UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Gabriel Wanderley Albuquerque Silva

Redução de Dimensionalidade Aplicada a Sistemas de Radiolocalização por Regressão Direta em Regiões com Diferentes Níveis de Urbanização

Recife
2021

Gabriel Wanderley Albuquerque Silva

Redução de Dimensionalidade Aplicada a Sistemas de Radiolocalização por Regressão Direta em Regiões com Diferentes Níveis de Urbanização

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Áreas de Concentração: Redes de Computadores e Sistemas Distribuídos/Inteligência Computacional

Orientador: Daniel Carvalho da Cunha

Recife

2021

Catalogação na fonte
Bibliotecária: Mônica Uchôa, CRB4-1010

S586r Silva, Gabriel Wanderley Albuquerque.
Redução de dimensionalidade aplicada a sistemas de radiolocalização por regressão direta em regiões com diferentes níveis de urbanização / Gabriel Wanderley Albuquerque Silva. – 2021.
68 f.: il., fig., tab.

Orientador: Daniel Carvalho da Cunha.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn. Programa de Pós-graduação em Ciência da Computação. Recife, 2021.
Inclui referências e apêndices.

1. Radiolocalização. 2. Aprendizado de máquina. 3. Regressão direta. 4. Extração de características. 5. Tempo de processamento. I. Cunha, Daniel Carvalho da (Orientador). II. Título.

681.3 CDD (23. ed.) UFPE- CCEN 2021 - 190

Gabriel Wanderley Albuquerque Silva

“Redução de Dimensionalidade Aplicada a Sistema de Radiolocalização por Regressão Direta em Regiões com Diferentes Níveis de Urbanização”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Redes de Computadores e Sistemas Distribuídos.

Aprovado em: 10/09/2021.

BANCA EXAMINADORA

_____

Prof. Dr. Germano Crispim Vasconcelos
Centro de Informática / UFPE

_____

Prof. Dr. Waslon Terllizzie Araújo Lopes
Centro de Energias Alternativas e Renováveis / UFPB

_____

Prof. Dr. Daniel Carvalho da Cunha
Centro de Informática/ UFPE
(Orientador)

Dedico a todos que acreditam no poder da ciência, da educação e da cultura.

AGRADECIMENTOS

Agradeço a todos que contribuíram de alguma forma com o desenvolvimento desta dissertação. Principalmente à minha mãe, Rosileide Albuquerque, por ter me guiado no caminho da educação e por ter me ensinado que ter um bom caráter e comprometimento para com minhas obrigações é essencial. Agradeço ao meu orientador, Daniel Carvalho da Cunha, por ter me orientado com maestria, por ter dedicado seu tempo nas reuniões e revisões, e também por ser um amigo nesse complexo processo. Agradeço aos amigos e familiares pela compreensão e forças depositadas em mim quando eu precisava.

Obrigado.

RESUMO

A difusão do uso de dispositivos móveis (DMs) tem estimulado a adoção de inúmeros serviços baseados em localização que, por sua vez, dependem de técnicas de localização em redes sem fio. Apesar do sistema de posicionamento global ser uma das principais técnicas usadas para fornecer a localização de DM, sua acurácia depende fortemente da existência de linha de visada entre transmissor e receptor. Para evitar tal desvantagem, técnicas de radiolocalização baseadas nos níveis de potência do sinal de rádiofrequência (RF) recebidos são amplamente utilizadas. Uma dessas técnicas, chamada de método de localização por regressão direta (LRD), emprega algoritmos de aprendizado de máquina para fazer a predição das coordenadas geográficas do DM. Face ao exposto, este trabalho analisou a aplicação do método LRD em duas regiões com diferentes níveis de urbanização. Nas regiões consideradas, bases de dados contendo níveis de sinal de RF de três gerações de redes celulares foram construídas, de forma unificada, assim como segmentada por rede, a partir de coleta via *crowdsourcing*. O primeiro aspecto da análise foi a robustez do método de localização em função do nível de urbanização das regiões consideradas. O método LRD se mostrou mais estável (diminuição do erro médio de predição em função do aumento do conjunto de treinamento) na região com maior nível de urbanização e mais eficiente quando aplicado à rede 3G em ambas as regiões. Além de fatores relacionados aos diferentes níveis de urbanização das regiões investigadas, o aumento esperado da quantidade de estações rádiobase com a implantação de redes de próxima geração também é relevante para a aplicabilidade do método LRD. Assim, o segundo aspecto analisado foi o efeito da redução de dimensionalidade na acurácia e nos tempos de execução do método LRD. Para isso, cinco algoritmos de extração de características (AECs), três lineares e dois não-lineares, foram considerados. Resultados experimentais mostraram que os AECs não-lineares obtiveram melhores resultados que os AECs lineares. Dentre os AECs não-lineares, o algoritmo KPCA-Sigmoide diminuiu o erro médio do método LRD em até 15% quando comparado ao erro do método LRD sem o uso de AECs. Além disso, o algoritmo KPCA-Sigmoide causou uma diminuição aproximada de sete vezes no tempo de treinamento e de aproximadamente quatro vezes no tempo de predição do método LRD, sem prejudicar a acurácia da localização.

Palavras-chaves: radiolocalização; aprendizado de máquina; regressão direta; extração de características; tempo de processamento.

ABSTRACT

The popularization of mobile devices (MDs) has promoted the use of multiple location-based services which rely on location techniques in wireless networks. Although the global positioning system is one of the main strategies used to provide the location of MDs, its accuracy depends on the existence of a line of sight between a transmitter and a receiver. To prevent this, radiolocation techniques based on received radio frequency (RF) signal strength levels are widely used. One of these techniques, which is called the direct regression location method (DRL), employs machine learning algorithms to predict the geographic coordinates of MDs. Given the above, this work has analyzed the application of the DRL method in two regions with different levels of urbanization. In the considered regions, databases were built (via crowdsourcing) containing RF signal levels from three generations of cellular networks. At first, in a unified way, then in a segmented way distributed by network. The first aspect of the analysis was the location method robustness, taking into consideration the urbanization level of the chosen regions. The DRL method was more stable, as it decreased the prediction mean error according to the training set increasing in the region with the highest level of urbanization, as well as it has seemed to be more efficient when applied to the 3G network in both regions. In addition to issues related to the region's different levels of urbanization, the expected growth in the number of base stations that follows the implementation of next-generation networks is also relevant for the applicability of the DRL method. The second aspect this work has analyzed was the effect of dimensionality reduction on the accuracy and on the execution times of the DRL method. For this, five feature extraction algorithms (FEAs), three linear and two non-linear ones, were considered. Experimental results have shown that non-linear FEAs obtained better results than linear FEAs. Among the non-linear FEAs, the KPCA-Sigmoid algorithm reduced the mean error of the DRL method by up to 15% when compared to the error of the DRL method without the use of FEAs. Added to that, the KPCA-Sigmoid algorithm caused a decrease of approximately seven times in training time and approximately four times in the prediction time of the DRL method, without compromising the accuracy of the location.

Keywords: radiolocation; machine learning; direct regression; feature extraction; processing time.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

k-NN	*k-nearest neighbors*
t-SNE	*t-distributed stochastic neighbor embedding*
2G	segunda geração
3G	terceira geração
4G	quarta geração
AEC	algoritmo de extração de características
AM	aprendizado de máquina
ANATEL	Agência Nacional de Telecomunicações
BTS	*base transceiver station*
DM	dispositivo móvel
DTM	densidade de telefonia móvel
E-911	*Enhanced* 911
EMA	erro médio absoluto
ERB	estação rádio base
FCC	*Federal Communications Commission*
GPS	*global positioning system*
GSM	*global system for mobile communications*
ICA	*independent component analysis*
IoT	*internet of Things*
ISOMAP	*isometric mapping*
JSON	*JavaScript Object Notation*
KPCA	*kernel* PCA
LDA	*linear discriminant analysis*
LRD	localização por regressão direta
LRD/C-AEC	método LRD com algoritmo de extração de características
LRD/S-AEC	método LRD sem algoritmo de extração de características
MDS	*multi-dimensional scaling*
PC	*principal component*
PCA	*principal component analysis*
RF	rádiofrequência
RSSI	*received signal strength indicators*

SVD	*singular value decomposition*
SVM	*support-vector machine*
VANETs	*vehicle ad hoc networks*
Wk-NN	*weighted k-nearest neighbors*
Z-BNU	zona com baixo nível de urbanização
Z-ANU	zona com alto nível de urbanização

SUMÁRIO

1 INTRODUÇÃO

O crescimento do uso de dispositivo móvel (DM) vem possibilitando a utilização de processos digitais para tarefas básicas no dia-a-dia de usuários de redes móveis. As chamadas aplicações de Internet das Coisas (IoT, *Internet of Things*) utilizam dados gerados por dispositivos para simplificar e enriquecer atividades e experiências humanas (MAHDAVINEJAD et al., 2018). Dentre as principais aplicações de IoT está a capacidade de realizar a predição da localização de DMs por meio de dados extraídos a partir de sinais de rádiofrequência (RF). O sistema global de posicionamento (GPS, *global positioning system*) é considerado um dos métodos mais conhecidos de localização de DMs. Porém, o GPS apresenta algumas limitações, como, por exemplo, a necessidade de visada direta entre transmissor e receptor, bem como a alta demanda de energia do DM. Para suprir essas limitações, algumas técnicas foram propostas na literatura, incluindo abordagens baseadas em trilateração (OGUEJIOFOR et al., 2020) e em *fingerprinting* (APOSTOLO; SAMPAIO; VITERBO, 2019), ambas baseadas no processamento de indicadores de intensidade do nível de sinal (RSSIs, *received signal strength indicators*) relacionados às estação rádio base (ERB) detectadas pelo DM.

Um método de localização por regressão direta (LRD) foi proposto em (OLIVEIRA et al., 2019) e utiliza algoritmos de aprendizado de máquina (AM) para realizar a predição direta das coordenadas geográficas do DM com base em seus valores de RSSI coletados. Em outros termos, as ERBs e seus respectivos valores de RSSI, detectados pelo DM, são transformados em características da regressão e as coordenadas geográficas do DM são os alvos da regressão. A etapa de treinamento do método LRD é mais simples que a etapa de treinamento do método *fingerprinting*, pois dispensa a construção de um mapa de rádio. Além disso, o método LRD oferece uma maior compatibilidade com a região de aplicação, em oposição a métodos de trilateração. Esta compatibilidade é justificada pela forma que a base de dados de RSSIs é construída, a qual utiliza métodos de *crowdsourcing* para inclusão de registros. Métodos de *crowdsourcing* obtêm recursos ou informações através de usuários da própria rede e, no caso deste trabalho de mestrado, é aplicado para compor a base utilizada para treinamento do sistema.

A acurácia e o tempo de execução do método LRD estão diretamente relacionados à disponibilidade de ERBs na região de aplicação do método (OLIVEIRA et al., 2019). Face a esta particularidade inerente a métodos baseados em algoritmos de AM, torna-se necessário investigar a aplicação do método LRD em regiões com diferentes densidades de ERBs. Um dos fatores que afeta a densidade de ERBs é o nível de urbanização da região. Por exemplo, zonas com alto nível de urbanização (Z-ANUs) são ambientes com densidade demográfica elevada e infraestrutura de redes de comunicação abrangente e densa. No entanto, zonas com baixo nível de urbanização (Z-BNUs) apresentam densidade demográfica reduzida e, além disso, possuem infraestrutura de redes de comunicação que

não engloba todo o território. Outra característica que distingue essas duas regiões é a quantidade e a proporção entre as ERBs de diferentes gerações de rede celular. Enquanto as Z-ANUs tendem a ter infraestrutura de rede de melhor qualidade, ou seja, uma maior proporção de ERBs de gerações de rede mais recentes, as Z-BNUs tendem a levar mais tempo para atualizar as tecnologias disponíveis (HONG; THAKURIAH, 2018). Outro fato interessante é que há a possibilidade de alguns DMs não possuirem infraestrutura de rede para detectar ERBs de todas as gerações de rede disponíveis. Todos esses fatores apresentados afetam a acurácia e o tempo de execução do método LRD.

Outro aspecto importante é o aumento do número de ERBs com o decorrer dos anos. Estima-se que haverá mais de 500 bilhões de DMs conectados à Internet até o ano de 2030. Essa tendência tem motivado a implantação das redes 5G, bem como o estudo das redes 6G, tecnologias com altas taxas de transmissão de dados e baixas latências. Para tornar esse potencial disponível para todos os DMs, novas ERBs precisarão ser disponibilizadas. A expectativa é que esse número de novas ERBs relativas à rede 5G, por exemplo, seja muito maior que o número de ERBs 4G implantadas (SHAFIQUE et al., 2020). O aumento do número de ERBs significa crescimento do número de características no método LRD, o que impacta diretamente no tempo de execução de algoritmos de AM (ANOWAR; SADAOUI; SELIM, 2021). Adicionalmente, algumas ERBs podem ter RSSIs proporcionais ou até mesmo equivalentes entre si, pois podem estar localizadas na mesma torre de transmissão, respeitando, assim, regras semelhantes de propagação de sinais de RF. Neste caso, técnicas de redução de dimensionalidade podem ser aplicadas para diminuir a quantidade de características na base de entrada do método LRD, sem comprometer a sua acurácia (QI; JIN; YAN, 2018). Dentre as técnicas de redução de dimensionalidade, os AECs são comumente utilizados. Além de reduzirem a quantidade de características, os AECs promovem uma combinação ou cruzamento entre as características existentes, gerando um conjunto menor de novas características com preservação da variância e covariância das características do conjunto original. Manter esses aspectos no novo conjunto de características contribui para o bom desempenho de algoritmos de AM (ZHAO et al., 2013).

1.1 MOTIVAÇÃO E OBJETIVOS DO TRABALHO

É fundamental que novos métodos de radiolocalização de DMs, como o método LRD, obtenham bons resultados de acurácia e de tempo de execução em diversas configurações de ambientes, como as zonas de diferentes níveis de urbanização, além de serem compatíveis com as diferentes gerações de rede celular disponíveis. Por exemplo, a obtenção de boa acurácia é um requisito primário para situações de localização de DMs originadores de chamadas de emergência. Além disso, é interessante promover a escalabilidade do método LRD em relação à quantidade de características, visto que o número de ERBs apresenta uma tendência de crescimento com a adoção de redes de próxima geração, como as redes

5G.

Desta forma, o objetivo geral deste trabalho de mestrado é avaliar a aplicação do método LRD em zonas com diferentes níveis de urbanização para avaliar a adaptabilidade do método às características dessas regiões. Com isso em mente, os objetivos específicos do trabalho estão apresentados a seguir:

- Avaliar o emprego do método LRD a partir de bases de dados unificadas (todas as gerações de rede celular adotadas) para cada tipo de região;
- Analisar a utilização do método LRD a partir de bases de dados segmentadas por geração de rede celular e por tipo de região;
- Investigar o uso dos AECs selecionados nas bases de dados de entrada do método LRD de cada região considerada com o intuito de determinar o algoritmo com a configuração mais adequada;
- Obter resultados de acurácia e tempos de execução (treinamento/predição) do método LRD sem e com o emprego de AECs para verificar os benefícios da redução de dimensionalidade no problema de localização estudado.

1.2 ESTRUTURA DA DISSERTAÇÃO

Nesta Seção, é listada a organização deste trabalho. No Capítulo 2, são introduzidos e discutidos os conceitos e as terminologias referentes ao método LRD. Nele, são apresentadas as fases de treinamento (*off-line*) e predição (*on-line*), bem como a abordagem do algoritmo de AM utilizado. Ao fim do capítulo, trabalhos relacionados à localização de DMs são discutidos.

O Capítulo 3 apresenta os cinco AECs selecionados para aplicação nas bases de dados geradas neste trabalho, sendo três deles classificados como AECs lineares e dois como AECs não-lineares. Após as breves descrições dos algoritmos, trabalhos relacionados à redução de características e localização são abordados.

O Capítulo 4 aborda os experimentos e retrata as características das duas regiões escolhidas, assim como descreve a metodologia de coleta, pré-processamento e geração das bases de dados. Em seguida, a análise comparativa dos dados é realizada considerando os erros médios gerados pela aplicação do método LRD com e sem o emprego de AECs nas regiões determinadas.

Por fim, o Capítulo 5 apresenta as principais conclusões derivadas deste trabalho e as perspectivas de trabalhos futuros.

1.3 RESUMO DAS PRINCIPAIS CONTRIBUIÇÕES

As principais contribuições desta dissertação estão listadas a seguir:

- O método LRD apresentou uma acurácia maior na região com alto nível de urbanização. Considerando o caso em que bases de dados unificadas são empregadas, o erro médio de localização obtido na região Z-ANU é aproximadamente cinco vezes menor do que o erro obtido na região Z-BNU.
- O método LRD aplicado à base da rede 3G obtém o maior valor de eficiência dentre os valores analisados. Além disso, o método LRD aplicado à base da rede 4G atinge uma melhor eficiência na região Z-ANU em comparação com a Z-BNU. O método LRD aplicado à base da rede 2G obteve eficiências semelhantes em ambas as regiões;
- Dentre os AECs investigados neste trabalho, os AECs não-lineares obtêm melhores resultados que os AECs lineares em ambas as regiões. Em alguns casos, a aplicação do método LRD com o uso de AECs diminuiu em aproximadamente 15% o erro médio da aplicação do método LRD sem o uso de AECs;
- A redução de dimensionalidade causou uma diminuição do tempo de processamento do método LRD, sendo de aproximadamente sete vezes no tempo de treinamento e de aproximadamente quatro vezes no tempo de predição, sem prejudicar a acurácia da localização.

1.4 PRODUÇÃO BIBLIOGRÁFICA

- Publicação do artigo "*An RSS-based regression model for user equipment location in cellular networks using machine learning*" no periódico **Springer Wireless Networks** (Qualis A2 - Computação).
- Publicação do artigo "*Localização de usuários móveis baseada em fingerprint de rádio frequência: redução de espaço de busca usando parâmetros de atraso de onda das redes celulares*" nos Anais do **XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2020)** (Qualis A4 - Computação).

2 LOCALIZAÇÃO DE USUÁRIOS MÓVEIS EM REDES SEM FIO

A capacidade de localizar DMs tornou-se algo indispensável nas aplicações de IoT (MISTRY; MISTRY, 2015). Dentre algumas aplicações, estão aquelas voltadas à saúde do idoso e outras direcionadas a redes *ad-hoc* veiculares (VANETs, *vehicle ad hoc networks*) (SHUKRI et al., 2017),(KUUTTI et al., 2018). O monitoramento do comportamento do movimento dos usuários móveis durante a pandemia da Covid-19 teve grande impacto no controle do contágio da doença, desde que limitar a mobilidade humana certamente reduz a transmissão do vírus, assim como também dá suporte a diversos serviços em cidades inteligentes em situações de *lockdown* (DONG; YAO, 2021),(SAEED et al., 2020),(FAZIO et al., 2020). O GPS é uma das técnicas mais utilizadas para esta finalidade, porém apresenta diversas limitações, como a necessidade de linha de visada direta para satélites (o que impede a localização em ambientes *indoor*), além de drenar a bateria do DM rapidamente (RIZK; SHOKRY; YOUSSEF, 2019). Diante dessas limitações, técnicas baseadas no processamento de RSSIs se apresentam como alternativa, pois a obtenção dos RSSIs não possui as limitações relacionadas ao GPS.

No contexto das redes celulares, o DM consegue detectar sinais emitidos por diversas ERBs e registra o RSSI de cada um desses emissores. As ERBs estão situadas em torres de transmissão que, por sua vez, ficam posicionadas em locais estratégicos, de forma a maximizar a propagação do sinal em uma determinada região. A taxa de atualização do RSSI não é fixa, porém apresenta uma variação que auxilia diversos modelos de predição de localização de DMs . Cada DM é associado a um usuário da rede, independente da tecnologia de transmissão. Nesta dissertação, admite-se que um DM está associado a um usuário móvel da rede de telefonia celular.

Dentre as técnicas de localização baseadas em RSSI, a técnica de trilateração é a mais simples e a que exige menor processamento computacional, pois envolve apenas cálculos com complexidade lineares e é baseada em modelos de sinais propagação (OGUEJIOFOR et al., 2020). Contudo, esses modelos de propagação variam de acordo com o tipo de ambiente, seja ele interno, externo, urbano, com obstáculos etc. Os modelos também são afetados por diversos fatores, como altura, ganhos em potência e padrão de irradiação das antenas, distância entre o transmissor e o receptor, reflexão ou transmissão por múltiplos caminhos (MORAVEK et al., 2011). Todas essas características tornam a implantação da técnica de trilateração complexa e arriscada, visto que a calibração desses parâmetros pode não refletir as características reais de propagação.

