Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica
Pós-graduação em Engenharia Elétrica

Felipe Augusto Machado Corrêa

Proposta de um Método de Identificação de Anomalias em Redes Móveis Com Base na Representação Dimensional de KPIs usando PCA e Clustering

Uberlândia
2020

Felipe Augusto Machado Corrêa

PROPOSTA DE UM MÉTODO DE IDENTIFICAÇÃO DE ANOMALIAS EM REDES MÓVEIS COM BASE NA REPRESENTAÇÃO DIMENSIONAL DE KPIS USANDO PCA E CLUSTERING

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de mestre em Ciências.

Orientador: Prof. Dr. Alan Petrônio Pinheiro

Banca Examinadora:

Alan Petrônio Pinheiro (orientador), Dr. – UFU

Daniel Fernandes Macedo, Dr. – UFMG

Gilberto Arantes Carrijo, Dr. – UFU

Lorenço Santos Vasconcelos, Dr. – UFU

Uberlândia, 20 de fevereiro de 2020

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

C824 Corrêa, Felipe Augusto Machado, 1992-
2020 Proposta de um Método de Identificação de Anomalias em Redes Móveis Com Base na Representação Dimensional de KPIs Usando PCA e Clustering [recurso eletrônico] / Felipe Augusto Machado Corrêa. - 2020.

Orientador: Alan Petrônio Pinheiro.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Engenharia Elétrica.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2020.264
Inclui bibliografia.
Inclui ilustrações.

1. Engenharia elétrica. I. Pinheiro, Alan Petrônio,1982-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Engenharia Elétrica. III. Título.

CDU: 621.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica
Av. João Naves de Ávila, 2121, Bloco 3N - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4707 - www.posgrad.feelt.ufu.br - copel@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Engenharia Elétrica				
Defesa de:	Dissertação de Mestrado Acadêmico, 733, PPGEELT				
Data:	Vinte de fevereiro de dois mil e vinte	Hora de início:	14:00	Hora de encerramento:	17:00
Matrícula do Discente:	11722EEL004				
Nome do Discente:	Felipe Augusto Machado Corrêa				
Título do Trabalho:	Proposta de um método de identificação de anomalias em redes móveis com base na representação dimensional de KPIs usando PCA e clustering				
Área de concentração:	Processamento da informação				
Linha de pesquisa:	Processamento digital de sinais				
Projeto de Pesquisa de vinculação:	Título: Estudo e desenvolvimento piloto de novos modelos de serviços e infraestrutura de TIC voltados ao uso de antenas de telecomunicações da rede de distribuição da CEB alinhados ao cenário de SG e IoT MESTRADO Agência Financiadora: CEB/Aneel Início 23/11/18 Término 22/5/2021 No. do Projeto na agência: PD-05160-1805/2018 Professor Coordenador: Alan Petrônio Pinheiro				

Reuniu-se no Anfiteatro 1E, Campus Santa Mônica, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Engenharia Elétrica, assim composta: Professores Doutores: Gilberto Arantes Carrijo - FEELT/UFU; Lorenço Santos Vasconcelos - FEELT/UFU; Daniel Fernandes Macedo - UFMG; Alan Petrônio Pinheiro - FEELT/UFU, orientador(a) do(a) candidato(a).

Iniciando os trabalhos o(a) presidente da mesa, Dr(a). Alan Petrônio Pinheiro, apresentou a Comissão Examinadora e o candidato(a), agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

Aprovado(a).

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Gilberto Arantes Carrijo**, **Presidente**, em 20/02/2020, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alan Petronio Pinheiro**, **Professor(a) do Magistério Superior**, em 20/02/2020, às 18:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Lorenco Santos Vasconcelos**, **Professor(a) do Magistério Superior**, em 20/02/2020, às 23:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Daniel Fernandes Macedo**, **Usuário Externo**, em 27/02/2020, às 13:58, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1827433** e o código CRC **918A4890**.

Referência: Processo nº 23117.004450/2020-11 SEI nº 1827433

Agradecimentos

Aos meus pais e a minha irmã, pelo apoio ao longo da pós-graduação e pela paciência em lidar com a distância e a minha ausência em alguns momentos importantes.

Aos professores que contribuíram para a construção do meu conhecimento. Em especial, ao Professor Alan, pela oportunidade de desenvolver e compartilhar um projeto pesquisa na pós-graduação, pelos anos de orientação e suporte, e pelos ensinamentos e incentivos ao longo deste período.

Aos meus amigos, pelo suporte e parceria ao longo do desenvolvimento do projeto. Agradeço também aos meus colegas de trabalho que se esforçaram para que eu tivesse um tempo a mais para me dedicar aos estudos e, assim, desenvolver o projeto. Em especial, registro o meu agradecimento aos colegas Murilo Henrique e Diego Castanheira, que foram fundamentais para a conclusão do trabalho.

A Universidade Federal de Uberlândia, pela oportunidade de participar do Programa de Pós-graduação e pela estrutura disponibilizada. Do mesmo modo, a Algar Telecom, por disponibilizar os recursos necessários para o desenvolvimento deste projeto.

Resumo

A complexidade, abrangência e quantidade de dados produzidos pelas redes móveis têm exigido novos métodos para uma análise mais automatizada e capaz de buscar por padrões de anomalias não-óbvios. Neste sentido, esta pesquisa propôs um método de identificação de anomalias baseado em técnicas de redução de dimensionalidade e *clustering*, aplicados a um conjunto de dados composto por indicadores de desempenho. Este método emprega as técnicas PCA e DBSCAN, que são aplicados aos dados (indicadores) que foram segmentados por hora. Para efeitos de comparação, foi implementado um segundo método de referência que emprega as técnicas SOM e LoOP. Os resultados de ambos os métodos, o aqui proposto e o de referência, foram comparados. Foi usado um acervo de dados com seis indicadores de uma rede LTE em operação. Para averiguar o desempenho e a sensibilidade dos métodos, foram inseridas anomalias sintéticas baseadas em alguns cenários de degradação características das redes móveis. Além disto, as classificações de ambos os métodos foram, posteriormente, avaliadas por dois profissionais qualificados. O método utilizado obteve um bom desempenho na identificação das anomalias sintéticas inseridas se comparado com o método padrão. Neste sentido, os resultados encontrados encorajam o uso do método proposto como ferramenta complementar de análise para nortear a correção de falhas, otimizar recursos e indicar a expansão de capacidade da rede.

Palavras-chave: *Clustering*, *Data Analytics*, detecção de anomalias, redes móveis, redução de dimensionalidade.

Abstract

The complexity, coverage and amount of data produced by mobile networks have required new methods for a more automated analysis and searching for patterns of non-obvious anomalies. In this sense, this research proposed an anomaly identification method based on dimensionality reduction and clustering techniques, applied to a data set composed of performance indicators. This method employs the PCA and DBSCAN techniques, which are applied to a hour segmented real dataset. A second method, based on the structure of the proposed method, but with SOM and LoOP techniques was presented. This method can be seen as a reference method for comparing the proposed method, in order to validate the results found. The methods were evaluated using a data set with six indicators of an LTE network in operation. To investigate the performance and sensitivity of the methods, synthetic anomalies were inserted based on some degradation scenarios characteristic of mobile networks. In addition, network elements with a high number of anomalies were assessed by two qualified professionals. The methods performed well in the identification of the inserted synthetic anomalies, however, some limitations in the reference method regarding parameterization and the definition of a threshold of the techniques hinder their use with other data sets. Finally, the results encourage the use of methods as a tool to guide the correction of failures, optimize resources and indication of network capacity expansion.

Keywords: Clustering, Data Analytics, Anomaly detection, mobile networks, dimensionality reduction.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

Abreviaturas:

3GPP	*3rd Generation Partnership Project*
3G	*3rd Generation*
4G	*4th Generation*
5G	*5th Generation*
AVA	*Automation, Virtualized and Analytics*
BMU	*Best Matching Unit*
CS	*Circuit Switch*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DC	*Density-Connected*
DDR	*Directly Density-Rechable*
DR	*Density-Rechable*
DRB	*Data Radio Bearer*
ERB	Estação Rádio Base
FCM	*Fuzzy C-means*
ICA	*Independent Component Analysis*
kNN	*k-Nearest Neighbor*
KPI	*Key Performance Indicator*
LOCI	*Local Correlation Integral*
LOF	*Local Outlier Factor*
LoOP	*Local Outlier Probabilities*
LTE	*Long Term Evolution*
MDT	*Minimization of Drive Testing*
MIMO	*Multiple Input Multiple Output*
NGMN	*Next Generation Mobile Networks*
NNC	*Nearest Neighbor Clustering*
OPTICS	*Ordering Points to Identify the Clustering Structure*
OSS	*Operations Support System*
OTT	*Over The Top*
PCA	*Principal Component Analysis*
PRB	*Physical Resource Block*
PUCCH	*Physical Uplink Control Channel*
PUSCH	*Physical Uplink Shared Channel*
QoE	*Quality of Experience*
RAN	*Radio Access Network*
RRC	*Radio Resource Control*
RRU	*Radio Resource Utilization*

RSSI	*Received Signal Strength Indicator*
SBS	*Sequential Backward Selection*
SFS	*Sequential Forward Selection*
SINR	*Signal-to-Interference-plus-Noise Ratio*
SOM	*Selj-Organizing Map*
SON	*Selj-Organizing Networks*
SRB	*Signaling Radio Bearer*
SVM	*Support* Vector Machine
UE	*User Equipment*

Siglas:

CAPEX	*Capital Expenditure*
OPEX	*Operational Expenditure*

Sumário

1 Introdução

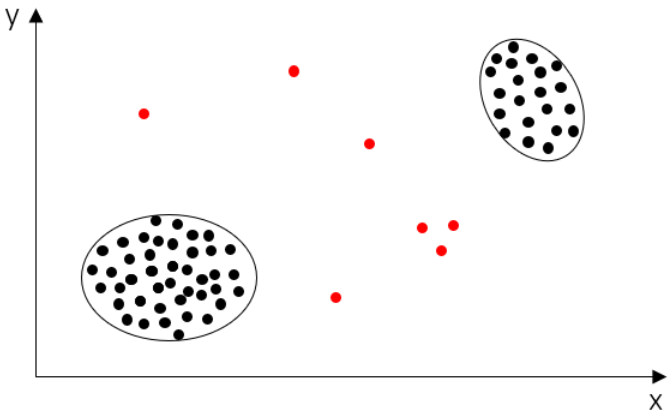
1.1 Apresentação

As redes celulares têm como uma de suas características a geração de uma grande quantidade de dados de gerenciamento de desempenho da rede. Isso se deve, em grande parte, ao aspecto de mobilidade e à transmissão de dados via interface aérea para um número, geralmente, grande e crescente de usuários. Esses dados são gerados continuamente a partir dos elementos da rede, como estações rádio base (ERB), e contabilizam, por exemplo, a quantidade de tentativas de requisição de conexão [1]. Estas informações são associadas a contadores que são usados no cálculo dos principais indicadores de desempenho, também nomeados por KPIs (*Key Performance Indicators*), e possuem um papel fundamental no planejamento, na otimização, no monitoramento, na detecção e na recuperação de falhas da rede.

Em um ambiente operacional, a degradação no desempenho é uma ocorrência comum em redes celulares e pode estar relacionada a vários fatores e causas. No monitoramento do desempenho da rede, a degradação é refletida principalmente nos KPIs e é caracterizada por apresentar padrões que não estão em conformidade com o comportamento esperado ou considerado como normal. Estes padrões são geralmente encontrados sob o nome de anomalias e podem ser achados em diversos domínios de aplicação [2]. No contexto de redes celulares, as anomalias podem ser geradas ou induzidas em um conjunto de dados por diversas razões, como a sobrecarga na utilização da rede, a interferência externa, as falhas na rede de transmissão (*backhaul*), dentre outras.

A Figura 1 ilustra as anomalias em um conjunto de dados com duas dimensões. As regiões com os dados normais estão delimitadas por uma elipse e os dados dispersos são as anomalias.

Figura 1 - Exemplo de anomalias em um gráfico de dispersão de um conjunto de dados com duas dimensões



Fonte: autoria própria

A detecção destes padrões anormais pode ser realizada de diversas maneiras. A seleção da técnica adequada depende diretamente das características do conjunto de dados e dos requisitos do problema [2]. Nas redes celulares, as técnicas de detecção de anomalias podem ser usadas em várias aplicações, uma das que se destacam é a detecção de falhas, contida nas funções de recuperação do *Self-Organizing Network* (SON) [3]. O SON é uma coleção de funções de autonomia de configuração, de otimização e de recuperação, considerada como essencial para as redes celulares atuais e futuras. O objetivo, neste caso, é prover resiliência e simplificação dos procedimentos de configuração, otimização e operação da rede, enquanto reduz o capital investido (CAPEX – *Capital Expenditure*) e o custo operacional para se manter (OPEX - *Operational Expenditure*) [4].

Este trabalho está inserido no contexto de detecção de anomalias aplicado em redes celulares, que foi explorado ao longo das próximas seções.

1.2 Objetivos da pesquisa

O monitoramento de desempenho das redes celulares é parte essencial para a qualidade de serviço prestado pelas operadoras e experiência dos usuários. Dessa forma, detectar os comportamentos anormais que causam esta degradação, de forma automatizada, contribui para a manutenção e melhoria do desempenho da rede. Nesse sentido, esta pesquisa teve como objetivo principal criar um método capaz de localizar os comportamentos anormais em um conjunto de KPIs de redes celulares e prover um meio *complementar* para que os operadores da rede se orientem melhor e identifiquem padrões nem sempre visíveis a uma análise imediata.

Vale destacar que o método proposto não tem a intenção de substituir outros existentes, mas, sim, de complementar o leque de opções. Isto é viável, pois, na prática, cada método produz resultados diferentes baseados nas características do método e no conjunto de dados em análise. Por isto, os métodos podem ser complementares na expectativa de buscar mais a fundo padrões de normalidade, usando diferentes meios para isto. Em virtude de todas estas questões e em decorrência do objetivo primário desta pesquisa, seguem outros objetivos derivados:

- avaliar e classificar, na prática, a existência de comportamento anormal nos elementos de rede que apresentam um elevado número de instâncias categorizadas como anômalas;
- comparar o resultado do método proposto com outro usado como referência para fins de validação de desempenho;
- avaliar a assertividade do método de detecção apresentado por meio da inserção artificial de anomalias que seguem alguns comportamentos característicos de degradação de redes celulares.

1.3 Limitações, convenções e escopo

Para que os propósitos básicos desta pesquisa pudessem ser alcançados, foi necessário delimitar algumas questões e estabelecer algumas convenções para que a metodologia proposta pudesse ser realizada dentro da realidade vigente. Dentre as principais limitações, pode-se citar:

- ainda que a quantidade de dados fosse julgada como relativamente grande, foram usados dados de uma única operadora;
- o conjunto de elementos de rede avaliados separadamente por dois profissionais foi limitado a trinta e quatro elementos, de forma a não prejudicar as atividades cotidianas na operadora, uma vez que esta análise é complexa e demanda considerável tempo.
- o método proposto foi desenvolvido com o intuito de apenas categorizar as instâncias de dados como anômala ou não. A classificação do tipo da anomalia não foi feita, pois necessita de um atributo rotulado previamente para uma etapa de treinamento, que permite fazer a distinção dos tipos das anomalias. Neste caso, o método proposto cria um atributo rotulado por meio da categorização, a partir de uma técnica não supervisionada.

Com relação às convenções adotadas, podem-se citar como principais:

- os métodos apresentados não foram desenvolvidos para aplicações em tempo real, visto que os KPIs são disponibilizados no OSS (*Operations Support System*) com pelo menos uma hora de atraso;
- não foi escopo do trabalho avaliar o tempo de execução dos algoritmos e otimizar os algoritmos tendo o tempo de execução como parâmetro;
- a análise visual e gráfica dos dados foi importante para o operador, pois como se trata de uma quantidade relevante de KPIs que devem ser analisados

simultaneamente, técnicas para redução de dimensionalidade puderam contribuir para este requerimento;

- o método apresentado foi desenvolvido para identificar anormalidades, isto é, o método pôde identificar tanto uma degradação ou falha quanto um desempenho excelente em algum elemento de rede.

1.4 Justificativa

Tradicionalmente, o monitoramento de desempenho da rede é baseado em limiares estabelecidos com base na experiência e no conhecimento do operador para identificar alguma degradação ou comportamento anormal. Como a rede possui diferentes configurações e apresenta um comportamento dinâmico que varia ao longo do dia, o monitoramento se torna limitado. Além disso, existe um grande número de KPIs armazenados nos OSS's (*Operations Support Systems*), e com a introdução de novas tecnologias pela quinta geração de redes celulares (5G) e a interoperabilidade com redes celulares de gerações anteriores, a quantidade de KPIs gerados tende a aumentar significativamente [5]. Analisar toda essa quantidade de indicadores em um ambiente de operação é impraticável. Neste cenário, uma alternativa eficiente é automatizar o processo de monitoramento dos KPIs, desenvolvendo métodos ou modelos de análises que identifiquem padrões nos dados que não estão em conformidade com o comportamento esperado.

1.5 Contribuição

Dentro do contexto previamente exposto, percebe-se que esta pesquisa pretendeu contribuir com a área de detecção de anomalias em redes celulares propondo um método para identificação de padrões de comportamento anormais em base de dados multivariada, considerando, simultaneamente, vários indicadores de rede. O método se

baseia em KPIs de redes celulares na forma de séries temporais, independente de qual medição represente, levando em conta o comportamento dinâmico da rede ao longo do dia. Além disso, fornecerá subsídios de análise para correção de eventuais falhas, otimização de recursos, planejamento e expansão de capacidade da rede móvel.

1.6 Organização do texto

No Capítulo 2 é apresentada a fundamentação teórica sobre o tema aqui pesquisado. Nele é dado enfoque ao gerenciamento do desempenho de redes celulares, abordando as medições e os indicadores, o monitoramento do desempenho, a detecção de falhas e o SON. Além disso, é destacado a identificação de anomalias e as abordagens utilizadas. Por fim, as técnicas de reconhecimento de padrões utilizadas nos métodos apresentados neste trabalho são detalhadas.

No Capítulo 3 é mostrado uma análise do estado da arte e apresentados os trabalhos mais relevantes encontrados na literatura acadêmica sobre *Data Analytics* no contexto de redes celulares e a constatação de anomalias aplicadas de forma ampla e restrita ao uso de conjunto de dados com indicadores de desempenho, também no contexto de redes celulares.

No Capítulo 4, a metodologia abordada nesta pesquisa é detalhada. Nele são apresentados a seleção de alguns conjuntos de dados para análise e a descrição da estrutura do método proposto. É também tratado como foi implementado um segundo método usado como referência. Em todos os casos são descritos as definições de parametrização adotadas, o detalhamento dos experimentos e os cenários de testes, e as validações e análises realizadas, na quais incluem a avaliação individual prática de alguns elementos de rede e a assertividade dos métodos.

O Capítulo 5 traz os resultados em forma de gráficos e tabelas com base nos experimentos, testes, validações e análises definidas. As discussões acerca dos resultados alcançados e suas limitações também foram tratadas neste capítulo.

Finalmente, o Capítulo 6 apresenta as conclusões obtidas e apresenta sugestões de trabalhos futuros.

2 Fundamentação teórica

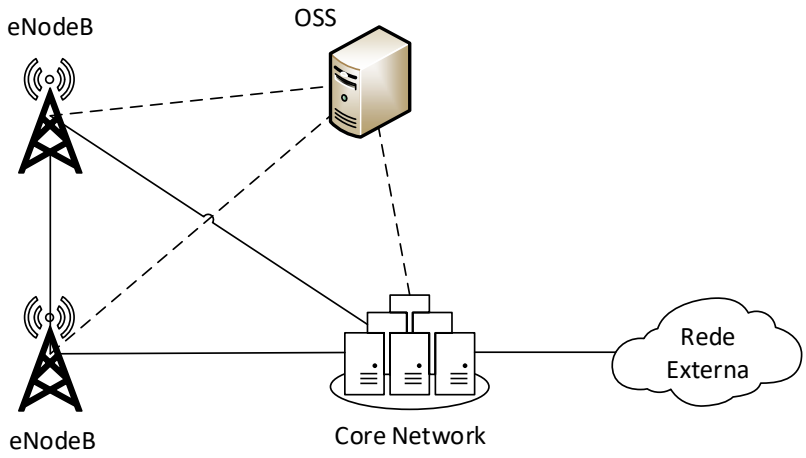
Neste capítulo são disponibilizadas informações sobre: (i) o gerenciamento do desempenho de redes celulares, (ii) a detecção de anomalias e (iii) o reconhecimento de padrões de anomalias. Nestes tópicos são abordados aspectos, técnicas e algoritmos no contexto de detecção de anomalias em redes celulares. Tais conhecimentos formam uma base técnica para o método proposto.

2.1 Gerenciamento do desempenho de redes celulares

2.1.1 Medições e indicadores

As redes celulares, por apresentarem as características de mobilidade e transmissão via interface aérea, geram uma grande quantidade de informações de desempenho. Estas informações são usadas para várias finalidades, como monitoramento de desempenho, detecção de falhas, otimização de recursos, planejamento futuro da rede, dentre outras. De forma simplificada, cada elemento da rede realiza medições periodicamente, associa a medição a um contador específico e envia esta informação em sistemas de suportes de operações OSS, como ilustrado na Figura 2. Neste caso, tanto os elementos da rede de acesso (RAN – *Radio Access Network*), representada pelas ERBs do 4G (eNodeB), quanto da rede núcleo (*Core Network*), enviam as informações.

Figura 2 - Arquitetura simplificada dos elementos da rede LTE e o OSS



Fonte: autoria própria

Os contadores, por sua vez, são organizados em categorias ou famílias definidas pelo 3GPP (*3rd Generation Partnership Project*) [6]. Especificamente, para a rede de acesso LTE (*Long Term Evolution*), os contadores de gerenciamento de desempenho são organizados em famílias, como *Data Radio Bearer* (DRB), *Radio Resource Control* (RRC), *Radio Resource Utilization* (RRU), entre outras. Estes agrupamentos de contadores citados, por sua vez, podem estar relacionadas ao número de UEs (*User Equipment*) ativos, as falhas de estabelecimento de conexão e a utilização de PRBs (*Physical Resource Blocks*), respectivamente, por exemplo [1].

O volume de dados de gerenciamento de desempenho gerado depende da quantidade de elementos ativos na rede; e geralmente é elevado. Neste sentido, são criados indicadores chaves de desempenho ou KPIs a partir dos contadores. Dessa forma, é possível criar indicadores de taxa de sucesso ou taxa de queda, por exemplo, o que facilita o monitoramento da rede de forma macro, sem a necessidade de analisar todos os contadores de falhas, a princípio.

De forma geral, os contadores podem ser gerados com periodicidade de 5, 15, 30 e 60 minutos, e serem agregados por dia, semana, mês e assim por diante. A habilitação

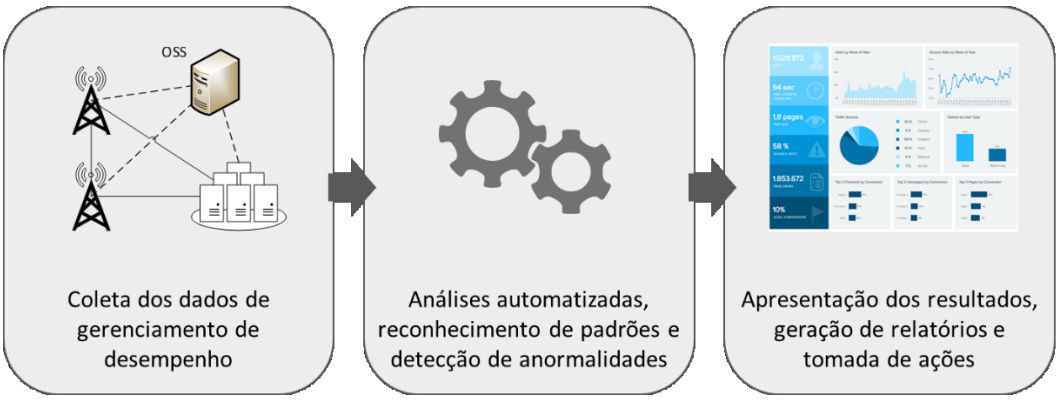
da periodicidade das medições fica a cargo das operadoras. Quanto menor a periodicidade, maior o volume de dados gerados por dia.

2.1.2 Monitoramento do desempenho e detecção de falhas

As informações de gerenciamento de desempenho são usadas para várias finalidades. Uma das principais é o monitoramento do desempenho e a detecção de falhas na rede. Tradicionalmente, o monitoramento do desempenho da rede depende do serviço de especialistas qualificados para identificar alguma falha ou comportamento anormal na rede. Para tanto, são definidos limiares estáticos para os KPIs; e os elementos que atingem este limiar são analisados individualmente para se determinar a falha [7]. Por ser uma atividade manual, o processo é ineficiente, dado o elevado número de elementos na rede de acesso e dos KPIs disponíveis.

Neste cenário, uma alternativa eficiente é automatizar o processo de monitoramento, desenvolvendo métodos ou modelos que analisam todos os elementos da rede (células, no caso), identificando padrões nos dados que não estejam em conformidade com o comportamento esperado, detectando e reportando estas informações para tratativa ou tomando ações automaticamente, como representado na Figura 3.

Figura 3 - Modelo de monitoramento de desempenho e detecção de falhas automatizado



Fonte: autoria própria

Neste contexto, várias soluções desenvolvidas na indústria são disponibilizadas comercialmente, como: o AVA (*Automation, Virtualized and Analytics*) (Nokia), que identifica e corrige proativamente as falhas e configura os elementos de rede para lidar com as mudanças nas condições da rede; o NetWarden (Bwtech), que correlaciona todas as fontes de dados coletadas na rede para fornecer automaticamente informações acionáveis para monitoramento, otimização e solução de problemas da rede; e o *Deep Network Analytics* (Amdocs), que combina informações de RAN com OSS e com dados de clientes para implantar a rede de maneira proativa. Outras soluções da indústria são as pesquisas acadêmicas relacionadas ao monitoramento e otimização da rede, a previsão, a detecção, a recuperação e a detecção de falhas, como descrito em [8].

2.1.3 Self-Organizing Cellular Networks

Na medida que a rede celular evolui e novos requerimentos e funcionalidades são inseridos, a complexidade de se configurar, operar e monitorar tais redes aumenta significativamente. Uma possível alternativa para este cenário é a introdução de inteligência na rede, conhecida como *Self-Organizing Networks* (SON). O conceito de SON nas redes celulares fora definida como uma rede que opera de forma autônoma e adaptativa, que seja escalável, estável e ágil o suficiente para manter os objetivos desejáveis, provendo resiliência e simplificação dos procedimentos de configuração, otimização e operação da rede. Ao mesmo tempo, deve reduzir o capital investido (CAPEX) e o custo operacional para se manter (OPEX) a estrutura constituída [9].

Portanto essas redes não apenas podem decidir independentemente quando e como determinadas ações serão acionadas, com base em sua interação contínua com o ambiente, mas também aprender e melhorar seu desempenho com base nas ações anteriores realizadas pelo sistema. O conceito de SON em redes móveis também pode ser dividido em três categorias principais: *self-configuration*, *self-optimization* e *self-healing* [9-10].

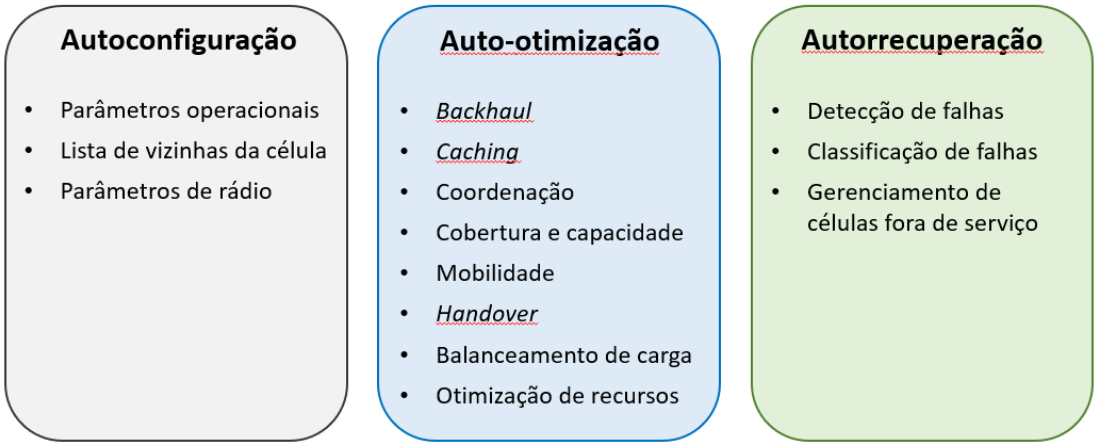
O *self-configuration* ou autoconfiguração pode ser definido como todos os procedimentos de configurações necessários para tornar a rede operacional. Sempre que um novo elemento é implantado ou existe alguma mudança na rede em operação (como falhas, mudanças de políticas e de *baseline*, por exemplo), os procedimentos de *self-configuration* são iniciados. No caso dos parâmetros de rádio da célula, por exemplo, é definido um *baseline* de forma que todas as células implantadas e em operação estejam de acordo com os valores definidos.

O *self-optimization* ou auto-otimização são formados por funções que otimizam continuamente os parâmetros de rede e se baseiam nas medições (contadores e KPIs) para obtenção do melhor resultado. O *self-optimization* é acionado após uma etapa de configuração e pode ser usado em várias aplicações, como otimização de cobertura, balanceamento de recursos de rádio, gerenciamento de interferência e mobilidade, eficiência energética, dentre outras.

O *self-healing* ou autorrecuperação é acionado sempre que ocorre alguma falha e tem como objetivo garantir uma rápida recuperação, de forma a minimizar os impactos. Como as falhas podem ocorrer inesperadamente, a rede é monitorada continuamente e as funções de *self-healing* além de detectar a falha devem também diagnosticar (determinar a causa) e acionar os mecanismos adequados para a solução. O uso do *self-healing* está relacionado com detecção de falhas, classificação de falhas e gerenciamento de interrupção de célula.

As funções SON são divididas em algumas subfunções, conhecidas como *use cases* ou casos de uso, representados na Figura 4. O desenvolvimento e a padronização dos casos de uso ficam a cargo de entidades como o 3GPP e o NGMN (*Next Generation Mobile Networks*) Alliance, que em conjunto com outras inciativas têm desenvolvido novas soluções no cenário de SON. Os trabalhos elaborados recentemente e as informações adicionais sobre SON podem ser consultados em [4] e [3].

Figura 4 - Casos de uso mais comuns de cada função SON



Fonte: adaptado de [4]

2.2 Detecção de anomalias

2.2.1 Anomalias

As anomalias são padrões que não seguem um comportamento definido como normal ou em conformidade com o comportamento esperado. Estes padrões são frequentemente encontrados sob o nome de *outliers* e a sua detecção é localizada em uma ampla gama de aplicações, como falhas em redes de telecomunicações, invasões em sistemas de segurança e fraudes em consumo de energia elétrica e cartões de crédito [2] e [10].

A formulação do problema e a escolha das técnicas apropriadas para detecção das anomalias são determinados por alguns fatores, como a natureza dos dados de entrada e a disponibilidade de rótulos (conhecidos também como classes) [2]. Os dados de entrada são formados por um conjunto de instâncias, que contém registros dos atributos (conhecidos também como variáveis ou dimensões). Os atributos, por sua vez, podem ser numéricos ou categóricos e determinam as técnicas aplicadas na identificação das anomalias. Por exemplo, para as técnicas de agrupamento baseadas em medidas de distâncias, os atributos de entrada devem ser numéricos. Um outro fator que influencia

na escolha das técnicas é a disponibilidade de rótulos nos dados de entrada. Nos casos em que os rótulos estão disponíveis, as técnicas de classificação com uma etapa de treinamento podem ser aplicadas, como regressão, árvore de decisão e redes neurais.

As anomalias podem ser encontradas em vários contextos e, portanto, categorizadas de várias formas, como as [10]:

- Anomalia de ponto: se uma instância de dados específica apresenta um comportamento anômalo em relação ao conjunto de dados, pode ser considerada uma anomalia de ponto. Por exemplo, quando ocorre uma queda instantânea aleatória na disponibilidade de serviço de um equipamento de uma rede telecomunicações.
- Anomalia contextual: se uma instância de dados apresenta um comportamento anômalo em um contexto específico, é denominada anomalia contextual ou condicional. Por exemplo, quando acontece um aumento no tráfego de dados de uma ERB que cobre uma região em que está ocorrendo um evento esportivo acompanhado por muitos torcedores.
- Anomalia coletiva: se uma coleção de instâncias de dados semelhantes apresenta um comportamento anômalo em relação a todo o conjunto de dados, o grupo de instâncias de dados é denominado anomalia coletiva. Por exemplo, quando ocorre uma queda na disponibilidade de serviço de várias ERBs devido a um problema na rede de transporte.

As anomalias podem ser encontradas como resultado das técnicas de detecção, tipicamente, em dois tipos: pontuação ou rótulo [2]. A pontuação representa o grau de anormalidade das instâncias avaliadas, por este motivo podem ser ordenadas e um analista seleciona as instâncias com os maiores valores ou define um limiar para categorização das anomalias. No caso do rótulo, as técnicas categorizam as instâncias

como anômalas ou não e, geralmente, são computacionalmente mais eficientes por não apresentarem a necessidade de atribuírem uma pontuação para cada instância.

2.2.2 Técnicas de detecção

As técnicas de detecção das anomalias são selecionadas com base em vários fatores e dependem diretamente das características do conjunto de dados e requisitos do problema. Neste sentido, várias abordagens têm sido propostas para identificar as instâncias normais e anômalas. A seguir é apresentado um resumo das técnicas e algoritmos comumente utilizados. Informações adicionais sobre a detecção de anomalias podem ser encontradas em [2] e [10].

As técnicas de detecção baseadas em classificação operam em duas fases (treinamento e teste) e dependem de um conjunto de instâncias rotuladas para aprender um modelo. Na fase de treinamento, as técnicas aprendem as características dos atributos a partir das instâncias identificadas como anômalas ou não anômalas e, então, na fase de teste, a instância é classificada com estes rótulos conforme a abordagem da técnica utilizada.

As técnicas baseadas em agrupamento que não dependem de um conjunto de instâncias rotuladas, baseiam-se em algumas suposições, como: (i) as instâncias de dados normais pertencem a um *cluster* ou grupo e as anomalias a nenhum *cluster*, (ii) as instâncias normais ficam próximas ao centroide do *cluster* mais próximo, enquanto as anomalias estão longe do centroide do *cluster* mais próximo, e (iii) as instâncias normais pertencem a *clusters* grandes e densos, enquanto anomalias pertencem a *clusters* pequenos ou esparsos.

Outra abordagem para a detecção de anomalias é o de vizinhos mais próximos, quando as técnicas requerem uma medida de distância entre as instâncias de dados, por exemplo, a distância euclidiana. As instâncias normais ocorrem em regiões densas em que a distância entre os vizinhos mais próximos é pequena, enquanto as anomalias ocorrem

em regiões dispersas em que a distância entre os vizinhos mais próximos é elevada.

Outro exemplo de medidas usadas na detecção de anomalias são as medidas teóricas da informação, como entropia, entropia condicional, entropia relativa, ganho e custo de informação. Estas medidas são utilizadas na análise do conjunto de dados para a identificação dos padrões de comportamento anormais.

As técnicas de detecção de anomalias estatísticas se baseiam na suposição de que as instâncias normais ocorrem nas regiões de alta probabilidade de um modelo estocástico e as anomalias ocorrem nas regiões de baixa probabilidade de um modelo estocástico. As técnicas estatísticas são divididas em paramétricas e não paramétricas, que são aplicadas para se ajustar a um modelo estatístico para determinar se uma instância pertence ou não a esse modelo.

Além desta abordagem, algumas técnicas exploram um subespaço de dimensão inferior e buscam uma diferença significativa entre as instâncias normais e anormais. Para tanto, tentam encontrar uma aproximação dos dados usando uma combinação de atributos que capturam a maior parte da variabilidade nos dados.

As anomalias apresentam um padrão de comportamento que não segue o que se conhece ou define como normal. A identificação destes padrões está relacionada com uma área de pesquisa conhecida como Reconhecimento de Padrões, na qual se concentra grande parte das técnicas e algoritmos de detecção de anomalias. Na próxima seção, alguns algoritmos são apresentados de forma detalhada com o intuito de prover uma base teórica para o entendimento do trabalho.

2.3 Reconhecimento de padrões

O reconhecimento de padrões tem como propósito o estudo de como as máquinas podem observar o ambiente, aprender a distinguir padrões de interesse e tomar decisões sobre as categorias dos padrões. Esclarecer os mecanismos de tomada de decisões é essencial no problema de reconhecimento de padrões, que pode ser entendido como uma

tarefa de classificação ou categorização, em que os rótulos ou classes são definidas pelo projetista do sistema, conhecida como classificação supervisionada, ou são aprendidas com base na semelhança dos padrões, conhecida como classificação não supervisionada [11].

No projeto de um modelo de reconhecimento de padrões, algumas características chaves devem ser observadas, como a definição das classes e a representação dos padrões, a extração e seleção das características do conjunto de dados, a análise de *clusters*, o projeto e aprendizado de classificadores, a seleção do conjunto de dados para treinamento e o teste e avaliação do desempenho do modelo. No contexto de detecção de anomalias, o reconhecimento de padrões é aplicado integralmente na identificação dos comportamentos normal e anômalo das instâncias de dados. Nessa perspectiva, nas próximas seções deste capítulo, as técnicas e algoritmos usados no desenvolvimento do trabalho são apresentados com algum detalhamento.

