Aplicando Estatística Multivariada para Detecção e Diagnóstico de Anomalias em Dados Urbanos

Thiago I. A. Souza, Deborah M.V. Magalhães, Danielo G. Gomes

[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

[thiagoiachiley, deborah, dgomes]@great.ufc.br

Abstract. *By 2020 around 40 ZB of data will be generated per year. In a such scenario of smart cities, for example, the analysis and mining of a large amount of data generated by its residents can help public managers to improve and deploy services aimed at the welfare of the citizens. In this context, the anomalies detection gains importance in the effective environmental monitoring of an intelligent urban space. Here we apply multivariate statistics in urban data (e.g. temperature, humidity, pollutant gases, noise level) in the following methodological sequence: (i) Principal Component Analysis (PCA) to reduce data dimensionality; (Ii) D- and Q-statistics for anomalies detection and (iii) CDC (Complete Decomposition Contribution) method for diagnosis (causes) of the anomalies found. We analyzed two real databases of the Smart Citizen platform and the results point to the efficiency of our proposal, indicating which environmental variables have the highest impact on the dataset anomalous pattern*

Resumo. *Estima-se que até 2020 cerca de 40 ZB (Zettabytes) de dados serão gerados por ano. Em um cenário de cidades inteligentes, por exemplo, a análise e mineração de um grande volume de dados gerados pelos seus moradores pode ajudar os gestores públicos na melhoria e implementação de serviços voltados ao bem-estar do cidadão. Neste contexto, a detecção de anomalias (valores discrepantes) ganha importância no monitoramento ambiental eficaz de um espaço urbano. Neste artigo, aplicamos estatística multivariada em dados urbanos (temperatura, umidade, gases poluentes, nível de ruído sonoro) na seguinte sequência metodológica: (i) PCA (Principal Component Analysis) para redução da dimensionalidade dos dados; (ii) Estatísticas D e Q para detecção de anomalias e (iii) Método CDC (Complete Decomposition Contribution) para diagnóstico (causas) das anomalias encontradas. Analisamos duas bases de dados reais da plataforma Smart Citizen e os resultados apontam para a eficiência da nossa proposta, indicando quais variáveis ambientais apresentaram maior impacto no comportamento anômalo dos dados.*

1. Introdução

De acordo com a Organização das Nações Unidas (ONU), atualmente há mais pessoas vivendo nas cidades do que no campo e a estimativa é de que a população mundial urbana, que não passava dos 30% em 1950, atingirá um patamar de 66% em 2050

[United Nations and Social Affairs 2015]. Esse adensamento dos aglomerados urbanos implica evidentemente em um aumento dos problemas típicos das cidades, sobretudo das capitais e metrópoles. Por conseguinte, a infraestrutura e os serviços atuais providos pelo poder público podem revelar-se insuficientes para lidar com questões cada vez mais críticas de mobilidade urbana, segurança pública, acesso à saúde, poluição ambiental e gestão de ativos (energia elétrica, água e gás). Apesar de indesejáveis obstáculos do nosso cotidiano, estes problemas geram oportunidade para um planejamento urbano alinhado com o conceito moderno de cidades inteligentes, agregando soluções baseadas em Internet das Coisas (*Internet of Things*, IoT) [Gomes and Forster 2015], bem como análise e mineração de dados sensoriados por diversos e heterogêneos objetos inteligentes distribuídos em um espaço urbano [Rathore et al. 2016] [Abellá-García et al. 2015].

Nesta perspectiva de cidades populosas e inteligentes, nas quais uma vasta quantidade de dispositivos, sejam eles fixos (e.g. nós sensores, prédios) ou móveis (e.g. *smartphones*, veículos), sensoriam o ambiente urbano e trocam informações entre si, podemos considerar um volume massivo e heterogêneo de dados gerados, processados e em trânsito. Monitorar, analisar e minerar esse volume de dados são desafios da chamada Computação Urbana [Kamienski et al. 2016] [Silva and Loureiro 2016]. O desafio aumenta criticamente quando esses dados apresentam anomalias ou *outliers*, i.e. valores discrepantes.

Detecção de anomalias é uma das tarefas fundamentais na mineração de dados, juntamente com modelagem preditiva, análise de *cluster* e análise exploratória [Zhang et al. 2010]. Anomalias podem ter uma influência considerável nos resultados de uma análise através de estimativas tendenciosas de parâmetros ou mesmo previsões incorretas. Entretanto, apesar das anomalias em geral serem causadas por erros de medições, algumas vezes elas podem indicar eventos de interesse (e.g. detecção de fraude em cartões de crédito, intrusão em rede, rastreamento de alvos, monitoramento de sinais vitais [Cucina et al. 2014]). No contexto deste artigo, cujo escopo é de cidades inteligentes, as anomalias podem indicar congestionamentos, níveis elevados de poluição (do ar, sonora), ilhas de calor, dentre outros eventos de interesse e que estejam fora do padrão normal.