Uma segunda técnica de localização é a abordagem por *fingerprinting*, que é baseada na associação entre os valores de RSSI previamente coletados em um determinado local com os valores de RSSI medidos pelo DM a ser localizado. Esta técnica possui duas fases, denominadas de *off-line* e *on-line*. Na fase *off-line*, amostras de referência contendo valores de RSSI reais são coletadas e associadas a um determinado conjunto de coordenadas

geográficas. Tal conjunto de valores de referência, chamado de mapa de rádio, constitui um mapa de *fingerprints* da região de interesse. Na fase *on-line*, a localização do DM é predita por meio da associação entre os valores de RSSI medidos no próprio DM e os valores de RSSI contidos no mapa de rádio (APOSTOLO; SAMPAIO; VITERBO, 2019). Diferente da trilateração, a técnica de *fingerprinting* fornece bons resultados não apenas em ambientes *outdoor* (VO; DE, 2016), mas também em ambientes *indoor* (KHALAJMEHRABADI; GATSIS; AKOPIAN, 2017). Por outro lado, a técnica de *fingerprinting* ainda pode fazer uso de modelo de propagação ou algum método de estimativa de sinais de RF para locais onde não há medições reais de RSSI, trazendo para a abordagem alguns dos problemas e complicações inerentes a esses modelos.

2.1 LOCALIZAÇÃO POR REGRESSÃO DIRETA DE USUÁRIOS MÓVEIS EM REDES SEM FIO

Um novo método de localização baseado em RSSI chamado de localização por regressão direta (LRD) foi proposto em (OLIVEIRA et al., 2019). Além do método LRD ter as vantagens da técnica de localização por *fingerprinting*, como, por exemplo, bom desempenho em ambientes *indoor* e realização da predição baseando-se em dados reais, o método LRD elimina as dependências com os modelos de propagação. O método LRD consiste no uso de algoritmos de AM como um modelo de regressão com base nos valores de RSSI detectados pelo DM para fazer a predição direta de suas coordenadas geográficas e, assim, obter sua localização. Assim como a técnica de *fingerprinting*, o método LRD possui duas fases, também chamadas de *off-line* e *on-line*. No entanto, por fazer a regressão direta dos valores alvos, o método LRD não necessita do conhecimento prévio de características das antenas de transmissão, sequer de modelos de propagação, como a técnica de *fingerprinting*. Em (OLIVEIRA et al., 2019), além da proposição do método LRD, foi realizado um estudo comparativo entre os algoritmos de AM usados na regressão e concluiu-se que o uso do algoritmo *k-nearest neighbors* (k-NN) como regressor gerou um melhor desempenho no cenário com um número maior de ERBs. Além disto, o algoritmo k-NN é um algoritmo com processos claros e de pequena complexidade. Por esta razão, o regressor k-NN aplicado ao método LRD será considerado neste trabalho.

2.1.1 Algoritmo k-NN

O algoritmo k-NN é uma técnica de AM que compara a amostra atual com as k amostras mais próximas, também chamadas de vizinhos, existentes na base de dados de treinamento (COVER; HART, 1967). O k-NN pode ser usado tanto em tarefas de classificação quanto de regressão e é descrito conforme a seguir. Dada a i-ésima instância de teste

X_i, o primeiro passo é encontrar os k vizinhos desta instância X_i dentre as amostras pertencentes à base de treinamento. O desempenho do algoritmo está diretamente associado ao método que mede essa distância entre amostras. Neste trabalho, temos as amostras definidas como um vetor de características m-dimensional $X_i = [X_{i1}, X_{i2}, ..., X_{im}]$ e a métrica escolhida foi a distância Euclidiana, uma das métricas de distância mais usadas na literatura (KUHN; JOHNSON, 2013). A distância Euclidiana entre duas amostras é definida como

$$d\left(X_i, X_j\right) = \sqrt{\sum_{r=1}^{m} \left(|X_{ir} - X_{jr}|\right)^2} \,, \tag{2.1}$$

em que X_i pertence às amostras de teste, X_j pertence às amostras de treinamento e m é a quantidade de características dessas instâncias.

Depois da seleção das k amostras mais próximas de X_i, o valor predito ou valor alvo de X_i, denotado por $\hat{f}(X_i)$, é dado pela média de valores das k instâncias mais próximas de X_i. Desta forma, o valor alvo de X_i é dado por

$$\hat{f}(\mathbf{X}_i) \leftarrow \frac{\sum_{j=1}^{k} f(\mathbf{X}_j)}{k} \,, \tag{2.2}$$

em que $f(X_j)$ é o valor alvo da instância de treinamento X_j. Maiores informações sobre o algoritmo k-NN podem ser encontradas em (COVER; HART, 1967).

2.1.2 Fase *off-line*

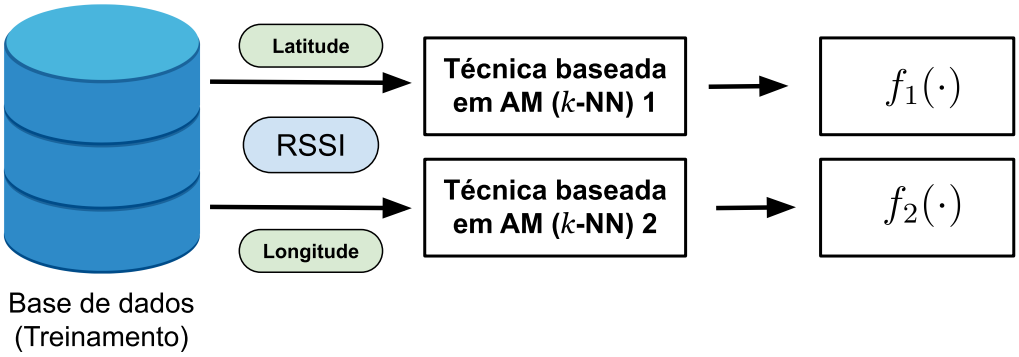
A fase *off-line* do método LRD, também chamada de fase de treinamento, acontece previamente à execução da predição da localização do DM, uma vez que prepara o sistema para lidar com os dados que serão inseridos na fase de predição. Em AM, a fase de treinamento refere-se a estimativa dos parâmetros do modelo ou do algoritmo de AM. Porém, nesta dissertação, a fase de treinamento refere-se ao processo de ajuste ou busca do melhor valor do parâmetro k. O Algoritmo 1 descreve em detalhes a fase *off-line* do método LRD.

No passo 1, uma coleta prévia é feita e todos os dados necessários para treinamento são obtidos, isto é, a posição real (coordenadas geográficas) do DM , assim como os valores de

Algoritmo 1 Descrição da fase *off-line* do método LRD.

1: Coletar as amostras formadas pelos RSSI e coordenadas para construir a base de dados de treinamento.
2: Treinar o algoritmo baseado em AM considerando a latitude como valor alvo, obtendo a função de hipótese $f_1(\cdot)$.
3: Treinar o algoritmo baseado em AM considerando a longitude como valor alvo, obtendo a função de hipótese $f_2(\cdot)$.

Figura 1 – Fase *off-line* do método LRD: obtenção das funções de hipótese $f_1(\cdot)$ e $f_2(\cdot)$.



Fonte: Elaborada pelo autor (2021)

RSSI para todas as ERBs que aquele DM consegue detectar naquele determinado local. A forma como esta coleta ocorreu neste trabalho será apresentada na Seção 4.2, assim como detalhes das gerações de redes celulares abordadas e também das condições de coleta. Os passos 2 e 3 consistem em obter as funções de hipótese $f_1(\cdot)$ e $f_2(\cdot)$ usando um modelo baseado em AM, neste caso, o k-NN, que serão usadas para predizer a latitude e longitude alvos do DM. As funções de hipóteses $f_1(\cdot)$ e $f_2(\cdot)$ representam, cada uma, a instância do algoritmo k-NN para latitude e longitude, respectivamente, após a busca do melhor valor de k para aquele determinado conjunto de validação. Desta forma, assuma que as bases de dados S^{ϕ} e S^{λ} são dadas por

$$S^{\phi} = \left\{(\phi_i, \mathbf{q}_i) \in \mathbb{R} \times \mathbb{R}^{N_i}\right\} \tag{2.3}$$

e

$$S^{\lambda} = \left\{(\lambda_i, \mathbf{q}_i) \in \mathbb{R} \times \mathbb{R}^{N_i}\right\}, \tag{2.4}$$

em que ϕ_i e λ_i são, respectivamente, a latitude e a longitude do i-ésimo ponto de validação, $\mathbf{q}_i$ é o vetor de medição de RSSI no i-ésimo ponto e N_i denota a quantidade de ERBs presentes na i-ésima instância. Assim, S^{ϕ} e S^{λ} são as bases de dados destinadas à execução dos passos 2 e 3, respectivamente. Os valores de ϕ_i e λ_i são considerados como alvos, enquanto os valores contidos em $\mathbf{q}$ representam as características. A fase *off-line* do método LRD é ilustrada na Fig. 1.

O processo de otimização dos parâmetros do algoritmo de AM é fundamental. Por exemplo, para o k-NN, a variação do número de vizinhos impacta diretamente no desempenho final do algoritmo, pois um valor de k pequeno contribui para o aumento da influência de dados ruidosos ou dados que não seguem o padrão de comportamento da base de dados, assim como um valor de k grande pode aumentar a influência de vizinhos com valores díspares ou que pertencem a outras classes, além de implicar no aumento da complexidade computacional.

Algoritmo 2 Descrição da fase *on-line* do método LRD.

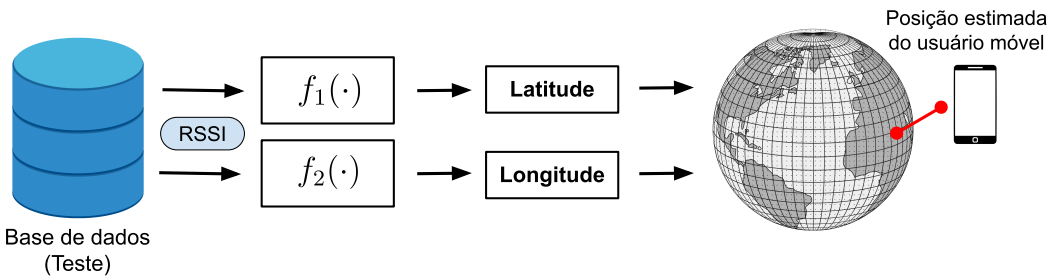
1: Coletar valores de RSSI do DM.
2: Dadas as medições de RSSI do DM, estimar latitude e longitude usando $f_1(\cdot)$ e $f_2(\cdot)$, respectivamente.
3: Calcular a posição do DM usando as coordenadas geográficas preditas.

2.1.3 Fase *on-line*

Após a conclusão da fase de treinamento, o sistema está preparado para realizar a fase de predição, também conhecida como fase *on-line*. Esta fase é composta pelos passos descritos no Algoritmo 2.

No primeiro passo, o DM deve estar conectado à rede celular. Este passo é semelhante ao primeiro passo da fase *off-line* e coleta os valores de RSSI das ERBs que este aparelho consegue sensoriar, porém, nesta fase, este não precisa estar conectado ao GPS. Após o sensoriamento, esses valores são usados como características de entrada para as funções de hipóteses $f_1(\cdot)$ e $f_2(\cdot)$ na fase 2, obtendo, respectivamente, a latitude e longitude alvos do DM . No passo 3, a posição final é estimada usando as coordenadas geográficas obtidas no passo anterior. A Fig. 2 ilustra a fase *on-line* do método LRD.

Figura 2 – Fase *on-line* do método LRD: execução da predição da posição do DM .



Fonte: Elaborada pelo autor (2021)

2.2 TRABALHOS RELACIONADOS

Nesta Seção, alguns trabalhos relacionados à localização de DMs em ambientes de diferentes níveis de urbanização serão discutidos. Alguns trabalhos abordam a aplicação de diversas técnicas de localização. Porém, em sua maioria, não apresentam diversidade de cenários, como análises particulares para cada geração de rede celular e variação no número de ERBs (características) na apresentação dos resultados. Devido ao recente e contínuo crescimento do número de DMs , principalmente em áreas com baixos níveis de urbanização, há poucos trabalhos que analisam o impacto da aplicação de técnicas de

localização nessas regiões. Porém, é importante considerar a disponibilidade de dados de treinamento e, principalmente, a densidade de ERBs em áreas pouco urbanizadas.

Meniem et al. propuseram um dos primeiros trabalhos a comparar o desempenho de técnica de localização baseada em valores de RSSI em ambientes com diferentes níveis de urbanização (MENIEM; HAMAD; SHAABAN, 2013). O trabalho apresenta uma técnica baseada no valor relativo do sinal recebido, a qual aplica um sistema de regras de correlações para indicar a posição final do DM. Apesar de também possuir fases *off-line* e *on-line*, a proposta não envolve técnicas de AM. No entanto, por ser um dos trabalhos pioneiros na avaliação de técnicas de localização em ambientes com diferentes níveis de urbanização, apresenta apenas dados baseados na segunda geração de redes celulares (GSM, *Global System for Mobile Communications*). Por fim, os resultados sugerem que, como esperado, na maior parte dos casos, há um aumento do erro médio na região rural (área de menor urbanização) em comparação com a região urbana (área de maior urbanização).

Elbakly e Youssef propuseram uma técnica de localização com o objetivo de não necessitar da complexidade de métodos tradicionais de localização como calibrações, coletas prévias de dados ou qualquer suporte adicional da infraestrutura de redes celulares existente (ELBAKLY; YOUSSEF, 2019). A técnica é baseada no diagrama de Voronoi, o qual traça regiões ou polígonos com linhas equidistantes entre esses pontos centrais (AURENHAMMER, 1991). Apesar de não depender de coletas de dados ou calibrações constantes, o método proposto necessita da localização real de todas as ERBs e torres de transmissão, assim como das informações de setorização dessas ERBs. Os resultados mostram que a técnica obteve menor acurácia (maior erro) em relação a técnicas como *fingerprinting* e LRD. Por outro lado, a proposta obtém bons resultados quando comparada a outras técnicas de baixo custo computacional, como *Cell ID* (DUFKOVá et al., 2008) e métodos baseados em centroides, nos quais a distância é estimada com base no centro de massa das localizações das ERBs detectadas pelo DM . Embora este trabalho compare a aplicação desta técnica em ambientes com diferentes níveis de urbanização e conclua também que há uma degradação da acurácia nos ambientes menos urbanizados ($152\,m$ e $224\,m$ de erro médio para as áreas urbana e rural, respectivamente), não houve um paralelo entre os resultados obtidos e os resultados de métodos complexos e que apresentam maior acurácia para esses ambientes em específico.

Recentemente, foi publicado um trabalho que apresenta uma técnica de *fingerprinting* baseada na combinação de duas técnicas de AM, quais sejam, k-NN com pesos (Wk-NN, *weighted k-nearest neighbors*) e redes neurais de múltiplas camadas (ABDALLAH; SAAB; KASSAS, 2018). O método proposto é aplicado em três ambientes com diferentes níveis de urbanização. Assim como o método LRD, a técnica proposta por Abdallah et. al. não necessita de informações prévias das antenas e suas ERBs. Ainda que o trabalho apresente resultados congruentes com outros estudos semelhantes, indicando uma menor acurácia em ambientes com baixo nível de urbanização em relação aos ambientes densamente urba-

nizados, tais análises são focadas na comparação entre o método proposto e a abordagem que aplica apenas o Wk-NN. Por consequência, o trabalho é insuficiente na comparação dos resultados entre os ambientes, não dispondo de uma análise completa das características dos dois locais, como correlação entre os erros médios e a densidade de ERBs das regiões. O estudo também não apresenta diversidade de cenários, como variação da base de dados por geração de rede celular.

3 REDUÇÃO DE DIMENSIONALIDADE

Segundo dados da Agência Nacional de Telecomunicações (ANATEL), o número de ERBs cresce ano após ano. Entre os anos de 2018 e 2020, a quantidade de ERBs no Brasil passou de 93.218 para 102.462, um crescimento de aproximadamente 10% em um período de três anos (ANATEL, 2021). Com a implantação da quinta geração das redes móveis celulares, esse número tende a crescer ainda mais. O aumento na quantidade de ERBs disponíveis impacta diretamente no número de características de algoritmos de AM aplicados ao problema da radiolocalização, em particular no algoritmo k-NN. O aumento ou diminuição da quantidade de ERBs pode afetar tanto a acurácia quanto o custo computacional dos sistemas de localização, neste último caso, alterando os tempos de execução das fases de treinamento e predição (VIKRAM et al., 2019). Além disso, as ERBs que ficam localizadas na mesma torre de transmissão também podem apresentar comportamentos parecidos, caso apresentem características de transmissão semelhantes. Neste caso, uma características pode ser equivalente a outra, não tendo relevância no sistema de predição. Tal fato possibilita o emprego de técnicas de redução de características que, por consequência, reduzem o tempo de execução das fases do método LRD.

Então, com o propósito de analisar as consequências da redução de características na acurácia e no tempo de processamento, alguns AECs foram aplicados ao método LRD nos cenários de urbanização propostos. O objetivo dos AECs é reduzir a quantidade de características, gerando o menor impacto possível na variância ou representatividade de tais características como um todo.

3.1 EXTRAÇÃO DE CARACTERÍSTICAS

AECs são usados para reduzir a quantidade de dados de entrada de sistemas de predição, mantendo as informações que são mais relevantes. O conjunto de dados resultante deste processo preserva boa parte da variância da base de dados original sem perder informações de uma determinada características por completo, como fazem os métodos de seleção de características (JOSHI; MACHCHHAR, 2014). Por este motivo, neste trabalho, serão aplicados apenas métodos de extração de características.

Todos os AECs têm em comum a transformação das características originais em novas características que são mais significativas ou importantes para o processo de predição. Tais métodos reduzem a complexidade do conjunto e geram novas características a partir da combinação (linear ou não-linear) das características originais, tendo como resultado uma representação simples da base original (KHALID; KHALIL; NASREEN, 2014). Assim como em (REDDY et al., 2020), o objetivo é evidenciar que a redução da dimensionalidade não degrada o desempenho de algoritmos de AM. Todos os algoritmos apresentados nas

Subseções seguintes são descritos mais detalhadamente em (ANOWAR; SADAOUI; SELIM, 2021).

3.1.1 Análise de Componentes Principais

A técnica de análise de componentes principais (PCA, *principal component analysis*) é um algoritmo de transformação não supervisionado e linear que produz novas características chamadas de componentes principais, (*principal components*, PCs), resultando em um alto valor de variância correspondente àquele dado. O algoritmo PCA projeta a base de dados de alta dimensão em um novo espaço onde os eixos dessas novas características, ou PCs, são ortogonais e apresentam variância máxima. O algoritmo PCA constrói uma matriz $\mathbf{W}$, dk-dimensional, que mapeia uma matriz original $\mathbf{X}$, de dimensão d (colunas), em uma nova matriz $\mathbf{Y}$ de dimensão k ($k \leq d$).

A decomposição em autovalores e autovetores pode ser definida da seguinte forma, onde a matriz de covariância é decomposta em três outras matrizes:

$$\mathbf{X}\mathbf{X}^T \rightarrow \mathbf{B}\mathbf{D}\mathbf{B}^T \ , \tag{3.1}$$

em que $\mathbf{B}$ é a matriz quadrada ($d \times d$) que contém os autovetores; $\mathbf{D}$ é a matriz diagonal ($d \times d$) onde todos os elementos são zero, exceto os elementos da matriz diagonal, os quais representam os autovalores; $\mathbf{B}^T$ é a matriz transposta de $\mathbf{B}$.

Os passos de transformação da técnica PCA estão indicados no Algoritmo 3.

Algoritmo 3 Descrição dos passos da técnica PCA.

1: Construir a matriz de covariância ($\mathbf{X}.\mathbf{X}^T$).
2: Aplicar a decomposição em autovalores e autovetores da matriz de covariância (BROWN, 2018).
3: Dispor os autovalores em ordem decrescente para ordenar os autovetores.
4: Construir a matriz dk-dimensional $\mathbf{W}$ com os k principais autovetores.
5: Transformar a matriz $\mathbf{X}$ usando $\mathbf{W}$ para obter o novo subespaço $\mathbf{Y} = \mathbf{X}.\mathbf{W}$.

3.1.2 Decomposição em Valores Singulares

O algoritmo de decomposição em valores singulares (SVD, *singular value decomposition*) disponibiliza uma representação exata de um determinado conjunto de dados em uma matriz com um número qualquer de dimensões. Quanto menor o número de dimensões, menor será a precisão da ilustração desses dados. O algoritmo SVD pode ser representado pelo teorema

$$\mathbf{X} \rightarrow \mathbf{N}\mathbf{S}\mathbf{Z}^T \ , \tag{3.2}$$

em que:

- **X** representa a matriz original.
- **N** é uma matriz com colunas de vetores unitários ortogonais. Esta matriz respeita a seguinte definição:

$$\mathbf{N}\mathbf{N}^T = \mathbf{N}^T\mathbf{N} = \mathbf{I}\mathbf{N} \,, \tag{3.3}$$

  em que **I** é a matriz identidade.
- **S** é uma matriz diagonal $(k \times k)$ em que os elementos da diagonal são denominados como valores singulares.
- **Z** é uma matriz ortogonal ($\mathbf{Z}\mathbf{Z}^T = \mathbf{Z}^T\mathbf{Z} = \mathbf{I}$) de tamanho $(k \times d)$ conhecida como vetor singular à direita.

