Análise e predição de ocorrência de malária no Estado do Amazonas: resultados preliminares utilizando modelo LSTM

Matheus Félix Xavier Barboza[1], Guto Leoni Santos[2,3], Iago R. R. Silva[2,3], Theo Lynn[3], Vanderson de Souza Sampaio[4,5], Patricia Takako Endo[1,3]

[1]Universidade de Pernambuco (UPE)
Pernambuco – Brasil

[2]Universidade Federal de Pernambuco (UFPE)
Pernambuco – Brasil

[3]Dublin City University (DCU)
Dublin – Irlanda

[4]Fundação de Vigilância em Saúde do Amazonas (FVS-AM)
Amazonas – Brasil

[5]Fundação de Medicina Tropical Dr. Heitor Vieira Dourado (FMT-HVD)
Amazonas – Brasil

matheus.barboza@upe.br, {guto.leoni, iago.silva}@gprt.ufpe.br

theo.lynn@dcu.ie, vandersons@gmail.com, patricia.endo@upe.br

Resumo. *Através de dados fornecidos pelo Sistema de Informação de Agravos de Notificação (SINAN), de 2003 à 2018, foi constatado que o Estado do Amazonas possui um altíssimo índice de casos diagnosticados de malária. O presente artigo apresenta na análise desses dados por meio da aplicação do algoritmo de clusterização k-means e modelo de predição utilizando deep learning. O objetivo principal é predizer a quantidade de ocorrências de malária na região, apresentando modelos específicos por cluster. Inicialmente, este trabalho apresenta os resultados referentes a cidade de Manaus, que por sua particularidade, está isolada das demais cidades, em um cluster. Resultados mostram que o modelo Long-Short Term Memory (LSTM) proposto apresenta um bom resultad, com RMSE de 0,0362.*

1. Introdução

A malária é uma doença que proporciona transtornos na saúde pública em diversas zonas tropicais do mundo [Cowman et al. 2016]. Lugares como países africanos (como os localizados no sul do deserto do Saara), sudeste asiático e Amazônia apresentam problemas graves na saúde pública devido a forte incidência da doença nestas regiões. Pesquisas apontam que mais de 200 milhões de pessoas por ano são acometidas por esta doença, resultando em aproximadamente 600 mil mortes [Cowman et al. 2016]. Um paciente que contém a doença apresenta sintomas tais quais febre intermitente, dor muscular e cefaleia [Waitumbi et al. 2010]. Segundo a Organização Mundial da Saúde (OMS), no ano de 2015, o Brasil registrou o maior número de casos de malária entre todos os países das

Américas[1] e 99% dos casos naturais do território são registrados na Região Amazônica[2], dado esse que preocupa não somente a população local, como também os órgãos de saúde pública.

A ONU possui um programa chamado Agenda 2030 para o Desenvolvimento Sustentável[3] que tem como objetivo um plano de ação global que visa melhorar o mundo com os seus países membros e indica 17 Objetivos de Desenvolvimento Sustentável (ODS). Dentre os 17 ODS, o ODS 3 - Saúde e Bem-Estar - tem como objetivo assegurar uma vida saudável e promover o bem-estar para todos, em todas as idades. Um dos objetivos do ODS 3 até 2030 é proporcionar políticas de saúde pública que acabem com epidemias como AIDS, tuberculose, malária, dentre outras doenças.

A criação de estratégias de combate a doenças pode tornar-se mais eficaz caso o acesso à informações relevantes permaneça disponível para profissionais da saúde e cientistas de dados. Assim, buscou-se criar um sistema de apoio a decisão que visa predizer o número de ocorrências de malária com base em dados históricos. Para isso, utilizou-se uma base de dados do Sistema de Informação de Agravos e Notificação (SINAN), que registra uma série de informações relacionadas aos casos de malária em todo Brasil. Com base nos dados recuperados do SINAN, é possível realizar uma série de análises estatísticas com o intuito de identificar padrões e criar modelos para predizer o comportamento das ocorrências de malária. Essa pesquisa focou no estado do Amazonas, dado a grande quantidade de ocorrências da doença nessa região.

Verificou-se que o número de ocorrências de malária não é constante entre as cidades do Amazonas, pois as cidades possuem diferentes características, tais como população, proximidade da floresta, o fato de possuir rios, etc. Assim, após análises estatísticas, foi definida uma metodologia baseada em modelos de aprendizado de máquina, uma vez que são capazes de extrair padrões de grandes conjuntos de dados [Anzai 2012]. Primeiramente, foi utilizada a técnica de *clusterização k-means* para categorização das cidades com características similares. Uma vez que os grupos forem definidos, um modelo de predição baseado em redes *long-short term memory* (LSTM) foi utilizado para estimar o número de casos de malária das cidades, com base no número de ocorrências dos últimos anos [Valenca 2010].

Este trabalho tem como objetivo apresentar: *(a)* uma análise exploratória preliminar dos dados sobre malária notificados no Sistema de Informação de Agravos de Notificação (SINAN) e; *(b)* a metodologia adotada para a criação de modelos de predição específicos por cidade.

2. Metodologia

2.1. Base de dados

O SINAN[4] armazena dados de notificações de casos de doenças e agravos que se encontram na lista nacional de doenças de notificação compulsória. O sistema foi implantado em 1993, mas foi distribuído de forma homogênea por todo o território brasileiro apenas

[1]https://www.who.int/malaria/publications/world-malaria-report-2016/report/en/
[2]https://portal.fiocruz.br/noticia/malaria-regiao-amazonica-concentra-99-dos-casos-no-brasil
[3]http://www.agenda2030.com.br
[4]http://portalsinan.saude.gov.br

em 1998[5]. Desde lá, sua utilização tem permitido o fornecimento de dados para análise por meio de uma rede informatizada com o objetivo de dar suporte aos profissionais de saúde para medidas de intervenção.

A base utilizada para este trabalho contém dados cadastrados desde Janeiro de 2003 até Dezembro de 2018. Isso totaliza aproximadamente 6 milhões de registros de malária em todo o Estado do Amazonas. A Figura 1 apresenta a série temporal referente a quantidade de ocorrências de malária por mês no Estado do Amazonas.

É possível observar que entre os anos 2003 e 2007, o número de casos foi mais alto que nos anos subsequentes, chegando a 30.000 registros em Julho de 2005. No entanto, após 2008, o número de registros diminuiu de forma considerável. Também é possível observar que entre os meses de Julho e Agosto, há um crescimento no número de registros de malária na região.

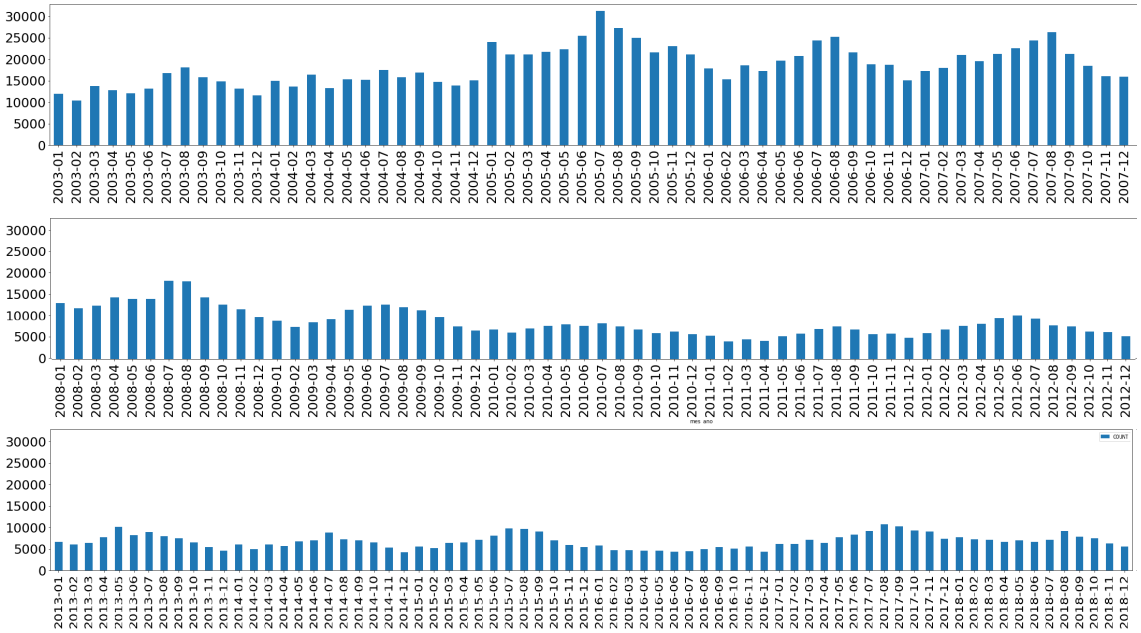

Figura 1. Série temporal referente ao número mensal de ocorrências de malária no Estado do Amazonas durante os anos de 2003 à 2018.

A Figura 2 apresenta o número total de registros por cidade a partir do ano de 2003. Pode-se perceber que a quantidade de notificações tem uma grande variação por município. Manaus, por ser a cidade mais populosa do estado, é a cidade com o maior número de registros. No entanto, mesmo em cidades menores, o número de registros não segue um padrão.

Essa heterogeneidade do número de registros de malárias das cidades do Amazonas afeta fortemente no processo de predição utilizando *deep learning*. Por exemplo, um modelo que consiga predizer o número de casos da cidade de Manaus pode não conseguir predizer bem para a cidade de Anorí, onde o número de ocorrências é muito baixo. Assim, dividir as cidades em *clusters* com base em características em comum pode facilitar a criação de modelos de predição.

[5]http://portalsinan.saude.gov.br/o-sinan

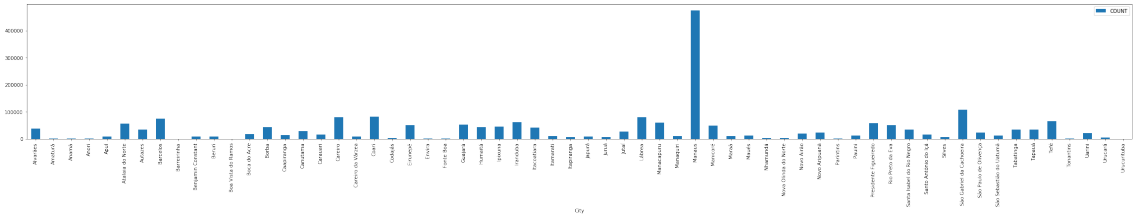

Figura 2. Quantidade de ocorrências de malária no Estado do Amazonas por município

2.2. Clusterização

A clusterização de dados é uma técnica da mineração de dados na qual não se aplica classes previamente definidas para identificação de grupos (*clusters*). Esta é uma técnica de aprendizado não-supervisionado, considerando somente informações das variáveis para reconhecer e agrupar automaticamente dados com características estatisticamente semelhantes. Estes dados podem seguir o mesmo padrão [Saxena et al. 2017].

2.2.1. K-means

O algoritmo *k-means* é um dos métodos mais conhecidos para dar procedimento à técnica de clusterização. Ele consiste em particionar uma quantidade predefinida de *clusters* contendo dados que possuem características estatisticamente similares, para fornecer uma classificação de informações não-supervisionadas de acordo com os próprios dados. O algoritmo faz a comparação baseando-se na proximidade entre o valor médio dos objetos de acordo com a distância euclidiana [Arora et al. 2016].

Após o agrupamento, são definidos os centroides, pontos centrais que são inicialmente relacionados à um único *cluster* de forma aleatória. A cada etapa do algoritmo, novas atribuições são definidas ao grupo de acordo com a posição do centroide e dos objetos relacionados. A Eq. 1 define a primeira etapa do algoritmo [Arora et al. 2016], durante a atribuição do objeto ao centroide mais próximo:

$$S_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \leq \|x_p - \mu_j^{(t)}\|^2 \; \forall j, 1 \leq j \leq k\} \tag{1}$$

onde S é o valor da distância entre o objeto e o centroide; i indica a quantidade de iterações; x_p é o número de atribuições ao ponto mais aproximado; $\mu^{(t)}$ é o valor do centroide; j é o valor da medida de dissimilaridade; e k é referente ao número de *clusters*.

Após esta etapa, os pontos centrais são reposicionados, calculando a média das observações atribuídas aos respectivos pontos centrais, conforme a Eq. 2:

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2}$$

O processo de clusterização citado anteriormente foi necessário para a classificação das médias de registros de malária de cada município do estado.

2.3. Long-Short Term Memory

O processo de análise de dados sequenciais executado por redes neurais artificiais tradicionais possui uma limitação: dados passados não são considerados durante o processo. Isso pode se tornar um problema em casos de bases de dados em que o seu histórico é importante para predição.

As Redes Neurais Recorrentes (do inglês, *Recurrent Neural Networks* (RNN)) são mais eficazes para este tipo de situação, pois são capazes de trabalhar com conexões anteriores; assim, as decisões tomadas são baseadas simultaneamente em informações precedentes e recentes [Zen et al. 2016].

As redes LSTM são tipos especiais de RNN, capazes de trabalhar com maiores períodos de tempo [Kalchbrenner et al. 2015]. Esta característica permite uma vantagem em relação a outros tipos de redes neurais artificiais por geralmente obter um resultado mais eficiente. Isso se deve ao contexto dos dados analisados, baseando-se em seu histórico para contornar o problema de dependência de longo prazo.

Dessa forma, este trabalho utiliza redes LSTM para a proposição do modelo preditivo, devido à necessidade de verificar o comportamento dos dados anteriores a fim de se obter um resultado satisfatório.

3. Resultados preliminares

3.1. Clusterização das cidades

Neste trabalho, o objetivo principal da clusterização dos dados é definir grupos de cidades do Estado do Amazonas que estejam estatisticamente mais próximas, para que assim seja possível propor modelos *deep learning* mais específicos, por *cluster*, ao invés de se utilizar outro tipo de organização, por exemplo político-geográfico. Para tanto, calculou-se a média e a mediana de ocorrências por cidade, e aplicou-se o algoritmo *k-means* [Jain 2010].

A Figura 3 apresenta o mapa do Estado do Amazonas, incluindo os 62 municípios, com os cinco grupos de cidades definidos pelo *k-means*. Cada *cluster* contém a média e mediana das ocorrências por município durante dezesseis anos, de Janeiro de 2003 à Dezembro de 2018. Como notado anteriormente, pode-se ratificar que Manaus é uma cidade com ocorrência bastante elevada, ficando isolada em um cluster sem outras cidades. De acordo com os dados levantados, a cidade tem uma média de 845,9 notificações por mês durante o período analisado, ficando com a maior taxa de casos diagnosticados em todo o estado.

3.2. Resultados de predição utilizando LSTM

Considerando os resultados da clusterização apresentados anteriormente, como ponto de partida, este trabalho considerou apenas a cidade de Manaus para criar o modelo de predição baseado em *deep learning*. A justificativa é devido ao fato de Manaus ter sido classificada como o município com maior índice de ocorrências durante o intervalo de tempo estudado, de acordo com os dados obtidos, o número de casos diagnosticados foi de aproximadamente 470 mil casos apenas nesta cidade, correspondendo à 22,37% das ocorrências em todo o Estado do Amazonas. Durante os anos de 2003 a 2007 o número de casos foi mais alto, chegando a aproximadamente 8.000 casos por mês. No decorrer

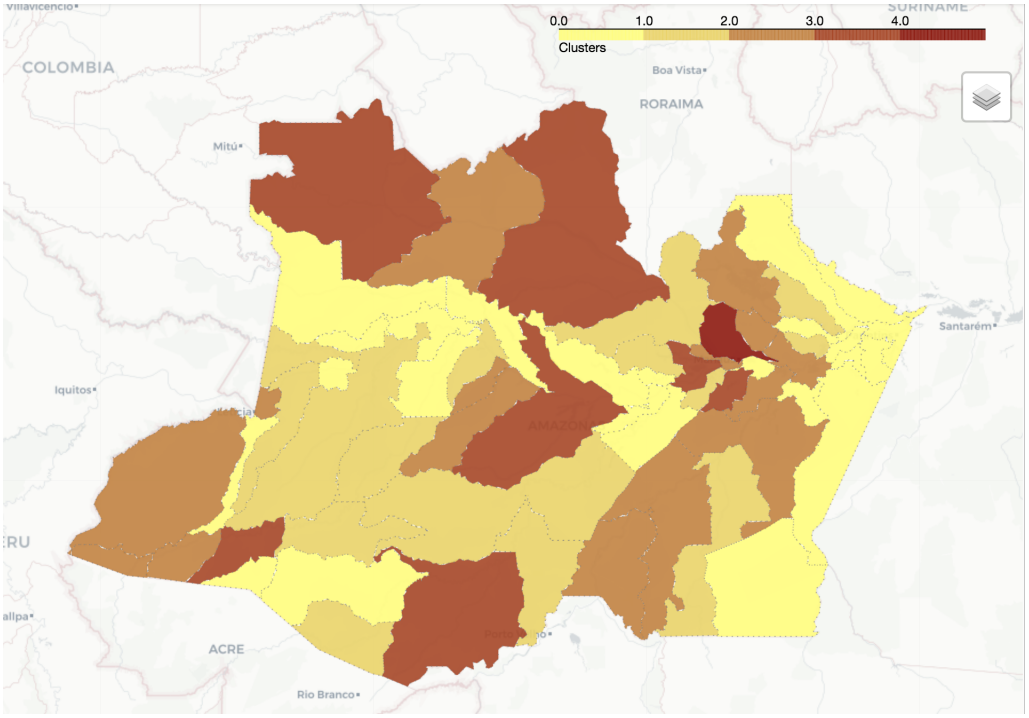

Figura 3. Gráfico mostra a incidência de malária diagnosticada na população amazonense entre este período

dos anos seguintes o número foi decrescendo consideravelmente. A Figura 4 apresenta a série temporal referente ao número de registros de malária apenas na cidade de Manaus.

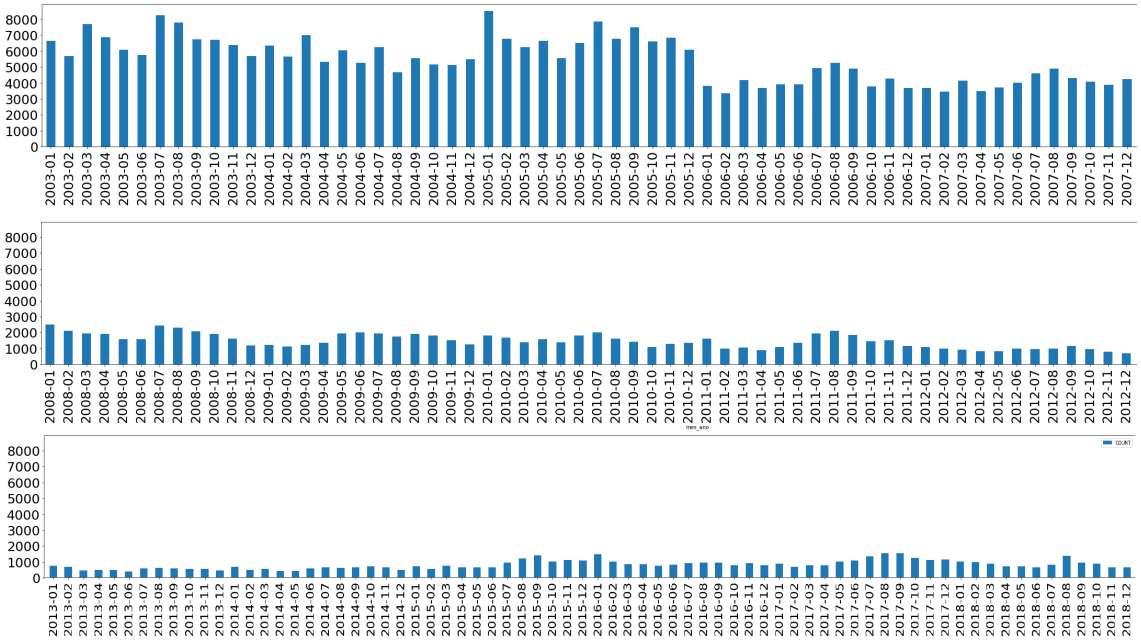

Figura 4. Série temporal referente a ocorrência de malária na cidade de Manaus

A Figura 5 apresenta o modelo *deep learning* utilizado para predição de ocorrências de malária na cidade de Manaus. Foram utilizadas duas camadas de redes

neurais do tipo LSTM com cinquenta unidades por camada, com $droupout = 20\%$, e uma camada dense. 80% dos dados históricos (de Janeiro de 2003 a Outubro de 2015) foram utilizados para treinamento do modelo, enquanto 20% (outubro de 2015 a dezembro de 2018) foram utilizados para teste.

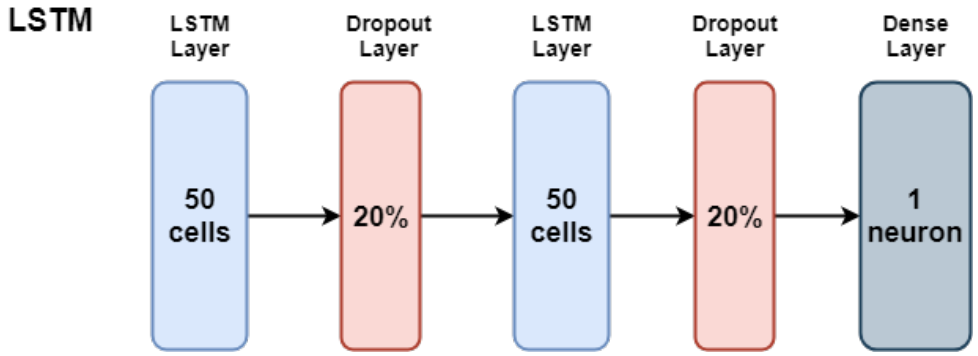

Figura 5. Modelo *deep learning* referente a ocorrência de malária na cidade de Manaus

A Figura 6 apresenta os resultados preditos pelo modelo LSTM (em laranja) em comparação com os dados reais (em azul). É possível notar que os dados preditos seguem um padrão bastante semelhante aos dados reais.

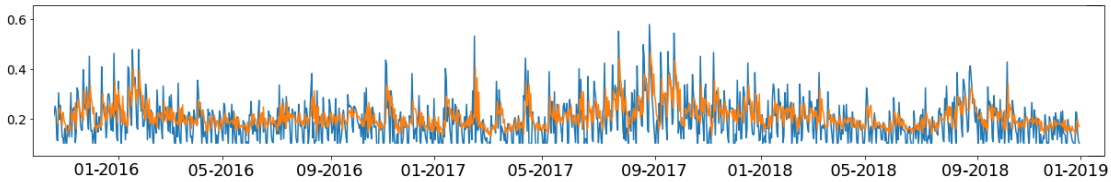

Figura 6. Resultado referente a predição da ocorrência de malária na cidade de Manaus.

Para avaliar quantitativamente o modelo LSTM proposto, utilizou-se métrica *root-mean-square error* (RMSE), definida pela Eq. 3:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{3}$$

A métrica aplicada no modelo é análoga ao desvio padrão, medida utilizada dentro da área de estatística para expressar o grau de dispersão de um conjunto de dados e verificar a sua homogeneidade [Leonardi et al. 2009]. Quanto mais seu valor se aproxima de 0, maior é a sua aproximação da média. Esta métrica leva em consideração o valores reais da série temporal e a diferenciação em relação aos valores que foram preditos pelo modelo proposto. Ao final, o número referente à métrica RMSE obtido pelo modelo foi de aproximadamente 0,0362.

4. Conclusões e próximos passos

A malária é uma doença tropical que ainda causa transtornos na saúde pública. O estado do Amazonas, por exemplo, enfrenta graves problemas devido a forte incidência desta

doença em suas cidades, chegando a registrar 30.000 casos em Julho de 2005, e a apresentar o maior registro de casos de malária entre todos os países da América em 2015. Apesar dos casos terem reduzido consideravelmente nos últimos 9 anos, a malária ainda é considerada epidêmica na região.

Neste trabalho, foi apresentado um modelo de predição de ocorrências de casos de malária, considerando dados do estado do Amazonas. Foi utilizado o algoritmo *k-means* para agrupamento e clusterização das cidades com características similares. E uma rede neural LSTM foi proposta para predizer as ocorrências de malária na cidade de Manaus. Os resultados mostram-se satisfatórios, uma vez que observando-se a métrica adotada (RMSE), o resultado obtido (erro) foi baixo, e apresenta um potencial de capacidade de predizer novos casos da doença.

Como próximos passos, planeja-se propor e comparar um modelo para cada *cluster* (um conjunto de cidades) e analisar a hipótese de que ter modelos por *cluster* apresentará resultados melhores do que ter um único modelo para o estado do Amazonas. Também objetiva-se a utilização e análise comportamental de outras redes neurais e algoritmos de aprendizado de máquina, tais quais Redes Neurais Recorrentes, *Gated Recurrent Unit* (GRU) e *Random Forest*.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, da Fundação de Amparo a Ciência e Tecnologia de Pernambuco (FACEPE) - Pernambuco - Código de Financiamento IBPG-0059-1.03/19 e do Irish Institute of Digital Bussiness (dotLAB), Irlanda.

Referências

Anzai, Y. (2012). *Pattern recognition and machine learning*. Elsevier.

Arora, P., Varshney, S., et al. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512.

Cowman, A. F., Healer, J., Marapana, D., and Marsh, K. (2016). Malaria: biology and disease. *Cell*, 167(3):610–624.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Kalchbrenner, N., Danihelka, I., and Graves, A. (2015). Grid long short-term memory. *arXiv preprint arXiv:1507.01526*.

Leonardi, F., Oliveira, C., Fonseca, L. M. G., and Almeida, C. d. (2009). Fusão de imagens cbers 2b: Ccd-hrc. *Simpósio Brasileiro de Sensoriamento Remoto (SBSR)*, 14:6951–6958.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681.

Valenca, M. J. S. (2010). Fundamentos das redes neurais: Exemplos em java. 2ª edição. ed. *Olinda: Livro Rápido*.

Waitumbi, J. N., Kuypers, J., Anyona, S. B., Koros, J. N., Polhemus, M. E., Gerlach, J., Steele, M., Englund, J. A., Neuzil, K. M., and Domingo, G. J. (2010). Outpatient upper respiratory tract viral infections in children with malaria symptoms in western kenya. *The American journal of tropical medicine and hygiene*, 83(5):1010–1013.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., and Szczepaniak, P. (2016). Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. *arXiv preprint arXiv:1606.06061*.