2.3.1 Redução de dimensionalidade

A redução da dimensionalidade de um conjunto de dados tem como resultado a extração e a seleção de características e visa dois pontos principais: custo de medição e precisão da classificação. Um número pequeno de dimensões ou atributos simplifica a representação de padrões, permite a criação de classificadores mais rápidos e reduz o problema da maldição da dimensionalidade[1]. Além disso, uma redução no número de dimensões pode evidenciar padrões "escondidos" no conjunto de dados. Por outro lado, pode levar também a uma perda na representatividade das dimensões e reduzir o desempenho dos classificadores.

[1] A maldição da dimensionalidade ou *curse of dimensionality* refere-se aos fenômenos que ocorrem com dados de alta dimensão. Dimensões extras podem causar uma queda no desempenho dos classificadores, mesmo que estas dimensões tenham informações úteis.

A extração de características resultante da redução de dimensionalidade se refere à criação de novos traços com base em transformações ou combinações do conjunto de dados original. A seleção de características concerne à seleção de uma parcela das dimensões geradas na redução de dimensionalidade do conjunto de dados original. A escolha entre a extração ou a seleção de características depende do contexto de aplicação e dos dados de treinamento disponíveis (no caso de classificadores supervisionados). Como exemplo de algoritmos de extração de características, pode-se citar: *Principal Component Analysis* (PCA) [12], *Independent Component Analysis* (ICA) [13] e *Selj-Organizing Map* (SOM) [14]. No caso dos algoritmos de seleção de características, menciona-se o *Sequential Forward Selection* (SFS) e o *Sequential Backward Selection* (SBS) [15]. Neste trabalho, foram usados algoritmos de extração de características, que são abordados especificamente a seguir.

2.3.1.1 Principal Component Analysis

A PCA é uma técnica simples e não paramétrica de redução de dimensionalidade, que transforma um conjunto de dados multivariados de possíveis variáveis correlacionadas em um conjunto ortogonal de componentes [16]. As componentes PCA, por sua vez, quantificam a importância de cada variável para descrever a variabilidade de um conjunto de dados, o que garante à técnica a capacidade de extrair informações relevantes.

A primeira etapa para encontrar as componentes é calcular a matriz de covariância A de um conjunto de dados X com n instâncias e p variáveis, ou seja, uma matriz n x p. O resultado é uma matriz p x p, conforme a equação abaixo.

$$A_{p,p} = cov(X_{n,p}) \tag{2.1}$$

A partir da matriz de covariância A são calculados os autovalores u_p e autovetores $v_{p,p}$, que representam a magnitude e a direção da variância dos dados, respectivamente.

Os autovetores são unitários, ortogonais entre si e formam uma matriz p x p (Equação (2.2)). Os autovalores são associados aos autovetores e indicam o quanto da variância do conjunto de dados é representada ou explicada pelos autovetores. As componentes são combinações lineares de todas as variáveis formadas pela associação entre os autovalores e autovetores, produzindo novas variáveis não correlacionadas, de forma que as componentes principais são determinadas pelos autovalores com alta magnitude.

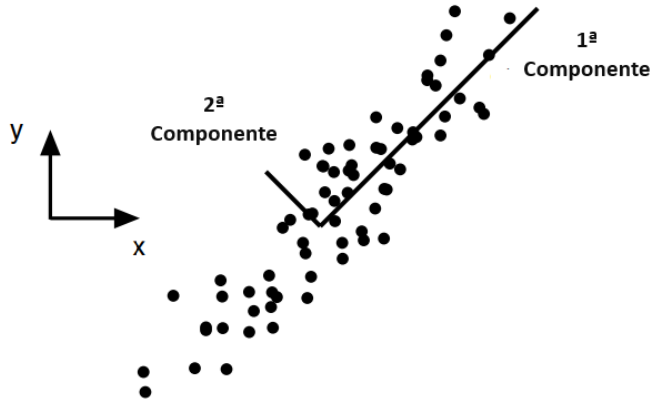
$$autovalores = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \dots \\ u_p \end{pmatrix} \quad autovetores = \begin{pmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \dots & v_{1,p} \\ v_{2,1} & v_{2,2} & v_{2,3} & \dots & v_{2,p} \\ v_{3,1} & v_{3,2} & v_{3,3} & \dots & v_{3,p} \\ \dots & \dots & \dots & & \dots \\ v_{p,1} & v_{p,2} & v_{p,3} & \dots & v_{p,p} \end{pmatrix} \tag{2.2}$$

Uma vez que i componentes são selecionadas, o conjunto de dados final X' é o resultado da multiplicação entre o conjunto de dados X e a matriz de autovetores transposta $v_{i,p}^T$, conforme a Equação (2.3).

$$X'_{n,i} = X_{n,p} . v_{i,p}^T \tag{2.3}$$

A Figura 5 ilustra as duas componentes PCA geradas a partir de um conjunto de dados de duas variáveis. A maior variância por qualquer projeção dos dados fica ao longo da primeira componente e a próxima maior variância fica ao longo da segunda componente e assim por diante, nos casos em que o conjunto de dados contenha mais de duas variáveis.

Figura 5 – Gráfico de dispersão de duas variáveis e a direção das duas componentes PCA geradas

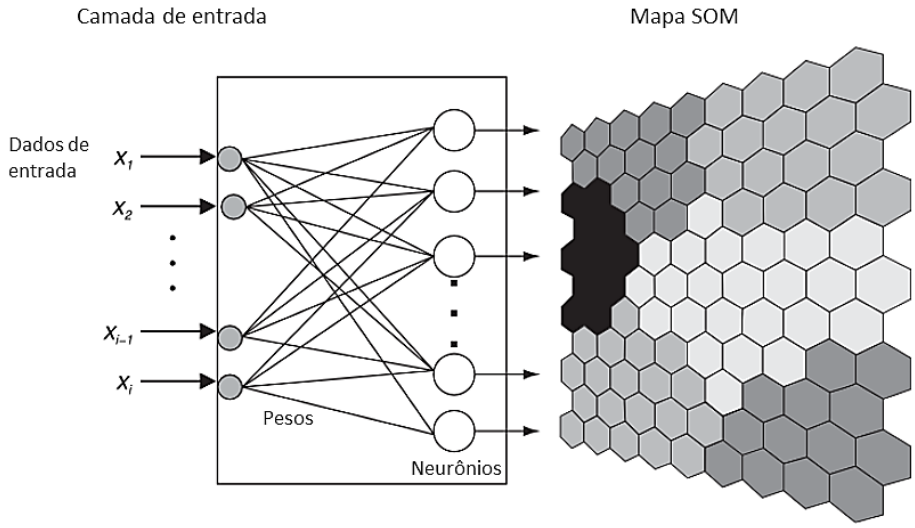


Fonte: adaptado de [16]

2.3.1.2 Self-Organizing Map

O SOM é um tipo de rede neural não supervisionada, usada, principalmente, no agrupamento de dados e extração de características de um conjunto de dados [14]. O SOM possui uma camada de neurônios ou 'nós', na qual cada nó tem sua posição no mapa, e um vetor de 'livro de códigos' (*codebook*), formados pelos pesos nas conexões dos neurônios, como mostra a Figura 6.

Figura 6 – Representação de uma rede SOM



Fonte: adaptado de [17]

Os *codebooks* têm as mesmas dimensões que os dados de entrada e são conectados a todas as x_i unidades de entrada. A saída da rede SOM é uma representação de baixa dimensão (geralmente duas) dos dados de entrada, de forma que os dados mais semelhantes são associados em nós adjacentes, enquanto dados menos similares são situados mais afastados um do outro na rede. Essa distribuição permite obter uma visão das relações topográficas dos dados, especialmente dos itens de dados de alta dimensão.

O SOM pode ser inicializado de forma aleatória ou linear, em que os valores são atribuídos para as dimensões de cada vetor *codebook*. Na fase de treinamento, os vetores de entrada (ou instâncias) *x(t)* são apresentados à rede em uma ordem aleatória e são comparados com os vetores *codebook* m_i, com base em uma medida de distância (geralmente Euclidiana), para encontrar o vetor *codebook* mais próximo m_c (Equação (2.4)), conhecido como BMU (*Best Matching Unit*).

$$\|x - m_c\| = min_i \{\|x - m_i\|\} \tag{2.4}$$

Em seguida, os pesos dos neurônios são atualizados de forma que seus vetores *codebook* se movam na direção do vetor de entrada. Para tanto, é calculado o raio da vizinhança R_n que decresce de acordo com o número de épocas ou iterações, conforme a Equação (2.5), em que σ_o é o raio inicial definido para a rede e λ é o tempo dado pela razão entre o número de épocas e o raio do mapa.

$$R_n\big(\sigma(t)\big) = \sigma_o . e^{-t/\lambda} \tag{2.5}$$

Os pesos são atualizados conforme a Equação (2.6) para se assemelharem ao vetor de entrada, dessa forma os neurônios na vizinhança próxima ao BMU sofrem alta mudança nos seus pesos. Neste caso, α(t) é a taxa de aprendizado que decresce ao longo das épocas e h_{ci}(t) é uma função gaussiana para o cálculo da distância.

$$m_i(t+1) = m_i(t) + \alpha(t).h_{ci}(t).(I(t) - w(t)) \quad (2.6)$$

O treinamento finaliza quando o número de épocas é atingido ou algum parâmetro de erro é minimizado. Após o término desta etapa, os vetores de *codebook* dos neurônios vizinhos na rede representam padrões de entrada próximos do espaço original e a topologia dos dados é preservada. Além dos vetores *codebook*, a rede SOM tem como saída algumas informações importantes, como a distância entre nós (matriz U) e o número de instâncias de entrada atribuídas em cada neurônio.

2.3.2 Clustering

O *clustering* ou agrupamento, conhecido também como classificação não supervisionada, é uma técnica empregada para construir limites de decisão com base em dados de treinamento não rotulados, visto que em várias aplicações é difícil (ou mesmo impossível) rotular de forma confiável uma amostra de treinamento com sua verdadeira categoria. A definição dos grupos ou *clusters* é baseada em uma medida de similaridade, como a distância entre os dados. Os dados que fazem parte do mesmo *cluster* tendem a possuir alguma característica ou propriedade em comum e no cenário de detecção de anomalias podem representar o padrão normal ou anormal dos dados.

Na literatura, várias técnicas e algoritmos de agrupamentos foram propostos e ainda continuam sendo desenvolvidos. Abaixo segue um resumo dos tipos de algoritmos e/ou abordagens comumente utilizados.

- Particionais, que produzem *clusters* pela otimização de uma função critério definida localmente ou globalmente (exemplos: LBG – Linde-Buzo-Gray [18], K-*means* [19] e K-*medoids* [20]);
- Hierárquicos, que organizam os dados em uma sequência aninhada de grupos que podem ser exibidos na forma de uma árvore (exemplos: *Single Linkage* [21] e *Complete Linkage* [22]);

- *Fuzzy* ou difuso, que associa cada padrão a cada *cluster* usando uma função de associação, de forma que os novos *clusters* formados podem ser sobrepostos (exemplo: FCM – *Fuzzy* C-*means* [23]);
- Baseado em densidade, que organizam os *clusters* com base na proximidade dos dados a partir do raio de vizinhança e do número mínimo de pontos em cada *clusters* (exemplos: DBSCAN - *Density-Based Spatial Clustering of Applications with Noise* [24] e OPTICS - *Ordering Points to Identify the Clustering Structure* [25]);
- *Mixture-resolving*, em que os *clusters* são formados a partir de várias distribuições dos dados, com o objetivo de identificar os parâmetros e números de cada uma (exemplo: *Model-based* [26]);
- Vizinhos mais próximos, em que os dados não processados são atribuídos ao *cluster* do vizinho mais próximo processado (exemplo: NNC – *Nearest Neighbor Clustering* [27]).

2.3.2.1 Agrupamento baseado em densidade

No agrupamento baseado em densidade, um *cluster* é formado em qualquer direção conduzido pela densidade dos pontos (leia-se dados). Os algoritmos baseados em densidade são capazes de identificar *clusters* de formas arbitrárias e procuram por regiões de alta densidade separados por regiões de baixa densidade. Um dos primeiros algoritmos implementados neste contexto foi o DBSCAN [24].

A ideia central do DBSCAN está associada à noção de *density-reachability* (acessibilidade em densidade) e depende dos parâmetros *Eps* (raio da vizinhança de um ponto p) e *MinPts* (número mínimo de pontos da vizinhança). Tais parâmetros são matematicamente associados como:

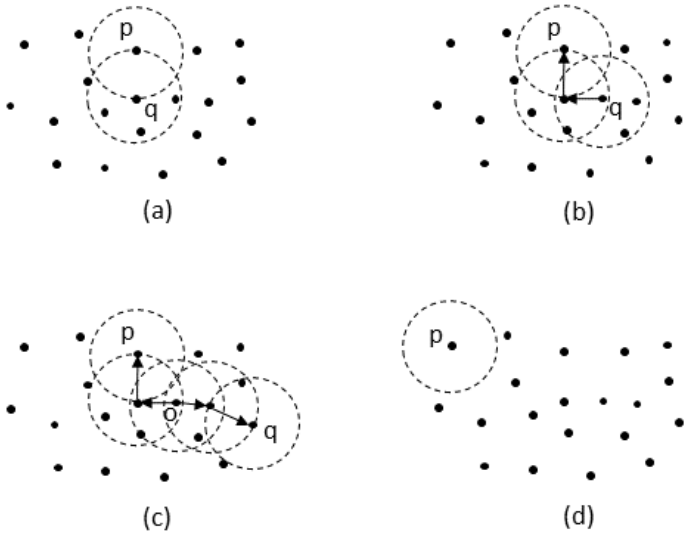
$$V_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\} \tag{2.7}$$

$$p \in V_{Eps}(q) \tag{2.8}$$

$$|V_{Eps}(q)| \geq MinPts \tag{2.9}$$

A vizinhança de um ponto p é definida conforme a Equação (2.7), em que D é o conjunto de pontos com distância (Euclidiana, por exemplo) menor ou igual ao parâmetro *Eps*. Um ponto p é considerado *directly density-rechable* (DDR - diretamente alcançável em densidade) de um ponto q com base nas Equações (2.8) e (2.9). Se o número de pontos DDR de um ponto p é maior que *MinPts*, este ponto é um *core point* (ponto de centro), como é o caso do ponto q na Figura 7a. Caso o número de pontos seja menor que *MinPts*, este é um *border point* (ponto de borda), ilustrado na Figura 7a pelo ponto p. Um ponto p é *density-rechable* (DR - alcançável em densidade) de um ponto q caso exista uma sequência de pontos $p_1, p_2, ..., p_n, p_1 = q, p_n = p$ em que $p_1 + 1$ é DR de p_i, como mostra a Figura 7b. Dois pontos p e q são *density-connected* (DC - conectado em densidade) se são DR de um ponto o, como são ilustradas na Figura 7c.

Figura 7 - Representação das definições de diretamente alcançável em densidade (a), alcançável em densidade (b), conectado em densidade (c) e um ponto categorizado como ruído (d)

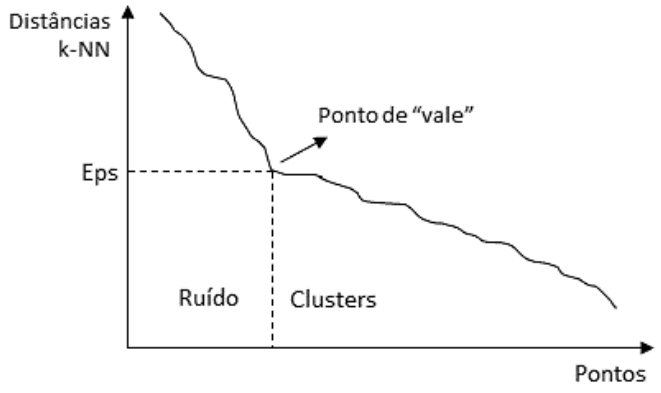


Fonte: adaptado de [24]

Um *cluster* é formado se o número de pontos na vizinhança *Eps* de p for maior que *MinPts*. Neste cenário, os pontos DC de p são adicionados ao *cluster*. Caso contrário, um ponto não processado na base de dados é selecionado. Um ponto será considerado como ruído (*noise point*) se não for DC de outro ponto e, portanto, não pertencer a algum *cluster*, como mostra a Figura 7d. Ao fim do processo, os pontos categorizados como ruídos tendem a apresentar diferentes propriedades dos pontos pertencentes a um *cluster*. No contexto de detecção de anomalias, estes pontos são interpretados como anomalias.

O algoritmo DBSCAN depende diretamente da escolha dos parâmetros *Eps* e *MinPts* para a formação dos *clusters* e categorização dos pontos ruídos. Geralmente um estimador razoável de densidade deve contar com as distâncias dos k-vizinhos mais próximos, em que k é igual ao *MinPts*. O valor de *Eps* pode ser encontrado com base no gráfico das distâncias ordenadas, a partir do ponto do primeiro "vale", representado na Figura 8. Os pontos com uma distância do vizinho mais próximo menor que o valor de *Eps* são atribuídos a algum *cluster*, enquanto que os outros pontos são categorizados como ruído [24].

Figura 8 – Gráfico das distâncias dos k-vizinhos mais próximos ordenadas e o ponto de "vale" definido como o valor do parâmetro Eps



Fonte: adaptado de [24]

2.3.3 Detecção de outliers locais

Uma instância que parece ser inconsistente ou que desvia de forma acentuada das outras instâncias de um conjunto de dados apresenta um comportamento definido como *outlier*. Esta característica remete ao conceito de anomalia, que na literatura, os termos *outlier* e anomalia, são geralmente encontrados no contexto de detecção de anomalias [2]. Inclusive a detecção de *outliers* abrange aspectos de uma ampla variedade de técnicas, que são aplicadas em abordagens fundamentalmente idênticas, como detecção de anomalias, detecção de ruído, detecção de novidade e detecção de desvio ou mineração de exceção [28].

As abordagens para se identificar *outliers* se resumem em métodos de detecção (i) globais e (ii) locais, que, respectivamente, (i) produzem rótulo binários e analisam o conjunto de dados como um todo e (ii), geralmente, atribuem pontuações, que representam um "grau de comportamento *outlier*", além de analisar uma seleção local de instâncias do conjunto de dados, baseada na vizinhança. Comumente os métodos locais são mais flexíveis, pois também podem produzir rótulos. Como exemplo de técnicas de detecção de *outliers* locais, pode-se citar: *Local Outlier Probabilities* (LoOP) [29], *Local Outlier Factor* (LOF) [30] e *Local Correlation Integral* (LOCI) [31].

2.3.3.1 Local Outlier Probabilities

O LoOP é um método desenvolvido para detecção de *outliers* baseado na densidade local dos dados, que combina a ideia de uma pontuação referente ao grau que uma instância de dados possa ser identificada como um *outlier*, como LOF e LOCI, junto a conceitos probabilísticos. O resultado é uma pontuação que varia de 0 a 1, que pode ser diretamente interpretada como uma probabilidade de uma instância de dados ser um *outlier* [29].

O primeiro passo do método de detecção de *outliers* é calcular a distância probabilística *pdist(o,S)* de uma instância de dados o de um conjunto de dados S dos k-vizinhos mais próximos de o, conforme a Equação (2.10), em que σ é o desvio padrão e

λ controla a aproximação da densidade, mas é um fator de normalização e não afeta a identificação dos *outliers*.

$$pdist(o,S) = \lambda.\sigma(o,S) \tag{2.10}$$

A seguir é calculado o valor de PLOF (*Probabilistic Local Outlier Factor*), que é a razão entre a distância probabilística das instâncias em torno de o e o valor esperado das estimativas para as distâncias em torno de todos os objetos no conjunto de dados S, de acordo com a equação abaixo:

$$PLOF_S(o) = \frac{pdist(o,S(o))}{E_{s\epsilon S(o)}.[pdist(s,S(s))]} - 1 \tag{2.11}$$

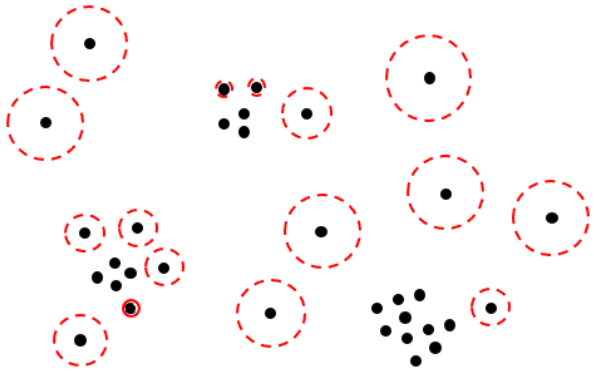
Por fim, os valores de PLOF são normalizados e convertidos em valores de probabilidade, conforme as equações a seguir:

$$LoOP_S(o) = \max\left\{0, erf\left(\frac{PLOF_S(o)}{nPLOF.\sqrt{2}}\right)\right\} \tag{2.12}$$

$$nPLOF\ \sqrt{2} = \lambda.\sqrt{E[(PLOF)^2]} \tag{2.13}$$

O valor de LoOP será próximo a 0 para instâncias localizadas em regiões densas e próximo a 1 para instância localizadas em regiões dispersas, isto é, distantes de outras instâncias do conjunto de dados, conforme o raio do círculo exemplificado na Figura 9.

Figura 9 – Representação gráfica da dispersão de algumas instâncias de dados (ou pontos), em que o raio do círculo pontilhado representa a pontuação ou probabilidade do ponto ser um *outlier*



Fonte: autoria própria

2.4 Resumo do capítulo

Neste capítulo foram apresentados conceitos, definições e aspectos importantes no contexto em que o trabalho se situa, como o gerenciamento de desempenho de redes celulares, desde a geração até a aquisição dos dados de medição aos modelos de SON, mostrando como são essenciais para a melhoria e manutenção do desempenho das redes. Ademais, foram descritas as características das anomalias, as abordagens e as técnicas utilizadas na sua detecção. Por fim, foram descritas as técnicas selecionadas para o desenvolvimento deste trabalho, inseridas no contexto de Reconhecimento de Padrões.

3 Estado da arte e pesquisas correlatas

Neste capítulo é apresentado uma visão geral das pesquisas desenvolvidas no contexto em que o presente trabalho se insere. Desta maneira, são citados trabalhos importantes publicados tendo como aplicação a detecção de anomalias em redes celulares, a partir de uso de técnicas multidisciplinares do campo de *Data Analytics*. Como o trabalho aqui descrito se baseia em indicadores de desempenho gerados pelos elementos da rede de acesso, foi destinada uma seção para a citação de pesquisas desenvolvidas neste mesmo contexto.

3.1 Data Analytics no contexto de redes celulares

A introdução de novas funcionalidades e tecnologias, como MIMO (*Multiple Input Multiple Output*) massivo, virtualização de rede, propagação em ondas milimétricas (*mmWaves*) e densificação da rede de acesso, têm tornado as redes celulares cada vez mais complexas. Consequentemente as tarefas de manutenção, operação, otimização e planejamento têm se tornado cada vez mais desafiadoras. Neste cenário, as análises estatísticas básicas se tornam limitadas e o uso de técnicas de reconhecimento de padrões, inteligência computacional, aprendizado de máquina, estatística avançada e teoria de sinais, apresentam-se como uma boa alternativa. Todos estes campos interdisciplinares convergem para a área de *Data Analytics*, que concentra várias disciplinas científicas na análise de conjuntos de dados para o suporte na tomada de decisões [32].

No contexto de rede celulares, várias abordagens têm sido propostas com base em técnicas de *Data Analytics*. A área de SON engloba boa parte dos trabalhos desenvolvidos nos casos de uso das funções de autoconfiguração, auto-otimização e auto-recuperação,

que podem ser consultados em [4]. Especificamente para a auto-recuperação, os autores em [3] apresentam uma revisão, demonstrando os desafios e as pesquisas relacionadas com a detecção e classificação de falhas, e o gerenciamento de interrupção de célula para redes celulares emergentes.

Para lidar com a quantidade massiva de dados brutos gerados nas redes celulares e extrair informações úteis, várias abordagens têm sido propostas sob a ótica de *Big Data Analytics*. Os autores em [33] apresentam uma visão geral de *Big Data Analytics* baseada na teoria de matriz aleatória e numa arquitetura de *framework* para aplicações em redes celulares. Além disso, os autores apresentam alguns desafios e perspectivas de pesquisas para redes celulares da próxima geração. No contexto de *Big Data Analytics* e detecção de anomalias, em [34] é desenvolvida uma plataforma de identificação de anomalias e análise de causa raiz, fundamentada na mineração de regras de associação combinadas com abordagens estatísticas para identificar as anomalias. Por fim, uma revisão detalhada dos atuais esforços acadêmicos e industriais direcionados para previsão, detecção, recuperação e prevenção de falhas, monitoramento e otimização de rede, *cache* e entrega de conteúdo usando *Big Data Analytics* podem ser encontradas em [8].

3.2 Detecção de anomalias

A detecção de anomalias pode ser realizada de diversas formas, seja por meio de análises estatísticas, mineração de dados, aprendizado de máquina, dentre outras. No contexto de redes móveis, diferentes abordagens têm sido propostas. Em [35], por exemplo, são comparados cinco algoritmos de aprendizado não supervisionado (K-*means*, FCM, LOF, LoOP e SOM) com o intuito de detectar falhas na rede. Os resultados mostraram que o SOM apresentou desempenho superior ao K-*means* e FCM. Os autores em [36] avaliaram o desempenho de três técnicas de aprendizado supervisionado (árvore de decisão, kNN - *k-Nearest Neighbor* e SVM - *Support Vector Machine*) para prever o nível de QoE (*Quality of Experience*) do usuário final, usando as informações obtidas

para detectar anomalias em ERBs. A avaliação das técnicas foi realizada a partir de um cenário de teste criado em um simulador de rede.

Outras abordagens usam análises ou métodos estatísticos na detecção de anomalias, por exemplo, um comparativo de perspectivas baseadas em entropia com um método de distribuições empíricas de probabilidade para detectar e diagnosticar anomalias em grande escala de uma rede celular real, causada por serviços específicos OTT (*Over The Top*) e dispositivos de *smartphone* [37]. Os autores em [38] propõem um método com mecanismos baseados em correlação para identificar anomalias na evolução temporal de indicadores. Na avaliação do método foram definidos alguns padrões sintéticos de degradação que foram comparados com os dados de entrada. Por fim, em [39] é apresentado um módulo de análise de eventos originados pelo equipamento do usuário (UE) e a ERB servidora, baseado nas técnicas kNN, SOM e estruturas de dados probabilísticas (*Locai-Sensitive Hashing*, *Probabilistic Anomaly Detection*) para a detecção de *sleeping cells* (células fora de serviço que não reportam alarmes deste estado). A validação é realizada por meio de um simulador, usando a funcionalidade *Minimization of Drive Testing* (MDT).

3.2.1 Detecção de anomalias baseada em indicadores de desempenho

O método de detecção de anomalias desenvolvido neste trabalho se baseou no uso de indicadores de desempenho (KPIs) de redes celulares. Neste contexto, alguns autores [7] combinam várias funções estatísticas para criar medidas como pontuação (*i.e., score*) e probabilidade para determinar a importância e a certeza das anomalias detectadas. Em [40] algumas funções estatísticas também são usadas como média e desvio padrão, mas com o objetivo de determinar o perfil de comportamento do KPI. A detecção das anomalias neste caso é realizada numa etapa de *clustering* baseado em densidade.

Em [41], [42], por exemplo, os autores propõem um *framework* com um método *ensemble* adaptativo que analisa tanto conjuntos de dados univariados (um KPI por vez)

como multivariados (mais de um KPI por vez), indicando a severidade da degradação das células por meio de diferentes métodos preditivos em conjunto com informações de gerenciamento de configuração da rede. Outra abordagem de detecção de anomalias [43] apresenta uma estrutura de identificação de anomalias e análise de causa raiz em várias camadas, fundamentada em um conjunto de dados de serviço de vídeo chamada. A partir destes dados é calculado um fator de anormalidade de QoE, e com base em alguns KPIs é aplicado o método *Sequence Pattern Mining* para predizer a degradação de QoE com precisão e detectar uma incompatibilidade de células vizinhas.

Um aspecto importante na detecção de anomalias em KPIs é que, ao longo do dia, estes valores podem apresentar variação, especialmente em virtude da dinâmica que estas redes têm com seus usuários. Além disso, a anomalia pode estar relacionada ao perfil de comportamento entre os KPIs e a análise multivariada pode identificar anomalias não encontradas na análise individual dos KPIs (univariada). Neste cenário, este trabalho propôs uma nova abordagem para a detecção de anomalias, apoiada na estratificação temporal e na redução da dimensionalidade do conjunto de dados, de forma que a dinâmica da rede e as relações entre os KPIs sejam consideradas e a identificação das anomalias realizadas com base na densidade dos dados em cada período.

3.3 Resumo do capítulo

Neste capítulo foram apresentadas as pesquisas desenvolvidas no contexto em que o presente trabalho está inserido. Inicialmente foram citadas algumas pesquisas que apresentam uma revisão das publicações no contexto de *Data Analytics* aplicado em redes celulares. Em seguida, foram apresentadas algumas pesquisas relacionadas à detecção de anomalias em redes celulares, que comparam várias técnicas de *Data Analytics* para obterem o resultado mais assertivo para o caso de uso em estudo. Por fim, foram mencionadas as pesquisas baseadas em KPIs, que se encontram em contextos aproximados deste trabalho, mostrando como é uma área rica em esforços

para pesquisa, pois oportuniza que variados meios podem ser utilizados para se conseguir resultados.

4 Metodologia e desenvolvimento

Neste capítulo é apresentada a ideia básica do método proposto para detecção das anomalias com base na análise de múltiplos KPIs da rede móvel. É importante destacar que, como descrito no Capítulo 1, o objetivo deste trabalho foi o de propor um meio *complementar* (e não substituto) para se identificar comportamentos anômalos em células de redes móveis.

Neste sentido, este capítulo trata da descrição do método proposto, identificando e descrevendo seu procedimento e a forma como analisa o problema e o tratamento dos dados. Para tanto, realizou-se alguns experimentos nas etapas de redução de dimensionalidade e agrupamento, seguidos pela validação do método. Essa abordagem possui um caráter didático e ilustrativo para que o leitor consiga avaliar o funcionamento das etapas do método. Além disso, com o intuito de melhor caracterizar e validar os resultados do método aqui proposto, um segundo método foi apresentado, e os desempenhos na detecção de anomalias de ambos os métodos foram comparados.

Tendo por base este elenco de conteúdos, este capítulo pôde ser subdividido em três partes, a saber: a primeira descreve o método e seus procedimentos, a segunda parte descreve alguns experimentos usados para avaliar os resultados alcançados pelo método proposto, e a terceira parte compara o desempenho com um segundo método de detecção de anomalias.

4.1 Materiais: dados empregados

4.1.1 Base de dados de KPIs

Os dados usados neste trabalho foram extraídos diretamente do OSS e são formados por KPIs ao nível de célula. A periodicidade desta amostragem de indicadores

da rede é horária. Isto quer dizer que cada célula gera um valor por hora para cada KPI. Ao todo foram coletados KPIs de 1.054 células de uma rede 4G e 3.339 células de uma rede 3G, ambas em operação. Cada linha da tabela (arquivo no formato .csv) tem registado, além dos KPIs do período de registro, o dia e horário que foram feitas as coletas dos indicadores das células e os códigos das células. Estes dados foram cedidos pela operadora telefônica Algar Telecom e coletados entre os dias 01/01/19 à 14/03/19. A Figura 10 ilustra um exemplo de parte de um conjunto de dados extraído diretamente do OSS.

Figura 10 - Exemplo de um conjunto de dados com três células, três KPIs e sete instâncias extraído do OSS

	A	B	C	D	E	F
1	Dia	Hora	Célula ID	KPI 1	KPI 2	KPI 3
2	03.01.2019	00:00:00	1	3	9130.66	14.72
3	03.01.2019	00:00:00	2	2	5569.75	14.55
4	03.01.2019	00:00:00	3	6	9055.49	14.75
5	03.01.2019	01:00:00	1	2	9360.88	13.72
6	03.01.2019	01:00:00	2	1	4418.65	15.09
7	03.01.2019	01:00:00	3	4	7687.78	14.85

Fonte: autoria própria

Como foram usadas diferentes tecnologias, os KPIs foram organizados em categorias de três grupos, que foram numerados de 1 a 3. São eles:

- **Grupo 1**: conglomera alguns KPIs essenciais, geralmente analisados para avaliação da utilização de rede e da qualidade de sinal da tecnologia 4G, como mostrados na Tabela 1. Essencialmente, incluem (para cada célula): número de usuários ativos, taxa de dados média (no *downlink*), a relação sinal-ruído-interferência (SINR – *Signal-to-Interference-plus-Noise Ratio*) média e os indicadores de intensidade do sinal recebido (RSSI - *Received Signal Strength Indicator*) médio no canal físico de controle no *uplink* (PUCCH - *Physical Uplink Control Channel*) e no canal físico compartilhado no *uplink* (PUSCH – *Physical Uplink Shared Channel*).

Tabela 1 – Atributos do conjunto de dados 1 (de redes móveis 4G)

Atributo	Formato	Faixa
Dia	Data	[01/03/2019 a 14/03/2019]
Hora	Hora	[00:00 a 23:00]
Célula ID	Categórico	[1 a 1054]
Número médio de usuários ativos	Numérico	[0 a 397]
Taxa de dados média no downlink	Numérico	[0.00 a 68693.46]
SINR para PUCCH médio	Numérico	[-8.74 a 27.00]
SINR para PUSCH médio	Numérico	[-10.00 a 30.00]
RSSI para PUCCH médio	Numérico	[-119.90 a -80.05]
RSSI para PUSCH médio	Numérico	[-120.00 a -80.00]

- **Grupo 2**: formado por KPIs da tecnologia 4G, em especial os relacionados ao *Radio Bearer* da rede 4G. O *Radio Bearer* são canais que transportam dados de controle (*Signaling Radio Bearer* - SRB) e do usuário (*Data Radio Bearer* - DRB) para camadas superiores. Os KPIs selecionados para este conjunto de dados são referentes ao processo de configuração e estabelecimento do *Radio Bearer*. Este conjunto de dados está representado na Tabela 2.

Tabela 2 - Atributos do conjunto de dados 2 (de redes móveis 4G)

Atributo	Formato	Faixa
Dia	Data	[07/01/2019 a 13/01/2019]
Hora	Hora	[00:00 a 23:00]
Célula ID	Categórico	[1 a 1054]
Número de tentativas de configuração de DRBs	Numérico	[1 a 247553]
Número de DRBs estabelecidos com sucesso	Numérico	[1 a 247403]
Número de falhas de configuração de DRBs	Numérico	[0 a 17686]
Tempo máximo de confi. de Enhanced-Radio Access Bearer	Numérico	[22 a 5433]
Tempo médio de confi. de Enhanced-Radio Access Bearer	Numérico	[22 a 1446]
Número de DRBs liberados devido ao Detach	Numérico	[0 a 1220]
Número de tentativas de configuração de SRB 1	Numérico	[1 a 59455]
Número de falhas de configuração de SRB 1	Numérico	[0 a 1032]
Número de conclusões de configuração de SRB 1	Numérico	[1 a 59427]
Número de tentativas de configuração de SRB 2	Numérico	[1 a 57975]
Número de falhas de configuração de SRB 2	Numérico	[0 a 1119]
Número de conclusões de configuração de SRB 2	Numérico	[1 a 57967]

- **Grupo 3**: dizem respeito aos indicadores de controle de recursos de rádio (*Radio Resource Control* - RRC) da rede 3G. O RRC é responsável pelo estabelecimento das conexões e funções de liberação (*release*) de sinalização na interface aérea. Como existem vários tipos de fluxos de sinalização para estabelecimento da conexão, os contadores são associados a cada etapa dos processos para facilitar a análise de desempenho e identificação de desvios na rede. O conjunto de dados 3 com seus KPIs selecionados está representado na Tabela 3.

Tabela 3 - Atributos do conjunto de dados 3 (de redes móveis 3G)

Atributo	Formato	Faixa
Dia	Data	[01/01/2019 a 07/01/2019]
Hora	Hora	[00:00 a 23:00]
Célula ID	Categórico	[1 a 3339]
Liberação de acesso RRC de reestabelecimento de chamada	Numérico	[0 a 25]
Liberação de acesso RRC de CS Fallback	Numérico	[0 a 96]
Liberação de acesso RRC de detach	Numérico	[0 a 22]
Liberação de acesso RRC de chamada de emergência	Numérico	[0 a 3]
Liberação de acesso RRC de nova seleção na mesma tecnologia	Numérico	[0 a 464]
Liberação de acesso RRC de chamada em segundo plano originada	Numérico	[0 a 3423]
Liberação de acesso RRC de conversação de chamada originada	Numérico	[0 a 120]
Liberação de acesso RRC de sinalização de alta prioridade originada	Numérico	[0 a 59]
Liberação de acesso RRC de chamada interativa originada	Numérico	[0 a 210]
Liberação de acesso RRC de sinalização de baixa prioridade originada	Numérico	[0 a 30]
Liberação de acesso RRC de chamada de tráfego registrado originada	Numérico	[0 a 288]
Liberação de acesso RRC de chamada em segundo plano terminada	Numérico	[0 a 358]
Liberação de acesso RRC de conversação de chamada terminada	Numérico	[0 a 98]
Liberação de acesso RRC de sinalização de alta prioridade terminada	Numérico	[0 a 3]
Liberação de acesso RRC de sinalização de baixa prioridade terminada	Numérico	[0 a 102]
Liberação de acesso RRC de registro	Numérico	[0 a 175]

Este conjunto de dados foi submetido ao método proposto, que foi implementado no ambiente de programação RStudio, que opera na linguagem R. As funções e pacotes desta linguagem foram usadas para implementação dos métodos desenvolvidos neste trabalho. Todos os testes foram executados em computadores comuns (1 processador desktop com memória RAM de 8GB de RAM). O custo computacional foi variável e não é objeto de consideração neste trabalho.