O Teorema (3.2) remonta a matriz de entrada **X** em uma matriz de dimensão menor **Y**, tal que

$$\mathbf{Y} = \mathbf{N}_m\mathbf{S}_m\mathbf{Z}_m^T \,, \tag{3.4}$$

em que $\mathbf{N}_m$, $\mathbf{S}_m$ e $\mathbf{Z}_m^T$ são versões truncadas de **N**, **S** e $\mathbf{Z}^T$, respectivamente. Para esta transformação, apenas os m melhores valores singulares são retidos em **Y**. Os passos da técnica SVD estão resumidos no Algoritmo 4.

Algoritmo 4 Descrição dos passos da técnica SVD.

1: Decompor a matriz **X** em três matrizes **S**, **N** e **Z**.
2: Construir a matriz **Y**, selecionando os m valores singulares principais.

3.1.3 Análise de Componentes Independentes

A técnica de análise de componentes independentes (ICA, *independent component analysis*) é um AEC linear que gera novas características que são estatisticamente independentes entre si. O que diferencia o algoritmo ICA dos outros AECs é que o ICA busca por características que são não-Gaussianas e estatisticamente independentes. Por exemplo, o algoritmo PCA busca por um subespaço que melhor represente os dados, porém o ICA busca dados que são independentes uns dos outros, em sua maioria.

Primeiramente, o algoritmo ICA decompõe a matriz original **X** tal que

$$\mathbf{X} \rightarrow \mathbf{A}\mathbf{S} \,, \tag{3.5}$$

em que **A** é a matriz de mixagem desconhecida e **S** é o coeficiente básico em que as características maximizam a independência mútua. Para obter a matriz de dimensão m,

o algoritmo ICA produz a matriz $\mathbf{Y}$ definida por

$$\mathbf{Y} = \mathbf{A}_m \mathbf{S}_m \,, \tag{3.6}$$

em que $\mathbf{A}_m$ e $\mathbf{S}_m$ são versões truncadas de $\mathbf{A}$ e $\mathbf{S}$, respectivamente. O Algoritmo 5 reúne as etapas da técnica ICA.

Algoritmo 5 Descrição dos passos da técnica ICA.

1: Decompor a matriz $\mathbf{X}$ nas matrizes $\mathbf{A}$ e $\mathbf{S}$.
2: Selecionar os m componentes independentes principais.
3: Construir a matriz $\mathbf{Y}$ usando os m componentes selecionados no passo anterior.

3.1.4 *Kernel* PCA

Uma vez que o algoritmo PCA não apresenta bom desempenho com dados não-lineares, a técnica *Kernel* PCA (KPCA) foi desenvolvida (JOLLIFFE; CADIMA, 2016). Assim, o algoritmo KPCA usa núcleos para projetar os dados em um espaço de maior dimensão. Desta forma, os dados tornam-se linearmente separáveis. Esses núcleos podem ser definidos como

$$K(x_a, x_b) = \phi(x_a)^T \cdot \phi(x_b) \,, \tag{3.7}$$

em que x_a e x_b são dois pontos arbitrários e a função $\phi(\cdot)$ mapeia os pontos x_a e x_b em um espaço de maior dimensão utilizando coordenadas polares. Nesta dissertação, são considerados os núcleos definidos por

$$K(x_a, x_b) = \begin{cases} x_a^T x_b \,, & \text{linear} \\ (\alpha x_a^T x_b + c_1)^{\Theta} \,, & \text{polinomial} \\ \exp\left(-\frac{|x_a - x_b|^2}{2\sigma^2}\right), & \text{RBF(\textit{radial basis function}) ou Gaussiano} \\ \tanh(x_a^T x_b + c_2) \,, & \text{sigmoide} \\ \exp\left(\frac{x_a^T x_b}{\|x_a\| \|x_b\|}\right), & \text{cosseno} \end{cases} \tag{3.8}$$

em que α é o coeficiente do polinômio, Θ é o grau do polinômio e c_1 e c_2 são constantes.

Com a aplicação da projeção com o núcleo selecionado, a matriz de entrada não-linear $\mathbf{X}$ é então transformada em uma matriz linear $\mathbf{X}'$. Desta forma, o método PCA é utilizado no novo espaço linearmente separável para reduzir a dimensionalidade. A técnica KPCA é descrita no Algoritmo 6.

Algoritmo 6 Descrição dos passos da técnica *Kernel* PCA.

1: Obter a matriz linearmente separável $\mathbf{X}'$ com o núcleo de mapeamento.
2: Aplicar o algoritmo PCA em $\mathbf{X}'$ para obter o subespaço reduzido Y (ver Algoritmo 3).

3.1.5 Mapeamento Isométrico

Escalas clássicas para calcular dissimilaridades entre dados, como a distância Euclidiana entre dois pontos, são geralmente usadas em métodos como PCA ou SVD, porém não consideram a distribuição da vizinhança desses pontos. Isso significa que os métodos de escala clássica não capturam padrões não-lineares em conjuntos de dados. Então um novo método, chamado de mapeamento isométrico (ISOMAP, *isometric mapping*), soluciona o problema mantendo a distância geodésica em pares no subespaço de dimensões reduzidas. O algoritmo ISOMAP é um AEC não-linear, não-supervisionado e baseado no algoritmo de escala multidimensional (MDS, *multi-dimensional scaling*) (COX; COX, 2008). O primeiro passo do algoritmo é gerar um grafo de vizinhança determinando a distância geodésica. No passo seguinte, o algoritmo ISOMAP calcula a distância mínima entre todos os pares (matrix d^G). No último passo, o algoritmo ISOMAP executa o algoritmo MDS no resultado do passo anterior. O Algoritmo 7 apresenta as etapas da técnica ISOMAP.

Algoritmo 7 Descrição dos passos da técnica ISOMAP.

1: Construir o grafo de vizinhança de X.
2: Calcular a distância geodésica e construir d^G.
3: Aplicar o algoritmo MDS em d^G para obter o novo subespaço Y (COX; COX, 2008).

3.2 TRABALHOS RELACIONADOS

Nesta Seção, são apresentados trabalhos relacionados à análise comparativa de AECs, em especial aqueles voltados para a aplicação de AECs em técnicas de localização de DMs em redes sem fio. Há uma gama de trabalhos que analisam qualitativamente alguns AECs, destacando vantagens e desvantagens de cada algoritmo, sem comparações quantitativas (RAMACHANDRAN; RAVICHANDRAN; RAVEENDRAN, 2020), (JOSHI; MACHCHHAR, 2014), (KHALID; KHALIL; NASREEN, 2014). Também há trabalhos que analisam os AECs do ponto de vista da visualização de dados, sem análise de aplicações práticas e cenários reais (TERUEL; CANOVAS; GARCIA, 2017),(AYESHA; HANIF; TALIB, 2020). De maneira a facilitar a comparação com a análise apresentada nesta dissertação, é interessante considerar a aplicação de AECs sobre bases de dados semelhantes às bases de dados aqui consideradas em termos de quantidade de características e registros.

Anowar et. al. desenvolveram uma comparação empírica dos principais AECs, incluindo todos aqueles aqui abordados (ANOWAR; SADAOUI; SELIM, 2021). O estudo inclui análises detalhadas de desempenho dos AECs, tais como comparações de acurácia, custo computacional, parâmetros utilizados em cada algoritmo e as principais limitações de cada um deles. Para a análise quantitativa, são considerados a acurácia da aplicação de AECs no algoritmo de máquina de vetores de suporte (SVM, *support-vector machine*), a

qual é gerada a partir do erro de classificação dos dados, e o desempenho dos AECs. A acurácia é medida por meio do teste estatístico *F-Score*, o qual calcula a proporção entre os verdadeiros positivos, os falsos negativos e o total de resultados positivos. O desempenho é calculado em *ms* e representa o tempo de execução do algoritmo. Outra métrica apresentada é a qualidade dos dados, calculada por meio da correlação entre os componentes gerados (em termos de valor-p). Os resultados indicam que o algoritmo KPCA obteve melhor acurácia na maioria dos casos envolvendo as três bases consideradas. Todas as três bases utilizadas são relacionadas a problemas de classificação, têm tamanhos diferentes (10.000, 561 e 96 registros) e são de domínio público. Os algoritmos ICA e PCA também obtiveram resultados semelhantes. Além disso, concluiu-se que, em alguns casos, as técnicas de classificação com aplicação de AECs obtiveram um aumento na acurácia em relação às técnicas de classificação sem a aplicação de AECs. Também notou-se que, em geral, os AECs não-lineares obtiveram melhores resultados que os AECs lineares. Apesar de Anowar et. al. terem desenvolvido um trabalho empírico bem completo, a análise foi aplicada apenas a problemas de classificação, não avaliando o impacto dos AECs em problemas de regressão. Também não foram apresentados resultados baseados em uma variação progressiva do número de componentes em conjunto com os valores de acurácia dos AECs, o que é de suma importância para a definição do fator ou percentual de redução em uma aplicação real.

Teruel et. al. propuseram a aplicação de alguns AECs, dentre eles o PCA, o LDA (*linear discriminant analysis*) e o t-SNE (*t-distributed stochastic neighbor embedding*) em problemas de localização *indoor* (TERUEL; CANOVAS; GARCIA, 2017). Embora a pesquisa inclua algoritmos lineares e não-lineares na análise, o trabalho não inclui alguns AECs não-lineares importantes como o KPCA e o ISOMAP, pois os autores justificam que esses algoritmos não são adequados para a aplicação em bases de dados pequenas (bases de 100 a 200 registros) e de alta dimensionalidade. A abordagem foca na visualização e separação dos registros em *clusters*, ou seja, em conjuntos agrupados e não faz uma análise de acurácia ou impacto no tempo de processamento após as aplicações do PCA, LDA e t-SNE. Face aos resultados expostos, concluiu-se que o algoritmo t-SNE superou os algoritmos PCA e LDA, uma vez que obteve melhor desempenho com dados esparsos. Porém, como já mencionado, seria interessante comparar os resultados obtidos pelos AECs analisados na pesquisa com outras técnicas não-lineares.

Qi et. al. também aplicaram alguns AECs em problemas de localização *indoor* (QI; JIN; YAN, 2018). O estudo propõe uma técnica de combinação de classificadores em conjunto com aplicação do algoritmo PCA. Foi feita uma comparação entre a técnica proposta e a técnica com classificador único, ambas com o propósito de realizar a predição do andar em que o DM se encontra. Neste caso, foram utilizadas redes neurais de apenas uma camada como classificadores. Algumas análises foram realizadas, como, por exemplo, o impacto da variação do tamanho do conjunto de treinamento na acurácia da localização, assim como a

variação do fator de redução da base, ou seja, do percentual de componentes utilizados em relação à acurácia de localização. É interessante ressaltar que a técnica proposta obteve melhores resultados do que a aplicação do classificador individual (com e sem redução de dimensionalidade) para fatores de redução a partir de 55%. Porém, o estudo não avaliou os impactos da aplicação do algoritmo PCA nos tempos de execução das fases *on-line* e *off-line* da localização *indoor*, além de focar apenas em problemas de classificação.

4 EXPERIMENTOS E RESULTADOS

Neste Capítulo, serão detalhadas as informações de cada zona de interesse de onde foram coletados os dados da rede celular, assim como os procedimentos de coleta e pré-processamento desses dados. Logo após, o processo de separação desses dados nas diferentes bases e cenários também é descrito em detalhes e, por fim, os resultados relativos a esses cenários são apresentados e discutidos.

4.1 AMBIENTES INVESTIGADOS

A disposição do sinal de RF em redes móveis celulares pode variar de acordo com a densidade de ERBs em determinados tipos de ambientes. Por exemplo, zonas com alto nível de urbanização são ambientes com densidade demográfica elevada, infraestrutura de redes de comunicação abrangente e um número maior de construções (prédios e casas). Tal ambiente será denotado por Z-ANU, acrônimo de zona com alto nível de urbanização, deste ponto em diante. Em contrapartida, as zonas com baixo nível de urbanização apresentam densidade demográfica reduzida em relação às Z-ANUs, infraestrutura de redes de comunicação que não abarca todo o território e, por fim, um número reduzido de construções. De modo equivalente, utilizaremos a sigla Z-BNU para nos referirmos a uma zona com baixo nível de urbanização. Os diferentes níveis de urbanização definidos anteriormente possuem influência direta na propagação e disponibilidade dos sinais de RF nos ambientes mencionados.

Com o intuito de analisar a disposição das redes móveis celulares nos ambientes definidos anteriormente, algumas regiões com diferentes níveis de urbanização serão consideradas neste trabalho. Duas regiões de interesse das cidades de Recife-PE e Sirinhaém-PE foram escolhidas para representar as Z-ANUs e Z-BNUs, respectivamente. A Tabela 1 apresenta algumas características geográficas e populacionais das duas cidades mencionadas. Tais características estão diretamente relacionadas com a distribuição de ERBs e disponibilização de sinais de RF nas regiões de interesse. Além disso, os valores de densidade demográfica apresentados evidenciam a escolha de Recife como uma Z-ANU e de

Tabela 1 – Características das regiões com diferentes níveis de urbanização consideradas neste trabalho (IBGE, 2021).

	Recife (Z-ANU)	Sirinhaém (Z-BNU)
Área total (km^2)	218,43	374,32
População total (hab.)	1.645.727	46.361
Densidade demográfica (hab./km^2)	7.534,2	123,9

Fonte: Elaborada pelo autor (2021)

Sirinhaém como uma Z-BNU.

Pode-se perceber que, embora a cidade de Sirinhaém tenha uma área 71% maior que a área da cidade de Recife, esta tem uma população que representa apenas 2,82% da população total da Z-ANU escolhida. Isto faz com que a densidade demográfica de Sirinhaém seja de apenas 123 habitantes por quilômetro quadrado, enquanto a densidade do Recife chega a aproximadamente 7.534 habitantes por quilômetro quadrado, um número cerca de 61 vezes maior.

Segundo dados da ANATEL, a cidade do Recife tem cobertura de sinais de rede móvel celular em 99,7% do território total, enquanto a cidade de Sirinhaém apresenta apenas 63,9% do seu território coberto (ANATEL, 2021). Tal percentual pode ser explicado pela grande área inabitada do município, característica comum entre as Z-BNUs. Outro dado representativo importante é o percentual de moradores e domicílios cobertos por sinais de redes de telefonia celular em cada uma das cidades. Ainda segundo a ANATEL, enquanto Recife possui cobertura de 100% dos moradores e domicílios da cidade, Sirinháem se limita a ter 90,5% dos moradores e 91,3% dos domicílios cobertos pelas redes, deixando parte da população sem cobertura de sinal.

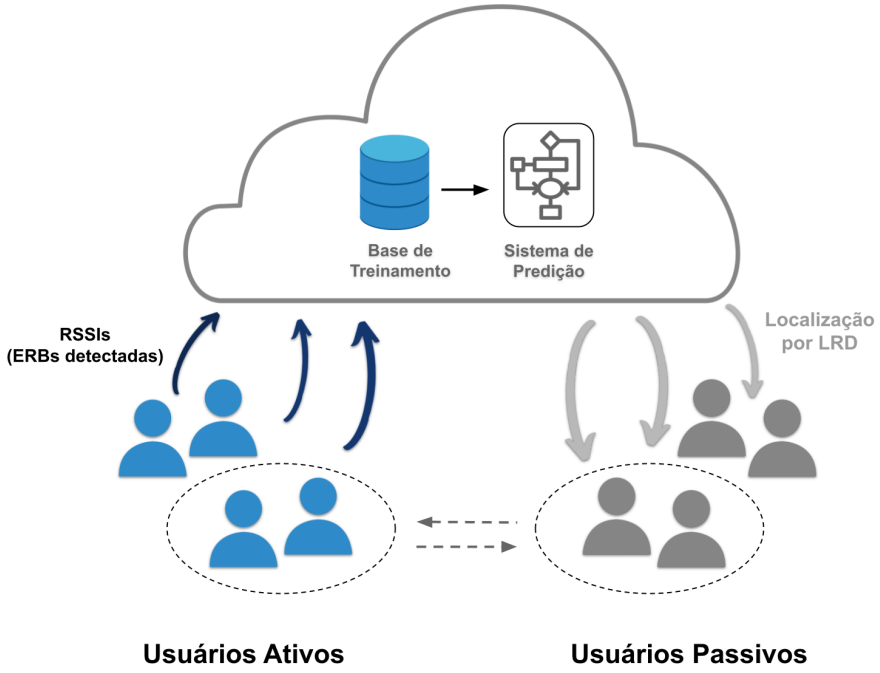
Face ao exposto, os usuários móveis das Z-ANUs tendem a ser mais ativos que os usuários móveis das Z-BNUs. Esta percepção é reforçada por um indicador chamado de densidade de telefonia móvel (DTM), ou seja, a quantidade de acessos a cada 100 habitantes. A DTM referente à cidade de Sirinhaém é de 66,2 acessos por cada 100 habitantes, enquanto a DTM do Recife é de 147,2 acessos por cada 100 habitantes (ANATEL, 2021). Tais valores mostram que, de fato, existe maior tráfego de dados nas redes celulares da cidade do Recife, em virtude do seu maior nível de urbanização. Essas informações serão úteis nas análises dos resultados apresentados no decorrer do trabalho.

4.2 COLETA DE DADOS

O processo de coleta de dados se deu por meio da simulação do procedimento de construção de bases de dados do tipo contribuição coletiva, ou *crowdsourcing* (LECA et al., 2017). A Fig. 3 ilustra o diagrama que representa um sistema de localização baseado em AM, com destaque para o procedimento de construção de uma base de dados de treinamento *crowdsourcing* a partir da contribuição dos usuários ativos. Por definição, usuários ativos são aqueles que fornecem informações (em nosso caso, os RSSIs medidos a partir de cada ERB da região de interesse) que serão usadas como características. Além disso, uma vez que os usuários ativos estão sempre com o GPS ligado, as coordenadas geográficas são coletadas como valores alvos para uso em um modelo de aprendizagem supervisionada. Desta forma, os modelos preditivos (sistema de predição) são beneficiados com o fornecimento de informações de maneira contínua e progressiva.

Há um segundo tipo de usuário, chamado de usuário passivo, que usufrui da predição

Figura 3 – Diagrama representativo de um sistema de localização baseado em aprendizado de máquina com ênfase na construção da base de treinamento via *crowdsourcing*.



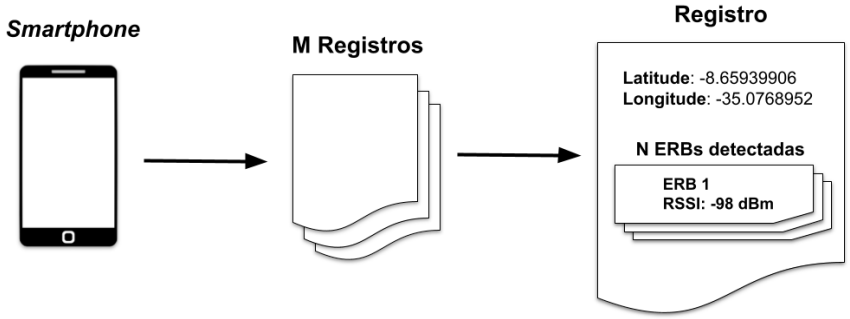
Fonte: Elaborada pelo autor (2021)

do sistema de localização. Um ponto importante a se observar no cenário ilustrado na Fig. 3 é que os usuários passivos podem eventualmente se tornar usuários ativos e vice-versa. Com isso, fica evidente que o procedimento de construção da base de dados via *crowdsourcing* é dinâmico e retroalimentado. Nesta pesquisa, a construção da base de dados se deu pela utilização de um DM de uso pessoal (*smartphone*), que simulou o comportamento de diversos usuários ativos.

A coleta dos dados oriundos das redes celulares presentes na Z-ANU e na Z-BNU foi realizada por meio de uma aplicação para sistemas *Android*. Um *smartphone Android* Motorola Moto G5 Plus com dois *chips* e sistema operacional *Android* 8.1 foi usado para coletar os sinais de RF provenientes das ERBs. Este DM é compatível com as bandas de frequência das redes celulares de 2G, 3G e de 4G. Por ser um celular com dois *chips*, ou seja, que permite o escaneamento de redes de duas operadoras de telefonia móvel simultaneamente, foram consideradas duas operadoras de telecomunicações ativas e com cobertura nas duas cidades, quais sejam, a TIM e a Oi. O sistema GPS do *smartphone* permaneceu ativo durante toda a coleta, em virtude da necessidade de preenchimento dos valores alvos para treinamento e para o cálculo da acurácia da predição.

