UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

THIAGO IACHILEY ARAUJO DE SOUZA

MÉTODOS DE DETECÇÃO DE *OUTLIERS* PARA O MONITORAMENTO AMBIENTAL DE ESPAÇOS URBANOS INTELIGENTES VIA ANÁLISE MULTIVARIADA E MULTIDIMENSIONAL

FORTALEZA

2020

THIAGO IACHILEY ARAUJO DE SOUZA

MÉTODOS DE DETECÇÃO DE *OUTLIERS* PARA O MONITORAMENTO AMBIENTAL DE ESPAÇOS URBANOS INTELIGENTES VIA ANÁLISE MULTIVARIADA E MULTIDIMENSIONAL

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Engenharia Elétrica

Orientador: Prof. Dr. Danielo Gonçalves Gomes

Coorientador: Prof. Dr. André Luiz Lins de Aquino

FORTALEZA

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S236m Souza, Thiago Iachiley Araujo de.
MÉTODOS DE DETECÇÃO DE OUTLIERS PARA O MONITORAMENTO AMBIENTAL DE ESPAÇOS URBANOS INTELIGENTES VIA ANÁLISE MULTIVARIADA E MULTIDIMENSIONAL / Thiago Iachiley Araujo de Souza. – 2020.
101 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2020.
Orientação: Prof. Dr. Danielo Gonçalves Gomes.
Coorientação: Prof. Dr. André Luiz Lins de Aquino.

1. Detecção de Outliers. 2. Internet das Coisas. 3. Análise Multivariada. 4. Análise Multidimensional. 5. Cidades Inteligentes. I. Título.

CDD 621.38

THIAGO IACHILEY ARAUJO DE SOUZA

MÉTODOS DE DETECÇÃO DE *OUTLIERS* PARA O MONITORAMENTO AMBIENTAL DE ESPAÇOS URBANOS INTELIGENTES VIA ANÁLISE MULTIVARIADA E MULTIDIMENSIONAL

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Engenharia Elétrica

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Danielo Gonçalves Gomes (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. André Luiz Lins de Aquino (Coorientador)
Universidade Federal do Alagoas (UFAL)

Prof. Dr. Heitor Soares Ramos Filho
Universidade Federal de Minas Gerais (UFMG)

Prof. Dr. Thiago Henrique Silva
Universidade Tecnológica Federal do Paraná (UTFPR)

Prof. Dra. Michela Mulas
Universidade Federal do Ceará (UFC)

À minha família, por sua capacidade de acreditar, motivar, incentivar e investir em mim. Mainha, seu exemplo, cuidado e dedicação foi o que me deram, em todos os momentos, a esperança, o fôlego e a força para seguir. Minhas irmãs, pelo apoio incondicional durante toda esta longa e árdua caminhada.

AGRADECIMENTOS

Gratidão a Deus, pelo dom da vida e pela saúde ao longo deste doutorado.

Agradeço a minha mãe, minha mainha (Francisca Monteiro), minhas irmãs (Márcia, Suely, Hanne), meus primos (Joatan e João Neto) e minha família em geral, pelo amor e o apoio incondicional que recebi durante toda esta minha jornada. Gratidão àquela que dilata o meu coração (Carol), meu amor, pelo companheirismo, apoio e compreensão. Meus agradecimentos também aos meus importantes amigos: Arthur, Manuel, Rafael Braga e Mateus.

Minha profunda gratidão ao meu orientador Prof. Danielo G. Gomes, não apenas pela sábia e prudente orientação, mas também pelos conselhos, pelas longas conversas de orientação (não apenas para o doutorado, mas para a vida), pela paciência e confiança com que me conduziu ao longo deste doutorado.

Gratidão ao meu co-orientador Prof. André L. L. Aquino (ALLA), pelas discussões, pelas conduções, pelas sugestões e valiosas revisões, e por sempre estar disponível em todas as empreitadas propostas neste doutorado. Enfim, muito obrigado professor.

Agradeço à coordenação do PPGETI, bem como a todos os professores e à secretaria do pelo apoio ao longo de todo o doutorado e pela contribuição direta ou indireta em minha formação acadêmica.

“Posso todas as coisas naquele que me fortalece.”

(Filipenses 4:13)

RESUMO

Desde 2007, pela primeira vez na história da humanidade, mais pessoas vivem nas cidades do que no campo. Segundo projeções da Organização das Nações Unidas (ONU), a população mundial deverá ser cerca de 70% urbana em 2050. Este crescimento urbano exponencial traz consigo problemas críticos e típicos das cidades, tais como os de mobilidade urbana, saúde, segurança pública e de poluição ambiental. Dado que as cidades são um vasto e heterogêneo repositório de dados em potencial, cuja complexidade é proporcional ao seu tamanho e população, uma possível solução para tratar estes problemas parte do monitoramento de eventos associados a dados urbanos. Entretanto, um dos desafios ao lidar com estes dados é reconhecer quais deles estão "fora do padrão"(*outliers*). Esta tese investiga a detecção de *outliers* em dados urbanos sob três abordagens diferentes: (i) abordagem multivariada *offline*, através da qual modelamos os dados como matrizes, suprimindo uma de suas dimensões, e realizamos uma análise multivariada através da técnica multivariada Análise Fatorial Exploratória (AFE); (ii) abordagem multidimensional *offline*, em que modelamos os dados como um tensor de terceira ordem, e realizamos uma análise multidimensional *offline* através da técnica multidimensional HOSVD (*Higher-Order Singular Value Decomposition* - HOSVD); e (iii) abordagem multidimensional *online,* na qual modelamos os dados como um tensor de terceira ordem e realizamos uma análise multidimensional *online* combinando a técnica multidimensional HOSVD com a estratégia da janela deslizante. Foram coletados dados reais da plataforma de monitoramento ambiental urbano *Smart Citizen*, configurando-se como conjuntos de dados multidimensionais dada as suas dimensões: temporal (instantes da ocorrência dos eventos), variáveis ambientais (medidas físicas coletadas pelos sensores) e espacial (cidades analisadas). Os resultados obtidos revelaram para a abordagem multivariada offline quais os fatores mais influentes nos padrões de *outliers* detectados (com uma AUC de 75%), enquanto que para a abordagem multidimensional offline um modelo de detecção de outliers foi gerado (com uma AUC de 91%), e por fim, para a abordagem multidimensional online variações instantâneas de ocorrência de eventos específicos foram extraídas, identificando com eficiência a dinâmica do processo (com uma AUC de 95%).

Palavras-chave: Detecção de Outliers. Internet das Coisas. Análise Multivariada. Análise Multidimensional. Monitoramento Ambiental. Cidades Inteligentes. AFE. HOSVD.

ABSTRACT

Since 2007, for the first time in human history, more people live in cities than in the countryside. According to projections by the United Nations (UN), the world population is expected to be about 70% urban by 2050. This exponential urban growth brings with it critical and typical problems in cities, such as urban mobility, health, public safety and security. environment pollution. Given that cities are a vast and heterogeneous repository of potential data, the complexity of which is proportional to their size and population, a possible solution to address these problems comes from monitoring events associated with urban data. However, one of the challenges when dealing with this data is to distinguish it as "in the pattern"and "out of the pattern"(outlier) which, in the final analysis, can help or hinder the decision making, for example, of a manager public in a city. This thesis investigates the detection of outliers in smart city applications under three different approaches: offline multivariate approach, where we model the data as matrices, suppressing one of its dimensions, and perform a multivariate analysis using the exploratory factor analysis (EFA) multivariate technique; offline multidimensional approach, where we model the data as a third order tensor, and perform an offline multidimensional analysis using the multidimensional technique Higher-Order Singular Value Decomposition (HOSVD); and online multidimensional approach, where we model the data as a third order tensor and perform an online multidimensional analysis combining the multidimensional HOSVD technique with the sliding window strategy. In order to carry out this research, real data were collected from the Smart Citizen urban environmental monitoring platform, configuring themselves as multidimensional data sets given their dimensions: temporal (moments of occurrence of events), environmental variables (physical measures collected by the sensors) and spatial (analyzed cities). The results obtained revealed for the offline multivariate approach which factors were most influential in the patterns of detected outliers (with an accuracy of 75%), whereas for the multidimensional offline approach an outlier detection model was generated (with an accuracy of 91%), and finally, for the online multidimensional approach, instantaneous variations of the occurrence of specific events were extracted, efficiently identifying the dynamics of the process (with an accuracy of 95%).

Keywords: Outlier Detection. Internet of Things. Multivariate Analysis. Multidimensional Analysis. Environmental Monitoring. Smart Cities. EFA. HOSVD.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

$\mathbf{x}$	Vetor
$\mathbf{X}$	Matriz
$\underline{\mathbf{X}}$	Tensor
$\underline{\mathbf{S}}$	Tensor núcleo
I_i	Dimensão
Λ	Fatores de carregamento
$\mathbf{e}$	Vetor dos termos residuais
$\mathbf{F}$	Matriz dos fatores comuns
h_i	Comunalidade
λ_i	autovalor
$\mathbf{S}$	Matriz de covariância
$\mathbf{R}$	Matriz de correlação
$\mathbf{U}_i$	Matriz fator de carregamento
R	Rank
$\mathbf{U}, \mathbf{V}$	Matrizes ortonormais
S	Conjunto de nós de observação
$\mathbf{V}'$	Vetor real
Ψ_i	Operador transformação
ψ_F	Operador redução de dimensionalidade via AFE
ψ_O	Operador detecção de outliers
ψ_H	Operador redução de dimensionalidade via HOSVD
ψ_C	Operador análise de agrupamento

SUMÁRIO

1 INTRODUÇÃO

Esta tese investiga a aplicação de ferramentas matemáticas da análise multivariada e multidimensional na detecção de *outliers* no contexto do monitoramento ambiental urbano.Ao longo deste capítulo, destacamos alguns aspectos iniciais relacionados ao aumento da população mundial em zonas urbanas, abordando como este fenômeno tem modificado a conjuntura estrutural das grandes cidades e afetado a oferta dos serviços públicos oferecidos aos seus cidadãos, bem como o fenômeno da Tecnologia da Informação e Comunicação (TIC) tem impactado a vida dessas pessoas por meio do monitoramento de variáveis e detecção de eventos sensíveis à saúde e bem estar dos cidadãos. Trataremos sobre a problematização deste estudo que aborda a visão restrita de se levar em conta não apenas a natureza e os atributos dos dados coletados a partir do monitoramento ambiental urbano, mas também por não considerar o aspecto multidimensional dos dados analisados. As principais motivações, assim como os objetivos gerais e específicos serão abordados com o propósito de demonstrar a aplicação de técnicas de análise multivariada e multidimensional em dados coletados a partir do monitoramento ambiental urbano.

1.1 Visão geral e contextualização

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; NATIONS; AFFAIRS, 2015). Aumento na população urbana implica em maior estresse na infraestrutura das cidades e consequente aumento de seus problemas correlatos (por exemplo, mobilidade urbana, segurança, saúde pública, etc.). Ao mesmo tempo que a urbanização cresce em um ritmo acelerado modificando o tecido urbano das grandes cidades juntamente com sua geometria, o fenômeno da Tecnologia da Informação e Comunicação (TIC) também cresce de forma extraordinária, o que viabiliza a integração de múltiplas disciplinas, incluindo redes celulares de quinta geração (5G), redes *ad roc* heterogêneas, redes móveis híbridas, redes de sensores sem fio, dentre outras, que englobam a chamada Internet das Coisas (*Internet of Things*, IoT). A IoT pode ser uma rede de dispositivos físicos, juntamente com coisas como *smartphones*, eletrodomésticos, veículos e outros, que se conectam para uma troca de informações em computadores. A IoT representa uma idéia geral para a flexibilidade dos dispositivos de rede para detectar e coletar informações em todo o mundo e compartilhar essas

informações em uma rede na qual serão processadas, analisadas e utilizadas. Com o crescimento de dispositivos conectados, a CISCO[1] indicou que entre 2015 e 2020 o número de dispositivos *mobile* conectados sofreu um aumento de 2,5 vezes, ou seja, um salto de 4,9 bilhões em 2015 para 12,2 bilhões em 2020, prevendo ainda que até o final de 2020 haverá cerca de 50 bilhões de dispositivos conectados (VEERAMANIKANDAN *et al.*, 2020). Como resultado da evolução dessa infraestrutura tecnológica, surge o paradigma da cidade inteligente, que integra todos os serviços urbanos habilitados pela TIC em um sistema combinado, de modo que a cidade possa ser inteligente e capaz de facilitar o acesso a diferentes tipos de serviços, compartilhamento de informações, monitoramento urbano em tempo real, e assim por diante (BI *et al.*, 2017).

O objetivo de uma cidade inteligente é fornecer aos habitantes uma qualidade de vida promissora, usando as tecnologias avançadas de informação, comunicação e controle para melhorar a eficiência dos serviços e atender às demandas dos habitantes (LIU *et al.*, 2017). Como resultado da combinação dessas tecnologias uma cidade inteligente pode oferecer várias aplicações, incluindo transporte inteligente, energia inteligente, assistência médica inteligente, residências inteligentes e monitoramento ambiental inteligente. Esta cidade conectada não só pode identificar rapidamente as demandas de pessoas em uma cidade, mas também pode manipular operações urbanas para melhorar a qualidade de vida urbana de maneira inteligente e sustentável. Espera-se que o mercado global de cidades inteligentes ultrapasse US 1200 bilhões até 2020, o que é quase o triplo do que em 2014 (NEIROTTI *et al.*, 2014).

Muitos estudos propuseram a implantação de cidades inteligentes baseadas em IoT e nas TIC para um propósito específico, tais como gerenciamento de resíduos, monitoramento de poluição sonora e monitoramento da qualidade do ar (LI *et al.*, 2009; NUORTIO *et al.*, 2006). Em todos esses trabalhos, o monitoramento de áreas urbanas ganha enfoque, o que fortalece o monitoramento ambiental inteligente, que desempenha um papel relevante na identificação de tendências nas mudanças no padrão de comportamento em variáveis ambientais (MILLS, 2007). Além disso, em todo o mundo diferentes iniciativas permitiram que os ambientes urbanos se tornassem mais inteligentes, por exemplo, Amsterdã, São Francisco e Barcelona, (CICIRELLI *et al.*, 2017) e a cidade de Santander (SANCHEZ *et al.*, 2013). Em todos estes casos, as TIC combinadas com IoT, ajudam a melhorar a qualidade de vida dos cidadãos e a eficiência das infraestruturas urbanas, que a partir do monitoramento de variáveis ambientais tais como, qualidade do ar, níveis de ruído, resíduos, iluminação pública, tráfego de veículos, ilhas de calor

[1] http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

e outras aplicações relacionadas a uma visão de cidade inteligente, auxiliam os pesquisadores à entender melhor os novos desafios dos grandes centros urbanos, fornecendo assim informações úteis para a tomada de decisão por parte dos gestores públicos. Portanto, nos últimos anos tem aumentado acentuadamente o grande volume de dados gerados por cidades inteligentes, na perspectiva do monitoramento ambiental urbano, sendo uma fonte de ricas e complexas estruturas de dados.

Em uma cidade inteligente, os dados são gerados de forma explosiva a partir de muitas fontes, desde que um grande número de tecnologias, infraestruturas e processos impulsionadores sejam incorporados à prática cotidiana. Em particular, a Internet móvel generalizada, a computação ubíqua e o software incorporado transformaram as coisas para serem mais inteligentes e criam enormes quantidades de dados (ZHOU *et al.*, 2014). Além disso, alguns aplicativos, como mídia social, fotos e vídeos, transações comerciais, publicidade e jogos, aceleraram a geração de dados. Por exemplo, algumas plataformas famosas de redes sociais (Facebook e Twitter) servem bilhões de visualizações de página todos os meses, armazenam um grande número de novas fotos todos os meses e gerenciam bilhões de conteúdos (BI *et al.*, 2017). Nessa perspectiva, (ZIKOPOULOS; EATON, 2012) em seu trabalho relatou que 2^{50} bytes de dados foram armazenados no mundo em 2000. Já (MANYIKA *et al.*, 2011), estimou que as empresas armazenariam globalmente mais de 2^{60} bytes de novos dados em unidades de disco até 2010, acumulando um crescimento de 40% de dados gerados globalmente por ano. Além disso, apontando para o futuro (ZIKOPOULOS; EATON, 2012) projetou que os volumes de dados esperados devem atingir 2^{70} bytes até 2020. Essa disponibilidade de dados mostra que o paradigma IoT (DEMPSEY *et al.*, 2009), combinado com as TIC (MAYER-SCHONBERGER; CUKIE, 2013), pode contribuir muito para tomadas de decisões cada vez mais informadas (ALAM *et al.*, 2014).

Neste contexto, 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 IoT tem se tornado uma facilitadora-chave para a viabilização das 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,

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

energia, dentre outros (ZHANG *et al.*, 2017). Um desafio significativo é o problema de monitorar, minerar e analisar os dados massivos e heterogêneos.

Na perspectiva das cidades inteligentes, o monitoramento ambiental se destaca, uma vez que a observação do comportamento de variáveis, tais como, gases poluentes, temperatura, ruído e luminosidade, dentre outras, podem fornecer informações vitais para a saúde das pessoas a partir da análise dos padrões desses dados (RATHORE *et al.*, 2016). Com o drástico aumento da urbanização nos últimos anos (ZHANG *et al.*, 2017), o monitoramento ambiental nos grandes centros urbanos desses serviços tem gerado uma grande quantidade de dados, sendo necessária a realização de uma análise eficiente para que esses dados, uma vez coletados, possam fornecer uma melhor compreensão da situação presente para que o futuro da sociedade possa ser planejado. Entretanto, com o aumento do volume de dados, as técnicas tradicionais de processamento e procedimentos analíticos para a análise de dados apresentam desempenho muito limitado (BABAR; ARIF, 2017) (STEED *et al.*, 2013). Esse problema se torna mais crítico à medida que mais dados são coletados e surgem *outliers* (valores discrepantes) (GAO *et al.*, 2020; KARANJIT; SHUCHITA, 2012).

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 (por exemplo, altos níveis de poluição do ar, ruído ambiental, ilhas de calor (IBRAHIM *et al.*, 2016)). Portanto, detectar esses *outliers* é uma tarefa fundamental da mineração e ciência de dados. Entretanto, tais observações podem não ser detectadas com a aplicação de certos métodos analíticos bidimensionais tradicionais, permanecendo invisíveis, como acontece, por exemplo, com a aplicação da 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 uma estrutura 2D capturando apenas as variações bidimensionais, desconsiderando a estrutura 3D natural dos dados.

As aplicações práticas da detecção de *outliers* no contexto do monitoramento ambiental urbano são amplas, tais como, a identificação de padrões de eventos incomuns no fluxo de tráfego urbano, tendências na mudança da qualidade do ar, monitoramento da qualidade da água,

dentre outras (GUARDIOLA *et al.*, 2014; LEE *et al.*, 2014; ENGLE *et al.*, 2014). De maneira geral, podemos ter modelos vetoriais (unidimensional), matriciais (multivariada) ou tensoriais (multidimensional). Os modelos vetoriais são as soluções mais básicas para analisar dados de monitoramento ambiental. Para esse tipo de abordagem, geramos uma série temporal de fluxo para cada variável ambiental e, em seguida, aplicamos um método de série temporal como a Média Móvel Integrada Autoregressiva (*Autoregressive Integrated Moving Average*, ARIMA) (ZAFRA *et al.*, 2017) ou um modelo de regressão (ORDONEZ *et al.*, 2012). A aplicação dessas técnicas é limitada, pois elas não levam em consideração a correlação entre as variáveis ambientais, portanto, não é ideal para lidar com ruídos e valores ausentes nos dados.

Alternativamente, os modelos matriciais têm a capacidade de modelar todos as variáveis ambientais simultaneamente e, portanto, não enfrentam o problema dos modelos vetoriais. Assim, com os métodos matriciais, um arranjo matricial é construído a partir do conjunto completo de variáveis ambientais (CHOI *et al.*, 2012) gerando uma estrutura bidimensional, em que a primeira dimensão pode representar a variação temporal (o instante da coleta do parâmetro ambiental analisado) e, a segunda dimensão pode representar as medidas físicas sensoriadas. O passo seguinte é aplicar uma solução de decomposição matricial, como o PCA (HOTELLING, 1933), por exemplo, na matriz de variáveis ambientais. O PCA é capaz de interpretar dados em termos de um pequeno número de variáveis linearmente independentes (ou componentes), o que consequentemente nos permite identificar *outliers* e padrões irregulares. Entretanto, embora o PCA forneça uma melhor qualidade de modelo em relação aos métodos baseados em vetores, ele sofre de um problema principal. Uma vez que, no método PCA a estrutura tridimensional (multidimensional) natural dos dados é relaxada em uma forma bidimensional (multivariada), muita das vezes, a aplicação do método não é capaz de capturar flutuações espaço-temporais existentes nos dados de monitoramento ambiental.

Os modelos multidimensionais se baseiam nos métodos de decomposição tensorial, que não incluem a maioria das limitações mencionadas acima, tornando-se potenciais ferramentas analíticas para a modelagem de dados no contexto do monitoramento ambiental urbano. A necessidade de tensores na modelagem de dados de monitoramento ambiental surgiu nos últimos anos com aplicações em dados de controle de poluentes do ar (LEE *et al.*, 2014; STANIMIROVA; SIMEONOV, 2005), dados de monitoramento de qualidade de água (SINGH *et al.*, 2007; ENGLE *et al.*, 2014), dados de qualidade do solo (ANDRADE *et al.*, 2007), bem como outros tipos de dados, em duas comunidades de pesquisa, a saber: mineração de dados (FANAEE-T; GAMA,

2015b) e sistemas de detecção de *outliers* de variáveis ambientais (FANAEE-T; GAMA, 2016).

A decomposição tensorial é uma ferramenta poderosa para a análise de dados de várias dimensões com muitas aplicações (MORUP, 2011) em psicometria, quimiometria, neurociência, processamento de sinais, bioinformática, visão computacional e mineração de dados. As soluções tensoriais, ao contrário de métodos, unidimensional e bidimensional, são capazes de modelar flutuações espaço-temporais (FANAEE-T; GAMA, 2015a). Portanto, eles são capazes de gerar um modelo mais natural a partir dos dados coletados e, consequentemente, descobrir padrões mais realistas. Essa grande flexibilidade e qualidade dos modelos tensoriais, aponta para o sucesso de sua aplicação, especialmente no monitoramento ambiental urbano.

Como o objetivo da decomposição tensorial é reproduzir eficientemente as dependências complexas e as interações de ordem superior entre diferentes modos/dimensões nos dados matriciais (multivariados), usando estruturas simples com relativamente poucos parâmetros, seu sucesso atrai cada vez mais atenção no campo da detecção de *outliers* na análise do monitoramento ambiental urbano. Por exemplo, (SOUZA *et al.*, 2019a) propõe um método que explora a natureza multidimensional de dados coletados a partir do monitoramento ambiental urbano, utilizando a decomposição tensorial HOSVD (*Higher-Order Singular Value Decomposition*) em um tensor de três dimensões (tempo vs. variáveis ambientais vs. espaço) combinada com a distância de Mahalanobis para a detecção de *outliers*.

1.2 Questões de pesquisa e Problema Geral

O problema geral abordado nesta tese consiste em como identificar padrões de *outliers* em grandes quantidades de dados heterogêneos de monitoramento ambiental urbano e, em especial, como extrair informação útil a partir destes dados. Para nortear nossa busca de solução, formulamos três questões de pesquisa (QP):

QP #1: É possível detectar e agrupar padrões de *outliers* de dados das variáveis sensoriadas ao longo do monitoramento ambiental urbano inteligente?

QP #2: É possível gerar um modelo que caracteriza as interações multidimensionais de dados monitorados de ambientes urbanos inteligentes para uma melhor detecção de *outliers*, a partir do acesso ao histórico destes dados?

QP #3: É possível extrair variações abruptas instantâneas no comportamento de dados capturados, a partir do monitoramento de espaços urbanos inteligentes, que representam ocorrências de eventos específicos relacionados as variáveis ambientais sensoriadas?

1.3 Hipóteses

A partir do problema geral e das três questões de pesquisa levantadas na seção anterior, derivamos três hipóteses:

Hipótese #1 - associada à **QP #1**: é possível detectar e agrupar *outliers* em dados urbanos, de forma satisfatória, a partir da extração de modelos fatoriais latentes via aplicação da análise multivariada;

Hipótese #2 - associada à **QP #2**: é possível gerar um modelo multidimensional capaz de caracterizar e revelar padrões multidimensionais de *outliers* ocultos às técnicas bidimensionais tradicionais de análise de dados.

Hipótese #3 - associada à **QP #3**: a partir da associação da técnica da janela deslizante ao método multidimensional HOSVD, é possível obter um modelo latente em tempo real que capta o instante em que determinado *outlier* surge, sinalizando a ocorrência de um evento fora do comportamento padrão de normalidade dos dados sensoriados.

1.4 Objetivo principal e objetivos específicos

O objetivo principal desta tese é investigar a aplicação da análise multivariada e multidimensional na detecção de *outliers*, como apoio ao monitoramento ambiental urbano, considerando tanto a natureza multivariada quanto a natureza multidimensional dos dados analisados. Os objetivos específicos desta tese são:

a) Obter uma estrutura fatorial base, a fim de analisar e nomear os fatores, a partir de padrões multivariados identificados e assim, aplicar a estatística de detecção de *outliers* em cada fator em uma abordagem offline de detecção;
b) Obter uma estrutura multidimensional base, a fim de obter uma série temporal estatística e assim, aplicar a estatística de detecção de *outliers* sobre os coeficientes da matriz fator resultante da decomposição multidimensional em uma abordagem offline de detecção;
c) Obter uma estrutura multidimensional base, a fim de obter uma série temporal estatística e assim, aplicar a estatística de detecção de *outliers* sobre os coeficientes da matriz fator resultante da decomposição multidimensional em uma abordagem online de detecção;
d) Identificar o potencial conjunto dessas técnicas de decomposição de dados,

multivariada e multidimensional, na detecção de *outliers* no contexto do monitoramento ambiental urbano.

Deste modo, esta pesquisa relaciona dados de monitoramento ambiental urbano distintos, por meio da análise multivariada e multidimensional de dados, na perspectiva da detecção de *outliers*, levando em conta as relações existentes entre as variáveis envolvidas no problema.

1.5 Contribuições

Nesta tese, investigamos a análise multivariada e multidimensional na detecção de *outliers* a partir de dados coletados de uma plataforma de monitoramento ambiental urbano denominada Smart Citizen (CITIZEN, 2016). A nossa proposta de detecção de *outliers* explora, em três abordagens, as dependências complexas e as interações de ordem superior entre as três dimensões dos dados coletados (tempo, variáveis ambientais e espaço). Desta forma, destacamos a seguir, as três contribuições principais desta tese:

Contribuição #1 - Uma nova abordagem de detecção de outliers para dados de monitoramento de ambientes urbanos inteligentes baseada na análise multivariada, via aplicação da Análise Fatorial Exploratória (SOUZA *et al.*, 2018) (**CP #1**).

Contribuição #2 - Uma nova abordagem de detecção de outliers em dados de espaços urbanos inteligentes baseada na análise multidimensional, via aplicação do método multidimensional de decomposição de dados HOSVD (SOUZA *et al.*, 2017; SOUZA *et al.*, 2019a) (**CP #2**).

Contribuição #3 - Uma nova abordagem de detecção de outliers para dados coletados a partir do monitoramento ambiental urbano inteligente baseada na análise multidimensional, via aplicação do método multidimensional de decomposição de dados HOSVD associado à estratégia da janela deslizante (SOUZA *et al.*, 2019b) (**CP #3**).

Para a primeira contribuição, modelamos os dados como matrizes, suprimindo a dimensão modo espacial, e realizamos uma análise multivariada *offline* considerando a dimensão modo temporal e a dimensão modo variáveis ambientais sensoriadas, para a detecção de *outliers*. Para a segunda contribuição, modelamos os dados como um tensor de terceira ordem, e realizamos uma análise multidimensional *offline* na detecção de *outliers*, considerando todas as três dimensões, a saber, dimensões: modo temporal, modo variáveis ambientais e modo espacial. Por fim, para a terceira contribuição, modelamos os dados como um tensor de terceira ordem

(considerando as dimensões: modo temporal, modo variáveis ambientais e modo espacial), e realizamos uma análise multidimensional *online* na detecção de *outliers*.

1.6 Estrutura da Tese

As contribuições diversas desta tese estão organizadas e apresentadas em cinco capítulos, além da introdução. No **Capítulo 2**, serão apresentados os principais conceitos dos fundamentos analíticos matemáticos utilizados nesta pesquisa, bem como uma revisão na literatura apresentando os principais aspectos que caracterizam a detecção de *outliers* e seu delineamento no monitoramento ambiental. No **Capítulo 3**, será apresentada a metodologia da pesquisa e a caracterização de todo o contexto desta pesquisa, tratando sobre alguns aspectos acerca do monitoramento ambiental urbano, bem como caracterizando os dados coletados e a aplicação das técnicas de decomposição e de detecção de *outliers*. No **Capítulo 4**, será apresentada a análise e discussão dos resultados obtidos pela aplicação das abordagens propostas. Por fim, no **Capítulo 5**, apresentamos as conclusões desta pesquisa e discutimos algumas implicações para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Neste capítulo serão abordados os principais tópicos que embasam os fundamentos teóricos desta tese de doutorado. Na primeira seção, serão apresentadas as duas principais técnicas de análise de dados que serão utilizadas nesta tese, a saber: a análise multivariada via análise fatorial exploratória; e a análise multidimensional via decomposição HOSVD. Na segunda seção, serão apresentados os trabalhos relacionados sobre a detecção de outliers fornecendo uma visão geral estruturada e abrangente, apontando as similaridades e diferenças com a pesquisa desenvolvida nesta tese.

2.1 Análise Multivariada e Multidimensional

2.1.1 Definições e notações

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; BADER, 2009). De acordo com a variação da ordenação dos dados multidimensionais, Kolda estabeleceu a seguinte divisão, em termos de notação:

- Tensor de ordem zero (Figura 1a): $x \in \Re$, representando um escalar;
- Tensor de primeira ordem (Figura 1b):

  $\mathbf{x} = \{x_{1,1}, x_{1,2}, \ldots, x_{1,i_n}\}$,

  com $\mathbf{x} \in \Re^{I_1}$ representando um vetor, onde $i_n \in 1,2,...,I_n$, $1 \leq n \leq N$;

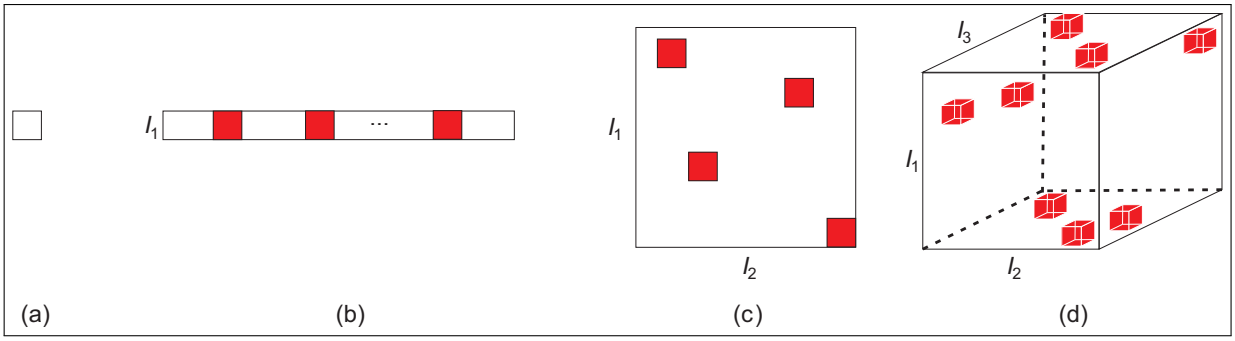
- Tensor de segunda ordem (Figura 1c):

$$\mathbf{X} = \{(x_{1,1}, x_{1,2}, \ldots, x_{1,i_n}),$$
$$(x_{2,1}, x_{2,2}, \ldots, x_{2,i_n}),$$
$$\vdots$$
$$(x_{i_n,1}, x_{i_n,2}, \ldots, x_{i_n,i_n})\}$$

  com $\mathbf{X} \in \Re^{I_1 \times I_2}$ representando uma matriz, onde $i_n \in 1, 2, ..., I_n$, $1 \leq n \leq N$;

- Tensor de terceira ordem (Figura 1d): $\underline{\mathbf{X}}^{I_1 \times I_2 \times I_3}$, representando um tensor de ordem superior. Para um tensor de terceira ordem (Figura 1d), um elemento individual deste tensor é denotado por x_{I_1,I_2,I_3}, seguindo a notação anterior. Os elementos destacados em vermelho ilustram os possíveis *outliers* do conjunto de dados.