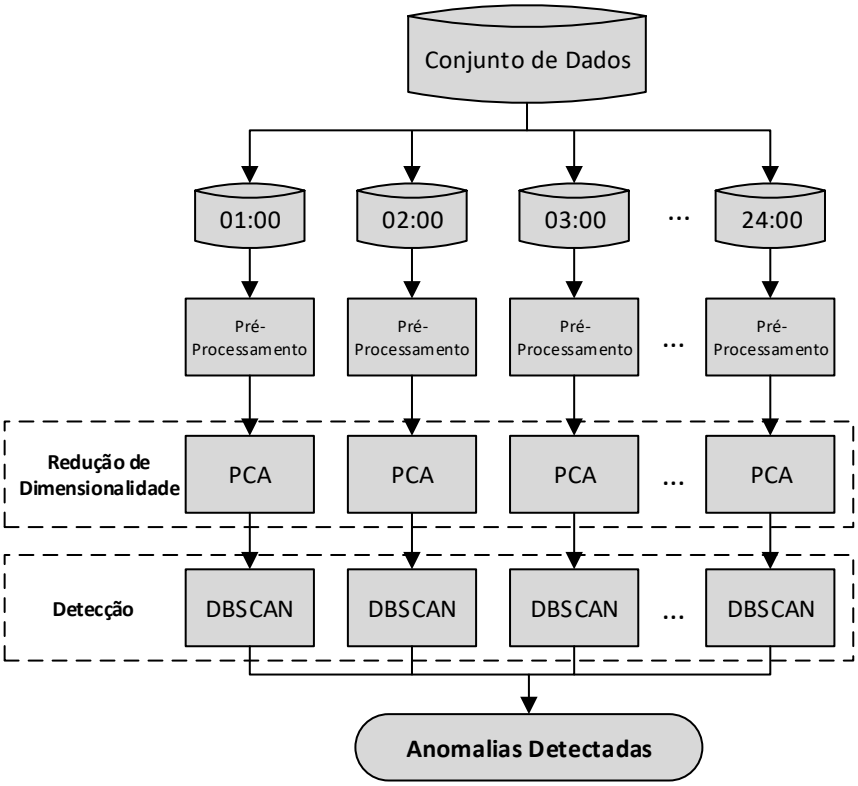
4.2 Descrição do método proposto

O processo de identificação de padrões (neste caso aqui, de anomalias) envolve algumas etapas que vão desde a seleção dos atributos, que no caso são os KPIs, até a detecção em si. Neste trabalho, o método proposto para detecção de anomalias é baseado, a grosso modo, na redução de dimensionalidade e no agrupamento do conjunto de dados formado na nova dimensão construída. A Figura 11 apresenta esquematicamente a estrutura do método de detecção de anomalias utilizado. Observe que nele consta a estratificação dos dados por hora, que foi feita para tentar identificar comportamentos temporais, uma vez que a dinâmica da célula está sujeita a diferentes condições que variam no dia.

O modelo proposto não exige dados de treinamento rotulados (*i.e.*, é não supervisionado). A detecção das anomalias é realizada com base na vizinhança dos dados. Como mostrado na estrutura do método, o conjunto de dados é analisado (ou fragmentado) por hora[2]. Este grupo de dados passa por um pré-processamento para remover registros incompletos. Depois de feito isto, os KPIs selecionados (por serem potencialmente muitos) são aplicados à técnica de PCA para tentar reduzir a dimensão de dados.

[2] Ela foi feita a cada período de uma hora pois as amostras de KPI coletadas também eram tomadas a cada período de uma hora. Contudo, dependendo do conjunto de dados, esta estratificação pode assumir outros valores.

Figura 11 – Estrutura do método de detecção de anomalias proposto



Fonte: autoria própria

Por se tratar potencialmente de muitos KPIs, seria interessante simplificar a análise desta quantidade de dados sem perder, preferencialmente, muita informação. Neste sentido, a técnica PCA transforma todos estes KPIs em componentes (ou eixos de análise), que mantêm a representação dos dados em um espaço "geométrico" limitado de dimensão menor do que aquele visto na base de dados originais.

Depois de reduzida a dimensão de análise espacial dos dados de KPI, o conjunto de dados é agrupado no espaço geométrico composto pelas componentes principais estimadas no passo anterior. Este agrupamento é feito através de uma técnica de *clustering*, conhecida como DBSCAN, e tratada no Capítulo 2 deste trabalho. Com base em sua lógica, descrita anteriormente, este algoritmo tenta fazer a identificação espacial dos dados e, com isto, promover a separação e rotulação de potenciais anomalias. A saída

do método proposto é a classificação das instâncias de dados como anômala ou não anômala.

Por se tratar de potenciais anomalias, a última etapa do método é a avaliação dos dados por um profissional. Isto deve acontecer, pois o método não tem por objetivo identificar todas as anomalias, ou classificá-las nominalmente, mas, sim, prover suporte para identificação de potenciais padrões suspeitos, para que ajudem o operador da rede a buscar sua otimização com base nesta ferramenta proposta e em caráter complementar a outras existentes. Com isto, espera-se encontrar a maior quantidade possível de padrões no grande conjunto de dados ou KPIs que é produzido pelo OSS, ajudando o operador da rede a edificar uma assertividade melhor nas suas avaliações.

Na implementação do método no ambiente de programação RStudio foram utilizadas algumas funções nativas [44] e pacotes desenvolvidos pela comunidade, que podem ser resumidos em cada uma das sub-etapas da seguinte forma:

- Leitura do arquivo: função nativa *read.csv()*;
- Fragmentação temporal: funções nativas *sapply()* e *lapply()*;
- Pré-processamento: funções nativas *subset()* e *scale()*;
- Redução de dimensionalidade (PCA): função nativa *prcomp()*;
- Detecção das anomalias (DBSCAN): função *dbscan()* do pacote "dbscan" [45];
- Visualização: função *ggplot()* do pacote "ggplot2" [46].

As sub-etapas do método são descritas em mais detalhes em cada uma das subseções que seguem.

4.2.1 Fragmentação dos dados e pré-processamento

O consumo de dados e a utilização da rede móvel apresentam um comportamento dinâmico ao longo do dia. Em geral, o pico do tráfego de voz ocorreu em torno das 11:00 e o pico do tráfego de dados em torno das 20:00. Neste caso, analisar os horários separadamente permitiu a identificação de comportamentos que não são comuns no

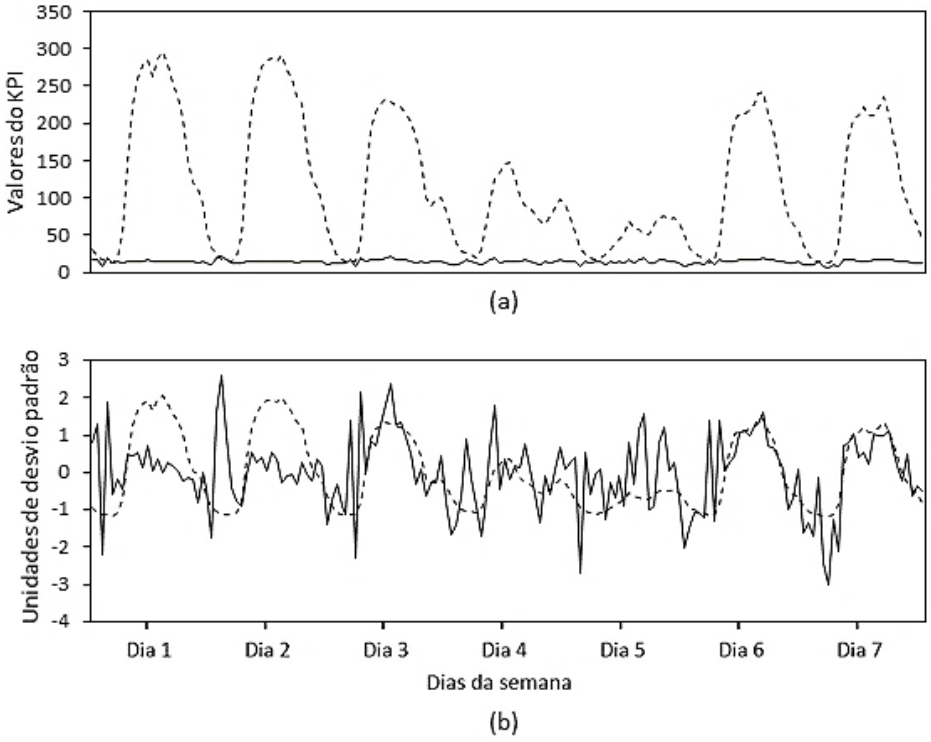
período. Dessa forma, após a seleção dos KPIs (ou atributos) foi realizada a fragmentação dos dados em horários.

Antes de iniciar qualquer análise foi importante que fosse realizado um pré-processamento dos dados com o intuito de se obter resultados significativos e úteis. Nesta aplicação isto é importante, porque as bases de dados puderam apresentar valores inconsistentes, nulos e *outliers*. No caso dos valores inconsistentes e nulos, as instâncias de dados que apresentaram pelo menos um valor destes casos foram removidas. No caso dos *outliers*, os valores foram mantidos por fazerem parte do escopo de estudo deste trabalho.

Um ponto relevante que foi verificado foi a variação dos dados. Os KPIs que possuíam alta variância dominam as técnicas baseadas em distância e variância, como as que foram utilizadas no método proposto. Assim, foi necessário fazer um dimensionamento adequado para tornar os KPIs iguais em sua importância. Uma forma de se fazer isso foi normalizando os dados com base na média aritmética como referência (*i.e.*, valor zero do eixo) e o desvio padrão como medida de variação do eixo/dimensão em questão. Este tipo de normalização manteve a forma da curva dos KPIs sem perder informações importantes.

A Figura 12a mostra a variação temporal de dois KPIs. Nota-se que o KPI 1 apresenta uma variação de amplitude maior que o KPI 2. Após a normalização (vista na Figura 12b), a variação dos KPIs se aproxima e é possível observar com mais detalhes o comportamento do KPI 2 sem distorções de magnitude inerentes à própria grandeza do KPI em questão. No caso desta figura, o valor 0 (zero) do eixo vertical indica os valores que estão na média aritmética. Cada B unidades que se move neste eixo indica uma variação de B vezes o desvio padrão deste KPI em questão.

Figura 12 - Representação gráfica de dois KPIs de variações e magnitudes diferentes: (a) valores originais e (b) normalizados



Fonte: autoria própria

4.2.2 Redução de dimensionalidade

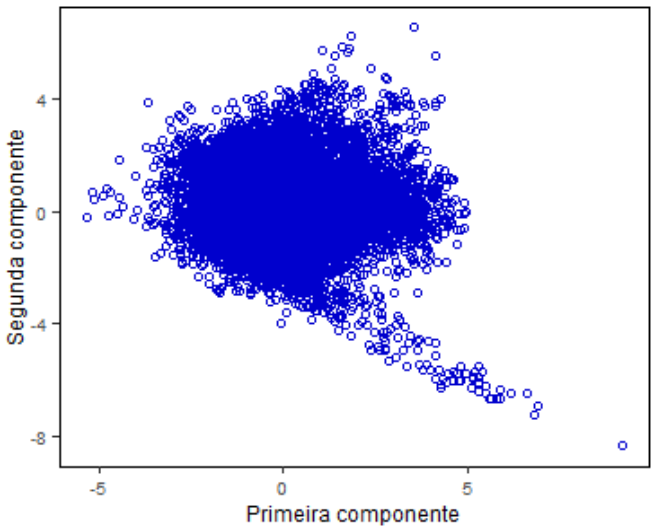
O método proposto foi desenvolvido de forma a ser aplicado em forma de conjuntos de dados que contenham mais de um KPI. Neste caso, a seleção de um grande número de KPIs pôde gerar uma grande dimensão de análise, especialmente quando esta quantidade ultrapassava três[3] KPIs. Para evitar isto, foi importante ter uma redução de dimensionalidade a fim de potencialmente evidenciar de forma mais clara informações "escondidas" nos dados brutos. Adicionalmente, a transformação do conjunto de dados obtidos com as técnicas evidencia também a relação e as características entre os atributos.

[3] Em outros termos, com mais de três KPIs não é possível plotar graficamente o arranjo espacial destes KPIs. Como a análise visual é importante para avaliação, fez-se necessário produzir gráficos, ao mesmo tempo que se deve permitir o uso de mais do que três KPIs. Neste sentido, a redução de dimensionalidade proporcionou ambas condições: trabalhar com muitos KPIs simultaneamente e, ainda assim, produzir gráficos para análise visual. O uso de projeções de espaço superiores ao tridimensional foi descartado nesta pesquisa pelo desafio de apresentar os resultados de forma ilustrativa e dificultar a análise do usuário operador de rede.

Neste caso, esta transformação permitiu não só uma redução no número de dimensões, mas uma análise multivariada dos KPIs.

O método aqui proposto usou a técnica de PCA (tratada na Seção 2.3.1.1) para transformar um conjunto de dados composto por KPIs em um conjunto ortogonal de componentes. Por ser puramente estatística, baixo custo computacional e largamente utilizada em modelos não supervisionados, optou-se pelo o seu uso na etapa de redução de dimensionalidade. Em geral, o número de componentes consideradas devem representar juntos no mínimo 80% da variância total das amostras. A Figura 13 mostra um gráfico de dispersão das duas primeiras componentes PCA geradas de um conjunto de dados com seis atributos.

Figura 13 - Gráfico de dispersão de duas primeiras componentes geradas pela técnica PCA



Fonte: autoria própria

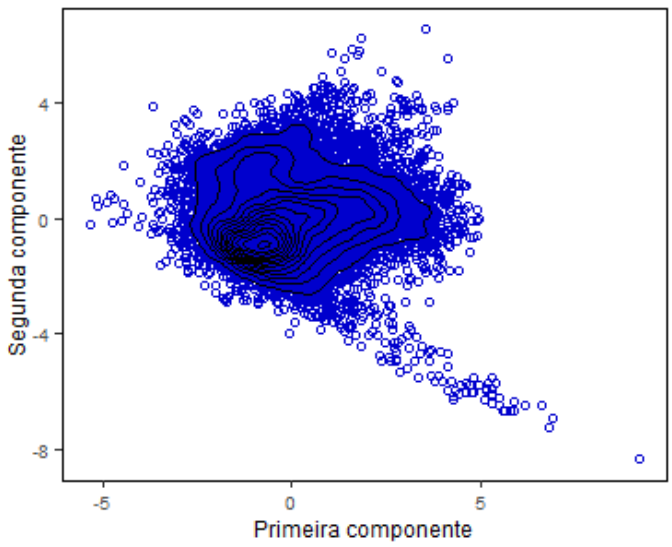
Observa-se na figura que grande parte das amostras se concentram numa região centralizada com algumas amostras dispersas ao redor. Este comportamento despertou o interesse de realização de estudos mais detalhados para estas regiões. A princípio, o comportamento tido como normal da rede é representado pela região densa e as amostras dispersas representam um comportamento anormal. A diferenciação destas regiões foi feita com base na proximidade das amostras e está descrita a seguir.

4.2.3 Clustering

O *clustering* ou agrupamento é uma técnica geralmente não supervisionada empregada para agrupar dados com base em uma medida de dissimilaridade. A distância entre os dados, por exemplo, pode ser uma das métricas adotadas para se medir esta dissimilaridade e formar grupos comumente chamados de *clusters*. Os dados que fazem parte do mesmo *cluster* tendem a possuir alguma característica ou propriedade em comum, e no cenário de detecção de anomalias podem representar o padrão normal ou anormal dos dados.

No agrupamento baseado em densidade, um *cluster* é formado em qualquer direção conduzida pela densidade dos pontos (leia-se dados). Os algoritmos baseados em densidades são capazes de identificar *clusters* de formas arbitrárias e procuram por regiões de alta densidade separados por regiões de baixa. A Figura 14 mostra as curvas de distribuição da densidade dos pontos. A região de alta densidade se localiza próximo ao centro e as regiões de baixa densidade estão no entorno, representadas pelos pontos dispersos.

Figura 14 - Representação da distribuição da densidade dos pontos de componentes geradas pela técnica PCA. As curvas em preto representam as linhas de densidade dos pontos



Fonte: autoria própria

Um dos primeiros algoritmos implementados neste contexto foi o DBSCAN (ver Seção 2.3.2.1), devido à sua capacidade de lidar com *clusters* de várias formas e tamanhos e robustez na detecção de ruídos, adotado como técnica de agrupamento baseado em densidade. Neste caso, a capacidade de classificação de pontos ruídos é essencial para a identificação das anomalias. Basicamente, o algoritmo depende da definição do raio da vizinhança (*Eps*) e do número mínimo de pontos da vizinhança (*MinPts*) para se formar um *cluster*.

No método proposto, adotou-se a estratégia de definir os parâmetros *Eps* e *MinPts* com o objetivo de se obter um *cluster* único, entendendo que ele representará o comportamento predominante da célula. Ao redor deste, os pontos ou 'ruídos' foram os candidatos a potenciais anomalias. Para tanto, o valor de *MinPts* deveria ser alto e indexado ao desvio padrão (σ) dos dados, uma vez que o comportamento dos KPIs varia durante o dia. Dessa forma, considerou-se $MinPts = 10.\sigma$ e o valor de *Eps* é o ponto de "vale" das distâncias dos k-vizinhos mais próximos, conforme Figura 8.

4.3 Experimentos para o desenvolvimento do método proposto

Na sequência são descritos uma série de experimentos realizados durante o trabalho, que justificaram a escolha das técnicas usadas ou nortearam o seu desenvolvimento. Destacando que os resultados apresentados e suas consequentes discussões estão apresentadas em um capítulo à parte (Capítulo 5).

4.3.1 Redução de dimensionalidade

As técnicas de redução de dimensionalidade têm por objetivo transformar um conjunto inicial de atributos em um novo conjunto, com um número menor de atributos e, ainda sim, preservando a maior parte possível da informação presente nos dados. Na

técnica PCA, a informação a ser preservada é a variância do conjunto de atributos, que pode ser representada pelas componentes geradas. Além disso, como as componentes resultam de uma transformação do conjunto de dados, a relação e as características entre as variáveis são preservadas e permitem uma análise multivariada.

A representação da variância de cada componente PCA se altera conforme os atributos. Como os conjuntos de dados 1, 2 e 3 possuem características distintas em termos de quantidade de atributos, tipos de KPI, data de coleta e tecnologia da rede, as componentes PCA foram geradas para cada conjunto de dados, de forma a ilustrar a diferença na representatividade das componentes ou o quanto da variância dos dados é explicada em cada componente.

4.3.2 Variações de agrupamento

Nas redes móveis, apesar da variação dos KPIs ser influenciada pelo comportamento dinâmico da rede, a correlação dos KPIs pode variar bastante e apresentar diferentes distribuições. Isto ocorre principalmente em KPIs que representam medições de tabelas diferentes e que são influenciados por outros fatores, como o número de falhas de requisição e a taxa de dados média. Além disso, a etapa de redução de dimensionalidade produz componentes ortogonais que apresentam distribuições distintas de acordo com o conjunto de dados selecionados.

Neste sentido, a definição de qual técnica de agrupamento seria utilizado no projeto foi baseada na distribuição espacial e no significado prático dos *clusters*, visando a análise de desempenho das redes móveis. Para tanto, foram selecionados alguns KPIs com diferentes correlações e as principais técnicas dos métodos de agrupamento mais utilizados, que foram os seguintes:

- Particionamento: K-*means*;
- Hierárquico: *Single Linkage* e *Complete Linkage*;
- *Fuzzy*: *Fuzzy* C-*Means*;

- Baseado em densidade: DBSCAN;
- Baseado em modelo: *Model-based*.

A utilização destas técnicas requereu, inicialmente, a definição de alguns parâmetros: (a) nas técnicas em que se usa uma medida de distância, foi considerada a distância euclidiana; (b) para as técnicas em que é necessário informar o número de *clusters*, foram usados os métodos do "cotovelo" e da silhueta para definir este número; (c) para a técnica baseada em densidade, foi usado o método do "vale" para definir o raio da vizinhança (*Eps*) e o número mínimo de pontos da vizinhança (*MinPts*) foi considerado igual a 10; (d) para os demais parâmetros, foi considerado o valor *default* comumente adotados nas funções utilizadas da linguagem R.

4.3.3 Análise do conjunto de dados

Os conjuntos de dados selecionados contêm KPIs que representam medições de diversas etapas, desde o estabelecimento da conexão à utilização da rede. Cada conjunto de dados possui uma particularidade, uma vez que os KPIs representam diferentes medições da rede. Por este motivo, foi esperado que a distribuição espacial dos dados fosse diferente para cada conjunto de dados. Além disso, a causa do comportamento anormal poderia estar associada aos KPIs, pois, por exemplo, o comportamento anormal de sobrecarga da rede poderia estar associado aos KPIs de número de usuários ativos e taxa de dados média.

A seleção do conjunto de dados para estudo foi baseada nos KPIs selecionados e na variedade de possíveis causas na degradação dos KPIs. Assim, o conjunto de dados 1 foi selecionado por refletir comportamentos anormais relacionados à qualidade de sinal e utilização da rede. Contudo estas características não são limitantes para o uso dos outros conjuntos de dados. O método proposto foi desenvolvido para ser aplicado em conjuntos de dados com mais vários KPIs, não tendo um limite máximo como limitação.

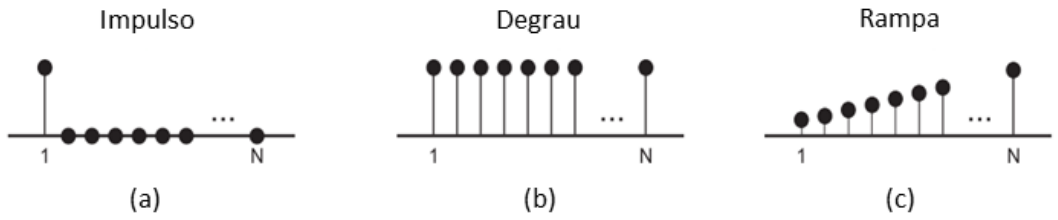
4.4 Validação e análises

Aqui são apresentadas as análises realizadas para fazer a comparação com outro método (também não supervisionado) fundamentado em redes neurais de detecção de anomalias. Esta comparação visou tentar indicar a qualificação do método como complementar e não, necessariamente, como superior, uma vez que a análise deste tipo de problema não é totalmente determinística.

4.4.1 Análise da assertividade do método proposto

A assertividade de modelos ou métodos não-supervisionados impõem certa dificuldade, visto que os dados não são rotulados. No método proposto, uma alternativa para estimar sua assertividade foi a de inserir **dados sintéticos** que seguem um dado padrão de degradação comum nas redes móveis. Estes padrões "simulam" as principais anomalias da rede e estão ilustrados na Figura 15. Eles são os mesmos encontrados em outros trabalhos [7] e [38].

Figura 15 – Padrões de degradação usados na inserção das anomalias sintéticas



Fonte: autoria própria

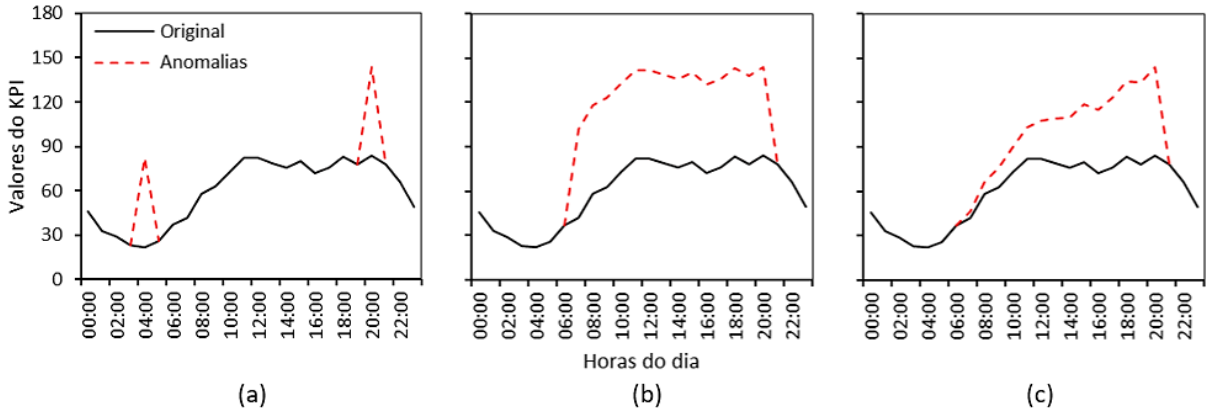
O comportamento que corresponde a um evento curto, como um pico de tráfego esporádico ou um bloqueio/desbloqueio de serviço de uma célula, é ilustrado na Figura 15a. A Figura 15b representa uma degradação constante, como uma alteração de um parâmetro ou falha no *backhaul* (enlace entre ERB e controlador). A Figura 15c reproduz uma degradação progressiva, como uma variação da sobrecarga ao longo do dia ou uma interferência externa.

Estes padrões foram adicionados aos valores originais dos KPIs de número médio de usuários ativos e RSSI médio, e subtraídos dos KPIs de taxas de dados médios no *downlink* e SINR médio de determinadas células com amplitude máxima de 3.σ, onde σ é o desvio padrão dos KPIs. Este valor foi definido de forma que as instâncias que estavam distantes em mais de 3.σ da média em uma distribuição normal são comumente consideradas como *outlier* na técnica de detecção de *outlier* mais simples [2]. Outros trabalhos [7] e [38], usaram valor similar. Os cenários de inserção das anomalias sintéticas foram os seguintes:

1. Impulso: a anomalia foi inserida no horário de menor movimento da rede (04:00) e no horário de pico (20:00), com amplitude de 3.σ;
2. Degrau: a anomalia foi inserida entre os horários 07:00 e 20:00. Durante este período a utilização da rede móvel é maior e os cenários de degradação costumam ocorrer com mais frequência. A amplitude das anomalias inseridas foi de 3.σ em todos os horários;
3. Rampa: assim como no cenário anterior, a anomalia foi inserida entre os horários 07:00 e 20:00. A amplitude das anomalias foi incrementada linearmente às 07:00 em parcelas de 1/14 de 3.σ até alcançar 3.σ às 20:00.

A Figura 16 mostra um exemplo da inserção das anomalias no KPI de número médio de usuários ativos em determinada célula do conjunto de dados. A inserção das anomalias foi realizada conforme os cenários descritos anteriormente e estão representadas pela linha tracejada.

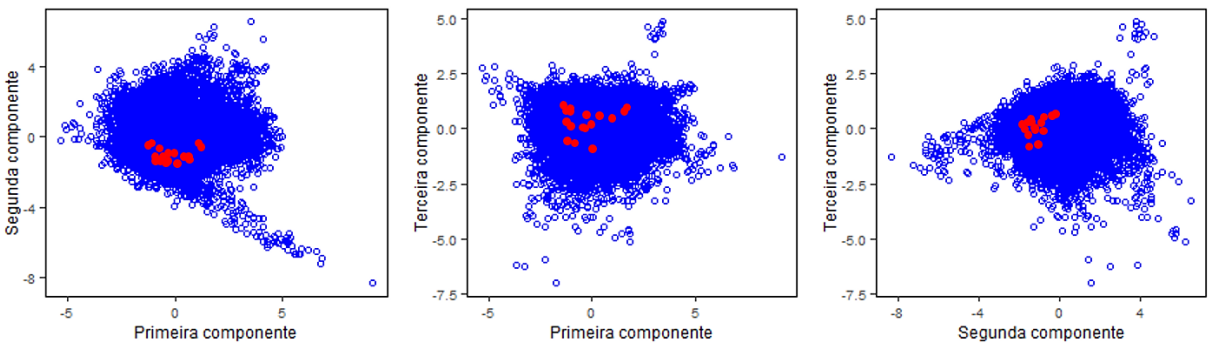
Figura 16 - Exemplo das anomalias sintéticas inseridas conforme os padrões de degradação impulso (a), degrau (b) e rampa (c)



Fonte: autoria própria

As células (elementos da rede móvel) consideradas para a inserção das anomalias sintéticas foram selecionadas com base na quantidade de instâncias categorizadas como anomalia (zero, no caso) pelo método proposto e na pontuação média de anormalidade das instâncias (menor possível, no caso) dada pelo método de referência, ambos apresentados na seção a seguir. Neste caso, as instâncias de dados das células selecionadas se localizam na região central, onde é formado o *cluster* principal, como ilustrado pela Figura 17. Além disso, os valores dos KPIs das células selecionadas estavam dentro dos valores considerados como normais operacionalmente, segundo os profissionais que avaliaram as células que apresentaram instâncias anômalas.

Figura 17 - Representação da gráfica da localização das instâncias de dados das 20:00h das células selecionadas para a inserção das anomalias sintéticas



Fonte: autoria própria

Dessa forma, foi adicionado um comportamento anormal em células que, originalmente, apresentavam um comportamento categorizado como normal pelo método proposto. Para tanto, foram selecionados dez células e dois dias (06/03 e 13/03) do conjunto de dados 1 para a inserção das anomalias. Com base nos resultados, foi possível avaliar o desempenho do método nos diferentes cenários em que a anomalias foram inseridas.

4.4.2 Análise individual das anomalias

O resultado do método proposto é a classificação das instâncias do conjunto de dados como anômala ou não, como ilustra a Figura 18. Neste caso, a Célula ID 2 possui duas instâncias anômalas que estão em diferentes horários. A partir dessas informações é possível calcular a quantidade ou o somatório de anomalias que uma célula teve ao longo das horas do dia e ranqueá-las.

Figura 18 – Exemplo de classificação das instâncias de um conjunto de dados pelo método proposto. Neste caso, só a Célula ID 2 possui instâncias anômalas, que estão em diferentes horários

	A	B	C	D	E	F	G
1	Dia	Hora	Célula ID	KPI 1	KPI 2	KPI 3	Anomalia
2	03.01.2019	00:00:00	1	3	9130.66	14.72	0
3	03.01.2019	00:00:00	2	2	5569.75	14.55	1
4	03.01.2019	00:00:00	3	6	9055.49	14.75	0
5	03.01.2019	01:00:00	1	2	9360.88	13.72	0
6	03.01.2019	01:00:00	2	1	4418.65	15.09	1
7	03.01.2019	01:00:00	3	4	7687.78	14.85	0

Fonte: autoria própria

Com o intuito de observar se as células classificadas com anomalias possuíam um comportamento considerado anormal, foram selecionadas as 20 (vinte) células com a maior quantidade para avaliação de 2 (dois) profissionais que trabalham com otimização RF e análise de desempenho da rede móvel na operadora que disponibilizou os dados. A partir destes resultados foi possível confirmar, na visão dos especialistas, se as células selecionadas apresentavam um comportamento que não segue o comportamento normal

da rede. Para tanto, foram adotadas algumas premissas que nortearam e delimitaram a avaliação individual das células, quais foram:

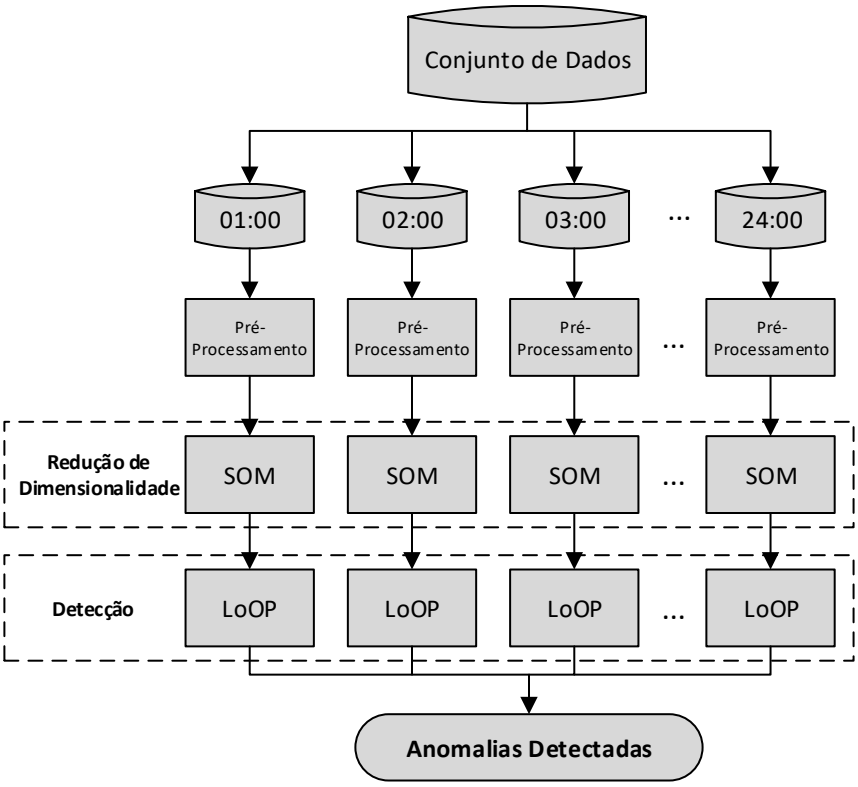
- A causa primária da anormalidade da célula seria considerada como a avaliação final, caso a célula apresentasse mais de uma causa;
- As ferramentas e dados disponíveis na operadora para pesquisa de falha poderiam ser usadas como analisador de espectro, *drive-test*, histórico de alarmes e de outros KPIs, dentre outros;
- As informações providas pelos autores foram o nome das células a serem avaliadas e os KPIs selecionados para aplicação do método.

4.4.3 Método de referência e análise comparativa

As validações e análises descritas até o momento foram realizadas exclusivamente com o método proposto para a detecção de anomalias. No contexto de detecção de anomalias baseado em modelos ou métodos não-supervisionados existe uma técnica muito usada para redução de dimensionalidade e evidenciação de padrões "escondidos", que é a *Self-Organizing Map* (SOM) (ver Seção 2.3.1.2). Assim, com o intuito de melhor caracterizar o método aqui proposto, seu desempenho na detecção de anomalias foi comparado com um método de referência baseado nesta técnica. A estrutura do método está ilustrada na Figura 19. Esta comparação foi feita com o intuído de verificar e validar o desempenho do método proposto com base em um método empregado e bem conhecido na área. Em virtude disto, este método é, algumas vezes, referenciado como "método padrão" ou "método de referência".

A estrutura de fluxo de processamento do método de referência foi adaptada para poder se aplicar ao conjunto de dados usados. Esta adaptação foi feita para preservar a estratificação por hora, além de facilitar a comparação entre o resultado dos dois métodos (o que proposto e o método padrão de referência).

Figura 19 – Estrutura do método de referência de detecção de anomalias baseado em redes neurais



Fonte: autoria própria

Os métodos são similares nas etapas iniciais de segmentação temporal do conjunto de dados e pré-processamento. Nas etapas seguintes, foi usado o SOM para a redução de dimensionalidade e a técnica *Local Outlier Probabilities* (LoOP) para a detecção das anomalias. A combinação entre SOM e LoOP foi utilizada para realizar o comparativo, pois apresentou bons resultados na identificação de padrões de degradação no desempenho da rede em comparação com outras técnicas não-supervisionadas, segundo os autores de [35]. Por isto, ele foi escolhido como método de referência para comparações.

Os principais parâmetros para treinamento do SOM foram definidos da seguinte forma:

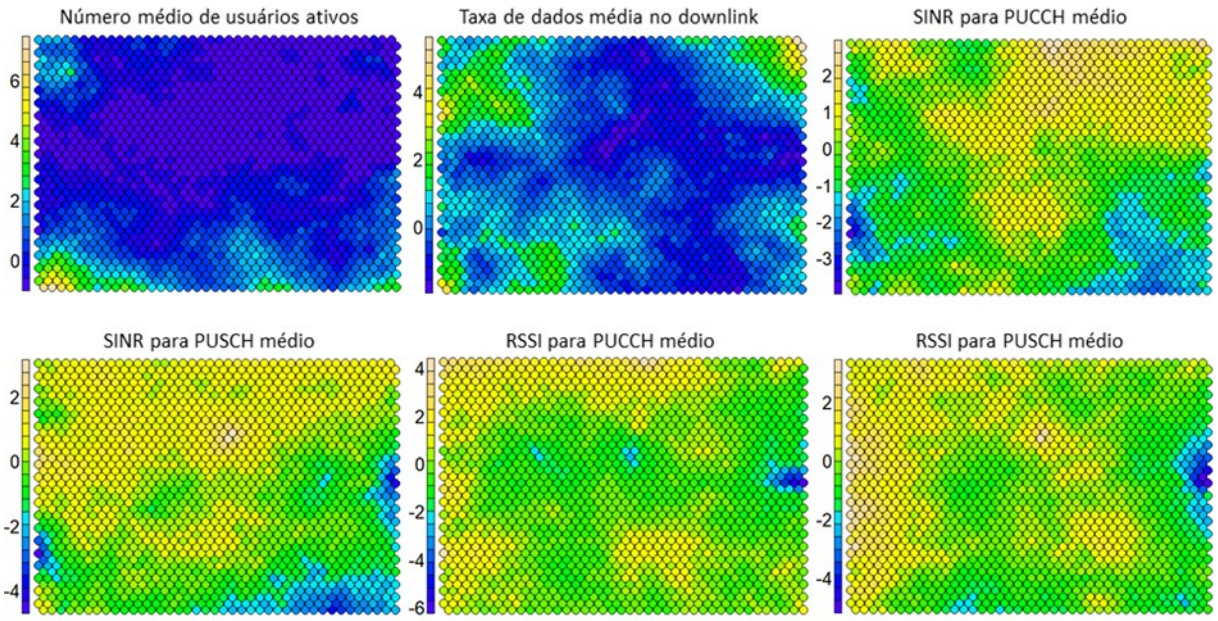
- Tamanho da rede: o número de neurônios ou 'nós' foi definido de forma que o número de amostras em cada nó estivesse entre 5 e 10. Assim, evitou-se vazios ou muitas amostras. O tamanho da rede foi definido em 40x40 nós;

- Iterações ou épocas: a distância dos pesos de cada nó às amostras por esse nó foi reduzida à medida que as iterações de treinamento progridem. O número de iterações ideal ocorreu quando se atingiu um valor mínimo seguido de uma estabilização. Dessa forma, o número de iterações foi definido como 250;
- Taxa de aprendizado: representou o tamanho do vetor de pesos na etapa de aprendizado. Neste caso, foi considerado o valor de 0,05 a 0,01;
- Raio da vizinhança: representou o tamanho do raio em que os BMUs foram atribuídos. A cada iteração, o raio da vizinhança diminuiu até atingir o valor mínimo. Neste caso foi considerado o valor de 10 a 0,1.