Dentre os diversos dados disponibilizados pela plataforma *Android*, apenas alguns são

Figura 4 – Processo da coleta de dados e detalhamento dos conteúdos dos registros captados.



Fonte: Elaborada pelo autor (2021)

de interesse desta pesquisa. Como entrada do método LRD, somente são necessários os RSSIs de cada ERB e as coordenadas geográficas do registro. A cada registro efetuado, ou seja, a cada medição realizada, os dados são coletados e armazenados em um arquivo que será posteriormente pré-processado e usado no treinamento e predição da localização.

A Fig. 4 ilustra o processo de coleta dos dados, assim como o detalhamento do conteúdo armazenado em registros. A base de dados coletada pelo *smartphone* é formada por M registros. Cada registro contém as coordenadas geográficas do ponto de coleta, em graus decimais, e os RSSIs das ERBs detectadas. Para cada uma das N ERBs sensoriadas, um valor de RSSI é armazenado. A taxa usada para a captura dos dados foi de um registro por segundo.

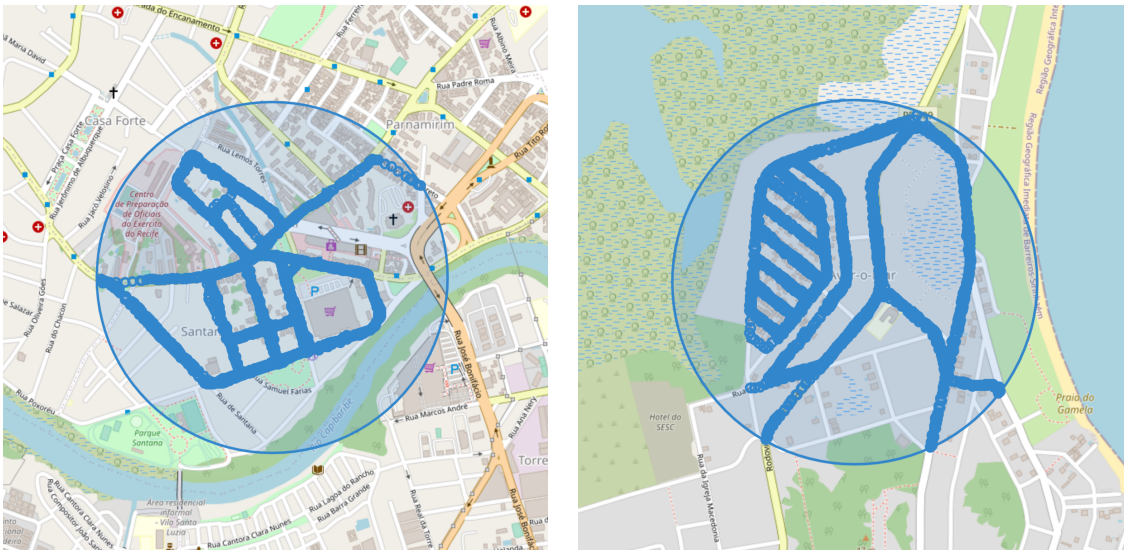
4.2.1 Pré-processamento

A etapa de pré-processamento possui dois objetivos principais neste trabalho. O primeiro é eliminar dados ruidosos que possam interferir negativamente na análise dos dados. O segundo objetivo é deixar as bases referentes às duas regiões em iguais condições para que a comparação entre elas seja justa e o mais transparente possível.

É preciso ressaltar que a taxa de atualização do RSSI para cada ERB pode variar. Desta forma, é possível que cada registro coletado possua valores repetidos que podem não corresponder com o valor real naquelas determinadas coordenadas. Apesar da taxa de amostragem ser fixa para o experimento (1 registro/s), os registros sequenciais que contêm os mesmos conjuntos de ERBs e os mesmo valores de RSSI para cada ERB foram removidos na fase de pré-processamento. Por outro lado, uma determinada ERB pode ter dois ou mais valores de RSSI detectados no mesmo registro. Nesta situação, a média dos valores de RSSI sensoriados é adotada como valor final para aquela ERB naquele registro.

Problemas com o sensoriamento das ERBs também podem acontecer, como, por exem-

Figura 5 – Ilustração das coordenadas dos dados coletados (em cor azul) nos mapas das regiões de interesse: (a) Z-ANU: Recife. (B) Z-BNU: Sirinhaém.



(a) (b)

Fonte: Elaborada pelo autor (2021)

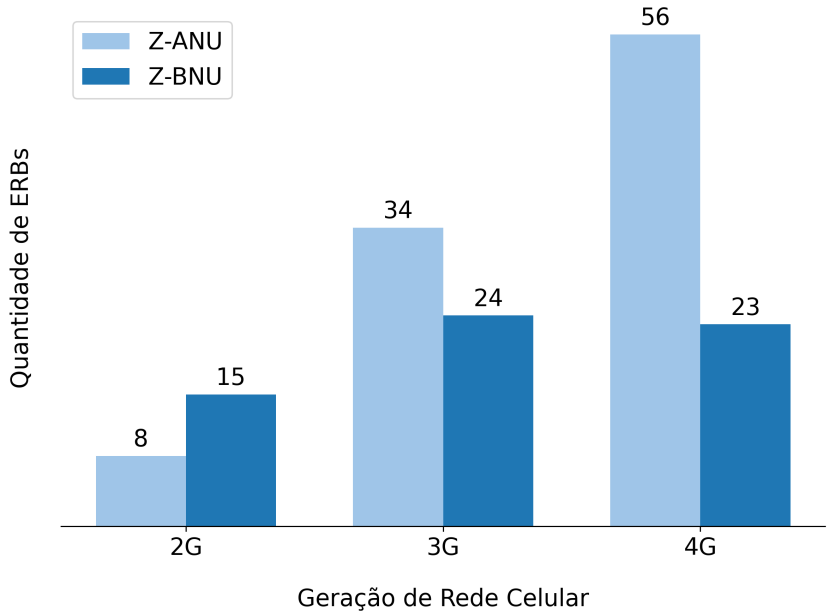
plo, a obtenção de código da célula ou código da rede móvel inválidos durante a identificação da ERB. Outra dificuldade é a captura do valor de RSSI, que, por vezes, pode ser inválido ou mesmo zerado. Todas essas ocorrências foram detectadas e removidas das bases de dados finais.

Ao final do processo de remoção de ERBs e registros inválidos, uma nova verificação é realizada e são retirados todos os registros que não contém nenhuma ERB no seu vetor de RSSI. Após a fase de pré-processamento, os dados foram contabilizados e analisados.

4.2.2 Resultado da Coleta de Dados

A coleta de dados foi realizada no decorrer de vias e ruas das duas regiões de interesse, quais sejam, Z-ANU e Z-BNU. Então, a escolha das regiões de coleta teve como base a semelhança da disposição destas vias e ruas, a fim de promover uma comparação justa. Desta forma, duas áreas específicas foram selecionadas em cada uma dessas regiões para a coleta de dado. As Figs. 5(a) e 5(b) ilustram as coordenadas dos dados coletados (em cor azul) nos mapas das regiões de interesse das duas cidades. Na cidade de Sirinhaém, foi escolhida uma área plana distante do centro da cidade, localizada em um povoado chamado Aver-o-mar. Neste local, foram coletados 1.920 registros em um raio pré-definido de $500\,m$. Na cidade do Recife, o objetivo foi encontrar uma área com topologia semelhante à topologia apresentada no povoado de Aver-o-mar. Por este motivo, a área selecionada para a coleta dos dados fica localizada entre os bairros de Casa Forte e Santana. Neste local, também foram coletados 1.920 registros no mesmo raio de $500\,m$.

Figura 6 – Quantidade de ERBs em cada uma das gerações de rede celular por região de interesse.



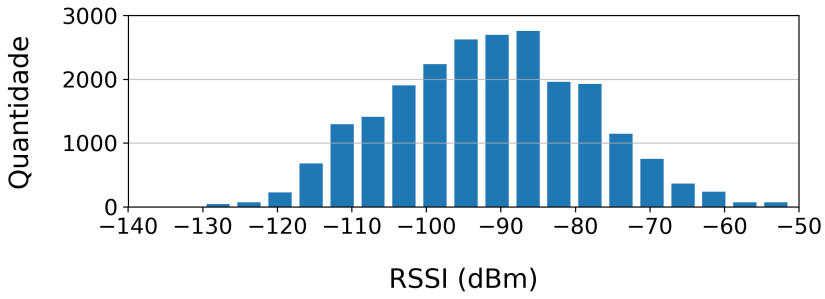
Fonte: Elaborada pelo autor (2021)

A Fig. 6 apresenta a quantidade de ERBs detectadas por geração de rede celular nos dois ambientes. No total, foram detectadas 98 ERBs [1] na Z-ANU, envolvendo as duas operadoras consideradas. Das 98 ERBs, apenas oito são da rede 2G, 34 são da rede 3G e 56 são da rede 4G. Dessa forma, as ERBs 4G são predominantes na Z-ANU, enquanto as ERBs 2G estão praticamente escassas. Em contrapartida, na Z-BNU, foram detectadas 62 ERBs no total, um número 36,7% menor que a quantidade de ERBs detectadas na Z-ANU. Dessas 62 ERBs, 15 são da rede 2G, 24 delas são da rede 3G e, por fim, 23 são da rede 4G.

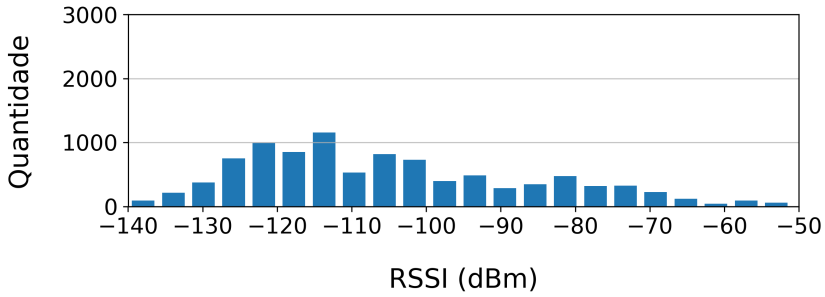
Apesar da Z-BNU apresentar uma quantidade menor de ERBs que a Z-ANU, o sensoriamento indicou a detecção de um número maior de ERBs 2G na Z-BNU em comparação com a Z-ANU. Este fato está ligado à infraestrutura de rede de melhor qualidade estar comumente disponível nas Z-ANUs em detrimento das Z-BNUs, tendo estas uma predisposição a levar mais tempo para atualizar as tecnologias disponíveis (HONG; THAKURIAH, 2018). Comparando a quantidade de ERBs 2G e 4G na Z-BNU, observamos que o número de ERBs 4G é 50% maior que o número de ERBs 2G. Se fizermos este mesmo comparativo na Z-ANU, as ERBs 4G superam as ERBs 2G em 700%. Também observamos que enquanto a quantidade de ERBs apresenta uma progressão à medida que a geração da rede celular evolui na Z-ANU, na Z-BNU isso até acontece na passagem da segunda para a terceira geração, mas logo depois apresenta um platô na passagem da terceira para a

[1] Embora as ERBs sejam identificadas de forma diferente nas redes 2G (BTS, *base transceiver station*), 3G (Node B) e 4G (e-Node B), neste documento, a sigla ERB será usada em todos os casos.

Figura 7 – Histogramas dos RSSIs coletados em relação a cada ERB nas duas regiões de interesse. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



(a)



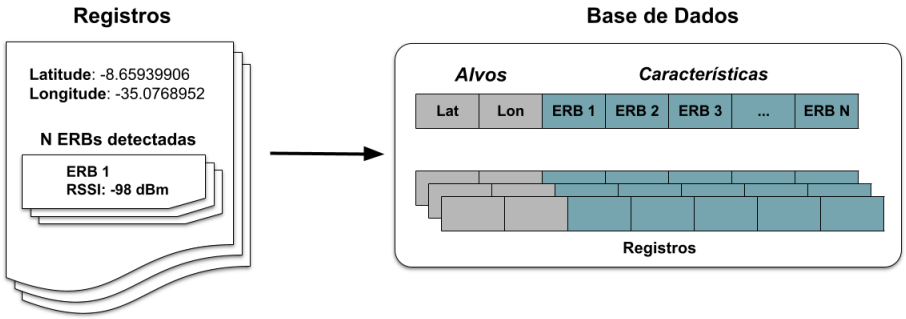
(b)

Fonte: Elaborada pelo autor (2021)

quarta geração.

A quantidade de ERBs detectadas pode variar entre os registros da base de dados. A média de ERBs detectadas por registro, ou seja, a quantidade média de ERBs que cada registro consegue sensoriar impacta diretamente no desempenho da regressão, pois quanto maior a quantidade de sinais detectados por registro, menor a quantidade de valores faltantes que precisarão ser preenchidos com algum valor padrão. Esses valores faltantes, também chamados de *missing values*, ocorrem quando o DM não consegue detectar RSSIs relativos às ERBs disponíveis na região. Para o preenchimento desses valores faltantes em base de dados compostas por RSSIs, usa-se o menor valor de RSSI detectado na base (ANAGNOSTOPOULOS; KALOUSIS, 2019). As Figs. 7(a) e 7(b) ilustram os histogramas dos RSSIs para as regiões Z-ANU e Z-BNU, respectivamente. É possível perceber que o menor RSSI detectado na Z-ANU foi de -130 dBm, enquanto na Z-BNU, o menor RSSI foi de -140 dBm. Dessa forma, esses foram os valores utilizados para preencher os RSSIs faltantes das duas bases, respectivamente. Continuando a análise dos histogramas, é interessante ressaltar que, na Z-BNU, a maioria dos RSSIs detectados está situada em torno de -120 dBm, enquanto na Z-ANU, este valor é por volta de -90 dBm. Isso indica que as ERBs na Z-ANU estão mais próximas do DM , pois os valores detectados são maiores, e as ERBs na Z-BNU estão mais distantes do DM devido à quantidade alta de RSSIs próximos de -120 dBm.

Figura 8 – Representação da transformação dos registros coletados em uma base de dados unificada no formato de entrada do método LRD.



Fonte: Elaborada pelo autor (2021)

Os registros coletados na Z-ANU detectam, em média, 11,7 ERBs das 98 disponíveis por registro. Isso significa que, para cada registro, por volta de 12% das ERBs disponíveis são detectadas. Na Z-BNU, cada registro detecta, em média, cinco ERBs das 62 disponíveis no total, o que remete a aproximadamente 8% das ERBs. Deste modo, cada registro na Z-ANU conseguiu detectar em média o dobro de ERBs detectadas pelos registros da Z-BNU, o que significa que uma quantidade menor de valores padrão será inserida na base de dados da Z-ANU. Essa característica impacta diretamente na acurácia da localização.

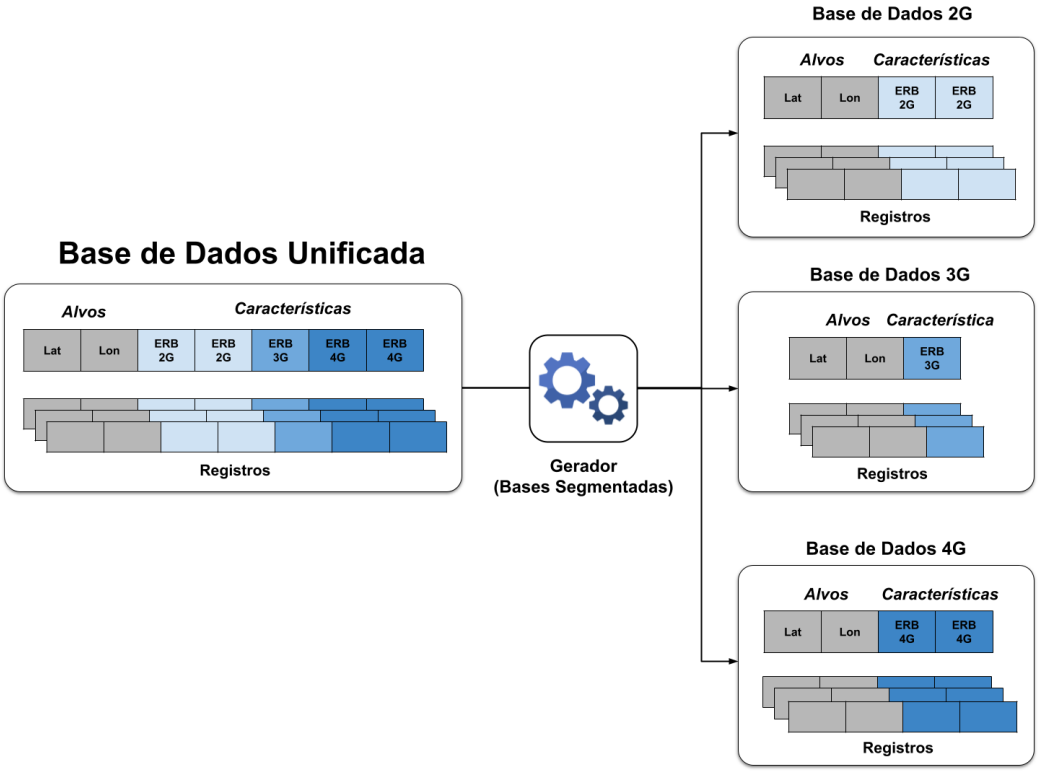
4.2.3 Geração das Bases de Dados

Para a composição das bases que foram utilizadas no método LRD, primeiramente foi preciso transformar as ERBs e os valores de RSSI associados a essas ERBs em características. A característica é composta por um rótulo e por um conjunto de valores. No caso da base de entrada para o método LRD, o rótulo de cada característica é formado pela descrição da ERB, ou seja, pela identificação da célula concatenada à geração da ERB, e os RSSIs detectados relacionados àquela ERB compõem o conjunto de valores da característica. De forma análoga à composição das características, as coordenadas geográficas são transformadas em valores alvos nesta base de entrada.

O experimento foi desenvolvido em duas fases. Na primeira fase, foi realizado um comparativo entre a localização do DM nas regiões considerando uma base unificada, ou seja, usando apenas uma base de treinamento contendo os dados de todas as três gerações de rede celular disponíveis nas regiões. Por fim, na segunda fase do experimento, foi obtida a predição da localização do DM por geração de rede celular. Assim, foi possível simular a situação em que o DM do usuário está apto a operar em apenas uma das gerações de rede celular disponíveis.

Por definição, denominamos de base unificada aquela gerada a partir de todos os

Figura 9 – Representação da segmentação da base de dados unificada em três bases secundárias, cada uma contendo informações de uma rede celular específica (2G, 3G ou 4G).



Fonte: Elaborada pelo autor (2021)

dados envolvidos no sensoriamento, incluindo as características formadas por todos os RSSIs e ERBs detectadas. As coordenadas geográficas obtidas a partir do sinal de GPS do DM compõem as duas primeiras colunas. Os valores que compõem essas colunas são considerados valores alvos. As demais colunas, também chamadas de características, são formadas pelos valores de RSSI captados em relação a cada uma das ERBs. A Fig. 8 indica o processo de transformação dos registros colhidos em campo em uma base de dados unificada no formato adequado à entrada do método LRD. Como descrito na Subseção 4.2.2, os valores faltantes são preenchidos com o valor padrão predefinido.

De maneira semelhante à construção da base de dados unificada, as bases segmentadas por geração de rede celular também contêm previamente todos os registros sensoriados. Todavia, as características serão separadas em três bases de dados secundárias pelo gerador de bases segmentadas, uma para cada tipo de rede celular. A Fig. 9 ilustra o processo de segmentação da base de dados unificada em três bases secundárias. O gerador replica todos os registros da base unificada nas três bases segmentadas e remove as colunas que não pertencem a uma determinada base secundária, ou seja, remove características que representam dados de ERBs de outras gerações de rede celular. Por exemplo, para gerar a base de dados 3G, o gerador, após replicar os dados da base unificada, retira as

características referentes às redes 2G e 4G. Esta configuração permitiu o treinamento de três instâncias paralelas do método LRD referentes aos três tipos de gerações de rede celular disponíveis. Por fim, os registros que não apresentam nenhum valor real de RSSI são removidos das respectivas bases de dados. É importante ressaltar que, ao final do processo, as bases de dados referentes à cada geração de rede celular podem ter quantidades de registros diferentes uma das outras. Complementarmente, as bases geradas não são mutuamente exclusivas, isto é, há a replicação de registros durante o processo de geração.

4.3 RESULTADOS

Neste trabalho, o desempenho do método LRD, definido na Seção 2.1, é avaliado por meio de simulações computacionais usando a linguagem Python com ênfase na biblioteca *scikit-learn* (HACKELING, 2014). O método LRD realiza a predição das coordenadas geográficas do DM baseando-se nos valores de RSSI que ele captura. Para isso, são consideradas duas regiões com diferentes níveis de urbanização, sendo a primeira com alto nível e denotada por Z-ANU, enquanto a segunda é de baixo nível e denotada por Z-BNU. As características de ambas as regiões foram descritas na Seção 4.1. O regressor utilizado no método LRD foi o algoritmo k-NN. No processo de treinamento deste algoritmo, foi aplicado o processo de *tuning* no parâmetro k (quantidade de vizinhos) por meio de uma busca exaustiva no conjunto de valores $\{1, 3, 5, 7, 9, 11\}$. O valor de k responsável pelo menor erro em um determinado conjunto de validação (subconjunto de registros do conjunto de treinamento) foi utilizado para treinamento do método LRD. Em cada execução do método LRD, dois regressores k-NN são utilizados, um para predição da latitude e outro para a predição da longitude. Para cada um desses regressores, um determinado valor de k foi assumido para cada execução do método LRD. O método LRD foi vastamente executado para obtenção dos resultados desta dissertação e os valores de k assumidos estão contidos no conjunto de valores $\{1, 3, 5, 7, 9, 11\}$.

Com o intuito de calcular o erro obtido nessa predição, foi utilizado o algoritmo proposto em (KARNEY, 2013), que adota como métrica a distância geodésica em superfície elipsoide. Este algoritmo calcula a distância, em metros, entre dois pontos. Neste trabalho, assumimos que um dos pontos é dado pelas coordenadas geográficas reais e o outro ponto representa as coordenadas geográficas preditas. Logo, a distância obtida é considerada o erro de predição e será denotada por d_g.

Para avaliar o método LRD, iremos considerar o erro médio obtido na predição dos registros da base de dados de teste. Este erro médio representa a acurácia do método LRD obtida em determinado cenário, que é caracterizado por região (Z-ANU ou Z-BNU), base de dados (unificada ou segmentada) e pela aplicação ou não de AECs. Para analisar o esforço computacional do método LRD, o tempo de processamento de cada fase (*on-line* e *off-line*) do algoritmo de regressão será medido.

Tabela 2 – Análise estatística do método LRD aplicado às regiões Z-ANU e Z-BNU para a base unificada de RSSIs (dados das redes 2G, 3G e 4G em conjunto).

	$\bar{\mu}\,(m)$	$\sigma\,(m)$	$\mu_{min}\,(m)$	$\mu_{max}\,(m)$
Z-ANU	16,24	26,27	0,001	297,87
Z-BNU	80,51	97,97	0,017	585,86

Fonte: Elaborada pelo autor (2021)

Para calcular a acurácia do método LRD nas regiões de interesse, definimos o erro médio absoluto (EMA), denotado por $\bar{\mu}$, tal que

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} d_g(y_i, p_i)\,, \tag{4.1}$$

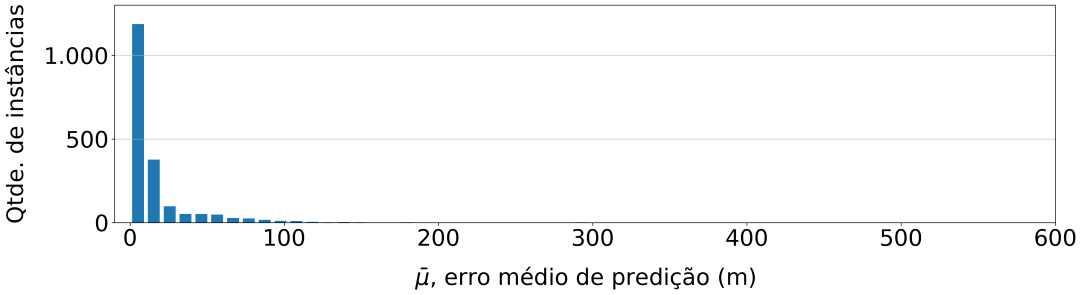
em que N é a quantidade de instâncias do conjunto de teste empregado na fase *on-line*, $d_g(\cdot,\cdot)$ é o erro de predição de cada instância, y_i representa as coordenadas geográficas (latitude a longitude) reais da i-ésima instância do conjunto de testes e, finalmente, p_i corresponde às coordenadas geográficas (latitude a longitude) preditas da i-ésima instância do conjunto de testes. É importante ressaltar que, neste trabalho, o EMA é considerado a métrica de acurácia de localização, porém não deve ser confundida com a acurácia relacionada diretamente aos valores alvos da aplicação dos algoritmo de AM. Esta acurácia relativa à divergência entre os valores alvos e os valores preditos das coordenadas geográficas não é avaliada de forma direta na apresentação dos resultados.

Para reduzir o impacto da aleatoriedade na divisão da base de dados em conjuntos de treinamento e teste, a técnica de validação cruzada com o uso do método K-fold é geralmente empregada (REFAEILZADEH; TANG; LIU, 2009). A validação cruzada tem o propósito de minimizar o impacto da aleatoriedade do particionamento dos registros entre os conjuntos de treinamento e teste. Na técnica de validação cruzada, a base de dados é dividida em K partes iguais. O conjunto de treinamento é constituído pelas primeiras $(K-1)$ partes, enquanto o conjunto de teste corresponde à parte restante. Este processo é repetido K vezes, alternando sempre o conjunto de teste por um dos K conjuntos que ainda não foi empregado como conjunto de teste. A acurácia final corresponde à média do erro das K iterações (YADAV; SHUKLA, 2016). Neste trabalho, foi escolhido o valor $K = 10$ para cada execução do algoritmo. Em outras palavras, para todos os cenários disponíveis, incluindo os cenários que aplicam as técnicas descritas no Cap. 3, os algoritmos de regressão foram executados dez vezes.

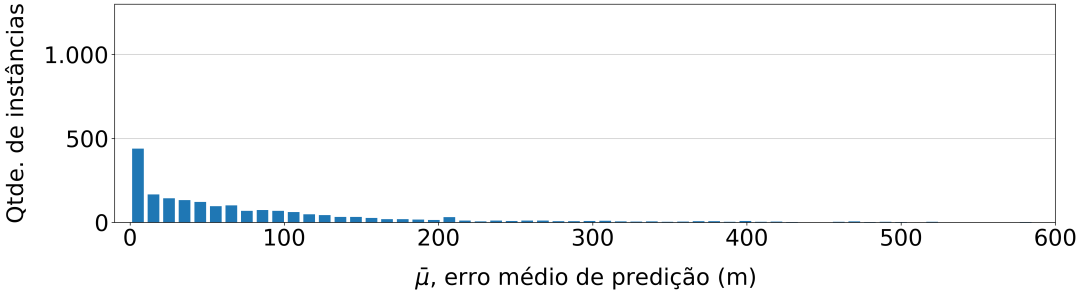
4.3.1 Localização por Regressão Direta em Base Unificada

A Tab. 2 apresenta a análise estatística do método LRD aplicado às regiões Z-ANU e Z-BNU considerando a base unificada (dados das redes 2G, 3G e 4G em conjunto). A

Figura 10 – Histograma do erro médio de predição do método LRD para as bases unificadas das duas regiões. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



(a)



(b)

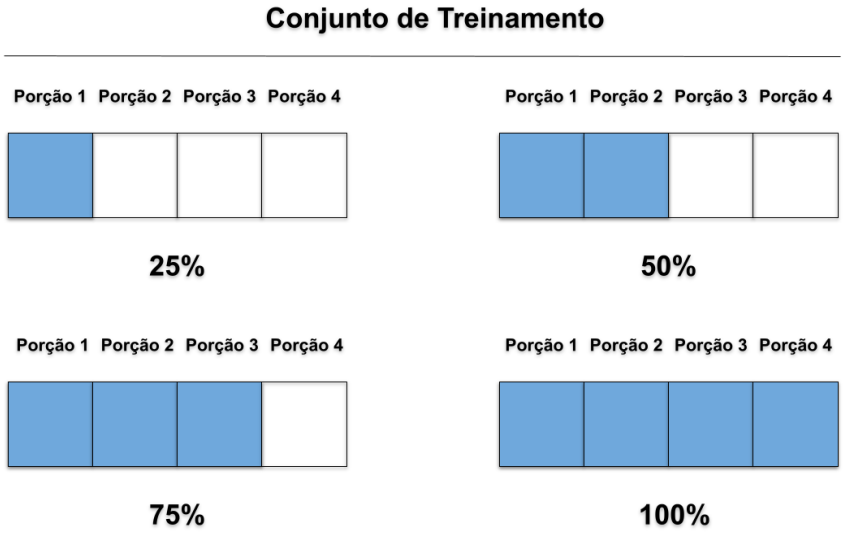
Fonte: Elaborada pelo autor (2021)

análise foi conduzida considerando todas as ERBs detectadas pelo DM nas duas regiões. Os parâmetros apresentados são o erro médio de predição $\bar{\mu}$, o desvio padrão σ, além dos erros mínimo e máximo, representados por μ_{min} e μ_{max}, respectivamente.

A partir das informações apresentadas na Tab. 2, temos que o erro médio obtido na região Z-ANU é aproximadamente cinco vezes menor do que o erro médio obtido na região Z-BNU. Tal fato é uma consequência direta do número de ERBs presentes em cada região e da quantidade de ERBs que o DM consegue detectar em cada coordenada geográfica. Como apresentado na Subseção 4.2.2, os registros da Z-ANU possuem, em média, o dobro das ERBs que os registros da Z-BNU. Analisando o desvio padrão obtido nas duas regiões, é interessante observar que o desvio padrão para a Z-ANU é maior que o da Z-BNU. Deste modo, as predições na Z-BNU se mostram mais instáveis e divergentes. Esse fato pode ser percebido observando os valores de erro máximo, onde o método LRD obteve 585,86 m de erro máximo quando aplicado na Z-BNU, e 297,87 m, quando aplicado na Z-ANU.

As Figs. 10(a) e 10(b) indicam os histogramas do erro médio de predição do método LRD aplicado às regiões Z-ANU e Z-BNU, respectivamente, para as bases unificadas de RSSIs. Nele, podemos notar que a distribuição dos dados da Z-ANU concentra a sua maioria abaixo dos 50 m de erro, enquanto a distribuição dos dados da Z-BNU dispõe os dados numa decrescente constante até por volta dos 200 m de erro. Esses resultados mostram que o método LRD apresentou melhor desempenho na Z-ANU, onde obteve

Figura 11 – Representação do fracionamento dos conjuntos de treinamento e das quatro etapas do experimento.



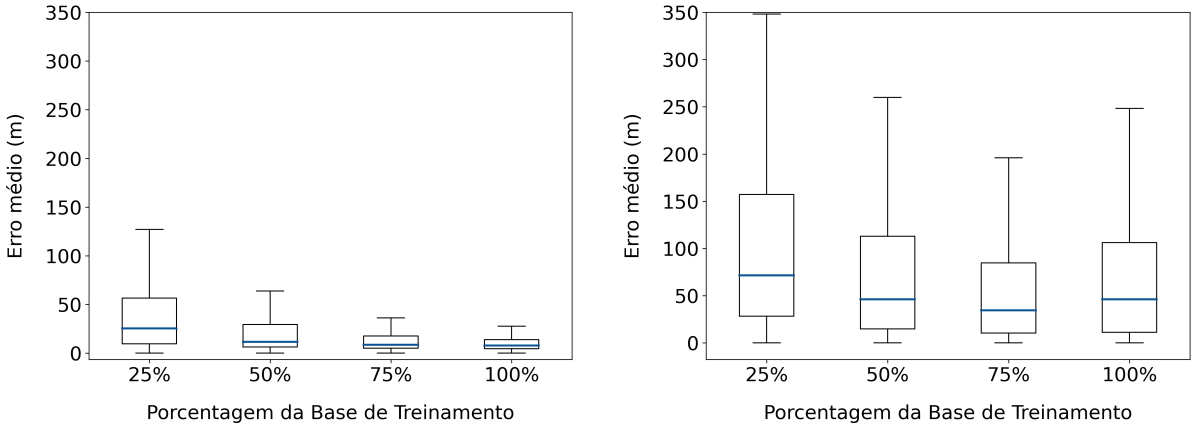
Fonte: Elaborada pelo autor (2021)

um erro médio e desvio padrão de 16,24 m e 26,27 m, respectivamente. Esse valores são menores que os apresentados em (OLIVEIRA et al., 2019), que foram de 34,90 m de erro médio e 33,38 m de desvio padrão. Enquanto na Z-BNU, os valores de erro médio e desvio padrão obtidos pelo método LRD foram de 80,51 m e 97,97 m, valores acima dos apresentados em (OLIVEIRA et al., 2019).

É importante considerar a variação da disponibilidade dos dados para treinamento, uma vez que estamos assumindo a construção de bases por meio de *crowdsourcing*. Para analisar o impacto da variação da quantidade dos dados nas duas regiões consideradas, o conjunto de treinamento foi fracionado em quatro porções e, a cada execução do experimento, os dados de uma porção foram adicionados cumulativamente ao conjunto de treinamento final. A Fig. 11 ilustra a divisão dos conjuntos de treinamento em cada execução do experimento. Deste modo, a primeira execução usou apenas uma porção, ou seja, 25% dos dados disponíveis para treinamento. Na segunda execução, foram usadas duas das quatro porções disponíveis, ou seja, 50% dos dados para treinamento. Da mesma forma, as terceira e quarta execuções do experimento usaram, respectivamente, três (75% dos dados disponíveis) e todas as quatro porções (100% dos dados disponíveis). Vale ressaltar que a divisão desses dados é feita de forma aleatória e que a técnica de validação cruzada foi realizada para cada execução do experimento. O tamanho do conjunto de testes foi definido em 25% dos dados disponíveis para treinamento (uma porção) e permaneceu o mesmo para todas as quatro execuções deste experimento a fim de proporcionar uma comparação justa dos resultados.

Além de fazer a análise do erro médio para cada uma das execuções do experimento

Figura 12 – Diagramas de caixa do erro médio do método LRD em função do tamanho do conjunto de treinamento empregado em cada região. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



(a) (b)

Fonte: Elaborada pelo autor (2021)

previamente definidas, é importante avaliar também o desvio padrão deste erro, o qual indica o quão preciso o método LRD pode ser em determinado cenário. Outro aspecto relevante é observar o valor da mediana, que auxilia na análise de dados distorcidos, ou seja, quando há o acúmulo de dados em um dos lados da mediana. Caso a mediana não esteja próxima ao centro da caixa, pode-se atestar uma assimetria nos erros médios obtidos e é possível concluir que o método apresenta inconstância ao realizar predições. Para isto, a Fig. 12 apresenta os diagramas de caixa para cada um das quatro execuções do experimento definidas anteriormente (25%, 50%, 75% e 100% dos dados de treino), em cada uma das regiões consideradas (Z-ANU e Z-BNU).

Com o diagrama de caixa, podemos analisar a estabilidade do método LRD observando o tamanho da caixa de cada uma das execuções. Quanto menor o tamanho da caixa, maior é a estabilidade do método. Outra característica importante é a distância entre os limites inferior e superior. Quanto maior a distância, mais esparsos serão os valores dos erros obtidos na localização, tornando o método mais instável naquele cenário. Observando os dados da Z-ANU, notamos que à medida que aumentamos o tamanho do conjunto de treinamento, o método vai se tornando mais estável, pois os tamanhos das caixas diminuem, e a mediana apresenta um comportamento decrescente em relação a todas as quatro execuções. Outro fato que consolida isto é que para a execução com 25% do conjunto de treinamento, temos um erro médio próximo a $40\,m$, enquanto para 100% temos um erro médio próximo a $16\,m$. Para a região Z-BNU, assim como na Z-ANU, também nota-se um comportamento decrescente no erro médio obtido pelo método LRD nas três primeiras execuções. Em contraste com o comportamento decrescente do erro médio de predição nas quatro execuções do experimento na Z-ANU, nas execuções com 75% e

Tabela 3 – Métricas referentes à aplicação do método LRD nas duas regiões com diferentes bases de treinamento e predição, cada uma representando um tipo de rede celular, além da base unificada, que inclui dados de todas as redes. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.

Rede	η_{ERB}	η_{base}	$\bar{\mu}_{ERB}$	$\bar{\mu}_{RSSI}$ (dBm)	P_{reg}	$\bar{\mu}$ (m)	σ (m)
Todas	98	1920	11,69	-91,47	100%	16,24	26,27
Rede 2G	8	74	2,91	-80,84	3,85%	23,70	51,65
Rede 3G	34	1892	5,90	-84,51	98,54%	18,84	29,77
Rede 4G	56	1919	5,75	-98,72	99,99%	23,60	41,86

(a)

Rede	η_{ERB}	η_{base}	$\bar{\mu}_{ERB}$	$\bar{\mu}_{RSSI}$ (dBm)	P_{reg}	$\bar{\mu}$ (m)	σ (m)
Todas	62	1920	5,04	-105,07	100%	80,51	97,97
Rede 2G	15	283	2,20	-101,70	14,74%	73,95	105,47
Rede 3G	24	1763	2,60	-90,94	91,82%	96,48	117,01
Rede 4G	23	897	4,98	-120,04	46,72%	158,99	141,02

(b)

Fonte: Elaborada pelo autor (2021)

100% do conjunto de treinamento da região Z-BNU, houve um aumento na instabilidade do método LRD. Isso pode ser explicado pela heterogeneidade dos registros gerados neste ambiente.

Após as análises apresentadas, podemos concluir que o método LRD obteve melhor desempenho na região Z-ANU, onde as predições das coordenadas geográficas foram mais estáveis e precisas.

4.3.2 Localização por Regressão Direta em Bases Segmentadas

Alguns DMs não suportam as três gerações de rede celular abordadas neste estudo de forma simultânea. Sendo assim, é interessante analisar o comportamento da aplicação do método LRD nas três bases secundárias representadas na Fig. 9, cada uma contendo apenas registros de RSSI de suas respectivas gerações de rede celular. As Tabs. 3(a) e 3(b) apresentam os dados para cada um dos cenários analisados. Para cada base, são informadas a quantidade de ERBs e a quantidade de registros armazenados, denotados, respectivamente, por η_{ERB} e η_{base}. É importante destacar que as bases não são mutuamente exclusivas. Logo, a soma das quantidades de registros das bases segmentadas não corresponde à quantidade de registros presentes na base unificada, que é igual a 1920 registros. Adicionalmente, são indicados a média de ERBs detectadas por cada registro, representada por $\bar{\mu}_{ERB}$, em virtude do número de ERBs detectadas pelo DM variar a cada registro coletado, e a média de valores de RSSI, dada em dBm e denotada por $\bar{\mu}_{RSSI}$. O

percentual de registros de uma determinada base em relação à base unificada, denotado por P_{reg}, é definido pela razão do número de registros na base de interesse pelo número de registros presentes na base unificada. Por exemplo, na Tab. 3(a), temos que $P_{reg} = 3,85\%$ para a rede 2G na região Z-ANU, resultado da divisão entre 74 e 1920. Por fim, valores do erro médio $\bar{\mu}$ e do desvio padrão σ do método LRD também são apresentados para cada uma das bases de dados considerada.

Para a Z-ANU, a princípio, pode-se notar que a quantidade de registros obtidos a partir das redes 3G e 4G são bem próximos da quantidade total de registros para todas as redes (base unificada). Porém, para a rede 2G, apenas 3,85% dos registros totais contém valores de RSSI. Essa quantidade reduzida de registros é decorrente da baixa disponibilidade de ERBs 2G na Z-ANU. Em relação à acurácia, o método LRD obteve o melhor desempenho quando aplicado à base segmentada da rede 3G, com um erro médio aproximado de 18,8 m. Este fato pode ser explicado pela alta quantidade de registros, o que interfere no tamanho da base de treinamento. Além disso, temos uma média de 5,9 ERBs por registro 3G, tendo apenas 34 ERBs no total. Por consequência, 28,1 características (resultado da diferença entre 34 e 5,9), em média, serão preenchidas com valores padrão. Isto indica que a base segmentada da rede 3G tem, em média, aproximadamente 82% de suas características preenchidas com valores padrão, enquanto a rede 4G possui, em média, algo em torno de 89% de suas características preenchidas com valores padrão. Quanto menor esse percentual, mais informações reais são disponibilizados para o método LRD, o que impacta positivamente na acurácia da localização. A base da rede 2G tem um percentual médio de 63% de suas características preenchidas com valores padrão, um valor menor do que aqueles obtidos para as bases da redes 3G e 4G. Porém, a base da rede 2G é reduzida em relação às outras duas redes, impactando no tamanho do conjunto de treinamento e, consequentemente, na acurácia do método LRD.

No caso da região Z-BNU, a quantidade de registros 2G também é menor quando comparada ao número de registros das demais redes, porém a redução não é tão significativa quanto na Z-ANU. Essa divergência na quantidade de registros 2G está relacionada à maior disponibilidade de ERBs da rede 2G na Z-BNU, como citado na Seção 4.2.2. Diferente da Z-ANU, a porcentagem de registros 4G na Z-BNU foi de 46,72% do total de registros. É interessante ressaltar que, diferente da acurácia da base da rede 4G, as acurácias das bases 2G e 3G se aproximaram da acurácia da base unificada. O método LRD aplicado à base segmentada 4G teve um erro médio de 158,99 m, ou seja, aproximadamente o dobro do erro médio obtido quando a base unificada é utilizada.

Apesar de a base segmentada 4G ter apresentado uma boa média de ERBs detectadas por registro em relação às outras redes, os registros 4G obtiveram valores de RSSI considerados baixos, com média de -120,04 dBm, valor considerado insatisfatório ou até com status de desconectado (3GPP, 2018). Além disso, como apresentado na Seção 4.2.2, o valor padrão de preenchimento escolhido para os *missing values* é o valor mínimo en-

contrado pelo DM na detecção de RSSIs, que foi de -140 dBm. Uma vez que a média dos valores de RSSI da base segmentada 4G está próxima desse valor mínimo, o método LRD pode presumir que alguns desses valores de RSSI estão muito próximos ao valor padrão, desconsiderando essa informação ao treinar o algoritmo de AM.

A detecção das ERBs pertencentes às diversas gerações de rede celular e seus respectivos RSSIs pelo DM depende diretamente da disponibilidade dessas tecnologias no DM. Alguns DMs não oferecem suporte às três redes analisadas neste trabalho. Desta forma, o número de ERBs detectadas é reduzido. Por este motivo, torna-se interessante analisar a aplicabilidade do método LRD em cada uma das redes celulares disponíveis na localidade. Contudo, não seria justo comparar apenas os erros médios referentes a cada uma das bases segmentadas, pois a quantidade de registros difere entre elas. Sendo assim, a fim de proporcionar uma comparação coerente entre as bases segmentadas, definimos a eficiência do método LRD aplicado às bases segmentadas, denotada por E_{rede}, tal que

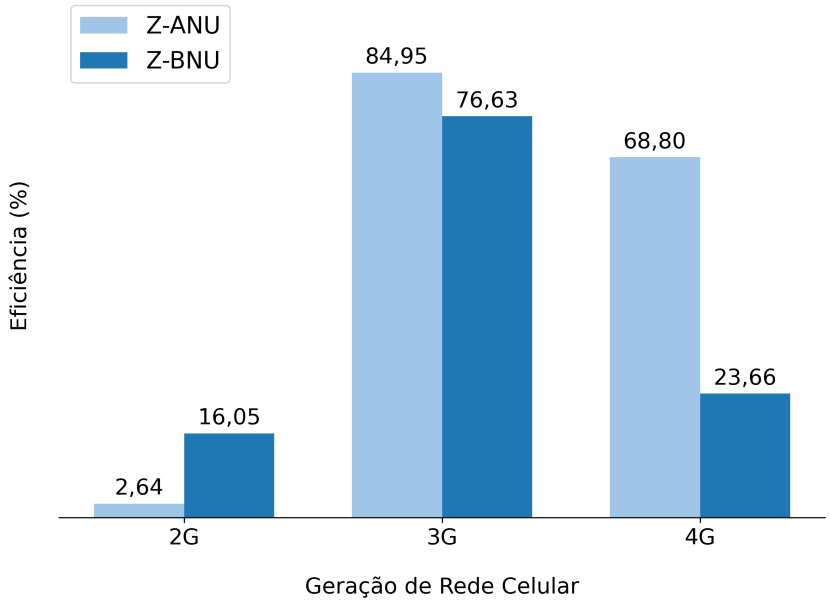
$$E_{rede} = P_{reg}\left(\frac{\bar{\mu}_{BU}}{\bar{\mu}_{BS}}\right), \tag{4.2}$$

em que P_{reg} é o percentual de registros definido anteriormente, $\bar{\mu}_{BU}$ é o erro médio de localização do método LRD aplicado à base unificada, e $\bar{\mu}_{BS}$ é o erro médio de localização do método LRD aplicado a uma determinada base de dados segmentada. Desta maneira, podemos relacionar a capacidade do método LRD fazer a predição dos registros com a acurácia da predição. Aplicando essa métrica em todos os cenários relativos às bases segmentadas por tipo de rede celular, é possível termos uma visão ampla da atuação do método LRD.

A Fig. 13 indica a eficiência do método LRD aplicado às bases de dados segmentadas em cada uma das regiões consideradas. É possível observar que o método LRD apresentou os menores valores de eficiência na rede 2G tanto na região Z-ANU, quanto na região Z-BNU. Tal fato se justifica pela baixa disponibilidade de ERBs 2G em ambas as regiões. Na região Z-BNU, a eficiência foi maior, uma vez que a quantidade de registros obtidos na rede 2G foi superior. Para a rede 3G, os maiores valores de eficiência foram obtidos nas duas regiões, sendo 84,95% na Z-ANU e 76,63% na Z-BNU. A maior diferença entre as eficiências do método LRD ocorreu na rede 4G. Na região Z-ANU, a eficiência obtida foi cerca de 68,8%, enquanto na Z-BNU, o valor obtido foi de 23,66%. Tal disparidade deve-se tanto à quantidade de registros obtidos quanto ao erro médio de localização do método LRD referentes às duas regiões. Também é importante ressaltar que, em nenhum dos casos analisados em particular, o método LRD obteve uma eficiência acima de 100%, o que indicaria que o uso da respectiva base de dados segmentada (2G, 3G ou 4G) geraria um melhor resultado do que a base unificada.

Após analisar os resultados de acurácia relativos à cada base de dados segmentada, é importante avaliar se esta acurácia está dentro de requisitos de tolerância. A FCC (*Federal Communications Commission*), órgão regulador das telecomunicações nos EUA,

Figura 13 – Eficiência do método LRD aplicado às bases segmentadas por tipo de rede celular.



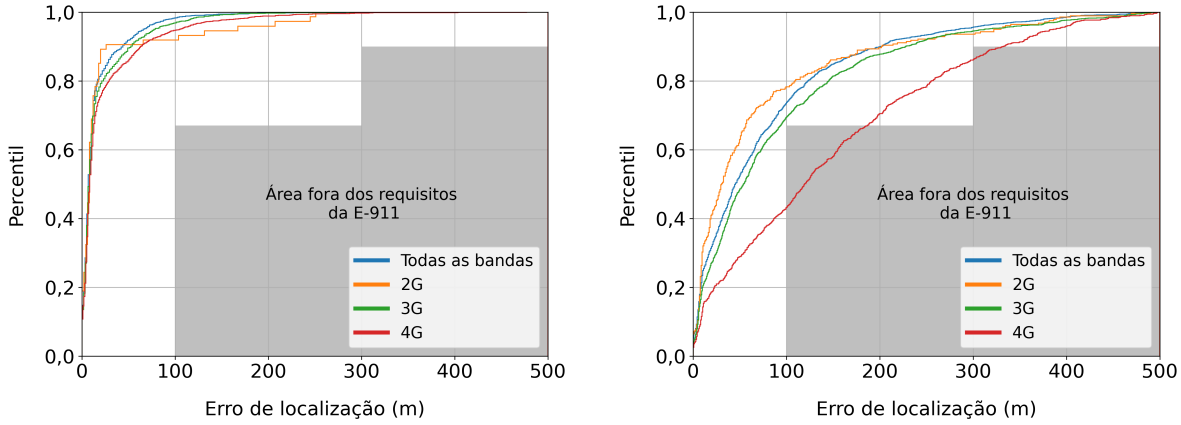
Fonte: Elaborada pelo autor (2021)

é responsável por determinar valores mínimos de erro na localização de um DM em caso de uma chamada de emergência (E-911). Este órgão determina que, em 67% dos casos, a acurácia deve ser de até 100 m, e para 90% dos casos, esta acurácia deve ser de até 300 m. Para verificar o cumprimento dessas determinações na aplicação do método LRD nas duas regiões de interesse, a função de distribuição cumulativa do erro da predição é mostrada na Fig. 14. As áreas em cinza representam as zonas fora dos requisitos de localização impostos pela FCC. É possível notar que, para todas as bases consideradas na Z-ANU, a predição do método LRD consegue atender aos requisitos da FCC, pois a maioria dos erros concentram-se abaixo de 100 m. Além disso, as curvas estão relativamente distantes das áreas em cinza. Porém, na região Z-BNU, o método LRD não atende aos requisitos da FCC para o caso da rede 4G. Além disso, as curvas das outras configurações se aproximam da área fora dos requisitos da E-911. O método LRD aplicado à base de dados 2G obtém melhor resultado para alguns limites de erros, mas vale ressaltar que a base referente a esta rede é reduzida em relação às demais bases.

4.3.3 Localização por Regressão Direta com Aplicação de AECs

A quantidade de características presentes nas bases de dados impacta diretamente no tempo de execução de algoritmos de AM (ANOWAR; SADAOUI; SELIM, 2021). A aplicação de AECs pode reduzir o tempo de execução sem comprometer os resultados de acurácia (QI; JIN; YAN, 2018). Com o objetivo de reduzir o tempo de processamento das fases *on-*

Figura 14 – Função de distribuição cumulativa do erro de localização do método LRD aplicado nas bases segmentadas por geração de rede celular. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



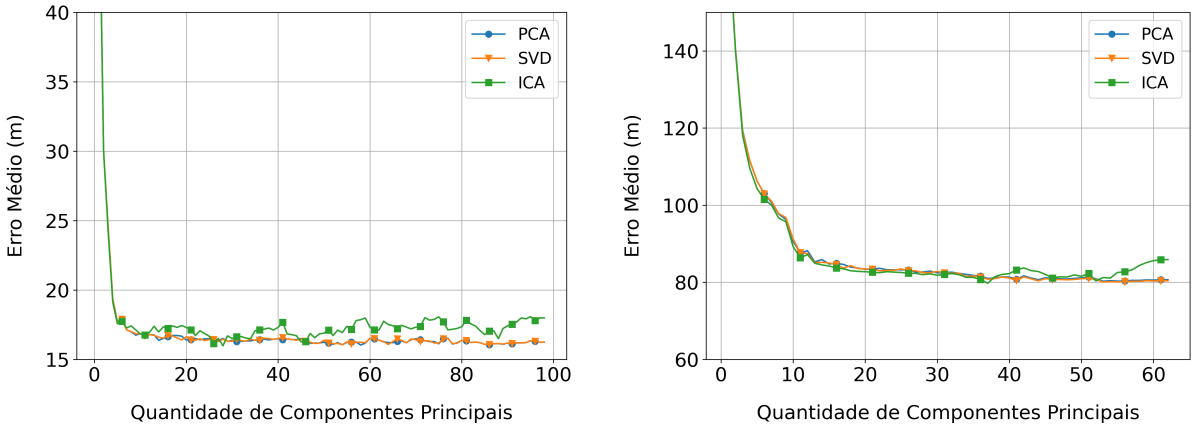
Fonte: Elaborada pelo autor (2021)

line e *off-line* do método LRD, os AECs apresentados no Cap. 3 foram aplicados na base unificada. A base unificada apresenta o maior número de características dentre as bases abordadas, fornecendo uma quantidade maior de dados para os AECs, além de conter ERBs de todas as gerações. Dentre esses AECs, há os algoritmos lineares e não-lineares. Os AECs lineares apresentam melhor desempenho em dados que podem ser linearmente separáveis, enquanto os AECs não-lineares apresentam melhor desempenho em dados que possuem padrões não-lineares. Com o propósito de observar o comportamento dessas duas categorias de AECs em dados de localização por RSSI, ambas foram aplicadas e analisadas.

Todos os AECs considerados neste trabalho possibilitam a variação do número de componentes (ou características) resultantes da redução. Uma vez que o comportamento de cada AEC depende diretamente do número de componentes utilizadas, denotado por C, uma avaliação do erro médio do método LRD foi realizada para as categorias de AECs consideradas nesta pesquisa. O valor de C variou no intervalo $[1, 2, \ldots, Q]$, em que Q representa o número total de características (ERBs) para cada região de interesse. Para a região Z-ANU, $Q = 98$, enquanto para a Z-BNU, $Q = 62$.

Primeiramente, os AECs lineares PCA, SVD e ICA foram analisados. As Figs. 15(a) e 15(b) ilustram os erros médios de localização do método LRD em função da quantidade de componentes principais nas regiões Z-ANU e Z-BNU, respectivamente. Vale salientar que, para auxiliar na visualização dos dados, as escalas dos eixos das ordenadas dos gráficos de cada região não possuem o mesmo intervalo. Além disto, os eixos das abscissas também utilizam intervalos distintos, uma vez que o valor de Q é diferente para cada região. Um comportamento comum nas duas regiões é que o valor do erro médio praticamente se

Figura 15 – Erro médio do método LRD em função da quantidade de componentes principais para os algoritmos lineares PCA, SVD e ICA. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



(a) (b)

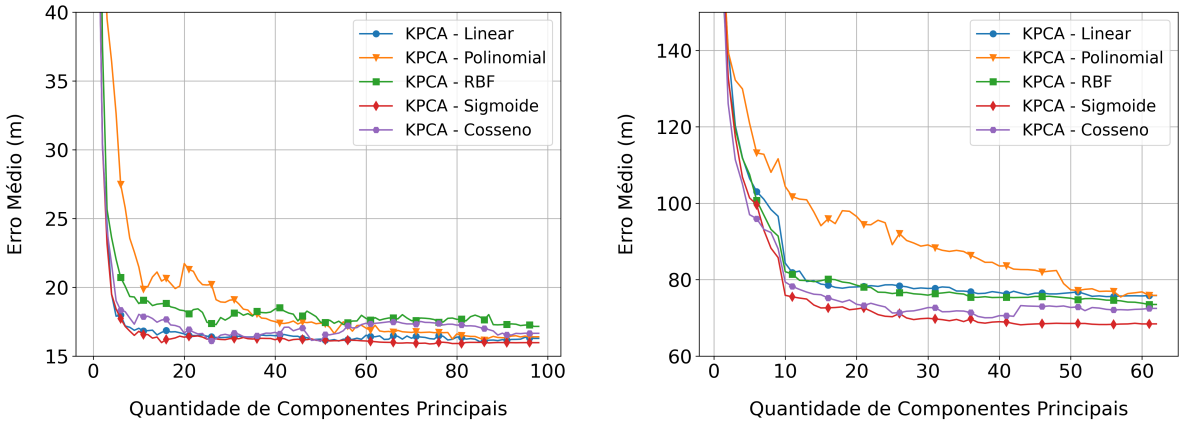
Fonte: Elaborada pelo autor (2021)

estabiliza após uma determinada quantidade de componentes. Para a Z-ANU, o valor do erro torna-se estável a partir de aproximadamente dez componentes. Já para a Z-BNU, pelo fato de as bases relacionadas a esta região possuírem dados mais instáveis, esse comportamento apenas é notado a partir de 20 componentes. Isso pode ser explicado pela quantidade total de características das bases geradas nas duas regiões. A base da região Z-ANU disponibiliza 98 características para a geração dos novos componentes, enquanto a base da Z-BNU disponibiliza apenas 62 características. Isto reflete na quantidade de informações fornecidas para os AECs e, consequentemente, na qualidade e nível de variância dos componentes gerados. Os algoritmos PCA e SVD apresentaram comportamentos semelhantes em ambas as regiões, o que pode ser explicado pelo fato de serem baseados em metodologias semelhantes. Esses dois AECs tiveram melhor desempenho na Z-ANU, com erros médios entre 16 e 17 m, menores que o algoritmo ICA. Por outro lado, na região Z-BNU, praticamente não há distinção entre os resultados dos algoritmos lineares (PCA, SVD e ICA) aplicados na regressão.

Dentre os AECs não-lineares disponíveis na literatura, os algoritmos KPCA e ISOMAP foram os selecionados para análise neste estudo. Além de não necessitarem de iterações, o que deixaria o processo de extração de características mais custoso, o algoritmo KPCA tem melhor desempenho em dados não-lineares e o algoritmo ISOMAP consegue lidar melhor com dados esparsos em relação aos outros AECs (ANOWAR; SADAOUI; SELIM, 2021). Na análise de ambos, foi necessário investigar cada um individualmente para verificar como o erro médio iria se comportar em função de alguns parâmetros importantes, como, por exemplo, o núcleo no algoritmo KPCA e o número de vizinhos no algoritmo ISOMAP.

As Figs. 16(a) e 16(b) indicam o erro médio do método LRD por quantidade de com-

Figura 16 – Erro médio do método LRD em função da quantidade de componentes principais para o algoritmo não-linear KPCA e seus diversos núcleos. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.

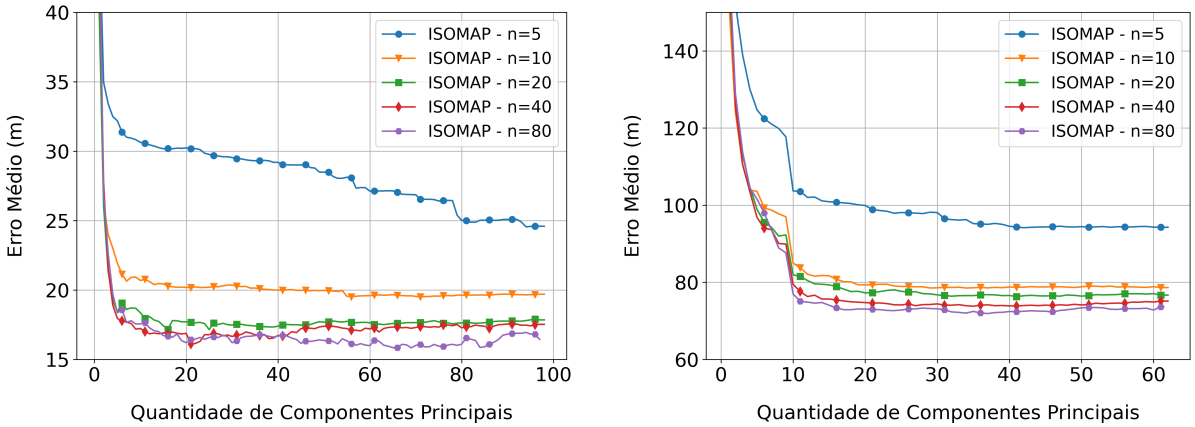


Fonte: Elaborada pelo autor (2021)

ponentes principais considerando a utilização do algoritmo KPCA para as regiões Z-ANU e Z-BNU, respectivamente. Foram adotados os núcleos linear, polinomial, RBF, sigmoide e cosseno. É possível observar que, para as duas regiões, o algoritmo KPCA-Sigmoide obteve o menor erro após o período de estabilização. Este período de estabilização acontece quando as características são reduzidas a poucos componentes. Na Z-ANU, o algoritmo KPCA-Linear conseguiu resultados semelhantes ao KPCA-Sigmoide. O mesmo não aconteceu na Z-BNU, sendo o algoritmo KPCA-Cosseno o que obteve resultados mais próximos ao KPCA-Sigmoide. O algoritmo KPCA-Polinomial apresentou uma estabilização mais lenta nas duas regiões consideradas, obtendo o pior resultado dentre todos os núcleos investigados. De maneira análoga ao que acontece aos AECs lineares, a estabilização do erro médio no algoritmo KPCA também acontece mais rapidamente para a região Z-ANU. Um fato relevante é que, para certos valores de C, a aplicação de algumas modalidades do algoritmo KPCA resultou em valores de erro médio menores do que os obtidos sem a extração de características. Por exemplo, na aplicação do KPCA-Sigmoide, o erro médio atingiu valores inferiores a $16\,m$ a partir de $C = 60$ componentes na região Z-ANU. A redução do erro médio abaixo daqueles obtidos sem a extração de características aconteceu principalmente na Z-BNU. Neste caso, houve uma redução de aproximadamente 12% no erro médio em virtude da aplicação do KPCA-Sigmoide para $C = 43$ componentes. Esse comportamento também foi relatado em (QI; JIN; YAN, 2018), em que resultados melhores de acurácia também foram obtidos após a implantação da extração de características.

As Figs. 17(a) e 17(b) ilustram o erro médio do método LRD em função da quantidade de componentes principais considerando a utilização do algoritmo ISOMAP para as regiões Z-ANU e Z-BNU, respectivamente. Em (ANOWAR; SADAOUI; SELIM, 2021), o número de

Figura 17 – Erro médio do método LRD em função da quantidade de componentes principais para o algoritmo não-linear ISOMAP com 5, 10, 20, 40 e 80 vizinhos. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



Fonte: Elaborada pelo autor (2021)

vizinhos permuta entre os valores do conjunto $\{2, 5\}$. O intervalo escolhido foi pequeno pois a base de dados usada para avaliar o ISOMAP tem apenas 200 registros. Dado que a base unificada considerada neste trabalho de dissertação possui 1920 registros, o intervalo do número de vizinhos escolhido inicia em um número pequeno e dobra a cada iteração. Assim, os valores assumidos para o número de vizinhos foram 5, 10, 20, 40 e 80. Os resultados indicam que o erro médio diminui com o aumento do número de vizinhos empregados no algoritmo ISOMAP. Na Z-ANU, por exemplo, o erro médio, após a estabilização, ficou próximo a $20\,m$ para o algoritmo ISOMAP ($n = 10$). Porém, quando o número de vizinhos aumentou para 80, o erro médio diminuiu para aproximadamente $17\,m$. Um comportamento semelhante acontece na região Z-BNU, onde há a defasagem de aproximadamente $23\,m$ no erro médio entre as modalidades ISOMAP ($n = 5$) e ISOMAP ($n = 80$), quando usados 40 ou mais componentes. Semelhante ao comportamento do algoritmo KPCA, também foram obtidos erros médios abaixo dos valores obtidos sem a extração de características.

O objetivo primário da extração de características nesta pesquisa é reduzir os tempos das fases *on-line* e *off-line* do método LRD sem prejudicar a acurácia da localização. Contudo, conforme mencionado anteriormente, a extração de características também pode promover uma melhoria da acurácia do sistema de localização, o que pode ser considerado um objetivo secundário. Dito isso, a Tab. 4 mostra a consolidação de resultados oriundos da aplicação dos principais AECs estudados, com ênfase na quantidade de componentes C e na minimização do erro médio do método LRD, denotados por $\bar{\mu}_{min}$, nas regiões Z-ANU e Z-BNU. Observa-se que o algoritmo KPCA-Sigmoide obteve $15{,}88\,m$ de erro médio na Z-ANU, $0{,}36\,m$ a menos que o erro obtido pelo método LRD sem a extração

Tabela 4 – Métricas de desempenho (número de características C e erro médio mínimo $\bar{\mu}_{min}$) do método LRD obtidas pela utilização de AECs. Os valores em negrito representam os algoritmos com os erros médios mínimos nas categorias linear e não-linear (KPCA e ISOMAP) para cada região.

AEC	Z-ANU		Z-BNU	
	$\bar{\mu}_{min}$	C	$\bar{\mu}_{min}$	C
PCA	16,03	58	80,28	53
SVD	16,05	54	80,06	53
ICA	**15,99**	**28**	**79,74**	**37**
KPCA-Linear	16,05	85	75,55	55
KPCA-Polinomial	16,16	86	75,46	57
KPCA-RBF	17,22	94	73,83	60
KPCA-Sigmoide	**15,88**	**74**	**68,17**	**43**
KPCA-Cosseno	16,07	50	70,02	38
ISOMAP ($n = 5$)	24,55	94	94,15	42
ISOMAP ($n = 10$)	19,51	56	78,50	34
ISOMAP ($n = 20$)	17,14	25	76,28	41
ISOMAP ($n = 40$)	16,10	21	73,78	35
ISOMAP ($n = 80$)	**15,83**	**69**	**71,88**	**37**

Fonte: Elaborada pelo autor (2021)

de características, usando 74 componentes para treinamento e predição. Ainda na região Z-ANU, o algoritmo ICA obteve um erro de 15,99 m, 0,25 m abaixo do erro do método LRD sem a extração de características, usando 28 componentes. Por fim, o algoritmo ISOMAP ($n = 80$) gerou um erro de 15,83 m usando 69 componentes, também superando o desempenho do método LRD original (*vide* Tab. 2). Uma vez que a maior diferença de erro médio em favor da aplicação dos AECs foi de 0,41 m, um valor considerado pequeno para a acurácia de sistemas de localização *outdoor*, é possível afirmar que, a princípio, o objetivo primário da utilização dos AECs foi atingido na região Z-ANU sem melhorias significativas na acurácia.

Prosseguindo com a análise da Tab. 4, podemos observar que a redução do erro médio foi mais significativa na Z-BNU. O algoritmo KPCA-Sigmoide atingiu um erro de 68,17 m, 12,34 m a menos que o erro médio sem a redução da dimensionalidade, usando 43 componentes. Já no algoritmo ISOMAP ($n = 80$), o menor erro foi 8,63 m abaixo do erro médio original, com 37 componentes. Finalmente, o menor erro dos algoritmos lineares, neste caso, o ICA, foi apenas 0,77 m menor do que o desempenho do método LRD sem a extração de características, com o mesmo número de componentes que o algoritmo ISOMAP ($n = 80$). Diante das métricas apresentadas para a região Z-BNU, nota-se que a aplicação dos AECs não-lineares resultou, em um primeiro momento, não apenas em

Tabela 5 – Métricas de desempenho (número mínimo de características C_{min} e erro médio tolerável $\bar{\mu}_*$) do método LRD obtidas pela utilização de AECs. Os valores em negrito representam os algoritmos com número mínimo de características nas categorias linear e não-linear (KPCA e ISOMAP) para cada região.

AEC	Z-ANU		Z-BNU	
	C_{min}	$\bar{\mu}_*$	C_{min}	$\bar{\mu}_*$
PCA	**8**	**16,99**	18	83,92
SVD	**8**	**17,00**	17	83,67
ICA	10	16,83	**14**	**84,48**
KPCA-Linear	9	16,88	10	84,32
KPCA-Polinomial	53	16,64	38	84,53
KPCA-RBF	-	-	10	82,09
KPCA-Sigmoide	**8**	**16,83**	**10**	**75,90**
KPCA-Cosseno	20	16,68	10	79,37
ISOMAP ($n = 5$)	-	-	-	-
ISOMAP ($n = 10$)	-	-	11	83,74
ISOMAP ($n = 20$)	-	-	10	81,92
ISOMAP ($n = 40$)	**11**	**17,01**	10	79,45
ISOMAP ($n = 80$)	14	17,00	**10**	**76,96**

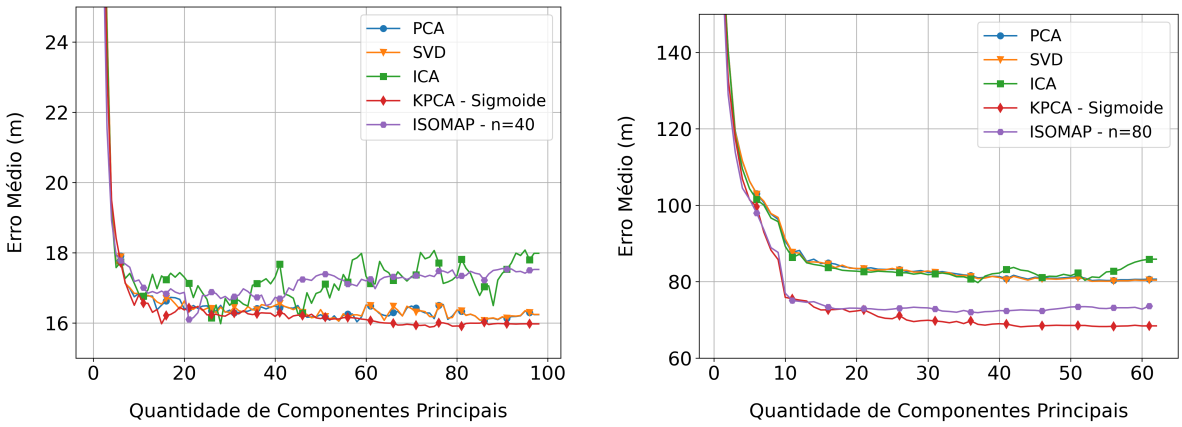
Fonte: Elaborada pelo autor (2021)

redução do número de componentes, mas também na melhoria da acurácia do método LRD. Em outras palavras, a redução de dimensionalidade na região Z-BNU permitiu que os objetivos primário e secundário fossem atingidos.

Tendo em vista o objetivo primário da aplicação de redução de dimensionalidade ao método LRD e sabendo também que a diminuição de características pode causar perda de acurácia, definimos uma métrica denominada erro médio tolerável, denotado por $\bar{\mu}_*$, tal que $\bar{\mu}_* \leq 1,05\bar{\mu}$, ou seja, um limiar de tolerância para o aumento do erro médio em 5%. Tal limiar foi escolhido pelo fato deste aumento não prejudicar o desempenho de bons sistemas de localização *outdoor* de uma maneira geral. Dessa forma, os valores de $\bar{\mu}_*$ para as regiões Z-ANU e Z-BNU são de 17,05 m e 84,53 m, respectivamente. Para esta análise, o AEC com melhor desempenho baseado no objetivo primário desta pesquisa é aquele que, com o menor número de componentes, fornece um erro igual ou menor ao erro médio tolerável.

A Tab. 5 mostra o número mínimo de componentes, denotado por C_{min}, correspondente ao erro $\bar{\mu}_*$, para cada um dos AECs investigados neste trabalho, em cada região de interesse. Para facilitar o entendimento dos valores apresentados, considere os casos dos algoritmos PCA, SVD e KPCA-Sigmoide na região Z-ANU. Os três AECs em questão obtiveram $C_{min} = 8$, o que representa aproximadamente 8% da quantidade inicial de

Figura 18 – Erro médio do método LRD em função da quantidade de componentes principais para os algoritmos lineares (PCA, SVD e ICA) e os melhores não-lineares (KPCA-Sigmoide e ISOMAP ($n = 40$ e $n = 80$)). (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



(a) (b)

Fonte: Elaborada pelo autor (2021)

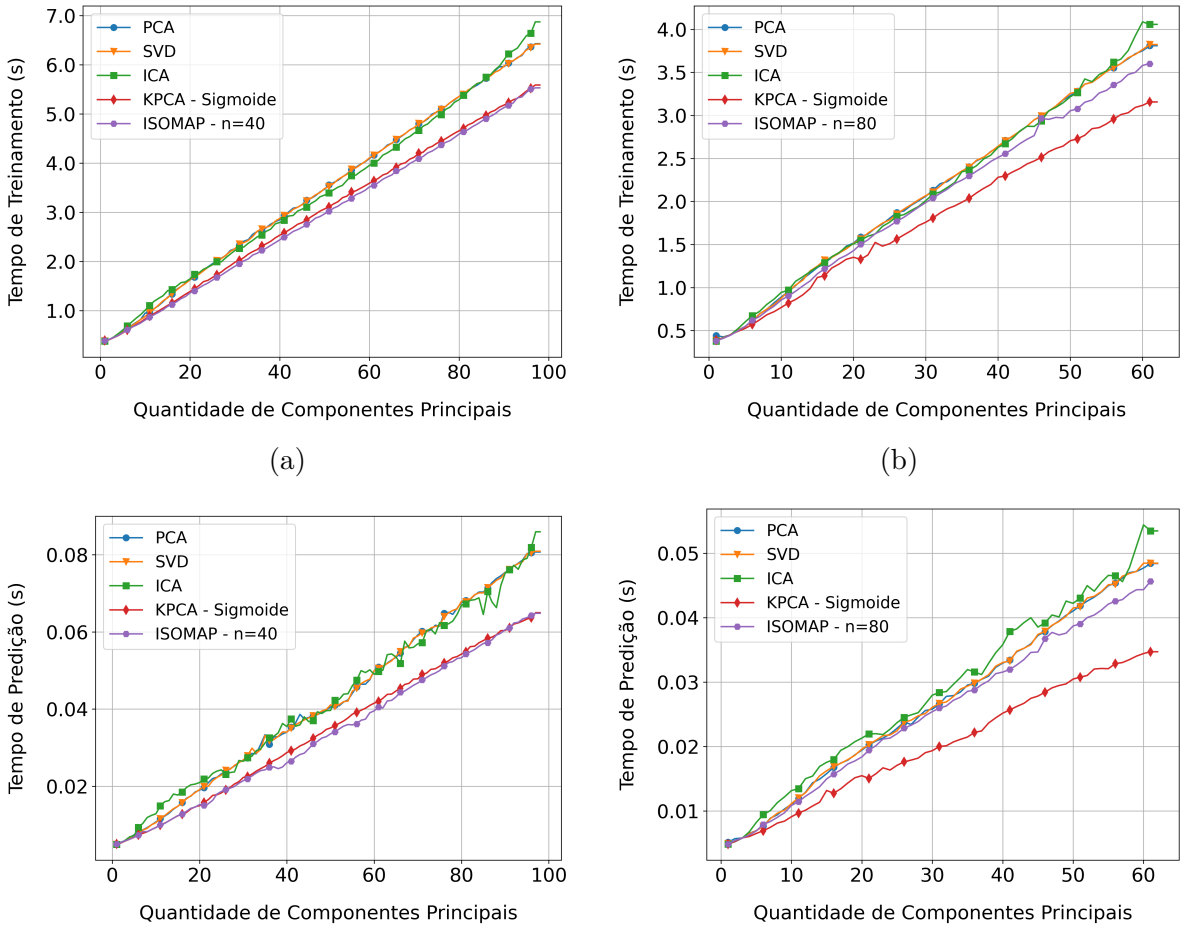
características da base de dados da região, respeitando o critério estabelecido por $\bar{\mu}_*$.

Enquanto isso, na região Z-BNU, os AECs não-lineares obtiveram o melhor resultado para C_{min} (dez características), o que representa algo em torno de 16% da quantidade inicial de características, para satisfazer o erro médio tolerável. Algumas configurações de AECs não-lineares não atenderam ao critério de $\bar{\mu}_*$ e, por esta razão, ficaram sem valor definido (representados por hífen).

A análise da Tab. 5 permitiu a seleção das melhores configurações para os algoritmos não-lineares KPCA e ISOMAP. Para a região Z-ANU, os algoritmos KPCA-Sigmoide e ISOMAP ($n = 40$) geraram os melhores resultados, usando oito e 11 componentes, respectivamente. Na região Z-BNU, houve um empate entre os algoritmos KPCA-RBF, KPCA-Sigmoide e KPCA-Cosseno, todos usando dez componentes. O mesmo aconteceu para o algoritmos ISOMAP com 20, 40 e 80 vizinhos. Como critério de desempate, foram escolhidos os algoritmos com menores valores de $\bar{\mu}_*$, quais sejam, o KPCA-Sigmoide e o ISOMAP ($n = 80$).

As Figs. 18(a) e 18(b) mostram, respectivamente para as regiões Z-ANU e Z-BNU, o erro médio do método LRD em função da quantidade de componentes principais, considerando a utilização dos AECs lineares, além dos algoritmos KPCA-Sigmoide e ISOMAP. No caso do algoritmo ISOMAP, foram utilizados 40 vizinhos na região Z-ANU, enquanto na Z-BNU, 80 vizinhos foram considerados. Em um contexto geral, o algoritmo KPCA-Sigmoide apresentou o melhor resultado, dado o critério de erro médio tolerável, para ambas as regiões. Na região Z-BNU, a diferença de erro médio entre o algoritmo KPCA e os demais é mais visível, atingindo aproximadamente $5\,m$ de diferença (a partir de 40 componentes) para o algoritmo ISOMAP com 80 vizinhos, a segunda melhor opção. Contudo,

Figura 19 – Tempos de processamento das fases *off-line* (treinamento) e *on-line* (predição) em função da quantidade de componentes principais utilizada nos AECs lineares (PCA, SVD e ICA) e não-lineares (KPCA-Sigmoide e ISOMAP ($n = 40$ e $n = 80$)) para cada região de interesse. Treinamento: (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém. Predição: (c) Z-ANU: Recife. (d) Z-BNU: Sirinhaém.



Fonte: Elaborada pelo autor (2021)

na região Z-ANU, os AECs não-lineares não superaram os lineares de forma tão nítida quanto no caso da região Z-BNU. Desta forma, podemos então afirmar que os AECs não-lineares KPCA-Sigmoide e ISOMAP geraram resultados melhores em comparação aos AECs lineares de uma maneira mais evidente na região com menor nível de urbanização.

Como mencionado anteriormente, a redução dos tempos de processamento das fases *off-line* e *on-line* é o objetivo primário da extração de características nesta pesquisa. Dito isso, as Figs. 19(a) e 19(b) mostram o tempo de treinamento, medido em s, em função da quantidade de componentes principais para os AECs lineares, KPCA-Sigmoide e ISOMAP, considerando as regiões Z-ANU e Z-BNU, respectivamente. Para a região Z-ANU, foi assumido $n = 40$ vizinhos para o algoritmo ISOMAP, ao passo que, para a Z-BNU, 80 vizinhos foram utilizados. Inicialmente, podemos notar, para ambas as regiões, um au-

mento do tempo de treinamento com comportamento próximo ao linear, à medida que a quantidade de componentes aumenta. Cabe ressaltar que isto é válido para todos os AECs considerados. Os tempos de treinamento do método LRD, para as duas regiões, exibem uma taxa de crescimento menor nos AECs não-lineares em comparação com os AECs lineares. Esse mesmo comportamento é observado nos tempos de predição mostrados nas Figs. 19(c) e 19(d).

Segundo os resultados apresentados na Fig. 19 e o propósito de diminuição de tempo por meio de redução de dimensionalidade, o algoritmo KPCA-Sigmoide obteve um dos melhores tempos de processamento para as fases *off-line* e *on-line* do método LRD, enquanto também obteve os menores erros toleráveis. Por este motivo, os algoritmos KPCA-Sigmoide com oito e dez componentes serão adotados como os AECs de melhor desempenho para as regiões Z-ANU e Z-BNU, respectivamente.

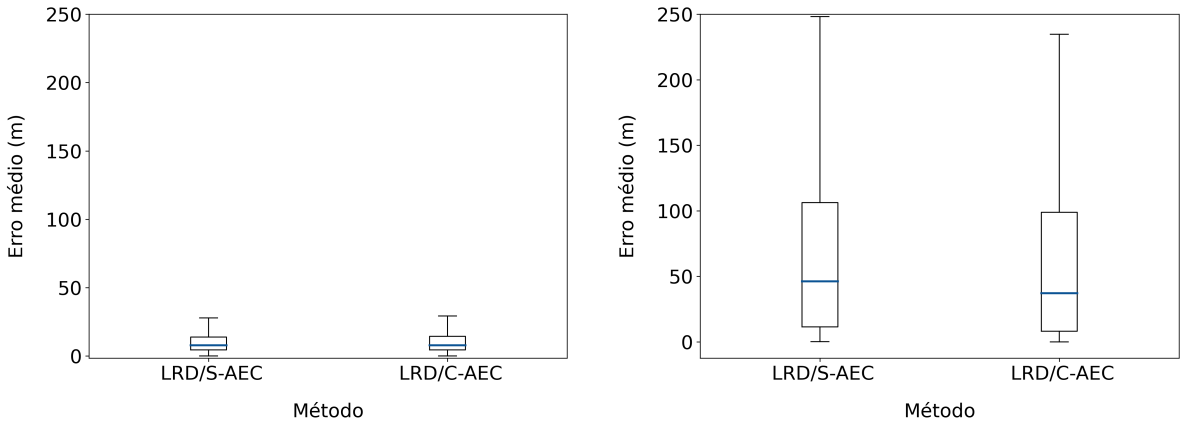
Para a extração de características ser considerada eficiente, o método LRD com aplicação do AEC não-linear KPCA-Sigmoide, que será denominado de sistema LRD/C-AEC, deve ter uma acurácia equivalente ou maior do que a acurácia do método LRD sem a aplicação do AEC não-linear KPCA-Sigmoide, chamado de sistema LRD/S-AEC. Para o sistema LRD/C-AEC, foi escolhido o algoritmo KPCA-Sigmoide com oito e dez componentes para as regiões Z-ANU e Z-BNU, respectivamente.

A fim de avaliar a similaridade dos sistemas LRD/C-AEC e LRD/S-AEC, é necessário analisar a equivalência entre os erros gerados por ambos. Caso a equivalência seja constatada, pode-se considerar que os erros gerados pelos dois sistemas, ou seja, a acurácia dos sistemas com e sem extração de características também são equivalentes. Entretanto, caso a equivalência não seja constatada, pode-se considerar que o sistema que obteve maior acurácia (menor erro médio) é mais eficiente que o sistema que obteve menor acurácia (maior erro médio). Para isto, alguns testes estatísticos foram realizados nos conjuntos formados pelos erros obtidos pelos sistemas LRD/C-AEC e LRD/S-AEC.

Inicialmente, para verificar a normalidade da distribuição, foi aplicado o teste de Shapiro-Wilk (SHAPIRO; WILK, 1965) aos sistemas LRD/S-AEC e LRD/C-AEC de cada região. Duas hipóteses são assumidas para o teste de Shapiro-Wilk, quais sejam, H_0, hipótese em que os dados seguem uma distribuição normal, chamada de hipótese nula; e H_1, hipótese em que os dados não seguem uma distribuição normal, denominada de hipótese alternativa. Para o teste de normalidade, os valores-p obtidos para as duas regiões foram menores do que 10^{-183}, sendo considerados infinitesimais, ou seja, muito próximos a zero. Assim, podemos rejeitar a hipótese nula e constatar que os dados não seguem uma distribuição normal nas duas regiões. Desta forma, a aplicação de um teste de hipótese não-paramétrico é mais apropriada (GIBBONS; CHAKRABORTI, 2011). Para isso, foi escolhido o teste de Wilcoxon (WILCOXON, 1945) para amostras independentes.