A partir de conjuntos de dados sensoriados nas cidades de Quito e Hong Kong, este artigo propõe a análise e mineração dos dados de cidades inteligentes aplicando estatística multivariada baseada na Análise de Componentes Principais (PCA). Nosso objetivo central é a detecção de anomalias de variáveis ambientais urbanas e respectivo diagnóstico de suas causas. Para tal, seguimos os seguintes passos metodológicos: (i) validação das bases de dados coletadas para a utilização do método multivariado PCA (subseção 3.3.1); (ii) redução de ambos os conjuntos de dados analisados para extração das componentes principais mais representativas (subseção 3.3.2); (iii) detecção de anomalias no espaço reduzido das componentes principais selecionadas da PCA (subseção 3.3.3); (iv) diagnóstico das causas das anomalias (subseção 3.4).

Até onde sabemos, este é o primeiro artigo a aplicar estatística multivariada [Green 2011] e o método *Complete Decomposition Contribution* [Acala and Qin 2011] para detecção e diagnóstico de anomalias, respectivamente, no contexto de cidades inteligentes e da computação urbana.

2. Estatística Multivariada Baseada em PCA

Os métodos de análise multivariada têm surgido com o intuito de extrair do conjunto de dados os mais significativos padrões de informações. A estatística multivariada abrange um determinado conjunto de métodos estatísticos, que tornam em estudos mais robustos e apreciáveis, a análise da informação .

Diante deste contexto a estatística multivariada, no que concerne a análise exploratória dos dados, busca encontrar padrões de informações intrínsecos dos complexos conjuntos de dados. A Análise de Componentes Principais (PCA) tem se destacado como uma das principais técnicas exploratória de dados na redução do volume de dados e identificação de tendências.

A PCA tem como objetivo identificar a relação entre características extraídas dos dados visando sua redução, eliminação de sobreposições desprezíveis e a escolha das formas mais relevantes das mesmas a partir de combinações lineares das variáveis originais [Machado et al. 2004], [Camacho 2010]. Desta forma, a PCA transforma variáveis discretas em coeficientes descorrelacionados através de uma transformação linear aplicada aos dados, de modo que, os dados transformados tenham suas projeções mais relevantes preservadas (componentes principais) para análise, em detrimento daquelas menos relevantes, os quais são desprezados. Para atingir esse objetivo, a PCA envolve o cálculo da decomposição em autovalores de uma matriz de covariância de dados ou da decomposição em valores singulares de uma matriz de dados.

Para o cálculo da PCA considere uma matriz de dados $\mathbf{X}$ $^{m\times n}$ formada por n vetores coluna, cada um de dimensionalidade m, $\mathbf{X} = [\mathbf{x}_1 \vdots \mathbf{x}_2 \vdots \cdots \vdots \mathbf{x}_n]$. Desta forma, PCA pode ser expressa da seguinte forma:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}, \tag{1}$$

em que t denota o símbolo de transposição da matriz, $\mathbf{T}$ é a matriz dos scores ($m \times n$), $\mathbf{P}$ ($n \times n$), também conhecida como matriz de carregamento, é a matriz formada pelos autovetores da matriz de covariância , e $\mathbf{E}$ é a matriz de resíduos ($m \times n$). A matriz de covariância é definida como se segue

$$= \frac{1}{m-1}\mathbf{X}^t\mathbf{X} = \mathbf{P} \ \ \mathbf{P}^t, \quad com \quad \mathbf{P}\mathbf{P}^t = \mathbf{P}^t\mathbf{P} = \mathbf{I}, \tag{2}$$

em que $= diag(\ _1, ..., \ _n)$ é uma matriz diagonal contendo os autovalores da matriz de covariância em ordem decrescente, e $\mathbf{I}$ é a matriz identidade. Cada autovalor codifica a variação relacionada ao componente correspondente. Então, a partir dos autovalores a variância explicada (VE) e a variância explicada cumulativa (VEC) associada a cada k-ésima componente extraída do modelo PCA podem ser calculadas:

$$VE_k = \frac{\ _k}{\sum_{j=1}^{n} \ _j}, \tag{3}$$

$$VEC_k = \frac{\sum_{j=1}^{k} \ _j}{\sum_{j=1}^{n} \ _j}. \tag{4}$$

Considerando que apenas k componentes significativas (com $k < n$) sejam retidas pela técnica PCA, então a matriz de carregamento $\mathbf{P}$ diminui de $n \times n$ para $n \times k$ e as amostras são projetadas para um espaço dimensional inferior definido pelas componentes principais significativas dada pela seguinte expressão:

$$\mathbf{T}_k = \mathbf{P}_k^t \mathbf{X}_k. \tag{5}$$

Dentre as principais aplicações da PCA destaca-se a redução de dimensionalidade [Li 2016], o que é muito útil para lidar com conjuntos de dados com elevado grau de correlação cruzada entre as variáveis [Harrou et al. 2016]. Esta capacidade é de extrema relevância para a detecção de anomalias, haja vista que um grande número de variáveis das mais diversas fontes podem ser analisadas ao mesmo tempo [Camacho and Ferrer 2014]. Além disso, padrões de comportamento anômalos podem ser detectados e interpretados a partir da inspeção da contribuição das variáveis envolvidas [Acala and Qin 2011].