A etapa de redução de dimensionalidade com o SOM tem como saída os vetores de pesos (*codebook*) para cada nó da rede e algumas informações importantes, como a distância entre nós (matriz U) e o número de amostras atribuídas em cada nó. Neste caso, por conter o valor de peso de cada atributo do conjunto de dados atribuído em cada nó da rede SOM, a detecção das anomalias foi realizada com base no *codebook*.

Os vetores de pesos apresentaram a mesma dimensão da quantidade de KPIs do conjunto de dados 1, ou seja: 6. Como a rede SOM possui 40x40 nós, o número total de vetores de peso é 1600. A Figura 20 exemplifica o *codebook* representado em mapas de calor para cada atributo do conjunto de dados 1. O resultado da etapa de redução de dimensionalidade é um *codebook* para cada período do conjunto de dados.

Figura 20 – Exemplos de mapas de calor do codebook da rede SOM para os atributos do conjunto de dados 1 no horário das 20:00h



Fonte: autoria própria

A detecção das anomalias foi realizada pela técnica LoOP, que, com base na densidade local, comparada à densidade dos k-vizinhos mais próximos e calcula a probabilidade de que aquela amostra seja um *outlier* ou uma anomalia. Os valores variam de 0 a 1, sendo 1 o maior *outlier*. Com relação aos parâmetros a serem definidos, o número de k-vizinhos mais próximos (k) para se comparar à densidade foi considerado igual a 10.σ, igual ao valor do parâmetro *MinPts* do DBSCAN no método proposto, e o fator de multiplicação para o desvio padrão (λ) igual a 3, que é valor *default* considerado pelos autores da técnica apresenta em [29].

A Figura 21 mostra um exemplo dos resultados obtidos com o método de referência. A rede SOM foi criada em cada hora do conjunto de dados, dessa forma, tem-se os valores do *codebook* dos atributos (três KPIs, no exemplo) e a pontuação de anormalidade de cada nó da rede SOM, que estão representados na Figura 21a. Além disso, foi possível obter em qual nó as instâncias do conjuntos de dados eram atribuídas (Figura 21b) a partir de um identificador da posição da instância. Com base nos campos

"Hora" e "Nó SOM" foi possível encontrar a pontuação de anormalidade de cada instância.

Figura 21 - Exemplo do resultado gerado pelo método de referência; em (a) são apresentados o nó SOM, o *codebook* (CB) com os valores para três KPIs e a pontuação de anormalidade; e em (b) as instâncias de dados de um conjunto de dados com três componentes principais (PC) geradas a partir de três KPIs normalizados

	A	B	C	D	E	F
1	Hora	Nó SOM	CB KPI 1	CB KPI 2	CB KPI 3	Pontuação
2	00:00:00	1	-0.91	-1.53	-2.14	0.99
3	00:00:00	2	-0.69	-1.41	-2.12	0.61
4	00:00:00	3	-0.87	-1.19	-3.22	0.78
5	00:00:00	4	0.64	-1.22	-2.49	0.53
6	00:00:00	5	1.56	-1.22	-2.5	0.32
7	00:00:00	6	0.77	-1.03	-2.35	0.00

(a)

	A	B	C	D	E	F	G
1	Dia	Hora	Nó SOM	Célula ID	KPI 1	KPI 2	KPI 3
2	03.01.2019	00:00:00	2	1	1.62	-1.43	-0.4
3	03.01.2019	00:00:00	5	2	-0.85	-1.38	-0.15
4	03.01.2019	00:00:00	3	3	-0.46	-0.96	-0.19
5	03.02.2019	00:00:00	4	1	-1.86	-1.31	-0.58
6	03.02.2019	00:00:00	4	2	-0.86	-1.38	-0.22
7	03.02.2019	00:00:00	2	3	-0.94	-0.86	0.25

(b)

Fonte: autoria própria

Os métodos proposto e de referência foram avaliados utilizando o conjunto de dados 1 na sua forma original e com as anomalias sintéticas, conforme os padrões de degradação ilustrados na Figura 15. As células (elementos da rede móvel) consideradas para a inserção das anomalias sintéticas foram selecionadas com base na quantidade de instâncias categorizadas como anomalia (zero, no caso) pelo método proposto e na pontuação média de anormalidade das instâncias (menor possível, no caso) dada pelo método de referência. Neste caso, foram selecionadas as mesmas dez células e os dias da semana da seção anterior do conjunto de dados 1 para a inserção das anomalias, representadas na Figura 16. Dessa forma, foi possível avaliar o desempenho dos métodos nos cenários em que a anomalias foram inseridas.

Como proposto, o método foi implementado no ambiente de programação RStudio. As funções nativas e os pacotes desenvolvidos pela comunidade usados podem ser resumidos da seguinte forma:

- Leitura do arquivo: função nativa *read.csv()*;
- Fragmentação temporal: funções nativas *sapply()* e *lapply()*;
- Pré-processamento: funções nativas *subset()* e *scale()*;
- Redução de dimensionalidade (SOM): função *som()* do pacote "kohonen" [47];
- Detecção das anomalias (LoOP): função *LOOP()* do pacote "DDoutlier" [48];
- Visualização: função nativa *plot()*.

4.4.4 Análise individual das anomalias incluindo o método de referência

As vinte células que apresentaram a maior quantidade de amostras classificadas como anomalias nos dois métodos foram selecionadas para a avaliação individual de dois profissionais habilitados, conforme as premissas citadas na Seção 4.4.2. Com base nos resultados, esperou-se que algumas células estivessem presentes nas vinte primeiras células dos dois métodos, o que confirmaria o comportamento anormal. Por outro lado, esperou-se que algumas células estivessem presentes nas vinte primeiras células em somente um dos métodos. Neste caso, a avaliação dos profissionais foi importante para a validação dos resultados, uma vez que o método proposto não traça diagnóstico, mas, sim, somente anomalia. Neste sentido, a avaliação dos profissionais produziram, além do diagnóstico, a indicação do que realmente se trata de anomalia ou não dentro dos resultados encontrados pelo método proposto. Isto faz sentido, porque a proposta do método foi justamente varrer toda uma imensidão de dados e selecionar apenas os casos de comportamentos estranhos para o avaliador humano. É deste a avaliação final, de qualquer maneira.

4.5 Resumo do capítulo

Neste capítulo foram apresentados detalhes de como o trabalho foi desenvolvido. Os conjuntos de dados selecionados foram formados por indicadores de desempenho das redes móveis e se caracterizaram por estar na forma de série temporal. Estas características foram exploradas no método proposto, que estratificou os dados numa etapa inicial dado o comportamento dinâmico das redes móveis. Na seção de experimentos e testes foram apresentadas as análises realizadas que justificam a escolha das técnicas e as considerações adotadas. As validações e análises da assertividade do método proposto e de um método complementar foram apresentadas ao final.

5 Resultados e discussões

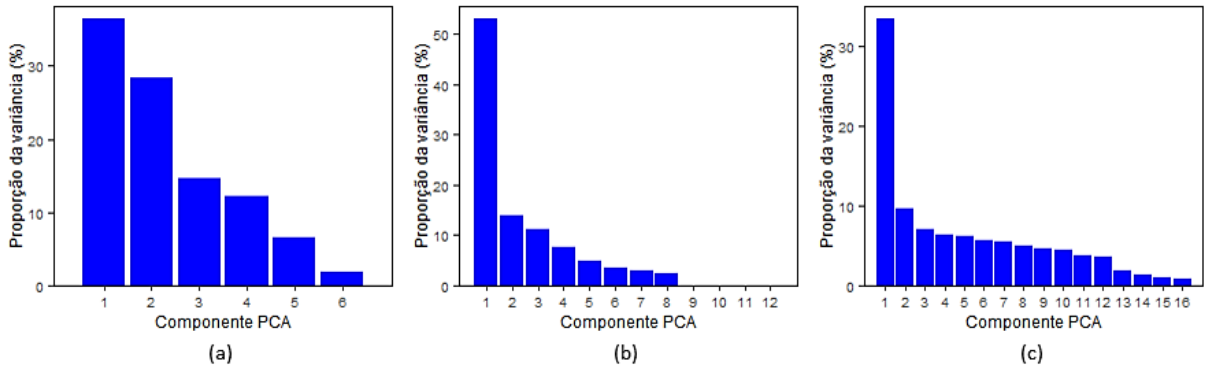
Este capítulo apresenta os resultados do método proposto para detecção das anomalias. Inicialmente são apresentados os resultados dos experimentos realizados para o desenvolvimento do método proposto e, em seguida, as validações e análises. Por fim, as discussões gerais são apresentadas com base nos resultados e no uso do método como ferramenta para nortear a correção de falhas, otimização de recursos e planejamento da rede de acesso móvel.

5.1 Resultados do desenvolvimento do método proposto

5.1.1 Redução de dimensionalidade

Os gráficos da Figura 22 demonstram como se comportam as variâncias dos grupos de dados estudados nesta pesquisa, segundo a quantidade de componentes (*i.e.*, dimensões) PCA. Os conjuntos de dados 1, 2 e 3 apresentaram diferentes proporções na representatividade das componentes. Estes resultados estão relacionados às características distintas dos conjuntos de dados, a quantidade de atributos, aos tipos de KPI, a data de coleta e a tecnologia da rede.

Figura 22 - Representação da porcentagem da variância explicada pelas componentes PCA para o conjunto de dados 1 (a), conjunto de dados 2 (b) e conjunto de dados 3 (c)



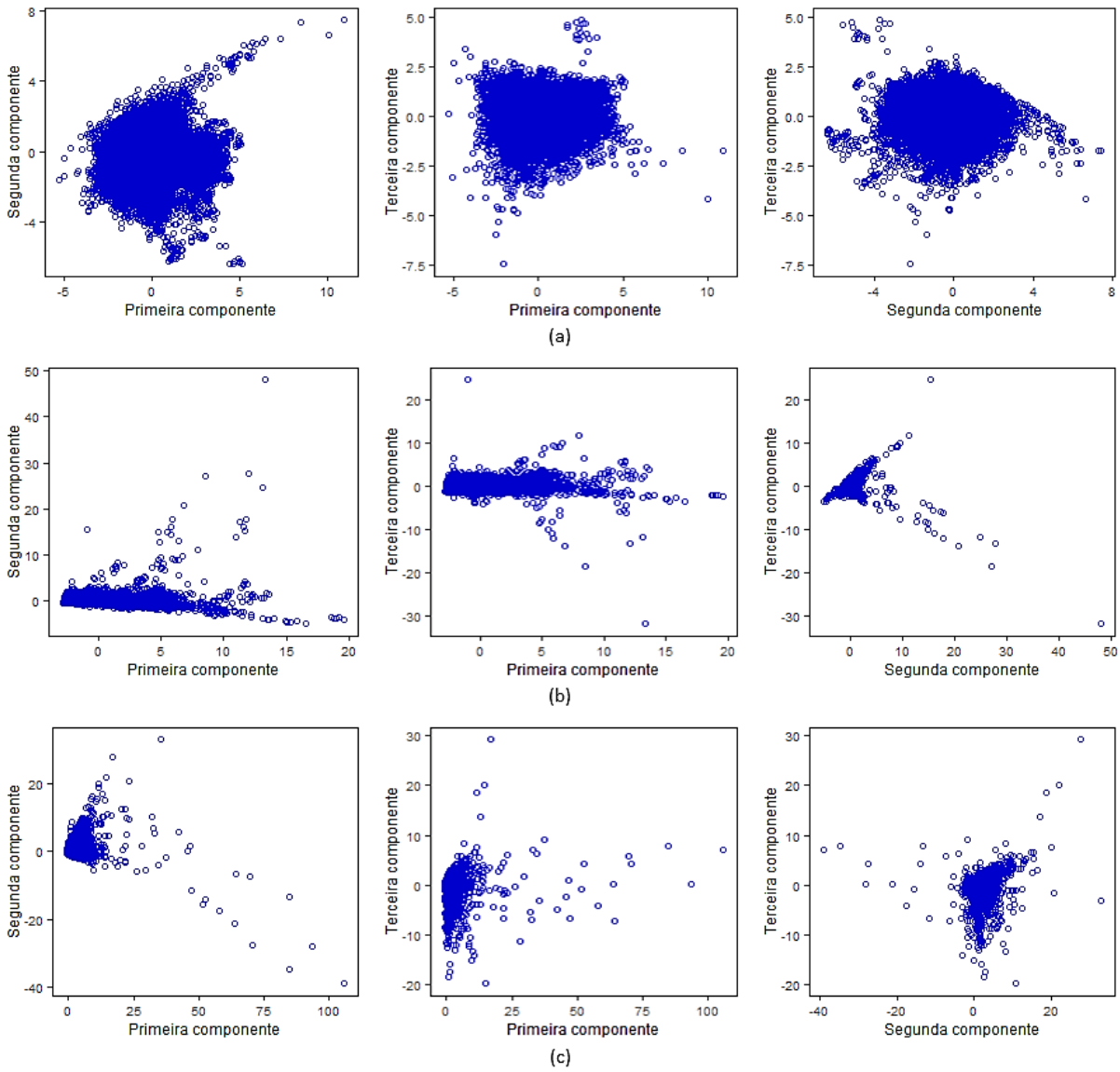
Fonte: autoria própria

Usualmente, o número de componentes PCA considerado deve representar no mínimo 80% da variância do conjunto de dados. Contudo isto não significa que o número de componentes consideradas possa ser menor. Para o conjunto de dados 1, as 3 primeiras componentes representam aproximadamente 80% da variância. Para o conjunto de dados 2, seriam necessárias 4 componentes para se alcançar patamar próximo aos 80%. Neste caso, 4 componentes representam cerca de 86% da variância. Para o conjunto de dados 3, as 9 primeiras componentes representam aproximadamente 80% da variância dos dados. Nos casos em que o número de componentes é grande o suficiente para afetar o desempenho do modelo, é necessário avaliar a possibilidade de considerar uma quantidade menor de componentes em detrimento da representatividade[4] da variância dos dados.

Para ilustrar a diferença na dispersão dos dados dos três conjuntos, a Figura 23 mostra os gráficos gerados com as 3 primeiras componentes PCA. Foram analisadas as projeções de cada plano que compõem o espaço tridimensional criado pelo método de PCA. Nota-se que em todos os conjuntos de dados, a maioria das amostras se concentram em uma região densa e ao redor se localizam algumas amostras dispersas. Este comportamento foi objeto de estudo deste trabalho, especialmente na diferenciação destes pontos de concentração (ou regiões) de amostras, assim como nos casos de amostras dispersas.

[4] Neste sentido, trata-se de representatividade gráfica, uma vez que não se pode produzir gráficos com dimensões maiores que três. Como a maior parte das análises feitas aqui são de natureza gráfica, é inviável o uso de dimensões maiores que estas.

Figura 23 - Gráfico de dispersão das duas primeiras componentes PCA geradas para o conjunto de dados: (a) Conjunto de dados 1, (b) Conjunto de dados 2, e (c) Conjunto de dados 3.



Fonte: autoria própria

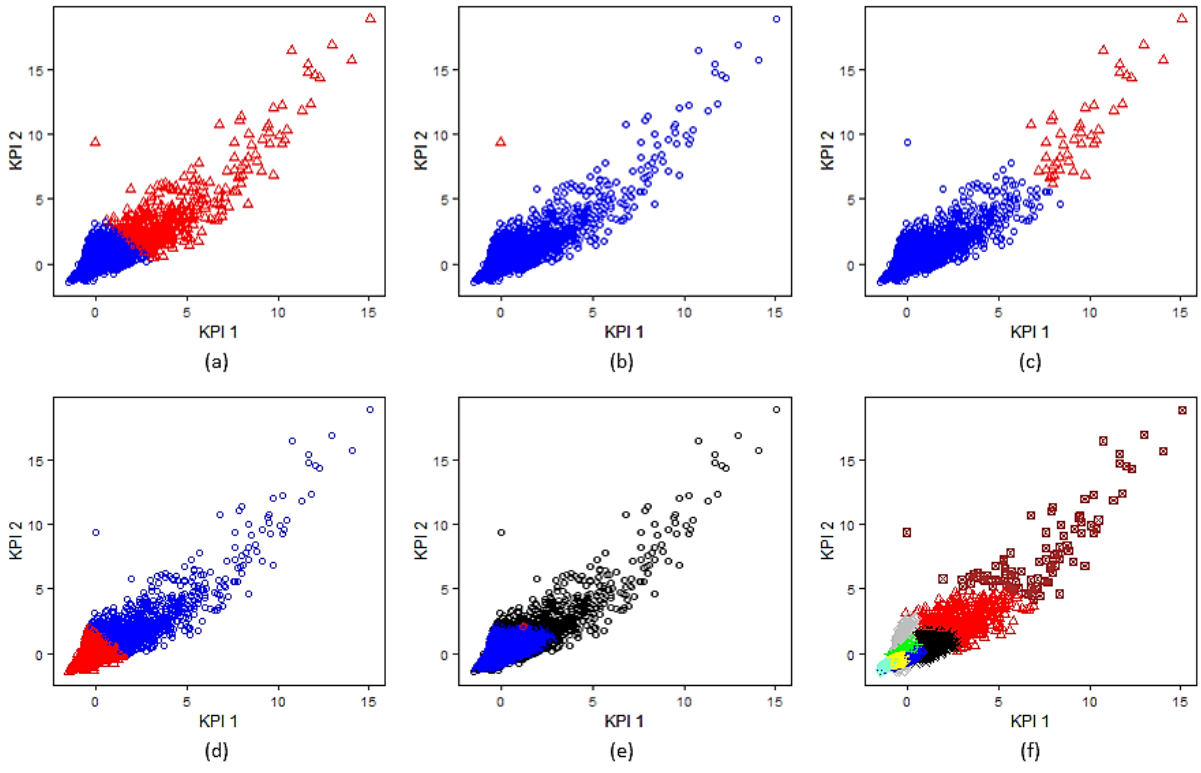
5.1.2 Variações de agrupamento

As figuras a seguir apresentam os resultados das diferentes técnicas de agrupamento utilizando dois KPIs com alta (KPI 1: taxa de dados no *uplink* na visão do usuário; KPI 2: taxa de dados no *uplink* da célula, apresentados na Figura 24), média (KPI 1: taxa de dados no *downlink* na visão do usuário; KPI 2: SINR para PUCCH médio, apresentados na Figura 25) e baixa correlação (KPI 1: taxa de dados no *uplink* na visão do usuário; KPI 2: SINR para PUCCH médio, apresentados na Figura 26). As

técnicas K-*means*, *Single Linkage*, *Complete Linkage* e *Fuzzy C-means* requerem que o número de *clusters* seja informado previamente. Com base nos métodos do "cotovelo" (*elbow method*) e da silhueta (*silhouette method*), o número de *clusters* ideal encontrado foi igual a 2 para alta e média correlações e 3 para baixa correlação. As técnicas DBSCAN e *Model-based* não dependem do número de *clusters* e o agrupamento é realizado com base na densidade e na mistura finita de distribuições de probabilidade, respectivamente.

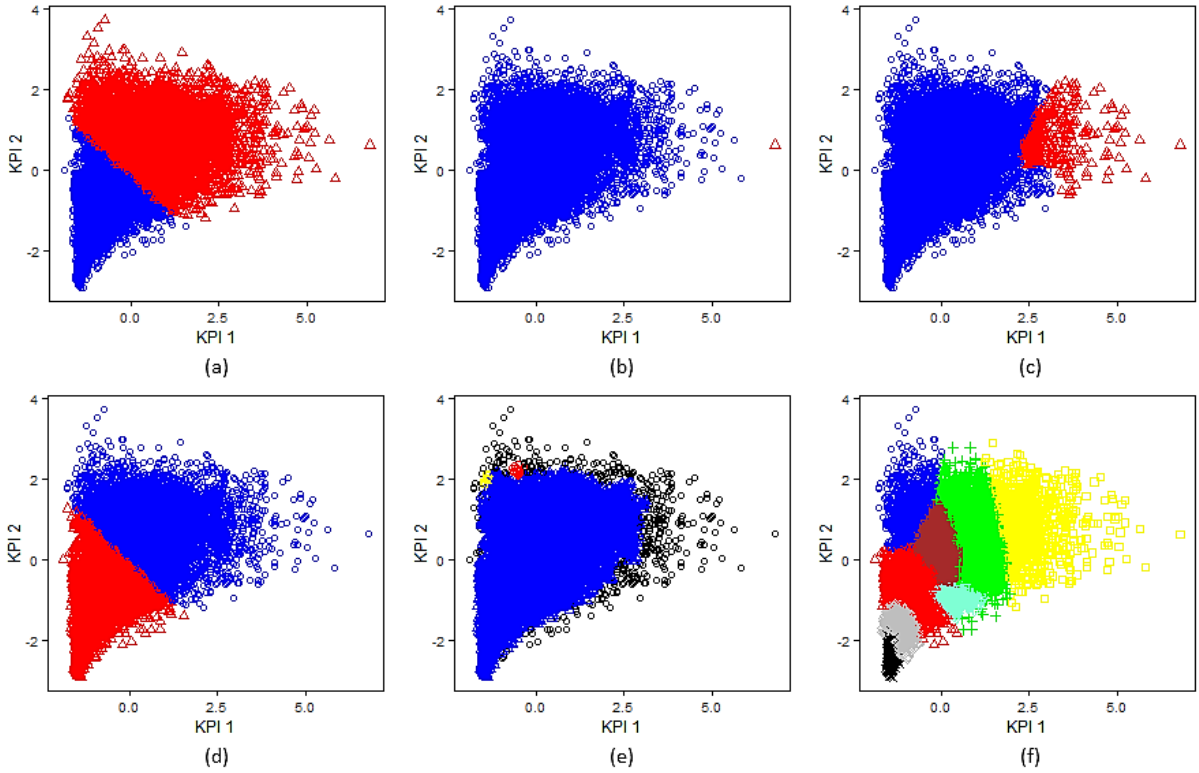
A seleção da técnica de agrupamento utilizada no método proposto foi baseada no significado prático dos *clusters* visando a análise de desempenho das redes móveis. Por meio da inspeção visual dos resultados encontrados, a diferenciação das regiões pela densidade despertou o interesse, visto que uma pequena parcela dos dados se localiza em regiões dispersas e foram classificadas como ruído, o que pode caracterizar um comportamento anormal. Neste sentido, a técnica que apresentou os melhores resultados foi o DBSCAN.

Figura 24 - Gráficos de dispersão de dois KPIs com alta correlação e os *clusters* criados pelas técnicas: (a) K-*means*, (b) *Single Linkage*, (c) *Complete Linkage*, (d) *Fuzzy* C-*means*, (e) DBSCAN e (f) *Model-based*.



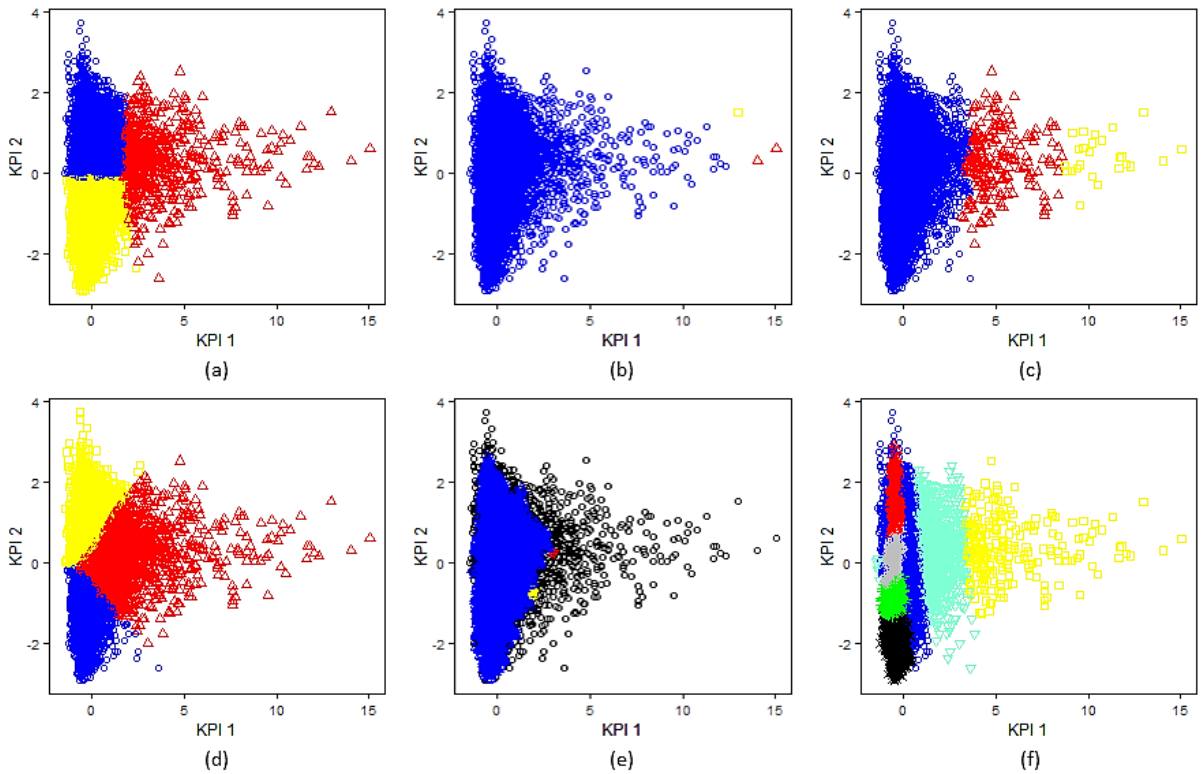
Fonte: autoria própria

Figura 25 - Gráficos de dispersão de dois KPIs com média correlação e os *clusters* criados pelas técnicas: (a) K-*means*, (b) *Single Linkage*, (c) *Complete Linkage*, (d) *Fuzzy* C-*means*, (e) DBSCAN e (f) *Model-based*



Fonte: autoria própria

Figura 26 - Gráficos de dispersão de dois KPIs com baixa correlação e os *clusters* criados pelas técnicas: (a) K-*means*, (b) *Single Linkage*, (c) *Complete Linkage*, (d) *Fuzzy* C-*means*, (e) DBSCAN e (f) *Model-based*



Fonte: autoria própria

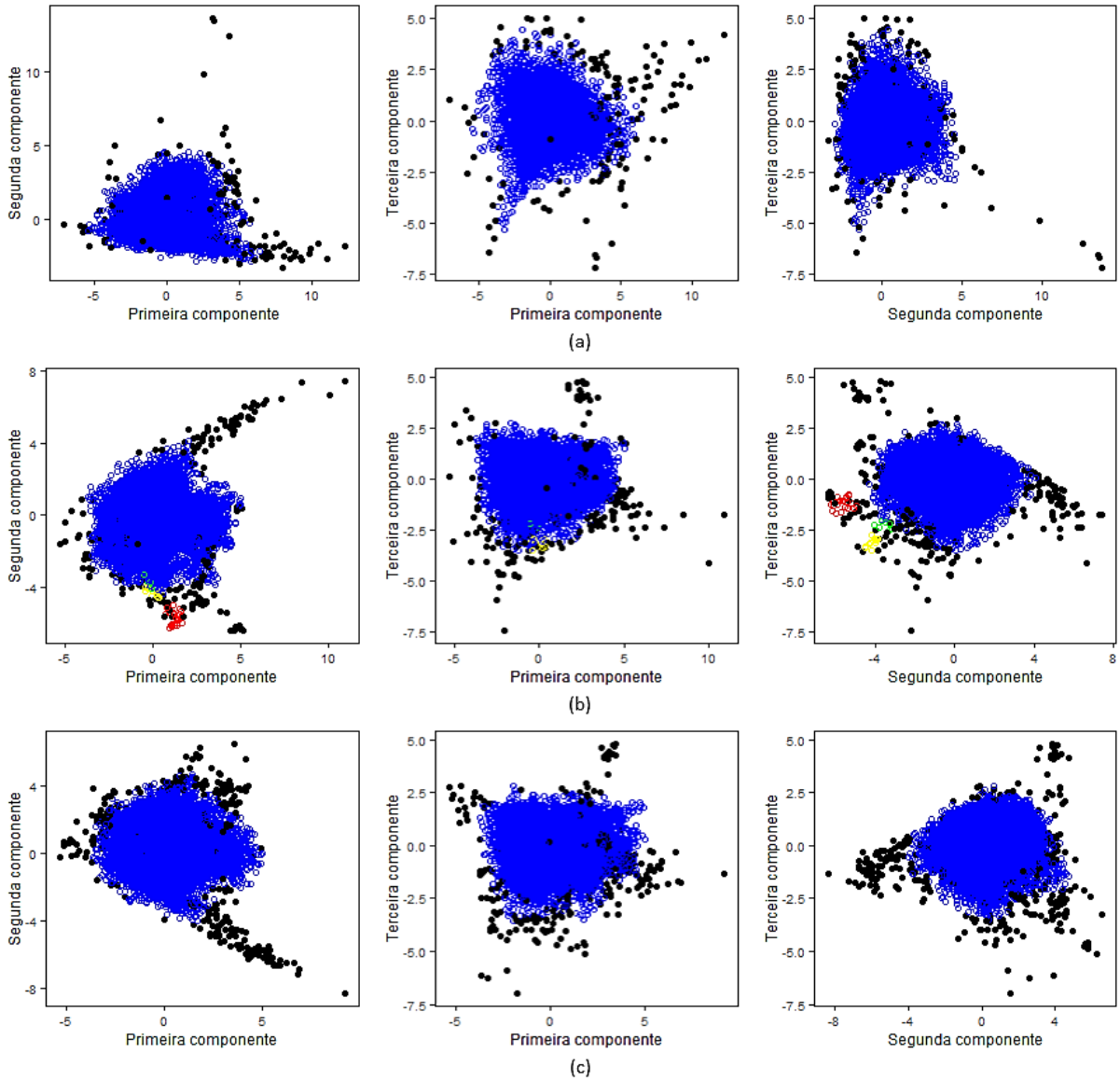
5.2 Resultados e avaliações dos métodos

5.2.1 Método proposto e análise da assertividade

O método proposto foi aplicado no conjunto de dados 1. Na etapa inicial, os dados foram segmentados conforme a periodicidade da série temporal, que no caso foi hora. As próximas etapas foram aplicadas em cada hora separadamente, concluindo com as anomalias identificadas. O número de componentes PCA consideradas foi 3 (que abrange em média cerca de 80% da variância dos dados). A Figura 27 mostra os resultados para três horas diferentes do dia: 04:00h, 12:00h e 20:00h.

As amostras categorizadas como 'anomalias' estão representadas pelo círculo preenchido preto. As demais amostras, categorizadas como comportamento normal, estão representadas pelo círculo não preenchido em azul ou outras cores (casos em que se foi criado mais de um *cluster*). Os gráficos das demais horas foram disponibilizados no Apêndice A para visualização.

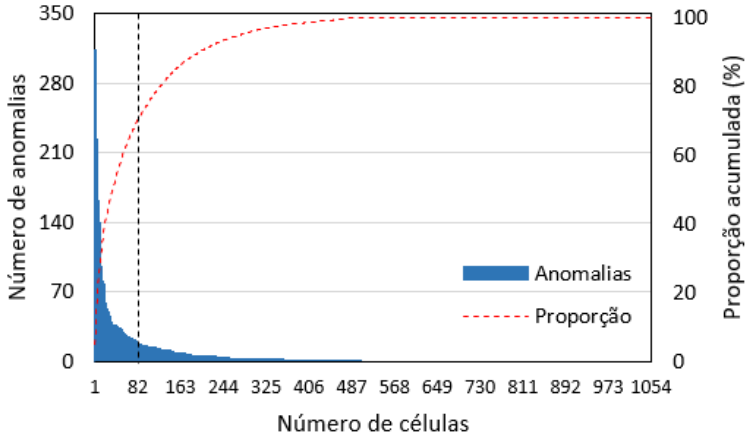
Figura 27 - Gráficos de dispersão das componentes PCA e o resultado da técnica DBSCAN. Os gráficos são referentes a três horas diferentes: (a) 04:00, (b) 12:00 e (c) 20:00



Fonte: autoria própria

O método de detecção de anomalias proposto apontou pelo menos uma potencial anomalia, considerando todos os horários, em 501 células (ou 47,5% do total de células) do conjunto de dados 1. A Figura 28 mostra a quantidade de anomalias detectadas por célula, ordenada de forma decrescente, e a proporção acumulada. Observa-se que uma grande parcela das anomalias se concentra em um número pequeno de células. Por exemplo, a linha vertical tracejada mostra que cerca de 70% de todas as anomalias foram detectadas em 82 células (ou 7,8% do total de células).

Figura 28 - Representação gráfica da quantidade de anomalias detectadas por célula (barras em azul) e a proporção acumulada (linha tracejada)

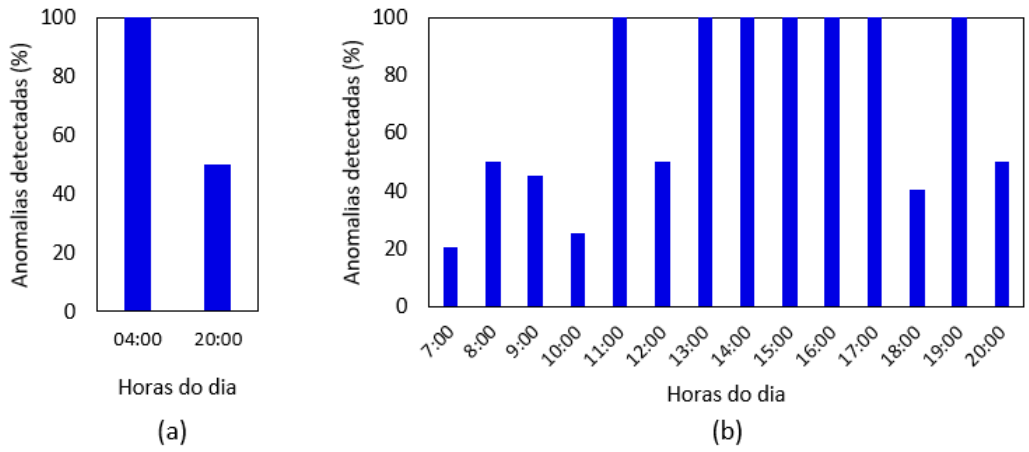


Fonte: autoria própria

Para avaliar a assertividade do método proposto, foram considerados alguns padrões característicos de degradação na rede celular. A partir destes padrões, as amostras foram alteradas "emulando" uma degradação com amplitude máxima de três vezes o desvio padrão (3σ). As anomalias foram inseridas em dez células da rede que não apresentaram amostras categorizadas como anomalia pelo método proposto e com a menor pontuação de anormalidade dada pelo método de referência. A Figura 29 apresenta os resultados obtidos pelo método proposto nos padrões impulso e degrau. As anomalias foram inseridas com amplitude de 3σ no impulso e em todos os horários no degrau.

Os resultados para o impulso (Figura 29a) mostram que às 04:00 foram detectadas 100% das anomalias e às 20:00 detectadas 50% das anomalias. No período de baixo tráfego (04:00) não é comum uma célula ter valores próximo a três vezes o desvio padrão, mas este comportamento anormal foi identificado pelo método. Por outro lado, nos períodos de alto tráfego (20:00), uma parcela das células apresenta degradação próxima das células em que foram inseridas as anomalias. Esta característica associada à forma como método categoriza as anomalias, que será discutido ao longo desta seção, pode ter dificultado a detecção total das anomalias inseridas.

Figura 29 - Gráficos da porcentagem de anomalias detectadas pelo método proposto nos padrões de degradação impulso (a) e degrau (b). Neste último, a linha tracejada representa a degradação inserida nos diferentes horários



Fonte: autoria própria

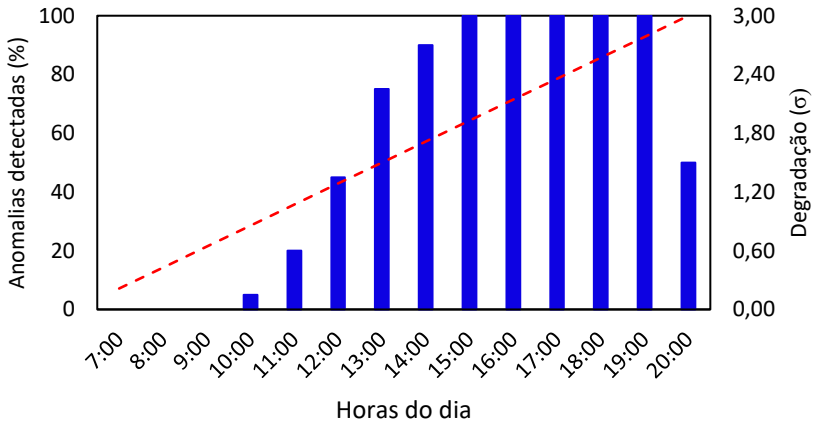
No caso das anomalias inseridas seguindo o padrão degrau para inserção das anomalias, a amplitude considerada para todos os horários foi de 3σ e o resultado da detecção está representado na Figura 29b. Observa-se que em 50% dos horários a porcentagem de detecção das anomalias foi de 100%. Por outro lado, nos outros 50% dos horários a detecção foi menor ou igual a 50%.