O teste de Wilcoxon também apresenta duas hipóteses, sendo a hipótese nula H_0, a que considera os dois sistemas equivalentes, e a hipótese alternativa H_1, aquela que assume

Figura 20 – Diagrama de caixa do erro médio de localização dos sistemas LRD/S-AEC e LRD/C-AEC. (a) Z-ANU: Recife. (b) Z-BNU: Sirinhaém.



Fonte: Elaborada pelo autor (2021)

que os dois sistemas não são equivalentes. A definição do valor do nível de significância também é de α=0,05 e a análise dos dados é semelhante ao teste de Shapiro-Wilk. Observando os resultados da Tab. 6, podemos constatar que o valor-p para a região Z-ANU foi 0,31, ou seja, $p > \alpha$. Logo, não é possível rejeitarmos a hipótese nula, não havendo evidências para afirmar que um método apresente melhor resultado ou que seja mais eficiente que outro na região Z-ANU. Em contrapartida, na região Z-BNU, o valor-p foi igual a $2,03 \cdot 10^{-4}$, ou seja, $p << \alpha$, o que significa que podemos rejeitar a hipótese nula e afirmar que os resultados apresentados pelo sistema LRD/C-AEC não são equivalentes aos apresentados pelo sistema LRD/S-AEC. Em outras palavras, uma vez que mostramos que os sistemas não são equivalentes, podemos afirmar, conforme foi definido anteriormente, que o sistema com extração de características (LRD/C-AEC) é mais eficiente que o sistema sem extração de características (LRD/S-AEC) para a região Z-BNU, pois é o sistema com maior acurácia dentre os sistemas analisados.

As Figs. 20(a) e 20(b) apresentam os diagramas de caixa do erro médio de localização dos sistemas LRD/S-AEC e LRD/C-AEC para as regiões Z-ANU e Z-BNU, respectivamente. É possível observar que, em ambas as regiões, as acurácias dos dois sistemas são equivalentes. Em outras palavras, a mediana dos erros são próximas. Ainda podemos ob-

Tabela 6 – Valores-p do teste estatístico de Wilcoxon aplicado aos sistemas LRD/S-AEC e LRD/C-AEC.

	Wilcoxon
Z-ANU	0,31
Z-BNU	$2,03 \cdot 10^{-4}$

Fonte: Elaborada pelo autor (2021)

Tabela 7 – Tempos de treinamento e de predição (em s) para os sistemas LRD/S-AEC e LRD/C-AEC aplicados aos dados das regiões Z-ANU e Z-BNU.[2]

	Tempo de treinamento		Tempo de predição	
	LRD/S-AEC	LRD/C-AEC	LRD/S-AEC	LRD/C-AEC
Z-ANU	5,06	0,71	0,06	0,0081
Z-BNU	3,09	0,77	0,04	0,0091

Fonte: Elaborada pelo autor (2021)

servar que os limites inferiores e superiores também são equivalentes. Em alguns casos, como na aplicação do LRD/C-AEC na região Z-BNU, o limite superior apresentou inclusive uma melhora de aproximadamente dez metros. Em suma, pode-se concluir, através da análise da Fig. 20, que a aplicação do método LRD com o uso de AECs para ambas as regiões obteve resultados de acurácia compatíveis com os resultados da aplicação do método LRD sem o uso de AECs.

Comparando os tempos de execução das fases do método LRD entre os sistemas LRD/S-AEC e LRD/C-AEC, pode-se concluir, a partir dos dados da Tab. 7, que houve uma redução no tempo de treinamento de aproximadamente sete vezes para a Z-ANU e de aproximadamente quatro vezes para a Z-BNU. Para os tempos de predição, também houve reduções de tempo semelhantes aos valores da fase de treinamento, tendo a Z-ANU e a Z-BNU diminuições aproximadas pouco acima de sete e quatro vezes, respectivamente.

Tendo em vista os resultados apresentados, pode-se constatar que o objetivo de reduzir o tempo de execução das fases do método LRD com a aplicação de AECs foi alcançado, pois houve uma redução significativa em ambos os tempos de treinamento e predição do método de localização. Contudo, considerando os EMAs mínimos obtidos por cada um dos AECs, ainda pode-se afirmar que, em alguns casos, ocorreu o aumento da acurácia, ou seja, a redução do erro médio da localização, ao mesmo tempo que houve a diminuição do número de componentes e, consequentemente, a redução dos tempos de execução das fases do método LRD. Esse aspecto configura o melhor cenário para sistemas de localização.

[2] As simulações foram executadas em um sistema com processador Core i5-3330 3 GHz e memória RAM de 8 GB.

5 CONCLUSÃO

Este trabalho analisou a aplicação de um método de radiolocalização de dispositivos móveis por regressão direta, denotado por método LRD, em regiões com diferentes níveis de urbanização. O primeiro aspecto da análise foi a robustez do método LRD em função do nível de urbanização das regiões consideradas. Uma vez que o método LRD emprega algoritmos de AM, o segundo aspecto analisado foi o efeito da redução de dimensionalidade na acurácia e nos tempos de execução do método de localização. A literatura aborda essas análises, porém não apresenta um trabalho com resultados quantitativos relacionados à densidade de ERBs das regiões, tampouco avalia regiões de diferentes níveis de urbanização no tocante à extração de características.

Para a avaliação dos aspectos anteriormente mencionados, regiões circulares com raios aproximados de $500\,m$ foram consideradas nas cidades de Recife-PE e Sirinhaém-PE, respectivamente, como de alto nível de urbanização (denotada por Z-ANU) e de baixo nível de urbanização (identificada por Z-BNU). Três gerações de rede celular foram consideradas neste trabalho, quais sejam, as redes 2G, 3G e 4G. Após a coleta das informações de rede, duas bases de dados unificadas foram geradas para cada região.

O método LRD aplicado à base unificada da região Z-ANU obteve um erro médio aproximadamente cinco vezes menor do que o erro médio referente à base unificada da região Z-BNU. Para verificar a estabilidade do método LRD nas regiões consideradas, a base de dados unificada de entrada foi utilizada não apenas em sua totalidade (100%), mas também foi fracionada em porções menores contendo 25%, 50% e 75% do conjunto de treinamento, para representar a disponibilidade da coleta via *crowdsourcing*. Diferente dos resultados da região Z-BNU, os resultados da Z-ANU mostraram-se mais estáveis, pois o erro médio de localização diminuiu com o aumento do tamanho do conjunto de treinamento.

As bases unificadas foram segmentadas por geração de rede celular, contendo apenas registros que detectaram as ERBs da rede celular em questão. Devido à incompatibilidade no número de registros das bases segmentadas, foi definida um métrica de eficiência do método LRD. Dentre as redes disponíveis, o método LRD apresentou melhor eficiência quando aplicado à rede 3G em ambas as regiões. Para todos os cenários considerados, a acurácia do método LRD ficou dentro dos requisitos exigidos pela FCC, exceto para o cenário da base segmentada 4G referente à região Z-BNU.

Em uma segunda fase do estudo, foi investigada a aplicação de AECs ao método LRD, considerando a utilização das bases unificadas das regiões Z-ANU e Z-BNU. Cinco AECs foram considerados neste trabalho, sendo três lineares e dois não-lineares. Em geral, os AECs não-lineares obtiveram um desempenho melhor que os AECs lineares. Para alguns AECs, como, por exemplo, o algoritmo KPCA-Sigmoide, o método LRD obteve erros médios aproximadamente 15% menores na região Z-BNU em comparação com os erros

médios obtidos sem a aplicação de AECs. Na região Z-ANU, a redução do erro médio foi de 2%. Embora o aumento da acurácia seja favorável à aplicação de AECs no método LRD, os fatores de redução alcançados não são significativos para acurácias de sistemas de localização *outdoor*.

O segundo efeito obtido pela utilização de AECs foi a diminuição do tempo de processamento (treinamento e predição) do método LRD. Para atingir uma redução máxima de componentes e, consequentemente, maximizar a diminuição dos tempos de treinamento e predição, foi definido como critério um limiar de degradação de 5% para o erro médio de localização do método LRD. Dado este critério, o algoritmo KPCA-Sigmoide obteve os melhores resultados para ambas as regiões, conseguindo realizar a predição usando aproximadamente 8% e 16% do número inicial de componentes empregadas nas regiões Z-ANU e Z-BNU, respectivamente. A redução de componentes promovida na base de dados da região Z-ANU pelo algoritmo KPCA-Sigmoide resultou em uma diminuição aproximada de sete vezes nos tempos de treinamento e predição do método LRD quando comparado ao tempo de treinamento obtido sem a redução de características. No caso da região Z-BNU, a diminuição do tempo de processamento foi de aproximadamente quatro vezes.

O desempenho geral do método LRD mostrou-se satisfatório para as regiões com diferentes níveis de urbanização consideradas neste trabalho. Além de obter resultados dentro dos requisitos da FCC para a maioria das análises apresentadas, o método LRD também mostrou-se adaptável ao emprego de AECs. A redução de dimensionalidade se mostrou efetiva na diminuição dos tempos de execução das fases *off-line* e *on-line* do método LRD, chegando a promover uma melhoria da acurácia, porém não tão significativa quanto a redução do esforço computacional da técnica de localização.

Por fim, como perspectiva de trabalhos futuros, sugere-se a aplicação de outros métodos de localização de DMs, como, por exemplo, a técnica de *fingerprinting*, em conjunto com AECs em regiões com diferentes níveis de urbanização. Outro ponto relevante a ser considerado é o uso de combinação de regressores no método LRD. O uso das técnicas de combinação ou seleção de regressores são apropriadas quando as bases de dados apresentam características não-determinísticas nos valores de suas características (AMEMIYA, 1980). Desta forma, a combinação de regressores pode trazer benefícios para as Z-BNUs, onde os dados apresentam maior heterogeneidade.

REFERÊNCIAS

3GPP. *Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management*. [S.l.], 2018. Version 15.3.0.

ABDALLAH, A. A.; SAAB, S. S.; KASSAS, Z. M. A machine learning approach for localization in cellular environments. In: *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*. [S.l.: s.n.], 2018. p. 1223–1227.

AMEMIYA, T. Selection of regressors. *International Economic Review*, [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], v. 21, n. 2, p. 331–354, 1980.

ANAGNOSTOPOULOS, G. G.; KALOUSIS, A. A reproducible analysis of RSSI fingerprinting for outdoor localization using Sigfox: Preprocessing and hyperparameter tuning. In: *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. [S.l.: s.n.], 2019. p. 1–8.

ANATEL. *Agência Nacional de Telecomunicações*. [S.l.], 2021. Acessado em 14 de maio de 2021. Disponível em: <https://www.gov.br/anatel/>.

ANOWAR, F.; SADAOUI, S.; SELIM, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, v. 40, p. 100378, 2021.

APOSTOLO, G. H.; SAMPAIO, I. G. B.; VITERBO, J. Feature selection on database optimization for wi-fi fingerprint indoor positioning. *Procedia Computer Science*, v. 159, p. 251–260, 2019.

AURENHAMMER, F. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 23, n. 3, p. 345–405, set. 1991. ISSN 0360-0300.

AYESHA, S.; HANIF, M. K.; TALIB, R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, v. 59, p. 44–58, 2020. ISSN 1566-2535.

BROWN, J. D. *Eigen Decomposition*. [S.l.]: Springer International Publishing, 2018. 117-148 p.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.

COX, M. A. A.; COX, T. F. *Multidimensional Scaling*. [S.l.]: Springer Berlin Heidelberg, 2008. 315–347 p.

DONG, Y.; YAO, Y.-D. IoT platform for COVID-19 prevention and control: A survey. *IEEE Access*, v. 9, p. 49929–49941, 2021.

DUFKOVá, K.; FICEK, M.; KENCL, L.; NOVAK, J.; KOUBA, J.; GREGOR, I.; DANIHELKA, J. Active GSM cell-id tracking: "where did you disappear?". In: . [S.l.: s.n.], 2008. p. 7–12.

ELBAKLY, R.; YOUSSEF, M. Crescendo: An infrastructure-free ubiquitous cellular network-based localization system. In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. [S.l.: s.n.], 2019. p. 1–6.

FAZIO, M.; BUZACHIS, A.; GALLETTA, A.; CELESTI, A.; VILLARI, M. A proximity-based indoor navigation system tackling the COVID-19 social distancing measures. In: *2020 IEEE Symposium on Computers and Communications (ISCC)*. [S.l.: s.n.], 2020. p. 1–6.

GIBBONS, J. D.; CHAKRABORTI, S. *Nonparametric Statistical Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. 977–979 p. ISBN 978-3-642-04898-2.

HACKELING, G. *Mastering Machine Learning With Scikit-Learn*. [S.l.]: Packt Publishing, 2014. ISBN 1783988363.

HONG, J.; THAKURIAH, P. V. Examining the relationship between different urbanization settings, smartphone use to access the internet and trip frequencies. *Journal of Transport Geography*, v. 69, p. 11–18, 2018.

IBGE. *Instituto Brasileiro De Geografia e Estatística - Censo Brasileiro de 2010*. [S.l.], 2021. Acessado em 14 de maio de 2021. Disponível em: <https://cidades.ibge.gov.br/>.

JOLLIFFE, I.; CADIMA, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 374, p. 20150202, 04 2016.

JOSHI, S. K.; MACHCHHAR, S. An evolution and evaluation of dimensionality reduction techniques — a comparative study. In: *2014 IEEE International Conference on Computational Intelligence and Computing Research*. [S.l.: s.n.], 2014. p. 1–5.

KARNEY, C. F. F. Algorithms for geodesics. *Journal of Geodesy*, v. 87, p. 43–55, 2013.

KHALAJMEHRABADI, A.; GATSIS, N.; AKOPIAN, D. Modern WLAN fingerprinting indoor positioning methods and deployment challenges. *IEEE Communications Surveys Tutorials*, v. 19, n. 3, p. 1974–2002, 2017.

KHALID, S.; KHALIL, T.; NASREEN, S. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 Science and Information Conference*. [S.l.: s.n.], 2014. p. 372–378.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. New York, NY: Springer, 2013. ISBN 9781461468493 1461468493 1461468485 9781461468486.

KUUTTI, S.; FALLAH, S.; KATSAROS, K.; DIANATI, M.; MCCULLOUGH, F.; MOUZAKITIS, A. A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications. *IEEE Internet of Things Journal*, v. 5, n. 2, p. 829–846, 2018.

LECA, C. L.; CIOTIRNAE, P.; RINCU, C. I.; NICOLAESCU, I. Characteristics of crowdsourcing for outdoor radio fingerprinting positioning. In: *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. [S.l.: s.n.], 2017. p. 1–4.

MAHDAVINEJAD, M. S.; REZVAN, M.; BAREKATAIN, M.; ADIBI, P.; BARNAGHI, P.; SHETH, A. P. Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, v. 4, n. 3, p. 161–175, 2018. ISSN 2352-8648.

MENIEM, M. H. A.; HAMAD, A. M.; SHAABAN, E. Relative RSS-based GSM localization technique. In: *IEEE International Conference on Electro-Information Technology , EIT 2013*. [S.l.: s.n.], 2013. p. 1–6.

MISTRY, H. P.; MISTRY, N. H. RSSI based localization scheme in wireless sensor networks: A survey. In: *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. [S.l.: s.n.], 2015. p. 647–652.

MORAVEK, P.; KOMOSNY, D.; SIMEK, M.; JELINEK, M.; GIRBAU, D.; LAZARO, A. Investigation of radio channel uncertainty in distance estimation in wireless sensor networks. *Telecommunication Systems*, v. 52, p. 1549–1558, 2011.

OGUEJIOFOR, O.; ANIEDU, A.; EJIOFOR, H.; OKOLIBE, A. Trilateration based localization algorithm for wireless sensor network. *International Journal of Science and Modern Engineering (IJISME)*, v. 1, p. 21–27, 2020.

OLIVEIRA, L. L.; JR., L. A. O.; SILVA, G. W. A.; TIMOTEO, R. D. A.; CUNHA, D. C. An RSS-based regression model for user equipment location in cellular networks using machine learning. *Wireless Networks*, v. 25, p. 4839–4848, 2019.

QI, G.; JIN, Y.; YAN, J. RSSI-based floor localization using principal component analysis and ensemble extreme learning machine technique. In: *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. [S.l.: s.n.], 2018. p. 1–5.

RAMACHANDRAN, R.; RAVICHANDRAN, G.; RAVEENDRAN, A. Evaluation of dimensionality reduction techniques for big data. In: *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. [S.l.: s.n.], 2020. p. 226–231.

REDDY, G. T.; REDDY, M. P. K.; LAKSHMANNA, K.; KALURI, R.; RAJPUT, D. S.; SRIVASTAVA, G.; BAKER, T. Analysis of dimensionality reduction techniques on big data. *IEEE Access*, v. 8, p. 54776–54788, 2020.

REFAEILZADEH, P.; TANG, L.; LIU, H. *Cross-Validation*. Boston, MA: Springer US, 2009. 532–538 p. ISBN 978-0-387-39940-9.

RIZK, H.; SHOKRY, A.; YOUSSEF, M. Effectiveness of data augmentation in cellular-based localization using deep learning. In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. [S.l.: s.n.], 2019. p. 1–6.

SAEED, N.; BADER, A.; AL-NAFFOURI, T. Y.; ALOUINI, M. S. *When wireless communication faces COVID-19: combating the pandemic and saving the economy*. 2020.

SHAFIQUE, K.; KHAWAJA, B. A.; SABIR, F.; QAZI, S.; MUSTAQIM, M. Internet of Things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. *IEEE Access*, v. 8, p. 23022–23040, 2020.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, [Oxford University Press, Biometrika Trust], v. 52, n. 3/4, p. 591–611, 1965. ISSN 00063444.

SHUKRI, S.; KAMARUDIN, L.; NDZI, D.; ZAKARIA, A.; AZEMI, S.; KAMARUDIN, K.; ZAKARIA, S. M. M. S. RSSI-based device free localization for elderly care application. In: *2nd International Conference on Internet of Things, Big Data and Security*. [S.l.: s.n.], 2017. p. 125–135.

TERUEL, P. E. Lopez-de; CANOVAS, O.; GARCIA, F. J. Using dimensionality reduction techniques for refining passive indoor positioning systems based on radio fingerprinting. *Sensors*, v. 17, n. 4, 2017.

VIKRAM, M.; PAVAN, R.; DINESHBHAI, N. D.; MOHAN, B. Performance evaluation of dimensionality reduction techniques on high dimensional data. In: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. [S.l.: s.n.], 2019. p. 1169–1174.

VO, Q. D.; DE, P. A survey of fingerprint-based outdoor localization. *IEEE Communications Surveys Tutorials*, v. 18, n. 1, p. 491–506, 2016.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, [International Biometric Society, Wiley], v. 1, n. 6, p. 80–83, 1945. ISSN 00994987.

YADAV, S.; SHUKLA, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. [S.l.: s.n.], 2016. p. 78–83.

ZHAO, Z.; ZHANG, R.; COX, J.; DULING, D.; SARLE, W. Massively parallel feature selection: an approach based on variance preservation. *Mach. Learn.*, v. 92, n. 1, p. 195–220, 2013.

APÊNDICE A – DESCRIÇÃO DOS REGISTROS DE REDE

A aplicação desenvolvida para coleta dos dados de rede deste trabalho armazena os dados no formato *JavaScript Object Notation* (JSON). Cada registro é representado por um objeto JSON, este contendo o par chave e valor para cada uma das informações coletadas. A descrição das chaves são apresentadas na Tab. 8. Um exemplo da representação de registro coletado é mostrado no Código A.1.

Tabela 8 – Descrição das informações contidas nos registros da coleta de dados.

Campo (Chave)	Descrição
date	Data e horário
position	Coordenadas geográficas
lat	Latitude (graus decimais)
lon	Longitude (graus decimais)
cellScans	Conjunto de informações de cada ERB detectada
networkType	Tipo de rede da ERB detectada
cellId	Identificação única da ERB detectada
areaCode	Código de área da ERB detectada (apenas para GSM)
RSSI	Indicador de intensidade do nível de sinal da ERB detectada

Código A.1 - Registro coletado no formato JSON.

```
{
  "date": Thu Jan 22 20:02:16 GMT-03:00 2021,
  "position": { "lat": -8.03902873, "lon": -34.91155472 },
  "cellScans": [
    {
      "networkType": GSM,
      "cellId": 32331,
      "areaCode": 6840,
      "RSSI": -87
    },
    {
      "networkType": LTE,
      "cellId": 343,
      "RSSI": -96
    },
    {
      "networkType": LTE,
      "cellId": 441,
      "RSSI": -92
    },
    {
      "networkType": LTE,
      "cellId": 305,
      "RSSI": -90
    },
    {
```

```
27          "networkType": LTE,
            "cellId": 75,
29          "RSSI": -93
          },
31        {
            "networkType": LTE,
33          "cellId": 441,
            "RSSI": -101
35        },
          {
37          "networkType": WCDMA,
            "cellId": 269,
39          "RSSI": -103
          },
41        {
            "networkType": WCDMA,
43          "cellId": 58,
            "RSSI": -103
45        }
      ]
47  }
```