge-Based Systems*, 98:130–147.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874.

Guardiola, I. G., Leon, T., and Mallor, F. (2014). A functional approach to monitor and recognize patterns of daily traffic profiles. *Transportation Research Part B*, 65:119–136.

Hotelling, H. (1947). *Multivariate quality control.* In: Techniques of statistical analysis. NewYork: McGraw-Hill.

Ibrahim, A. T. H., Victor, C., Nor, B. A., Kayode, A., Ibrar, Y., Abdullah, G., Ejaz, A., and Haruna, C. (2016). The role of big data in smart city. *International Journal of Information Management*, 36:748–758.

Khatib, E. J., Barco, R., Munoz, P., Bandera, I., and Serrano, I. (2016). Self-healing in mobile networks with big data. *IEEE Communications Magazine*, 54:114–120.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *Society for Industrial and Applied Mathematics*, 51:455–500.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. John Wiley and Sons.

Li, J., Han, G., Wen, J., and Gao, X. (2011). Robust tensor subspace learning for anomaly detection. *International Journal of Machine Learning and Cybernetics*, 2:89–98.

Osanaiye, O., Choo, K.-K. R., and Dlodlo, M. (2016). Distributed denial of service (ddos) resilience in cloud: Review and conceptual cloud ddos mitigation framework. *Journal of Network and Computer Applications*, 67:147–165.

Programme, U. N. H. S. (2016). *World Cities Report 2016: Urbanization and Development : Emerging Futures*. UN Habitat.

Rathore, M. M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Knowledge-Based Systems*, 101:63–80.

Singh, K. P., Malik, A., Singh, V. K., and Sinha, S. (2006). Multiway data analysis of soils irrigated with wastewater-a case study. *Chemometrics and Intelligent Laboratory Systems*, 83:1–12.

Slavakis, K., Giannakis, G. B., and Mateos, G. (2014). Modeling and optimization for big data analytics. *IEEE Signal Processing Magazine*, 31:18–31.

Souza, T. I. A., Magalhães, D. M. V., and Gomes, D. G. (2017). Aplicando estatística multivariada para detecção e diagnóstico de anomalias em dados urbanos. *Anais do I Workshop de Computação Urbana (CoUrb)*, 1:72–85.

Steed, C. A., Ricciuto, D. M., Shipman, G., Smith, B., Thornton, P. E., Wang, D., Shi, X., and Williams, D. N. (2013). Big data visual analytics for exploratory earth system simulation analysis. *Computers And Geosciences*, 61:71–82.

Suzhi, B., Rui, Z., Zhi, D., and Shuguang, C. (2015). Wireless communications in the era of big data. *IEEE Communications Magazine*, 53:190–199.

United Nations, D. o. E. and Social Affairs, P. D. (2015). World urbanization prospects: The 2014 revision, highlights.

Xu, Z., Yan, F., and Qi, Y. (2015). Bayesian nonparametric models for multiway data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:475–487.

Zhang, K., Ni, J., Yang, K., Liang, X., Ren, J., and Shen, X. (2017). Security and privacy in smart city applications: Challenges and solutions. *IEEE Communications Magazine*, 17:122–129.

Zhang, L., Zhang, L., Tao, D., and Huang, X. (2011). A multifeature tensor for remote-sensing target recognition. *IEEE Geoscience and Remote Sensing Letters*, 8:374–378.