3. Material e Métodos

Esta seção descreve os aspectos metodológicos da pesquisa que nortearam a obtenção dos dados, os métodos multivariados que foram aplicados no contexto do processamento, análise e extração de informações relevantes dos conjuntos de dados em estudo, bem como a detecção e o diagnóstico das anomalias.

3.1. Base de Dados e Contexto da Amostra

Neste estudo foram utilizados os dados obtidos da plataforma Smart Citizen[1], um projeto que tem como objetivo conectar as pessoas com seus ambientes e cidades. O Smart Citizen Kit[2] baseia-se na geolocalização para coleta e compartilhamento de dados e fornece dados em tempo real de temperatura, umidade, ruído, níveis de monóxido de carbono (CO), dióxido de nitrogênio (NO_2) e luminosidade (com uma variável que envolve a luminosidade captada do sol e uma outra variável, completando as 7, que capta a luminosidade de placas solares). Até o momento em que se deu esta pesquisa, a referida plataforma contava com 661 nós sensores (online, offline, indoor e outdoor) de monitoramento de variáveis de ambientes urbanos espalhados pelos cinco continentes. Entretanto, nem todos os nós sensores encontravam-se ativos. Além disso, foi obervado que grande parte das medições não são contínuas no tempo, havendo várias lacunas temporais nas leituras dos sensores. Desta forma, escolhemos como objeto de estudo os nós sensores das cidades de Quito e Hong Kong por possuírem nós sensores *online* e com medições capturadas em tempo real sem falhas nas leituras.

3.2. Organização e Simulação dos Dados Multivariados

A natureza do conjunto de dados obtidos nesta pesquisa é multidimensional [Prada et al. 2012]. Os dados capturados da plataforma Smart Citizen foram organizados em um tensor $\mathcal{X}$ (Figura 1) com dimensões 2160 tempo em horas (linhas) versus 7 variáveis monitoradas (colunas) versus 2 cidades analisadas (cada cidade uma matriz).

[1]https://smartcitizen.me/

[2]http://docs.smartcitizen.me/#/start/hardware

Assim, o tensor $\mathcal{X}$ foi matriciado de forma que possamos analisar individualmente as matrizes de cada cidade. As marcações em vermelho da Figura 1 representam as possíveis anomalias existentes no conjunto de dados que serão detectadas e diagnosticadas neste trabalho.

A modelagem de dados deste estudo tem a complexidade de lidar com um arranjo tensorial de dados 3D, mas ao mesmo tempo oferece a possibilidade de explorar uma abordagem matricial de dados através da aplicação da técnica multivariada PCA. Para tanto, cada matriz contendo as variáveis ambientais, correspondendo às cidades investigadas, foi analisada individualmente pela técnica multivariada.

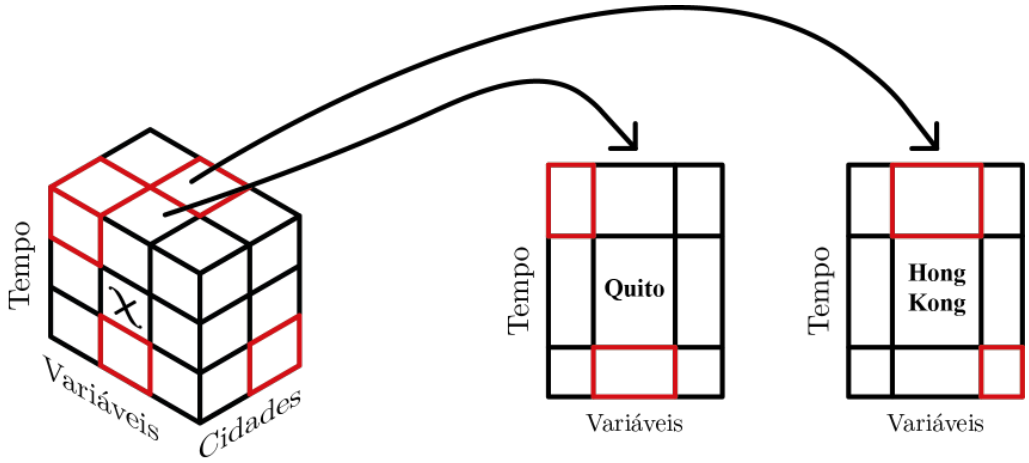

Figura 1. Arranjo tensorial 3D dos dados tratados neste artigo.

3.3. Aplicação da Estatística Multivariada Baseada na PCA

A estatística multivariada baseada em PCA utilizando as estatísticas D e Q no monitoramento de processos foi utilizada em uma gama de cenários, tais como: monitoramento de processos industriais [Zhao and Gao 2016], perdas de pacotes em redes de sensores sem fio [Magan-Carrion et al. 2015], monitoramento de redes [Camacho et al. 2016], dentre outros. Nossa pesquisa se diferencia de tais trabalhos, pelo fato de que utilizamos as estatísticas D e Q como complementares na análise e não de forma comparativa como usualmente acontece. Ou seja, é investigado o caráter exploratório de ambas as estatísticas em que uma pode revelar algum padrão anômalo não identificado pela outra, como também ambas corroborarem com o padrão encontrado. Além disso, as estatísticas D e Q são utilizadas para ampliar a capacidade de detecção de padrões anômalos no espaço reduzido gerado pela PCA. Desta forma, será apresentado a seguir os procedimentos para aplicação da estatística multivariada baseada na PCA.