Figura 1 – Tensor de ordem: (a) zero; (b) um; (c) dois; (d) três.



Fonte: elaborado pelo autor (2020).

Enquanto um tensor de segunda ordem é geralmente usado para representar problemas bidimensionais, explorados por técnicas multivariadas, um tensor de terceira ordem representa problemas tridimensionais comumente explorados por técnicas de decomposição tensorial ou multidimensionais.

2.1.2 Análise Multivariada de Dados

Análise Fatorial Exploratória (AFE), no que diz respeito à análise exploratória dos dados, busca encontrar padrões de informações intrínsecas ao conjunto de dados, oferecendo até mesmo suporte para a detecção e diagnóstico de *outliers* (LI *et al.*, 2017). A AFE é um dos métodos da estatística multivariada que tem como objetivo principal a redução de dimensionalidade e identificar as relações subjacentes entre as variáveis medidas, determinando o número e a natureza apropriada dos fatores comuns (fatores latentes) necessários para explicar a

matriz de correlação observada (BARTHOLOMEW; KNOTT, 1999). Assim, podemos distinguir ou classificar as variáveis de acordo com as contribuições dos fatores latentes para cada variável individualmente. O modelo expressa um vetor $\mathbf{x}$ m-dimensional (variáveis ambientais) como

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{e}, \tag{2.1}$$

em que Λ $(m \times k)$ é a matriz de coeficientes ou fatores de carregamento e $\mathbf{f}$ é o vetor k-dimensional dos fatores comuns (fatores latentes), com $k \leq m$, e $\mathbf{e}$ é um vetor m-dimensional dos termos residuais do modelo (fatores específicos).

Para simplificar ainda mais nosso modelo original, adotamos um modelo de fatores ortogonais com base em três pressupostos (BASILEVSKY, 2009): (i) a média do vetor dos fatores comuns $\mathbf{f}$ é zero e a matriz de covariância é a identidade; (ii) a média dos termos do vetor de erro $\mathbf{e}$ é zero e a matriz de covariância é diagonal; e (iii) os termos do vetor de erro não tem correlação com os fatores comuns.

Para extração dos fatores latentes baseados no modelo da análise fatorial, conforme detalhado em (SUNDBERG; FELDMANN, 2016) consideremos primeiro, a covariância entre $\mathbf{X}$ (matriz dos vetores m-dimensionais $\mathbf{x}$) e $\mathbf{F}$ como

$$Cov(\mathbf{X}, \mathbf{F}) = \Lambda, \tag{2.2}$$

se as variáveis observadas forem padronizadas antes da aplicação da análise fatorial, como realizado nesta pesquisa, então a equação pode ser reescrita como

$$Cor(\mathbf{X}, \mathbf{F}) = \Lambda, \tag{2.3}$$

onde Cor é a abreviação de correlação. A equação sugere que os coeficientes de carga de fatores representam a correlação entre as variáveis ambientais e os fatores comuns, e a soma dos quadrados dos coeficientes de carregamento para qualquer fator indica o grau de variabilidade explicada por ele.

Após a extração dos fatores latentes baseados no modelo da análise fatorial, um índice importante a ser obtido é a comunalidade. A comunalidade para qualquer variável pode ser interpretada como a proporção da variabilidade dessa variável explicada por todos os fatores comuns (BASILEVSKY, 2009). Assim, a comunalidade h_i para a i-ésima variável é definida como

$$h_i = \sum_{j=1}^{k} \lambda_{ij}^2, \tag{2.4}$$

na qual λ_{ij} é o coeficiente do fator de carregamento da *i*-ésima variável no *j*-ésimo fator comum.

Além das comunalidades, calculamos os escores dos fatores que são definidos como combinações lineares das variáveis observadas. A estimativa dos escores dos fatores com base na regressão é dada diretamente aqui sem derivação (BARTHOLOMEW; KNOTT, 1999)

$$\mathbf{S} = \Lambda \mathbf{R}^{-1} \mathbf{X}, \tag{2.5}$$

em que **S** denota os escores dos fatores e **R** é a matriz de correlação das variáveis observadas, representadas pela matriz **X**.

2.1.3 Análise Multidimensional de Dados

Um tensor é um objeto geométrico usado em matemática e física para extensão de conceitos como escalares, vetores e matrizes de dimensões superiores. A origem da palavra tensor é o latim tenere "alongar", que apareceu pela primeira vez na anatomia no século XVII, para denotar o alongamento muscular. Mais tarde, foi usado em meados do século XVIII por William Hamilton para descrever alguns conceitos na álgebra de quaternion. O cálculo do tensor, que se aproxima do significado atual da palavra, foi introduzido em 1900 pelo matemático italiano Gregorio Ricci-Curbastro. Em 1915, Albert Einstein usou o tensor na teoria da relatividade geral para explicar a estrutura geométrica e causal do espaço-tempo e definir conceitos como distância, volume, curvatura, ângulo, futuro e passado.

Os primeiros princípios das decomposições tensoriais foram fundados pelo matemático americano Frank Hitchcock em 1927 (SMILDE *et al.*, 2004) e, posteriormente, psicólogos como Raymond Cattell (CARROL; CHANG, 1970), Ledyard Tucker (TUCKER, 1966) e Richard Harshman (HARSHMAN, 1970) foram pioneiros em estender aplicações de decomposições tensoriais na análise de dados, particularmente, em psicologia entre as décadas de 1940 e 1970. Em 1981, a decomposição tensorial foi introduzida por Appellof e Davidson à comunidade de quimiometria. Somente uma década mais tarde surgiram as primeiras aplicações tensoriais na detecção de *outliers* nessa comunidade. O trabalho de (NOMIKOS; MACGREGOR, 1994) sobre o monitoramento de cargas por várias dimensões foi pioneiro na motivação de métodos tensoriais nos problemas de monitoramento e detecção de *outliers*.

A aplicação moderna de tensores na detecção de *outliers* apareceu há uma década em uma série de artigos de Jim Sun e colegas (SUN *et al.*, 2006a; SUN *et al.*, 2006b), que tiveram uma grande contribuição para o crescimento da pesquisa de detecção de *outliers* baseada em

tensores (DOT). Atualmente, a aplicação da DOT tem sido difundida em áreas mais amplas, incluindo monitoramento ambiental (LEE *et al.*, 2014), vigilância por vídeo (TRAN *et al.*, 2012), segurança de rede (SUN *et al.*, 2006b), redes sociais (KOUTRA *et al.*, 2012), sistemas baseados em texto (PANISSON *et al.*, 2014), neurociência (ACAR *et al.*, 2007), sensoriamento remoto (RENARD; BOURENNANE, 2008), sensores (SUN *et al.*, 2006a) e outros domínios (KOSANOVICH *et al.*, 1994; MU *et al.*, 2011; BAI *et al.*, 2013; WANG *et al.*, 2020; SONG *et al.*, 2020).

2.1.4 Operações e conceitos elementares

A seguir, são apresentadas as principais operações básicas utilizadas para a realização da fatoração na análise multidimensional dos dados desta tese, a saber, os produtos: externo, Kronecker, Hadamard e Khatri-Rao (KOLDA; BADER, 2009; SMILDE *et al.*, 2004; CICHOCKI *et al.*, 2009).

O produto externo ($\circ$) entre i_N vetores $\mathbf{x}$, gera um tensor $\underline{\mathbf{X}} \in \Re^{I_1 \times I_2 \times I_3 \times \ldots \times I_n}$,

$$\underline{\mathbf{X}} = \mathbf{x}_{I_1} \circ \mathbf{x}_{I_2} \circ \cdots \circ \mathbf{x}_{I_N} = \prod_{i_n=1}^{I_N} \circ \mathbf{x}_{i_n} \tag{2.6}$$

O produto de Kronecker entre dois tensores, ambos de ordem 2, ou seja, duas matrizes, $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ e $\mathbf{B} \in \mathbb{R}^{J_1 \times J_2}$, é definido como

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1I_2}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2I_2}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I_1 1}\mathbf{B} & a_{I_1 2}\mathbf{B} & \cdots & a_{I_1 I_2}\mathbf{B} \end{bmatrix} \tag{2.7}$$

O produto de Khatri-Rao entre dois tensores, ambos de ordem 2, ou seja, duas matrizes, $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ e $\mathbf{B} \in \mathbb{R}^{I_1 \times I_2}$, é definido da seguinte forma:

$$\mathbf{A} \diamond \mathbf{B} = (\mathbf{A}_1 \otimes \mathbf{B}_1 \ldots \mathbf{A}_k \otimes \mathbf{B}_k) \tag{2.8}$$

em que ambos os tensores de segunda ordem são arranjos matriciais com igual número de partições.

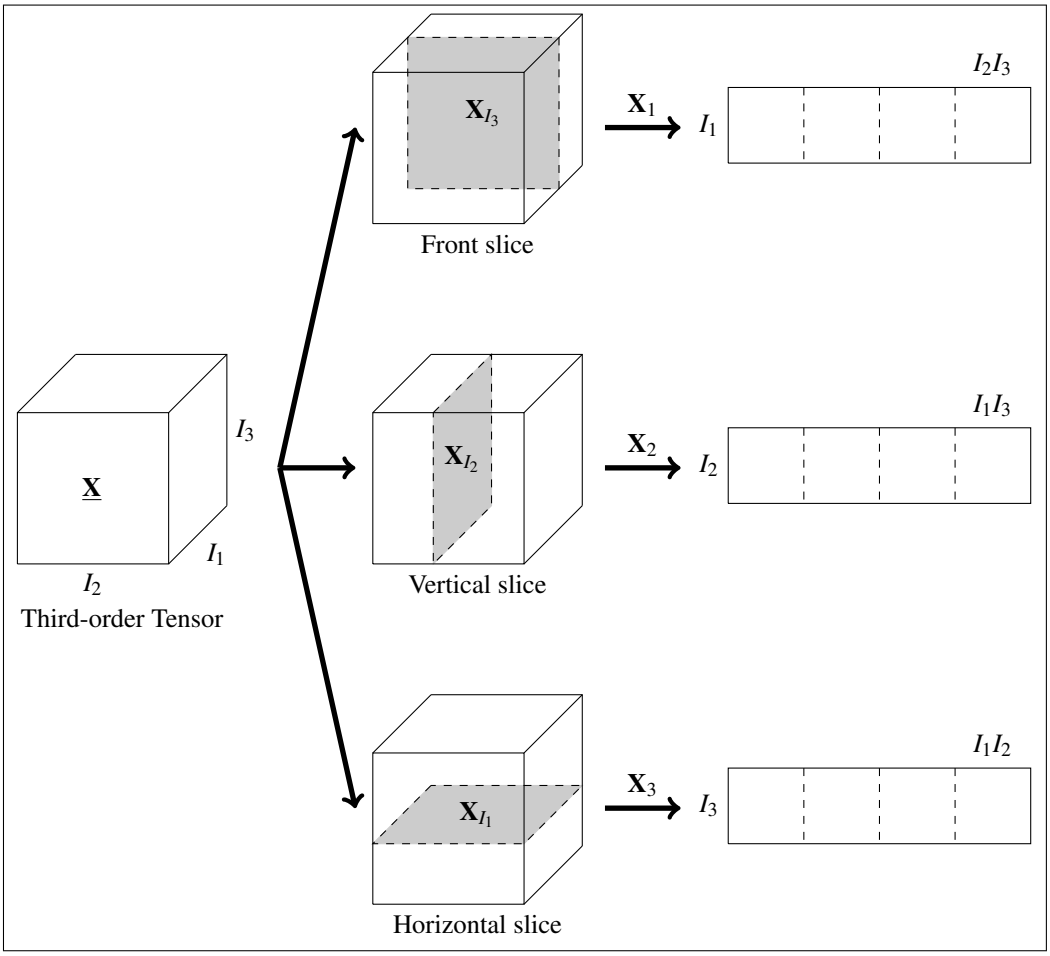
Além das operações entre tensores descritas anteriormente, podemos também representar os tensores através de diferentes maneiras. Em uma representação tensorial, os dados multivariados (representados por um tensor de segunda ordem) são organizados em arranjos

multidimensionais de ordem superior. Contudo, em diversas aplicações é conveniente que os tensores sejam representados por matrizes, nesta perspectiva, a organização de um tensor pode ser alterada transformando-o em matrizes. Este processo é chamado de matriciação, ou seja, é o processo de reordenação de um tensor em matrizes. Existem várias maneiras de realizar este ordenamento e, portanto, o processo de matriciação não é único. Assim, fixando um dos três índices de um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, são formadas as chamadas fatias ou fatias planas, que são secções bidimensionais do tensor:

- Fatia frontal $\mathbf{X}_{::I_3}$, ou apenas $\mathbf{X}_{I_3}$, representando a I_3-ésima fatia frontal;
- Fatia vertical $\mathbf{X}_{:I_2:}$, ou apenas $\mathbf{X}_{I_2}$, representando a I_2-ésima fatia vertical;
- Fatia horizontal $\mathbf{X}_{I_1::}$, ou apenas $\mathbf{X}_{I_1}$, representando a I_1-ésima fatia horizontal.

Portanto, uma matriz padrão que é matriciada no n-ésimo modo de um tensor $\underline{\mathbf{X}} \in \Re^{I_1 \times I_2 \times I_3 \times \ldots \times I_n}$ é representada por $\mathbf{X}_n$. Para um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, os procedimentos de matriciação são apresentados na Figura 2. Além disso, a operação chamada de desmatriciação corresponde ao processo inverso à matriciação.

Figura 2 – Matriciação tensorial.



Fonte: elaborado pelo autor (2020).

Outra operação importante é a multiplicação de um tensor por uma matriz em um modo/dimensão específica (KROONENBERG, 2008). Portanto, o n-ésimo modo de multiplicação de um tensor $\underline{\mathbf{X}}$ por uma matriz $\mathbf{U}$ é representada da seguinte forma $\underline{\mathbf{X}} \times_n \mathbf{U}$. Para facilitar a compreensão, o modo de multiplicação n é uma multiplicação da forma matricial do tensor no modo especificado pré-multiplicado pela matriz (KOLDA; BADER, 2009).

As operações e conceitos abordados anteriormente compreendem os fundamentos analíticos para a realização tanto da organização dos dados, quanto da aplicação da fatoração tensorial utilizada nesta pesquisa.

2.1.5 Fatoração tensorial

As técnicas tradicionais de análise de dados, como análise exploratória de dados, classificação, agrupamento, regressão etc, apenas modelam dados bidimensionais, isto é, desconsideram a natureza multidimensional dos dados e não consideram a interação entre mais de duas dimensões. No entanto, em vários fenômenos do mundo real, existe um relacionamento mútuo entre mais de duas dimensões, em particular quando a dimensão temporal é adicionada ao problema.

Neste contexto, a fatoração tensorial, também conhecida como decomposição tensorial, surge como um processo que considera todas as dependências mútuas entre as distintas dimensões e fornece uma representação compacta dos dados originais em espaços dimensionais inferiores (RIPOLL; PAJAROLA, 2015). O processo inverso é chamado de reconstrução tensorial. As técnicas de fatoração de um tensor são métodos de análise de dados capazes de resumir as informações contidas em tensores de ordem superior referentes à interação entre seus modos usando um número reduzido de componentes. Além disso, esses métodos geram o chamado tensor núcleo (ou arranjo central), que é uma estrutura de informações que captura e descreve os relacionamentos e interações contidos nos diferentes modos de dados, representando uma rica fonte de informações à ser analisada (KIERS, 2000).

Atualmente, as técnicas de fatoração tensorial são aplicadas em vários domínios, como controle de processo (CHEN; YEN, 2003), neurociências (CHEN *et al.*, 2015), sensoriamento remoto (RENARD; BOURENNANE, 2008) e aplicações médicas (HO *et al.*, 2014). As técnicas mais comuns de fatoração tensorial são Tucker (TUCKER, 1966) e CP/PARAFAC (HARSHMAN, 1970), que são versões generalizadas do PCA ou, mais especificamente, da Decomposição de Valor Singular (SVD) para tensores de ordem superior. Portanto, a técnica

de fatoração tensorial pertence a duas categorias principais de métodos (não restritos apenas a estes): PARAFAC e Tucker. A Tabela 1 abaixo apresenta as principais técnicas que abrangem as duas categorias.

Tabela 1 – Principais técnicas de fatoração tensorial

Categorias	Métodos
PARAFAC-based	PARAFAC (HARSHMAN, 1970)
	Non-negative PARAFAC (CARROLL *et al.*, 1989)
	PARAFAC2 (KIERS, 1993)
Tucker-based	Tucker1 (KROONENBERG, 2008)
	Tucker2 (KOLDA; BADER, 2009)
	Tucker3 (TUCKER, 1966)
	HOSVD (LATHAUWER *et al.*, 2000)

Fonte: elaborado pelo autor (2020).

Particularmente, a decomposição de Tucker aproxima um tensor de ordem superior através de um produto de um tensor de ordem inferior (chamado de tensor núcleo), com dimensões predeterminadas, multiplicado por matrizes fatores em cada dimensão. Formalmente, o problema pode ser definido como um problema de otimização (CHEN *et al.*, 2014): dado um tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$, encontramos um tensor do núcleo $\underline{\mathbf{S}} \in \mathbb{R}^{R_1 \times R_2 \times \ldots \times R_n}$ com números inteiros predefinidos R_i com $1 \leq R_i \leq I_i$ para $i = 1, 2, \ldots, n$, e matrizes fatores $\mathbf{U}_n \in \mathbb{R}^{I_n \times R_n}$ que otimizam a seguinte função

$$min||\underline{\mathbf{X}} - \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_n \mathbf{U}_n||, \tag{2.9}$$

No modelo acima, n representa a dimensão do tensor (por exemplo, para um tensor tridimensional, $n = 3$) e $R_1, R_2, \ldots, R_n$ ($i = 1, 2, \ldots, n$) são parâmetros de entrada do modelo (tamanho do núcleo). O modelo mais utilizado para encontrar as matrizes fatores $\mathbf{U}_n$ e o tensor núcleo $\underline{\mathbf{S}}$ é uma fatoração chamada HOSVD, em que o primeiro tensor é matriciado em matrizes de ordem inferior em todos os seus modos. Por exemplo, podemos matriciar o tensor com dimensões $I_1 \times I_2 \times I_3$ em matrizes $I_1 \times I_2 I_3$ ou $I_2 \times I_1 I_3$ ou $I_3 \times I_1 I_2$ e, em seguida, SVD é executado independentemente em cada matriz.

Assim, o teorema (apresentado no Apêndice B) nos diz que toda matriz $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2}$ pode ser decomposta no produto:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{i=1}^{R} s_{ii} \mathbf{u_i} \mathbf{u_i}^T \tag{2.10}$$

em que $\mathbf{U} \in \mathbb{R}^{I_1 \times I_1}$ e $\mathbf{V} \in \mathbb{R}^{I_2 \times I_2}$ são duas matrizes ortonormais e $\mathbf{S} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{I_1 \times I_2}$ é a matriz pseudo-diagonal com $\Sigma = diag(\sigma_1, \ldots, \sigma_R)$ e $R = rank(X)$. As entradas diagonais de $\mathbf{S}$ são valores singulares de $\mathbf{X}$, que são não-negativos e ordenados como $\sigma \geq, \ldots, \geq \sigma_R$. As colunas de $\mathbf{U}$ e $\mathbf{V}$ são os vetores singulares à esquerda e direita de $\mathbf{X}$, respectivamente.

Desta forma, podemos obter a decomposição HOSVD utilizando a notação de produto externo, como

$$\mathbf{X} = \mathbf{S} \times_1 \mathbf{U} \times_2 \mathbf{V} = \sum_{i=1}^{R} s_{ii} \mathbf{u_i} \circ \mathbf{u_i} \tag{2.11}$$

significando que a matriz $\mathbf{X}$ é decomposta em uma matriz núcleo diagonal $\mathbf{S}$ multiplicado por duas matrizes ortonormais $\mathbf{U}$ e $\mathbf{V}$ nos modos 1 (I_1) e 2 (I_2), respectivamente. A Figura 2 ilustra esses modos.

Analogamente, para um tensor de ordem superior $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$, a fatoração multivariada SVD aplicada a cada um dos modos deste tensor, é chamada de HOSVD (BERGQVIST; LARSSON, 2010). Assim, o HOSVD decompõe um tensor $\underline{\mathbf{X}}_{I_1 \times I_2 \times I_3 \times \ldots \times I_n}$ como

$$\underline{\mathbf{X}} = \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_n \mathbf{U}_n, \tag{2.12}$$

onde todas as matrizes $\mathbf{U}_n \in \mathbb{R}^{I_n \times I_n}$ (matrizes singulares modo n) são matrizes ortonormais e $\underline{\mathbf{S}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$ é um tensor núcleo que satisfaz todas as condições de ortogonalidade (BERGQVIST; LARSSON, 2010). Neste sentido, para um tensor de terceira ordem, a fatoraçao HOSVD é escrita como

$$\underline{\mathbf{X}} = \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \tag{2.13}$$

onde as matrizes $\mathbf{U}_1$, $\mathbf{U}_2$ e $\mathbf{U}_3$, são matrizes ortonormais e $\underline{\mathbf{S}}$ é o tensor núcleo da fatoração. Além disso, novamente $\times_1$ é um produto no modo-1, $\times_2$ é um produto no modo-2 e $\times_3$ é um produto no modo-3.

2.1.6 Seleção de componentes

A qualidade do modelo tensorial tem uma relação direta com a seleção de componentes do modelo (FANAEE-T; GAMA, 2016; CONG *et al.*, 2012). A ideia geral da seleção de componentes é expressar ou comprimir tensores com menos parâmetros, o que pode levar a uma enorme redução nos requisitos de memória computacional e a uma possível aceleração

computacional significativa (VONDREJC *et al.*, 2020). Neste sentido, na literatura algumas abordagens foram desenvolvidas para estimar o número ideal de componentes.

Neste sentido, a abordagem utilizada nesta tese para a seleção de componentes do método de decomposição multidimensional é o método da variância explicada. Nesse método de seleção de componentes, o número de componentes principais é escolhido com base na porcentagem acumulada de autovalores ou na porcentagem acumulada da variância explicada (FANAEE-T; GAMA, 2016). Portanto, caso a porcentagem acumulada dos primeiros componentes principais estiverem acima de um determinado limiar, selecionamos como o número adequado de componentes. Diversos trabalhos na literatura, como encontrados em (MORI; YU, 2014; URTUBIA *et al.*, 2012), utilizam o critério da variância explicada na seleção de ordem do modelo na detecção de *outliers* em fatorações tensoriais.

Particularmente, na fatoração HOSVD o percentual de variância explicada é calculada sobre cada matriz fatorial extraída da fatoração. Deste modo, podemos controlar seletivamente como cada modo especificamente explica o espaço original dos dados e assim eliminar o ruído de cada modo separadamente. Além disso, as dimensões também podem ser reduzidas, removendo os últimos vetores coluna do tensor ao longo do modo desejado.

2.2 Detecção de *outliers*

A detecção de outliers é um problema importante que é atualmente pesquisado em diversas áreas de pesquisa e domínios de aplicação. Muitas técnicas de detecção de outliers foram desenvolvidas especificamente para determinados domínios de aplicação, enquanto outras são mais genéricas.

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 muitas vezes indicar eventos de interesse, como por exemplo, no monitoramento ambiental urbano, representar altos níveis de poluição do ar, ruído ambiental, ilhas de calor (IBRAHIM *et al.*, 2016), bem como também, fraude no cartão de crédito, intrusão cibernética, atividade terrorista ou quebra do sistema (CHANDOLA *et al.*, 2009).

Originalmente, a detecção de outliers está relacionada com a remoção de ruído (TENG H., 1990) e acomodação de ruído (ROUSSEEUW; LEROY, 1987). O ruído pode ser definido como um fenômeno nos dados que não interessa ao analista, mas atua como um

obstáculo à análise dos dados. A remoção do ruído é motivada pela necessidade de remover os objetos indesejados antes que qualquer análise de dados seja realizada nos dados. A acomodação do ruído refere-se à imunização, ou seja, uma estimativa do modelo estatístico contra observações fora do padrão.

2.2.1 Natureza dos dados de entrada

Um dos principais aspectos de qualquer técnica de detecção de outliers é a natureza dos dados de entrada. 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; GAMA, 2016).

Nesta perspectiva, a natureza dos atributos extraídos determina a aplicabilidade das técnicas de detecção de outliers. Por exemplo, para técnicas estatísticas, diferentes modelos estatísticos devem ser usados para dados contínuos e categóricos. Da mesma forma, para técnicas baseadas em vizinhos mais próximos, a natureza dos atributos determina qual a medida de distância pode ser a melhor para ser utilizada. Nesses casos, em vez dos dados reais, a distância entre as instâncias em pares pode ser fornecida na forma de uma matriz de distância (ou similaridade).

Em geral, as instâncias de dados podem estar relacionadas entre si. Alguns exemplos são dados de sequência, dados espaciais e dados gráficos. Nos dados de sequência, as instâncias de dados são ordenadas linearmente, por exemplo, dados de séries temporais, sequências de genoma, sequências de proteínas. Dados sequenciais foram utilizados no trabalho de [], em que dados do tipo séries temporais de concentrações de poluentes atmosféricos foram considerados na detecção de *outliers*. Nesse trabalho, um modelo linear generalizado prevê as observações das variáveis analisadas com base em medições conhecidas das variáveis sensoriadas e então, são Nos dados espaciais, cada instância de dados está relacionada às instâncias vizinhas, por exemplo, dados de tráfego de veículos. Quando os dados espaciais têm um componente temporal (sequencial), são referidos como dados espaço-temporais, por exemplo, dados climáticos, dados de monitoramento ambiental, etc. Nos dados gráficos, as instâncias de dados são representadas como vértices em um gráfico e conectadas a outros vértices com arestas.

2.2.2 Tipos de outliers

Com relação aos tipos outliers, (AHMED *et al.*, 2016) os categorizou em três tipos distintos, a saber:

- outliers pontuais - quando um único elemento difere da normalidade dos dados. Como exemplo real, considere a detecção de fraudes no cartão de crédito. Considere um conjunto de dados que corresponde às transações com cartão de crédito de uma pessoa, supondo que os dados sejam definidos usando apenas um recurso: valor gasto. Uma transação cuja quantia gasta é muito alta em comparação com o intervalo normal de despesas para essa pessoa será um outlier pontual;
- outliers contextuais - quando um evento isolado e atípico ocorre. Como exemplo real, conjuntos de dados espaciais, por exemplo, longitude e latitude de um local são atributos contextuais. Em dados de séries temporais, o tempo é um atributo contextual que determina a posição de uma instância em toda a sequência;
- outliers coletivos - quando uma série de eventos atípicos ocorrem. Por exemplo, considere a saída de um eletrocardiograma humano, onde um padrão atípico pode ser observado durante um determinado período de tempo. Este valor, por si só, não seria um valor atípico, mas em conjunto representa outliers coletivos.

Além disso, segundo (CHANDOLA *et al.*, 2009) embora outliers pontuais possam ocorrer em qualquer conjunto de dados, os coletivos podem ocorrer apenas em conjuntos de dados nos quais as instâncias de dados estão relacionadas. Por outro lado, a ocorrência de outliers contextuais depende da disponibilidade de atributos de contexto nos dados. Nesta perspectiva, um outlier pontual ou coletivo também pode ser um outlier contextual, se analisado em relação à um contexto. Assim, um problema de detecção de outliers pontuais ou coletivo pode ser transformado em um problema de detecção de outliers contextuais incorporando as informações de contexto.

2.2.3 Rótulos de dados

Os rótulos associados a uma instância de dados indicam se essa instância é normal ou não. Com base na extensão em que os rótulos estão disponíveis, as técnicas de detecção de outliers podem operar em um dos três modos a seguir:

- detecção de outliers supervisionada - na abordagem supervisionada, rotulamos

a classe regular e os outliers, uma vez que o objetivo é construir um modelo preditivo para a classe regular contra a classe dos outliers e, portanto, qualquer instância de dados não observada é comparada ao modelo para determinar à qual classe pertence;

- detecção de outliers semi-supervisionada - na abordagem semi-supervisionado, pressupomos que as instâncias de rótulos de dados de treinamento sejam apenas para a classe regular, sem exigir rótulos para a classe dos outliers;
- detecção de outliers não-supervisionada - na abordagem não-supervisionada, não há necessidade de dados de treinamento e, portanto, é amplamente utilizada. Essa abordagem assume que os dados de entrada seguem um modelo estocástico. Assim, um teste de inferência estatística decide se a instância de dados pertence ao modelo, no nosso caso, regular ou não.

2.2.4 Saídas da detecção de outliers

Outro aspecto relevante que deve ser considerado na aplicação de qualquer técnica de detecção de outliers é a maneira como os outliers são relatados. Normalmente, as saídas produzidas pelas técnicas de detecção de outliers são um dos dois tipos a seguir (CHANDOLA *et al.*, 2009):

- Scores: as técnicas de detecção de outliers que geram scores, atribuem um score de anomalia a cada instância nos dados de teste, e dependendo do grau desta pontuação, essa instância pode ser considerada um outlier. Portanto, o resultado dessas técnicas é uma lista classificada de outliers. Um analista pode optar por analisar o pequeno número com os principais outliers ou usar um limite de corte para selecioná-los;
- Labels: as técnicas de detecção de outliers baseada em labels, atribuem um label (normal ou outlier) a cada instância de teste.

É importante destacar que as técnicas de detecção de *outliers* baseadas em scores permitem ao analista usar um limite específico do domínio de aplicação para selecionar os outliers mais representativos à pesquisa. Por outro lado, as técnicas que fornecem rótulos binários para as instâncias de teste não permitem diretamente que os analistas façam essa escolha, embora isso possa ser controlado indiretamente por meio de escolhas de parâmetros específicos em cada técnica.

2.2.5 Classificação das técnicas de detecção de outliers

Nos últimos anos, vários métodos foram introduzidos na literatura para a detecção de outliers em instâncias formadas por atributos multivariados e multidimensionais. Esses métodos podem ser agrupados em cinco categorias principais (CHANDOLA *et al.*, 2009):

- Métodos baseados em classificação - pressupõem que os rótulos de classe possam operar em um dos três modos discutidos acima, a saber: supervisionado, semi-supervisionado e não supervisionado;
- Métodos baseados em *cluster* - geram conjuntos de dados semelhantes onde instâncias de dados normais estão próximas ao centro de um determinado grupo e o mais externo é o elemento que não pertence a nenhum grupo, ou ocorre em um pequeno grupo com baixa densidade;
- Métodos baseados na teoria da informação - pressupõem que os outliers possam ser detectados, uma vez que resultam em irregularidades no conteúdo das informações do conjunto de dados;
- Métodos espectrais - mapeiam o espaço de dados para um subespaço menor, supondo que exista um subespaço cujos dados e outliers possam ser facilmente identificados;
- Métodos baseados em estatística - pressupõem que os dados de entrada seguem um modelo estocástico e, portanto, para cada instância, um teste de inferência estatística é aplicado para decidir se a instância de dados pertence ao modelo, portanto, normal ou não. Essas técnicas podem ser classificadas como: paramétrica e não-paramétrica. As técnicas paramétricas, geralmente assumem o conhecimento da distribuição subjacente do conjunto de dados. Por outro lado, as técnicas não-paramétricas não têm nenhuma premissa predefinida sobre a distribuição do conjunto de dados.

