UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Airton Martins Neris

Avaliação de métodos para interpolação espacial de dados de precipitação

Recife

2019

Airton Martins Neris

Avaliação de métodos para interpolação espacial de dados de precipitação

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientador: Glauco Estácio Gonçalves

Coorientador: Victor Wanderley Costa de Medeiros

Recife

2019

Dedico este trabalho aos meus pais, Marcia de Assis Neris Martins e Manoel Ferreira Martins Filho. O apoio e confiança dedicados por eles foi de extrema importância para concluir minha graduação. Sou eternamente grato pelos esforços empenhados por eles para me ajudar, e também pela compreensão que eles tiveram em alguns momentos difíceis dessa jornada. Então dedico todo o valor da minha graduação a eles. Obrigado por tudo.

Agradecimentos

Agradeço primeiramente a Deus por todas as oportunidades e sinais dados a mim para seguir os melhores caminhos.

Aos meus orientadores pelo exemplo de profissionais e pela orientação desde o projeto de pesquisa até este trabalho de conclusão de curso durante meus momentos de orientando no JuaLabs. A ajuda e compreensão deles foi de extrema importância para chegar até aqui, serei eternamente grato.

Aos meus pais pelo incentivo e confiança em mim em todas as minhas decisões. Vocês foram a principal causa para eu ter chegado até aqui.

Em especial a Taiane Viana, que me deu grande apoio em vários aspectos, sempre disposta a me ajudar, esteve comigo de perto principalmente nos momentos difíceis e finais deste trabalho, deixo aqui minha gratidão a ela.

Aos meus amigos do curso de Bacharelado em Sistema de informação na UFRPE. A ajuda de vocês foi essencial para esta conquista, obrigado por tudo.

Aos meus amigos do trabalho, que me incentivaram e me ajudaram a sanar dúvidas técnicas em alguns aspectos dos conteúdos de aprendizado de maquina.

“Agonia, êxtase, paz...cada passagem tem sua própria beleza”
(Karthus)

Resumo

Informação sobre a quantidade de precipitação de chuva é essencial para os mais variados setores, como agrícola e agroflorestal. Apesar dessa importância, muitas áreas ainda não possuem estações meteorológicas, o que ocasiona a falta de dados. Para suprir essa necessidade existem os métodos de interpolação espacial, que utilizam as informações de pontos correlatos para estimar o valor inexistente em determinada área. Assim, este trabalho tem como objetivo avaliar métodos para a interpolação de dados diários de precipitação. As técnicas de interpolação utilizadas nos experimentos foram os métodos: Ponderação pelo Inverso da Distância; Krigagem Ordinária; Floresta Aleatória. Para a Floresta Aleatória foram usadas duas configurações distintas, uma que recebe como entrada as coordenadas, e outra que recebe a distância de *buffer*, que é um dos mais recentes pré-processamentos utilizados na literatura para que a Floresta Aleatória estime seus valores com base no seu referencial geográfico. Foram utilizados dados de precipitações de chuva provenientes das 46 estações meteorológicas do estado de Pernambuco referentes ao período de 2013 a 2018, e para comparar a precisão da generalização dos métodos, foi utilizado a validação cruzada *leave-one-out*. Nos resultados, a Ponderação pelo Inverso da Distância apresentou um melhor desempenho em suas estimativas, para todas as métricas, e a Floresta Aleatória utilizando coordenadas obteve o segundo melhor resultado. A Floresta Aleatória utilizando a distância de *buffer*, teve um resultado inferior em termos de suas métricas, mas a qualidade da espacialização visual mostrou-se superior por oferecer um resultado visualmente mais suave do que aquele oferecido pela Floresta Aleatória utilizando coordenadas.

Palavras-chave: Precipitação, Aprendizagem de Máquina, Interpolação espacial, distância de *buffer*

Abstract

Information on the amount of rainfall is essential for the most varied sectors, such as agriculture and agroforestry. Despite this importance many areas are still not covered by meteorological stations, which causes the lack of data. To meet this need there are methods of spatial interpolation, which use the information of correlated points to estimate the value that does not exist in a certain area. Thus, this work aims to evaluate methods for the interpolation of daily precipitation data. The interpolation techniques used in the experiments were the methods: Inverse Distance Weighting; Ordinary Kriging; Random Forest. For the Random Forest two different configurations were used, one that receives as input the coordinates, and another that receives the *buffer* distance, which is one of the most recent pre-processing used in the literature for the Random Forest to estimate its values based on geographical reference. We used rainfall data from the 46 meteorological stations from the state of Pernambuco in the period from 2013 to 2018, and to compare the precision of the generalization of the methods, we used the *leave-one-out* cross validation. In the results, the Inverse Distance Weighting presented a better performance in its estimates, for all the metrics, and the Random Forest using coordinates obtained the second best result. Random Forest using *buffer* distance had a lower result in terms of its metrics, but the quality of visual spatialization proved to be superior by offering a visually smoother result than offered by Random Forest using coordinates.