3.3.1. Validação da Análise Exploratória

Para avaliar a adequação do conjunto de dados coletados nesta pesquisa à aplicação da PCA, utilizou-se a medida da adequação da Amostra de Kaiser-Meyer-Olkin (KMO) [Green 2011]. O teste indica que um valor acima de 0,6 é considerado adequado

para aplicação da técnica multivariada [Green 2011]. Além de utilizar o KMO, avaliou-se também a matriz de correlação dos parâmetros ambientais coletados através do teste de esfericidade de Bartlett [Green 2011] no intuito de testar a hipótese de que a matriz de correlação é uma matriz identidade. Caso seja verdadeira a hipótese, significa dizer que não há correlações entre as variáveis, e portanto, a análise não pode ser realizada [Dunteman 1989].

3.3.2. Modelagem da PCA

Conforme descrito na Seção 2, PCA tem sido amplamente aplicada no domínio de redução de dimensionalidade. Tal domínio, explorado neste trabalho, permite que a análise de dados concentre-se nas informações mais relevantes dos dados originais, contidas agora nas novas variáveis descorrelacionadas geradas pelo modelo multivariado. Para tanto, procedeu-se com a extração das componentes mais significativas através da variação de cada componente principal codificada por cada autovalor correspondente (Equações 3 e 4).

3.3.3. Detecção de Anomalias

Neste artigo, o monitoramento das grandezas físicas ambientais para a detecção de anomalias baseada na PCA consiste no cálculo de duas estatísticas, a saber: estatística D [Hotelling 1947] calculada a partir dos scores das componentes principais; e estatística Q [Jackson and Mudholkar 1979] calculada a partir dos resíduos das componentes principais. Assim, tanto a estatística D quanto a estatística Q são caculadas pelas seguintes expressões, respectivamente:

$$D = \sum_{i=1}^{k} \frac{p_i^2}{i}, \tag{6}$$

$$Q = \sum_{i=1}^{n} (e_i)^2, \tag{7}$$

em que p_i representa a i-ésima componente principal, $_i$ é o autovalor correspondendo à i-ésima componente principal, e e_i representa o valor residual correspondente à n-ésima variável. Ambas as estatísticas, além de serem complementares, geram gráficos de controle que permitem o monitoramento dos mais variados processos multivariados [Camacho et al. 2016].

3.3.4. Diagnóstico de Anomalias

O monitoramento de processos multivariados não consiste apenas em detectar eventos anormais. Além da detecção de anomalias, deseja-se um sistema dignóstico capaz de encontrar a fonte responsável por gerar anomalias [Shang et al. 2016]. Desta forma, a abordagem mais generalizada para diagnósticos em estatística multivariada é a

contribuição das parcelas [Nomikos and MacGregor 1994]. Os gráficos de contribuição identificam a influência das variáveis para um valor anômalo detectado pelas estatísticas de monitoramento (D e Q). Em outras palavras, trata-se de diagramas de barras em que a contribuição do conjunto de variáveis para uma estatística (D e/ou Q) será inspecionada. Neste trabalho foi utilizado o método diagnóstico denominado Contribuição de Decomposição Completa *Complete Decomposition Contribution* - CDC) [Acala and Qin 2011].

$$CDC_i = \mathbf{x}_m \mathbf{M}^{\frac{1}{2}} \; {}^{t}_{i} \; {}_{i}\mathbf{M}^{\frac{1}{2}} \mathbf{x}^t_m, \tag{8}$$

em que $_i$ é um vetor $1 \times n$ que corresponde a i-ésima coluna da matriz identidade, $_i = [00...1...0]^t$, e $\mathbf{M}$ é definido de acordo com a estatística dignosticada. Para a estatística D, $\mathbf{M}$ é definido como:

$$\mathbf{M} = \mathbf{P}_k \; {}^{-1}\mathbf{P}^t_k, \tag{9}$$

em que é uma matriz diagonal contendo as k primeiras componentes principais ou autovetores da matriz de covariância , definida na Seção 3 pela Equação 2. Para a estatística Q, $\mathbf{M}$ é definido como:

$$\mathbf{M} = \mathbf{P}_k \mathbf{P}^t_k, \tag{10}$$

em que $\mathbf{P}_k$ corresponde a uma matriz de k autovetores extraídos de $\mathbf{E}^t_k \mathbf{E}_k$, em que $\mathbf{E}$ é a matriz de resíduos apresentada na Seção 2 pela Equação 1.

4. Resultados

Nesta seção são apresentados os resultados obtidos a partir da aplicação da análise dos componentes principais considerando os testes estatísticos para validação da base de dados, seleção das componentes principais da PCA e, por fim, as etapas de detecção de anomalias e respectivo diagnóstico.