Embora as abordagens paramétricas sejam ideais quando a distribuição dos dados é conhecida e garantida, os métodos estatísticos de detecção que não fazem suposições a priori são conhecidos por serem mais precisos quando as distribuições de dados variam, são desconhecidos ou não se ajustam a nenhuma distribuição particular. Portanto, uma vez que as abordagens não paramétricas evitam erros potencialmente significativos devido à incompatibilidade de suposições de distribuição, nesta tese, fundamentamos nossa abordagem de detecção de *outliers* em técnicas baseadas em estatística não-paramétrica. Isso os torna mais flexíveis em geral no contexto da

aplicação da tarefa de detecção de *outliers* no monitoramento ambiental urbano. Por exemplo, (SOUZA *et al.*, 2019a) utiliza métodos estatísticos não-paramétricos para a detecção de padrões de *outliers* no monitoramento de espaços urbanos.

2.3 Detecção de outliers no monitoramento ambiental urbano

Com o crescente aspecto de *big data* nas cidades inteligentes, a detecção de outliers no contexto do monitoramento de dados ambientais tem sido uma área ativa de pesquisa, mas ainda são necessárias mais pesquisas devido à crescente demanda por objetos "inteligentes", bem como o aumento do fluxo de dados de variáveis sensoriadas. Nesta perspectiva, com o grande volume de informações coletadas, extrair informações úteis é de fundamental relevância para uma compreensão mais profunda acerca das dinâmicas das cidades e dos espaços urbanos monitorados. Portanto, por exemplo, informações sobre mobilidade urbana são úteis para sistemas de tráfego; aprender o comportamento humano e as relações sociais pode beneficiar a saúde pública, a segurança e o comércio; caracterizar regiões e cidades é fundamental para o planejamento urbano (PAN *et al.*, 2013). Logo, a detecção de outliers no monitoramento ambiental urbano pode auxiliar os gestores nas tomadas de decisões, a partir da extração de informações úteis desses ambientes podendo ajudar uma cidade em pelo menos oito de seus aspectos (PIRO *et al.*, 2014): transporte inteligente, planejamento urbano inteligente, saúde pública inteligente, segurança pública inteligente, comércio inteligente, energia inteligente, indústria inteligente e monitoramento ambiental urbano inteligente. Esses aspectos são abordados por (PAN *et al.*, 2013) e resumidos nesta seção.

Transporte Inteligente- Informações como condições de tráfego, mapas de estradas, oferta e demanda de transporte, além de informações e conhecimentos implícitos, como acidentes de trânsito, estratégia de direção e navegação de rotas, são informações vitais no monitoramento da dinâmica tráfego de veículos nas grandes cidades (ZHANG *et al.*, 2017). Os valores discrepantes detectados podem indicar certos picos na dinâmica do tráfego, por exemplo, acidentes de trânsito, alarmes de comportamentos inadequados de direção e até congestionamentos no trânsito.

Planejamento Urbano Inteligente - O planejamento urbano é um processo técnico e político relacionado ao controle do uso da terra e ao *design* de ambientes urbanos, que podem se beneficiar da detecção de outliers. Por exemplo, um grande número de veículos na saída de uma estrada ou de pacientes esperando do lado de fora de um hospital. Além disso, de acordo

com o monitoramento das obras de infraestrutura, o planejamento poderia ser investigado com o auxílio de métodos de detecção de outliers, retornando informações úteis sobre as demandas de serviços aos gestores públicos, ou seja, se os serviços são suficientes ou existem excessos.

Saúde pública inteligente - A detecção de outliers pode melhorar consideravelmente a saúde pública por meio do monitoramento de pacientes, indivíduos suscetíveis e pessoas não saudáveis. Discrepâncias observadas no acompanhamento da saúde de indivíduos nos resultados do monitoramento de exames, tais como, eletrocardiograma, pressão arterial, índices de glicemia, monitoramento de atividade física, em geral, podem revelar anormalidades na saúde dos indivíduos, auxiliando também na compreensão de indivíduos suscetíveis à patologias específicas e no controle da disseminação de doenças.

Segurança pública inteligente - A detecção de outliers pode ser usada para revelar muitos comportamentos humanos, como:

- Ocorrências de eventos sociais e encontros anormais de pessoas (a reunião e o movimento de multidões são frequentemente acompanhados por fluxos humanos anormais, que podem ser detectados e monitorados com a detecção maciça de outliers);
- Comportamento inadequado dos indivíduos (a maioria das pessoas têm rotas rotineiras todos os dias, por exemplo, um padrão repetitivo de viagem entre casa e trabalho durante a semana, e nesta perspectiva, os outliers poderiam representar um potencial comportamento inadequado dos indivíduos); e
- Movimentos de criminosos e comportamentos suspeitos, e pessoas perdidas em desastres (dados discrepantes de criminosos suspeitos podem inferir criminosos reais, encontrar gangues e detectar comunidades perdidas).

Comércio Inteligente - A publicidade é importante para as empresas aumentarem seus negócios. Assim, o comércio pode ser melhorado através da detecção de outliers. Por exemplo, os locais mais visitados pelas pessoas (identificados por outliers, representando um alto fluxo de clientes em potencial) podem ser locais estratégicos para colocar anúncios comerciais. Além disso, os valores discrepantes ainda podem indicar tendências fora do padrão de normalidade, como por exemplo, a anormalidade no comportamento de consumo rotineiro dos consumidores, ajudando a melhorar os serviços de compras, a partir, do conhecimento de comportamentos dos consumidores.

Energia inteligente - No campo energético, a detecção de outliers pode ser muito

útil, revelando quais eletrodomésticos apresentam maiores índices de energia. (DO; CETIN, 2018), em seu trabalho aplicou a detecção de outliers à um grande conjunto de dados de dados mensais de uso de energia de edifícios residenciais com três métodos diferentes, incluindo o método do desvio padrão, o método do quartil e o teste de Grubbs. Como resultado da aplicação da detecção de outliers, foram identificados não apenas os valores extremos, mas também as causas que os geraram, ou seja, os tipos de eletrodomésticos que aumentaram potencialmente os níveis de consumo energético das residências.

Indústria inteligente - Para um processo industrial, é de grande importância garantir a qualidade do produto e controlar o desempenho simultaneamente, dada à crescente demanda por sistemas com desempenho satisfatório, qualidade da produção e operação econômica. Nesta perspectiva, o trabalho de (WANG; MAO, 2019) propõe, em uma perspectiva industrial, um esquema de detecção externa que pode ser usada diretamente para o monitoramento de processos industriais ou controle de processos, a partir do desenvolvimento de algoritmos de detecção, dos quais a função média, função de covariância, função de probabilidade e o método de inferência, são especialmente desenvolvidos com base na regressão gaussiana tradicional.

Monitoramento ambiental urbano inteligente - Em particular, na perspectiva ambiental, a detecção de *outliers* nos espaços urbanos inteligentes ganha força na literatura, pois a detecção de ruído nas regiões, poluição da água e do ar, condições meteorológicas nas florestas e assim por diante, pode proporcionar um desenvolvimento sustentável e inteligente para as cidades (ZHANG *et al.*, 2017). Para que se entenda o uso das ferramentas de detecção de *outliers* no contexto do monitoramento ambiental urbano inteligente, em nossa pesquisa inicial (SOUZA *et al.*, 2019a) propomos um método *offline* para a detecção de *outliers*, explorando a natureza multidimensional dos dados monitorados nos espaços urbanos. Através de três estágios, nos quais, no primeiro estágio, os dados são modelados como um tensor de dados de terceira ordem (tensor 3D) para obter a redução de dimensionalidade do conjunto de dados, a fim de obter um melhor ajuste para o segundo estágio, que se configura em uma etapa de agrupamento e, finalmente, gera um modelo de detecção refinado, capturando padrões espaço-temporal de ambientes urbanos.

2.4 Sumário do capítulo

Neste capítulo foram apresentadas as principais características e definições matemáticas acerca dos fundamentos analíticos que serão utilizados para analisar os padrões extraídos

dos dados coletados nesta pesquisa. Nesta tese, foram escolhidos os modelos, Análise Fatorial Exploratória (análise multivariada) e HOSVD (análise multidimensional), uma vez que permitem a realização não apenas de uma análise exploratória dos conjuntos de dados analisados, mas também viabilizam a decomposição tanto de conjuntos de dados bidimensionais como multidimensionais. Além disso, tais modelos satisfazem o intuito desta tese que visa explorar tais ferramentas analíticas de modo inovador para a detecção de *outliers* no monitoramento ambiental urbano, contribuindo para a computação urbana. Porém, antes da aplicação dessas técnicas nos dados desta pesquisa, faz-se necessário compreender o conceito de *outliers* e como essas técnicas podem ser utilizadas no intuito de fornecer novos métodos de detecção de *outliers*. Desse modo, no capítulo seguinte abordaremos sobre a detecção de *outliers* de maneira abrangente, bem como focando na aplicação moderna de tensores na detecção de *outliers*, incluindo no monitoramento ambiental de espaços urbanos inteligentes.

3 DETECÇÃO DE *OUTLIERS* NO MONITORAMENTO AMBIENTAL URBANO

Neste capítulo, serão apresentados os procedimentos metodológicos e a caracterização de todo o contexto desta pesquisa. Inicialmente, iremos tratar sobre alguns aspectos acerca do monitoramento ambiental urbano, apresentando as principais plataformas de monitoramento que disponibilizam dados reais para análise. Em seguida, faremos uma caracterização dos dados coletados da plataforma escolhida para esta pesquisa. Por fim, as duas principais etapas de realização desta pesquisa serão caracterizadas por três abordagens distintas: (i) abordagem multivariada offline na detecção de *outliers*; (ii) abordagem multidimensional offline na detecção de *outliers*; e (iii) abordagem multidimensional online na detecção de *outliers*.

3.1 Contextualização

Os recentes avanços nas tecnologias de sensores conduziram ao desenvolvimento de pequenos monitores portáteis e de baixo custo para usos diversos e dinâmicos impactando, particularmente, o monitoramento ambiental urbano de grandes cidades (MCKERCHER *et al.*, 2017; THOMPSON, 2016; CARTON; ACHE, 2017). Essa tendência é motivada claramente por um desejo generalizado de criar avaliações mais precisas, por exemplo, da exposição à poluição do ar humano, identificar regiões críticas de temperaturas elevadas, etc. A quantificação dessas evidências é também útil para fins legislativos (THOMPSON, 2016), para que mudanças no planejamento urbano possam ser realizadas. Por exemplo, a criação de ciclovias "*off-road*"podem ser criadas ou mantidas para viagens não motorizadas mais seguras e saudáveis (CARTON; ACHE, 2017). Assim, a rápida aceleração das inovações tecnológicas na detecção de eventos ambientais oferece vastas oportunidades para melhorar a tomada de decisão individual e coletiva, e a capacidade de buscar maior igualdade ambiental.

Atualmente, massivos esforços de monitoramento relacionado às variáveis ambientais ganharam considerável popularidade em todo o mundo. Em nossa pesquisa identificamos pelo menos 4 esforços atuais ocorrendo em paralelo. Alguns desses esforços são discutidos em (THOMPSON, 2016). Basicamente os dados medidos pelos sensores são adquiridos e enviados via *bluetooth* ou conexão com fio para um *smartphone* ou *tablet*. Os dados do sensor são então combinados com um registro de data e hora e coordenadas GPS, e esses dados são enviados por uma conexão sem fio a um servidor centralizado, onde todos os dados são compilados e armazenados. Um usuário final pode adquirir os dados por meio de *download* e processamento

de dados posteriormente ou, alternativamente, o fluxo de dados pode ser retornado ao usuário via conexão sem fio para fornecer atualizações quase em tempo real do ambiente atual, sobrepondo os dados nos mapas ou fornecendo gráficos ou tabelas da experiência do usuário. O poder da abordagem reside em etapas muito rápidas de processamento e comunicação que fornecem dados ao usuário final e permitem que eles tomem decisões informadas.

Apresentamos, a seguir, os atuais esforços de monitoramento ambiental, categorizados como projetos acadêmicos ou comerciais. Os principais projetos acadêmicos são:

HAZEWATCH: Este projeto se origina de uma equipe de estudantes de engenharia elétrica, pesquisadores e professores da Universidade de *New South Wales*, na Austrália (SIVARAMAN *et al.*, 2013). A equipe do projeto constrói sensores móveis de poluição do ar que são conectados a veículos a motor e usados para coletar dados de qualidade do ar em Sydney e nas proximidades. As medidas incluem monóxido de carbono, ozônio, dióxido de enxofre e dióxido de nitrogênio. As medições do sensor são enviadas por meio de uma conexão *Bluetooth* para um *iPhone* dentro do veículo, e são coletadas marcações de tempo e espaço via GPS. Todos os dados são enviados para um servidor via conexão *WiFi* ou 4G. Os dados relatados são usados para criar mapas de poluição no *Google Maps* para uma interpretação rápida e fácil das leituras de poluição.

MAQUMON: A Rede Móvel de Monitoramento da Qualidade do Ar é um esforço coordenado pelo Laboratório de Sistemas Incorporados em Rede da Universidade Vanderbilt e apoiado pela Microsoft Research. A visão de rede consiste em vários nós de sensores montados em um veículo (ou portáteis) que medem níveis de O_3 (MiCs 2610), N_{O2} (MiCs 2710) e CO. O controle do instrumento é fornecido por um processador Intel 8051, e os dados podem ser armazenados a bordo via 2Mb de memória flash. Um receptor GPS EM-406 dedicado permite que o instrumento grave a localização sem a necessidade de usar o *smartphone* ou *tablet* para esta função. Os dados podem ser lidos diretamente pelo usuário final através de uma tela LCD ou, alternativamente, transferidos via USB ou *Bluetooth*.

Por outro lado, os principais projetos comerciais são:

TZOA: Este projeto combina plataformas de sensores com *upload* automático de dados e monitoramento em tempo real da qualidade do ar. Dois modelos de sensores estão disponíveis, o TZOA *consumer* (US $ 139) e a TZOA *Research* (US $ 600). Ambos os sensores incluem um contador de partículas, sensor UV, sensor de luz ambiente, umidade, temperatura e sensores de pressão. O esforço de TZOA foi estabelecido recentemente e sua equipe de projeto

continua avaliando e aprimorando os sensores.

Smart Citizen: Para o projeto *Smart Citizen*, o módulo sensor baseado em Arduino monitora gases poluentes como CO, N_{O2}, bem como, temperatura, umidade, intensidade da luz e níveis de ruído. Uma vez que os arquivos e os esquemas de *design* do dispositivo são de código aberto, os usuários podem transmitir os dados capturados para o site do projeto, além de possibilitar que os usuários criem seus próprios dispositivos sensores. O projeto *Smart Citizen* é um esforço colaborativo entre o *Fab Lab* de Barcelona no Instituto de Arquitetura Avançada da Catalunha. Segundo o site do projeto, atualmente existem mais de 800 módulos de sensores distribuídos em todos os continentes, exceto na Antártica.

Nesta perspectiva, os dados utilizados para o desenvolvimento desta tese foram extraídos da plataforma de monitoramento ambiental urbano *Smart Citizen*. A escolha dos dados desta plataforma é justificada pela possibilidade de se extrair não apenas séries temporais de variáveis ambientais sensoriadas, como também considerar a dimensão espacial desses dados, o que viabiliza considerar uma abordagem multivariada e construir uma abordagem multidimensional. Desta forma, apresentamos na Tabela 2 a seguir as variáveis consideradas juntamente com suas unidades para as três abordagens propostas nesta pesquisa. Em relação a quantidade de dados considerados nas abordagens propostas, apresentamos esses valores ao longo das seções seguintes, em cada abordagem respectiva.

Tabela 2 – Sumário das variáveis sensoriadas.

Variáveis Ambientais	Unidades
Temperatura	^{o}C
Umidade Relativa	%
Luminosidade	Lux
NO_2	ppm
CO	ppm
Ruído	dBA

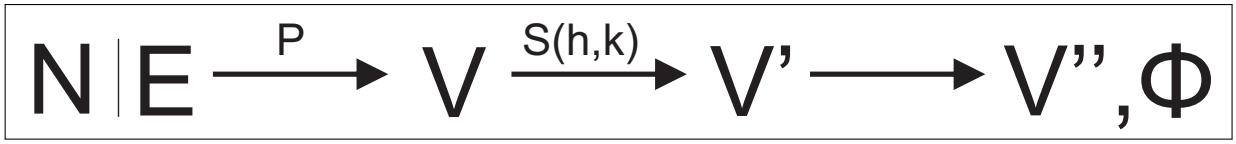
Fonte: elaborado pelo autor (2020).

3.2 Abordagens para Detecção de Outliers

Apresentamos três abordagens para o problema da detecção de outliers, a saber: abordagem multivariada *offline*, abordagem multidimensional *offline* e abordagem multidimensional *online*. Nesta perspectiva, modelamos o problema da detecção de *outliers*, nas respectivas abordagens, como um sistema. O diagrama da Figura 3, com base no apresentado por Aquino (AQUINO;

NAKAMURA, 2009) e também utilizado por (SOUZA *et al.*, 2019a), resume nossa modelagem.

Figura 3 – Diagrama da proposta de detecção de *outliers*.



Fonte: elaborado pelo autor (2020).

Um exemplo desse modelo é uma cidade (N), com nossa atenção restrita a uma área crítica E onde a aplicação captura de forma offline a ocorrência de eventos anômalos. O fenômeno de interesse pode ser de seis tuplas (temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído), com precisão infinita no espaço, tempo e medidas.

Um conjunto de nós de observação, $\mathbf{S} = (S_1, \ldots, S_o)$, é implantado para realizar uma amostragem sobre $\mathbf{V}$. Cada nó S_i está ciente de sua posição (h_i) e de uma função característica k_i que descreve as operações que ele pode executar. Na Figura 3, h denota a coleção de todas as posições e k denota o conjunto de todas as funções características.

A coleta de todas as observações multivariadas é da forma $(s_1, \ldots, s_o)(t)$, em que cada s_j representa a operação de S_j. A rede registra um ponto com as seis variáveis a cada instante. Considerando a coleta de dados ao longo de uma janela temporal de tamanho n, $t_1, \ldots, t_n$, as informações capturadas pela rede são

$$\begin{pmatrix} (s_1, \ldots, s_o)(t_1) \\ (s_1, \ldots, s_o)(t_2) \\ \vdots \\ (s_1, \ldots, s_o)(t_n) \end{pmatrix},$$

que representa um vetor com valor real $\mathbf{V}'_{6\times o\times n}$, e assumimos que os dados sejam estacionários em cada janela temporal.

Conforme descrito na Figura 3, a estratégia de processamento multivariado é uma transformação da forma

$$\Psi \colon \mathbb{R}^{p\times o\times n} \to \mathbb{R}^{p\times o\times n'},$$

onde $n' < n$ é o número de outliers sobre $\mathbf{V}'$, portanto mantemos o número de variáveis p e o número de nós o.

Por uma questão de simplicidade, descrevemos essa estratégia de processamento multivariado em termos de cada nó, isto é, toda a transformação Ψ é o resultado da aplicação de operações em cada nó $1 \leq i \leq o$: $\Psi = (\Psi_1, \ldots, \Psi_o)$. Cada transformação de modo semelhante produz uma transformação Ψ_i: $\mathbb{R}^{p \times n} \rightarrow \mathbb{R}^{p \times n'}$.

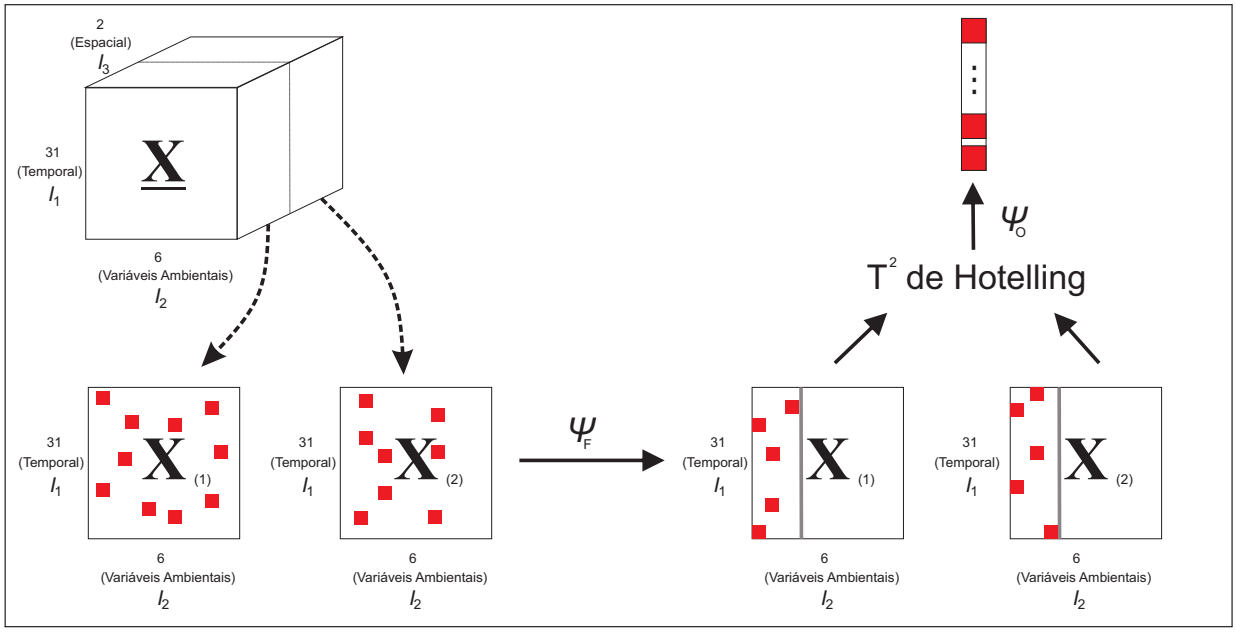
3.2.1 Detecção de Outlier Multivariado Offline

Modelando o problema da detecção de *outliers* como um sistema, podemos derivar a abordagem multivariada *offline*, conforme esboçado no diagrama da Figura 3. Nessa abordagem, as cidades escolhidas Elda e Rois compreendem os ambientes (N) a serem analisados, em que sobre o conjunto de todas as variáveis sensoriadas (V - temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído), é realizada a amostragem Ψ_i para cada N. Portanto, na abordagem multivariada *offline*, a amostragem Ψ_i é a composição de duas funções,

$$\Psi_i = \psi_F \circ \psi_O,$$

a saber, uma redução de dimensionalidade através da aplicação da análise fatorial exploratória (Ψ_F) sobre cada matriz de dados, seguida por uma detecção de *outliers* (Ψ_O). A Figura 4 ilustra essa etapa da modelagem com mais detalhes.

Figura 4 – Modelagem da análise multivariada offline.



Fonte: elaborado pelo autor (2020).

Os dados modelados nessa abordagem multivariada *offline* compreendem inicial-

mente um tensor de dados tridimensional $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ de dimensões $I_1 \times I_2 \times I_3$ (Figura 4). A partir deste tensor 3D ($\underline{\mathbf{X}}$) de dados, utilizamos a dimensão espacial representada pelas cidades de Elda e Rois, respectivamente, para a realização da análise fatorial em cada uma das matrizes multivariadas. Portanto, essa abordagem multivariada *offline* é apresentada no Algoritmo 1, em que a função ψ_F é aplicada sobre cada arranjo ($\mathbf{X}_1$ e $\mathbf{X}_2$) de dimensões $I_1 \times I_2$ cada. Após a aplicação da análise fatorial, o conjunto multivariado de dados é reduzido e sobre os fatores com maior variância, realizamos a detecção de *outliers* aplicando a função ψ_O. Como resultado da aplicação da estatística T^2 de *Hotelling*, temos um vetor contendo os *outliers* de cada arranjo matricial. A partir deste vetor de *outliers*, podemos realizar análises sobre os possíveis padrões de anormalidade extraídos da série temporal das variáveis ambientais de ambas as cidades. Portanto, fornecemos um quadro estatístico para combinar as dimensões, temporal e espacial, e assim detectar eventos no subespaço gerado a partir da aplicação do modelo fatorial.

Algoritmo 1: Abordagem multivariada offline de detecção de outliers

Entrada: Conjunto de dados multivariado X.

Amostragem Ψ_i.

1. $\psi_F \leftarrow$ calcular a AFE para um arranjo multivariado X.

2. $X_f \leftarrow$ selecionar os fatores mais representativos.

para $i \leftarrow 1$ *a* f **faça**
| $\psi_O \leftarrow$ Calcule sobre os fatores a distância de Mahalanobis.

fim

3. Reporte da estatística de detecção por fator.

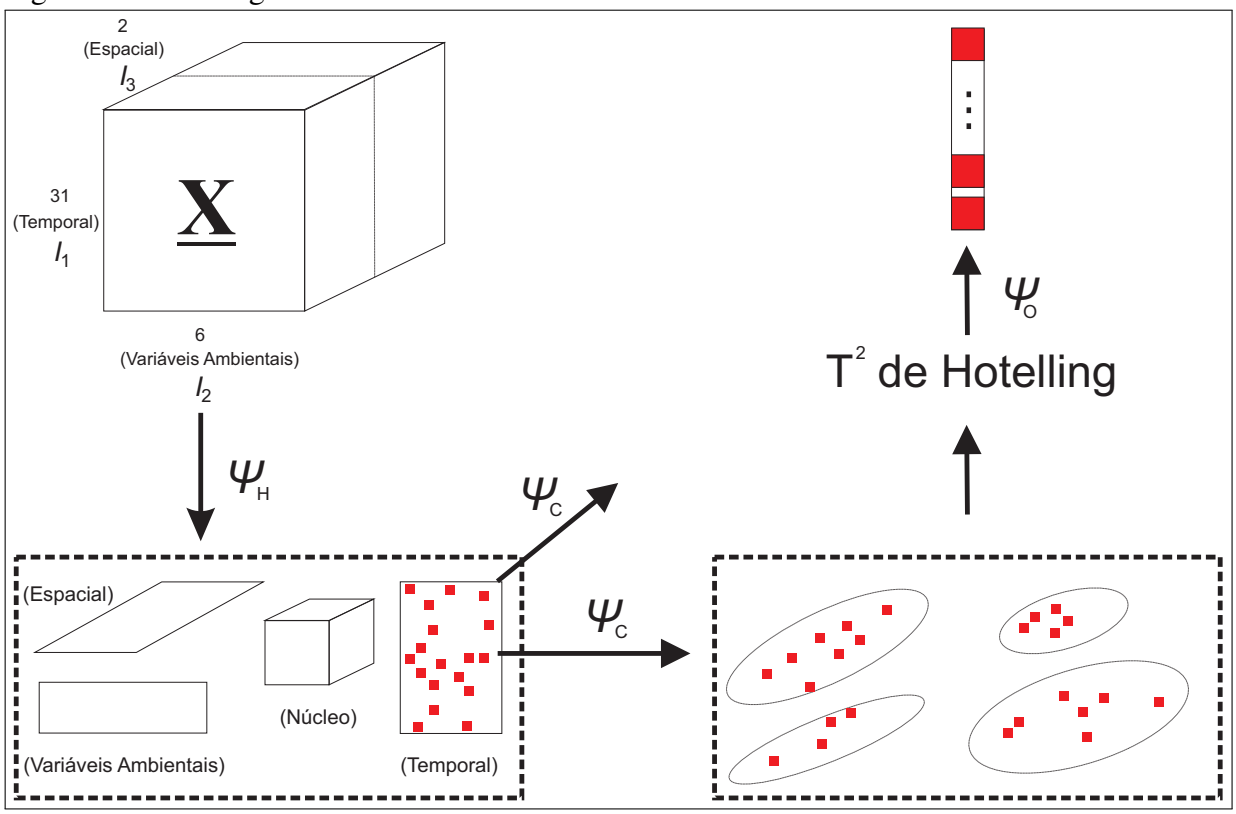
3.2.2 Detecção de Outlier Multidimensional Offline

Novamente, modelando o problema da detecção de outliers como um sistema, podemos derivar a abordagem multivariada offline, conforme esboçado no diagrama da Figura 3. Nessa abordagem, as cidades escolhidas Elda, Rois e Tallin compreendem os ambientes (N) a serem analisados, em que sobre o conjunto de todas as variáveis sensoriadas (V - temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído), é realizada a amostragem Ψ_i para cada N. Portanto, na abordagem multidimensional offline, a amostragem Ψ_i é a composição de três funções,

$$\Psi_i = \psi_H \circ \psi_C \circ \psi_O,$$

a saber, uma redução de dimensionalidade do HOSVD (Ψ_H), seguida por uma análise de agrupamento (Ψ_C) e, em seguida, a detecção de outliers (Ψ_O). A Figura 5 ilustra essa etapa da modelagem da análise multidimensional offline com mais detalhes.

Figura 5 – Modelagem da análise multidimensional offline.



Fonte: elaborado pelo autor (2020).

Assim, para detectar os outliers sobre $\mathbf{V}'$, aplicamos primeiro, o método multidimensional HOSVD para reduzir a dimensionalidade e revelar possíveis associações entre os componentes do tensor. Em seguida, realizamos a análise de cluster (ψ_C), com base no método *k*-means, apresentado no Algoritmo 2, para obter padrões de cluster semelhantes, a partir, dos componentes selecionados dos fatores matriciais retornados pelo modelo HOSVD.

Para a aplicação do método multidimensional HOSVD (ψ_H), organizamos o conjunto de todas as observações multivariadas $\mathbf{V}'$ em um tensor de terceira ordem $\underline{\mathbf{X}}_{I_1,I_2,I_3}$, onde I_1 corresponde à dimensão do tempo, I_2 às variáveis monitoradas e I_3 às cidades analisadas. Portanto, cada matriz $\mathbf{V}'$ (no total, quatro séries de observações multivariadas, cada uma representando uma cidade) é considerada como uma fatia do tensor $\underline{\mathbf{X}}$, na qual executamos o processo reverso de desdobramento no modo-*n*, denominado matriciação (como apresentado na Seção 2), para gerar um tensor genérico $\underline{\mathbf{X}}$, portanto: $\underline{\mathbf{X}}$, so: $\mathbb{R}^{I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_n} \rightarrow \mathbb{R}^{I_1 \times I_2 \ldots I_n}$. Assim, mapeamos o conjunto de todas as observações multivariadas para um tensor 3D.

Para o caso particular de um tensor de terceira ordem, foco deste trabalho, o método HOSVD decompõe um tensor $\underline{\mathbf{X}}_{I_1 I_2 I_3}$ nas matrizes fatoriais $\mathbf{U}_1$ ($I_1 \times w$), $\mathbf{U}_2$ ($I_2 \times q$), $\mathbf{U}_3$ ($I_3 \times r$) e no tensor núcleo $\underline{\mathbf{S}}_{wqr}$, onde $w < I_1$, $q < I_2$, e $r < I_3$.

Semelhante ao método multivariado SVD truncado para aproximação de baixo rank e redução de dimensionalidade de matrizes bidimensionais, usamos o método HOSVD para executar a aproximação de baixo rank e redução de dimensionalidade. Através da seleção das primeiras w colunas de $\mathbf{U}_1$, q colunas de $\mathbf{U}_2$ e r colunas de $\mathbf{U}_3$, a redução da dimensionalidade pelo HOSVD é eficiente. Os componentes do modelo foram selecionados usando o critério com base na variação explicada de cada componente. O número de componentes principais de cada matriz fatorial é escolhido com base na porcentagem cumulativa de variância explicada (KROONENBERG, 2008). Portanto, se a porcentagem acumulada dos primeiros componentes estiver acima de um limite (por exemplo, 75 % (FANAEE-T; GAMA, 2016)), o número apropriado de componentes será selecionado com os componentes que excederem esse limite. Assim, esta etapa apresenta uma maneira de reduzir drasticamente a dimensionalidade do conjunto de dados.

Uma vez que a função ψ_H extrai os fatores latentes, nossa análise se concentra no monitoramento das séries temporais extraídas da matriz de fatores $\mathbf{U}_1$ (modo temporal). Em seguida, aplicamos a função ψ_C com base no k-means (JAIN, 2010) para agrupar os fatores da matriz modo temporal. Esses componentes são selecionados anteriormente pelo critério de variação explicado, de modo que a semelhança entre os objetos de dados em cada cluster seja mais significativa do que a dos objetos de dados vizinhos aos clusters. Assim, minimizamos a soma dos quadrados das distâncias dentro de cada cluster, resultando em clusters com fatores de carregamento semelhantes. A medida de distância mais utilizada para calcular a semelhança é a distância euclidiana. O algoritmo padrão para o particionamento de dados k-means:

Algoritmo 2: Algoritmo função ψ_C.