Keywords: Precipitation, Machine Learning, Spatial Interpolation, Buffer Distance

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INMET	Instituto Nacional de Meteorologia
ANA	Agência Nacional de Águas
IDW	Inverse Distance Weighting (Ponderação pelo Inverso da Distância)
OK	Ordinary kriging (Krigagem Ordinária)
RF	Random forest (Floresta Aleatória)
RFsp	Random Forest for Spatial Predictions (Floresta Aleatória para Previsões Espaciais)
R	Coeficiente de Correlação
RMSE	Raiz do Erro Médio Quadrático
MAE	Erro Médio Absoluto
BIAS	Viés

Sumário

1 Introdução

Segundo Silva et al. (2019) a precipitação pluviométrica seria a fase do ciclo hidrológico que tem o papel de retornar as águas condensadas da atmosfera para a superfície em diversas formas, como chuva, granizo, neblina, neve, orvalho ou geada. Esse fenômeno acontece quando as nuvens estão carregadas de vapor de água, que ao atingirem uma altitude elevada ou encontrarem-se com massas de ar frio, condensam os vapores de água fazendo-os caírem na atmosfera em forma de chuva.

O mapeamento da precipitação pluviométrica e de outras variáveis ambientais é de fundamental importância no setor agrícola (VIOLA et al., 2010), para isso existem estações pluviométricas que fazem a coleta da quantidade de precipitação (mm por intervalo de tempo). Os dados dessas estações muitas vezes são armazenados e disponibilizados por agências de meteorologia na forma de serviços públicos digitais. Os dados de precipitação são organizados em séries históricas, referentes às medições horárias, diárias ou mensais das estações meteorológicas, de acordo com as normas definidas pela Organização Meteorológica Mundial (OMM) (JúNIOR, 2018).

Devido a sua extensão, a disponibilidade de estações pluviométricas no Brasil ainda é pequena (XAVIER; KING; SCANLON, 2016) fazendo-se necessário a utilização de técnicas de interpolação espacial para suprir os valores nas localidades onde esses equipamentos não estão presentes.

1.1 Motivação

Saber o volume de precipitação pluviométrica em uma determinada região geográfica é essencial para diversos setores, como por exemplo os setores agrícola e agroflorestal. Tais dados ambientais são usados para o monitoramento, análise e compreensão dos processos e dos fenômenos naturais de uma determinada região. A coleta de tais parâmetros é feita hoje essencialmente por órgãos públicos e instituições de pesquisa que dispõe de redes de estações pluviométricas espalhadas pelo território brasileiro tal como o Instituto Nacional de Meteorologia (INMET)[1] e a Agência Nacional de Águas (ANA)[2].

Apesar da existência dessas redes de estações pluviométricas e levando-se em conta a extensão do país, pode-se dizer que a área coberta por pluviômetros ou estações climáticas ainda é pequena no país. No estado de Pernambuco, por exemplo,

[1] http://www.inmet.gov.br/portal/
[2] https://www.ana.gov.br/

existem limitações na disponibilidade destes dados, tanto em termos de densidade da rede pluviográfica, como em relação ao pequeno período de observações disponível (SILVA et al., 2019). Essas condições ocasionam a falta de dados pluviométricos para determinadas áreas do estado e para sanar esse tipo de problema faz-se necessário a utilização de métodos para estimar a precipitação pluviométrica dessas localidades.

Existem diversas técnicas de interpolação espacial, muitas delas utilizadas para estimar valores para bases de dados pluviométricos. Os tipos mais utilizados podem ser classificados em três grandes grupos: as técnicas de interpolação não-geoestatísticas, aquelas baseadas em geoestatística e as baseadas em aprendizagem de máquina (LI; HEAP, 2014).

Este trabalho tem como foco a avaliação dos métodos de interpolação bastante utilizados na literatura e com bom desempenho nas avaliações. Os métodos utilizados nos experimentos desse trabalho foram: Ponderação pelo Inverso da Distância (*Inverse Distance Weighting* - IDW)(LU; WONG, 2008); Krigagem Ordinária (*Ordinary Kriging* - OK)(CRESSIE, 1988); Floresta Aleatória (*Random Forest* - RF)(BREIMAN, 2001); e Floresta Aleatória para Previsões Espaciais (*Random Forest for Spatial Predictions* - RFsp)(HENGL et al., 2018). Em particular, a técnica RFsp foi empregada neste trabalho por não ter sido até então, no melhor do nosso conhecimento, utilizada em experimentos mais extensos com dados de precipitação pluviométrica.

1.2 Objetivos

O objetivo principal deste trabalho é avaliar, através de experimentos, técnicas de interpolação para dados de precipitação de chuva diária no estado de Pernambuco, utilizando o recorte temporal do período de 2013 até 2018. Serão empregadas técnicas clássicas da área de interpolação espacial e técnicas mais recentes da aprendizagem de máquina. Com isso pretende-se atingir os seguintes objetivos específicos:

- Avaliar por meio de métricas de validação a capacidade de interpolação dos métodos para o cenário sugerido
- Comparar qualitativamente os gráficos gerados por cada uma das técnicas.

1.3 Organização do trabalho

O trabalho está organizado em cinco capítulos. Além deste capítulo de introdução, o Capítulo 2 explica o funcionamento dos métodos de interpolação dos dados de precipitação utilizados e a revisão dos trabalhos que realizaram experimentos correlatos aos efetuados neste trabalho. O Capítulo 3 descreve a metodologia utilizada

na montagem dos experimentos executados. O Capítulo 4 apresenta a discussão dos resultados. Por fim, as considerações finais e os trabalhos futuros são apresentados no Capítulo 5.

2 Referencial teórico

2.1 Interpolação de dados pluviométricos

Uma correta análise da distribuição espacial da precipitação pluviométrica é de bastante importância para o planejamento dos recursos hídricos nas bacias hidrográficas, e também são de grande apoio a estudos climatológicos e meteorológicos (MARCUZZO; ANDRADE; MELO, 2011). Um exemplo de setor que se utiliza desses dados seria o agroflorestal, no qual a utilização desses dados auxilia nas ações de modelagem da produção florestal, como nos modelos ecofisiológicos (VIOLA et al., 2010), onde esses modelos ajudam na compreensão, predição e controle de todo um sistema florestal, auxiliando no zoneamento florestal e mapeamento da produção. Dado que muitas vezes poucas áreas são cobertas por estações pluviométricas, faz-se necessário a utilização de métodos de interpolação espacial para suprir a falta desses dados. Essa interpolação pode ser feita por técnicas que utilizam recursos estatísticos ou até mesmo de aprendizagem de máquina.

Esta seção apresenta os quatro métodos de interpolação utilizados nas avaliações deste trabalho: IDW (Seção 2.1.1), OK (Seção 2.1.2), RF (Seção 2.1.3) e RFsp (Seção 2.1.4).

2.1.1 Ponderação pelo Inverso da Distância

A IDW é um dos métodos de interpolação espacial mais popularmente utilizados. Este método consiste na média ponderada espacial baseada nos valores dos pontos vizinhos ao ponto em que se deseja interpolar. Assim é possível combinar a mudança gradual no valor estimado de acordo com a distância (BABAK; DEUTSCH, 2009). O método se baseia no valor dos vizinhos e quanto maior a distância de um ponto a ser interpolado, menor será seu peso, i.e., menor será a sua influência no valor do ponto a ser estimado (SILVA et al., 2019).

O IDW pode ser formalmente apresentado como (LU; WONG, 2008)

$$\widehat{y}(S_0) = \sum_{i=1}^{n} \lambda_i y(S_i) \tag{2.1}$$

onde $\widehat{y}(S_0)$ é o valor que pretende-se interpolar em um ponto S_0, dado os valores observados $y(S_i)$ nos pontos S_i, por último, os valores λ_i são os pesos, deve-se observar que $\sum_{i=1}^{n} \lambda_i = 1$. Basicamente o valor a ser estimado em $\widehat{y}(S_0)$ é a combinação linear

entre os pesos e os valores nos pontos observados. Estes pesos são definidos como:

$$\lambda_i = d_{0i}^{-\alpha} / \sum_{i=1}^{n} d_{0i}^{-\alpha} \tag{2.2}$$

onde d_{0i} é o inverso da distância entre o ponto a ser estimado e os pontos observados e α é a potência. Note-se que a potência é um parâmetro deste método e que quanto maior o valor de α, maior será a contribuição dos pontos mais próximos para o valor estimado $\widehat{y}(S_0)$ e, consequentemente, menor será a contribuição dos pontos mais distantes.

2.1.2 Krigagem

Como definido em (MARCUZZO; ANDRADE; MELO, 2011), a Krigagem é um método geoestatístico que é baseado na Teoria das Variáveis Regionalizadas, que tem como suposição basilar que a variação espacial de um determinado fenômeno é estatisticamente homogênea em uma área. Neste método, a variação espacial é mensurada pelo semivariograma, que nada mais é que um gráfico de dispersão da semivariância em relação a distância dos pontos. Esse semivariograma é utilizado para a análise da dependência espacial entre as amostras, a Equação 2.3 (MARCUZZO; ANDRADE; MELO, 2011) mostra como o variograma é calculado.

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{s} \{Z(x_i) - Z(x_i + h)\}^2 \tag{2.3}$$

Nesta equação, h é a distância entre os pontos, $\gamma(h)$ é a semivariância para a distância h, n é o número de pontos amostrados separados pela distância h, s é a quantidade de pares de pontos separada pela distância h, $Z(x_i)$ é o valor da amostra na posição x_i, e $Z(x_i+h)$ é o valor da amostra na posição separada da posição x_i pela distância h. No trabalho de Silva et al. (2019) são apresentados diversos modelos matemáticos que são convencionalmente ajustados ao semivariograma, são eles: cúbico, exponencial, gaussiano, linear, logarítmico, pentasférico, potência (*Power*), quadrático, quadrático racional, esférico e Onda (*Wave*).

Neste trabalho utilizamos o modelo esférico ilustrado na Figura 1, e representado por MARCUZZO, ANDRADE e MELO (2011) no sistema de equação 2.4

$$\begin{aligned} \gamma(h) &= \begin{cases} c_0 + c\left(\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right), & \text{se } 0 < h \leq a \\ c_0 + c, & \text{se } h > a \end{cases} \\ \gamma(0) &= c_0 \end{aligned} \tag{2.4}$$

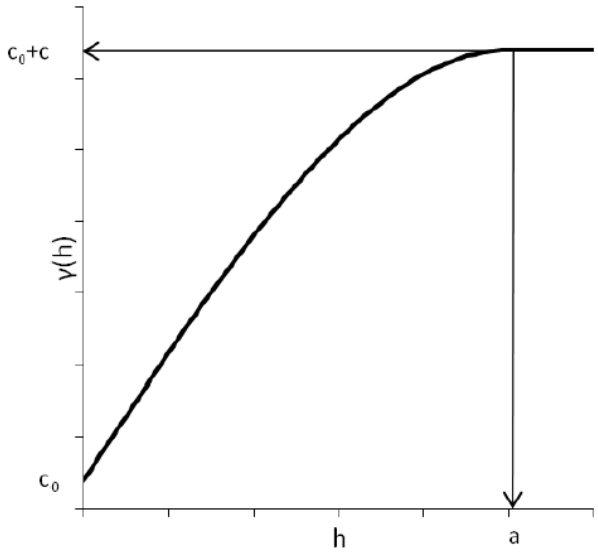


Figura 1 – Modelo esférico

em que a é a distância a partir da qual não há mais correlação espacial entre as variáveis, e o c_0 (chamado de *nugget*) é o valor de γ para distâncias iguais a zero, que indica as variações para distâncias muito pequenas, e $c_0 + c$ (chamado de *sill*) é o valor médio da semivariância além da distância a. O valor de c também é chamado de *partial sill*.

Existem vários tipos de Krigagem, como a Krigagem Simples (*Simple Kriging*), Krigagem Ordinária (*Ordinary Kriging*), Krigagem Universal (*Universal Kriging*) e outras. Para a Krigagem Ordinária o sistema de equação 2.5 é utilizado para determinar os pesos ω_i, cuja soma dos pesos é igual a 1. Neste sistema, ω_i é o peso para cada ponto, é uma variável temporária, $\gamma(h)$ é o valor de semivariância para pontos separados por uma distância h, h_{ij} é a distância entre x_i e x_j, h_j é a distância entre o ponto que se deseja calcular o valor de x sobre o ponto x_j.

$$\begin{array}{l} \sum_{i=1}^{N} \omega_i \gamma(h_{ij}) + \mu = \gamma(h_j), \quad j = 1, ..., n \\ \sum_{i=1}^{N} \omega_i = 1 \end{array} \tag{2.5}$$

No OK a interpolação de um ponto é representada pela equação 2.6, onde $Z(x)$ é o valor do ponto que se deseja estimar, n é a quantidade de pontos amostrados cujos valores serão usados na interpolação, $Z(x_i)$ é o valor do ponto, e ω_i é o valor do peso de $Z(x_i)$ sobre o ponto x (MARCUZZO; ANDRADE; MELO, 2011).

$$Z(x) = \sum_{i=1}^{n} \omega_i Z(x_i) \tag{2.6}$$

2.1.3 Floresta Aleatória

A RF é uma técnica de aprendizagem de máquina que combina o desempenho de diversas Árvores de decisão para classificar ou predizer o valor de uma variável (RODRIGUEZ-GALIANO et al., 2015). Árvores de decisão são algoritmos de aprendizagem de máquina supervisionados que podem ser utilizados tanto para problemas de classificação (Árvore de classificação) quanto de regressão (Árvore de regressão). A Árvore de regressão é uma variação da árvore de decisão feita para tornar possível a aproximação funções de valores reais(JúNIOR, 2018).

Dos diversos tipos de Árvores de Decisão existentes, tomamos como exemplo o método CART (Classification and Regression Tree). Como podemos ver no Algoritmo 1, empregando uma medida de impureza do nó com base na distribuição dos valores Y observados no nó, o Algoritmo 1 divide um nó pesquisando exaustivamente sobre todos os X e S para a divisão $\{X \in S\}$ que minimiza a impureza total de seus dois nós filhos (LOH, 2011). se X possui valores ordenados, o conjunto S é um intervalo do na forma $(-8, c]$. Do contrário, S é um subconjunto dos valores tomadas por X. O processo é feito recursivamente nos dados em cada nó filho (LOH, 2011).

Algoritmo 1: Pseudocódigo da árvore por busca exaustiva (LOH, 2011)

1. Inicie no nó raiz
2. Para cada X, encontre o conjunto S que minimiza a soma das impurezas do nó nos dois nós filhos e escolha a divisão $\{X^* \in S^*\}$ que dá o mínimo geral de X e S.
3. Se um critério de parada for atingido, pare. Senão aplique o passo 2 para cada nó filho, por sua vez

A RF representada por $\widehat{f}_{rf}^{K}(X)$ recebe um vetor de entrada X, contendo os valores das diferentes características analisadas para uma dada área de treinamento, consulta um número K de árvores de regressão $T(X)$ e com isso calcula a média dos resultados, como mostrado na Equação 2.7.

$$\widehat{f}_{rf}^{K}(X) = \frac{1}{K}\sum_{K-1}^{K} T(X) \tag{2.7}$$

Para evitar a correlação das diferentes árvores, a RF constrói uma diversidade de árvores, fazendo-as crescer a partir de diferentes subconjuntos de treinamento criados por meio de um procedimento chamado ensacamento (*Bagging*). O ensacamento é uma técnica usada para treinar a criação de dados, re-amostrando aleatoriamente o conjunto de dados originais com substituição (RODRIGUEZ-GALIANO et al., 2015).

Quando a RF faz uma árvore crescer, ela usa o melhor ponto de divisão dentro de um subconjunto de características evidenciais que foi selecionado aleatoriamente a partir do conjunto geral de recursos de evidência de entrada, isso pode diminuir a força de cada árvore, porém reduz a correlação entre as árvores, o que reduz o erro de generalização.

Como pode-se observar a RF é uma técnica aplicada para regressão e classificação em diversos contextos. Para interpolação espacial, como o caso de nosso trabalho, o vetor de entrada é comumente composto pela latitude e longitude das estações meteorológicas. Contudo, é possível modificar o vetor de entrada para considerar outras variáveis, como será apresentado na seção a seguir.

2.1.4 Floresta Aleatória para Previsões Espaciais

"Floresta aleatória, no entanto, ignora as localizações espaciais das observações e por isso qualquer autocorrelação espacial nos dados não são contabilizadas pelas covariáveis."(HENGL et al., 2018, p. 2). Em outras palavras o autor afirma que para a predição espacial com a RF, a correlação espacial dos pontos usados no treinamento é ignorada no processo de modelagem, o que pode levar a resultados enviesados e sub-ótimos. Assim, Hengl et al. (2018) propõe a técnica denominada Floresta aleatória para previsões espaciais (*Random Forest for Spatial Predictions* - RFsp). No RFsp são empregadas outras características relacionadas às distâncias entre os pontos de treinamento, ao invés de apenas utilizar as coordenadas como se faz na RF abordada na Seção 2.1.3, adicionando assim os efeitos de proximidade geográfica no processo de predição da RF.

Genericamente, a interpolação espacial $Y(s)$ da RFsp para uma coordenada s é dada por uma função f tal que

$$Y(s) = f(X_G, X_R, X_P) \tag{2.8}$$

onde X_G são as co-variáveis que representam a proximidade geográfica e relações espaciais entre os pontos observados, X_R são co-variáveis de refletância de superfície, como por exemplo imagens de sensoriamento remoto, e X_P são co-variáveis baseadas em processos, por exemplo o índice umidade topográfica(HENGL et al., 2018).

Para X_G podem ser empregadas diferentes co-variáveis tal como as coordenadas geofísicas, distância euclidiana para pontos de referência para uma determinada área de estudo, distância euclidiana sobre locais de amostragem, distâncias de declive, distância de resistência ou distância de *buffer* ponderadas. Neste trabalho, faremos $X_G = (d_{p1}, d_{p2}, ..., d_{pN})$, onde d_{pi} é a distância euclidiana entro local de amostragem para todos os locais de observação pertencentes à amostra. Para cada coordenada

pertencente aos pontos de observação, é criado um vetor de distâncias, cuja dimensão é igual à quantidade de pontos de amostragem no conjunto de treino.

Em seu trabalho, Hengl et al. (2018) demonstram, através de exemplos com diversas bases de interpolação espacial, que a incorporação do efeito de proximidade geográfica na RF melhora a a qualidade visual da interpolação espacial obtida quando comparada à RF convencional, obtendo uma qualidade próxima daquela vista em produtos obtidos por meio da Krigagem. Contudo, o autor conclui que dado alguns desafios metodológicos, como o cálculo da distância de *buffer* para um conjunto de pontos muito grandes, alguns métodos tradicionais podem ser mais adequados para promover a qualidade das interpolações.

2.2 Validação Cruzada

A validação cruzada (*cross validation*) é um método utilizado para avaliar a capacidade de generalização de determinada técnica, modelo ou algoritmo, dividindo os dados em duas partes, uma que utiliza para treinar o modelo e a outra fatia restante para validar a precisão de generalização do modelo, ajudando na escolha do melhor algoritmo para o problema em questão (REFAEILZADEH; TANG; LIU, 2009).

A divisão dos dados pode ser feita de formas diversas dando origem aos diferentes métodos de validação cruzada utilizados na literatura. A técnica base da validação cruzada é denominada *Hold-Out* e utiliza apenas um conjunto de teste independente. Os dados são divididos em duas partes não sobrepostas uma para treinamento e outra para testes, que não necessariamente possuem o mesmo tamanho. Os dados do teste são mantidos fora de todas as rodadas de treinamento efetuadas. Essa validação evita a sobreposição entre dados de treinamento e teste, tendo como desvantagem a não utilização de todos os dados disponíveis e possuindo resultados altamente dependentes da escolha da divisão dos dados de treinamento e teste (REFAEILZADEH; TANG; LIU, 2009).

Na técnica *k-fold* de validação cruzada, os dados são divididos em k grupos de tamanhos iguais, e a partir daí, são feitas interações para utilização de cada um dos k grupos no processo de validação do modelo, i.e., separa-se um dos grupos para a validação e treina-se o modelo com os outros grupos restantes, repetindo-se o processo com cada um dos grupos. Note-se que com esta técnica é possível não apenas estimar um único valor para a qualidade do modelo ou algoritmo testado, mas é possível obter uma distribuição de probabilidade acerca desta qualidade, o que pode ser usado para uma comparação mais profunda entre técnicas diversas (GUTTAG, 2013).

A técnica *Leave-One-Out* de validação cruzada é um caso de extremo da va-

lidação *k-fold*, onde k é igual ao número total de amostras disponíveis para o estudo. Em cada interação é utilizado para teste um único registro e o restante é posto no treinamento. