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

Gabriel W. A. Silva[1], **Daniel C. Cunha**[1]

[1]Centro de Informática (CIn) – Universidade Federal de Pernambuco (UFPE)
50.740-560 – Recife, PE – Brasil

{gwas,dcunha}@cin.ufpe.br

Abstract. *This work analyzes the application of the direct regression localization (DRL) method in two regions with different levels of urbanization. In addition, the effect of dimensionality reduction, through feature extraction algorithms (FEAs), is addressed on the accuracy and execution times of the radiolocalization method. Experimental results evidenced that the average prediction error of the DRL method decreased as a function of the increase in the training set in the region with the highest level of urbanization. Furthermore, the FEA kernel principal component analysis using sigmoid function provided an approximate seven-fold decrease in training time and approximately a four-fold decrease in the prediction time of the DRL method without impairing its accuracy.*

Resumo. *Este trabalho analisa a aplicação do método de localização por regressão direta (LRD) em duas regiões com diferentes níveis de urbanização, além de abordar o efeito da redução de dimensionalidade, por meio de algoritmos de extração de características (AECs), na acurácia e nos tempos de execução do método de radiolocalização. Resultados experimentais mostraram que o erro médio de predição do método LRD diminuiu em função do aumento do conjunto de treinamento na região com maior nível de urbanização. Adicionalmente, o AEC KPCA com núcleo Sigmóide proporcionou uma diminuição aproximada de sete vezes no tempo de treinamento e de cerca de quatro vezes no tempo de predição do método LRD sem prejudicar sua acurácia.*

1. Introdução

O crescimento do uso de dispositivos móveis (DMs) 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 os dados gerados pelos DMs para simplificar e enriquecer atividades e experiências humanas [Mahdavinejad et al. 2018]. Dentre as principais aplicações IoT está a capacidade de realizar a predição da localização de DMs por meio de dados extraídos a partir de sinais de 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 degradação de desempenho pelo ausência de visada direta entre transmissor e receptor, bem como a alta demanda de energia consumida pelo DM. Dessa forma, é importante buscarmos métodos de localização alternativos ao GPS.

Um método de localização baseado na intensidade do nível de sinal recebido (RSSI, *received signal strength indicator*) chamado de localização por regressão direta (LRD) foi proposto em [Oliveira et al. 2019]. O método 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 a partir das estações rádio base (ERBs) da rede celular. Dessa forma, a acurácia e o tempo de execução do método LRD estão diretamente relacionados à disponibilidade de ERBs na região de aplicação do método. Um dos fatores que afeta a densidade de ERBs é o nível de urbanização da região. Por exemplo, zonas com baixo nível de urbanização são ambientes com densidade demográfica reduzida e infraestrutura de redes com menor quantidade de ERBs. Por outro lado, zonas com alto nível de urbanização apresentam densidade demográfica elevada, além de infraestrutura de redes com maior número de ERBs. Considerando a expectativa do aumento da quantidade de ERBs com a adoção das redes 5G [Shafique et al. 2020], tal fato representa o crescimento do número de características no método LRD, o que impacta diretamente no tempo de execução de algoritmos de AM [Anowar et al. 2021]. Face a esta particularidade inerente a métodos baseados em algoritmos de AM, o objetivo deste trabalho é investigar a aplicação do método LRD em regiões com diferentes densidades de ERBs e como os algoritmos de extração de características (AECs) impactam o tempo de processamento do método de localização.

Os resultados deste estudo mostram que, para a zona com alto nível de urbanização, o erro médio de localização do método LRD diminui com o aumento do conjunto de treinamento utilizado. Além disso, dentre os AECs investigados, o algoritmo KPCA com núcleo Sigmóide proporcionou uma redução aproximada de sete e quatro vezes, nos tempos de treinamento e predição, respectivamente, do método LRD. Por fim, apesar do objetivo principal do emprego da redução de dimensionalidade seja a diminuição do tempo de processamento, o algoritmo KPCA proporcionou uma melhoria aproximada de 6% na acurácia do método LRD.

O artigo está organizado como se segue. Na Seção 2, são introduzidos conceitos e terminologias referentes ao método LRD, bem como a definição das fases que o compõem. A Seção 3 apresenta a configuração das zonas selecionadas, detalha o processo de coleta dos dados e evidencia informações relativas às ERBs de cada zona. Na Seção 4, a análise comparativa dos dados é realizada considerando acurácia e tempo de execução do método LRD. Por fim, a Seção 5 apresenta as principais conclusões deste trabalho.