Apesar da amplitude inserida ser suficiente para a amostra ser potencialmente considerada uma anomalia, em alguns horários o método não obteve bons resultados. Nestes casos, as amostras anômalas foram categorizadas no *cluster* principal ou em algum *cluster* pequeno adjacente, que representam o comportamento normal dos dados. No método proposto, as amostras categorizadas como anomalia são aquelas que não são conectadas em densidade com os pontos adjacentes, conforme o valor do parâmetro *Eps*, e, portanto, não pertencem a algum *cluster*. Por fim, o método conseguiu uma média de 70% de detecção das anomalias inseridas seguindo o padrão de degrau.

No padrão de degradação rampa (Figura 30), os resultados mostram que a identificação se iniciou às 10:00 com 5% das anomalias e cerca de 0,86σ. Às 13:00 foram detectadas 75% das anomalias inseridas, período no qual a amplitude das anomalias inseridas é 1,5σ. De 15:00 às 19:00, 100% das anomalias inseridas foram detectadas, o que representa um bom desempenho do método para as amplitudes de degradação

consideradas. De forma geral, o método apresentou um bom desempenho na detecção das anomalias inseridas, principalmente a partir das 12:00, período em que a amplitude das anomalias é maior que 1,5σ.

Figura 30 - Gráficos da porcentagem de anomalias detectadas pelo método proposto no padrão de degradação rampa



Fonte: autoria própria

Um caso interessante pode ser observado no horário das 18:00. Apesar da amplitude da anomalia ser menor no padrão de rampa (2,57σ), o método detectou 100% das anomalias. Entretanto, no padrão de degrau (3σ), o método detectou 40% das anomalias. Esta diferença está relacionada com a formação dos *clusters* explicada anteriormente. As amostras anômalas com amplitude menor foram identificadas pelo método por não pertencerem a algum *cluster*.

5.2.2 Análise individual das anomalias

As células que apresentaram um número elevado de anomalias exibem um comportamento anômalo recorrente, ou seja, a característica de anormalidade foi identificada pelo método na maioria dos dias contidos no conjunto de dados. A partir destes resultados, foram selecionadas as vinte células que exibiram mais anomalias para uma análise direta de dois profissionais da operadora habilitados para identificar qual

seria a possível anormalidade. Estas células apresentaram cerca de 40% do total de anomalias detectadas e os resultados estão representados na Tabela 4.

Os comportamentos identificados pelos profissionais para cada uma das células foram classificados em uma das categorias listadas na sequência:

- Normal: cenário em que não foi observado comportamento anômalo;
- *Overshooting*: quando a cobertura da célula forma uma área de cobertura descontínua, longa ou sobreposta em outras células adjacentes. Geralmente este comportamento reflete diretamente nos KPIs de SINR e RSSI.
- *Undershooting*: a cobertura da célula forma uma área de cobertura curta e insuficiente para a região planejada, considerando a cobertura das células adjacentes. Este comportamento geralmente também reflete diretamente nos KPIs de SINR e RSSI.
- Sobrecarga: nível elevado de utilização dos recursos da célula com degradação da experiência dos usuários. Este comportamento comumente afeta os KPIs de número de usuários ativos e taxa de dados média.
- Subutilização: nível baixo de utilização dos recursos da célula. Neste caso, os KPIs apresentam ótimos valores de SINR e RSSI e baixo número de usuários conectados.
- Interferência externa: existência de uma fonte externa que irradia um sinal na mesma frequência da célula e provoca interferência no sistema. Este comportamento reflete diretamente nos KPIs de SINR e RSSI.
- Falha no sistema irradiante: característica de algum problema no *hardware* do sistema irradiante (conectores, cabos e antena). Este comportamento é exposto diretamente nos KPIs de SINR e RSSI.

Tabela 4 - Avaliação prática das vinte células com o maior número de anomalias detectadas pelo método proposto

Célula ID	Número de Anomalias	Avaliação Profissional 1	Avaliação Profissional 2
1	313	*Overshooting*	Subutilização
2	223	Falha no sistema irradiante	Interferência externa
3	219	*Overshooting*	*Overshooting*
4	216	Sobrecarga	Sobrecarga
5	162	Sobrecarga	Sobrecarga
6	143	Sobrecarga	Sobrecarga
7	141	Normal	Normal
8	139	Normal	Normal
9	119	Interferência externa	Interferência externa
10	99	*Overshooting*	*Overshooting*
11	96	Falha no sistema irradiante	Normal
12	91	Sobrecarga	Sobrecarga
13	81	*Overshooting*	*Overshooting*
14	81	Sobrecarga	Sobrecarga
15	78	*Undershooting*	*Undershooting*
16	77	Normal	Subutilização
17	72	*Overshooting*	*Overshooting*
18	65	Subutilização	Subutilização
19	59	Normal	Normal
20	56	Subutilização	Subutilização

Destaca-se, novamente, que o método proposto não tem o objetivo de diagnóstico de comportamento de células, mas, sim, apenas a possível identificação de uma anomalia que, posteriormente, se confirmada, pode ser classificada como uma destas opções.

Os resultados da tabela anterior mostram que, segundo a avaliação dos profissionais, em 80% dos casos foi identificado um comportamento anormal nas células, dado que o objetivo principal do método é a detecção das anomalias. O comportamento foi considerado normal nas células 7, 8 e 19 por ambos os profissionais, enquanto que as células 11 e 16 foram consideradas normal por apenas um dos profissionais.

Além disso, em aproximadamente 80% das células, houve compatibilidade direta nos resultados da avaliação realizada pelos profissionais. Neste caso, as células 3, 10, 13 e 17 foram avaliadas como *overshooting*, as células 4, 5, 6, 12 e 14 foram avaliadas como sobrecarga, a célula 9 como interferência externa, a célula 15 como *undershooting* e as células 18 e 20 como subutilização. Por outro lado, não houve compatibilidade direta nas

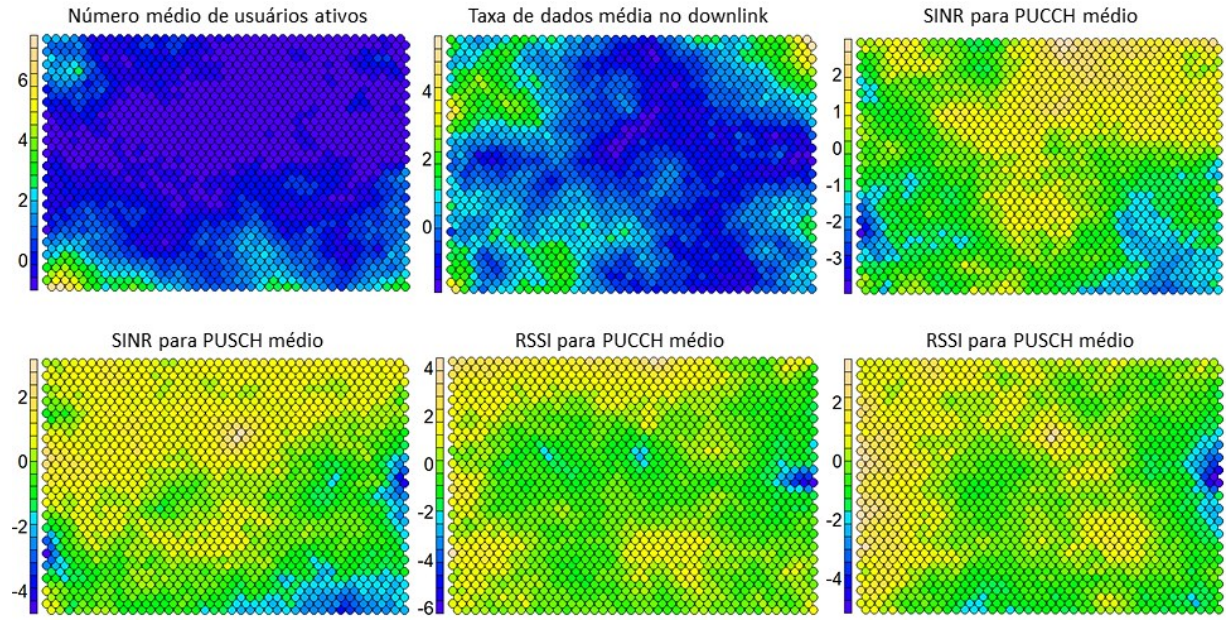
avaliações das células 1, 2, 11 e 16. Neste caso, existe a possibilidade de que a anormalidade da célula tenha mais de uma causa primária, de forma que a solução de somente uma das causas não fosse o suficiente para a que a célula tivesse um comportamento normal, como ocorre nas células 1 e 2. Além disso, existe a possibilidade de contradição em que um dos profissionais aponta alguma anormalidade e o outro não, como é o caso das células 11 e 16.

Embora este tipo de avaliação contenha algum grau de subjetividade, a seleção de dois profissionais foi uma alternativa adicional para ajudar a avaliar a assertividade de identificação de anomalia do método. Na sequência, a assertividade do método foi testada por meio da inserção de anomalias sintéticas em determinadas células.

5.2.3 Método de referência e análise comparativa

O método desenvolvido para comparação com o método proposto é baseado na combinação das técnicas SOM e LoOP. Os valores da magnitude de cada KPI no *codebook* (vetores de peso) da técnica SOM estão representados na Figura 31 em forma de mapas de calor. Neste caso, foram selecionados os resultados das 20:00, em específico, como exemplo. Os demais horários foram disponibilizados no Apêndice B.

Figura 31 – Mapas de calor do *codebook* da rede SOM para cada atributo do conjunto de dados 1 no horário das 20:00

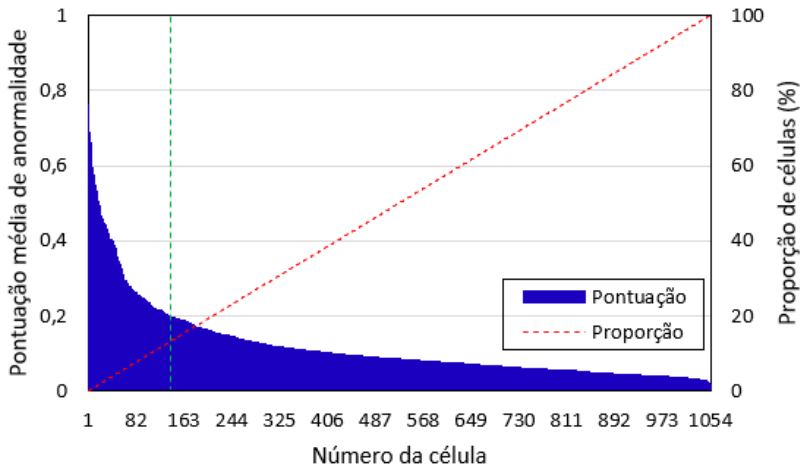


Fonte: autoria própria

Os mapas de calor gerados mostram a distribuição dos valores no *codebook* de cada KPI ao longo da rede SOM e são essenciais para visualizar a relação entre os KPIs nas diferentes regiões. A seguir, os dados do *codebook* foram usados na etapa de detecção das anomalias com a técnica LoOP. O resultado desta técnica é uma pontuação (*score*) de anormalidade de cada nó da rede SOM, que pode ser interpretado como uma probabilidade de anormalidade. A partir dessa informação, foi possível determinar a pontuação de anormalidade de cada célula, visto que as células são atribuídas aos nós da rede SOM. Por fim, o resultado obtido com este método foi a pontuação de anormalidade de cada célula do conjunto de dados em cada horário.

Como forma de se obter um resultado único por célula, as pontuações das instâncias foram agregadas usando a média aritmética. A Figura 32 mostra a pontuação média de anormalidade das células, ordenada de forma decrescente. Assim como no método proposto, uma pequena parcela de células apresentou um comportamento caracterizado por um elevado número de anomalias ou pontuação média de anormalidade, se comparado às demais células. Para exemplificar, 137 células (13% do total) apresentaram uma pontuação média de anormalidade maior ou igual a 0,2.

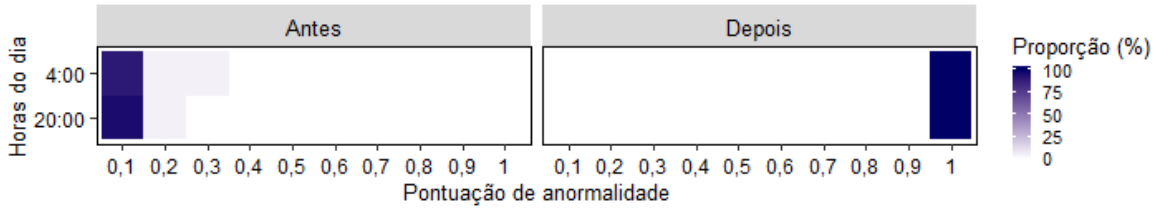
Figura 32 - Representação gráfica da pontuação média de anormalidade por célula (barras em azul) e a proporção acumulada do número de células (linha tracejada)



Fonte: autoria própria

A análise comparativa dos métodos proposto e de referência foram realizadas com base nos padrões de degradação para inserção das anomalias utilizados. As mesmas dez células selecionadas anteriormente para a inserção das anomalias foram consideradas neste caso. A inserção das anomalias sintéticas foi realizada em dois dias de um total de quatorze do conjunto de dados 1. Dessa forma, o número total de instâncias em que as anomalias foram inseridas é igual a vinte, em cada horário. A Figura 33 apresenta os resultados obtidos pelo método de referência no padrão impulso.

Figura 33 - Mapas de calor da proporção de pontuação de anormalidade das instâncias em diferentes horários no padrão de degradação impulso



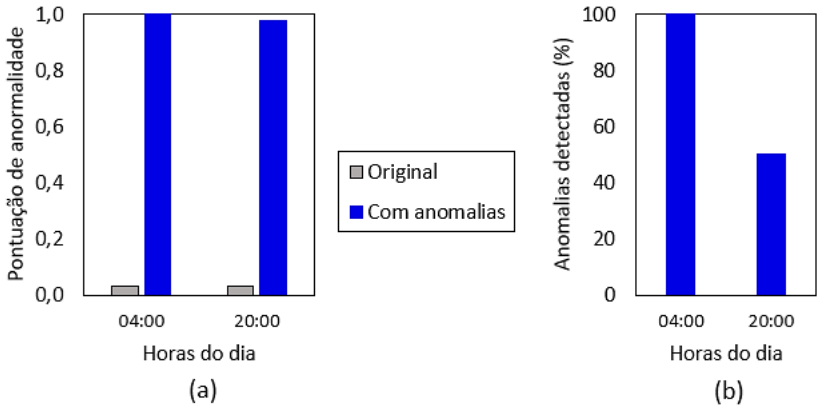
Fonte: autoria própria

No cenário em que método de referência foi aplicado nas instâncias originais (cenário "Antes"), a pontuação de anormalidade se concentra em grande parte (90% às 4:00 e 95% às 20:00) entre 0 e 0,1, e algumas pequenas parcelas de 5% em até 0,2 e 0,3. Com a inserção das anomalias (cenário "Depois"), 100% das instâncias apresentaram uma pontuação de anormalidade entre 0,9 e 1. Estes resultados mostram que o método de referência foi capaz de identificar as anomalias inseridas nos dois horários, visto que houve um aumento na pontuação de anormalidade.

A seguir, os resultados foram agregados usando a média aritmética, o que permitiu apresentar uma visão simplificada por horário, como uma forma de comparar com os resultados do método proposto. Como as células selecionadas para a inserção das anomalias são as mesmas selecionadas nos testes da seção anterior, os resultados obtidos com o método proposto são os mesmos e são apresentados novamente para serem comparados. A Figura 34 mostra a pontuação média de anormalidade pelo método de

referência e a porcentagem de anomalias detectadas pelo método proposto nos dois horários.

Figura 34 - Gráficos da pontuação média de anormalidade pelo método de referência (a) e da porcentagem de anomalias detectadas pelo método proposto (b) no padrão de degradação impulso

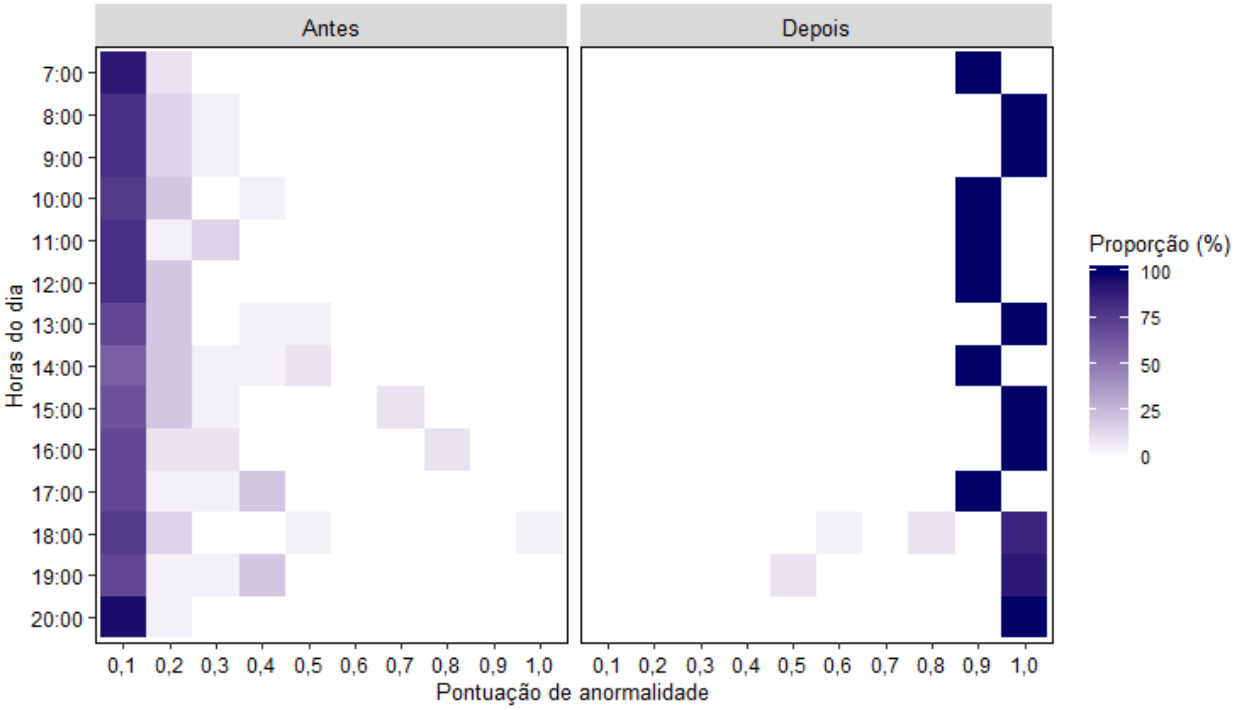


Fonte: autoria própria

Os resultados obtidos com o método de referência (Figura 34a) mostram que no horário das 04:00h a pontuação média de anormalidade das células passou de 0,03 para 1,00 com a inserção das anomalias. No horário das 20:00h, a pontuação média variou de 0,03 para 0,98. Os resultados encontrados mostram que o método de referência obteve um bom desempenho na identificação das anomalias inseridas, visto que a pontuação máxima de anormalidade possível é 1,00. Se comparado ao método proposto (Figura 34b), no horário das 20:00h, o método de referência conseguiu identificar um número maior de anomalias inseridas baseado na pontuação média de anormalidade.

No padrão de degradação degrau, as anomalias foram inseridas no período de 7:00h às 20:00h. A proporção da pontuação de anormalidade das instâncias está representada na Figura 35. As células selecionadas para a inserção das anomalias com as instâncias originais (cenário “Antes”) apresentam uma baixa pontuação de anormalidade, que se concentram majoritariamente entre 0 e 0,1. Com a inserção das anomalias (cenário “Depois”) de amplitude igual a 3σ em todos os horários, a pontuação de anormalidade aumenta significativamente.

Figura 35 - Mapas de calor da proporção de pontuação de anormalidade das instâncias em diferentes horários no padrão de degradação degrau

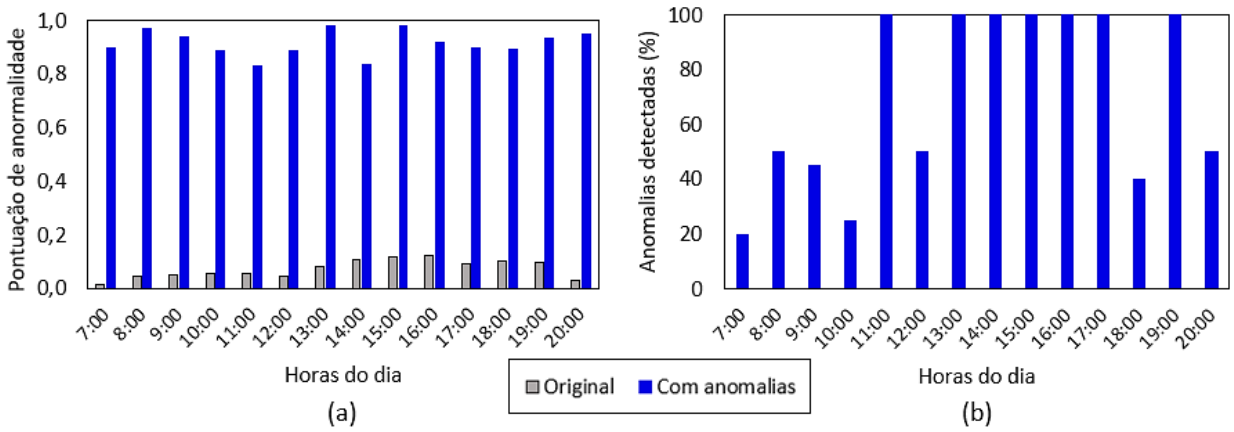


Fonte: autoria própria

A proporção de anormalidade é de 100% entre 0,8 a 0,9 e 0,9 a 1 nos diferentes horários, exceto nos horários das 18:00h e 19:00h em que existem uma pequena parcela das instâncias que obtiveram uma pontuação de anormalidade menor que 0,8. Estes resultados mostram que o método de referência também foi capaz de identificar as anomalias inseridas nos horários, dado o aumento na pontuação de anormalidade.

A seguir, a Figura 36 mostra a pontuação média de anormalidade e a porcentagem de anomalias detectadas pelo método proposto. Observa-se que a pontuação média de anormalidade dada pelo método de referência (Figura 36a) foi significativamente alta com a inserção das anomalias. De forma geral, a pontuação média nos diferentes horários foi superior a 0,8 e no máximo 0,98, que ocorreu às 13:00h e às 15:00h. O método proposto (Figura 36b) em determinados horários não foi capaz de identificar todas as anomalias inseridas, contudo apresentou uma taxa média de assertividade de 70%.

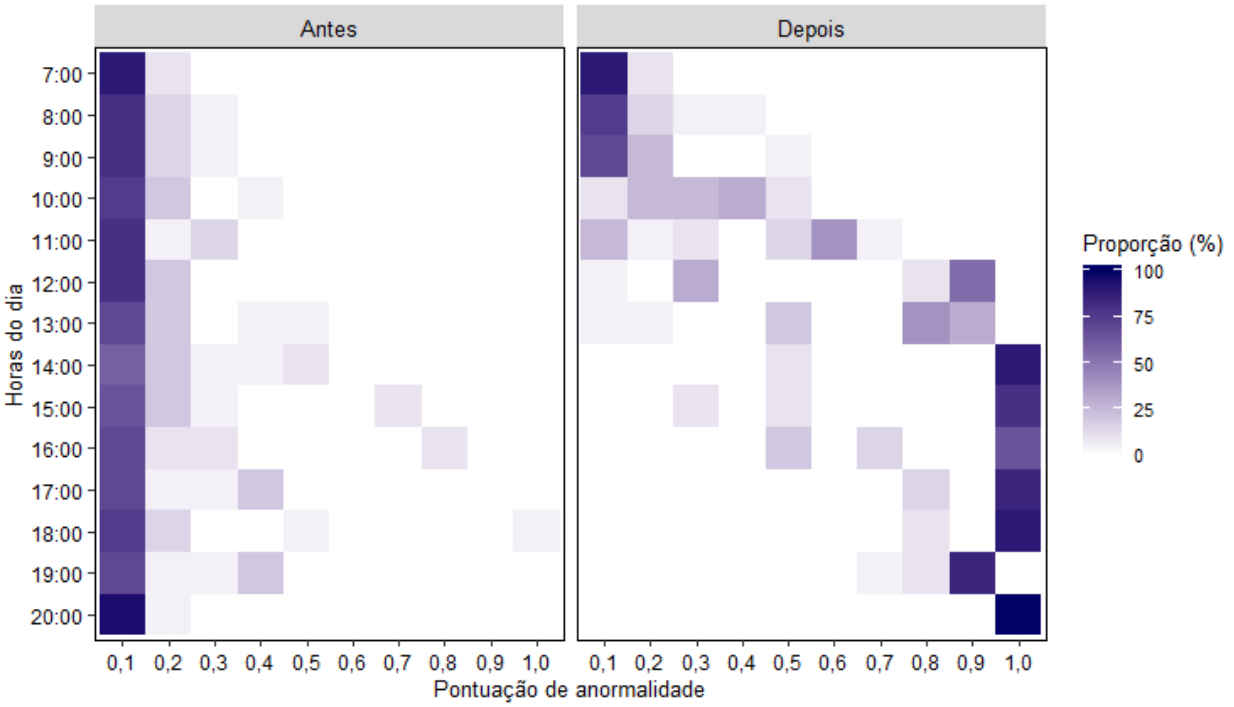
Figura 36 - Gráficos da pontuação média de anormalidade pelo método de referência (a) e da porcentagem de anomalias detectadas pelo método proposto (b) no padrão de degradação degrau nos diferentes horários



Fonte: autoria própria

Por fim, os resultados obtidos pelo método de referência com a inserção das anomalias seguindo o padrão de degradação rampa estão representados na Figura 37. Neste caso, as anomalias foram inseridas com uma amplitude que varia de forma linear e incremental, com parcelas de 1/14 de 3σ, até que o valor final seja de 3σ.

Figura 37 - Mapas de calor da proporção de pontuação de anormalidade das instâncias em diferentes horários no padrão de degradação degrau

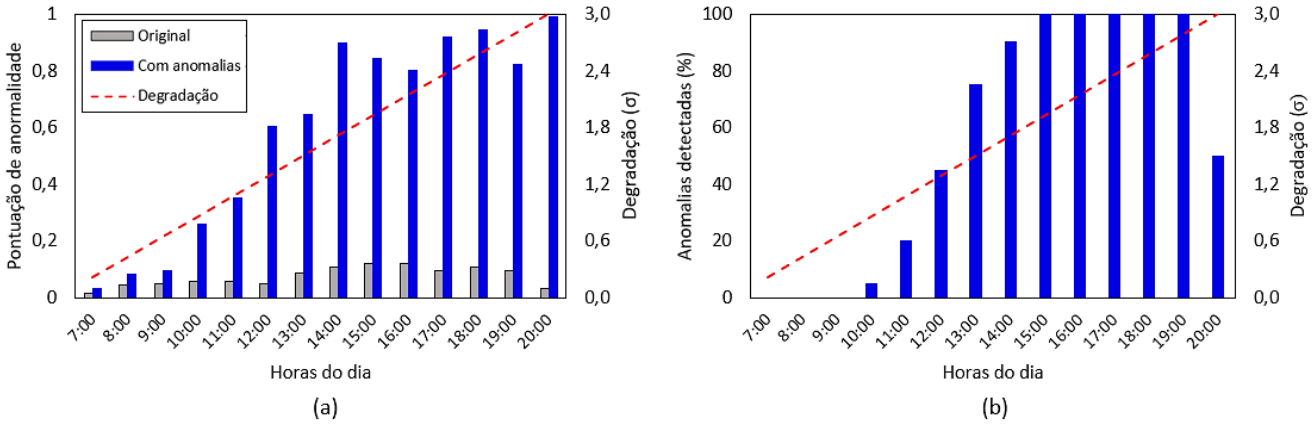


Fonte: autoria própria

Neste caso, observa-se com mais clareza o aumento da proporção da pontuação de anormalidade das instâncias à medida em que a amplitude das anomalias sintéticas inseridas aumentou. Nos primeiros horários é possível notar o aumento na proporção, que fica evidente a partir das 10:00h, horário em que a amplitude das anomalias inseridas foi de 0,86σ. A partir das 14:00h (1,71σ), a proporção da pontuação de anormalidade se concentrou em grande parte entre 0,8 a 0,9 e 0,9 a 1.

A Figura 38 mostra a pontuação média de anormalidade nos diferentes horários em conjunto com os resultados do método proposto. Além disso, no eixo vertical secundário dos gráficos foi adicionado o valor de amplitude das anomalias inseridas ao longo dos horários. Nos resultados do método de referência (Figura 38a), nota-se que em todos os horários a pontuação média de anormalidade aumentou com a inserção das anomalias, mesmo nos casos de menor amplitude, que se inicia às 07:00h (0,21σ). A partir das 10:00h (0,86σ), a pontuação média de anormalidade aumenta de forma acentuada, como visto anteriormente, passando de 0,06 para 0,26 neste horário e chegando a 0,98 às 20:00h, horário no qual a amplitude das anomalias inseridas foi de 3σ.

Figura 38 - Gráficos da pontuação média de anormalidade pelo método de referência (a) e da porcentagem de anomalias detectadas pelo método proposto (b) no padrão de degradação rampa. A linha tracejada representa a degradação inserida nas anomalias sintéticas nos diferentes horários



Fonte: autoria própria

Neste cenário foi possível notar que o comportamento dos métodos proposto (Figura 38b) e de referência na detecção das anomalias foram semelhantes nos horários

à medida em que a amplitude das anomalias era incrementada. O método proposto foi capaz de identificar 100% das anomalias das 15:00h às 19:00h, período em que a pontuação média de anormalidade foi de no mínimo 0,8. Uma diferença na capacidade de identificar as anomalias inseridas ocorreu no horário das 20:00, em que o método proposto identificou 50%, como apresentado na seção anterior.

Os resultados encontrados com a inserção das anomalias sintéticas seguindo os três padrões de degradação mostram que o método de referência apresentou uma maior capacidade de identificação das anomalias inseridas em determinados horários, nos casos em que a amplitude foi de 3σ. Por outro lado, possui como característica a necessidade da definição de um limiar indicando se o comportamento é anormal ou não, diferentemente do método proposto, que apresentou bons resultados na detecção das anomalias em todos os cenários, principalmente no cenário de degradação rampa em que identificou 100% das anomalias inseridas com amplitudes menores que 3σ no período de 15:00h às 19:00h.

5.2.4 Análise individual das anomalias incluindo o método de referência

A seguir foram selecionadas as vinte células que apresentaram a maior pontuação média de anormalidade (acima de 0,49) para uma análise direta de dois profissionais da operadora habilitados para identificar qual seria a possível anormalidade. Em conjunto, foram adicionadas as vinte células que apresentaram a maior quantidade de anomalias detectadas pelo método proposto (Seção 5.2.1). A inclusão destas células permitiu avaliar o cenário em que o método proposto detectou um grande número de anomalias e o método de referência indicou uma baixa pontuação média de anormalidade.

No conjunto de quarenta células consideradas para a avaliação, seis estavam presentes na seleção dos dois métodos. Portanto, ao todo, foram avaliadas trinta e quatro células e os resultados estão representados na Tabela 5. As células estão organizadas de forma decrescente em relação ao número de anomalias detectadas pelo método proposto.

Dessa forma, a avaliação das vinte primeiras células e a descrição dos comportamentos identificados pelos profissionais são apresentadas na Seção 5.2.1.

Tabela 5 - Avaliação prática de um grupo de trinta e quatro células que apresentaram a maior quantidade de anomalias detectadas pelo método proposto e a maior pontuação média de anormalidade dada pelo método de referência

Célula ID	Anomalias Método Proposto	Pontuação Média de Anormalidade	Avaliação Profissional 1	Avaliação Profissional 2
1	313	0,67	*Overshooting*	Subutilização
2	223	0,76	Falha no sistema irradiante	Interferência Externa
3	219	0,51	*Overshooting*	*Overshooting*
4	216	0,60	Sobrecarga	Sobrecarga
5	162	0,40	Sobrecarga	Sobrecarga
6	143	0,44	Sobrecarga	Sobrecarga
7	141	0,49	Normal	Normal
8	139	0,45	Normal	Normal
9	119	0,40	Interferência externa	Interferência externa
10	99	0,33	*Overshooting*	*Overshooting*
11	96	0,53	Falha no sistema irradiante	Normal
12	91	0,26	Sobrecarga	Sobrecarga
13	81	0,29	*Overshooting*	*Overshooting*
14	81	0,21	Sobrecarga	Sobrecarga
15	78	0,44	*Undershooting*	*Undershooting*
16	77	0,48	Normal	Subutilização
17	72	0,25	*Overshooting*	*Overshooting*
18	65	0,40	Subutilização	Subutilização
19	59	0,31	Normal	Normal
20	56	0,43	Subutilização	Subutilização
21	47	0,69	Falha no sistema irradiante	*Undershooting*
22	43	0,55	Normal	Subutilização
23	37	0,66	*Overshooting*	*Overshooting*
24	35	0,58	Subutilização	Subutilização
25	35	0,57	Normal	Normal
26	28	0,54	Subutilização	Subutilização
27	28	0,57	Normal	Normal
28	23	0,51	Subutilização	Subutilização
29	9	0,57	Subutilização	Subutilização
30	6	0,66	Subutilização	Subutilização
31	5	0,63	Subutilização	Subutilização
32	3	0,50	Subutilização	Subutilização
33	3	0,50	Subutilização	Subutilização
34	2	0,73	Sobrecarga	Sobrecarga

Os resultados da tabela mostram que, segundo a avaliação dos profissionais, em 80,9% dos casos foi identificado um comportamento anormal nas células, dado que o objetivo principal do método é a detecção das anomalias. O comportamento foi considerado normal nas células 7, 8, 19, 25 e 27 por ambos os profissionais, enquanto que as células 11, 16 e 22 foram consideradas normal por apenas um dos profissionais.

Além disso, em 82,3% das células houve compatibilidade direta nos resultados da avaliação realizada pelos profissionais. Neste caso, as células 3, 10, 13, 17 e 23 foram avaliadas como *overshooting*, as células 4, 5, 6, 12, 14 e 34 foram avaliadas como sobrecarga, a célula 9 como interferência externa, a célula 15 como *undershooting* e as células 18, 20, 24, 26 e 28 a 33 como subutilização.

Por outro lado, não houve compatibilidade direta nas avaliações das células 1, 2, 11, 16, 21 e 22. Neste caso, existe a possibilidade de que a anormalidade da célula tenha mais de uma causa primária, de forma que a solução de somente uma das causas não seja suficiente para a que a célula tenha um comportamento normal, como ocorre nas células 1 e 2. Além disso, existe a possibilidade de contradição em que um dos profissionais aponta alguma anormalidade e o outro não, como é o caso das células 11, 16, 21 e 22.

Como mencionado anteriormente, algumas células foram selecionadas em ambos os métodos, quais sejam: 1, 2, 3, 4, 7 e 11. Este resultado reforça a evidência de que as células apresentam um comportamento anormal, porém, segundo a avaliação dos profissionais, isto se aplica as células 1, 2, 3 e 4. No caso da célula 7, a avaliação apontou um comportamento normal e para a célula 11, apenas um dos profissionais, a avaliou como um comportamento normal. Como as avaliações contém um certo grau de subjetividade, existia a possibilidade deste cenário ocorrer.

Alguns outros casos merecem ser destacados; por exemplo, algumas células apresentaram relativamente um alto número de anomalias detectadas pelo método proposto e uma baixa pontuação média de anormalidade pelo método de referência. No cenário contrário isto também pode ocorrer, ou seja, algumas células apresentam

relativamente um baixo número de anomalias detectadas pelo método proposto e uma alta pontuação média de anormalidade pelo método de referência. Como cada método aplica técnicas diferentes, a detecção do comportamento anormal é realizada por abordagens diferentes. Portanto, as avaliações dos profissionais se fizeram importantes para determinar o comportamento da célula.

Os valores obtidos pelo método proposto (coluna 2 da Tabela 5) e pelo método de referência (coluna 3 da Tabela 5) foram avaliados por correlação, mas, antes, foram normalizados. O resultado foi de -0.2773 para o método de Spearman e 0.0103 para Pearson, ambos tidos como baixa correlação.

Por fim foi realizada uma breve análise do impacto do tempo de execução dos métodos proposto e de referência. Apesar de não ser objetivo do trabalho, estas informações permitem verificar a viabilidade dos métodos em aplicações próximas ao tempo de geração das medições. A Tabela 6 mostra o tempo de execução dos métodos à medida em que se aumenta o número de instâncias no conjunto de dados 1.

Tabela 6 - Tempo de execução dos métodos proposto e de referência pelo número de instâncias do conjunto de dados 1

Instâncias (x100.000)	Método proposto	Método de referência
3,5	00h 00m 21s	00h 26m 12s
6,9	00h 00m 43s	00h 51m 53s
10,4	00h 01m 04s	01h 18m 12s
13,8	00h 01m 24s	01h 43m 49s
17,3	00h 01m 38s	02h 09m 52s

Observa-se que o método proposto apresenta um tempo de execução baixo se comparado ao método de referência, se mostrando viável em aplicações próximas ao tempo de geração das medições de 5, 15, 30 e 60 minutos, por exemplo, com base no número de instâncias considerado. Porém o método de referência é inviável para este tipo de cenário, principalmente para as periodicidades de 5 e 15 minutos. Isto ocorre devido ao treinamento e ajuste dos pesos da rede SOM, que reprocessam as instâncias a cada iteração.