4.1. Testes Estatísticos e Validação

Conforme colocado na subseção 3.3.1, para a validação da base de dados visando aplicação da PCA, foram realizados os testes de Esfericidade de Bartlett e a medida da adequação da Amostra de Kaiser-Meyer-Olkin (KMO) [Green 2011]. Os resultados são apresentados na Tabela 1. O teste de KMO examina o ajuste dos dados tomando todas as variáveis simultaneamente e provê uma informação sintética sobre os dados indicando a proporção da variância dos dados que pode ser considerada comum a todas as variáveis. O teste indica que, quanto mais próximo da unidade, melhor o resultado. Já o teste de esfericidade de Bartlett, testa a hipótese de que a matriz de correlação é uma matriz identidade, isto é, que não há correlações entre as variáveis [Dunteman 1989].

Juntos, ambos os testes fornecem um padrão mínimo que deve ser observado antes que a análise de componentes principais seja relizada. Neste contexto, os valores obtidos pelos testes (KMO 0,6 e Esfericidade de Bartlett com rejeição da hipótese nula conforme recomendado por [Green 2011]) apontam para a validação sobre a utilização da PCA com respeito à matriz de correlação.

Teste	Base de Dados I - Quito	Base de Dados II - Hong Kong
Adequação da Amostra - KMO	0,70	0,65
Esfericidade de Bartlett	16267,027	38232,98

Tabela 1. Testes e Validação dos Dados.

4.2. Seleção das Componentes da PCA

Os procedimentos de seleção das componentes principais da PCA compreendem-se em duas etapas, a saber: (i) análise do percentual da variância total explicada por cada componente associado ao seu respectivo autovalor; (ii) e o critério de Kaiser, o qual diz que as componentes a serem consideradas devem apresentar autovalores superiores a unidade [Green 2011].

	Parâmetros Analisados		
Componentes	Autovalores	% Variância	% Variância Acumulada
1	3,87	55,34	55,34
2	**1,83**	**26,15**	**81,49**
3	0,60	8,60	90,10
4	0,36	5,11	95,20
5	0,22	3,20	98,39
6	0,074	1,06	99,45
7	0,55	0,54	100

Tabela 2. Variância Explicada - Base de Dados I - Quito.

	Parâmetros Analisados		
Componentes	Autovalores	% Variância	% Variância Acumulada
1	3,462	49,45	49,45
2	**1,676**	**23,94**	**73,40**
3	0,880	12,56	85,96
4	0,603	8,60	94,57
5	0,297	4,24	98,82
6	0,082	1,17	99,96
7	$2{,}6 \times 10^{-7}$	$3{,}7 \times 10^{-6}$	100

Tabela 3. Variância Explicada - Base de Dados II - Hong Kong.

Observando o percentual de variância explicada que cada componente principal preserva, pode-se reduzir o conjunto de dados original que engloba as 7 variáveis ambientais, para um número menor que apresente significativo percentual de informação útil para ser analisada. Desta forma, reduziu-se ambos os conjuntos originais de dados para 2 componentes principais, as quais representam um percentual de variância explicada em torno de 81%, para a base de dados I (Tabela 2), e de 73% de informação dos dados originais para a base de dados II (Tabela 3).

No intuito de fundamentar a escolha das componentes principais que serão objetos de análise, considerou-se além do valor do percentual de variância explicada retida por cada componente principal, o critério de Kaiser [Green 2011]. Portanto, este critério

(que toma os autovalores superiores a unidade) corrobora com o percentual de variância explicada retida pelas duas primeiras componentes, para ambas as bases de dados.

4.3. Detecção e Diagnóstico de Anomalias

4.3.1. Etapa de Detecção

Após a seleção das componentes principais para cada base de dados, aplicou-se as estatísticas D (Equação 6) e Q (Equação 7) para a detecção das anomalias nas duas componentes principais selecionadas. Nesta perspectiva, para a cidade de Quito tanto a estatística D quanto a estatística Q não identificaram nenhum comportamento anômalo nos dados para a componente principal #1 (Figura 2) com relação à janela de tempo analisada, indicando normalidade nos parâmetros ambientais coletados pelos sensores. O mesmo padrão de normalidade é observado em relação à componente principal #2, em que as estatísticas D (Figura 3(a)) e Q (Figura 3 (b)) não identificaram nenhuma anomalia.

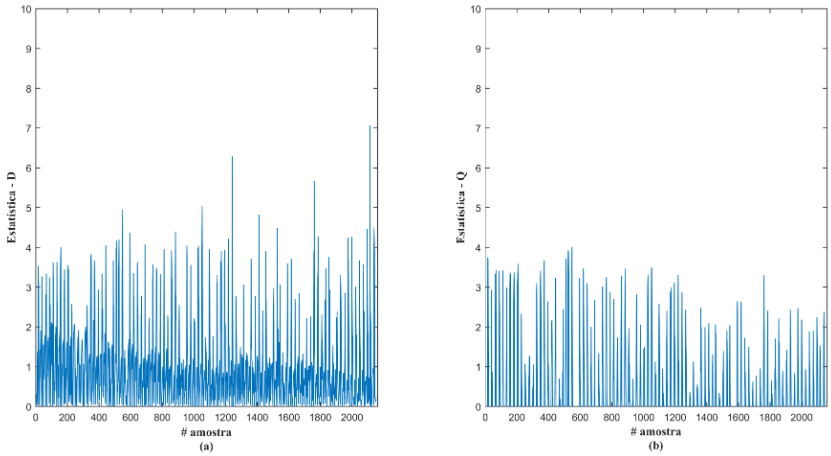

Figura 2. Componente principal #1 - Cidade de Quito.