2. Radiolocalização por Regressão Direta

O método LRD possui duas fases, chamadas de fases *off-line* e *on-line*, ilustradas, respectivamente, nas Figs. 1a e 1b. A fase *off-line*, também chamada de fase de treinamento, é formada por três passos e 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. No primeiro passo, uma coleta 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 RSSI para todas as ERBs que aquele DM consegue detectar naquele determinado local. No segundo passo, as funções de hipótese $f_1(\cdot)$ e $f_2(\cdot)$ são obtidas por meio de um modelo baseado em AM e usadas para predizer a latitude e longitude alvos do DM. As funções $f_1(\cdot)$ e $f_2(\cdot)$ representam, cada uma, a instância do algoritmo k-*nearest*

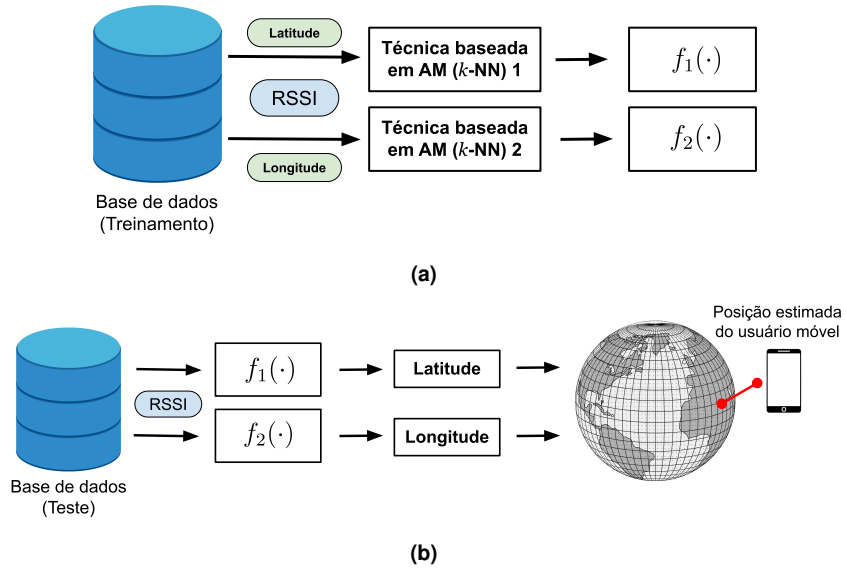

Figura 1. Representação do método de localização por regressão direta (LRD). (a) Fase *off-line*: obtenção das funções de hipótese $f_1(\cdot)$ e $f_2(\cdot)$. (b) Fase *on-line*: execução da predição da posição do usuário móvel.

neighbors (k-NN) para latitude e longitude, respectivamente. Desta forma, assuma que as bases de dados S^{ϕ} e S^{λ} são dadas por

$$S^{\phi} = \left\{(\phi_i, \mathbf{q}_i) \in \mathbb{R} \times \mathbb{R}^{N_i}\right\} \tag{1}$$

e

$$S^{\lambda} = \left\{(\lambda_i, \mathbf{q}_i) \in \mathbb{R} \times \mathbb{R}^{N_i}\right\}, \tag{2}$$

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 sensoriadas na i-ésima instância. Assim, S^{ϕ} e S^{λ} são as bases de dados destinadas à construção das funções $f_1(\cdot)$ e $f_2(\cdot)$, respectivamente. Os valores de ϕ_i e λ_i são considerados como alvos, enquanto os valores contidos em $\mathbf{q}_i$ representam as características.

A fase *on-line*, também conhecida como fase de predição, é iniciada após a conclusão da fase de treinamento e também é composta por três passos. 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. Após o sensoriamento, esses valores são usados como características de entrada para as funções $f_1(\cdot)$ e $f_2(\cdot)$ na segunda fase, obtendo, respectivamente, a latitude e longitude alvos do DM. No terceiro passo, a posição final é estimada usando as coordenadas geográficas obtidas no passo anterior. Maiores detalhes sobre o método LRD podem ser encontrados em [Oliveira et al. 2019].

3. Setup Experimental

A disposição do sinal de RF em redes 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). Neste artigo, tal ambiente será denominado *zona de alto nível de urbanização* e denotado por Z_A deste ponto em diante. Em contrapartida, as zonas com baixo nível de urbanização apresentam densidade demográfica reduzida em relação à zona Z_A, infraestrutura de redes de comunicação com menor número de ERBs e, por fim, um número reduzido de construções. De modo equivalente, denotaremos uma zona com baixo nível de urbanização por Z_B. Os diferentes níveis de urbanização definidos anteriormente possuem influência direta na propagação e disponibilidade dos sinais de RF nos ambientes mencionados.

Duas regiões de interesse das cidades de Recife-PE e Sirinhaém-PE foram escolhidas para representar as zonas Z_A e Z_B, respectivamente. Segundo a ANATEL, a cidade de Recife possui aproximadamente 7.534 habitantes por quilômetro quadrado, com cobertura de sinais de rede celular em 99,7% do seu território. Por outro lado, a cidade de Sirinhaém possui uma densidade demográfica de 123 habitantes por quilômetro quadrado, com apenas 63,9% do seu território coberto por sinais de telefonia celular. Por fim, enquanto a cidade de Recife possui cobertura de 100% dos moradores e domicílios da cidade, Sirinhaém se limita a ter 90,5% dos moradores e 91,3% dos domicílios cobertos pelas redes, deixando parte da população sem cobertura de sinal [Anatel 2021].