Data: Vetores de projeção no subespaço da matriz fator; k, o número de clusters
Result: k centróides
Defina o número desejado de clusters, k;
Escolha de k aleatoriamente, a partir, dos vetores de projeção do subespaço da matriz fatorial, como pontos de partidas iniciais;
repeat
  Atribua cada vetor de projeção de subespaço ao cluster com o centróide mais próximo;
  Calcule a média de todos os clusters e atualize o valor do centróide para o valor médio desse cluster;
until *se alguma partição foi modificada desde a última iteração*;
return k centróides.

Para avaliar o desempenho dos clusters gerados, usamos o coeficiente de Silhouette:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{3.1}$$

onde $a(i)$ é a medida média do objeto i em relação a todos os objetos no mesmo cluster k, $b(i)$ é a medida média entre o objeto i em relação a todos objetos no cluster vizinho mais próximo e $s(i)$ varia de $[-1,1]$ para que $-1 \leq s(i) \geq 1$.

Por fim, a função ψ_O é aplicada, com base na distância de Mahalanobis para a detecção de *outliers* sobre os *clusters* gerados, a partir dos fatores extraídos da matriz modo temporal. Uma vez que, medidas baseadas em distâncias são utilizadas para localizar automaticamente observações multivariadas que estão distantes do centro do conjunto de dados, aplicamos nesta tese a distância de Mahalanobis por considerar as correlações entre as variáveis, bem como as diferenças nas variâncias entre essas variáveis. O Algoritmo 3 apresenta a amostragem Ψ_i para a abordagem multidimensional *offline* de detecção de *outliers*.

Algoritmo 3: Abordagem multidimensional offline de detecção de outliers

Entrada: Conjunto de dados multidimensional $\underline{\mathbf{X}}$.

Amostragem Ψ_i.

1. $\psi_H \leftarrow$ calcular a decomposição HOSVD para um arranjo multidimensional $\underline{\mathbf{X}}$.
2. $X_f \leftarrow$ selecionar os componentes mais representativos.
3. $\psi_C \leftarrow$ aplicar a função sobre os componentes selecionados.

para $i \leftarrow 1$ *a* X_f **faça**
| $\psi_O \leftarrow$ Calcule sobre os fatores a distância de Mahalanobis.

fim

3. Reporte da estatística de detecção por fator.

3.2.3 Detecção de Outlier Multidimensional Online

Para a abordagem multidimensional *online*, adaptamos a modelagem *offline*, usando uma janela deslizante de comprimento fixo com intervalos de tempo bem definidos. Portanto, nesta abordagem o nosso método é composto por duas funções, pois não usamos a função de análise de agrupamento. Portanto, a amostragem foi composta pelas funções, redução de dimensionalidade utilizando o método multidimensional HOSVD (Ψ_H) e detecção de *outlier online* (Ψ_O), através da janela deslizante:

$$\Psi_i = \psi_H \circ \psi_O,$$

Para modelar os dados para usar o método multidimensional HOSVD (ψ_H), organizamos o conjunto de todas as observações multivariadas $\mathbf{V}'$ em um tensor de terceira ordem $\underline{\mathbf{X}}_{I_1,I_2,I_3}$, onde I_1 corresponde à dimensão do tempo, I_2 às variáveis detectadas e I_3 às cidades analisadas. Portanto, cada matriz $\mathbf{V}'$ (no total, três séries de observações multivariadas representando uma cidade respectiva) é considerada como uma fatia do tensor $\underline{\mathbf{X}}$. Portanto, através de uma janela deslizante, é gerado um modelo de observação dos fluxos de dados mais recentes.

Desta forma, a distância de projeção de cada vetor de dados decomposto pelo método multidimensional, no subconjunto definido por cada componente principal selecionado, é calculada à medida que novos dados chegam e um índice estatístico externo é calculado.

Assim, para identificar os outliers, aplicamos a função ψ_H para reduzir a dimensionalidade e descobrir possíveis associações entre os componentes do tensor multidimensional. Após a redução de dimensionalidade, a função Ψ_O é aplicada para a detecção de outliers. Dentro desta função, incorporamos o método da janela deslizante para capturar as mudanças contínuas das características estatísticas da similaridade de maneira rápida e oportuna. A idéia central é obter, a partir, da composição das duas funções (Ψ_i) uma melhoria na precisão da detecção no monitoramento on-line. Então, à medida que a janela move todo o processo de decomposição multidimensional de dados é realizado para cada unidade amostral considerada na série temporal analisada, a distância de Mahalanobis é calculada e um limiar é gerado para cada resultado da detecção.

Além disso, o número de componentes principais de cada matriz fatorial foi escolhido com base na porcentagem cumulativa de variância explicada, de forma similar à abordagem offline.

Portanto, na abordagem de detecção multidimensional online de *outlier*, nos baseamos em janelas deslizantes, ou seja, a cada instante é verificada uma quantidade constante de tempo, chamada de janela (NGUYEN *et al.*, 2015). Como o fluxo é atualizado continuamente com dados atualizados, é impossível manter todos eles na memória principal. Portanto, é usada uma janela que monitora os dados mais recentes e todas as tarefas de decomposição e detecção são realizadas com base no que é "visível"através da janela. Ou seja, usamos janelas deslizantes para restringir nossa atenção a dados recentes, porque as séries temporais são ruidosas e podem mudar seus comportamentos ao longo do tempo, ou seja, não estacionárias. Nesse contexto, dados antigos podem adicionar viés à inferência de dados recentes, inclusive no bojo do monitoramento ambiental urbano. Portanto, a aplicação da técnica da janela deslizante é útil para

rastrear a dinâmica do processo nos dados, não apenas lidando com a não estacionariedade, mas também reduzindo o custo computacional do algoritmo e dos requisitos de armazenamento, para que sejam adequados para detecção online. Portanto, para os dados dentro da janela, realizamos a estatística de detecção.

O tensor $\underline{\mathbf{X}}_{I_1 I_2 I_3}$ é amostrado periodicamente ao longo da série temporal da dimensão I_1, para as variáveis detectadas nas cidades analisadas. Portanto, um fluxo multidimensional é um fluxo de linhas de dados do tensor $\underline{\mathbf{X}}$ que abrange as três dimensões (I_1, I_2, I_3), isto é, ao definir a dimensão I_1 e variar as dimensões I_2 e I_3, temos a unidade de amostra atual de interesse. A Figura 6 mostra o esquema desta abordagem, onde, quando fixamos a dimensão I_1, por exemplo, t1 (veja linha tracejada na Figura 6), temos o primeiro instante da série temporal ao longo das variáveis (dimensão I_2) nas respectivas cidades (dimensão I_3). Nossa janela temporal tem uma duração de 24 horas, na qual, após o método ser aplicado à cada unidade amostral do tensor de dados, após percorrer as primeiras 24 horas, passamos para a segunda janela na qual descartamos o primeiro elemento (instante t1) da janela 1 e consideramos o instante de tempo seguinte (t25) para a janela 2 (Figura 6). Além disso, dentro de cada janela, a distância de Mahalanobis é calculada em cada unidade amostral retornada pela decomposição multidimensional à medida que a janela se move.

Algoritmo 4: Abordagem multidimensional online de detecção de outliers

Entrada: Conjunto de dados multidimensional $\underline{\mathbf{X}}$.

Amostragem Ψ_i.

1. [Xi, Xj, Xk]=size($\underline{\mathbf{X}}$);

2. width_window = 24; deslocamento = 1;

janela = floor(Xi-width_window)/deslocamento;

para $i \leftarrow 1$ *a janela* **faça**

  $\psi_H \leftarrow$ calcule a decomposição HOSVD para um arranjo multidimensional $\underline{\mathbf{X}}$

  $\psi_O \leftarrow$ Calcule sobre os fatores a distância de Mahalanobis.

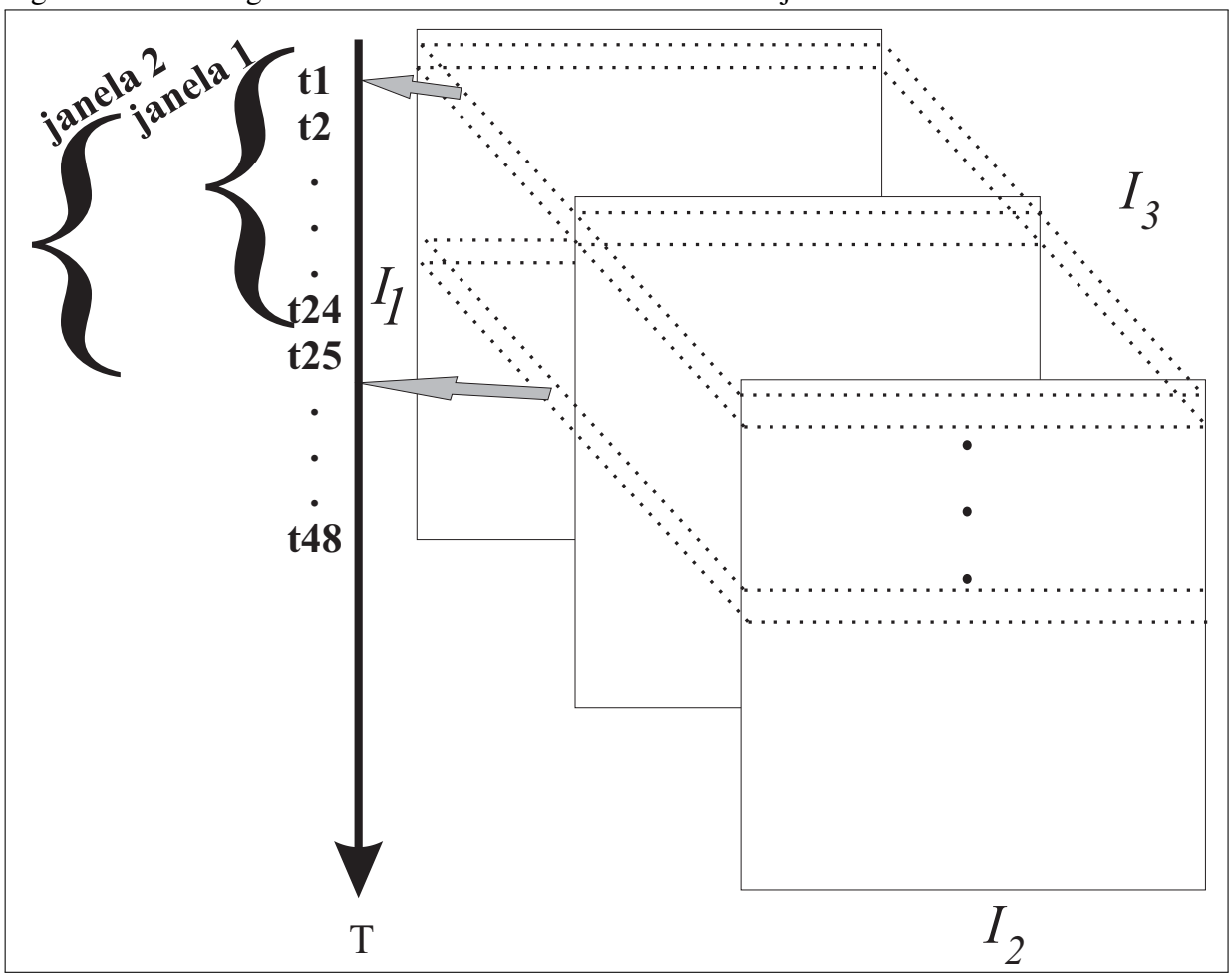
fim

3. Reporte da estatística de detecção por fator.

3.3 Detecção de *Outliers* e Limiar de Detecção

Na função de detecção de outliers ψ_O, calculamos a distância de projeção de cada vetor de grupo no subespaço definido pelo componente selecionado com maior variação. Para isso, a distância usada foi a de Mahalanobis, também conhecida como estatística T^2 de Hotel-

Figura 6 – Abordagem do monitoramento ambiental online: janela deslizante.



Fonte: elaborado pelo autor (2020).

ling, uma métrica comum para monitorar séries temporais, que é calculada da seguinte forma (MAHALANOBIS, 1936):

$$T_t^2 = (\mathbf{x}_t - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_t - \overline{\mathbf{x}})^T, \tag{3.2}$$

onde $\overline{\mathbf{x}}$ é a média, $\mathbf{x}_t$ é a observação multivariada no momento t e $\mathbf{S}$ é a matriz de covariância.

Com base no resultado dessa métrica, classificamos uma observação no instante t_i sobre $\mathbf{V}'_{t_i}$ como uma condição normal, se o valor calculado para a distância de Mahalanobis (T_t^2) estiver abaixo do limite de controle de distância de Mahalanobis (T_α), ou seja, $T_t^2 < T_\alpha$, por outro lado, o classificamos como um *outlier*, se a distância de Mahalanobis (T_t^2) for igual ou exceder o limite de controle de distância de Mahalanobis (T_α), ou seja, $T_t^2 \geq T_\alpha$. Desta forma, um *outlier* é definido a partir do cálculo dessa métrica. Os limites aproximados do controle de distância de Mahalanobis, com um nível de confiança α, podem ser determinados de diferentes

maneiras aplicando as premissas de distribuição de probabilidade (TRACY *et al.*, 1972):

$$T_\alpha = \frac{d(n^2-1)}{n(n-d)} F_\alpha(d, n-d), \tag{3.3}$$

onde $F_\alpha(d, n-d)$ é o limite superior do percentil da distribuição de F com graus de liberdade d e $n-d$. Portanto, se $T_t^2 > T_\alpha$, ou seja, maior que o limite superior, as observações são consideradas outliers, caso contrário, normais. É importante destacar que os rótulos dos *outliers* são definidos a partir do cálculo desta métrica, ou seja, para as instâncias que estiverem fora do limiar calculado.

3.4 Avaliação das abordagens propostas

Para avaliar a eficiência dos métodos aplicados nas abordagens propostas nesta tese, utilizamos a métrica *Receiver Operating Characteristic* (ROC), que é um método clássico aplicável a algoritmos de detecção de *outliers* não-supervisionados (GOLDSTEIN; UCHIDA, 2016). A curva ROC compreende uma representação gráfica da sensibilidade no eixo vertical e da especificidade no eixo horizontal (SWETS; PICKETT, 1982). O valor da "área sob a curva"(*Area Under Curve* - AUC), caracteriza a área sob a curva ROC, que varia entre 0 e 1.

Em nossa pesquisa, a sensibilidade ou taxa de verdadeiro positivo (*True-Positive Rate* - TPR) corresponde à razão entre o número de outliers corretamente identificados (ou seja, verdadeiro positivo, *True-Positive* - TP), e o número de todos os outliers (ou seja, TP + falso negativo, *False-Negative* - FN), onde FN são os outliers falsos identificados.

$$sensibilidade = \frac{TP}{TP+FN}. \tag{3.4}$$

Por outro lado, a especificidade ou taxa de falso positivo (*False-Positive Rate* - FPR) é a razão entre o número de eventos normais corretamente identificados (ou seja, falso positivo, *False-Positive* - FP) e o número de todos os eventos (ou seja, FP + verdadeiro negativo, *True-Negative* - TN), onde TN é o número de eventos erroneamente identificado como eventos normais.

$$especificidade = \frac{FP}{TN+FP}. \tag{3.5}$$

Nos últimos anos as curvas ROC têm sido uma das métricas mais populares para a avaliação de desempenho de diferentes abordagens de detecção de *outliers* (CAMPOS *et al.*, 2016). As abordagens propostas nesta tese ilustram uma detecção consistente de *outliers* quando comparado com outros trabalhos da literatura, uma vez que, apresenta uma pequena proporção de falsas detecções, enquanto que o sucesso de outros métodos concorrentes dependem mais do conjunto de dados.

3.5 Sumário do capítulo

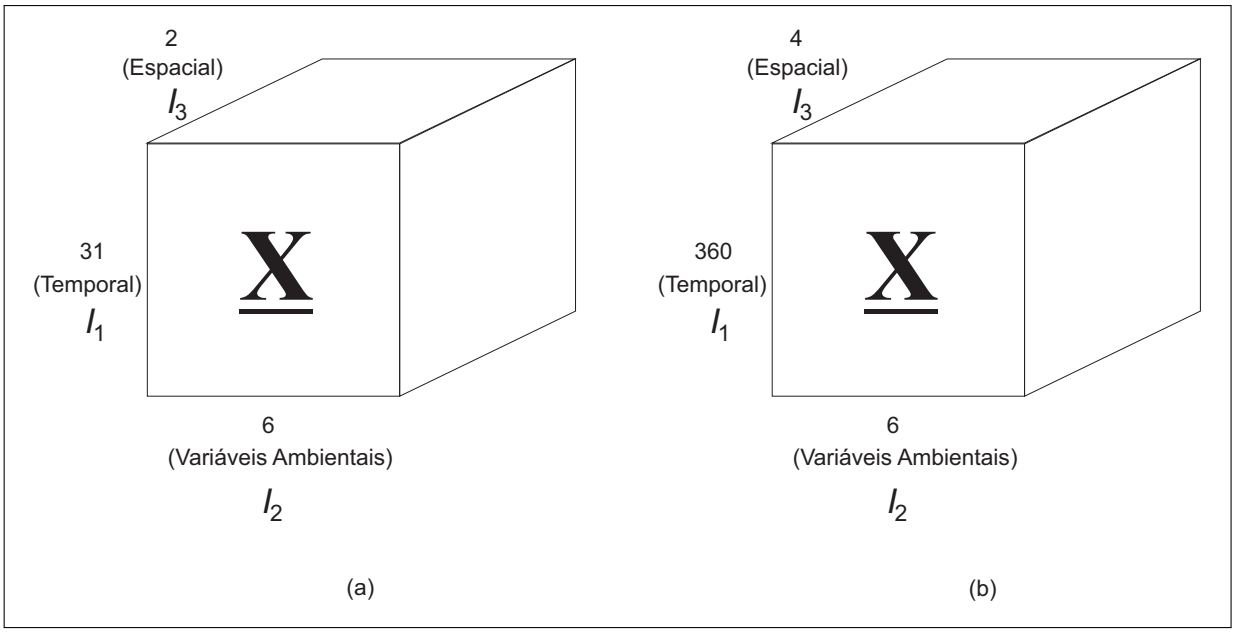
Neste capítulo, realizamos uma caracterização geral acerca dos conjuntos de dados que serão analisados nesta tese, bem como um detalhamento de sua estrutura. Além disso, realizamos uma breve discussão sobre as plataformas de monitoramento ambiental mais populares, assim como a plataforma de monitoramento ambiental utilizada nesta pesquisa. Por fim, caracterizamos as três bordagens metodológicas desta pesquisa, descrevendo e discutindo a modelagem para cada abordagem. No próximo capítulo, realizamos a análise e discussão dos resultados obtidos a partir da modelagem empregada neste capítulo, com o objetivo de obter padrões de comportamento de variáveis ambientais a partir da detecção de outliers.

4 RESULTADOS

Neste capítulo, são apresentados e discutidos os principais resultados obtidos pela aplicação dos métodos propostos. Na primeira parte, será realizada uma análise obtida a partir da análise multivariada, considerando a validação dos dados, seleção de componentes principais do modelo e os resultados da aplicação das estatísticas de detecção de *outliers* a partir da modelagem bidimensional em análise. Na segunda parte, será realizada uma análise obtida a partir da análise multidimensional offline, considerando a descrição dos dados, bem como a seleção de componentes principais do modelo, seguida da análise das matrizes fatores e da análise de agrupamento, finalizando com a detecção de *outliers*. Por fim, na terceira etapa, serão discutidos os resultados retornados a partir da análise multidimensional online, considerando a detecção de *outliers* através do método da janela deslizante.

De forma geral, a organização dos dados coletados, a partir das medidas capturas pelos sensores fornecidos pela plataforma de monitoramento ambiental Smart Citizen (CITIZEN, 2016), para as três abordagens propostas nesta pesquisa, são multidimensionais (3D) (Figura 7), com as seguintes dimensões: temporal (I_1), variáveis ambientais (I_2) e espacial (I_3). A Figura 7a ilustra o esquema do arranjo de dados para a abordagem multivariada, enquanto que a Figura 7b ilustra o arranjo de dados para a abordagem multidimensional, tanto *offline* quanto *online*.

Figura 7 – Representação do arranjo tensorial de dados: (a) arranjo da abordagem multivariada; (b) arranjo da abordagem multidimensional.

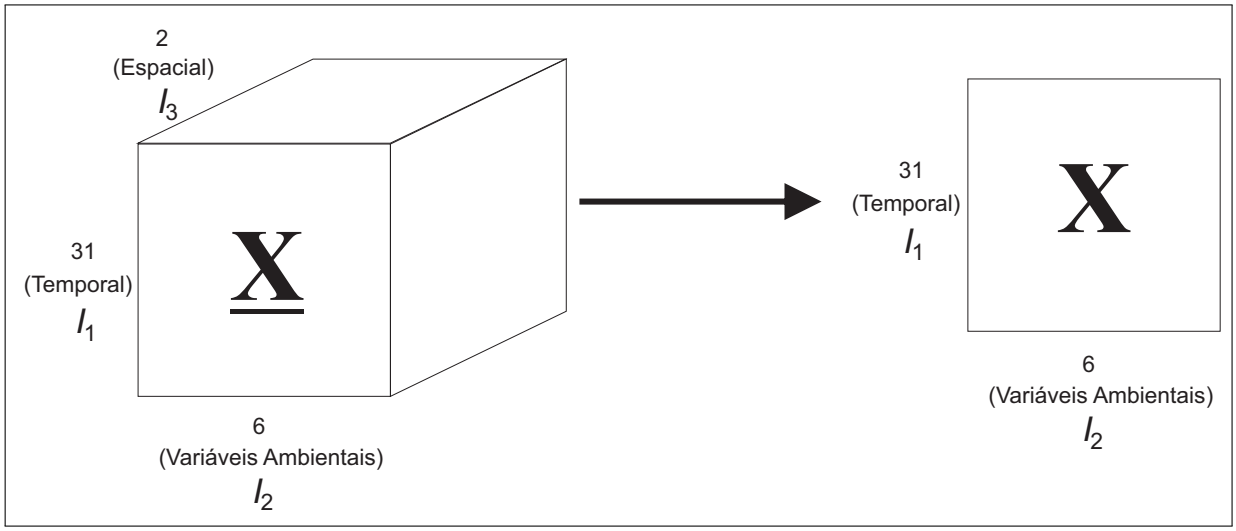


Fonte: elaborado pelo autor (2020).

4.1 Resultados da Análise Multivariada

Os dados coletados para a abordagem offline multivariada foram organizados em um tensor com dimensões 31 (tempo em dias) $\times$ 6 (variáveis ambientais) $\times$ 2 (cidades monitoradas). Os dados obtidos a partir dos sensores da plataforma Smart Citizen, fornecem as medidas de temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído (Tabela 2). Diante do exposto, as medidas de tais variáveis (registradas por um período de 31 dias do mês de Julho de 2017) de duas cidades da Espanha, Elda e Rois, foram transformadas a partir de um tensor 3D, em uma matriz 2D realizando o processo de matriciação (Figura 8). Assim, a análise fatorial exploratória é realizada sobre cada matriz gerada, cada uma com dimensões, 31 (tempo em dias) $\times$ 6 (variáveis ambientais).

Figura 8 – Representação do arranjo tensorial de dados: (a) arranjo da abordagem multivariada; (b) arranjo da abordagem multidimensional.



Fonte: elaborado pelo autor (2020).

Aplicamos a função ψ_F para realizar a Análise Fatorial Exploratória (AFE) sobre o conjunto de dados que engloba as 6 variáveis observadas durante um período de 31 dias do mês de Julho de 2017 de duas cidades da Espanha, Elda e Rois. Os resultados apontaram para uma solução de dois fatores, derivados para cada cidade.

Análises preliminares foram realizadas para examinar a adequação dos dados à AFE (Tabela 3). Para tanto, no intuito de verificar se as variáveis analisadas são correlacionadas entre si, gerando a hipótese de a matriz de correlação das variáveis ser identidade, os testes de Kaiser-Meyer-Olkin (KMO) e Bartlett foram aplicados sobre os dados de cada cidade. Conforme observado na Tabela 3, o valor do teste KMO se mostrou significativo tanto para os dados

da cidade de Elda (0,71), quanto da cidade de Rois (0,75), garantindo uma boa adequação da amostra a aplicação da AFE, uma vez que possui valor superior a 0,6 (BARTHOLOMEW; KNOTT, 1999). Já o teste de Bartlett rejeitou a hipótese de que a matriz de correlação seria a matriz identidade (BASILEVSKY, 2009). Desta forma, ambos os métodos mostraram que os dados são adequados para aplicação da AFE.

Tabela 3 – Testes de Validação da AFE

Cidade	Adequação da Amostra - KMO	Esfericidade de Bartlett
Elda	0,71	489,08
Rois	0,75	435,89

Fonte: elaborado pelo autor (2020).

A variância extraída de cada nova variável foi comparada com as variâncias das variáveis originais, verificando o quanto de variância comum (comunalidade), existe entre as variáveis observadas e as que foram obtidas através da AFE. Desta forma, conforme observado na Tabela 4, os valores de comunalidade das variáveis apresentam valores superiores a 0,6, indicando que todas as variáveis, para ambas as cidades, apresentam elevada representatividade dentro dos fatores extraídos pela AFE.

Tabela 4 – Testes de Validação da AFE

Cidade	Variáveis	Extração	Cidade	Variáveis	Extração
	Temperatura	0,84		Temperatura	0,74
	Umidade	0,89		Umidade	0,69
Elda	Luminosidade	0,81	Rois	Luminosidade	0,67
	NO_2	0,93		NO_2	0,85
	CO	0,95		CO	0,92
	Ruído	0,99		Ruído	0,74

Fonte: elaborado pelo autor (2020).

Para a seleção do número de fatores, foi utilizado o critério da variância explicada (SUNDBERG; FELDMANN, 2016; SOUZA *et al.*, 2017) cujos primeiros dois fatores explicam cerca de 73% da variância total para a cidade de Elda (Tabela 5), e cerca de 83% da variância para a cidade de Rois (Tabela 6). Além da variância explicada, foi utilizado também o critério de Kaiser (KAISER, 1966; SOUZA *et al.*, 2017), que diz que os fatores a serem considerados devem apresentar autovalores acima da unidade ($\lambda > 1$, segunda coluna da Tabela 5 e Tabela 6).

Os fatores de carregamento permitem que uma correlação possa ser estabelecida entre as variáveis observadas e os fatores extraídos. Desta forma, tanto para a cidade de Elda (Tabela 7) quanto para a cidade de Rois (Tabela 8), todas as cargas com valores superiores a 0,6

Tabela 5 – Distribuição da Variância Explicada da AFE - Cidade de Elda

Fatores	Autovalores (λ)	% Variância	% Variância Acumulativa
1	2,89	48,17	48,17
2	**1,51**	**25,18**	**73,35**
3	0,99	16,69	90,05
4	0,32	5,44	95,49
5	0,18	3,13	98,62
6	0,08	1,39	100

Fonte: elaborado pelo autor (2020).

Tabela 6 – Distribuição da Variância Explicada da AFE - Cidade de Rois

Fatores	Autovalores (λ)	% Variância	% Variância Acumulativa
1	2,17	46,23	46,23
2	**1,61**	**36,90**	**83,14**
3	0,98	7,41	90,55
4	0,57	5,66	96,22
5	0,47	3,13	99,35
6	0,17	0,67	100

Fonte: elaborado pelo autor (2020).

estão destacadas em negrito. Nesta pesquisa, a análise desses fatores se estabeleceu a partir do cruzamento de cargas elevadas com as demais variáveis. Assim, a partir do padrão observado deste cruzamento, nomeamos os fatores extraídos de cada cidade que obtiveram os maiores valores, conforme discutido a seguir:

- **Fatores de Carregamento da Cidade de Elda -** Analisando os valores dos fatores das cargas desta cidade (Tabela 7), ambos têm em comum o fato de se referirem prioritariamente às variáveis climáticas, uma vez que para as variáveis, temperatura e umidade, os fatores apresentaram os maiores valores para o Fator I. Dessa forma, para fins de análise nomeamos o Fator I como **Condições Climáticas**.
- **Fatores de Carregamento da Cidade de Rois -** Os fatores de carregamento para esta cidade (Tabela 8) apresentam características relacionadas à poluição da cidade. Isto devido ao fato das variáveis monóxido de carbono (CO) e dióxido de nitrogênio (NO_2) apresentarem os maiores fatores de carregamento para o Fator II. Assim, o Fator II recebeu a denominação de **Qualidade do Ar**.

Os fatores derivados do modelo exploratório fatorial obtidos neste estudo podem ser utilizados como ferramenta para identificar padrões de eventos relacionados as variáveis ambientais que estejam fora do padrão de normalidade. Como é amplamente conhecido que uma cidade com cidadãos pouco saudáveis dificilmente se torna uma cidade inteligente, uma

Tabela 7 – Fatores de carregamento da AFE - Cidade de Elda

Variáveis Ambientais	Fator I	Fator II
Temperatura	**0,99**	0,59
Umidade	**0,95**	0,53
Luminosidade	0,03	0,15
NO_2	0,11	0,167
CO	0,41	0,09
Ruído	0,20	0,11

Fonte: elaborado pelo autor (2020).

Tabela 8 – Fatores de carregamento da AFE - Cidade de Rois

Variáveis Ambientais	Fator I	Fator II
Temperatura	0,39	0,38
Umidade	0,43	0,56
Luminosidade	0,19	0,15
NO_2	0,54	**0,91**
CO	0,49	**0,95**
Ruído	0,01	0,32

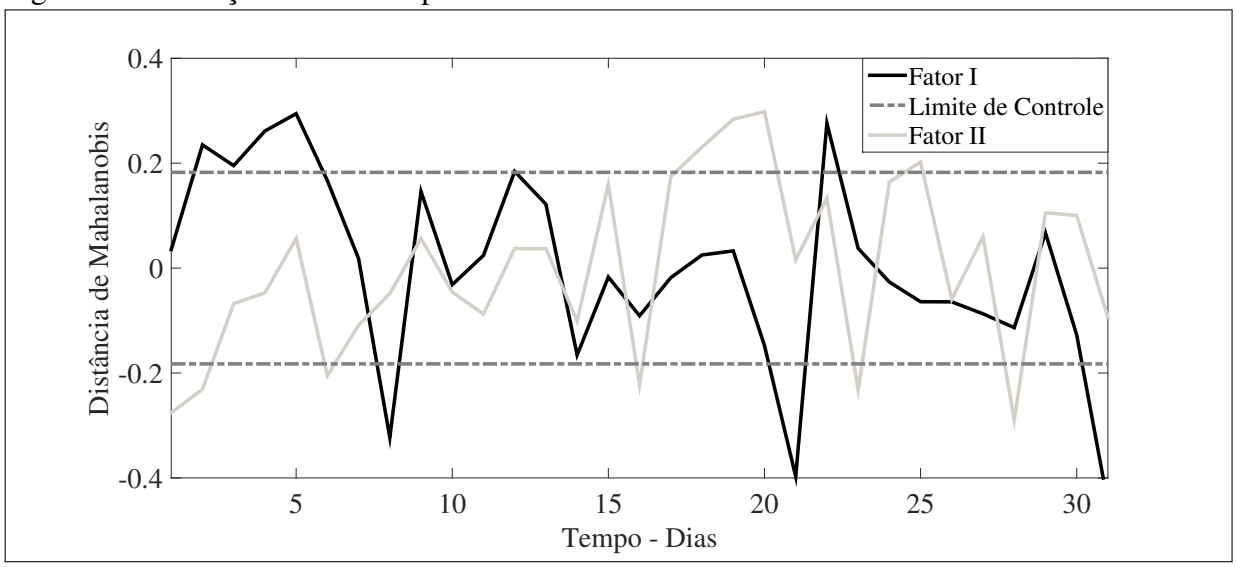
Fonte: elaborado pelo autor (2020).

vez que tais variáveis impactam diretamente na vida dos cidadãos, esses fatores podem revelar padrões de anormalidade que permaneceriam invisíveis frente a uma análise de dados que apenas explorasse a natureza descritiva dos dados.

Por fim, a partir dos fatores extraídos, de acordo com os critérios de variância explicada e Kaiser, aplicamos a função ψ_O e calculamos a distância de Mahalanobis para ambos os fatores extraídos de cada cidade, no intuito de serem utilizados como uma distribuição de referência empírica para estabelecer uma região gráfica de controle para o monitoramento do comportamento das variáveis ambientais. Assim, se os valores da estatística permanecem dentro das regiões de controle, não há evidências de que o processo em análise sofreu algum tipo de mudança. Entretanto, caso os valores em algum instante sejam traçados fora do limiar de controle, há evidências de que o processo sofreu algum tipo de alteração. Portanto, quando os fatores extraídos têm uma interpretação clara, o gráfico estatístico que descreve o comportamento dos fatores fornece uma ilustração visual útil para analisar os perfis das variáveis ambientais das cidades.

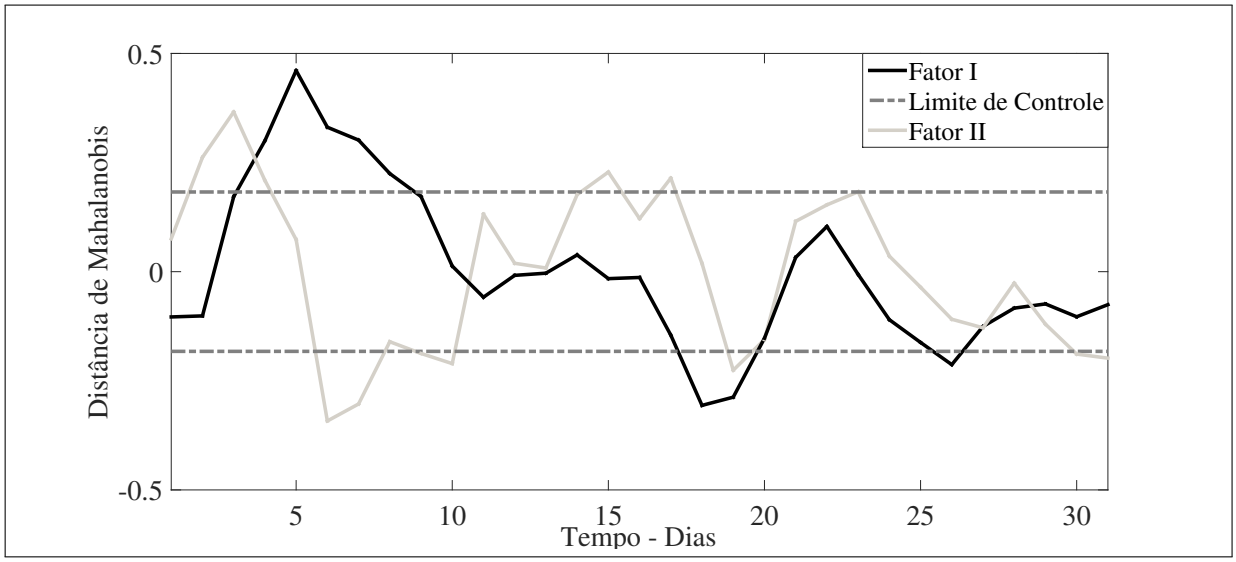
Os resultados da detecção de *outliers* através da distância de Mahalanobis aplicada sobre os dois fatores extraídos de cada cidade são apresentados na Figura 9 e Figura 10. É importante destacar que o parâmetro T_α, foi utilizado nesta pesquisa para delimitar a fronteira de monitoramento, no intuito de inferir qual região do gráfico está fora dos limites de controle. Para a cidade de Elda, verificamos conforme a Tabela 9 que $32,25\%$ e $29,03\%$ de eventos do Fator

Figura 9 – Detecção de *outliers* para a Cidade de Elda



Fonte: elaborado pelo autor (2020).

Figura 10 – Detecção de *outliers* para a Cidade de Rois



Fonte: elaborado pelo autor (2020).

I e Fator II, respectivamente, estavam fora das fronteiras estabelecidas pelo limite de controle. Já a cidade de Rois, para o Fator I e Fator II, respectivamente, os percentuais foram de $25,80\%$ e $41,94\%$ (Tabela 9) de eventos que ultrapassaram os limites de controle. Esses resultados permitem-nos analisar o dia em que determinado evento anormal aconteceu, e assim verificar o instante do ocorrido, podendo a informação ser utilizada pelo órgão público responsável por monitorar padrões ambientais bem como servir de *insights* para as tomadas de decisões futuras por parte dos gestores públicos.

Na Figura 9, T_α é o parâmetro delimitador dos valores da distância de Mahalanobis sobre os dois fatores extraídos. Observamos que na maioria dos dias não há alertas para eventos

Tabela 9 – Testes de Validação da AFE

	Fator I (%)	Fator II (%)
Elda	32,25	29,03
Rois	25,80	41,94

Fonte: elaborado pelo autor (2020).

fora dos limites, contudo o Fator I supera em cerca de 3% no número de *outliers* em relação ao Fator II. Este resultado permite-nos inferir que o Fator Condições Climáticas foi o responsável por influenciar na geração de *outliers* para a cidade de Elda, revelando um importante comportamento discrepante por parte das variáveis ambientais, temperatura e umidade, em determinados dias do mês de Julho. Este fator pode ser útil para revelar padrões de conforto ou desconforto climático da cidade.

Na Figura 10, observa-se novamente determinados instantes em que o valor da distância de Mahalanobis é superior ao limite de controle T_α, sinalizando alertas de ocorrência de eventos discrepantes. Verificamos que apesar do Fator I registrar o maior pico no dia 5 de Julho, o Fator II supera em cerca de 16% na quantidade de *outliers* detectados, sendo o fator responsável por influenciar significativamente na geração de tais eventos. Portanto, o Fator Qualidade do Ar, aponta para o comportamento discrepante das variáveis monitoradas relacionadas a poluição do ar, monóxido de carbono (CO) e dióxido de nitrogênio (NO_2), em determinados períodos do mês de Julho de 2017 da cidade de Rois. A análise deste fator pode ser útil para apontar tendências dos níveis de qualidade do ar, contribuindo para o monitoramento efetivo dos índices de poluição ambiental da cidade.

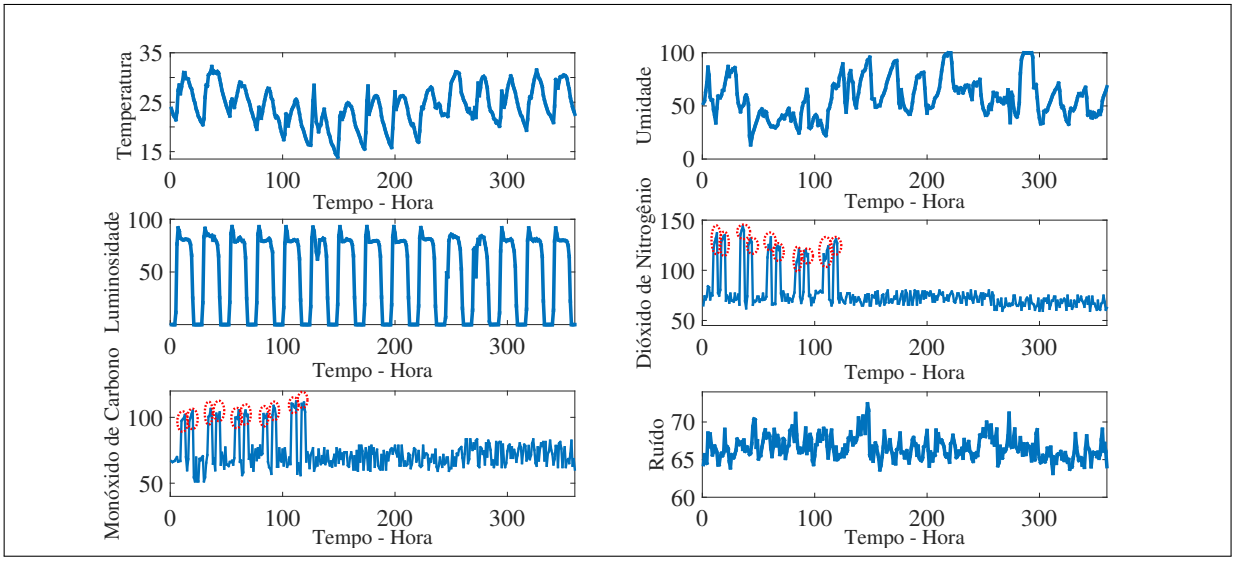
4.2 Resultados da Análise Multidimensional - offline

Os dados coletados para essa abordagem foram organizados em um tensor com dimensões 360 (tempo em dias) $\times$ 6 (variáveis ambientais) $\times$ 4 (cidades monitoradas). Para esta abordagem, as medidas consideradas foram semelhantes à abordagem multivariada, a saber: temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído. As variáveis foram monitoradas durante um período de 15 dias (01/07/2017 - 15/07/2017), para quatro cidades distintas (Elda e Rois da Espanha, Nuremberg da Alemanha, e Tallinn da Estônia), que produziram 360 observações discretizadas em horas. Além disso, para a realização da análise e aplicação dos métodos, padronizamos os dados coletados como *outdoor*. Assim, o modelo de decomposição multidimensional HOSVD é aplicado sobre o tensor de dados a fim de

se obter as relações existentes entre as variáveis de todas as suas dimensões.

Inicialmente realizamos uma análise dos dados com o objetivo de identificar os padrões das séries temporais de dados relacionados a cada cidade analisada. Para a cidade de Elda, a Figura 11 mostra dois conjuntos de picos com duração de quatro horas destacados em vermelho para as duas variáveis relacionadas aos gases poluentes, dióxido de nitrogênio e monóxido de carbono, pela manhã e outras pela tarde. No turno da manhã, observamos uma concentração de picos entre 10h e 13h, enquanto que à tarde há uma concentração de picos entre 17h e 20h nos dias 01/07 a 05/07, com destaque para o dia 05/07 em que o intervalo de desvio durou cinco horas.

Figura 11 – Série Temporal - Cidade de Elda



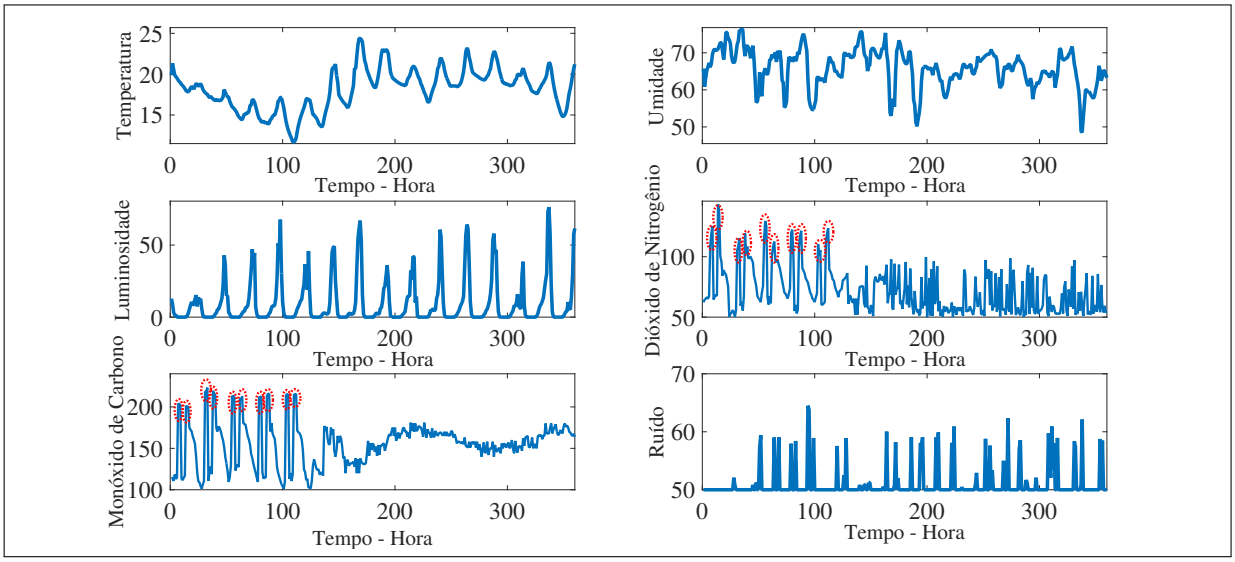
Fonte: elaborado pelo autor (2020).

Na Figura 12, temos a série temporal da cidade de Rois (agora com 30 picos destacados em vermelho, também para as variáveis dióxido de nitrogênio e monóxido de carbono). Semelhante à cidade de Elda, notamos uma concentração de picos nos primeiros cinco dias da manhã e da tarde, mas com duração de três horas. No turno da manhã a concentração dos picos ficou entre 7h e 9h, enquanto à tarde a concentração de picos ocorreu entre 14h e 16h dos dias 01/07 a 05/07.

A figura 13 mostra as séries temporais da cidade de Nuremberg, cujos 30 picos (marcados em vermelho) estão concentrados entre os dias 08/07 a 12/07 de manhã (6h às 8h) e pela tarde (13h às 15h), com duração de três horas também para os gases poluentes, dióxido de nitrogênio e monóxido de carbono.

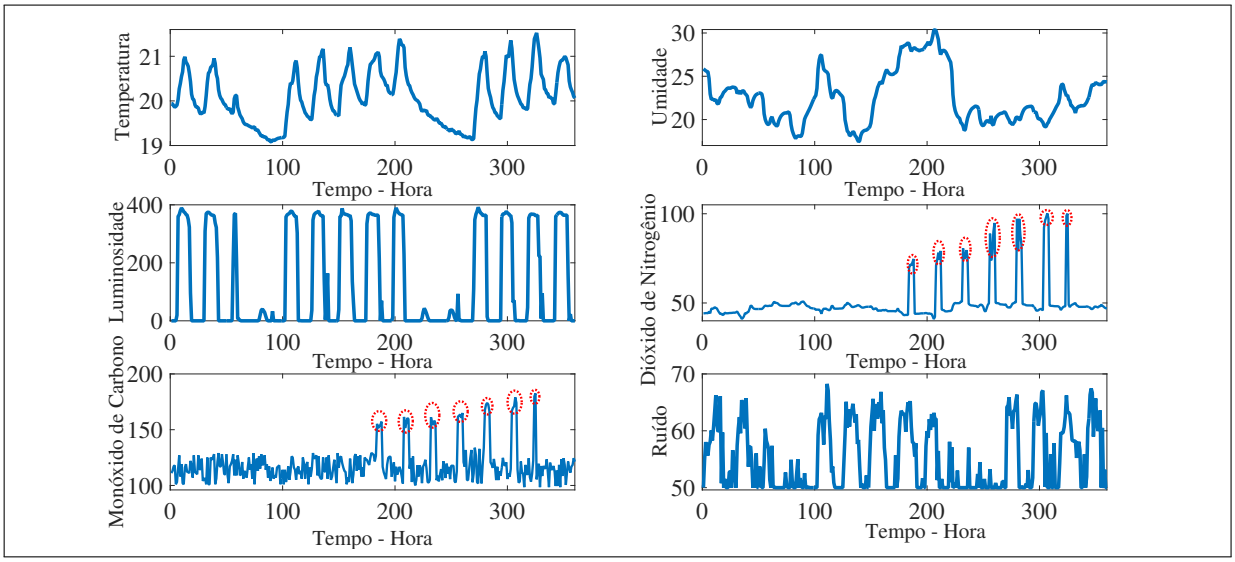
Por fim, a Figura 14 apresenta as séries temporais da cidade de Tallin, cujos 32 picos

Figura 12 – Série Temporal - Cidade de Rois



Fonte: elaborado pelo autor (2020).

Figura 13 – Série Temporal - Cidade de Nuremberg

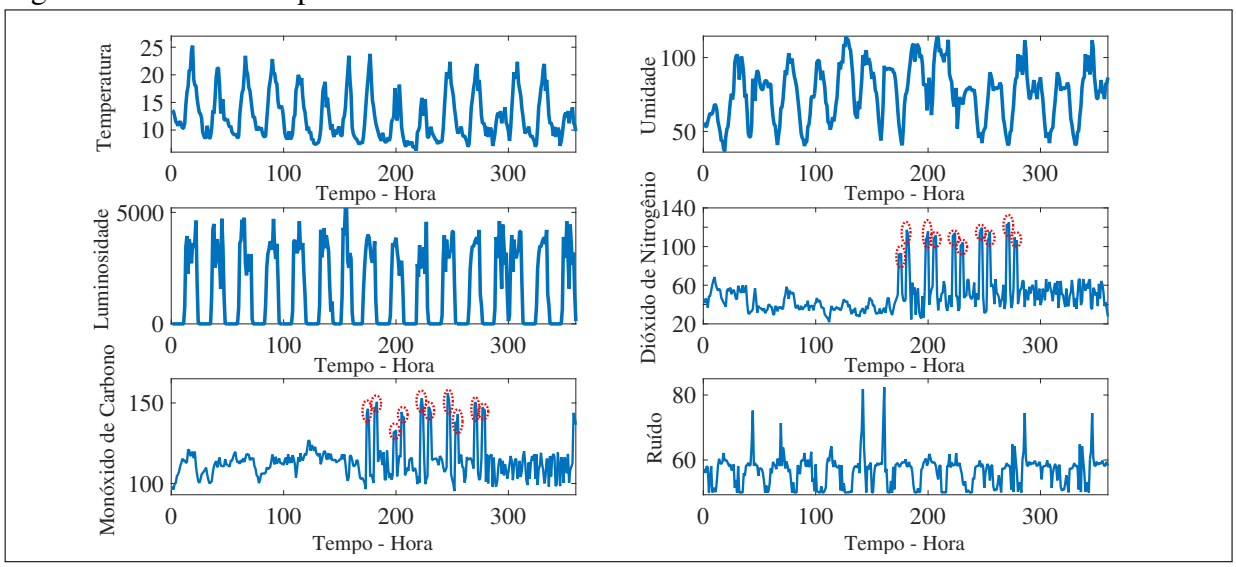


Fonte: elaborado pelo autor (2020).

(destacados em vermelho) estão concentrados nos últimos dias da série à tarde (08/07 à 14/07), com cinco horas de duração novamente para as variáveis, dióxido de nitrogênio e monóxido de carbono. Os picos da tarde estão concentrados entre as 16h às 20h, destacando os dois últimos desvios que apareceram isoladamente no décimo quarto dia. Como pode ser visto em todos os gráficos das Figuras 11, 12, 13 e 14, vários aumentos constantes e acentuados declínios ocorrem especialmente para as variáveis gases poluentes, dióxido de nitrogênio e monóxido de carbono, apresentando desvios periódicos em relação às outras variáveis das séries temporais. Nesta perspectiva, as demais variáveis não apresentaram desvios significativos no domínio do tempo.

De acordo com a função ψ_H apresentada anteriormente, calculamos os fatores

Figura 14 – Série Temporal - Cidade de Tallin



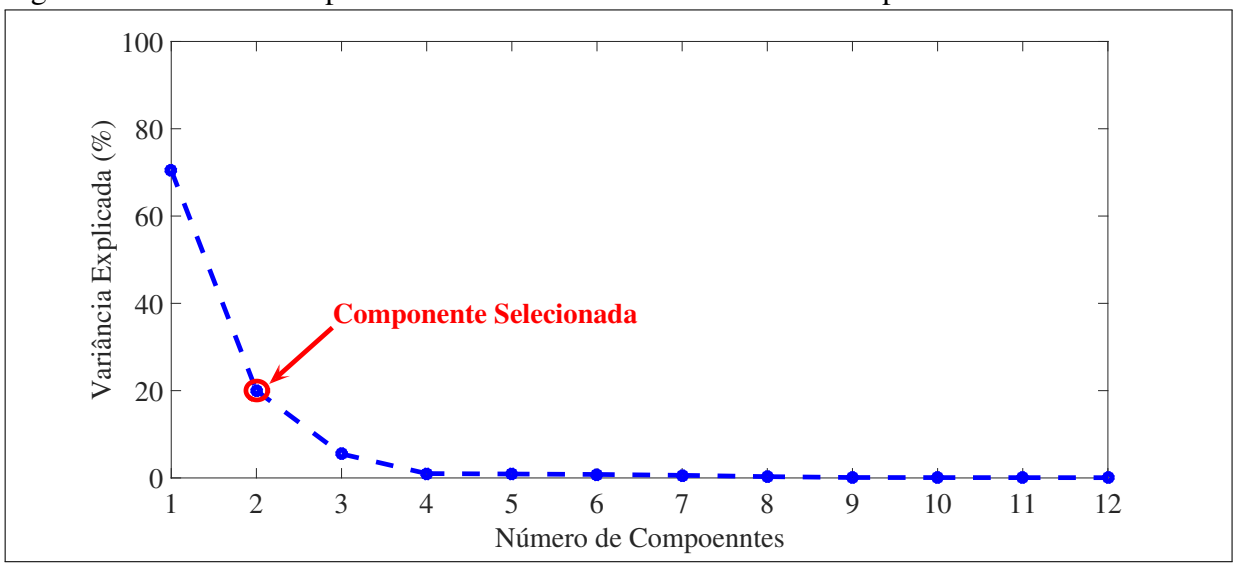
Fonte: elaborado pelo autor (2020).

matriciais $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ e o tensor núcleo $\underline{\mathbf{S}}$. A decomposição nos permite identificar diferentes padrões através da interpretação dos componentes ou fatores selecionados nas matrizes fatores. Esses componentes resumem os padrões de informações em cada dimensão. Uma maneira de selecionar o número apropriado de componentes é calcular a porcentagem cumulativa dos valores próprios ou a porcentagem acumulada da variação explicada na respectiva dimensão de interesse.

Para investigar a dinâmica temporal das medidas detectadas e a relação entre elas e as variáveis ambientais, analisamos a dimensão temporal e as variáveis detectadas. As Figuras 15 e 16 apresentam a porcentagem de variação explicada para os fatores da matriz $\mathbf{U}_1$, relacionados à dimensão do tempo (Figura 15) e $\mathbf{U}_2$, relacionado às variáveis ambientais medidas (Figura 16). Com uma porcentagem de 90,5 % de variação cumulativa para os dois primeiros componentes da matriz de fatores $\mathbf{U}_1$ e 84,5 % para a matriz de fatores $\mathbf{U}_2$ o modelo atinge uma variação explicada satisfatória.

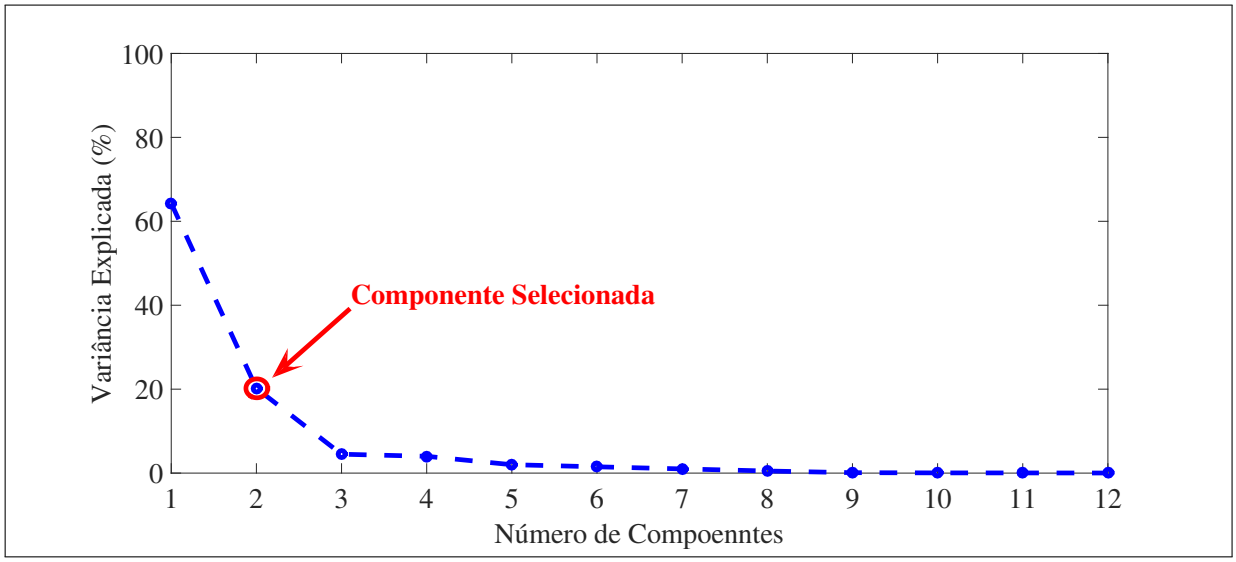
A Figura 17 mostra os perfis dos padrões temporais dos componentes selecionados na matriz fatorial $\mathbf{U}_1$. Observamos um padrão cíclico no comportamento de ambos os componentes, o principal componente I (CP I) e o principal componente II (CP II). O CP I apresenta valores máximos padrão dentro de uma periodicidade que ocorre a cada 12 horas, retornando logo após a origem do sistema de coordenadas. Assim, a partir da interpretação da PC I, podemos inferir que mudanças nos padrões das variáveis sensoriais ambientais ocorrem a cada metade do período do dia, retornando logo após seu estado normal. O CP II apresenta um padrão de periodicidade semelhante quando comparado ao CP I. Embora neste componente haja uma alternância entre picos máximos e mínimos de valores, permitindo inferir que nem sempre o

Figura 15 – Variância explicada da matriz de fatores - dimensão temporal



Fonte: elaborado pelo autor (2020).

Figura 16 – Variância explicada da matriz de fatores - dimensão variáveis ambientais



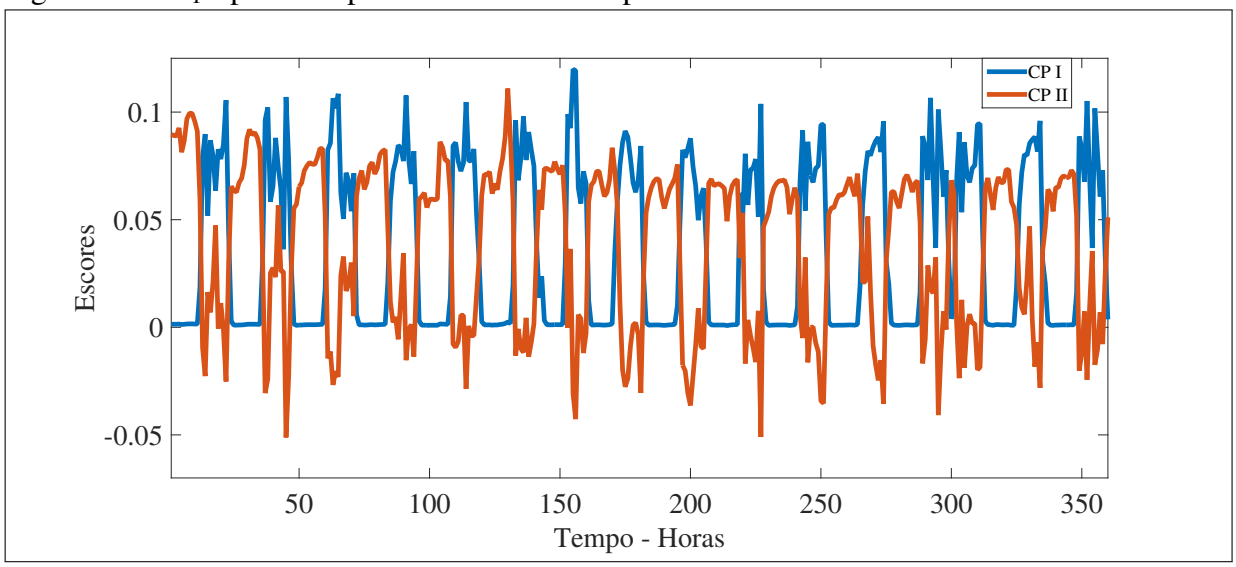
Fonte: elaborado pelo autor (2020).

padrão de normalidade das variáveis detectadas no ambiente é atingido novamente, esses padrões podem ir de uma extremidade à outra.

A fatoração também nos permite verificar as interações correspondentes em outros modos. A Figura 18 mostra o padrão de relacionamento das variáveis medidas e dos componentes selecionados do modelo multidimensional. As variáveis temperatura e umidade destacam-se com os valores mais altos para CP II, enquanto que para CP I as mesmas variáveis se destacam, permitindo inferir que as variáveis temperatura e umidade estão sempre alternando entre padrões cíclicos de comportamento.

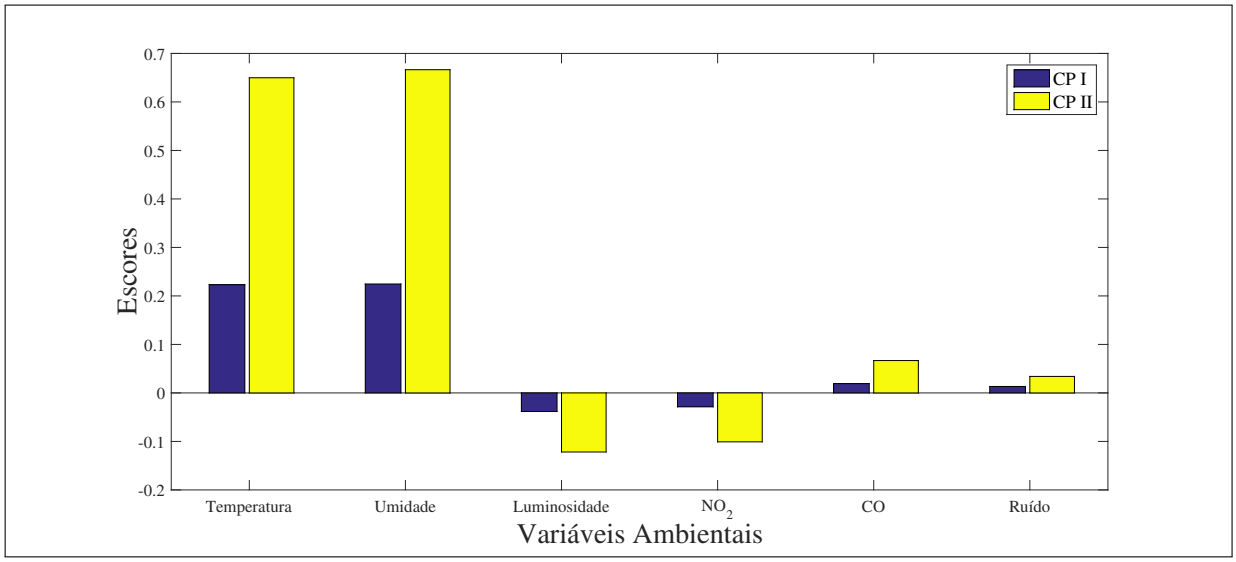
Após gerar os fatores matriciais da função ψ_H, aplicamos o algoritmo de *clustering*

Figura 17 – $\mathbf{U}_1$ - perfis de padrões do modo temporal



Fonte: elaborado pelo autor (2020).