5.3 Discussões gerais

O método proposto neste trabalho para a detecção de anomalias pode ser aplicado a outro conjunto de dados que contenham KPIs ou mesmo outros atributos na forma de séries temporais, sem a necessidade de adaptação, desde que apresentem a mesma periodicidade e desvio padrão diferente de zero, pois as técnicas empregadas exploram a variância dos dados. Neste caso, os dados de entrada para as etapas de redução de dimensionalidade e agrupamento são numéricos, por conta da característica das técnicas usadas. Os valores categóricos de "hora" são usados somente na estratificação do conjunto de dados.

Apesar dos conjuntos de dados apresentados conterem no máximo dezesseis KPIs, o método proposto pode ser aplicado em conjunto de dados com mais KPIs. A etapa de redução de dimensionalidade com a técnica PCA foi adicionada com o propósito de reduzir o número de dimensões e evidenciar as relações entre os KPIs com base na variância dos KPIs; neste caso, o seu uso foi essencial. O uso de um conjunto de dados com seis KPIs foi considerado pelos possíveis comportamentos anômalos encontrados e para se produzir resultados visuais para análise do avaliador, contudo outras técnicas de redução de dimensionalidade podem ser testadas.

A categorização das instâncias foi realizada com base na distinção de regiões de alta e baixa densidades, de forma que as instâncias que se concentraram em uma região de alta densidade foram categorizadas como apresentando um comportamento normal e as instâncias que se localizam em uma região de baixa densidade ou de forma dispersa, conforme a parametrização adotada, foram categorizadas como anomalias. Neste caso, a capacidade e robustez na classificação dos pontos dispersos e de lidar com *clusters* de diferentes formas e tamanhos foram essenciais para a identificação das anomalias. Dessa forma, foi considerada a técnica DBSCAN para a etapa de detecção.

Um aspecto importante na aplicação do método proposto é que, caso um grande número de células da rede móvel apresente um comportamento de degradação, que será

refletido nos KPIs, a categorização das instâncias com base na densidade das regiões deve ser revista, pois, por exemplo, a região de alta densidade pode representar um comportamento anormal. Além disso, a inserção de anomalias sintéticas de mesma amplitude para análise da assertividade do método pode criar uma nova região na dispersão dos dados, que está sujeita a ser categorizada como normal de acordo com a parametrização adotada. Uma forma de mitigar este problema é selecionar um conjunto de dados com um grande histórico.

As análises e os testes realizados com os métodos proposto e de referência mostraram que ambos apresentaram um bom desempenho na detecção das anomalias do conjunto de dados original, segundo as avaliações dos profissionais e a assertividade na detecção das anomalias sintéticas inseridas. Apesar do método de referência apresentar um desempenho superior em alguns cenários, a sua aplicação em outros conjuntos de dados depende de reajuste dos parâmetros, como o tamanho da rede SOM e o número de iterações. Além disso, necessita da definição de um limiar para a pontuação de anormalidade que aponte se a instância é anômala ou não.

A avaliação das células anômalas selecionadas foi realizada com base no conhecimento técnico dos profissionais e nas ferramentas disponíveis. Neste caso, devido ao grande número de demandas diárias dos profissionais, não foi possível verificar se realmente o comportamento anômalo apontado se confirmou na prática em todas as células. Sendo assim, não foi possível realizar uma avaliação da taxa de falsos positivos e falsos negativos. Um estudo futuro com "células de referência" e classificação supervisionada permitirão este tipo de avaliação.

A análise de correlação corrobora a ideia de que o método proposto não é igual ao de referência. Ressalta-se que não foi intenção desta pesquisa produzir método igual ou superior, uma vez que a avaliação final de células ainda deve ser feita por mão de obra especializada. A ideia aqui proposta foi a de desenvolver um método **complementar** que ajude a (i) identificar casos que não puderam ser descobertos por outros métodos, (ii)

reforçar indício de anomalia ou (iii) avaliar cenários específicos onde métodos convencionais podem falhar. De qualquer forma, o propósito foi dar ao usuário do sistema, em meio a tantos números (base de dados de KPIs), métodos para que este possa se orientar melhor e identificar padrões nem sempre visíveis a uma análise imediata.

A formulação do problema e a escolha das técnicas apropriadas para detecção das anomalias é determinada por vários fatores, como a natureza dos dados de entrada e a disponibilidade de rótulos. Apesar dos KPIs estarem na forma de séries temporais, o uso de técnicas desenvolvidas para análise deste tipo de dado não foi considerado, visto que as técnicas são aplicadas, geralmente, em um KPI por vez. A abordagem proposta nesta pesquisa teve o intuito de analisar os KPIs nos vários horários das séries temporais de forma multivariada, que possibilitasse evidenciar e identificar comportamentos anormais entre os KPIs.

5.4 Resumo e discussão geral do capítulo

Neste capítulo foram apresentados os resultados obtidos com os experimentos e testes das técnicas de redução de dimensionalidade e agrupamento para a definição da estrutura do método proposto para a detecção das anomalias. Os resultados obtidos com o conjunto de dados selecionado mostraram que o método proposto é capaz de identificar com frequência um comportamento anormal com base nos KPIs da rede móvel, inclusive quando este comportamento é inserido como uma forma de simular alguns cenários de degradação comumente encontrados num ambiente operacional.

Um segundo método, baseado na estrutura do método proposto, mas com técnicas diferentes, foi apresentado, e pôde ser visto como um procedimento de referência para a comparação do método proposto, com a finalidade de validar os resultados encontrados. O método de referência foi capaz de identificar as anomalias sintéticas inseridas, entretanto algumas limitações quanto a parametrização e a definição de um limiar das técnicas dificultam o seu uso com outros conjuntos de dados. Neste sentido foi destacado

que a proposta aqui não é de concorrência entre eficiência de métodos, mas, sim, de complementação (e validação, no caso).

6 Conclusão e trabalhos futuros

6.1 Conclusão

Este trabalho propôs um método de detecção de anomalias baseado em PCA e *clustering* que estratifica o conjunto de dados em períodos temporais e tem como objetivo fornecer subsídios de análise para correção de eventuais falhas, otimização de recursos, planejamento e expansão de capacidade da rede móvel.

Os cenários de testes e validações usaram um conjunto de KPIs de uma rede 4G em operação, relacionados à qualidade do sinal e a utilização da rede, aplicado a 1054 células para identificar quais tinham comportamentos potencialmente anômalos. As vinte células que apresentaram o maior número de anomalias detectadas foram analisadas e os resultados encontrados foram usados com sucesso na correção de falhas, otimização de recursos e expansão de capacidade da rede pela operadora nestas mesmas vinte células com base no método/ferramenta aqui proposto.

Com o intuito de reforçar e complementar as validações, um segundo método foi implementado com base em um conhecido índice na área. Este procedimento pôde ser visto como um método de referência para a comparação do método proposto e foi capaz de identificar as anomalias sintéticas inseridas. Entretanto algumas limitações quanto a parametrização e a definição de um limiar das técnicas dificultam o seu uso com outros conjuntos de dados.

Apesar de nos cenários de testes e validações se considerar seis KPIs, o método proposto pode ser aplicado em qualquer conjunto de KPIs na forma de séries temporais, independentemente da quantidade, desde que apresentem uma variância não nula. Os resultados devem mostrar o que pode ser considerado como comportamento normal e anormal, com base na proximidade matemática dos dados. Entre as vantagens do

método, pode-se citar sua capacidade de avaliar simultaneamente vários KPIs e poder ser executado periodicamente para validação das tratativas das falhas encontradas e detecção de novos comportamentos destoantes da média populacional avaliada.

6.2 Trabalhos futuros

Os métodos apresentados foram desenvolvidos com o intuito de fornecer subsídios de análise para correção de eventuais falhas, otimização de recursos, planejamento e expansão de capacidade da rede móvel. Neste sentido, alguns trabalhos futuros podem ser desenvolvidos:

- Adicionar uma etapa supervisionada no método proposto, que, com fundamento na classificação e validação prática das anomalias identificadas nos elementos de rede pelos profissionais, possa calcular a probabilidade ou apontar a provável causa de comportamento anormal nos elementos de rede;
- Desenvolver um método para previsão de falhas e limitações de capacidade da rede, com apoio da classificação e validação prática das anomalias;
- Investigar os horários em que ocorreram as anomalias e correlacionar com as avaliações dos profissionais, de forma a determinar se alguma das causas do comportamento anormal seguem algum padrão com relação aos horários;
- Desenvolver um método para diferenciar as células localizadas em regiões urbanas densas, urbanas de borda, rurais e de rodovias, com base na dispersão dos dados após a etapa de redução de dimensionalidade.

Referências

[1] 3GPP, "TS 32.425 V15.1.0, Performance Measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 15)," 2018.

[2] V. Chandola, A. Banerjee e V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys,* vol. 41, nº 3, pp. 1-58, 2009.
https://doi.org/10.1145/1541880.1541882

[3] A. Asghar, H. Farooq e A. Imran, "Self-Healing in Emerging Cellular Networks: Review, Challenges, and Research Directions," *IEEE Communications Surveys and Tutorials,* vol. 20, nº 3, pp. 1682-1709, 2018.
https://doi.org/10.1109/COMST.2018.2825786

[4] P. V. Klaine, M. A. Imran, O. Onireti e R. D. Souza, "A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks," *IEEE Communications Surveys & Tutorials,* vol. 19, nº 4, pp. 2392 - 2431, 2017.
https://doi.org/10.1109/COMST.2017.2727878

[5] M. S. e. al., "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE Journal on Selected Areas in Communications,* vol. 35, nº 6, pp. 1201-1221, 2017.
https://doi.org/10.1109/JSAC.2017.2692307

[6] 3GPP, "3rd Generation Partnership Program," [Online]. Available: https://www.3gpp.org/.

[7] M. Wang e S. Handurukande, "Anomaly Detection for Mobile Network Management," *International Journal of Next-Generation Computing,* vol. 9, nº 2, pp. 80-97, 2018.

[8] M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi e J. M. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey," *Computer Networks,* vol. 132, pp. 180-199, 2018.
https://doi.org/10.1016/j.comnet.2018.01.016

[9] O. G. Aliu, A. Imran, M. A. Imran e B. Evans, "A Survey of Self Organisation in Future Cellular Networks," *IEEE Communications Surveys & Tutorials,* vol. 15, nº 1, pp. 336-361, 2013.

https://doi.org/10.1109/SURV.2012.021312.00116

[10] M. Ahmed, A. N. Mahmood e J. Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications,* vol. 60, p. 19–31, 2016.
https://doi.org/10.1016/j.jnca.2015.11.016

[11] A. K. Jain, R. P. Duin e J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 22, nº 1, pp. 4-36, 2000.
https://doi.org/10.1109/34.824819

[12] I. T. Jolliffe, Principal Component Analysis, New York, USA: Springer, 2002.

[13] P. Comon, "Independent Component Analysis, a New Concept?," *Signal Processing,* vol. 36, nº 3, pp. 287-314, 1994.
https://doi.org/10.1016/0165-1684(94)90029-9

[14] T. Kohonen, "The Self-organizing Map," *Proceedingd of the IEEE,* vol. 78, nº 9, pp. 1464-1480, 1990.
https://doi.org/10.1109/5.58325

[15] P. Pudil, J. Novovicova e J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters,* vol. 15, nº 11, pp. 1.119-1.125, 1994.
https://doi.org/10.1016/0167-8655(94)90127-9

[16] J. Shlens, "A Tutorial on Principal Component Analysis," *Cornell University,* pp. 1-12, 2014.

[17] Y. Xiaohu, C. Zhang, X. Tan, S. Jin e H. Wu, "AI for 5G: Research Directions and Paradigms," *Science China Information Sciences 2,* 2018.

[18] Y. Linde, A. Buzo e R. Gary, "An algorithm for vector quantization design," *IEEE Tansaction on Communications,* vol. 28, nº 1, pp. 84-94, 1980.
https://doi.org/10.1109/TCOM.1980.1094577

[19] J. A. Hartigan e M. A. Wong, "A K-means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 28, nº 1, pp. 100-1008, 1979.
https://doi.org/10.2307/2346830

[20] L. Kaufman e P. Rousseeuw, "Clustering by means of Medoids, in Statistical Data Analysis Based on the L1–Norm and Related Methods," *Journal of computational and applied mathematics,* vol. 20, pp. 53-65, 1987.

[21] P. H. A. Sneath, R. R. Sokal e W. H. Freeman, "Numerical Taxonomy. The Principles and Practice of Numerical Classification," *Systematic Zoology,* vol. 24, nº 2, pp. 263-268, 1975.
https://doi.org/10.2307/2412767

[22] B. King, "Step-Wise Clustering Procedures," *Journal of the American Statistical Association,* vol. 62, nº 317, pp. 86-101, 1967.
https://doi.org/10.1080/01621459.1967.10482890

[23] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Philip Drive Assinippi Park Norwell, USA: Kluwer Academic Publishers, 1981.
https://doi.org/10.1007/978-1-4757-0450-1

[24] M. Ester, H.-P. Kriegel, J. Sander e X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the KDD,* vol. 96, p. 226–231, 1996.

[25] M. Ankerst, M. M. Breunig, H.-P. Kriegel e J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," *ACM SIGMOD Record,* vol. 28, nº 2, pp. 49 -60, 1999.
https://doi.org/10.1145/304181.304187

[26] J. D. Banfield e A. E. Raftery, "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics,* vol. 49, nº 3, pp. 803-821, 1993.
https://doi.org/10.2307/2532201

[27] S.-Y. Lu e K. S. Fu, "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 8, nº 5, pp. 381-389, 1978.
https://doi.org/10.1109/TSMC.1978.4309979

[28] V. J. Hodge e J. Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review: Kluwer Academic Publishers,* vol. 22, p. 85–126, 2004.
https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

[29] H.-P. Kriegel, P. Kröger, E. Schubert e A. Zimek, "LoOP: Local Outlier Probabilities," *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* p. 1649–1652, 2009.

https://doi.org/10.1145/1645953.1646195

[30] M. M. Breunig, H.-P. Kriegel, R. T. Ng e J. Sander, "LOF: Identifying Density-based Local Outliers," *ACM SIGMOD 2000 Int. Conf. On Management of Data,* vol. 29, nº 2, pp. 93-104, 2000.
https://doi.org/10.1145/335191.335388

[31] S. Papadimitriou, H. Kitagawa, P. Gibbons e C. Faloutsos, "LOCI: Fast Outlier Detection Using the Local Correlation Integral," em *Proceedings 19th International Conference on Data Engineering*, Bangalore, India, 2003.
https://doi.org/10.1109/ICDE.2003.1260802

[32] T. A. Runkler, Data Analytics: Models and Algorithms for Intelligent Data Analysis, Wiesbaden: New York: Springer Vieweg, 2012.
https://doi.org/10.1007/978-3-8348-2589-6

[33] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao e R. C. Qiu, "Big Data Analytics in Mobile Cellular Networks," *IEEE Access,* vol. 4, pp. 1985-1996, 2016.
https://doi.org/10.1109/ACCESS.2016.2540520

[34] K. Yang, R. Liu, Y. Sun, J. Yang e X. Chen, "Deep Network Analyzer (DNA): A Big Data Analytics Platform for Cellular Networks," *IEEE Internet of Things Journal,* vol. 4, nº 6, pp. 2019-2027, 2017.
https://doi.org/10.1109/JIOT.2016.2624761

[35] U. S. Hashmi, A. Darbandi e A. Imran, "Enabling Proactive Self-Healing by Data Mining Network Failure Logs," em *2017 International Conference on Computing, Networking and Communications (ICNC)*, Santa Clara, CA, USA, 2017.
https://doi.org/10.1109/ICCNC.2017.7876181

[36] C. V. Murudkar e R. D. Gitlin, "Machine Learning for QoE Prediction and Anomaly Detection in Self-Organizing Mobile Networking Systems," *International Journal of Wireless & Mobile Networks (IJWMN),* vol. 11, nº 2, pp. 1-12, 2019.
https://doi.org/10.5121/ijwmn.2019.11201

[37] P. Fiadino, A. D'Alconzo, M. Schiavone e P. Casas, "Challenging Entropy-based Anomaly Detection and Diagnosis in Cellular Networks," *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication,* p. 87–88, 2015.
https://doi.org/10.1145/2829988.2790011

[38] P. Muñoz, R. Barco, I. Serrano e A. Gómez-Andrades, "Correlation-Based Time-Series Analysis for Cell Degradation Detection in SON," *IEEE Communications Letters,* vol. 20, nº 2, pp. 396-399, 2016.

https://doi.org/10.1109/LCOMM.2016.2516004

[39] S. Chernov, M. Cochez e T. Ristaniemi, "Anomaly Detection Algorithms for the Sleeping Cell Detection in LTE Networks," *2015 IEEE 81st Vehicular Technology Conference (VTC Spring),* pp. 1-5, 2015. https://doi.org/10.1109/VTCSpring.2015.7145707

[40] L. Bodrog, M. Kajo, S. Kocsis e B. Schultz, "A Robust Algorithm for Anomaly Detection in Mobile Networks," *2016 IEEE 27th Annual International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC),* pp. 1-6, 2016. https://doi.org/10.1109/PIMRC.2016.7794573

[41] G. Ciocarlie, U. Lindqvist, S. Novaczki e H. Sanneck, "Detecting Anomalies in Cellular Networks Using an Ensemble Method," *2013 9th International Conference on Network and Service Management (CNSM),* p. 171–174, 2013. https://doi.org/10.1109/CNSM.2013.6727831

[42] G. Ciocarlie, U. Lindqvist, K. Nitz, S. Novaczki e H. Sanneck, "On the Feasibility of Deploying Cell Anomaly Detection in Operational Cellular Networks," *Network Operations and Management Symposium (NOMS),* p. 1–6, 2014. https://doi.org/10.1109/NOMS.2014.6838305

[43] W. Sun, X. Qin, S. Tang e G. Wei, "A QoE Anomaly Detection and Diagnosis Framework for Cellular Network Operators," *The 1st International Workshop on Quality of Experience-Based Wireless Communications 2015,* pp. 450-455, 2015. https://doi.org/10.1109/INFCOMW.2015.7179426

[44] "RDocumentation," [Online]. Available: https://www.rdocumentation.org/. [Acesso em 15 Março 2020].

[45] M. Hahsler, M. Piekenbrock, S. Arya e D. Mount, "The Comprehensive R Archive Network," 22 Outubro 2019. [Online]. Available: https://cran.r-project.org/web/packages/dbscan/dbscan.pdf. [Acesso em 15 Março 2020].

[46] H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani e D. Dunnington, "The Comprehensive R Archive Network," 05 Março 2020. [Online]. Available: https://cran.r-project.org/web/packages/ggplot2/index.html. [Acesso em 15 Março 2020].

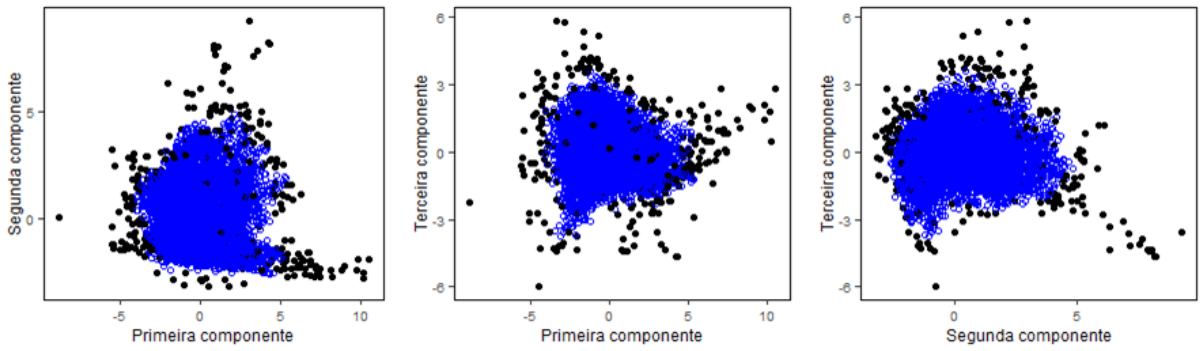
[47] R. Wehrens e J. Kruisselbrink, "The Comprehensive R Archive Network," 26 Novembro 2019. [Online]. Available: https://cran.r-project.org/web/packages/kohonen/kohonen.pdf. [Acesso em 19 Março 2020].

[48] J. H. Madsen, "The Comprehensive R Archive Network," 30 Maio 2018. [Online]. Available: https://cran.r-project.org/web/packages/DDoutlier/DDoutlier.pdf. [Acesso em 19 Março 2020].

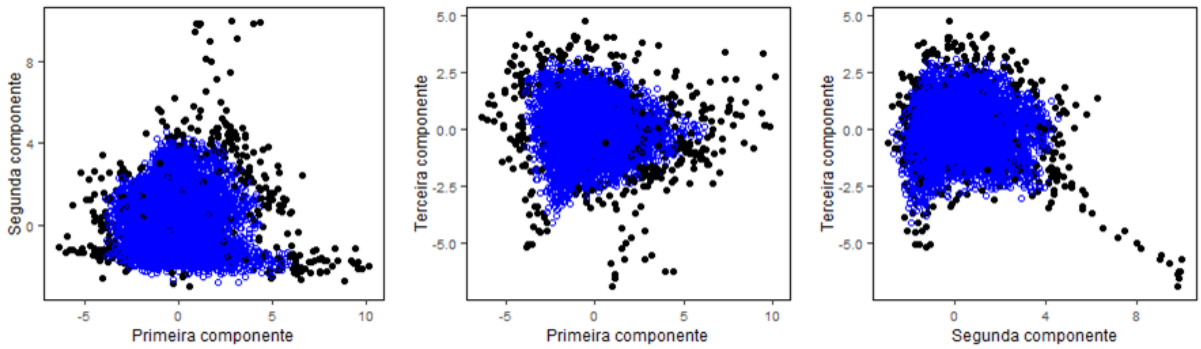
Apêndice A: Gráficos de dispersão dos resultados do método proposto

O método proposto estratifica o conjunto de dados de acordo com a periodicidade da série temporal dos atributos (no caso KPIs) numa etapa inicial e a seguir, são aplicadas as técnicas para detecção das anomalias em cada subconjunto. Os resultados gerados também seguem esta periodicidade e serão apresentados na forma de gráficos de dispersão para todos os períodos (no caso horas). As instâncias categorizadas como anomalias estão representadas pelo círculo preenchido preto, e as demais instâncias, categorizadas como comportamento normal pelo círculo não preenchido em azul ou outras cores (casos em que se foi criado mais de um *cluster*).