O processo de coleta de dados nas duas zonas urbanas mencionadas se deu por meio da simulação do procedimento de construção de bases de dados do tipo contribuição coletiva, ou *crowdsourcing*. A construção da base de dados de treinamento ocorre 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 usados como características. Além disso, considerando que os usuários ativos estão 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 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 do sistema de localização. Um ponto importante a se observar 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 nas regiões consideradas foi realizada por meio de uma aplicação desenvolvida para um DM *Android* compatível com as bandas de frequência das redes celulares 2G, 3G e 4G. O GPS do DM 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. Para cada uma das zonas selecionadas, foram definidas duas regiões de coleta de dados com base na semelhança da disposição de suas ruas, a fim de se promover uma comparação justa. A Fig. 2 ilustra as coordenadas dos dados coletados (em cor azul) nos mapas das regiões de coleta das duas cidades. Foram detectadas 98 ERBs na zona Z_A, sendo apenas oito da rede 2G, 34

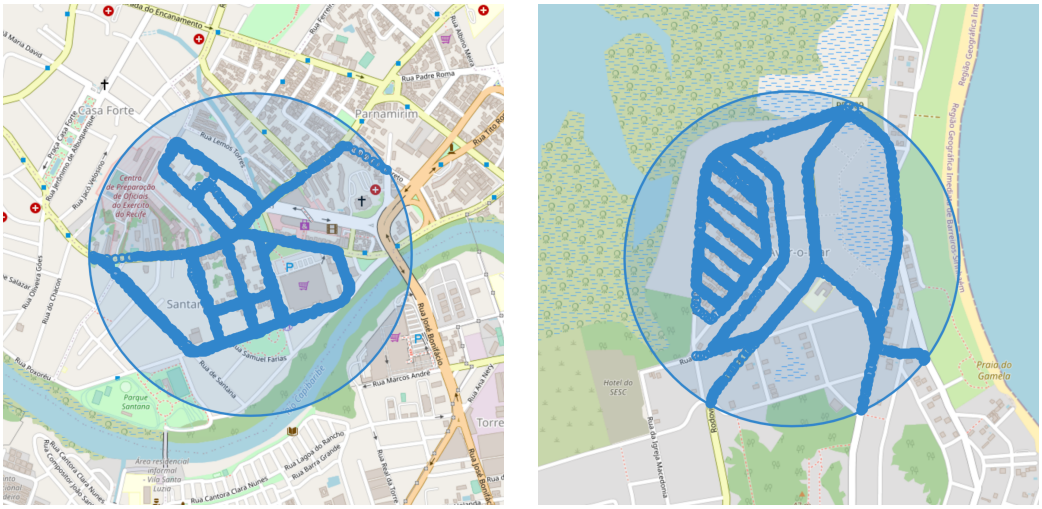

Figura 2. Distribuição dos dados coletados (em cor azul) nas zonas urbanas de Recife (Z_A, lado esquerdo) e de Sirinhaém (Z_B, lado direito).

da rede 3G e 56 da rede 4G. Em contrapartida, na zona Z_B foram detectadas 62 ERBs: 15 da rede 2G, 24 da rede 3G e 23 da rede 4G. A captura dos dados foi feita em forma de registros e em ambas as regiões, foram coletados 1.920 registros em uma área circular de raio igual a 500 m. Para a formação das bases de dados, os níveis de sinal medidos em relação às ERBs foram considerados como características e as coordenadas geográficas, os valores alvos. Assim, cada registro (instância) continha as coordenadas geográficas do ponto de coleta e os RSSIs das ERBs detectadas.

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, usa-se o menor valor de RSSI detectado na base [Anagnostopoulos and Kalousis 2019].

Os registros coletados em Z_A detectaram, em média, 11,7 ERBs das 98 disponíveis por registro. Isso significa que, para cada registro, por volta de 12% das ERBs disponíveis foram detectadas. Na zona Z_B, cada registro detectou, em média, cinco ERBs das 62 disponíveis no total, o que remete a aproximadamente 8% das ERBs. Deste modo, cada registro em Z_A conseguiu detectar em média o dobro de ERBs detectadas pelos registros de Z_B, o que significa que uma quantidade menor de valores padrão foi inserida na base de dados de Z_A. Por fim, os registros que não apresentaram nenhum valor real de RSSI foram removidos.

4. Resultados

Neste trabalho, o desempenho do método LRD é avaliado por meio de simulações computacionais[1] usando a linguagem Python com ênfase na biblioteca *scikit-learn* [Hackeling

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

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 (zonas Z_A e Z_B), cujas descrições foram realizadas na Seção 3. O regressor utilizado foi o algoritmo k-NN. Na fase de treinamento do método LRD, foi aplicado o processo de *tuning* no parâmetro k (quantidade de vizinhos) por meio de uma busca exaustiva no conjunto de valores $\{1, 3, 5, 7, 9, 11\}$ [Feurer and Hutter 2019]. O valor de k responsável pelo menor erro em um determinado conjunto de validação (subconjunto de registros do conjunto de treinamento) foi utilizado para a fase de predição 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 avaliar o método LRD, consideramos 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 zona urbana (Z_A ou Z_B) e pela aplicação ou não de AECs. Para analisar o esforço computacional do método LRD, o tempo de processamento de cada fase do algoritmo de regressão foi medido.