Figura 18 – $\mathbf{U}_2$ - perfis de padrões do modo variáveis ambientais

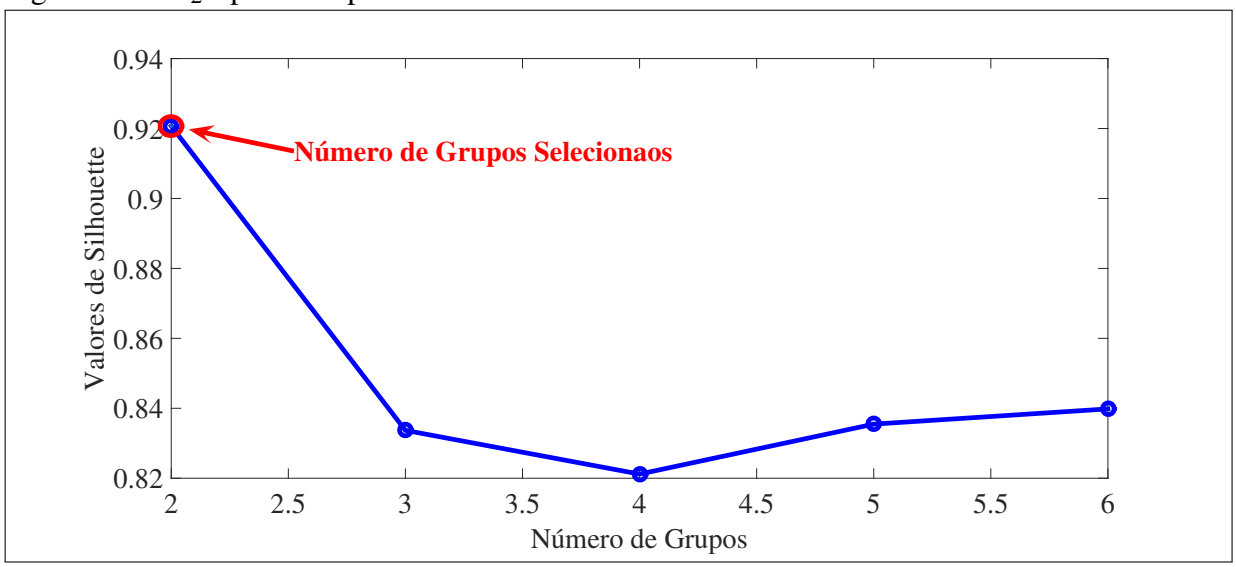


Fonte: elaborado pelo autor (2020).

aos componentes selecionados da matriz $\mathbf{U}_1$ do modelo multidimensional. Para estimar o desempenho dos algoritmos de *clustering* para os *clusters* gerados, adotamos o valor do coeficiente Silhouette (MUR *et al.*, 2016) para estimar a medida padrão para verificar a qualidade dos *clusters* e usá-la para determinar a melhor formação de *cluster*. A Figura 19 abaixo mostra as curvas Silhouette para o CP I e a Figura 20 para a CP II, que avaliam a qualidade de cada grupo. De acordo com a Figura 19, enquanto para o CP I o coeficiente mais alto de Silhouette obteve um valor de $0,92$ para dois *clusters*, para o CP II (Figura 20) o valor cai para $0,86$ para a mesma quantidade de *clusters*. Quando aumentamos o número de *clusters*, o coeficiente diminui, com um pequeno aumento de quatro *clusters*. Assim, pelo coeficiente Silhouette, temos dois *clusters*

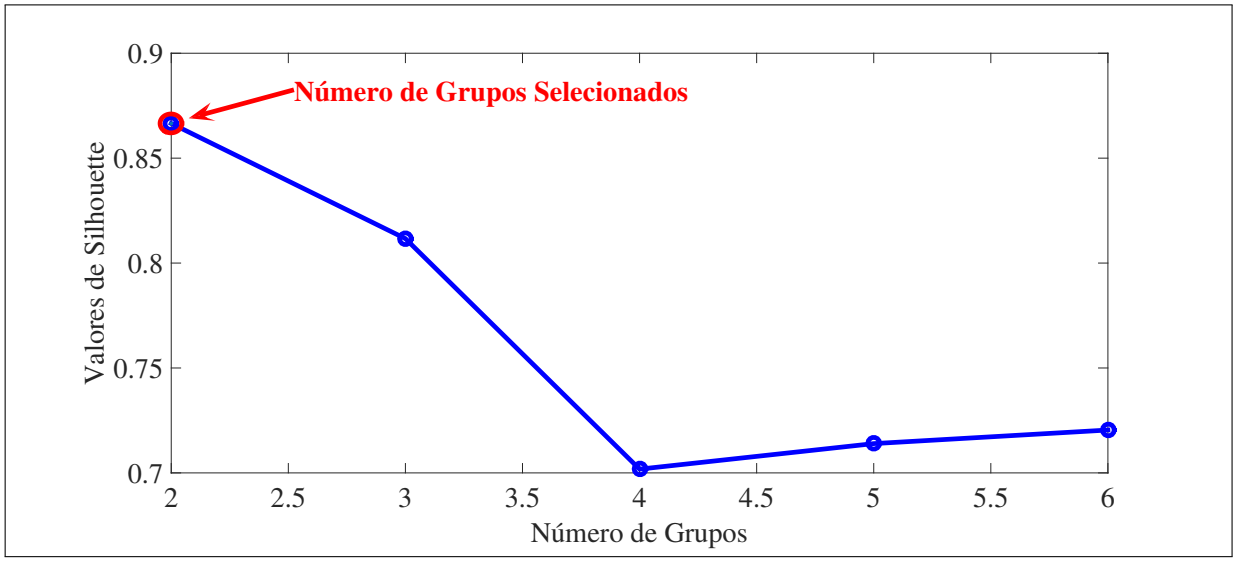
gerados pela função ψ_C para o CP I e o CP II.

Figura 19 – $\mathbf{U}_2$ - perfis de padrões do modo variáveis ambientais



Fonte: elaborado pelo autor (2020).

Figura 20 – $\mathbf{U}_2$ - perfis de padrões do modo variáveis ambientais



Fonte: elaborado pelo autor (2020).

Por fim, identificamos um conjunto de eventos isolados (Φ) de padrões regulares como resultado da função ψ_O executada sobre ψ_H combinada com as funções ψ_C. ψ_O calcula a distância de Mahalanobis sobre cada *cluster* gerado a partir do CP I e CP II. A Tabela 10 apresenta a distribuição de *cluster* para cada componente principal extraído do método HOSVD combinado com o algoritmo *k*-Means. Nesta tabela, temos a divisão de cada *cluster* gerada no respectivo componente principal selecionado. No primeiro componente principal (CP I), identificamos uma classificação de 360 observações em dois *clusters*; no *Cluster* 1 (com 155

eventos), o método proposto detecta 41 *outliers*, resultando em uma porcentagem de 26,45% de outliers para esse *cluster*. Por outro lado, para o *Cluster* 2 com um número mais substancial de eventos em *cluster* (205 eventos), o método detectou 30 *outliers* (produzindo uma porcentagem de 26,45% dos outliers detectados). Para o segundo componente principal (CP II), as 360 observações também classificamos em dois *clusters*, onde para o *Cluster* 1 (com 150 eventos), o método detectou 30 valores extremos, produzindo uma porcentagem de 20,00% de valores extremos detectados neste *cluster*. Por outro lado, para o *Cluster* 2 com 210 eventos em cluster, o método detectou 32 valores discrepantes (produzindo uma porcentagem de 15,24% dos valores discrepantes detectados).

Tabela 10 – Distribuição de clusters - média de HOSVD + k

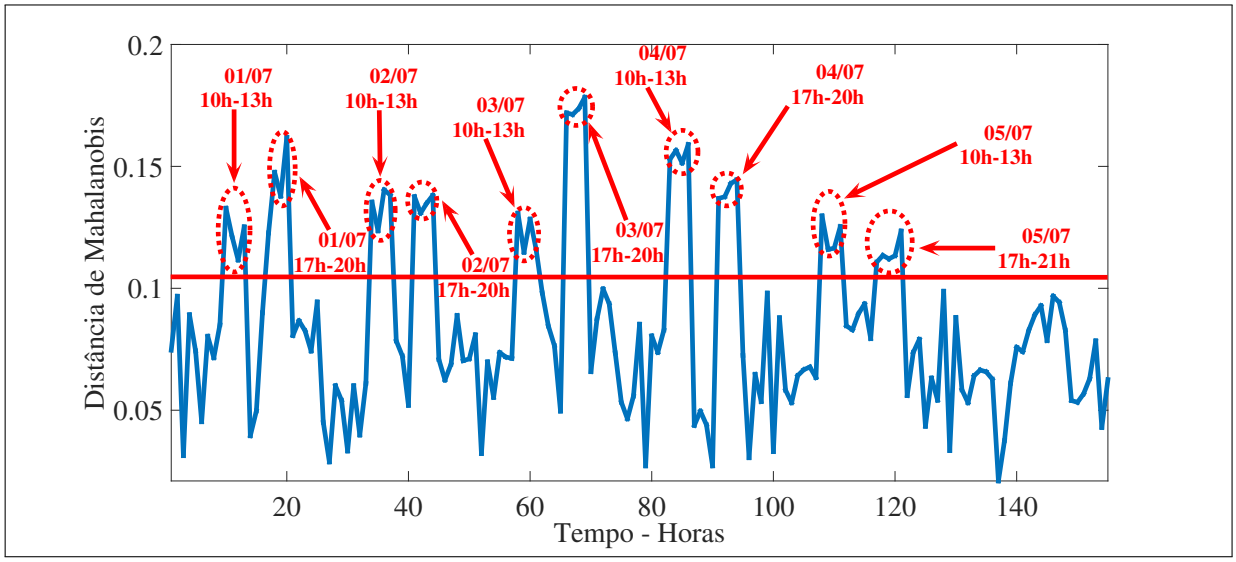
Comp.	Clusters	#events	$\lvert\Phi\rvert$	Percentage
	0	360	-	-
CP I	1	155	41	26.45%
	2	205	30	14.63%
	0	360	-	-
CP II	1	150	30	20.00%
	2	210	32	15.24%

Fonte: elaborado pelo autor (2020).

As Figuras 21, 22, 23 e 24 mostram os eventos para cada *cluster* com seus respectivos componentes principais. Para o CP I, os *clusters* I e II apresentam um padrão cujos valores extremos estão nos primeiros cinco dias do mês. No cluster I (Figura 21), que concentra 155 eventos (de acordo com a Tabela 10), 41 outliers (em vermelho) são identificados e divididos em dois momentos distintos: manhã (10h - 13h) e tarde (17h - 20h) entre 01/07 e 05/07, com destaque para o dia 05/07, que cobre uma janela de tempo estendida (17h - 21h). Ao verificar os dados originais, notamos que esses momentos ocorreram com as variáveis monóxido de carbono e dióxido de nitrogênio na cidade de Elda (Figura 11). Isso permite inferir que o *cluster* I do CP I está relacionado com os eventos ambientais ocorridos na cidade de Elda. Além disso, uma vez que o mesmo padrão encontrado nos dados transformados é observado nos dados originais, é possível observarmos também o sucesso do método na economia do custo computacional. Desta forma, como os dados de séries temporais apresentam um comportamento dinâmico no contexto do monitoramento ambiental urbano, o sucesso da proposta apresentada neste tese pode ser estendida para um conjunto de dados qualquer com um maior número de variáveis. No *cluster* II (Figura 22), foram classificados 205 eventos (como na Tabela 1), com 30 outliers destacados em vermelho (conforme Tabela 10) identificados e também divididos em dois grupos distintos.

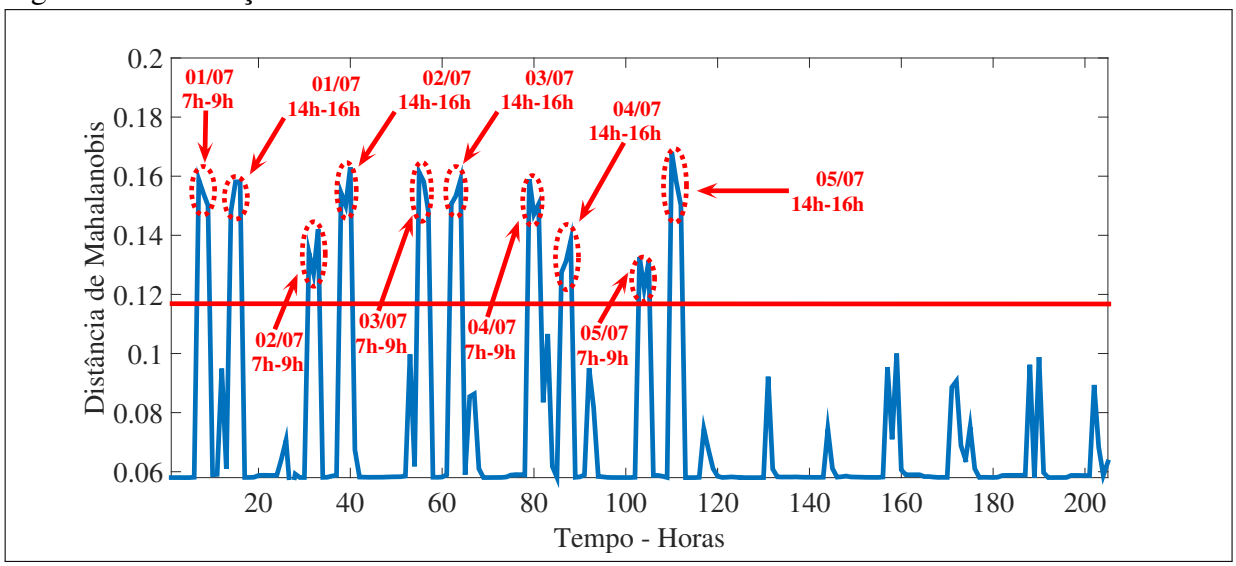
momentos, manhã (7h - 9h) e tarde (14h - 16h). A Figura 12 mostra os gases poluentes da cidade de Rois, o que significa que o *cluster* II está relacionado com os eventos ambientais ocorridos na cidade de Rois.

Figura 21 – Detecção de outliers - CP I - Cluster I



Fonte: elaborado pelo autor (2020).

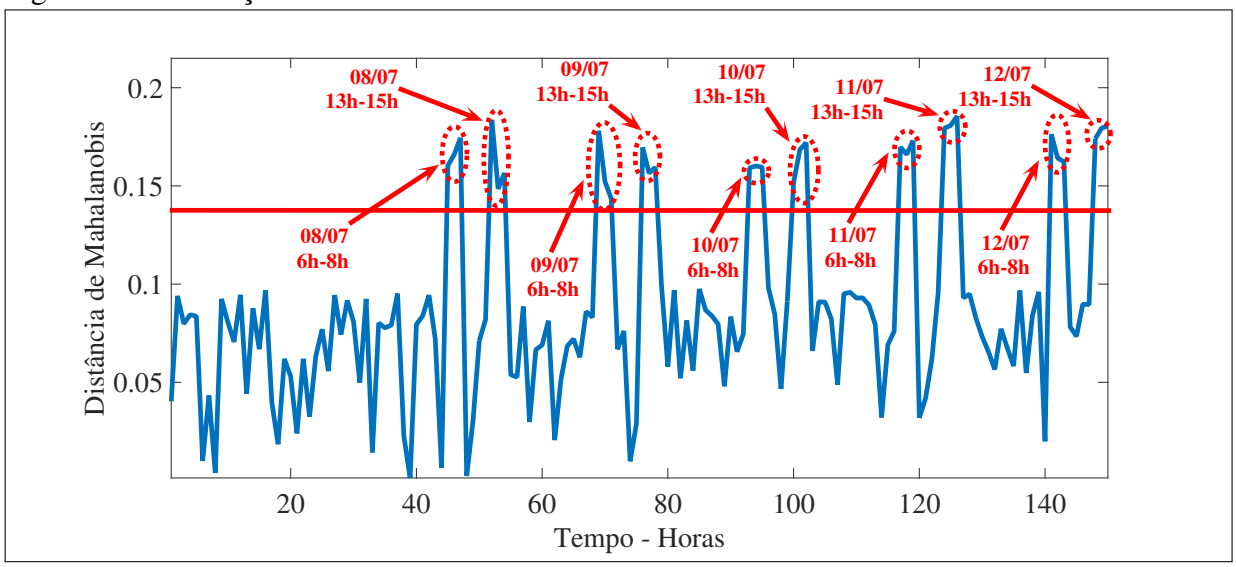
Figura 22 – Detecção de outliers - CP I - Cluster II



Fonte: elaborado pelo autor (2020).

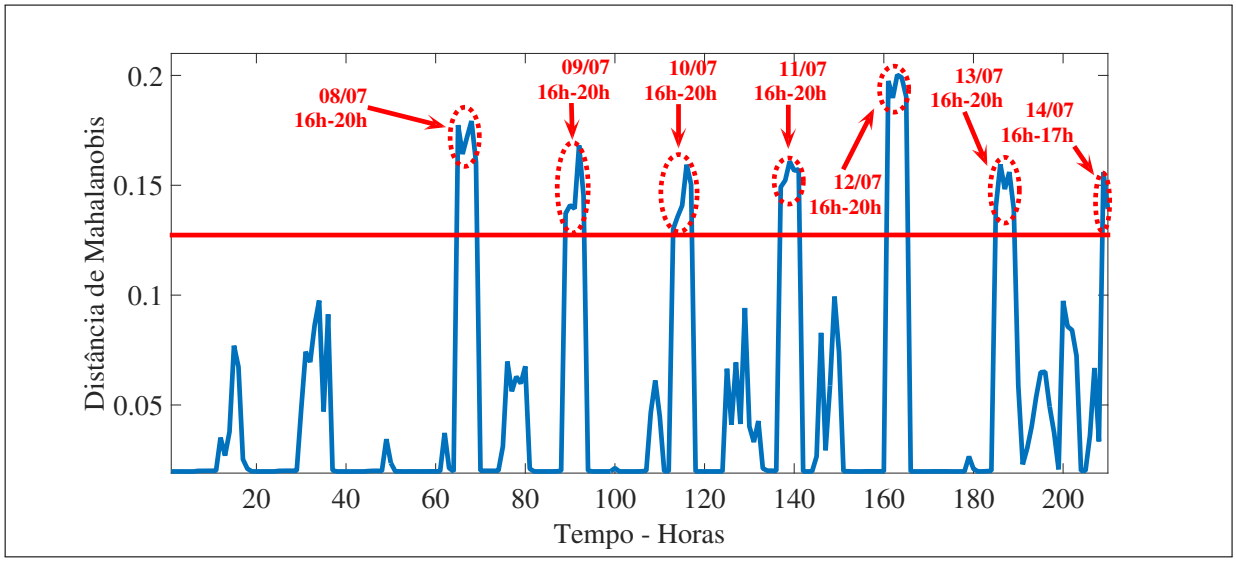
Para o CP II, os agrupamentos I e II apresentam um padrão cujos valores extremos estão entre 08/07 e 14/07. No *cluster* I (Figura 23), que inclui 150 eventos (de acordo com a Tabela 10), 30 outliers (marcados em vermelho) são identificados entre 08/07 e 12/07, em as manhãs (6h - 8h) e as tardes (13h - 15h). Ao verificar as séries temporais dos dados, observamos que esses períodos correspondem aos desvios relacionados às variáveis de monóxido de carbono

Figura 23 – Detecção de outliers - CP II - Cluster I



Fonte: elaborado pelo autor (2020).

Figura 24 – Detecção de outliers - CP II - Cluster II



Fonte: elaborado pelo autor (2020).

e dióxido de nitrogênio na cidade de Nuremberg (Figura 13). Isso permite inferir que o *cluster* I do CP II se encaixa nos eventos de Nuremberg. No *cluster* II (Figura 24), classificamos 210 eventos (de acordo com a Tabela 10), com 32 valores extremos marcados em vermelho, entre os dias 08/07 a 14/07 à tarde (16h - 20h), destacando os dois últimos desvios que apareceram isoladamente no décimo quarto dia (16h - 18h). Novamente, analisando as séries temporais dos dados, os desvios do *cluster* II ocorreram nos gases poluentes da cidade de Tallin (Figura 14), o que nos permite concluir que o *cluster* II se encaixa nos eventos de Tallin.

Calculamos a distância de Mahalanobis, através da aplicação da função ψ_O, em cada *cluster* gerado nos respectivos componentes principais com um valor limite de decisão

T_α calculado para cada *cluster*. Esse limite de decisão captura a estrutura geral dos eventos regulares, indicando os dados que, acima desse limite, divergem da normalidade do conjunto de observações temporais. Portanto, os valores limite encontrados para cada *cluster* são destacados com a linha vermelha nos gráficos das Figuras 21, 22, 23 e 24, a saber: CPI - *Cluster* I, $T_\alpha = 0,11$ (cidade de Elda); CPI - *Cluster* II, $T_\alpha = 0,12$ (cidade de Rois); CPII - *Cluster* I, $T_\alpha = 0,14$ (cidade de Nuremberg); CPII - *Cluster* II, $T_\alpha = 0,13$ (cidade de Tallin).

Além disso, os limites nos permitem quantificar o número de eventos anormais em cada cluster, bem como detectar padrões acima desses limites. Valores acima do limite calculado são discrepantes. É importante destacar que devemos detectar ou remover os valores discrepantes, porque um discrepante pode ser um evento real, por exemplo, um evento que indica um tsunami ou terremoto.

Entre os valores discrepantes encontrados, os mais significativos correspondem a eventos relacionados a momentos em que variáveis específicas, como dióxido de nitrogênio e monóxido de carbono, oscilaram abruptamente, provavelmente causadas por alguns erros de medição dos sensores ou pela ocorrência de algum fenômeno físico na região . Verificamos esse comportamento nos clusters dos dois componentes principais. Os outros eventos anormais estão relacionados a pequenas variações das variáveis em intervalos de tempo específicos.

4.3 Resultados da Análise Multidimensional - online

Nesta seção, ilustramos nosso método de detecção de *outlier online*, conforme apresentado com o objetivo de detectar *outliers* das variáveis ambientais monitoradas dos espaços urbanos das cidades. Uma comparação de desempenho foi realizada com base em simulações e os resultados são comparados com os resultados obtidos por uma pesquisa nossa (SOUZA *et al.*, 2019a), publicada anteriormente.

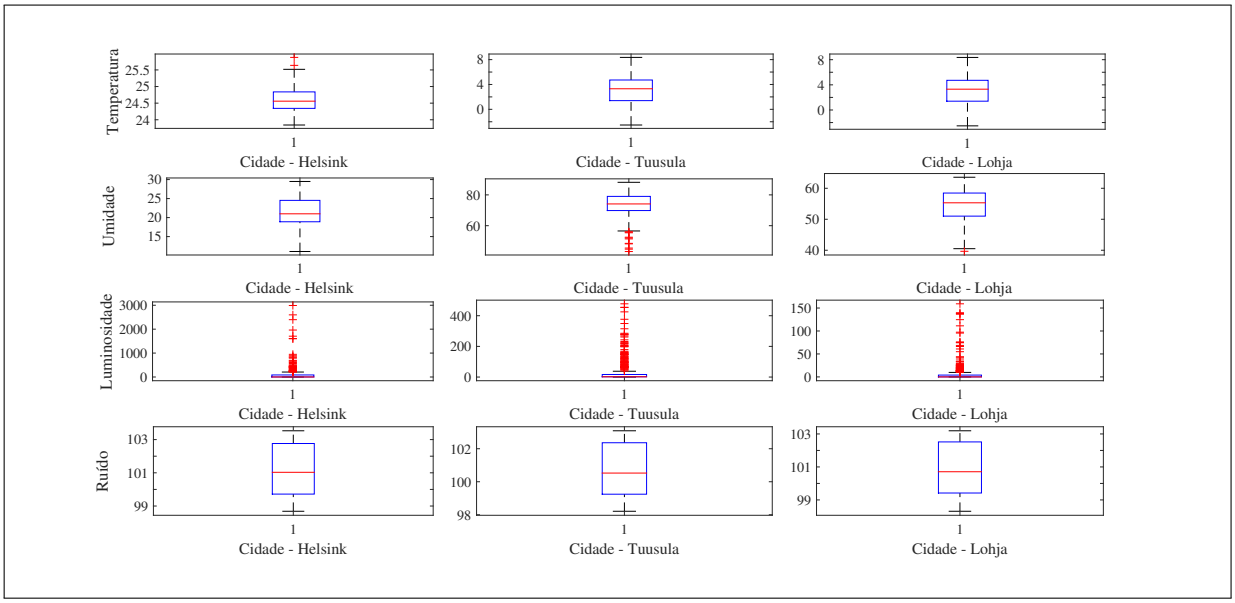
Os dados coletados para essa abordagem foram organizados em um tensor com dimensões 360 (tempo em dias) $\times$ 6 (variáveis ambientais) $\times$ 4 (cidades monitoradas). Para esta abordagem, as medidas consideradas foram semelhantes à abordagem multivariada, a saber: temperatura, umidade, luminosidade, monóxido de carbono, dióxido de nitrogênio e níveis de ruído. As variáveis foram monitoradas durante um período de 15 dias (01/07/2017 - 15/07/2017), para quatro cidades distintas (Elda e Rois da Espanha, Nuremberg da Alemanha, e Tallinn da Estônia), que produziram 360 observações discretizadas em horas. Portanto, através de uma janela deslizante, é gerado um modelo de observação com o fluxo de dados mais recentes, onde

o modelo é atualizado regularmente.

Os *boxplots* dos valores reais coletados pelos sensores da plataforma Smart Citizen das cidades de Helsinque, Tuusula e Lohja são apresentados nas Figuras 25 e 26. Foram considerados sensores externos por um período de 16 dias (1 de dezembro a 16 de dezembro de 2018), totalizando um banco de 381 horas de monitoramento das oito variáveis ambientais. Esses locais foram selecionados porque oferecem nós de sensores *online* onde as medições podem ser realizadas em tempo real sem dados faltantes.

Observando o padrão de comportamento dos dados, percebemos que as variáveis luminosidade e material particulado (PM 1, PM 10 e PM 2.5) foram as que apresentaram valores mais discrepantes em relação às demais variáveis. Esse comportamento da variável luminosidade é explicado pela alternância entre picos e vales de seus valores, pois, ao entardecer, a luminosidade nas cidades é reduzida consideravelmente. Por outro lado, para as variáveis PM 1, PM 10 e PM 2,5, o comportamento discrepante pode estar relacionado à áreas com poluição atmosférica significativa.

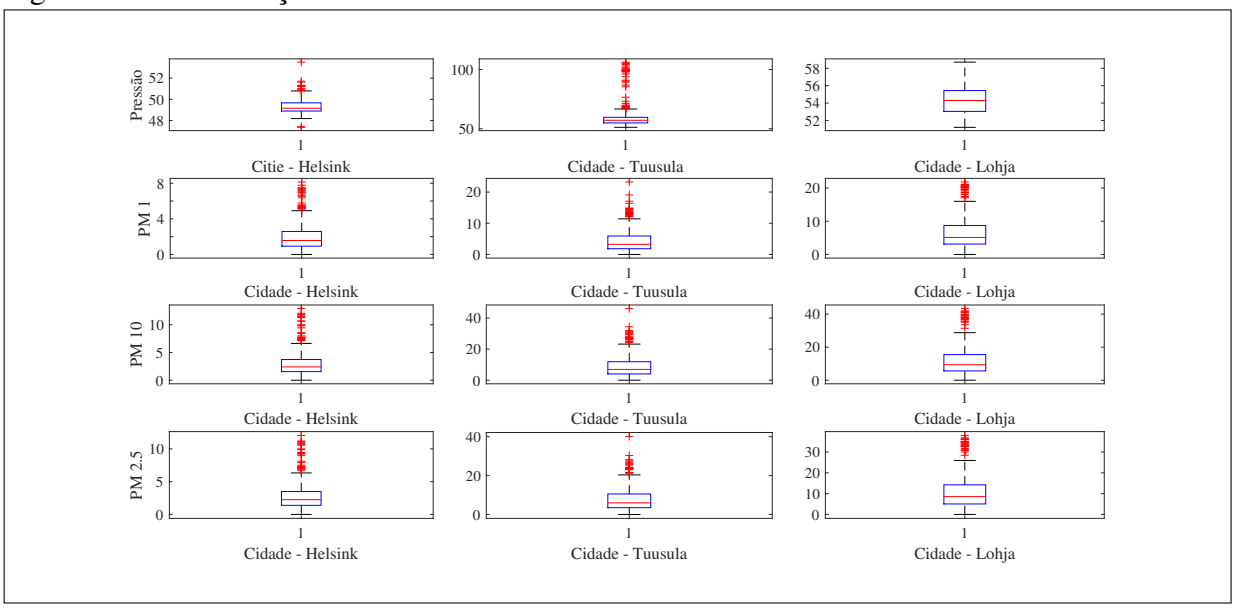
Figura 25 – Dados Originais



Fonte: elaborado pelo autor (2020).

A utilização da janela deslizante permite processar os dados em lotes menores por vez, geralmente para representar uma vizinhança de pontos nos dados. Portanto, usando uma banda fixa de 24 horas, atualizamos os dados a cada hora, ou seja, quando novos dados chegaram em um determinado instante, o método foi atualizado. A escolha de atualizar a janela deslizante a cada hora em uma janela fixa de 24 horas foi decidida, pois foi determinado que, com um aumento

Figura 26 – continuação.



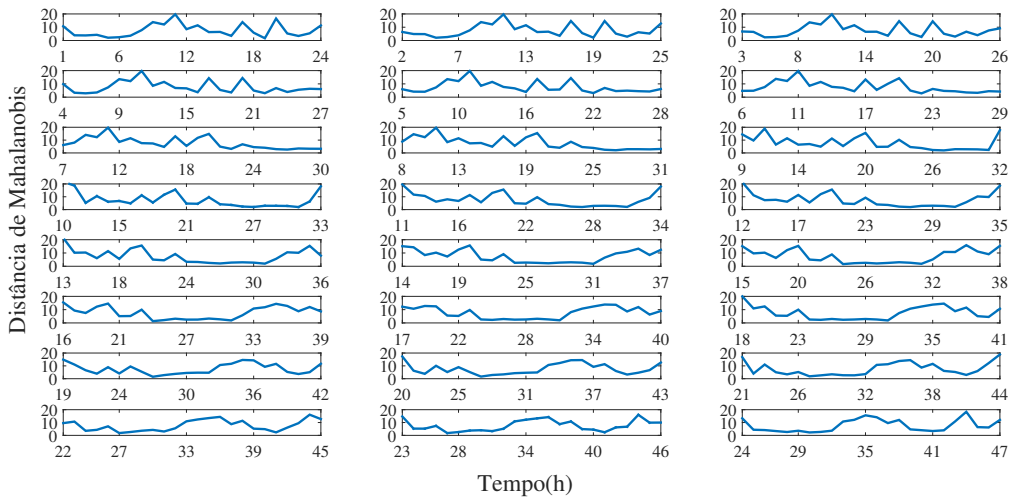
Fonte: elaborado pelo autor (2020).

nesse período, os picos mais curtos podem ser eliminados e os outliers podem ser camuflados. Ou seja, se a janela for muito grande, a janela poderá conter informações desatualizadas e a precisão do modelo diminuirá (NGUYEN *et al.*, 2015). Assim, a partir dos resultados da decomposição tensorial, analisamos a dimensão temporal uma vez que focamos na análise das séries temporais do modelo.

A Figura 27 apresenta o modelo de monitoramento online no qual, a cada hora, ou seja, a cada momento em que novos dados chegam, a decomposição do modelo multidimensional é atualizada. Assim, nossa janela deslizante se move ao longo de uma janela fixa de 24 horas e por toda essa janela, os dados são atualizados com uma granularidade de 1 hora até que toda a série temporal seja contemplada, onde observamos ao longo do processo a dinâmica temporal dos dados e seus efeitos no restante da faixa da janela deslizante enquanto ela se move. Por exemplo, a Figura 27 apresenta o monitoramento online para o primeiro e o segundo dia e a mudança na dinâmica do comportamento temporal à medida que novos dados são incorporados ao modelo multidimensional. Para um melhor entendimento, o primeiro *subplot* da Figura 27 mostra as primeiras 24 horas consideradas na análise do modelo, depois a janela se move (movendo-se para o segundo dia) a cada 1 hora e, quando um novo dado entra, o último dado da série temporal é descartado e, então, o modelo é atualizado. Esse processo é repetido até que toda a série temporal seja contemplada.

Para uma melhor visualização, a Figura 28 expande o primeiro *subplot* da Figura 27, onde observamos o monitoramento online (com a janela deslizante) e adicionamos uma compa-

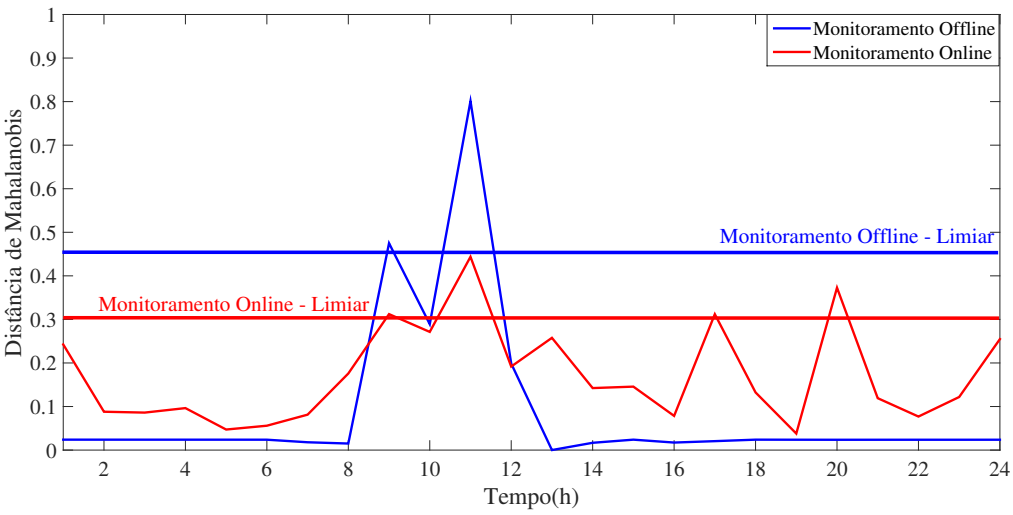
Figura 27 – Monitoramento online - Dia 1 e Dia 2.



Fonte: elaborado pelo autor (2020).

ração com o monitoramento *offline* (monitorando 24 horas), além de estabelecer os limites de detecção para ambas as abordagens. Assim, observamos que o monitoramento online através da janela deslizante (linha vermelha) identifica um número maior de picos, apontando para uma maior variação na dinâmica dos dados do que em relação ao monitoramento *offline* que identifica apenas um vale mais expressivo com seus respectivos dois picos (linha azul). Como os pontos do vale são cercados por dois vizinhos maiores (imediatamente anterior e posterior), esse resultado corrobora à hipótese de que o monitoramento *offline* não representa uma boa aproximação dos dados para monitoramento, ao contrário do monitoramento online que representa uma melhor aproximação desses pontos, revelando a granularidade dos outliers (NGUYEN *et al.*, 2015).

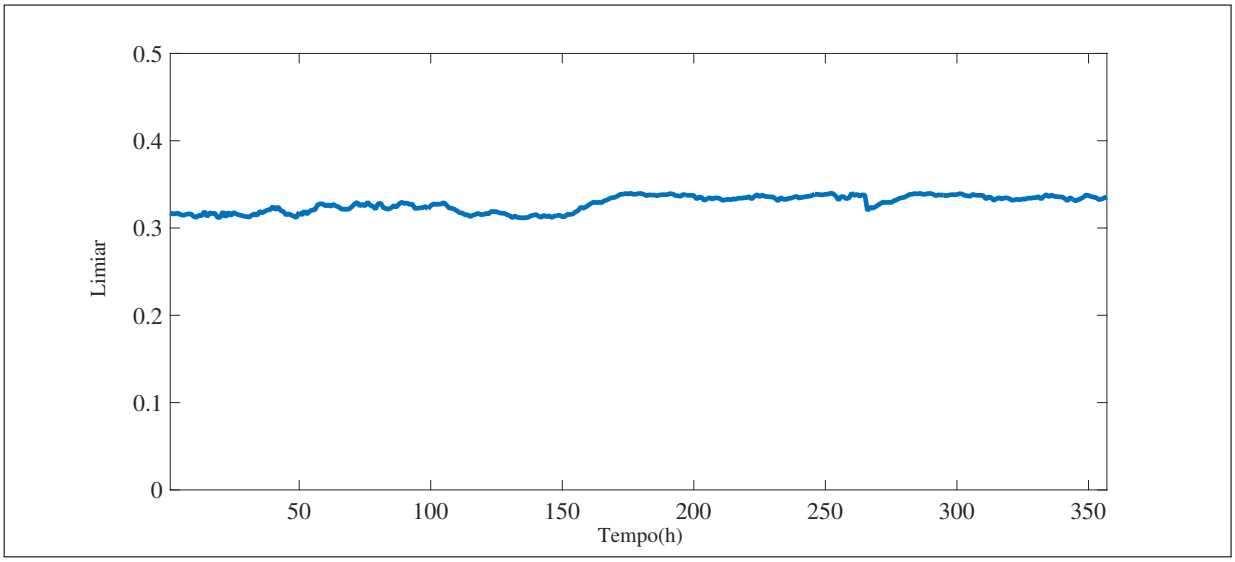
Figura 28 – Monitoramento online *versus* monitoramento offline - Dia 1.



Fonte: elaborado pelo autor (2020).

Outro aspecto importante está na análise da variação do limiar que é atualizado de acordo com o progresso da janela deslizante, ou seja, na perspectiva do monitoramento online gerando um limiar dinâmico à cada instante. Por outro lado, da perspectiva do monitoramento offline, é observado um limiar estático ao longo de toda a janela monitorada. Assim, a Figura 29 mostra a dinâmica do limiar na identificação dos outliers, em que flutuações são observadas ao longo da série temporal para o monitoramento online. A análise de um limiar dinâmico na detecção de outlier sob a perspectiva do monitoramento ambiental urbano ainda é escassa na literatura. No entanto, a detecção de outliers da perspectiva do tráfego de rede tem sido amplamente estudada. Por exemplo, é observado em alguns trabalhos como (BHUYAN *et al.*, 2015; JUN *et al.*, 2014) que limites dinâmicos melhoram a precisão de detecção de valores extremos, enquanto limites estáticos resultam em menor precisão de detecção. No contexto do monitoramento ambiental urbano, esse fato pode ser verificado quando se compara o limiar dinâmico utilizado nesta tese com o limiar estático utilizado em uma pesquisa nossa (SOUZA *et al.*, 2019a), publicada anteriormente, que apresentou menor precisão.

Figura 29 – Limiar do monitoramento online.

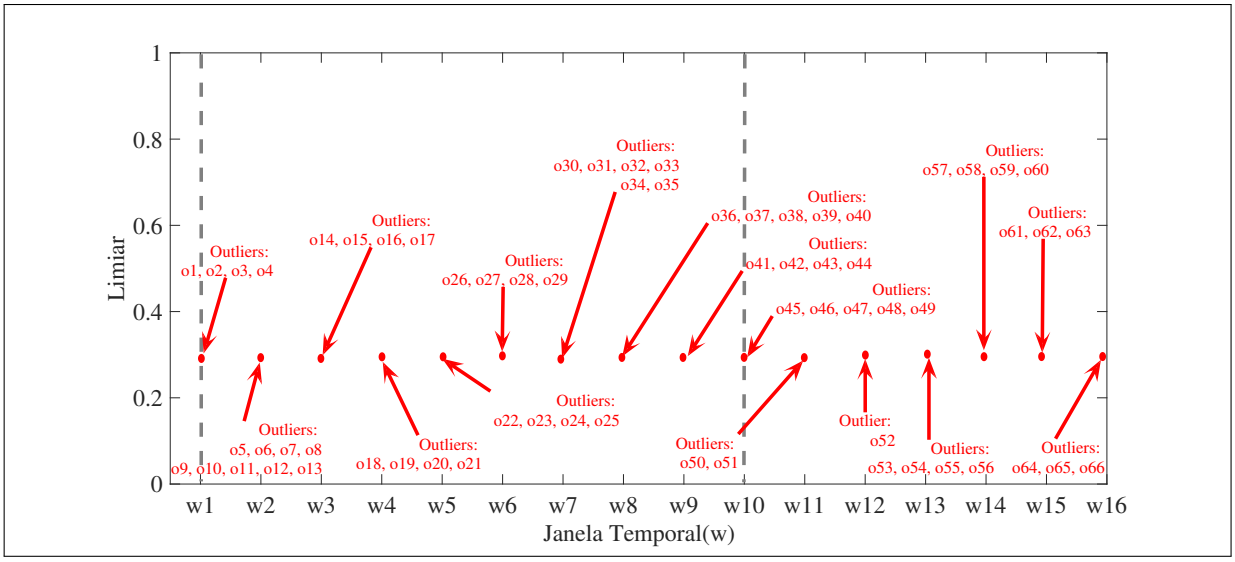


Fonte: elaborado pelo autor (2020).

À medida que a janela deslizante de um fluxo de dados se move e os dados antigos expiram e novos dados chegam, é possível descobrir os outliers dos fluxos de dados a qualquer momento. Nesta perspectiva, a Figura 30 apresenta as variáveis que estavam ao longo de todas as janelas (discriminadas no eixo x), a partir do fluxo de dados de entrada no modelo. Além disso, qual a janela que exibe uma quantidade maior de outliers e qual é a janela que recebe o menor número de outliers. Também podemos considerar uma série de janelas deslizantes e

observar a dinâmica do arranjo desses valores extremos ao longo do intervalo, de acordo com a evolução temporal. Considere a janela w2 como a que apresentou o maior número de outliers, na qual observamos um total de nove outliers, enquanto a janela w12 foi a que apresentou o menor número de valores discrepantes com apenas um outlier. Descobrimos que os intervalos entre as janelas w1 e w10 eram aqueles com maior concentração de outliers, enquanto as janelas restantes apresentaram uma redução no número de outliers (principalmente nas janelas w11 e w12), passando de uma média de 4,5 outliers por janela para uma média de 3,5. Como um todo, o monitoramento resulta em 16 janelas, totalizando 66 discrepantes. Além disso, o padrão de eventos gerado muda com a janela de dados deslizantes, portanto, é um modelo de detecção de padrão variável. Esse método pode capturar a dinâmica de um sistema variável no tempo e é adequado para descrever o comportamento dos dados a partir do monitoramento ambiental urbano variável no tempo.

Figura 30 – Detecção de outliers para janelas deslizantes.



Fonte: elaborado pelo autor (2020).

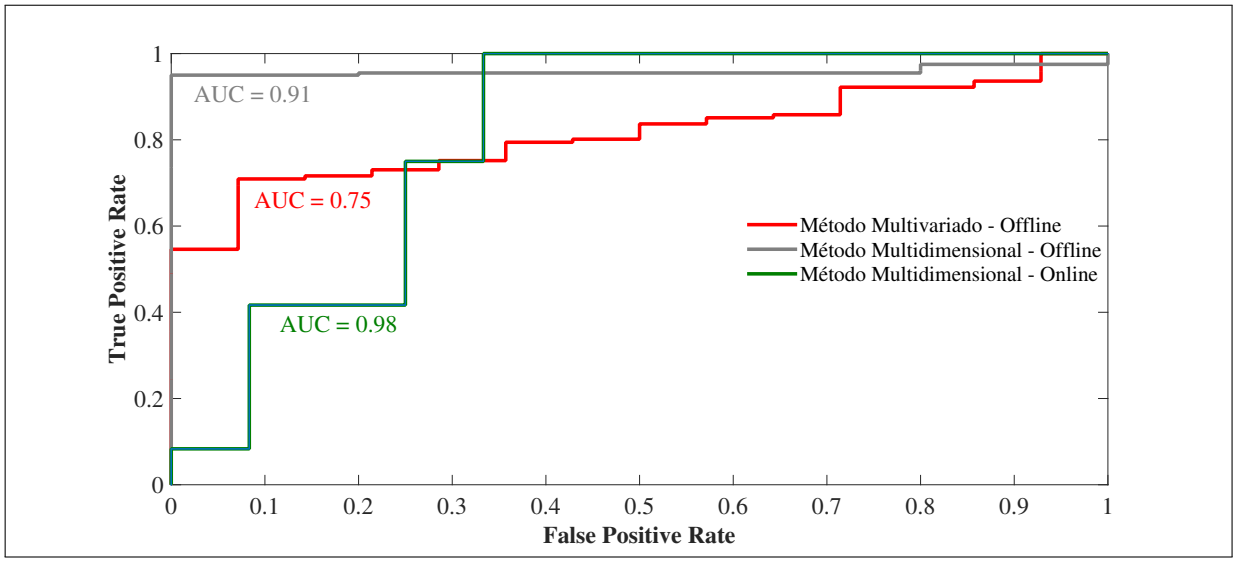
4.4 Avaliação de desempenho

Nesta seção, 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 plotar a taxa verdadeiro positivo (*True Positive Rate*, TPR), onde os rótulos dos *outliers* foram definidos na Seção 3.3 desta tese, contra a taxa de falso positivo (*False Positive Rate*, FPR). Além disso, uma abordagem para avaliar a significância estatística de uma curva ROC é calcular a ”área sob a curva”, que mensura o desempenho das abordagens - maiores

valores de AUC significam uma melhor abordagem.

Na Figura 31, comparamos o desempenho de detecção de *outlier* derivado do método multivariado *offline* (análise fatorial exploratória) com o método multidimensional *offline* (HOSVD + *kmeans*) e o método multidimensional *online* (HOSVD associado ao método da janela deslizante). Assim, os resultados mostrados na Figura 31 exibem uma AUC maior (0,91) para o método multidimensional offline e uma AUC inferior (0,75) para o método multivariado *offline*. As curvas revelam também que, como a AUC da abordagem multivariada *offline* é menor do que a AUC da abordagem multidimensional *offline*, esse resultado aponta para uma melhor eficiência do método HOSVD *offline* na detecção de *outlier* em comparação ao método multivariado *offline*. Entretanto, ambas as abordagens quando comparadas com o método multidimensional *online*, apresentam uma menor eficiência, uma vez que para o método HOSVD *online* a AUC foi superior (0,98) às outras duas abordagens. Portanto, observamos que o método *online* apresenta uma supremacia sobre os métodos *offline*. Além disso, o fenômeno de superioridade dos métodos multidimensionais se deve ao fato de que o modelo explora uma estrutura mais rica de informação, haja vista que, o método HOSVD decompôs os conjuntos de dados considerando as três dimensões do problema (tempo, variáveis ambientais e espaço).

Figura 31 – Avaliação de desempenho dos métodos de detecção de outliers: multivariado × multidimensional.



Fonte: elaborado pelo autor (2020).

Calculando a média das abordagens multidimensionais para o método HOSVD (Figura 31), temos uma AUC média de 95%, enquanto para a abordagem multivariada, temos uma AUC de 75%, com um ganho de 20% na precisão da detecção considerando a abordagem

multidimensional (*offline* e *online*). Dessa forma, os resultados mostraram maior eficiência para os métodos multidimensionais na detecção de *outliers* quando comparados à abordagem multivariada em termos da AUC.

4.5 Sumário do capítulo

Após a realização da análise dos dados neste capítulo, vislumbramos o potencial de aplicação da análise multivariada, via análise fatorial exploratória e, multidimensional, via decomposição HOSVD no contexto do monitoramento ambiental urbano inteligente para a detecção de *outliers*. Os resultados derivados das três abordagens propostas, indicaram perspectivas e contribuições distintas tanto para o modelo multivariado quanto para o modelo multidimensional, que vão além de simples visualizações de disposições gráficas dos fatores de carregamento, mas também por permitir investigar e explorar os padrões de *outliers* detectados. Para a abordagem multivariada, fatores latentes são extraídos e nomeados a partir do modelo fatorial; para a abordagem multidimensional *offline*, uma estrutura latente é derivada, revelando os padrões críticos do comportamento das variáveis ambientais sensoriadas a partir da série espaço-temporal retornada do modelo multidimensional; e finalmente, para a abordagem multidimensional *online*, através da associação do método multidimensional com a estratégia da janela deslizante, o modelo latente gerou uma estrutura em tempo real que capta o instante em que determinado evento muda ou um novo evento é introduzido na série temporal, indicando a ocorrência instantânea de um evento incomum. Por fim, o capítulo que se segue corresponde ao último desta tese e apresentará algumas considerações finais deste trabalho e perspectivas de trabalhos futuros.

5 CONCLUSÃO

Nesta tese, buscamos demonstrar a viabilidade da aplicação de ferramentas matemáticas da análise multivariada e multidimensional de dados na detecção de *outliers*, no contexto do monitoramento ambiental urbano com o intuito de contribuir para a extração de informações mais inteligentes e melhorar a confiabilidade da tomada de decisão por parte dos gestores públicos no gerenciamento dos ambientes urbanos das cidades. Para tanto, utilizando três abordagens distintas, a saber: abordagem multivariada *offline*, em que modelamos os dados como matrizes, suprimindo uma de suas dimensões, e realizamos uma análise multivariada *offline* para a detecção de *outliers*; abordagem multidimensional *offline*, em que modelamos os dados como um tensor de terceira ordem, e realizamos uma análise multidimensional *offline* na detecção de *outliers*; e abordagem multidimensional *online*, em que modelamos os dados como um tensor de terceira ordem, e realizamos uma análise multidimensional *online* na detecção de *outliers*.

Em nossa abordagem de detecção de *outliers offline* multivariada para dados de monitoramento de ambientes urbanos inteligentes, nos baseamos na Análise Fatorial Exploratória (contribuição #1), e obtivemos uma estrutura fatorial-base revelando os fatores latentes mais representativos, os quais foram nomeados, a saber: **Fator Condições Climáticas** e **Fator Qualidade do Ar**, (respostas à QP #1). A partir dos fatores latentes extraídos pelo modelo multivariado, a distância de Mahalanobis foi calculada sobre os escores dos fatores, tomando os valores da estatística como um recurso de identificação de eventos discrepantes, caso ultrapassem o limite de controle estabelecido. Padrões de *outliers* foram identificados para ambos os fatores: para o Fator Condições Climáticas, constatamos que as variáveis ambientais, temperatura e umidade, foram as responsáveis por gerar o comportamento discrepante; para o Fator Qualidade do Ar, as variáveis ambientais, monóxido de carbono e dióxido de nitrogênio, foram as que influenciaram no comportamento anômalo dos dados. Além disso, avaliando o desempenho da detecção de outliers dessa abordagem mensurando a significância estatística de uma curva ROC através da ”área sob a curva” (AUC), obtivemos uma AUC de cerca de 75%. Portanto, o método multivariado via análise fatorial exploratória aponta para a solução do problema destacado na Seção 1.2 e confirma a Hipótese #1 (vide Seção 1.3).

Já nossa abordagem de detecção de *outliers offline* multidimensional, com base na combinação de técnicas, de decomposição tensorial e multivariada para reconhecer padrões de dados coletados de sensores urbanos inteligentes (contribuição #2), geramos um modelo de detecção de *outliers* para o monitoramento ambiental urbano. Para este fim, caracterizamos

as interações de maneira multidimensional, através do método de fatoração tensorial HOSVD para extrair estruturas latentes. Como resultado, nosso método fornece uma nova abordagem de agrupamento que leva em consideração informações de diferentes dimensões e permite uma melhor interpretação dos padrões espaço-temporal dos dados subjacentes, a partir do acesso ao histórico de dados da série espaço-temporal decomposta pelo modelo multidimensional (resposta à QP #2). Portanto, nesta abordagem não há qualquer restrição a nenhuma dimensão considerada nos conjuntos de dados analisados, viabilizando uma análise mais profunda de todas as dinâmicas intrínsecas às variáveis latentes investigadas e promovendo uma interpretação mais clara das informações obtidas a partir dos dados originais, identificando padrões característicos e similaridades entre os atributos. Embora o modelo de fatoração tensorial retorne nas matrizes fatores decompostas, tanto informações espaciais como temporais, utilizamos implicitamente as informações espaciais, uma vez que as matrizes fatores englobam a influência dessa dimensão em seus componentes, enriquecendo na dimensão temporal analisada as informações relacionadas aos espaços urbanos inteligentes. Além disso, para esta abordagem a significância estatística retornada pela curva ROC através da ”área sob a curva” (AUC) foi de cerca de 91%. Enfatiza-se ainda que, as estratégias de detecção de *outliers* adotadas nesta tese contribuíram para solucionar uma das limitações encontradas em muitos trabalhos na literatura, que é meramente a aplicação de modelos multivariados, uma vez que esta pesquisa considerou não apenas a natureza multivariada dos dados coletados, mas explorou todas as dimensões reais do problema. Desta forma, o método multidimensional via a decomposição HOSVD aponta para a solução do problema destacado na Seção 1.2, sendo capaz de caracterizar e revelar padrões multidimensionais ocultos às técnicas bidimensionais tradicionais de análise de dados, e confirma a Hipótese #2 (vide Seção 1.3).

Por fim, em nossa abordagem de detecção de *outliers online* multidimensional, utilizamos a técnica da janela deslizante para fornecer a detecção online (contribuição #3), objetivando extrair as variações que são representações dos instantes de ocorrências de um evento específico, extraindo com eficiência a dinâmica do processo (resposta à QP #3). Ademais, para esta abordagem online a significância estatística retornada pela curva ROC através da ”área sob a curva” (AUC) foi de cerca de 98%. Além disso, contribuímos com a literatura incorporando uma nova análise tensorial baseada em janela (do inglês, *Window-based Tensor Analysis* - WTA). Assim, o método multidimensional via a decomposição HOSVD associado ao método da janela deslizante também mitiga o problema colocado na Seção 1.2 e confirma a Hipótese #3.

Com o número cada vez maior de fontes heterogêneas de dados, particularmente de

ambientes urbanos inteligentes, a presente tese contribui com uma nova abordagem de extração conhecimento e obtenção de informações úteis para o melhor monitoramento desses dados. Além disso, fornece uma nova possibilidade de identificação de padrões de dados que se desviam notavelmente do comportamento esperado, explorando suas interações espaço-temporais e suas complexidades cuja análise pode impactar na compreensão de problemas ambientais cada vez mais recorrentes em grandes centros urbanos, tais como mudanças climáticas, níveis de ruído e poluição do ar. Acreditamos que a presente tese pode auxiliar os gestores públicos das cidades a tomar decisões orientadas a evidências científicas.

5.1 Limitações e perspectivas

É importante mencionar que os métodos aplicados em cada abordagem precisam de um número considerável de amostras para detectar *outliers* com maior sucesso. Esse requisito causa impacto sobre aplicações em tempo real.

Além disso, apesar de capturar padrões usando matrizes fatores, as interações comuns também são caracterizadas no tensor do núcleo, o que pode ser considerado posteriormente como parâmetros de peso em modelos latentes. Uma abordagem adicional baseada em um modelo de fatoração probabilística combinado com uma abordagem tensorial caracterizaria as dependências e interações entre as diferentes dimensões em um ambiente de alta dimensão. Para além, em nossa pesquisa futura, planejamos continuar explorando nossa abordagem proposta nos quatro aspectos a seguir: primeiro, incorporar a janela deslizante nos outros modos de decomposição multidimensional, considerando o aspecto dinâmico das outras dimensões, bem como para explorar outros tipos de modelos de janelas, como janelas de marcos, janelas inclinadas e janelas desbotadas (NGUYEN *et al.*, 2015). Segundo, propor novas abordagens online usando outras decomposições multidimensionais. Terceiro, utilizar outras aplicações da vida real de nossa abordagem proposta, como detecção de injeção de dados falsos em cidades inteligentes, além de avaliar a redução de alarmes falsos. Por fim, além de trabalhar com a detecção de *outliers*, explorar o seu diagnóstico, identificando as potenciais variáveis responsáveis pela geração dos respectivos *outliers*.

Existem várias outras direções promissoras para pesquisas futuras na detecção de *outliers*, para além da abordagem em computação urbana. Por exemplo, a natureza multidimensional também pode ser evidenciada em dados coletados no contexto da apicultura de precisão (BRAGA, 2020), em que um arranjo multidimensional, com diferentes instantes de monito-

ramento (dimensão temporal) $\times$ variáveis físicas sensoriadas (dimensão atributos físicos) $\times$ apiários (dimensão espacial), pode ser explorado para a detecção de *outliers*, podendo indicar níveis críticos de bem estar das colmeias. Na área médica, a detecção de *outliers* também pode ser explorada, em que um modelo multidimensional (pacientes $\times$ *features* $\times$ cenários) pode ser utilizado para análise de registros médicos eletrônicos, onde cada cenário pode representar sinais de eletrocardiograma (ECG), ou parâmetros de exames de sangue, ou imagens (raios-x), etc. Desta forma, uma nova abordagem no monitoramento de saúde de pacientes pode ser derivada, podendo indicar inclusive novos surtos de doenças, bem como anomalias nos registros médicos de saúde das pessoas. No domínio das redes socais, a nossa abordagem também pode fornecer uma nova possibilidade de detecção de *outliers*. Dados provenientes de redes como Twitter ou Facebook, podem ser usados para construir um tensor multidimensional (usuário $\times$ localização $\times$ tempo), em que podemos detectar em cada instante graus elevados de proximidade entre pessoas, bem como também pessoas, links e comunidades anômalas. No monitoramento de qualidade de água, nossa abordagem também encontra forte possibilidade de aplicação, uma vez que construindo um tensor multidimensional (locais $\times$ variáveis $\times$ tempo), podemos descobrir não apenas localizações anormais, mas também instantes de tempo e medições que estão mais relacionadas a eventos discrepantes. Assim, os métodos de detecção de *outliers* desenvolvidos e derivados nesta tese, podem ser aplicados a uma ampla gama de dados complexos como séries temporais, padrões sequenciais, dados de redes, dados espaço-temporais (incluindo dados geoespaciais), dados de objetos móveis, dados de sistemas físicos cibernéticos, etc.

5.2 Publicações

Seguem as publicações decorrentes da pesquisa ao longo do doutorado:

(i) Artigos completos em periódicos:

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. An Online Method to Detect Urban Computing Outliers via Higher-Order Singular Value Decomposition. Sensors, v. 19, p. 4464, 2019.

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. A method to detect data outliers from smart urban spaces via tensor analysis. Future Generation Computer Systems, v. 92, p. 290-301, 2019

(ii) Artigos completos em anais de conferência:

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. Monitoramento Ambiental de

Cidades Urbanas: Detectando Outliers via Análise Fatorial Exploratória. XXXVIII Congresso da Sociedade Brasileira de Computação (CSBC), 2018. v. 1. p. 17-26.

SOUZA, T. I. A.; MAGALHAES, D. M. V.; AQUINO, A. L. L.; GOMES, D. G. Um Método para Detecção e Diagnóstico de Outliers em Dados Urbanos via Análise Multidimensional. Anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), 2018. v. 36. p. 1-14.

SOUZA, T. I. A.; GOMES, D. G.; MAGALHAES, D. M. V. Aplicando Estatística Multivariada para Detecção e Diagnóstico de Anomalias em Dados Urbanos. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), 2017, Belém. I Workshop de Computação Urbana (CoUrb), 2017. p. 72-85.

REFERÊNCIAS

ACAR, E.; BINGOL, C. A.; BINGOL, H.; BRO, R.; YENER, B. Seizure recognition on epilepsy feature tensor. In: **Proceedings of the 29th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, IEEE**. [S.l.: s.n.], 2007. p. 4273–4276.

AHMED, M.; MAHMOOD, A. N.; ISLAM, M. R. A survey of anomaly detection techniques in financial domain. **Future Generation Computer Systems**, v. 55, p. 278–288, 2016.

ALAM, J. R.; SAJID, A.; TALIB, R.; NIAZ, M. A review of the role of big data in business. **International Journal of Computer Science and Mobile Computing**, v. 3, p. 446–453, 2014.

ANDRADE, J.; KUBISTA, M.; CARLOSENA, A.; PRADA, D. 3-way characterization of soils by procrustes rotation, matrix-augmented principal components analysis and parallel factor analysis. **Analytica Chimica Acta**, v. 603, p. 20–29, 2007.

AQUINO, A. L. L.; JUNIOR, O. S.; FRERY, A. C.; ALBUQUERQUE, E. L.; MINI, R. A. F. Musa: Multivariate sampling algorithm for wireless sensor networks. **IEEE Transactions on Computers**, v. 63, n. 4, p. 968–978, 2014.

AQUINO, A. L. L.; NAKAMURA, E. F. Data centric sensor stream reduction for real-time applications in wireless sensor networks. **Sensors**, v. 9, n. 12, p. 9666–9688, 2009.

BABAR, M.; ARIF, F. Smart urban planning using big data analytics to contend with the interoperability in internet of things. **Knowledge-Based Systems**, v. 77, p. 65–76, 2017.

BAI, Y.; TEZCAN, J.; CHENG, Q.; CHENG, J. A multiway model for predicting earth-quake ground motion. In: **Proceedings of the 2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), IEEE**. [S.l.: s.n.], 2013. p. 219–224.

BARTHOLOMEW, D. J.; KNOTT, M. **Latent Variable Models and Factor Analysis.** [S.l.]: Arnold Publishers, 1999.

BASILEVSKY, A. T. **Statistical factor analysis and related methods.** [S.l.]: John Wiley and Sons, 2009.

BERGQVIST, G.; LARSSON, E. The higher-order singular value decomposition: theory and an application [lecture notes]. **IEEE Signal Processing Magazine**, v. 27, p. 151–154, 2010.

BHUYAN, M. H.; KALWAR, A.; GOSWAMI, A.; BHATTACHARYYA D.; KALITA, J. Low-rate and high-rate distributed dos attack detection using partial rank correlation. In: **In Proceedings of the fifth international conference on communication systems and network technologies (CSNT)**. [S.l.]: Gwalior, India, 2015. p. 706–710.

BI, Y.; LIN, C.; ZHOU, H.; YANG, P.; SHEN, X.; ZHAO, H. Time-constrained big data transfer for sdn-enabled smart city. **IEEE Communications Magazine**, v. 55, p. 44–50, 2017.

BRAGA, A. R. **Modelos de classificação para predição do bem estar de colônias da abelha *Apis Mellifera*.** 2020. 125 f. Tese (Doutorado) — Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Universidade Federal do Ceará, 2020.

CAMACHO, J.; VILLEGAS, A. P.; TEODORO, P. G.; FERNANDEZ, G. M. Pca-based multivariate statistical network monitoring for anomaly detection. **Computers and Security**, v. 59, p. 118–137, 2016.

CAMPOS, G. O.; ZIMEK, A.; SANDER, J.; CAMPELLO, R. J.; MICENKOVA, B.; E., S.; ASSENT, I.; HOULE, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. **Data Min. Knowl. Discov.**, v. 30, p. 891–927, 2016.

CARROL, J. D.; CHANG, J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. **Psychometrika**, v. 35, p. 283–319, 1970.

CARROLL, J. D.; SOETE, G. D.; PRUZANSKY, S. **Fitting of the latent class model via iteratively reweighted least squares candecomp with nonnegativity constraints**. [S.l.]: Multiway data analysis, 1989.

CARTON, L.; ACHE, P. Citizen-sensor-networks to confront government decision-makers: Two lessons from the netherlands. **Journal of Environmental Management**, v. 196, p. 234–251, 2017.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Computing Surveys**, 2009.

CHEN, B.; LI, Z.; ZHANG, S. On optimal low rank tucker approximation for tensors: the case for an adjustable core size. **Journal of Global Optimization**, v. 62, p. 811–832, 2014.

CHEN, D.; LI, X.; WANG, L.; KHAN, S.; WANG, J.; ZENG, K.; CAI, C. Fast and scalable multi-way analysis of massive neural data. **IEEE Transactions on Computers**, v. 64, p. 707–719, 2015.

CHEN, J.; YEN, J.-H. Three-way data analysis with time lagged window for online batch process monitoring. **Korean Journal of Chemical Engineering**, v. 20, p. 1000–1011, 2003.

CHOI, B. G.; PARK, H. S.; KIM, G. H.; JUNG, Y. M.; YI, K. B.; KIM, J.; HONG, W. H. Analysis of co2–nh3 reaction dynamics in an aqueous phase by pca and 2d ir cos. **Journal of Industrial and Engineering Chemistry**, v. 18, p. 98–104, 2012.

CICHOCKI, A.; ZDUNEK, R.; PHAN, A. H.; AMARI, S. **Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.** [S.l.]: John Wiley and Sons, 2009.

CICIRELLI, F.; GUERRIERI, A.; SPEZZANO, G.; VINCI, A. An edge-based platform for dynamic smart city applications. **Future Generation Computer Systems**, v. 76, p. 106–118, 2017.

CITIZEN. Smart citizen documentation. **URL. http://docs.smartcitizen.me/.**, 2016.

CONG, F.; H.PHAN, A.; ZHAO, Q.; HUTTUNEN-SCOTT, T.; KAARTINEN, J.; RISTANIEMI, T.; CICHOCKI, A. Benefits of multi-domain feature of mismatch negativity extracted by non-negative tensor factorization from eeg collected by low-density array. **International Journal of Neural Systems**, v. 22, p. 1250025, 2012.

DEMPSEY, N.; BRAMLEY, G.; POWER, S.; BROWN, C. The social dimension of sustainable development: Defining urban social sustainability. **Sustainable Development**, v. 19, p. 289–300, 2009.

DO, H.; CETIN, K. S. Evaluation of the causes and impact of outliers on residential building energy use prediction using inverse modeling. **Build. Environ.**, v. 138, p. 194–206, 2018.

ENGLE, M. A.; GALLO, M.; SCHROEDER, K. T.; GEBOY, N. J.; ZUPANCIC, J. W. Three-way compositional analysis of water quality monitoring data. **Environmental and Ecological Statistics**, v. 21, p. 565–581, 2014.

FANAEE-T, H.; GAMA, J. Eigenevent: an algorithm for event detection from complex data streams in syndromic surveillance. **Intelligent Data Analysis**, v. 19, p. 1–20, 2015.

FANAEE-T, H.; GAMA, J. Multi-aspect-streaming tensor analysis. **Knowledge-Based Systems**, v. 89, p. 332–345, 2015.

FANAEE-T, H.; GAMA, J. Tensor-based anomaly detection: An interdisciplinary survey. **Knowledge-Based Systems**, v. 98, p. 130–147, 2016.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, p. 861–874, 2006.

GAO, J.; JI, W.; ZHANG, L.; LI, A.; Y., W.; ZHANG, Z. Cube-based incremental outlier detection for streaming computing. **Information Sciences**, v. 517, p. 361–376, 2020.

GOLDSTEIN, M.; UCHIDA, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. **PLoS One**, v. 11, n. 4, p. e0152173, 2016.

GUARDIOLA, I. G.; LEON, T.; MALLOR, F. A functional approach to monitor and recognize patterns of daily traffic profiles. **Transportation Research Part B**, v. 65, p. 119–136, 2014.

HARSHMAN, R. A. Foundations of the parafac procedure: Models and conditions for an "explanatory"multi-modal factor analysis. **UCLA Working Papers in Phonetics**, v. 16, p. 84–104, 1970.

HO, J. C.; GHOSH, J.; STEINHUBL, S. R.; STEWART, W. F.; DENNY, J. C.; MALIN, B. A.; SUN, J. Limestone: high-throughput candidate phenotype generation via tensor factorization. **Journal of Biomedical Informatics**, v. 52, p. 199–211, 2014.

HOTELLING, H. Analysis of a complex of statistical variables in to principal components. **Journal of Educational Psychology**, v. 24, p. 417–441, 1933.

IBRAHIM, A. T. H.; VICTOR, C.; NOR, B. A.; KAYODE, A.; IBRAR, Y.; ABDULLAH, G.; EJAZ, A.; HARUNA, C. The role of big data in smart city. **International Journal of Information Management**, v. 36, p. 748–758, 2016.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, p. 651–666, 2010.

JUN, J.; AHN, C.; KIM, S. H. Ddos attack detection by using packet sampling and flow features. In: **In Proceedings of the twentyninth annual ACM symposium on applied computing**. [S.l.]: Gyeongju, Korea, 2014. p. 711–712.

KAISER, H. F. The application of electronic computers to factor analysis. **Educational and Psychological Measurement**, v. 20, p. 141–151, 1966.

KARANJIT, S.; SHUCHITA, U. D. Outlier detection: applications and techniques. **Int. J. Comput. Sci. Issue**, v. 9, p. 307–323, 2012.

KHATIB, E. J.; BARCO, R.; MUNOZ, P.; BANDERA, I.; SERRANO, I. Self-healing in mobile networks with big data. **IEEE Communications Magazine**, v. 54, p. 114–120, 2016.

KIERS, H. A. An alternating least squares algorithm for parafac2 and three-way dedicom. **Computational Statistics and Data Analysis**, v. 16, p. 103–118, 1993.

KIERS, H. A. L. Towards a standardized notation and terminology in multiway analysis. **Journal of Chemometrics**, v. 14, p. 105–22, 2000.

KOLDA, T. G.; BADER, B. W. Tensor decompositions and applications. **Society for Industrial and Applied Mathematics**, v. 51, p. 455–500, 2009.

KOSANOVICH, K.; PIOVOSO, M.; DAHL, K.; MACGREGOR, J.; NOMIKOS, P. Multi-way pca applied to an industrial batch process. In: **Proceedings of American Control Conference**. [S.l.: s.n.], 1994. p. 1294–1298.

KOUTRA, D.; PAPALEXAKIS, E. E.; FALOUTSOS, C. Tensorsplat: spotting latent anomalies in time. In: **Proceedings of the 16th Panhellenic Conference on Informatics (PCI), IEEE**. [S.l.: s.n.], 2012. p. 144–149.

KROONENBERG, P. M. **Applied Multiway Data Analysis**. [S.l.]: John Wiley and Sons, 2008.

LATHAUWER, L. D.; MOOR, B. D.; VANDEWALLE, J. A multilinear singular value decomposition. **SIAM Journal on Matrix Analysis and Applications**, v. 21, p. 1253–1278, 2000.

LEE, S.; LIU, H.; KIM, M.; KIM, J. T.; YOO, C. Online monitoring and interpretation of periodic diurnal and seasonal variations of indoor air pollutants in a subway station using parallel factor analysis (parafac). **Energy and Buildings**, v. 68, p. 87–98, 2014.

LI, J.; PEDRYCZ, W.; JAMAL, I. Multivariate time series anomaly detection: A framework of hidden markov models. **Applied Soft Computing**, v. 60, p. 229–240, 2017.

LI, X.; SHU, W.; LI, M.; HUANG, H. Y.; LUO, P. E.; WU, M. Performance evaluation of vehicle-based mobile sensor networks for traffic monitoring. **IEEE Trans. Veh. Technol.**, v. 58, p. 1647–1653, 2009.

LIU, Y.; WENG, X.; WAN, J.; YUE, X.; SONG, H.; VASILAKOS, A. V. Exploring data validity in transportation systems for smart cities. **IEEE Communications Magazine**, v. 55, p. 26–33, 2017.

MAHALANOBIS, P. C. On the generalised distance in statistics. **Proc. Natl. Inst. Sci. India**, v. 2, p. 49–55, 1936.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. **Big data: The Next Frontier for Innovation, Competition, and Productivity.** [S.l.]: McKinsey Global Institute, 2011.

MAYER-SCHONBERGER, V.; CUKIE, K. **Big data: A revolution that will transform how we live, work and think.** [S.l.]: John Murray, 2013.

MCKERCHER, G. R.; SALMOND, J. A.; VANOS, J. K. Characteristics and applications of small, portable gaseous air pollution monitors. **Environmental Pollution**, v. 223, p. 102–110, 2017.

MILLS, G. Cities as agents of global change. **International Journal of Climatology**, v. 27, p. 1849–1857, 2007.

MORI, J.; YU, J. Quality relevant nonlinear batch process performance monitoring using a kernel based multiway non-gaussian latent subspace projection approach. **Journal Process Control**, v. 24, p. 57–71, 2014.

MORUP, M. Applications of tensor (multiway array) factorizations and decompositions in data mining. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, p. 24–40, 2011.

MU, Y.; DING, W.; MORABITO, M.; TAO, D. Empirical discriminative tensor analysis for crime forecasting. In: **Knowledge Science, Engineering and Management, Springer**. [S.l.: s.n.], 2011. p. 293–304.

MUR, A.; DORMIDO, R.; DURO, N.; DORMIDO-CANTO, S.; VEGA, J. Determination of the optimal number of clusters using a spectral clustering optimization. **Expert Systems With Applications**, v. 65, p. 304–314, 2016.

NATIONS, D. o. E. U.; AFFAIRS, P. D. S. World urbanization prospects: The 2014 revision, highlights. New York, 2015.

NEIROTTI, P.; MARCO, A.; CAGLIANO, A. C.; MANGANO, G.; SCORRANO, F. Current trends in smart city initiatives: Some stylised facts. **Cities**, v. 38, p. 25–36, 2014.

NGUYEN, H. L.; WOON, Y. K.; NG, W. K. A survey on data stream clustering and classification. **Knowl. Inf. Syst.**, v. 45, p. 535–569, 2015.

NOMIKOS, P.; MACGREGOR, J. F. Monitoring batch processes using multiway principal component analysis. **AIChE Journal**, v. 40, p. 1361–1375, 1994.

NUORTIO, T.; KYTOJOKI, J.; NISKA, H.; BRAYSY, O. Improved route planning and scheduling of waste collection and transport. **Expert Syst. Appl: Int. J**, v. 30, p. 223–232, 2006.

ORDONEZ, C.; SESTELO, M.; ROCA-PARDINAS, J.; COVIAN, E. Variable selection in regression models used to analyse global positioning system accuracy in forest environments. **Applied Mathematics and Computation**, v. 219, p. 2220–2230, 2012.

OSANAIYE, O.; CHOO, K.-K. R.; DLODLO, M. Distributed denial of service (ddos) resilience in cloud: Review and conceptual cloud ddos mitigation framework. **Journal of Network and Computer Applications**, v. 67, p. 147–165, 2016.

PAN, G.; QI, G.; ZHANG, W.; LI, S.; WU, Z. Trace analysis and mining for smart cities: Issues, methods, and applications. **IEEE Communications Magazine**, v. 51, p. 120–126, 2013.

PANISSON, A.; GAUVIN, L.; QUAGGIOTTO, M.; CATTUTO, C. Mining concurrent topical activity in microblog streams. In: **Proceedings of the Workshop on Making Sense of Microposts Colocated with the International World Wide Web Conference**. [S.l.: s.n.], 2014. p. 3–10.

PIRO, G.; CIANCI, I.; GRIECO, L. A.; BOGGIA, G.; CAMARDA, P. Information centric services in smart cities. **Journal of Systems and Software**, v. 88, p. 169–188, 2014.

PROGRAMME, U. N. H. S. **World Cities Report 2016: Urbanization and Development : Emerging Futures**. UN Habitat, 2016. Disponível em: <http://cdn.plataformaurbana.cl/wp-content/uploads/2016/06/wcr-full-report-2016.pdf>.

RATHORE, M. M.; AHMAD, A.; PAUL, A.; RHO, S. Urban planning and building smart cities based on the internet of things using big data analytics. **Knowledge-Based Systems**, v. 101, p. 63–80, 2016.

RENARD, N.; BOURENNANE, S. Improvement of target detection methods by multiway filtering. **IEEE Transactions on Geoscience and Remote Sensing**, v. 46, p. 2407–2417, 2008.

RIPOLL, R. B.; PAJAROLA, R. Lossy volume compression using tucker truncation and thresholding. **The Visual Computer**, v. 32, p. 1433–1446, 2015.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust regression and outlier detection.** New York, NY.: John Wiley and Sons, 1987. v. 2.

SANCHEZ, L.; MUNOZ, L.; GALACHE, J.; SOTRES, P.; SANTANA, J.; GUTIERREZ, V.; RAMDHANY, R.; GLUHAK, A.; KRCO, S.; THEODORIDIS, E.; PFISTERER, D. Smartsantander: Iot experimentation over a smart city testbed. **Computer Network**, v. 61, p. 217–238, 2013.

SINGH, K. P.; MALIK, A.; BASANT, N.; SAXENA, P. Multi-way partial least squares modeling of water quality data. **Analytica Chimica Acta**, v. 2, p. 385–396, 2007.

SIVARAMAN, V.; CARRAPETTA, J.; HU, K.; LUXAN, B. G. Hazewatch: A participatory sensor system for monitoring air pollution in sydney. **Journal of Environmental Management**, v. 56, p. 23–29, 2013.

SMILDE, A.; BRO, R.; GELADI, P. **Multi-way Analysis - Applications in the Chemical Sciences.** [S.l.]: John Wiley and Sons, 2004.

SONG, J.; GAO, B.; WOO, W. L.; TIAN, G. Y. Ensemble tensor decomposition for infrared thermography cracks detection system. **Infrared Physics and Technology**, v. 105, p. 103–203, 2020.

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. Monitoramento ambiental de cidades urbanas: Detectando outliers via análise fatorial exploratória. **XXXVIII Congresso da Sociedade Brasileira de Computação (CSBC)**, v. 1, p. 17–26, 2018.

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. A method to detect data outliers from smart urban spaces via tensor analysis. **Future Gener. Comput. Syst.**, v. 92, p. 290–301, 2019.

SOUZA, T. I. A.; AQUINO, A. L. L.; GOMES, D. G. An online method to detect urban computing outliers via higher-order singular value decomposition. **Sensors**, v. 19, p. 44–64, 2019.

SOUZA, T. I. A.; MAGALHãES, D. M. V.; GOMES, D. G. Aplicando estatística multivariada para detecção e diagnóstico de anomalias em dados urbanos. **Anais do I Workshop de Computação Urbana (CoUrb)**, v. 1, p. 72–85, 2017.

STANIMIROVA, I.; SIMEONOV, V. Modeling of environmental four-way data from air quality control. **Chemometrics and Intelligent Laboratory Systems**, v. 77, p. 115–121, 2005.

STEED, C. A.; RICCIUTO, D. M.; SHIPMAN, G.; SMITH, B.; THORNTON, P. E.; WANG, D.; SHI, X.; WILLIAMS, D. N. Big data visual analytics for exploratory earth system simulation analysis. **Computers And Geosciences**, v. 61, p. 71–82, 2013.

SUN, J.; PAPADIMITRIOU, S.; PHILIP, S. Y. Window-based tensor analysis on high-dimensional and multi-aspect streams. In: **Proceedings of International Conference on Data Mining (ICDM)**. [S.l.: s.n.], 2006. p. 1076–1080.

SUN, J.; TAO, D.; FALOUTSOS, C. Beyond streams and graphs: dynamic tensor analysis. In: **Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM**. [S.l.: s.n.], 2006. p. 374–383.

SUNDBERG, R.; FELDMANN, U. Exploratory factor analysis—parameter estimation and scores prediction with high-dimensional data. **Journal of Multivariate Analysis**, v. 148, p. 49–59, 2016.

SUZHI, B.; RUI, Z.; ZHI, D.; SHUGUANG, C. Wireless communications in the era of big data. **IEEE Communications Magazine**, v. 53, p. 190–199, 2015.

SWETS, J. A.; PICKETT, R. M. **Evaluation of Diagnostic Systems: Methods from Signal Detection Theory**. [S.l.]: Academic Press, Inc. New York, 1982.

TENG H., C. K. a. L. S. Adaptive real-time anomaly detection using inductivelygenerated sequential patterns. **IEEE Computer Society Press**, v. 2, p. 278–284, 1990.

THOMPSON, J. E. Crowd-sourced air quality studies: A review of the literature and portable sensors. **Trends in Environmental Analytical Chemistry**, v. 11, p. 23–34, 2016.

TRACY, N. D.; YOUNG, J. C.; MASON, R. L. Multivariate control charts for individual observations. **Expert Systems With Applications**, v. 24, p. 88–95, 1972.

TRAN, L.; NAVASCA, C.; LUO, J. Video detection anomaly via low-rank and sparse decompositions. In: **Proceedings of Image Processing Workshop (WNYIPW), 2012, IEEE**. [S.l.]: Western, New York, 2012. p. 17–20.

TUCKER, L. R. Some mathematical notes on three-mode factor analysis. **Psychometrika**, v. 31, p. 279–311, 1966.

URTUBIA, A.; HERNáNDEZ, G.; ROGER, J. Detection of abnormal fermentations in wine process by multivariate statistics and pattern recognition techniques. **Journal Biotechnol.**, v. 159, p. 336–341, 2012.

VEERAMANIKANDAN; SANKARANARAYANAN, S.; RODRIGUES, J. J. P. C.; SUGUMARAN, V.; KOZLOV, S. Data flow and distributed deep neural network based low latency iot-edgecomputation model for big data environment. **Engineering Applications of Artificial Intelligence**, v. 94, p. 103785, 2020.

VONDREJC, J.; LIUA, D.; LADECKY, M.; MATTHIES, H. G. Fft-based homogenisation accelerated by low-rank tensor approximations. **Comput. Methods Appl. Mech. Engrg.**, v. 364, p. 1–21, 2020.

WANG, A.; JIN, Z.; TANG, G. Robust tensor decomposition via t-svd: Near-optimal statistical guarantee and scalable algorithms. **Signal Processing**, v. 167, p. 1–15, 2020.

WANG, B.; MAO, Z. Outlier detection based on gaussian process with application to industrial processes. **Appl. Soft Comput. J.**, v. 76, p. 505–516, 2019.

ZAFRA, C.; ANGEL, Y.; TORRES, E. Arima analysis of the effect of land surface coverage on pm10 concentrations in a high-altitude megacity. **Atmospheric Pollution Research**, v. 8, p. 660–668, 2017.

ZHANG, K.; NI, J.; YANG, K.; LIANG, X.; REN, J.; SHEN, X. Security and privacy in smart city applications: Challenges and solutions. **IEEE Communications Magazine**, v. 55, p. 122–129, 2017.

ZHOU, H.; LIU, B.; LIU, Y.; ZHANG, N.; GUI, L.; LI, Y.; SHEN, X. S.; YU, Q. A cooperative matching approach for resource management in dynamic spectrum access networks. **IEEE Transactions on Wireless Communications**, v. 13, p. 1047–1057, 2014.

ZIKOPOULOS, P.; EATON, C. **Understanding Big Data.** [S.l.]: McGraw-Hill, 2012.

APÊNDICE A – OUTRAS DECOMPOSIÇÕES TENSORIAIS

A.1 Fatoração PARAFAC

O PARAFAC é um método de decomposição tensorial de dados multidimensionais que desagrega o tensor multidimensional de dados em conjuntos de pontuações e carregamentos. Portanto, o modelo Parafac decompõe um tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$ como uma soma de produtos externos de vetores,

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{u_1}_r \circ \mathbf{u_2}_r \circ \ldots \circ \mathbf{u_n}_r = \sum_{r=1}^{R} \prod_{n=1}^{N} \circ \mathbf{u_n}_r = \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \ldots \times_n \mathbf{U}_n \tag{A.1}$$

onde $\mathbf{u_n}_r$ denota a r-ésima coluna de $\mathbf{U_n} \in R^{I_n \times R}$ para todo $n \in \{1, 2, \ldots, N\}$ e o tensor núcleo $\underline{\mathbf{S}}$ é o tensor identidade com 1's na superdiagonal e 0's nas demais posições fora da superdiagonal do arranjo tensorial.

Particularmente, o modelo PARAFAC para um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, é dado por três vetores de carregamento $\mathbf{u_1}_r$, $\mathbf{u_2}_r$, $\mathbf{u_3}_r$ que correspondem aos vetores de carregamento do modo I_1, I_2, I_3, respectivamente, e pode ser escrito em notação de produto externo

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{u_1}_r \circ \mathbf{u_2}_r \circ \mathbf{u_3}_r = \sum_{r=1}^{R} \prod_{i_n=1}^{R} \circ \mathbf{u_{i_n r}}, \tag{A.2}$$

O produto externo entre $\mathbf{u_1}_1$, $\mathbf{u_2}_1$ e $\mathbf{u_3}_1$ permite construir o tensor $\underline{\mathbf{X}}_1$, já o produto externo entre $\mathbf{u_1}_2$, $\mathbf{u_2}_2$ e $\mathbf{u_3}_2$ permite construir o tensor $\underline{\mathbf{X}}_2$ e, assim por diante, até que o produto externo entre $\mathbf{u_1}_R$, $\mathbf{u_2}_R$ e $\mathbf{u_3}_R$ permite construir o tensor $\underline{\mathbf{X}}_R$

Além disso, utilizando o processo de matriciação, o modelo PARAFAC para um tensor de terceira ordem, também pode ser escrito na forma matricial utilizando o produto de Khatri-Rao para cada modo conforme, apresentado a seguir:

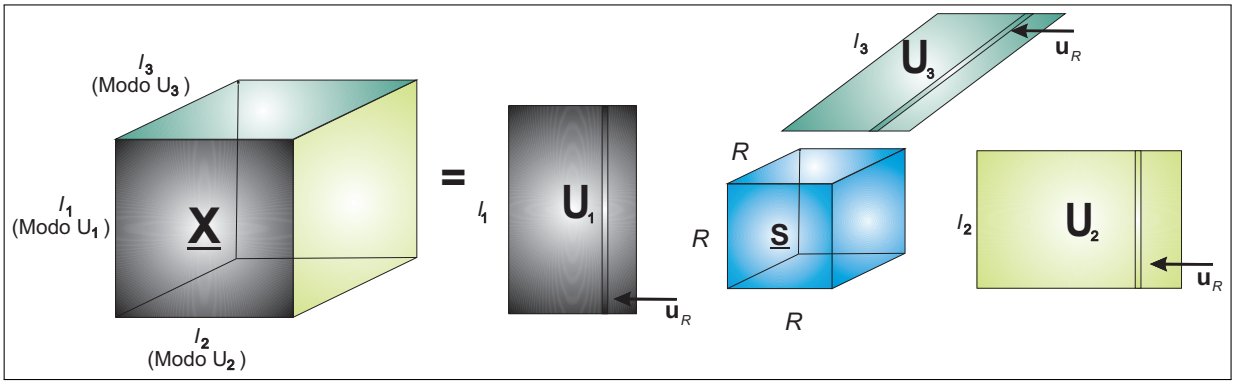
$$\mathbf{X}_{I_1} = \mathbf{U_1}(\mathbf{U_3} \diamond \mathbf{U_2})^T, \tag{A.3}$$

$$\mathbf{X}_{I_2} = \mathbf{U_2}(\mathbf{U_3} \diamond \mathbf{U_1})^T, \tag{A.4}$$

$$\mathbf{X}_{I_3} = \mathbf{U_3}(\mathbf{U_2} \diamond \mathbf{U_1})^T, \tag{A.5}$$

Uma representação pictórica do modelo PARAFAC na decomposição de um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, pode ser vista na Figura 32.

Figura 32 – Representação da fatoração PARAFAC para um tensor de terceia ordem.



Fonte: elaborado pelo autor (2020).

A.2 Fatoração Tucker

Alternativamente, temos a fatoração multidimensional de Tucker que, diferentemente da decomposição tensorial PARAFAC, incorpora dimensões de interação através de um tensor núcleo que pondera os diferentes modos da decomposição. Além disso, esse modelo de decomposição não requer o mesmo número de colunas para as matrizes de fatores $\{\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_n}\}$. Desta forma, analogamente à equação A.1, o modelo de fatoração de Tucker para um tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$ é

$$\underline{\mathbf{X}} = \sum_{r_1=1}^{R_1} \ldots \sum_{r_n=1}^{R_N} s_{r_1,\ldots,r_n} (\mathbf{u_{I_1 r_1}} \circ \ldots \circ \mathbf{u_{I_N r_n}}) = \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \ldots \times_n \mathbf{U}_n \tag{A.6}$$

onde $\mathbf{u_{I_N r_n}}$ denota a r_n-ésima coluna de $\mathbf{U_n} \in R^{I_n \times R_n}$. O tensor $\underline{\mathbf{S}} \in \mathbb{R}^{R \times R \times \ldots \times R}$ no modelo PARAFAC é substituído por um tensor núcleo mais geral $\underline{\mathbf{S}} \in \mathbb{R}^{R_1 \times R_2 \times \ldots \times R_n}$, o que viabiliza uma interação entre as matrizes fatores no modelo Tucker.

Particularmente, o modelo Tucker para um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, também chamado popularmente de Tucker3, pode ser escrito em notação de produto externo da seguinte maneira:

$$\underline{\mathbf{X}} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} (\mathbf{u_{I_1 r_1}} \circ \mathbf{u_{I_2 r_2}} \circ \mathbf{u_{I_3 r_3}}) = \underline{\mathbf{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \tag{A.7}$$

Além disso, utilizando o processo de matriciação, o modelo Tucker3 para um tensor de terceira ordem, também pode ser escrito na forma matricial utilizando o produto de Kronecker para cada modo conforme, apresentado a seguir:

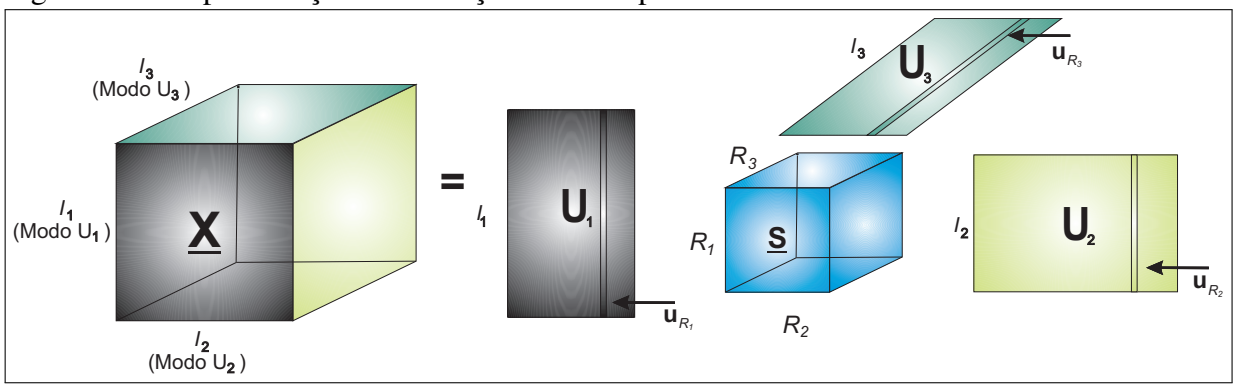
$$\mathbf{X}_{I_1} = \mathbf{U_1} \mathbf{S_{I_1}} (\mathbf{U_3} \otimes \mathbf{U_2})^T, \tag{A.8}$$

$$\mathbf{X}_{I_2} = \mathbf{U_2 S_{I_2}} (\mathbf{U_3} \otimes \mathbf{U_1})^T, \tag{A.9}$$

$$\mathbf{X}_{I_3} = \mathbf{U_3 S_{I_3}} (\mathbf{U_2} \otimes \mathbf{U_1})^T, \tag{A.10}$$

Uma representação pictórica do modelo Tucker3 na decomposição de um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, pode ser vista na Figura 33.

Figura 33 – Representação da fatoração Tucker3 para um tensor de terceia ordem.



Fonte: elaborado pelo autor (2020).

A.3 Relação entre as fatorações PARAFAC e Tucker3

Considerando os métodos de fatoração tensorais PARAFAC e Tucker3, observamos que ambos os métodos tensoriais guardam uma relação intrínseca entre si. O modelo Tucker3 para um tensor de terceira ordem $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, pode ser escrito em notação de soma como

$$x_{i_1 i_2 i_3} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} (u_{i_1 r_1} u_{i_2 r_2} u_{i_3 r_3}) + e_{i_1 i_2 i_3} \tag{A.11}$$

onde $u_{i_1 r_1}$ é o componente $I_1 \times R_1$ da matriz modo-1 $\mathbf{U}_1$, $u_{i_2 r_2}$ é o componente $I_2 \times R_2$ da matriz modo-2 $\mathbf{U}_2$, $u_{i_3 r_3}$ é o componente $I_3 \times R_3$ da matriz modo-3 $\mathbf{U}_3$, $s_{r_1 r_2 r_3}$ é o componente $R_1 \times R_2 \times R_3$ do tensor núcleo $\underline{\mathbf{S}}$, e $e_{i_1 i_2 i_3}$ é o componente $I_1 \times I_2 \times I_3$ do tensor de erro de aproximação $\mathbf{E}$.

Entretanto, conseidere agora que na equação 2.24, o termo $s_{r_1 r_2 r_3} = 1$, se e somente se $R_1 = R_2 = R_3 = R$. Neste caso, podemos reescrever a equação 2.24 como

$$x_{i_1 i_2 i_3} = \sum_{r=1}^{R} u_{i_1 r} u_{i_2 r} u_{i_3 r} + e_{i_1 i_2 i_3} \tag{A.12}$$

com as mesmas definições para $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ e $\underline{\mathbf{S}}$, exceto que a ordem das colunas é a mesma para as matrizes de componentes $R_1 = R_2 = R_3 = R$ e que o tensor núcleo $\underline{\mathbf{S}}$ $(R \times R \times R)$ é cúbico e

superdiagonal (s_{rrr} é diferente de zero se $r_1 = r_2 = r_3 = r$ e zero, caso contrário). Desta forma, a fatoração Tucker3 é compreendida como um modelo geral da fatoração PARAFAC.

APÊNDICE B – TEOREMA DO VALOR SINGULAR

Toda matriz $\mathbf{X}$ $(I_1 \times I_2)$ de posto r pode ser decomposta em

$$\mathbf{X} = \mathbf{USV}^T \tag{B.1}$$

em que $\mathbf{U} \in \mathbb{R}^{I_1 \times r}$ e $\mathbf{V} \in \mathbb{R}^{I_2 \times r}$ são ortonormais por coluna, ou seja, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ e $\mathbf{S} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{I_1 \times I_2}$ é a matriz pseudo-diagonal com $\Sigma = diag(\sigma_1, \ldots, \sigma_R)$ e $R = rank(X)$. As quantidades $\sigma_1, \sigma_2, \ldots, ...\sigma_r$ são os autovalores não-nulos das matrizes $\mathbf{XX}^T$ ou $\mathbf{X}^T\mathbf{X}$ e $\mathbf{U}$ e $\mathbf{V}$ correspondem às matrizes formadas pelos r autovetores das matrizes $\mathbf{XX}^T$ ou $\mathbf{X}^T\mathbf{X}$ dispostos em suas colunas, respectivamente.

Prova:

Vamos supor que podemos decompor a matriz $\mathbf{X}$, considerando

$$\mathbf{X} = \mathbf{USV}^T \tag{B.2}$$

da mesma forma como definida no enunciado do teorema acima. Por conveniência e sem perda de generalidade vamos assumir que $I_1 > I_2$. Assim, podemos construir as matrizes ortogonais $\mathbf{U}$ e $\mathbf{V}$ acrescentando $I_1 - r$ e $I_2 - r$ vetores, respectivamente, às colunas de $\mathbf{U}$ e $\mathbf{V}$, ortonormais aos primeiros r deles.

A matriz $\mathbf{X}^T\mathbf{X}$ possui autovalores $\sigma_1, \sigma_2, \ldots, ...\sigma_{I_2}$. Como $\mathbf{X}^T\mathbf{X}$ é simétrica, pois $(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$, então existe $\mathbf{V}$ ortogonal tal que

$$\mathbf{V}^T\mathbf{X}^T\mathbf{XV} = \mathbf{S}^2 \tag{B.3}$$

em que $\mathbf{V}$ é a matriz de autovetores de $\mathbf{X}^T\mathbf{X}$ formando suas colunas.

Assim, podemos verificar que:

$$(\mathbf{V}^T\mathbf{X}^T)\mathbf{XV} = (\mathbf{XV})^T(\mathbf{XV}) = \mathbf{S}^T\mathbf{S}. \tag{B.4}$$

Fazendo, $\mathbf{W} = \mathbf{XV}$, temos que $\mathbf{W}^T\mathbf{W} = \mathbf{S}^T\mathbf{S}$, sendo $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_r, ..., \mathbf{w}_{I_2}]$. Assim, podemos concluir que $\mathbf{w}_i^T\mathbf{w}_i = \sigma_i$, se $i \leq r$ ou $\mathbf{w}_i^T\mathbf{w}_i = 0$, se $i > r$, e $\mathbf{w}_i^T\mathbf{w}_j = 0$ se $i \neq j = 1, 2, ..., I_2$.

Se $\mathbf{w}_i = 0$ para i > j, as primeiras colunas de $\mathbf{W}$ são linearmente independentes. Logo, concluímos que $r \leq I_1$. Se definirmos,

$$\mathbf{u}_i = \frac{1}{\sqrt{\sigma_i}}\mathbf{w}_i = \frac{1}{\sqrt{\sigma_i}}\mathbf{Xv}_i, \tag{B.5}$$

com $i = 1, 2, ..., r$. Assim, tomando $\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, ..., \mathbf{u}_{I_2}$, se $r < I2$, ortogonais entre si e aos demais $\mathbf{u}_i$ tais que a matriz $I_1 \times I_1$, $\mathbf{U} = \mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{I_2}$, seja ortogonal. Desta forma, podemos escrever:

$$\mathbf{US} = \mathbf{XV}. \tag{B.6}$$

Como $\mathbf{V}$ é ortogonal, temos que $\mathbf{V}^{-1} = \mathbf{V}^T$,

$$\mathbf{XV} = \mathbf{US} \tag{B.7}$$

resultando em:

$$\mathbf{XVV}^T = \mathbf{USV}^T. \tag{B.8}$$

Logo, temos $\mathbf{X} = \mathbf{USV}^T$, como queríamos demonstrar.