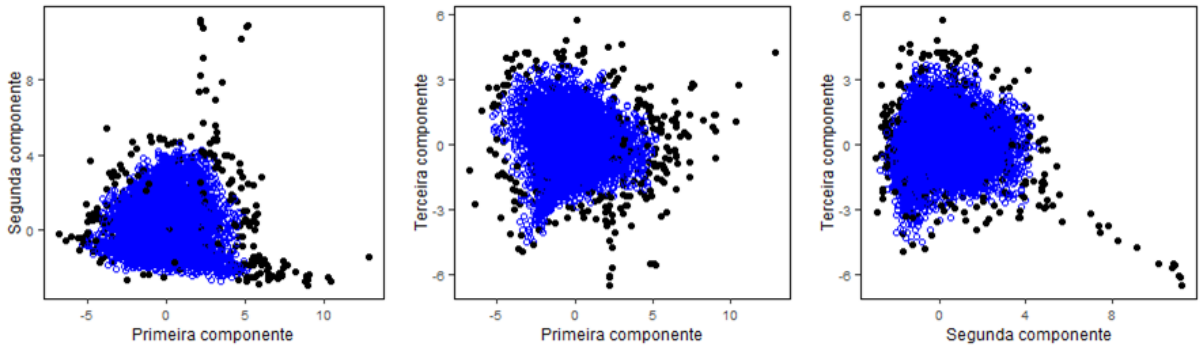
00:00h



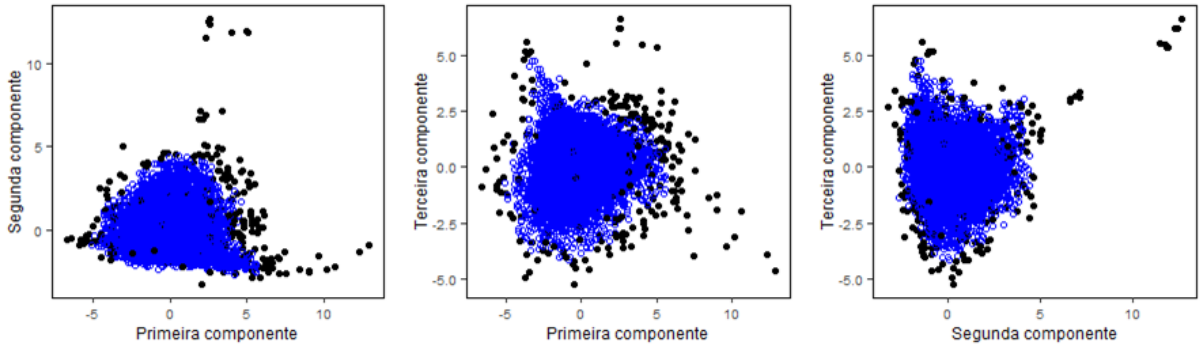
01:00h



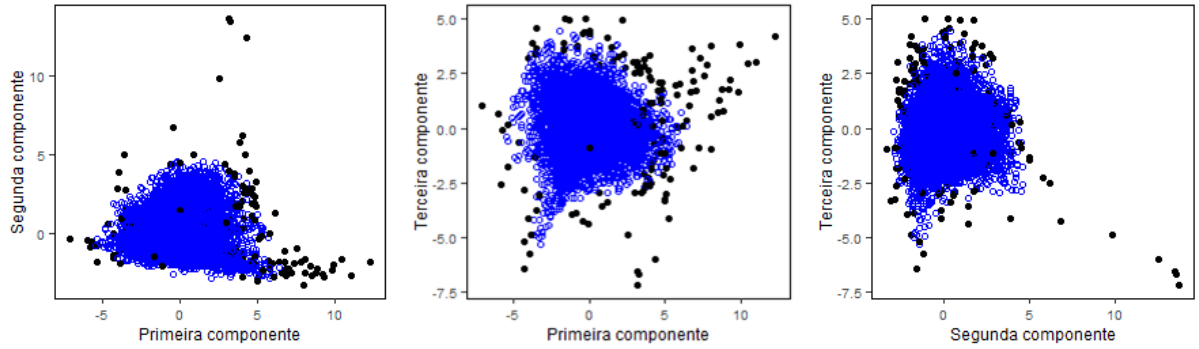
02:00h



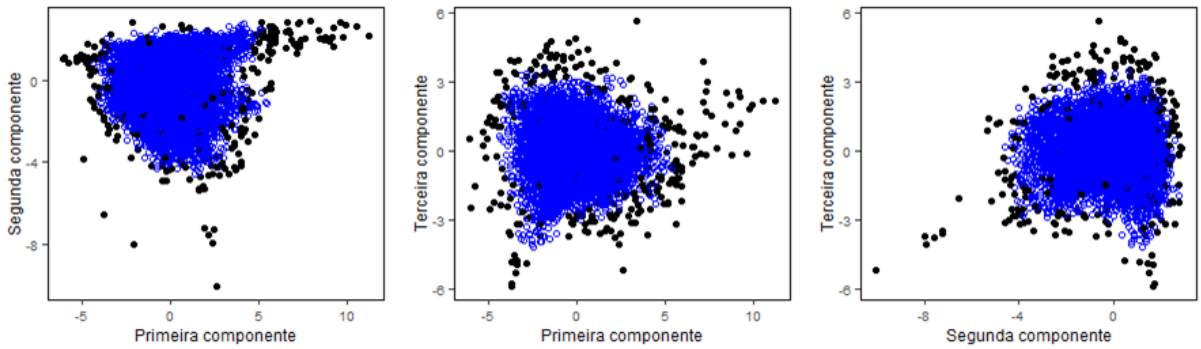
03:00h



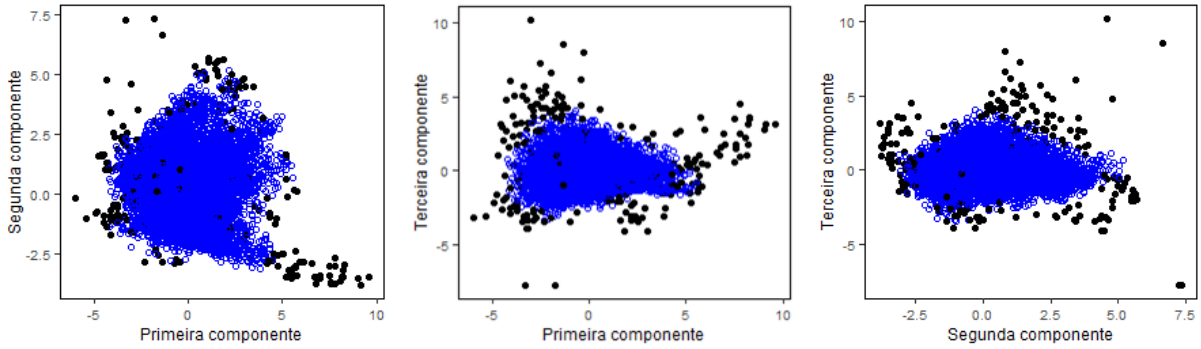
04:00h



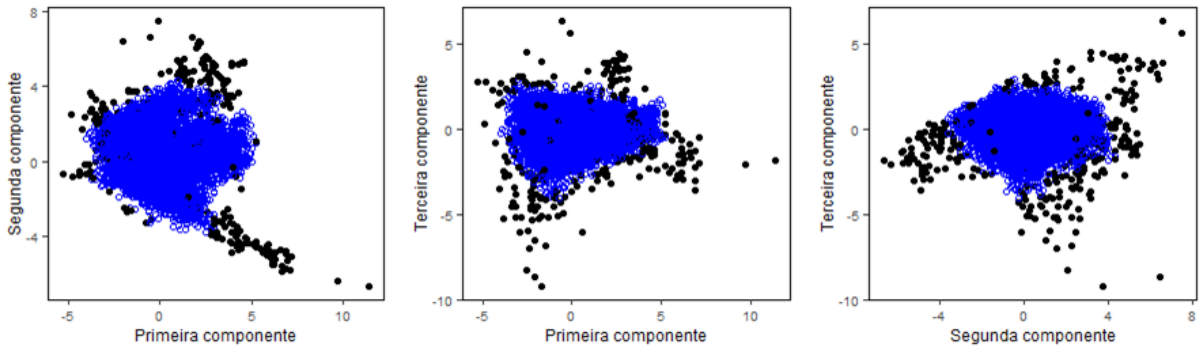
05:00h



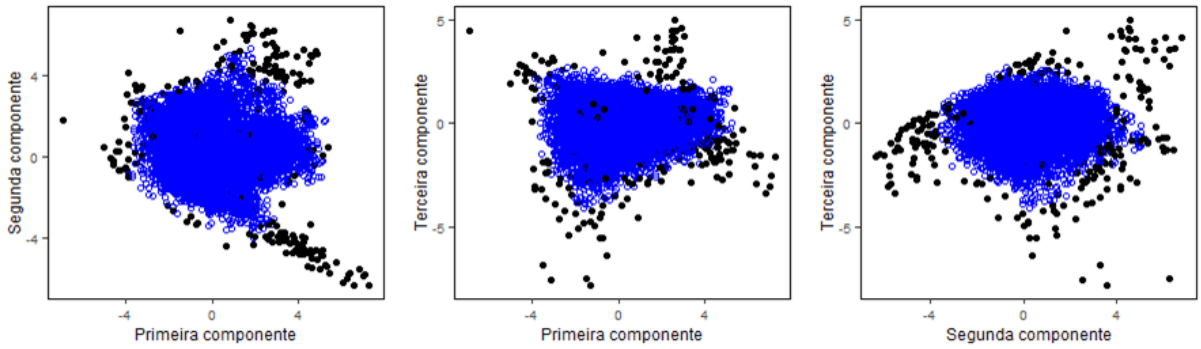
06:00h



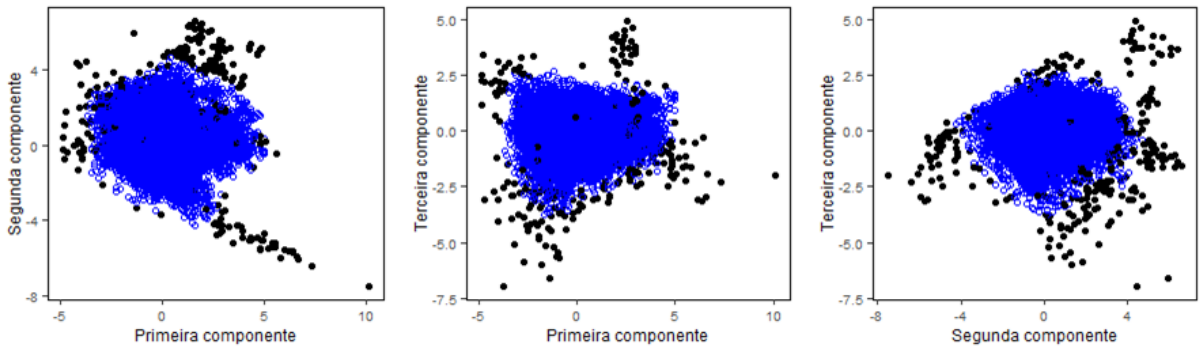
07:00h



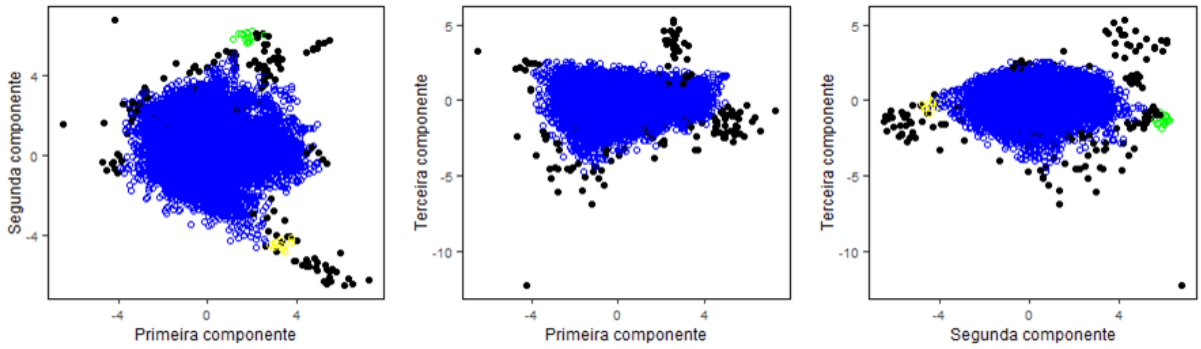
08:00h



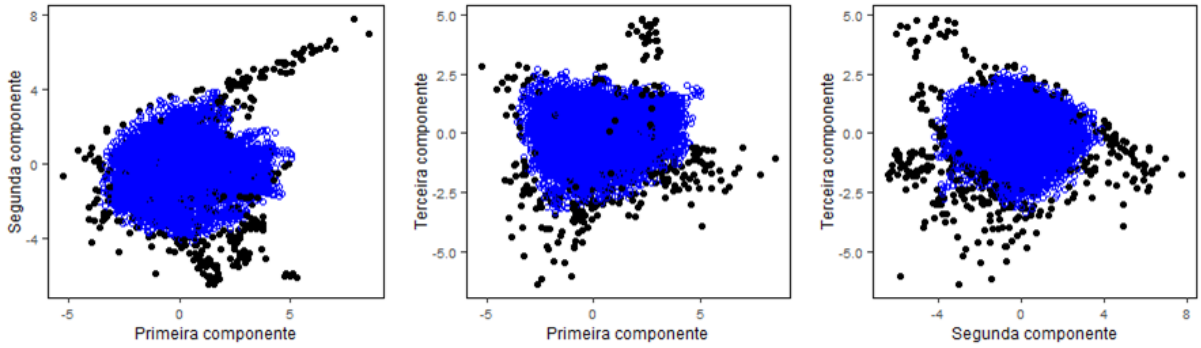
09:00h



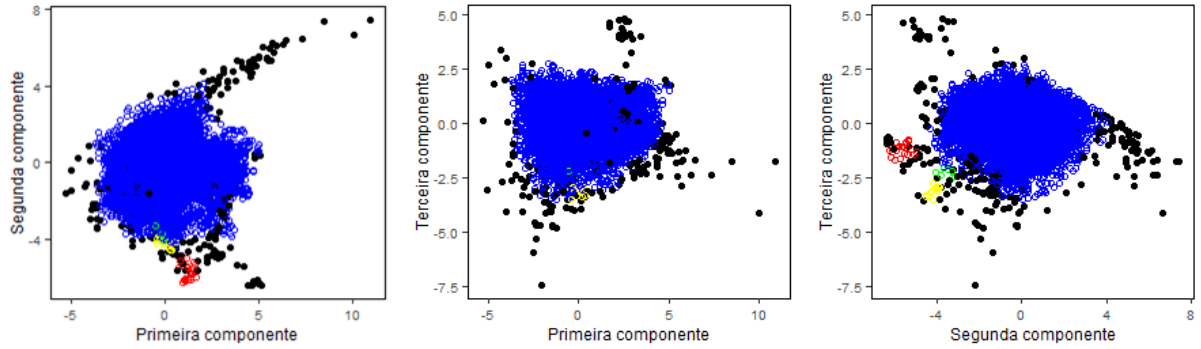
10:00h



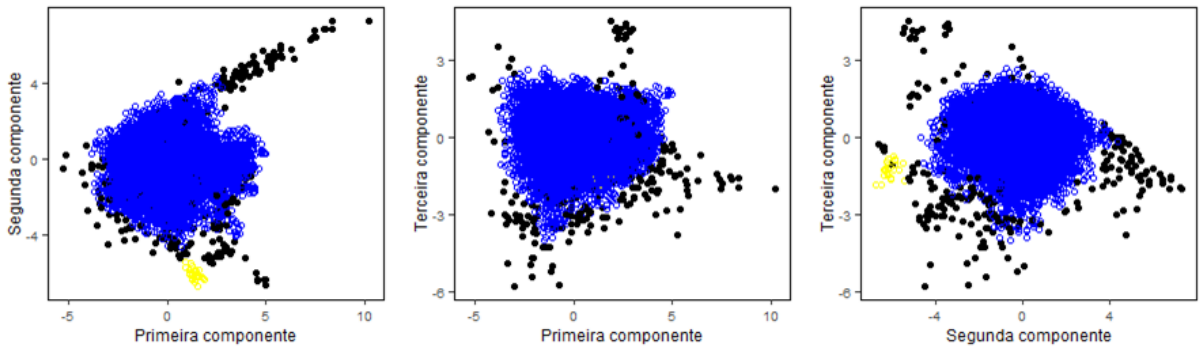
11:00h



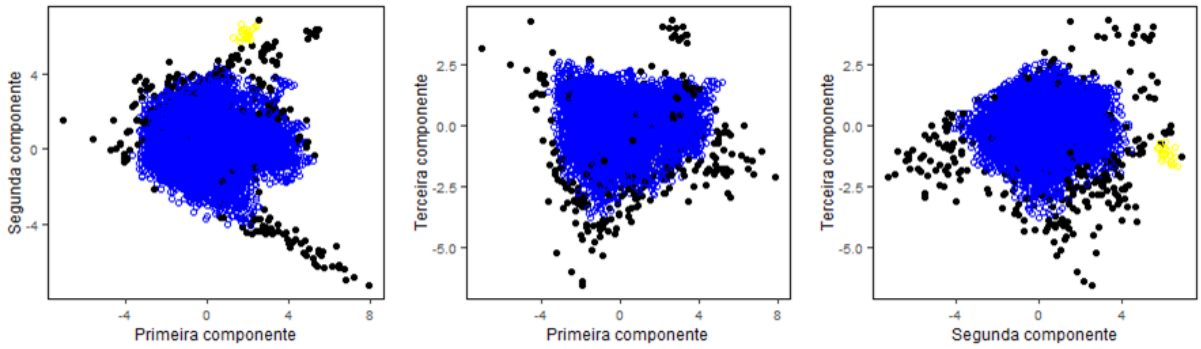
12:00h



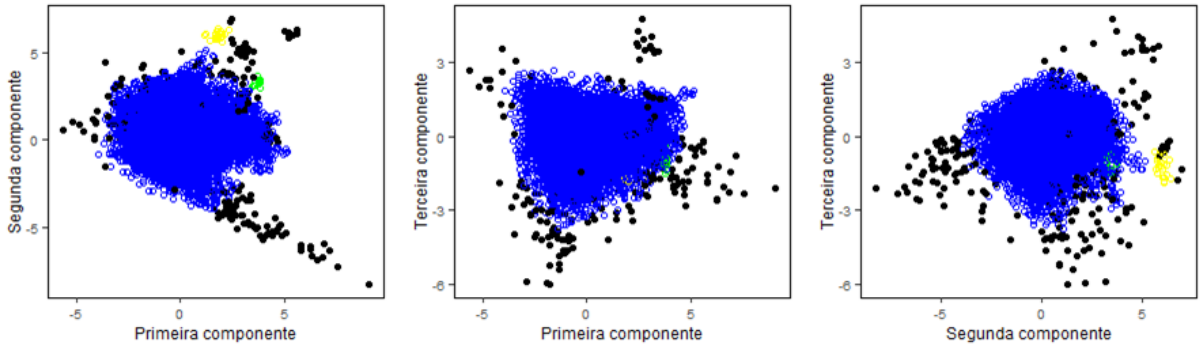
13:00h



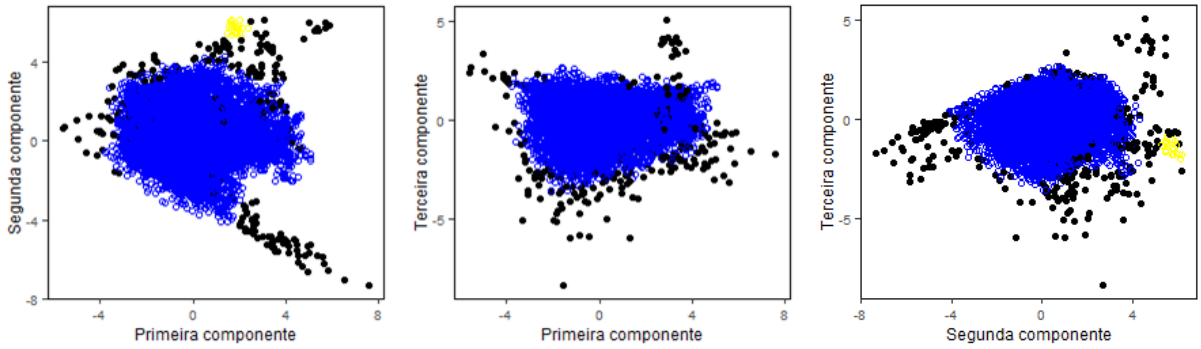
14:00h



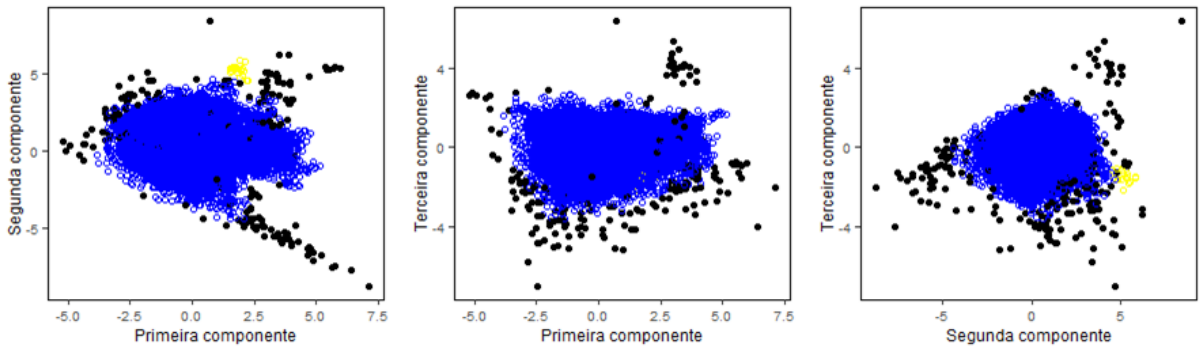
15:00h



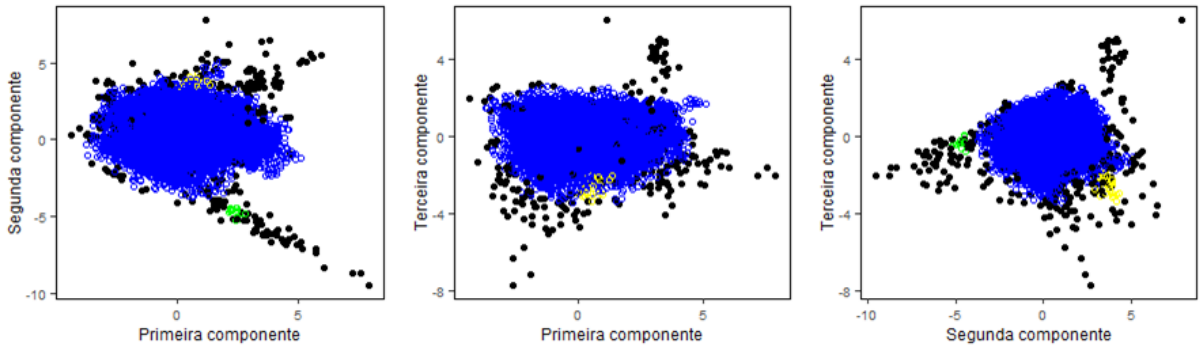
16:00h



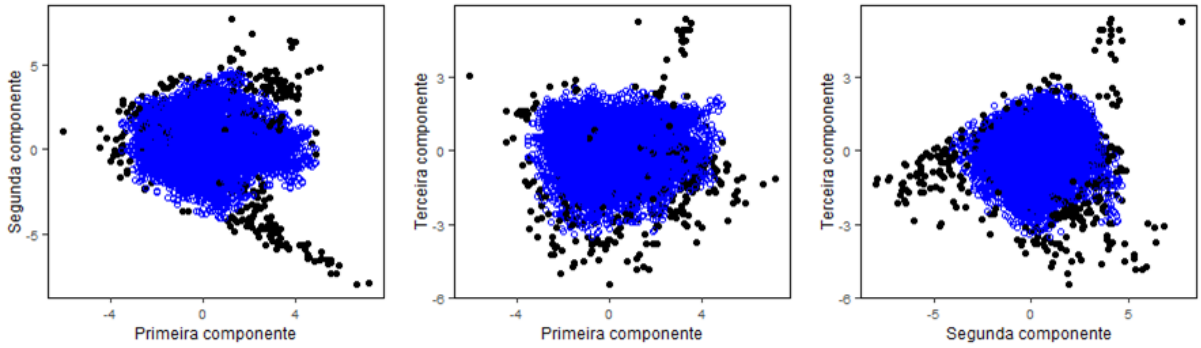
17:00h



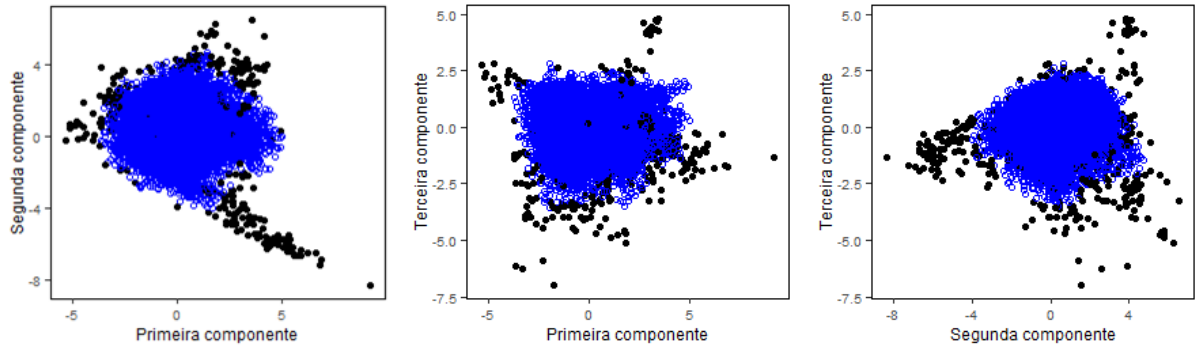
18:00h



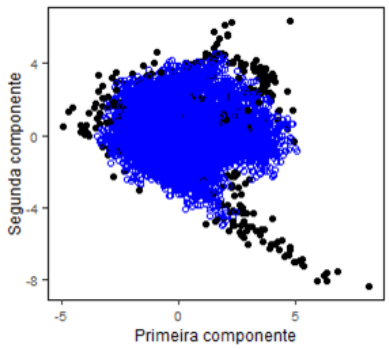
19:00h

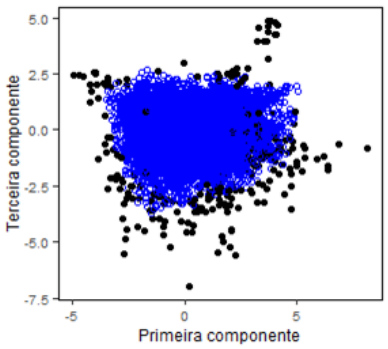


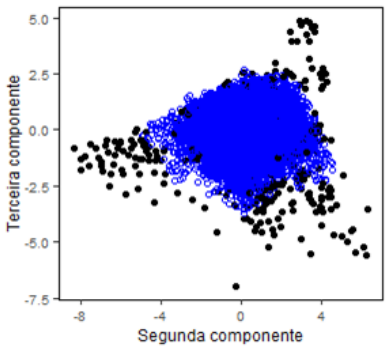
20:00h



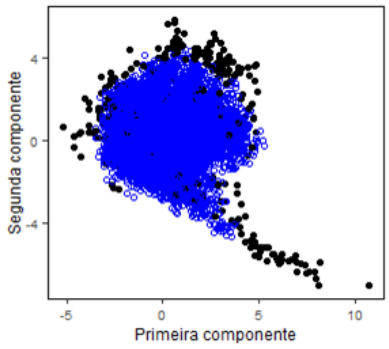
21:00h

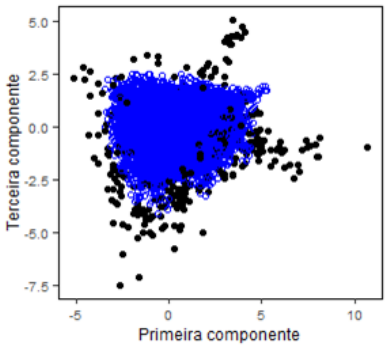


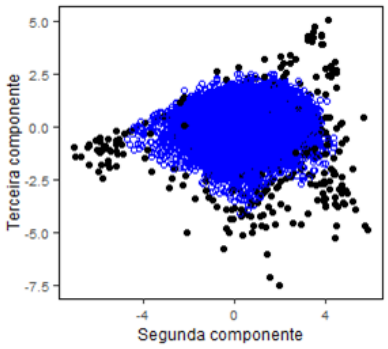




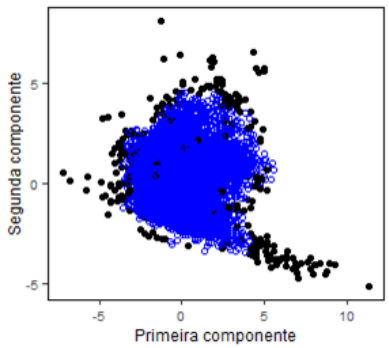
22:00h

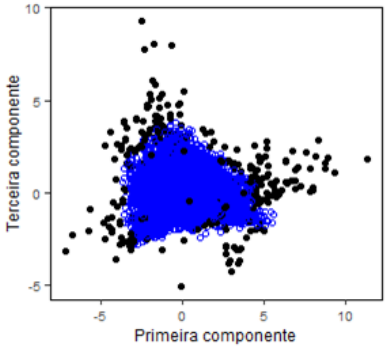


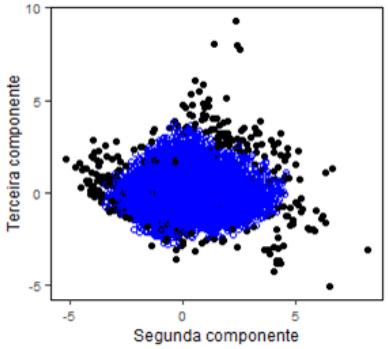




23:00h



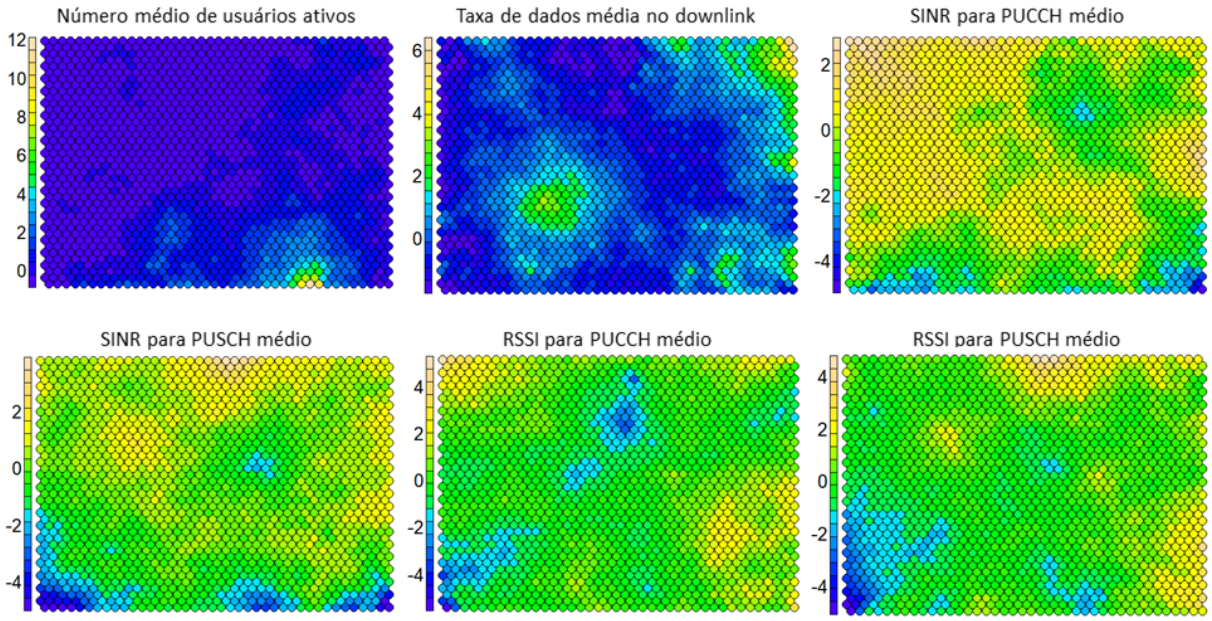




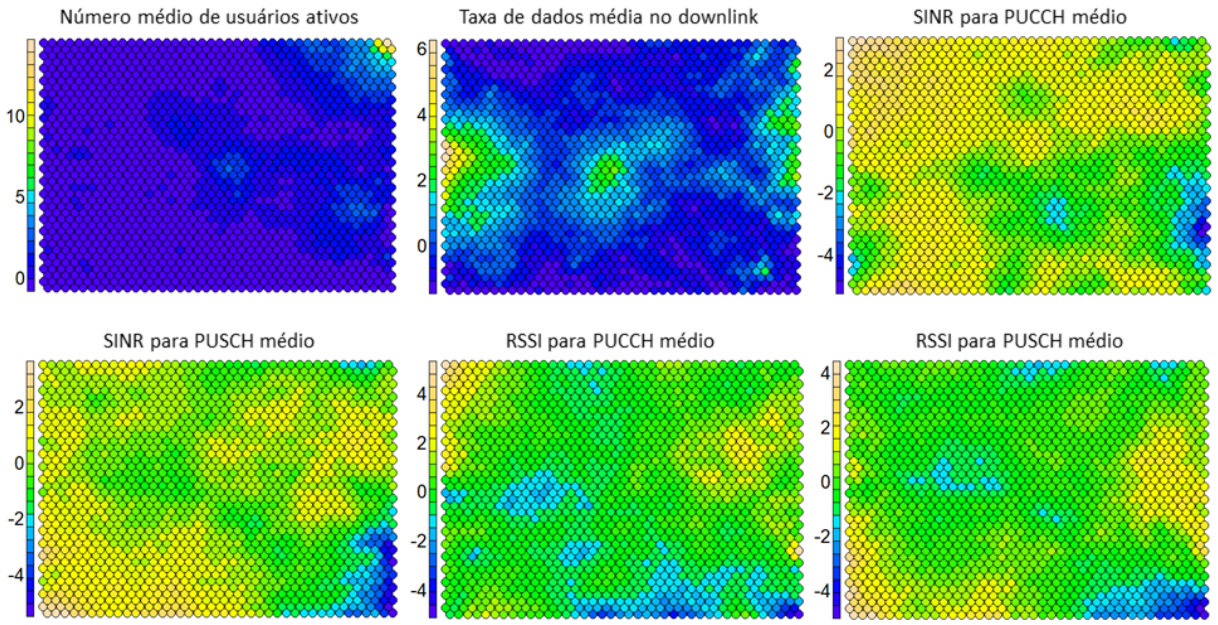
Apêndice B: Mapas de calor do codebook da rede SOM

O método de referência baseado em redes neurais e detecção de *outlier* local, assim como o método proposto, aplica as técnicas em subconjuntos conforme a periodicidades da série temporal dos atributos (no caso KPIs). Os resultados gerados também seguem esta periodicidade e os mapas de calor da etapa de redução de dimensionalidade com o SOM serão apresentados. Com base nestes gráficos é possível observar a mudança no comportamento dos KPIs ao longo dos diferentes horários.

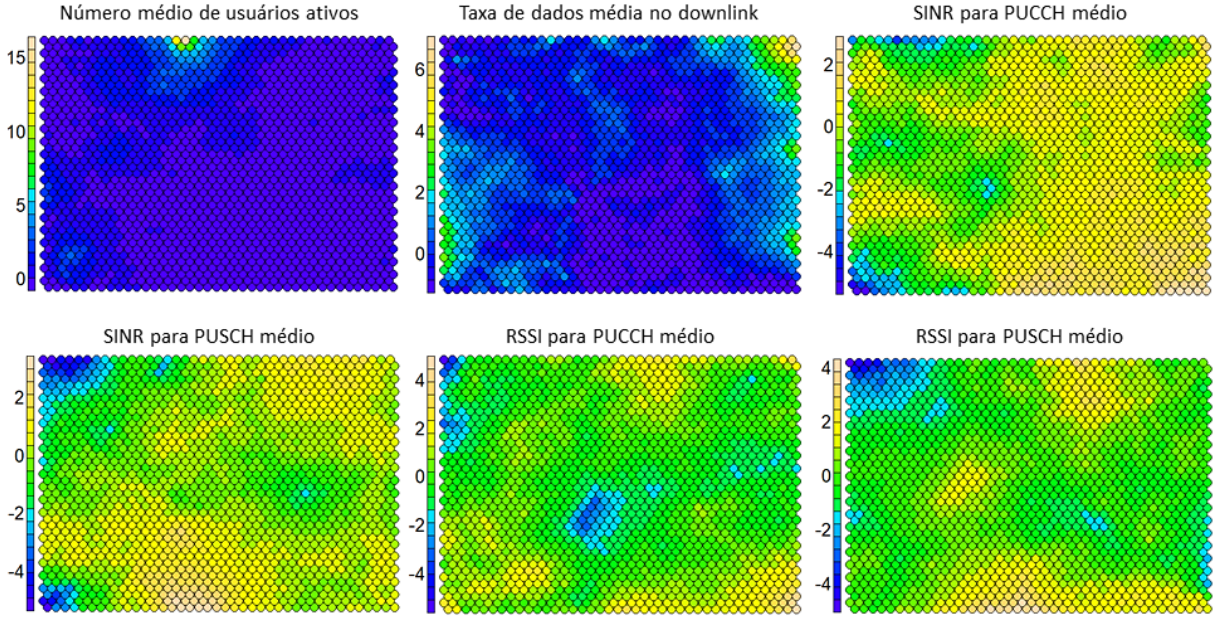
00:00h



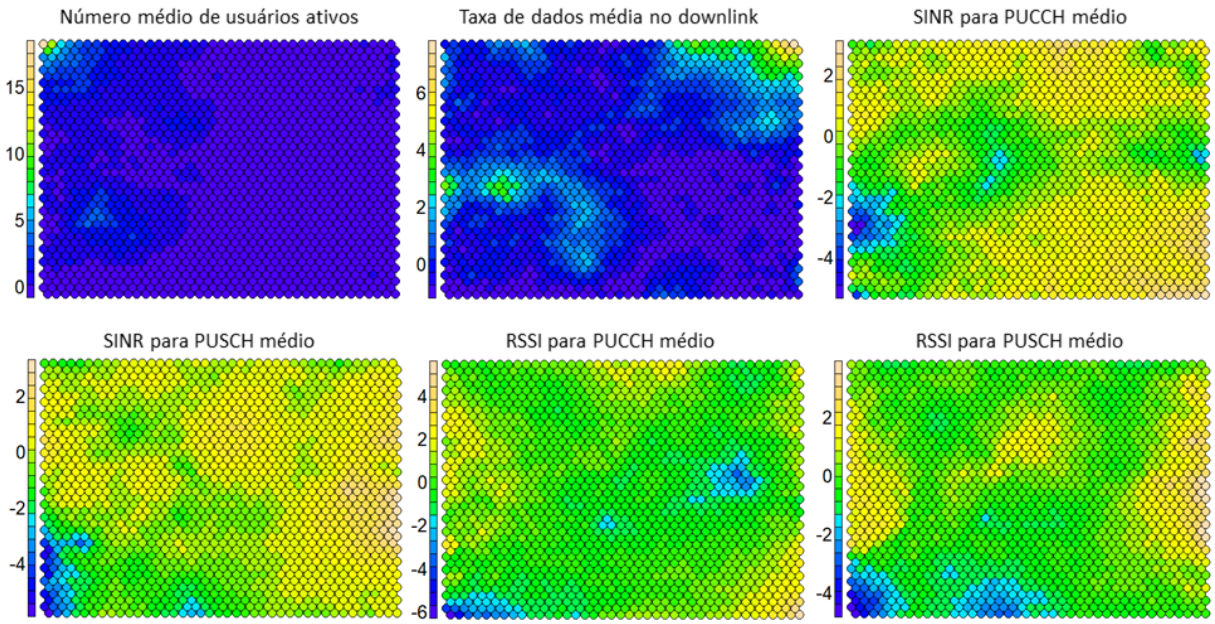
01:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

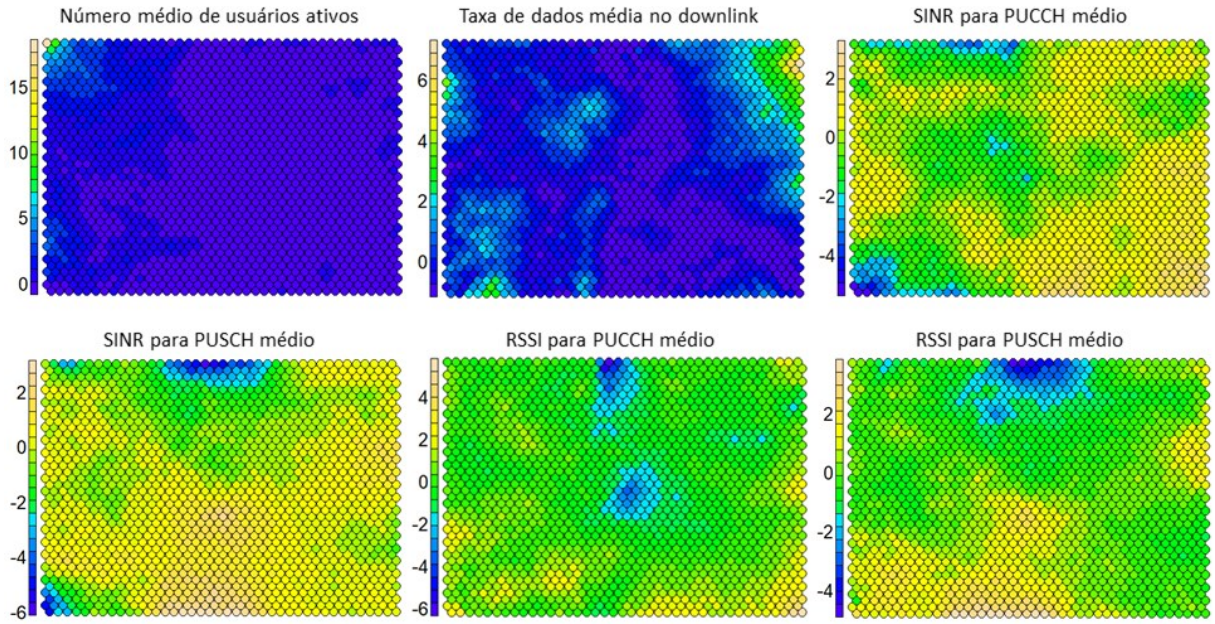
02:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

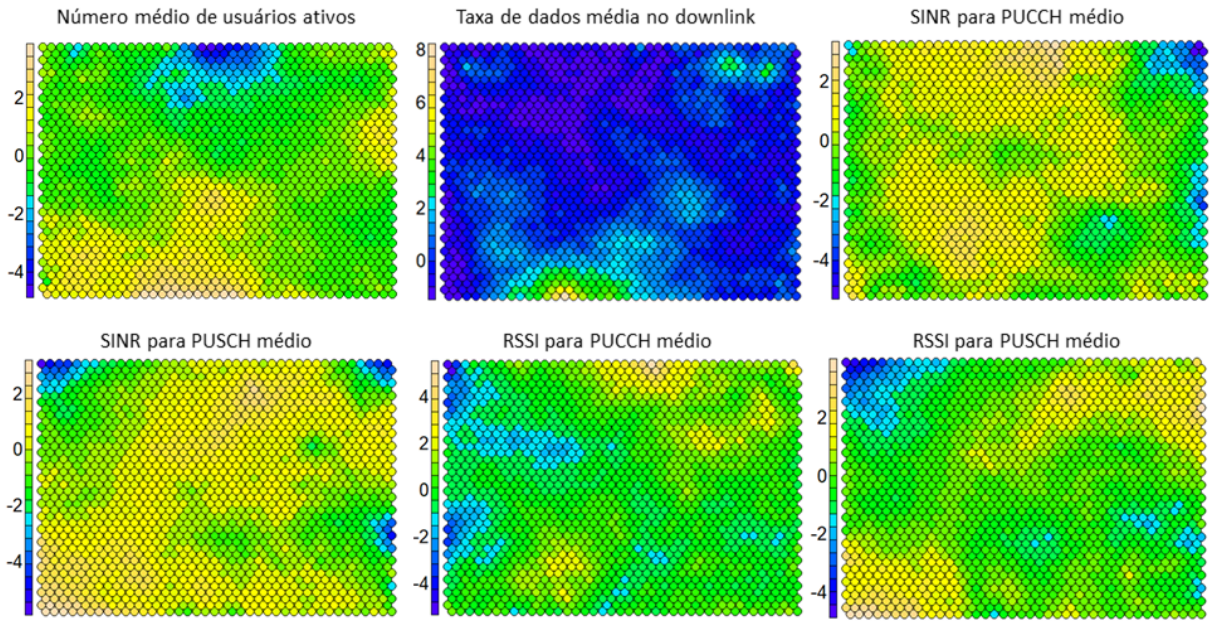
03:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

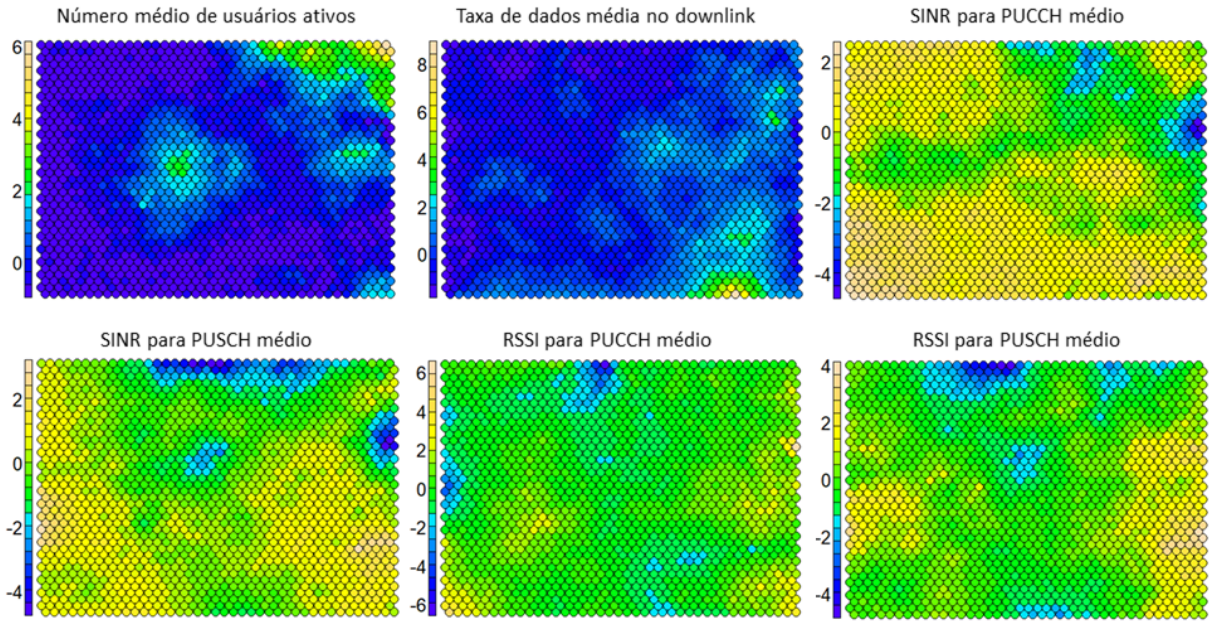
04:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

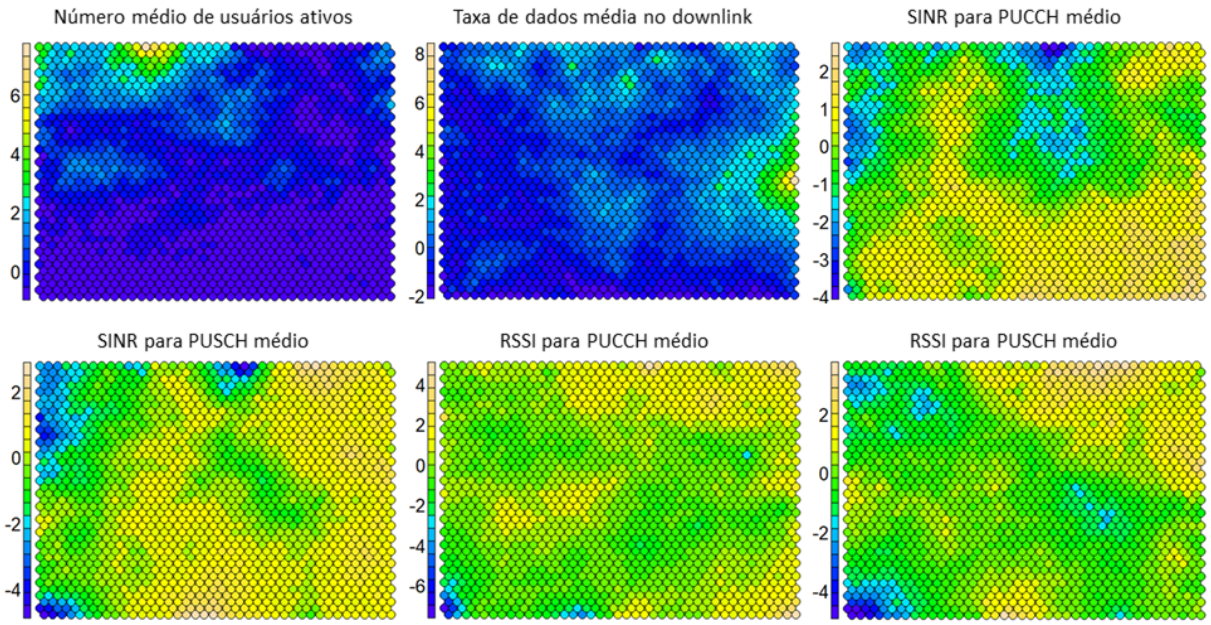
05:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

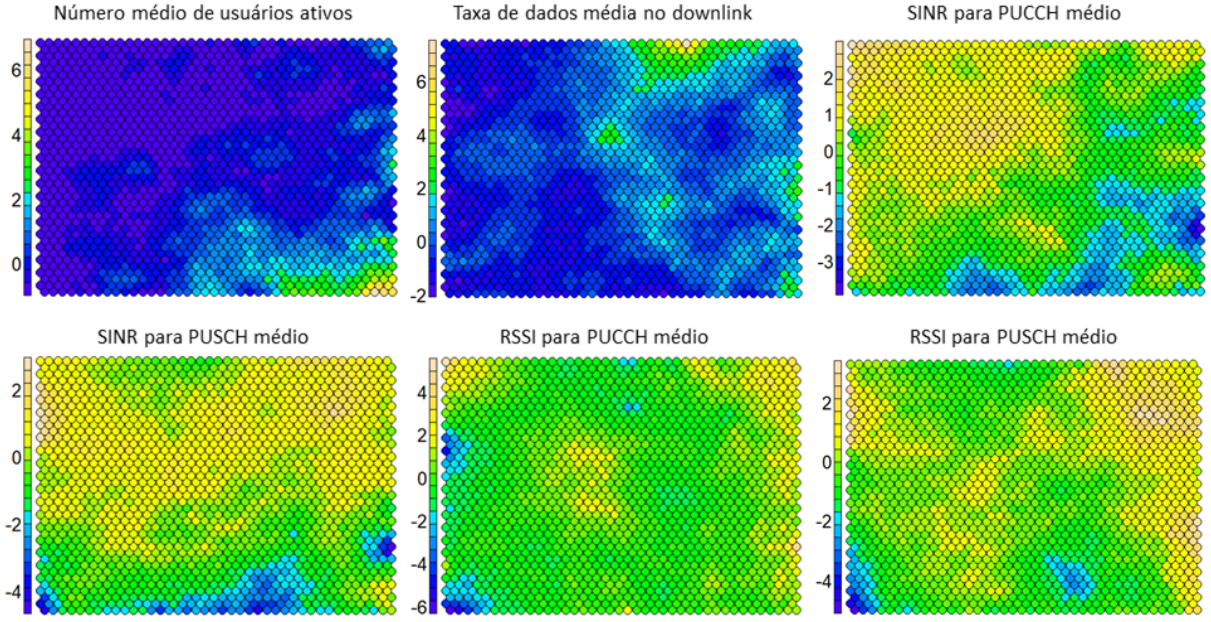
06:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

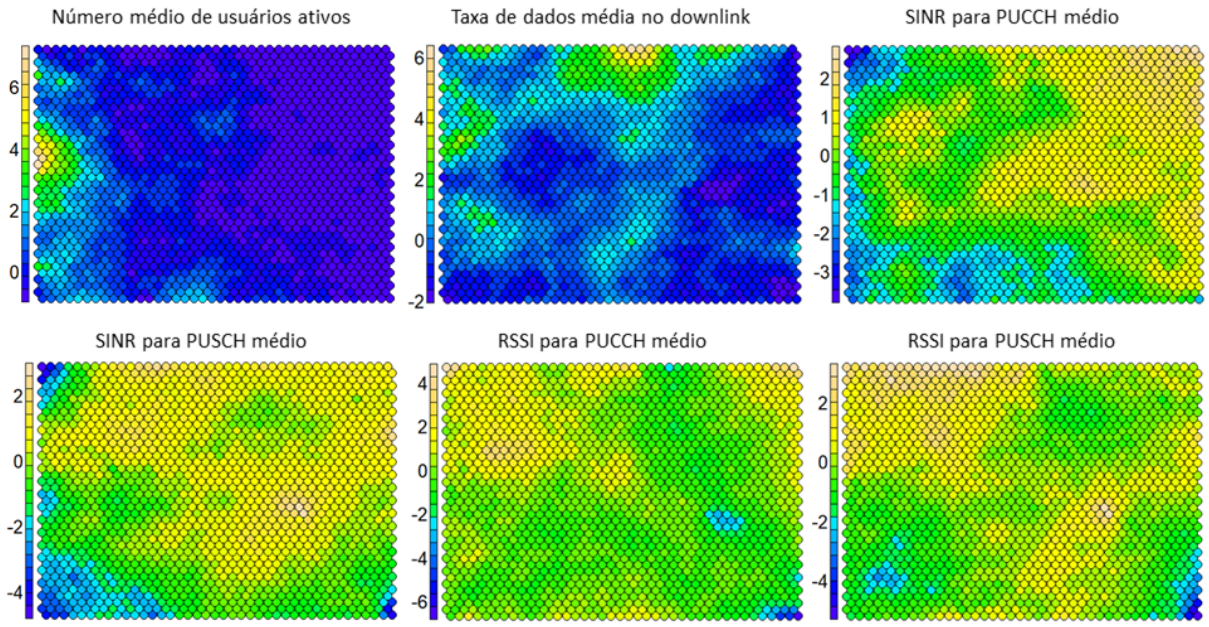
07:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

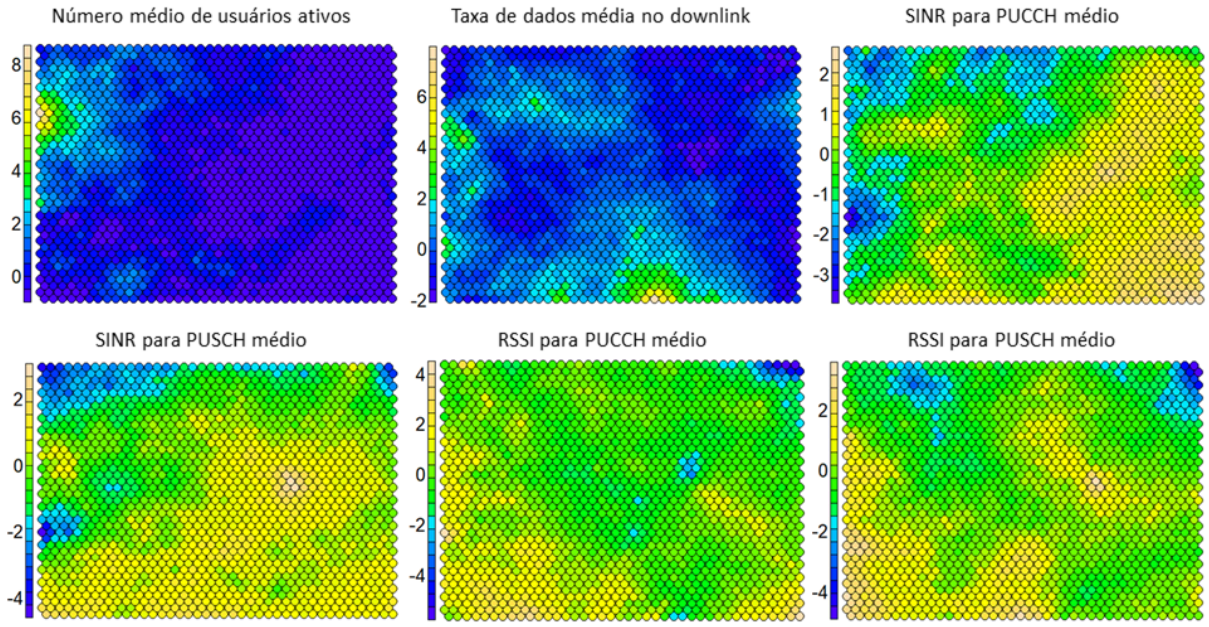
08:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

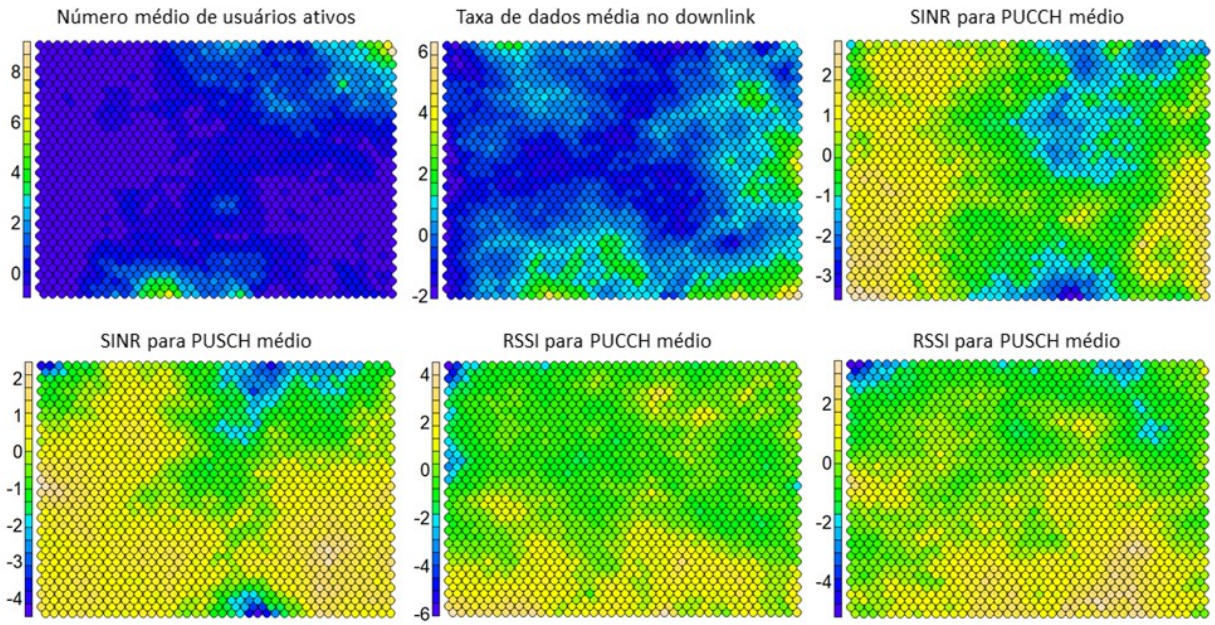
09:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

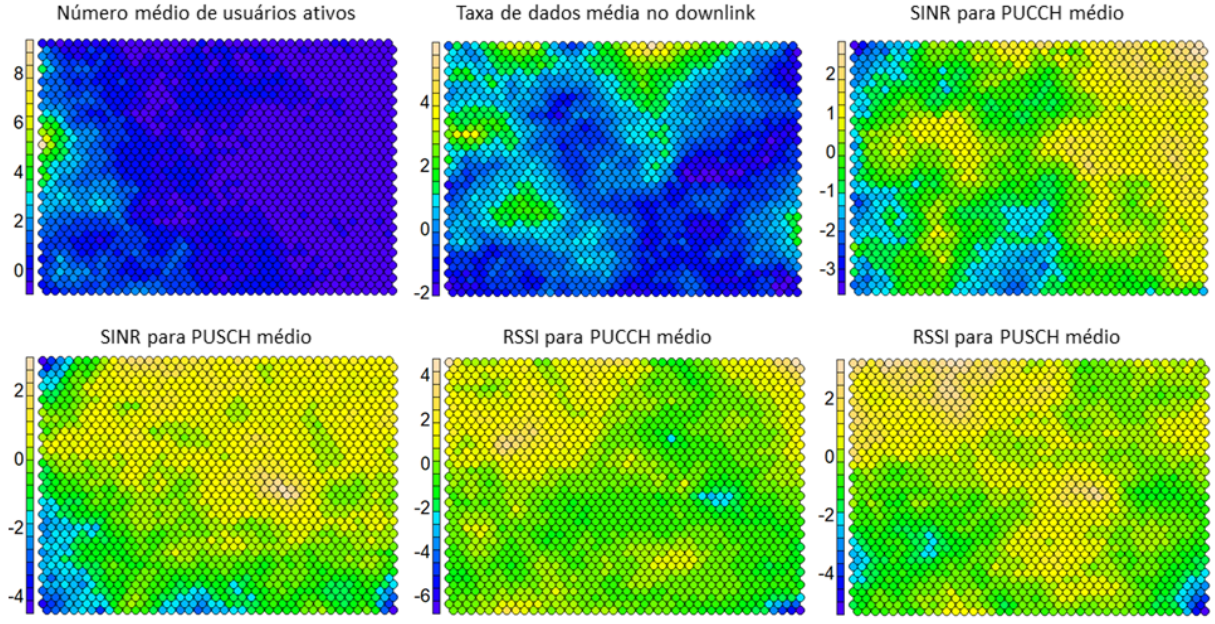
10:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

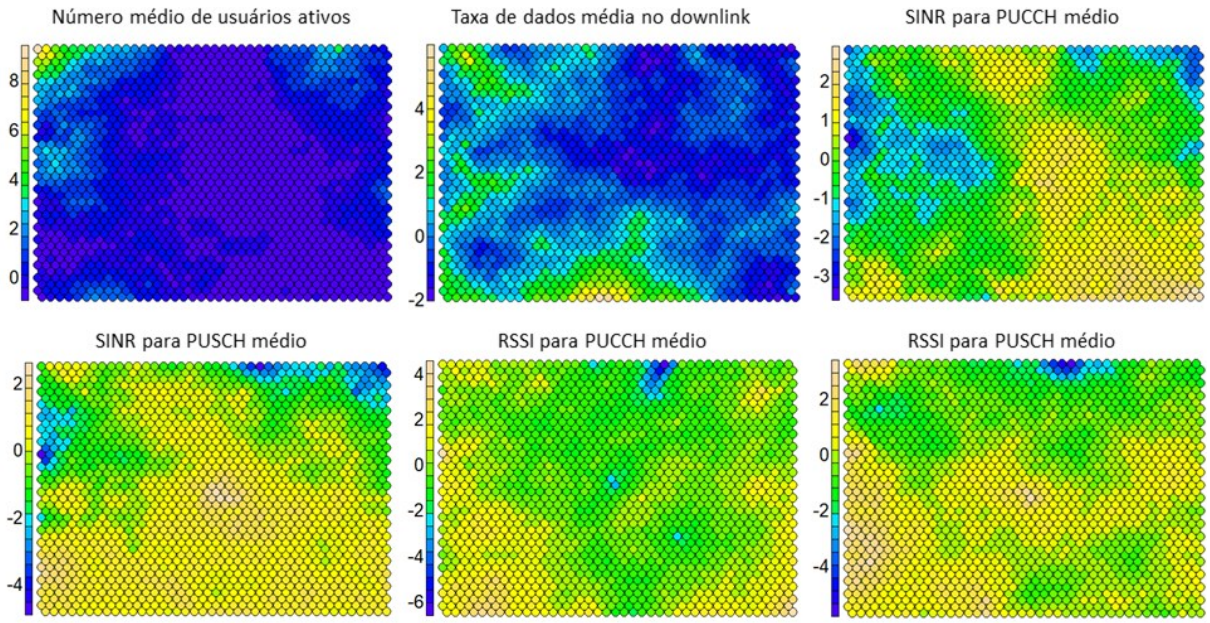
11:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

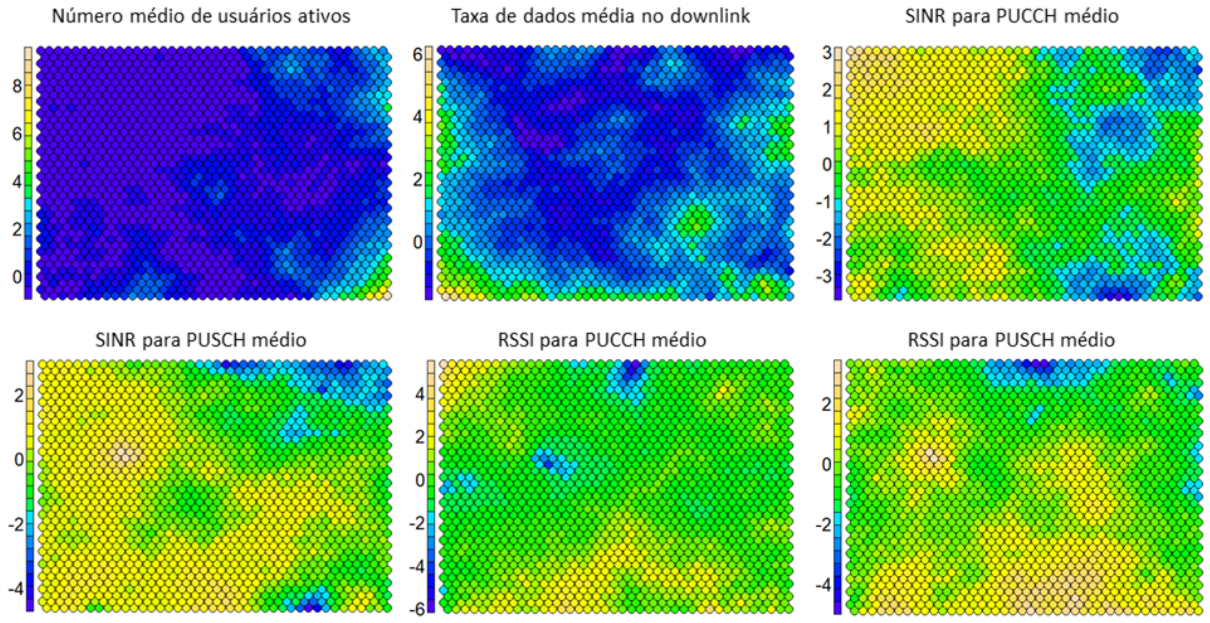
12:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

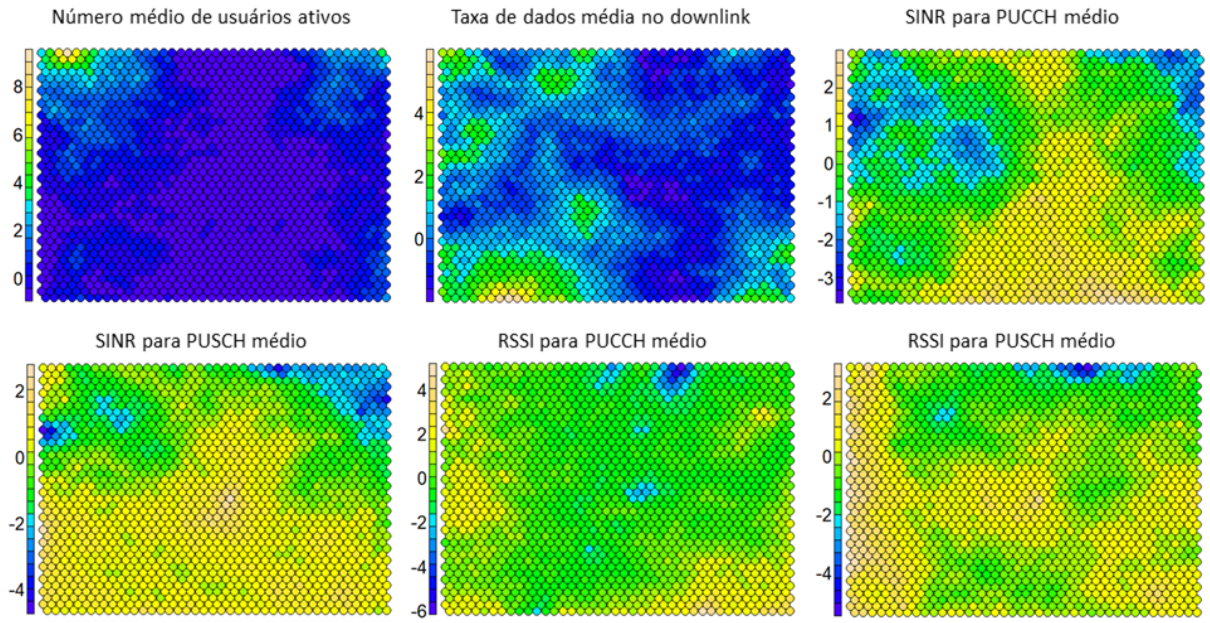
13:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

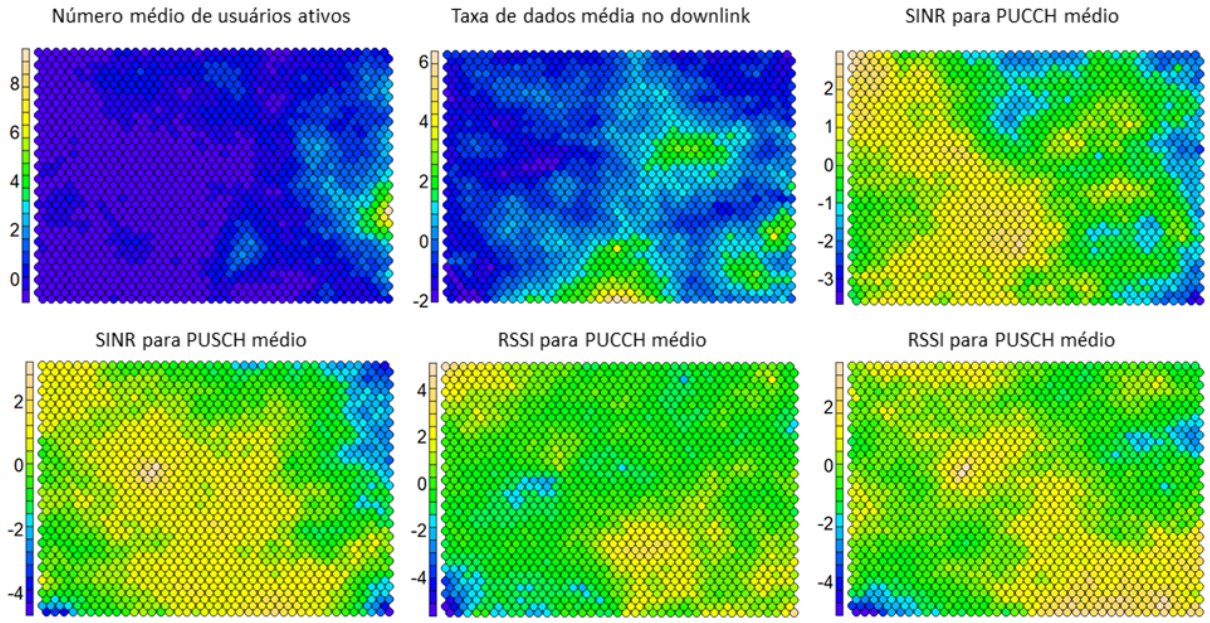
14:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

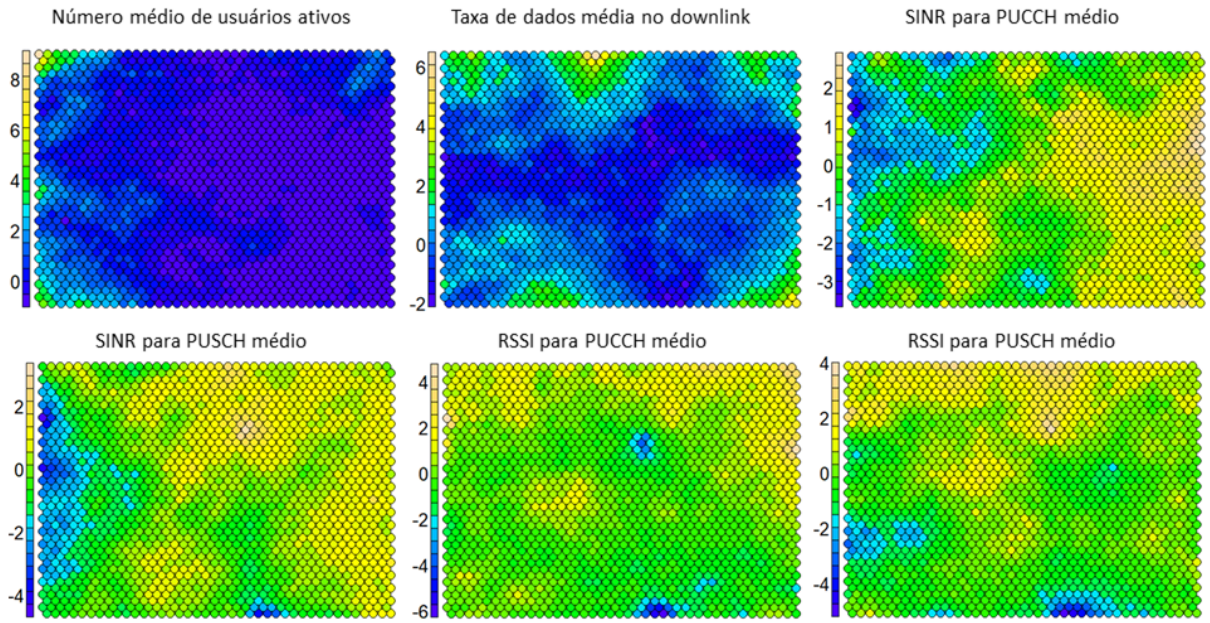
15:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

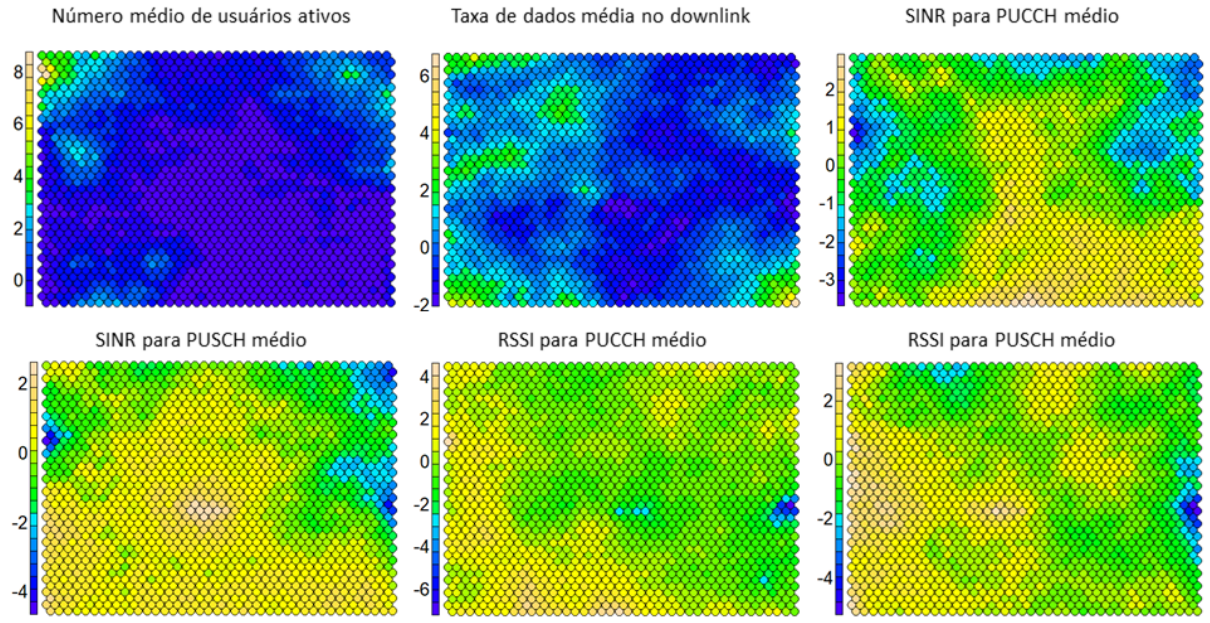
16:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

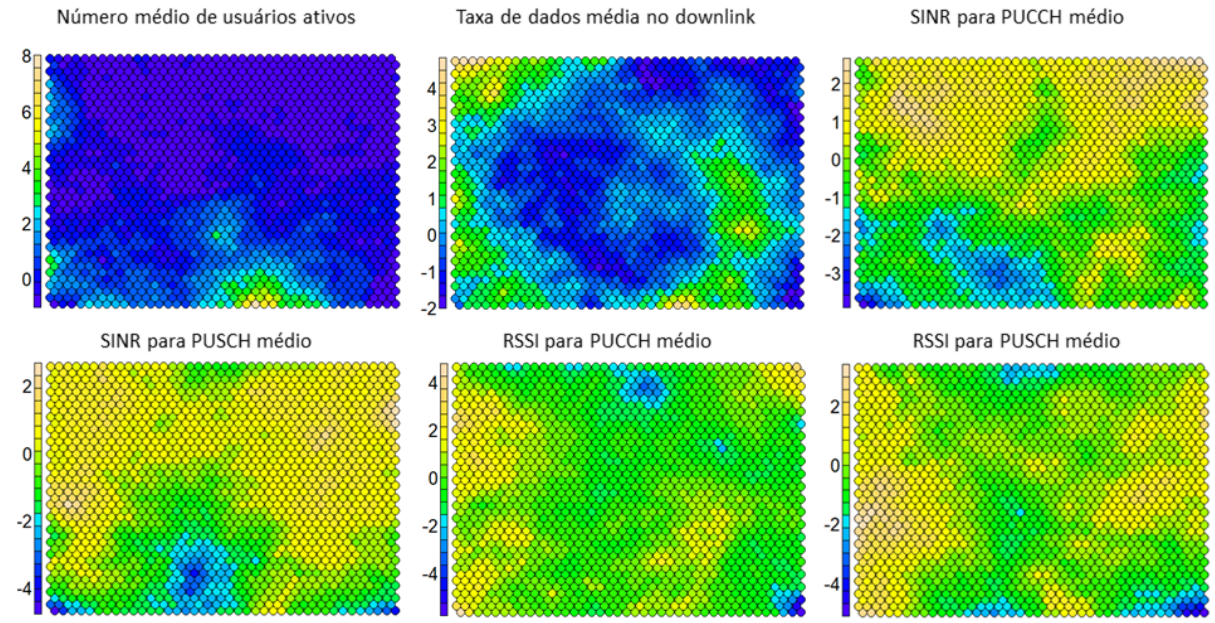
17:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

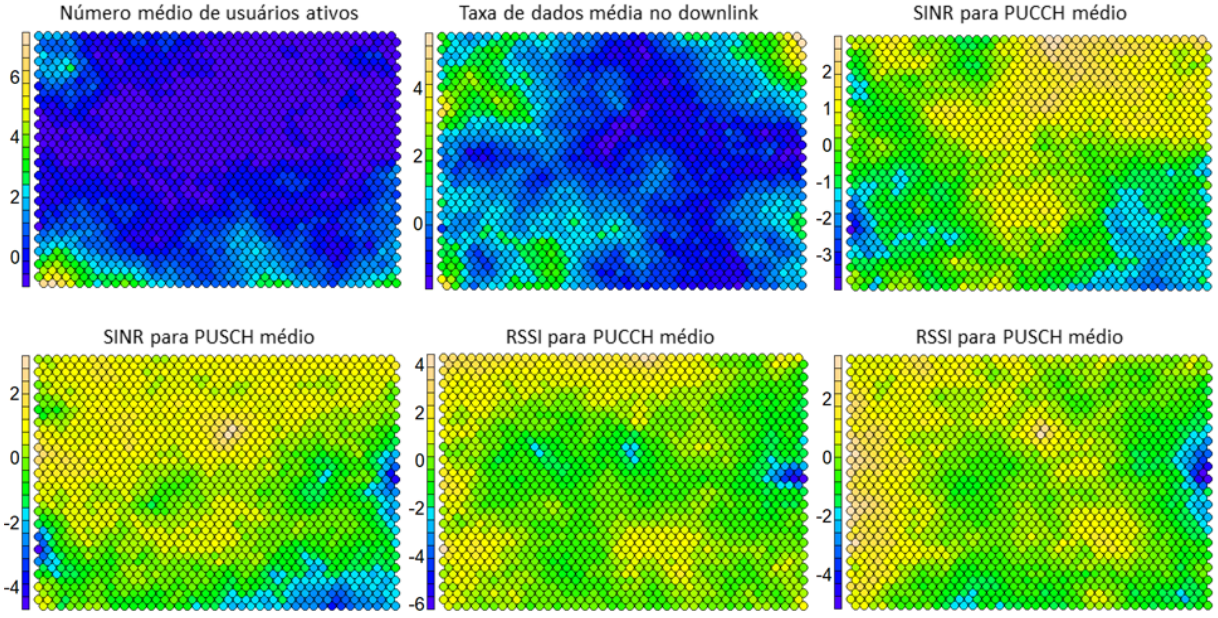
18:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

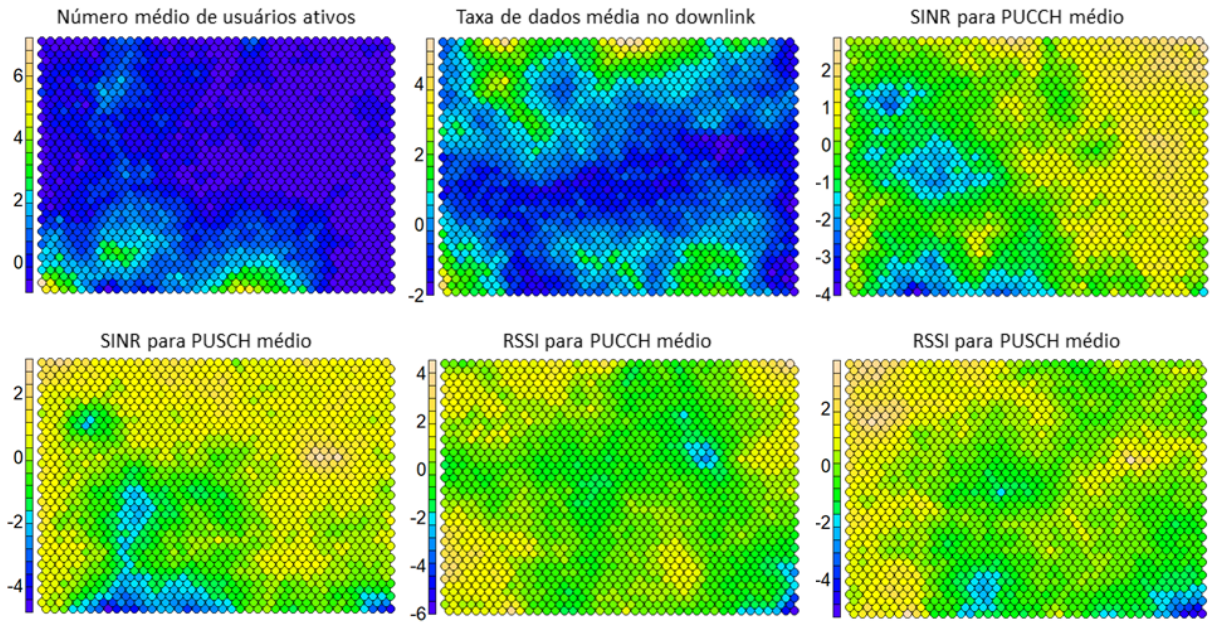
19:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

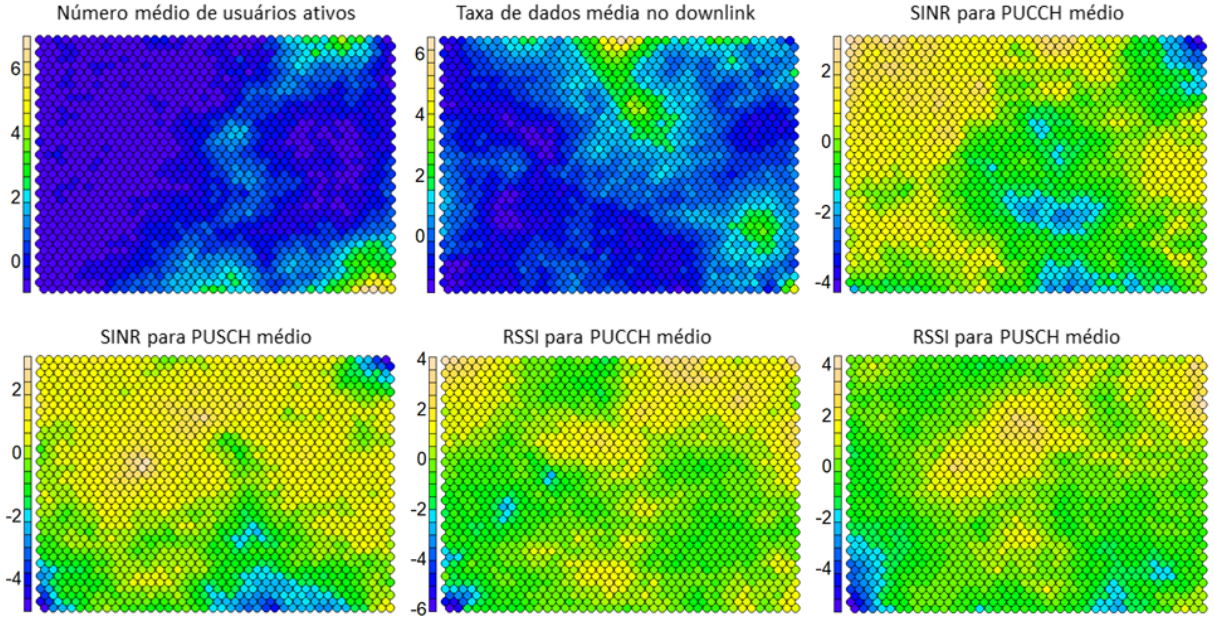
20:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

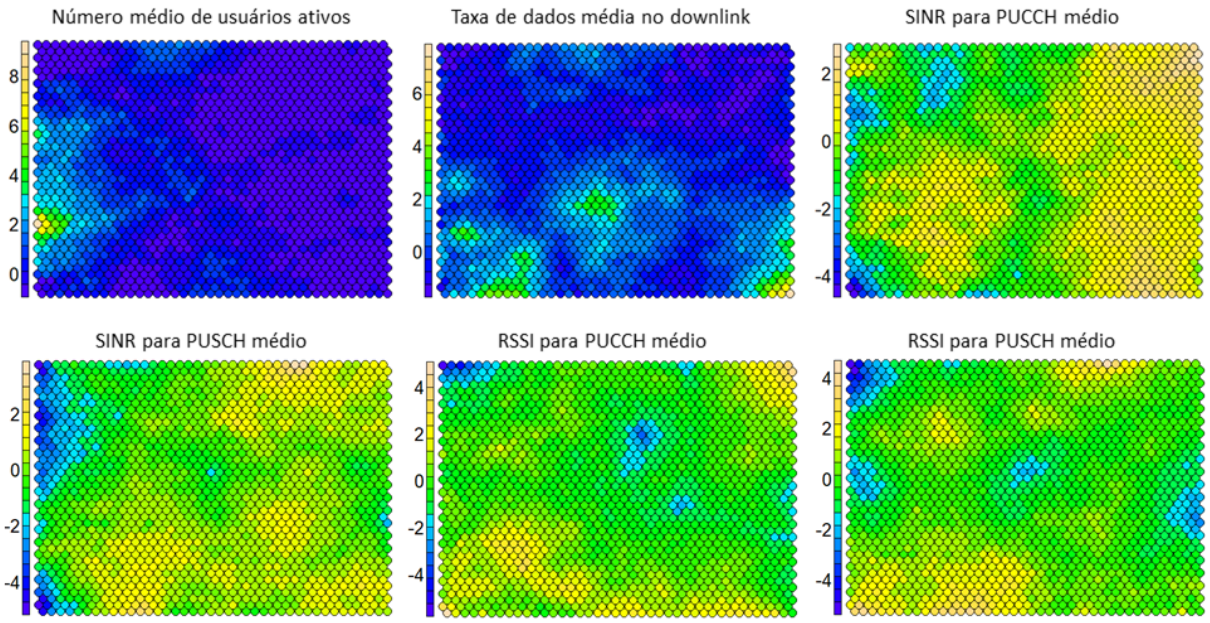
21:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

22:00h


Número médio de usuários ativos
Taxa de dados média no downlink
SINR para PUCCH médio
SINR para PUSCH médio
RSSI para PUCCH médio
RSSI para PUSCH médio

23:00h


Número médio de usuários ativos
8
6
4
2
0
Taxa de dados média no downlink
6
4
2
0
SINR para PUCCH médio
2
0
-2
-4
SINR para PUSCH médio
2
0
-2
-4
RSSI para PUCCH médio
4
2
0
-2
-4
RSSI para PUSCH médio
4
2
0
-2
-4

Apêndice C: Lista de publicações

Abaixo lista-se as produções científicas que foram frutos deste trabalho até a sua apresentação para a banca.

- Publicação do artigo “Proposta de um método de análise do desempenho de redes móveis com base na representação dimensional de KPIs usando PCA e clusterização” no “XXXVII Simpósio Brasileiro De Telecomunicações e Processamento de Sinais – SBrT2019” disponível em < http://sbrt.org.br/sbrt2019>.