Para calcular a acurácia do método LRD nas regiões de coleta que representam as zonas Z_A e Z_B, definimos o erro médio absoluto (EMA), denotado por $\bar{\mu}$, tal que

$$\bar{\mu} = \frac{1}{N} \sum_{i=1}^{N} d_g(\mathbf{y}_i, \mathbf{p}_i)\,, \tag{3}$$

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 posição de cada instância calculado com base na distância geodésica em superfície elipsoide, $\mathbf{y}_i$ representa as coordenadas geográficas (latitude e longitude) reais da i-ésima instância do conjunto de testes. Finalmente, $\mathbf{p}_i$ corresponde às coordenadas geográficas preditas da i-ésima instância do conjunto de teste.

Para reduzir o impacto da aleatoriedade na divisão da base de dados em conjuntos de treinamento e teste, a técnica de validação cruzada K-fold é geralmente empregada [Refaeilzadeh et al. 2009]. Neste trabalho, foi escolhido o valor $K = 10$ para cada execução da técnica K-fold.

Uma vez que estamos assumindo a construção de bases por meio de *crowdsourcing*, é importante considerarmos a variação da disponibilidade dos dados para treinamento. Para analisar o impacto da variação da quantidade dos dados nas zonas Z_A e Z_B, o conjunto de treinamento foi fracionado em quatro porções de 25% cada, e, a cada execução do experimento, os dados de uma porção foram adicionados cumulativamente ao conjunto de treinamento final. Deste modo, a primeira execução usou apenas 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 foi 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 e permaneceu o mesmo para todas as quatro execuções deste experimento a fim de proporcionar uma comparação justa dos resultados.

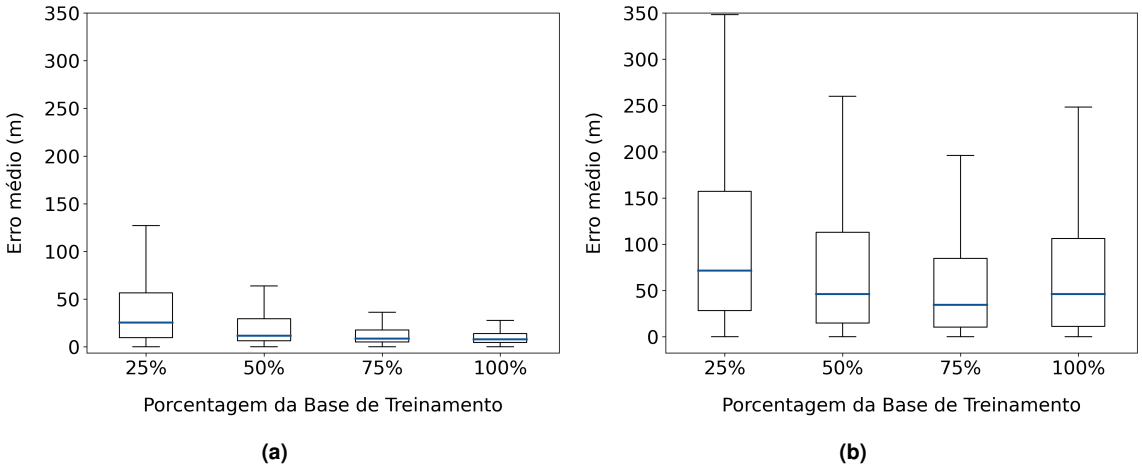

Figura 3. Diagramas de caixa do erro médio de predição de posição (acurácia) do método LRD em função do tamanho do conjunto de treinamento empregado. (a) Z_A: Recife. (b) Z_B: Sirinhaém.

As Figs. 3a e 3b apresentam os diagramas de caixa para cada uma das quatro execuções do experimento definidas anteriormente, nas zonas Z_A e Z_B, respectivamente. Observando os dados de Z_A, 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. Outro fato que consolida isto é que, para o experimento com 25% do conjunto de treinamento, temos um erro médio em torno de 40 m, enquanto que para 100%, temos um erro médio aproximado de 16,24 m. Para a zona Z_B, assim como em Z_A, também se percebe um comportamento decrescente no erro médio obtido pelo método LRD quando o conjunto de treinamento é formado por 25%, 50% e 75% dos dados disponíveis. No entanto, o uso total dos dados de Z_B para o conjunto de treinamento, que fornece um erro médio de 80,51 m, ocasionou uma reversão na tendência, o que possivelmente indica a instabilidade do método LRD em regiões com cobertura parcial da infraestrutura de redes. Tal fato pode ser explicado pela heterogeneidade dos registros gerados neste ambiente. Após as análises apresentadas, podemos concluir que o método LRD obteve melhor desempenho na zona Z_A, onde as predições das coordenadas geográficas foram mais estáveis e precisas.