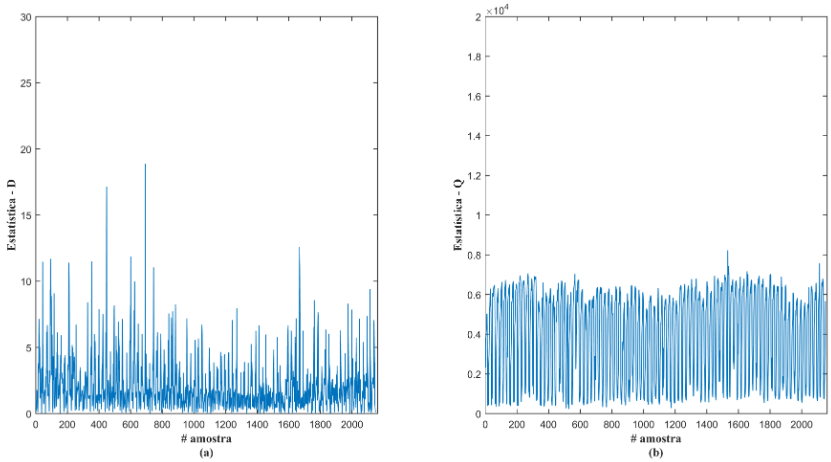

Figura 3. Componente principal #2 - Cidade de Quito.

Para o segundo cenário analisado, cidade de Hong Kong, as estatísticas identificaram uma alteração no comportamento dos dados no espaço das duas componentes selecionadas. Desta forma, a estatítica D apontou para uma anomalia presente nos parâmetros ambientais para a componente principal #1 (Figura 4 (a)), a saber a coleta $\#1208$. Para a estatística Q a observação anômala detectada corresponde à mesma coleta $\#1208$ (Figura 4 (b)). Neste caso, a anomalia detectada pela estatística D é também detectada pela estatística Q, corroborando ambas para o monitoramento dos parâmetros ambientais.

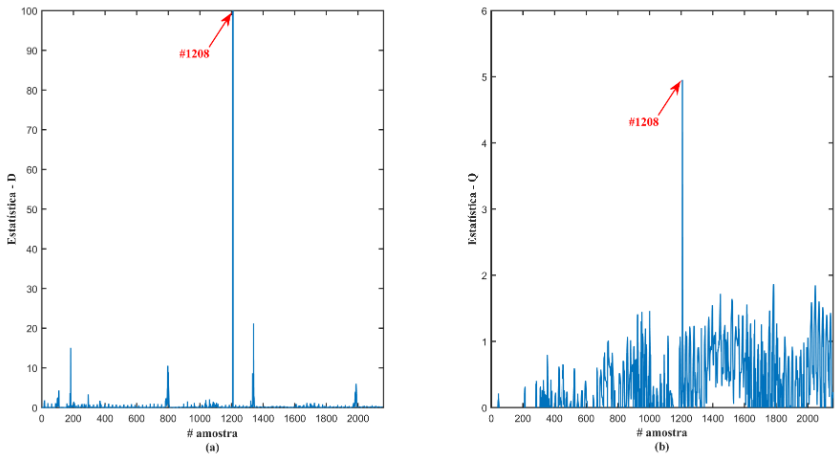

Figura 4. Componente principal #1 - Cidade de Hong Kong.

Em relação à componente principal #2, a estatística D também identifica a anomalia $\#1208$ (Figura 5 (a)), sendo corroborada com a estatística Q que identifica novamente o mesmo padrão anômalo (Figura 5 (b)). Desta forma, percebe-se que para ambas as duas componentes principais selecionadas, os gráficos de monitoramento das estatísticas D e Q apontam para uma mesma observação que se desvia das demais.

Comparando os resultados obtidos com os dados coletados da plataforma Smart Citizen, observamos que não há ocorrência de eventos anômalos na cidade de Quito (Figuras 2 e 3). Já com relação à cidade de Hong Kong, há destaque para um evento anômalo que ocorre na coleta #1208 (correspondendo às 17 horas do dia 06/06/2016, vide Figuras 4 e 5), momento em que ocorre a anomalia.

Os gráficos retornados pelas estatísticas D e Q se mostraram úteis no tocante ao monitoramento de tendências e padrões das variáveis ambientais de cidades inteligentes. Quando o objetivo é identificar anomalias, métodos clássicos de monitoramento estatístico podem ser ineficazes na identificação de tendências ou padrões cíclicos das métricas [O'Leary et al. 2016]. Assim, dentre as amostras medidas, identificamos a mais discrepante, permitindo estabelecer conclusões sobre o padrão das variáveis ambientais analisadas.

4.3.2. Etapa de Diagnóstico

O monitoramento de processos não consiste apenas em detectar eventos anormais, mas também em encontrar as variáveis que contribuam para o desvio. Desta forma, uma

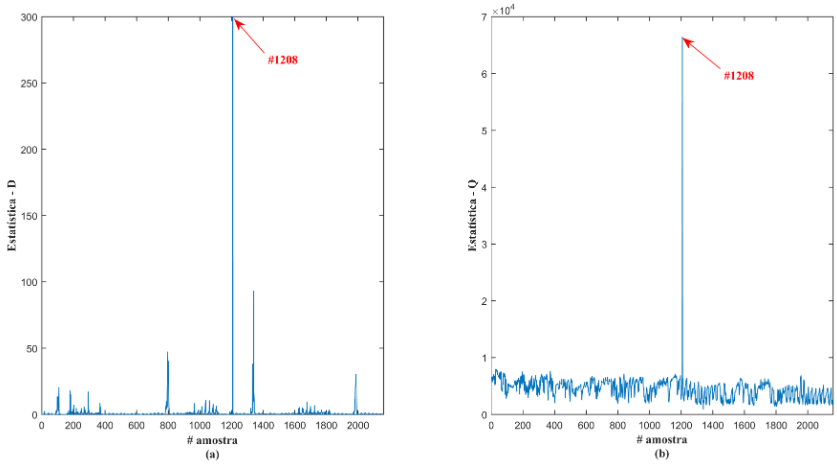

Figura 5. Componente principal #2 - Cidade de Hong Kong.

vez que uma anomalia foi detectada, as parcelas de contribuição do modelo PCA são calculadas para identificar as possíveis variáveis responsáveis pelo desvio. Assim, foram calculadas as parcelas de contribuição do modelo PCA no espaço das duas componentes principais selecionadas para a anomalia detectada (coleta $\#1208$) utilizando o método diagnóstico CDC, tanto para a estatística D quanto para Q (Figuras 6 e 7). É importante destacar que como nenhuma anomalia foi detectada pelas estatísticas para a base de dados da cidade de Quito, a etata de diagnóstico corresponde a apenas a cidade de Hong Kong.

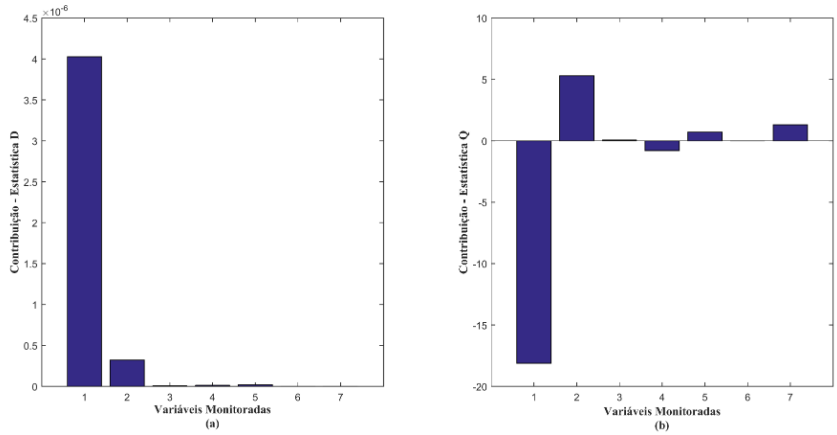

Figura 6. Plots Diagnósticos - Componente principal #1, cidade de Hong Kong.

No caso da Figura 6(a), observa-se que o método CDC calculado sobre a estatística D aponta para a variável temperatura como a responsável por causar o desvio no comportamento dos dados, no espaço da componente principal #1. Entretanto, quando calculado sobre a estatística Q, o método CDC aponta para uma influência negativa da temperatura (Figura 6(b)). Para tal padrão, detectado na coleta $\#1208$, observamos que quando verificado na base de dados original, este é o momento no qual foi registrado o menor valor de temperatura pelos sensores no intervalo de tempo monitorado, caracterizando uma anomalia. Portanto, o método CDC para a estatística D permitiu identificar que dentre as variáveis ambientais monitoradas, a que contribuiu para o desvio da coleta $\#1208$

foi a temperatura. Tal análise foi complementada pelo método CDC para a estatística Q, indicando o sentido (negativo) da contribuição da variável temperatura, ou seja, na realidade não houve um pico (um valor de máximo) no valor da temperatura como poderia se imaginar através da estatística D, mas sim um vale (um valor de mínimo) revelado pela estatística Q. Podemos inferir que este desvio influenciado pela temperatura pode indicar um erro instrumental do sensor, uma vez que até o momento do *outlier* detectado (amostra $\#1208$), não havia nenhuma variação discrepante dos dados.

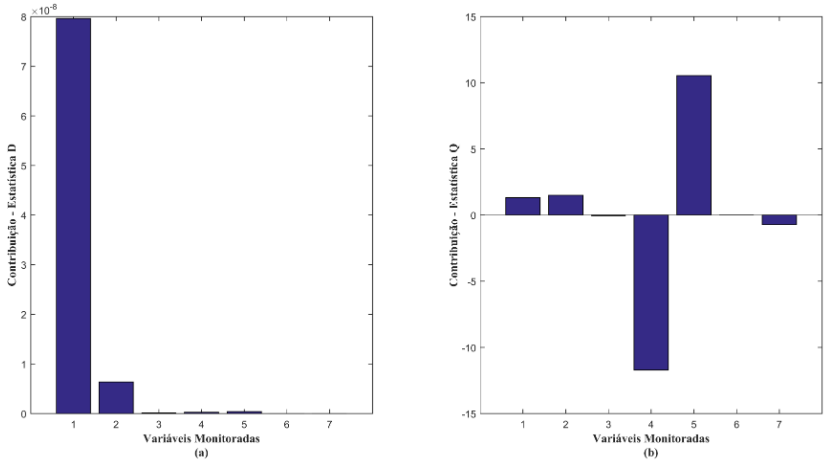

Figura 7. Plots Diagnósticos - Componente principal #2, cidade de Hong Kong.

Com relação ao espaço da componente principal #2, o método CDC calculado sobre a estatística D (Figura 7(a)) continua apontando para a variável temperatura como a responsável pelo desvio da amostra $\#1208$. Entretanto, quando aplicado à estatística Q (Figura 7(b)), o método destaca outras duas variáveis como sendo as que influenciaram no comportamento anômalo, a saber as variáveis: CO e NO_2. Ao se verificar a base de dados original, observamos que este é o momento no qual para ambas as variáveis (CO e NO_2) as medições excederam os padrões dos valores sensoriados no intervalo de tempo monitorado. Voltando ainda a verificar a Figura 5(a), verificamos que existem outros *outliers* (além da amostra $\#1208$), porém não tão significativos. Esta observação permite-nos concluir a inferência acerca do diagnóstico gerado pela Figura 7(b), em que o CO e NO_2 influenciaram o comportamento anômalo dos dados indicando que a poluição gerada por eles não se deveu a uma única medição anômala mas sim a uma série de medições que excederam determinados limites em um dado intervalo de tempo [Martinez et al. 2014], como a que ocorreu na amostra $\#1208$.

5. Conclusão

Neste artigo, aplicamos a técnica da estatística multivariada PCA no monitoramento de variáveis ambientais urbanas. Apesar dos dados serem coletados de diferentes e heterogêneos sensores, nossa proposta propicia a detecção de anomalias com identificação do exato momento de sua ocorrência. A partir da técnica de diagnóstico proposta, é possível afirmar quais variáveis causam o desvio da amostra.

A principal contribuição deste artigo é o diagnóstico das causas das anomalias detectadas pelas estatísticas D e Q, através das quais identificamos quais variáveis ambi-

entais contribuíram para o comportamento anômalo dos dados. Para tanto, propomos o uso do método estatístico *Complete Decomposition Contribution* para cálculo das parcelas de contribuição das variáveis para uma determinada anomalia detectada.

Os resultados alcançados podem oferecer indicadores para um planejamento urbano inteligente e sustentável. A partir destes indicadores, gestores públicos (e mesmo cidadãos participativos) podem tomar decisões mais apropriadas acerca de mudanças críticas de temperatura, umidade, ruído, níveis de monóxido de carbono (CO), dióxido de nitrogênio (NO2) e luminosidade.

Como perspectivas de trabalhos futuros, sugere-se (i) aumentar o número de observações e estender o estudo para outras cidades da plataforma Smart Citizen, (ii) considerar a natureza multidimensional dos dados e utilizar métodos tensorais de decomposição de dados, (iii) utilizar outras técnicas da estatística multivariada na detecção de *outliers* e outros métodos para diagnóstico de anomalias.

Referências

Abellá-García, A., de Urbina-Criado, M. O., and De-Pablos-Heredero, C. (2015). The ecosystem of services around smart cities: An exploratory analysis. *Procedia Computer Science*, 64:1075 – 1080.

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

Camacho, J. (2010). Missing-data theory in the context of exploratory data analysis. *Chemometrics and Intelligent Laboratory Systems*, 103:98–104.

Camacho, J. and Ferrer, A. (2014). Cross-validation in pca models with the element-wise k-fold (ekf) algorithm: Practical aspects. *Chemometrics and Intelligent Laboratory Systems*, 131:37–50.

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.

Cucina, D., Salvatore, A., and Protopapas, M. K. (2014). Outliers detection in multivariate time series using genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, 132:103–110.

Dunteman, G. H. (1989). *Principal components analysis (Quantitative Applications in the Social Science).* Sage Publications.

Gomes, D. G. and Forster, A. (2015). Introduction to the special issue on green engineering: Towards sustainable smart cities. *Computers & Electrical Engineering*, 45:141–142.

Green, P. E. (2011). *Multivariate Data Analysis.* Cengage Learning.

Harrou, F., Ramahaleomiarantsoa, J. F., Nounou, M. N., and Nounou, H. N. (2016). A data-based technique for monitoring of wound rotor induction machines: A simulation study. *Engineering Science and Technology, an International Journal*, 19:1424–1435.

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