Assumindo então que o método LRD é treinado com a totalidade dos dados do conjunto de treinamento, é importante avaliar se a acurácia do método está dentro de requisitos de tolerância. A FCC (*Federal Communications Commission*), órgão regulador das telecomunicações nos EUA, é responsável por determinar valores mínimos de erro de localização de um DM para serviços de chamadas de emergência (E-911). 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 [FCC 2010]. Para verificar o cumprimento dessas determinações na aplicação do método LRD nas duas zonas de interesse, a função de distribuição cumulativa do erro de predição é mostrada na Fig. 4. As áreas em cinza representam valores fora dos requisitos de localização impostos pela FCC. É possível notar que, para ambas as zonas urbanas, a predição do método LRD consegue atender aos requisitos da FCC, pois as funções de distribuição cumulativa do erro não interceptam a área em cinza.

A quantidade de características presentes na base de dados impacta diretamente no tempo de execução de algoritmos de AM [Anowar et al. 2021]. A aplicação de AECs

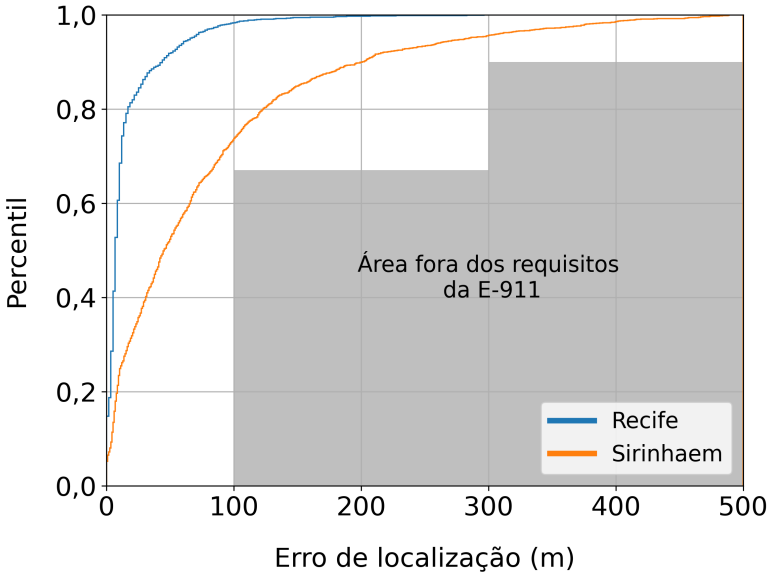

Figura 4. Função de distribuição cumulativa do erro de localização do método LRD aplicado nas zonas urbanas de interesse.

pode reduzir o tempo de execução sem comprometer os resultados de acurácia, assim como também pode melhorar a própria acurácia de algoritmos de AM devido ao pré-processamento adicional dos dados aplicado pelos AECs [Qi et al. 2018]. Com o objetivo de reduzir o tempo de processamento das fases *on-line* e *off-line* do método LRD, foi investigada a aplicação de alguns AECs lineares e não-lineares na base de valores de RSSI. Todos os AECs considerados neste trabalho possibilitam a variação do número de características (também chamadas de componentes principais) 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 os AECs considerados nesta pesquisa. O valor de C variou no intervalo $[1, 2, \ldots, Q]$, em que Q representa o número total de componentes principais (ERBs) para cada zona abordada, sendo $Q = 98$ para Z_A e $Q = 62$ para Z_B.

As Figs. 5a e 5b mostram, respectivamente para as zonas Z_A e Z_B, o erro médio do método LRD em função da quantidade de componentes principais, considerando a utilização dos AECs lineares PCA, SVD e ICA, bem como dos algoritmos não-lineares KPCA com núcleo Sigmóide (denotado por KPCA-Sig) e ISOMAP-n, sendo n o número de vizinhos. Foi escolhido um subconjunto de AECs representativos no que se refere à diversidade algorítmica. Maiores detalhes acerca dos AECs selecionados são descritos em [Anowar et al. 2021]. Cabe ressaltar que foram investigados outros núcleos (linear, polinomial, RBF – *radial basis function* e cosseno) para o algoritmo KPCA, assim como diferentes números de vizinhos no conjunto $\{5, 10, 20, 40, 80\}$ para o algoritmo ISOMAP. A escolha dos valores do número de vizinhos para o algoritmo ISOMAP teve como base as referências [Yousaf et al. 2021] e [Neto and Levada 2020], adotando uma abordagem hibrida dos valores definidos nestes trabalhos. Por conta das restrições de tamanho do artigo, focaremos apenas nos melhores casos de cada categoria de AEC não-linear abordada, quais sejam, os algoritmos KPCA-Sig, ISOMAP-40 e ISOMAP-80.

Em um contexto geral, o algoritmo KPCA-Sig apresentou o melhor resultado, para

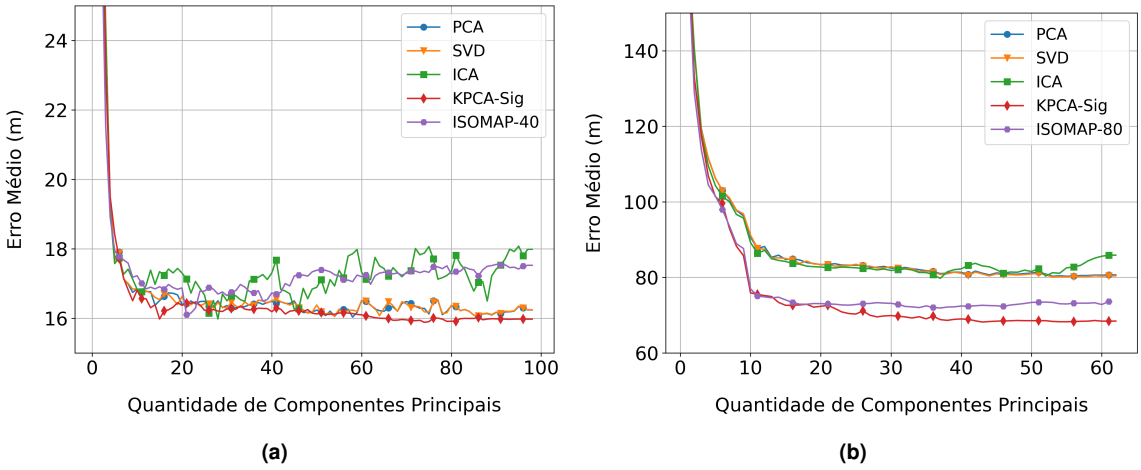

Figura 5. Erro médio do método LRD em função da quantidade de componentes principais para os AECs lineares (PCA, SVD e ICA) e não-lineares (KPCA-Sig, ISOMAP-40 e ISOMAP-80). (a) Z_A: Recife. (b) Z_B: Sirinhaém.

ambas as zonas urbanas. Em Z_B, os AECs não-lineares se mostraram claramente melhores do que os AECs lineares. Dentre os não-lineares, o algoritmo KPCA-Sig apresenta o menor erro médio até aproximadamente 20 componentes principais. Para números de componentes abaixo desse limiar, os erros médios dos algoritmos KPCA-Sig e ISOMAP-80 foram equivalentes. Contudo, na zona Z_A, os AECs não-lineares não superaram os lineares de forma tão nítida quanto no caso de Z_B. Para reforçar tal fato, a Tab. 1 apresenta o erro médio tolerável $\bar{\mu}_*$ e o número mínimo de componentes C_{min} para os AECs lineares e não-lineares considerados nas zonas Z_A e Z_B. O erro médio tolerável é definido como $\bar{\mu}_* \leq 1,05\bar{\mu}$, ou seja, é um limiar de tolerância para o aumento do erro médio em, no máximo, 5%. Tal limiar foi escolhido pelo fato deste aumento não prejudicar o desempenho de bons sistemas de localização *outdoor* de uma maneira geral. É possível observar que, em Z_A, $\bar{\mu}_*$ é aproximadamente o mesmo para todos os AECs considerados, enquanto em Z_B, os AECs não-lineares apresentaram uma diminuição em $\bar{\mu}_*$ na faixa de 7 a 8 m. Por fim, esta diminuição do erro reforça o fato de que a utilização de AECs pode resultar não apenas em redução dos tempos de processamento das fases que compõem o método LRD, mas também em melhoria de seu desempenho. Caso a minimização da quantidade de componentes não seja o foco, é possível obtermos reduções de até 15% no erro médio com a utilização dos AECs, porém esta investigação não fez parte do escopo deste trabalho.

Tabela 1. Métricas de desempenho (número mínimo de componentes C_{min} e erro médio tolerável $\bar{\mu}_*$) do método LRD obtidas pela utilização dos AECs lineares (PCA, SVD e ICA) e não-lineares (KPCA-Sig, ISOMAP-40 e ISOMAP-80) para as zonas urbanas Z_A e Z_B.

		AECs lineares			AECs não-lineares		
Zonas	Métricas	PCA	SVD	ICA	KPCA-Sig	ISOMAP-40	ISOMAP-80
Z_A	$\bar{\mu}_*$	16,99	17,99	16,83	16,83	17,01	17,00
	C_{min}	8	8	10	8	11	14
Z_B	$\bar{\mu}_*$	83,92	83,67	84,48	75,90	79,45	76,96
	C_{min}	18	17	14	10	10	10

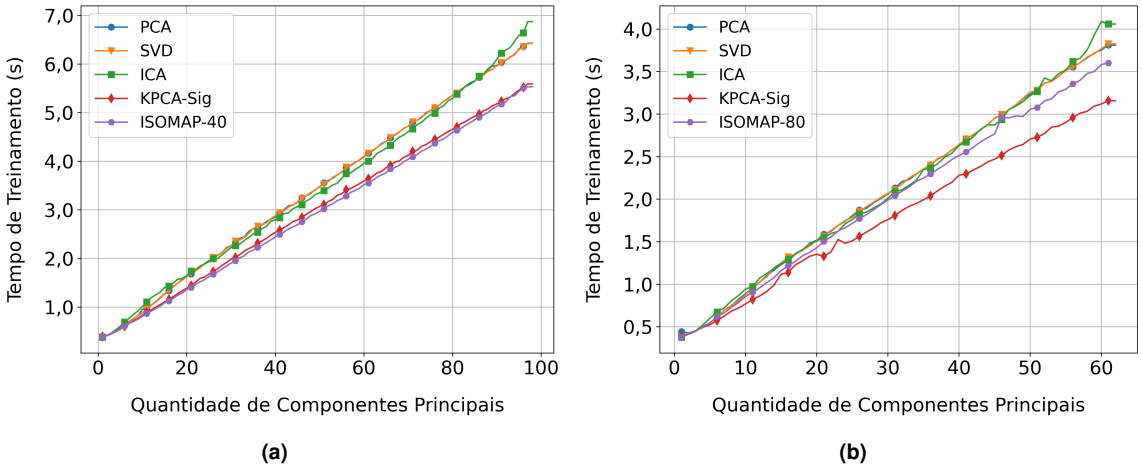

Figura 6. Tempos de treinamento em função da quantidade de componentes principais utilizada nos AECs lineares (PCA, SVD e ICA) e não-lineares (KPCA-Sig, ISOMAP-40 e ISOMAP-80). (a) Z_A: Recife. (b) Z_B: Sirinhaém.

As Figs. 6a e 6b mostram o tempo de treinamento (em s) em função da quantidade de componentes principais para os AECs lineares e os algoritmos KPCA-Sig, ISOMAP-40 e ISOMAP-80, considerando as zonas Z_A e Z_B, respectivamente. Inicialmente, podemos notar, para ambas as zonas, um aumento 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 Z_A e Z_B, exibem uma taxa de crescimento menor nos AECs não-lineares em comparação com os AECs lineares.

O mesmo comportamento é observado nos tempos de predição mostrados nas Figs. 7a e 7b. Segundo os resultados apresentados, o algoritmo KPCA-Sig 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-Sig com oito e dez componentes foram adotados como os AECs de melhor desempenho para as zonas Z_A e Z_B, respectivamente.

Com a finalidade de demonstrar e quantificar as vantagens da aplicação de AECs no método LRD, os tempos de execução das fases *on-line* e *off-line* foram medidos e normalizados, tendo como valor de referência o tempo do sistema LRD sem o uso de AECs, aqui representado pela sigla LRD/S-AEC. Para os tempos da fase *off-line*, o sistema LRD com utilização de AECs, denotado por LRD/C-AEC, apresentou um tempo de treinamento médio em torno de sete vezes menor em Z_A e aproximadamente quatro vezes menor em Z_B em comparação com o tempo de treinamento do sistema LRD/S-AEC. De forma semelhante, para a fase *on-line*, houve reduções aproximadas de mesma magnitude nos tempos médios de predição do sistema LRD/C-AEC em relação ao sistema LRD/S-AEC, ou seja, sete vezes para a zona Z_A e quatro vezes para a zona Z_B.

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 nos tempos de treinamento e predição. Ainda pode-se afirmar que, em alguns casos, ocorreu o aumento da acurácia, ou seja, a redução do erro médio de localização, ao mesmo tempo que houve a diminuição do

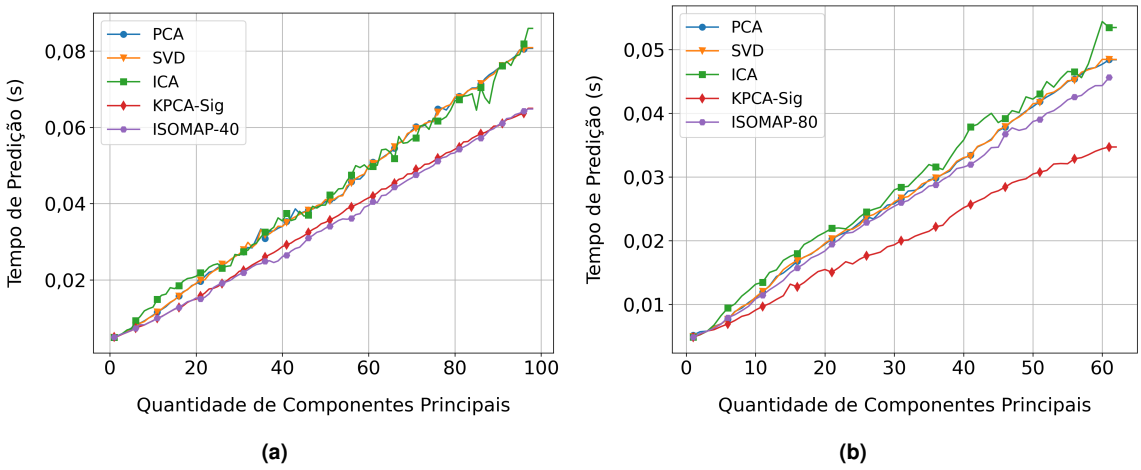

Figura 7. Tempos de predição em função da quantidade de componentes principais utilizada nos AECs lineares (PCA, SVD e ICA) e não-lineares (KPCA-Sigmoide, ISOMAP-40 e ISOMAP-80). (a) Z_A: Recife. (b) Z_B: Sirinhaém.

número de componentes e, consequentemente, a redução dos tempos de execução das fases do método LRD.

5. Conclusões

Uma vez que a diminuição do esforço computacional de técnicas de radiolocalização é um quesito relevante para otimizar o consumo de energia dos dispositivos móveis (DMs), este artigo apresentou uma investigação da utilização de algoritmos de extração de características (AECs) no método de localização por regressão direta (LRD), uma técnica baseada no emprego de algoritmos de aprendizado de máquina. Duas regiões, com raios aproximados de 500 m, foram consideradas nas cidades de Recife-PE e Sirinhaém-PE, respectivamente, como de alto nível de urbanização (zona Z_A) e de baixo nível de urbanização (zona Z_B). Primeiramente, para verificar a estabilidade do método LRD nas regiões consideradas, a base de dados 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. Diferente de Z_B, a zona Z_A apresentou resultados mais estáveis, uma vez que o erro médio de localização diminuiu com o aumento do tamanho do conjunto de treinamento.

O principal efeito obtido pela utilização de AECs no método LRD foi a diminuição do seu tempo de processamento (treinamento e predição). O algoritmo KPCA com núcleo Sigmóide (KPCA-Sig) obteve os melhores resultados para ambas as regiões, conseguindo realizar a predição usando aproximadamente 8% e 16% do número inicial de componentes empregadas nas zonas Z_A e Z_B, respectivamente. A redução de componentes promovida na base de dados de Z_A pelo algoritmo KPCA-Sig resultou em uma diminuição aproximada de sete vezes nos tempos de treinamento e predição do método LRD quando comparados aos tempos de treinamento e predição obtidos sem a redução de características. No caso de Z_B, a diminuição do tempo de processamento foi de aproximadamente quatro vezes. Dessa forma, a redução de dimensionalidade se mostrou efetiva na diminuição dos tempos de execução das fases *off-line* e *on-line* do método LRD.

Como perspectiva de trabalhos futuros, cabe avaliar a diminuição do erro médio do

método LRD em função da quantidade de características da base de entrada, considerando que a redução do tempo de processamento das fases do método LRD não seja o principal objetivo da utilização dos AECs.

Referências

Anagnostopoulos, G. G. and Kalousis, A. (2019). A reproducible analysis of RSSI fingerprinting for outdoor localization using Sigfox: Preprocessing and hyperparameter tuning. In *2019 Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8.

Anatel (2021). Technical report, ANATEL – Agência Nacional de Telecomunicações. Acessado em 14 de maio de 2021.

Anowar, F., Sadaoui, S., and Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput Sci Rev*, 40:100378.

FCC (2010). Second Report and Order on Wireless E911 Location Accuracy Requirements. Technical report, Federal Communications Commission, Washington-US. Rep. 10-176.

Feurer, M. and Hutter, F. (2019). Hyperparameter Optimization. In *Automated Machine Learning*, pages 3–33. Springer, Cham.

Hackeling, G. (2014). *Mastering Machine Learning With Scikit-Learn*. Packt Publishing.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., and Sheth, A. P. (2018). Machine learning for Internet of Things data analysis: a survey. *Digit Commun Netw*, 4(3):161–175.

Neto, A. C. and Levada, A. L. M. (2020). Isomap-KL: a parametric approach for unsupervised metric learning. In *2020 33rd SIBGRAPI Conf. on Graphics, Patterns and Images (SIBGRAPI)*, pages 287–294.

Oliveira, L. L., Jr., L. A. O., Silva, G. W. A., Timoteo, R. D. A., and Cunha, D. C. (2019). An RSS-based regression model for user equipment location in cellular networks using machine learning. *Wireless Netw*, 25:4839–4848.

Qi, G., Jin, Y., and Yan, J. (2018). RSSI-based floor localization using principal component analysis and ensemble extreme learning machine technique. In *2018 IEEE 23rd Int. Conf. on Digital Signal Processing (DSP)*, pages 1–5.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-Validation*. Springer US, Boston, MA.

Shafique, K., Khawaja, B. A., Sabir, F., Qazi, S., and Mustaqim, M. (2020). 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*, 8:23022–23040.

Yousaf, M., Khan, M. S. S., Rehman, T. U., Ullah, S., and Li, J. (2021). NRIC: A noise removal approach for nonlinear isomap method. *Neural Process. Lett.*, 53(3):2277–2304.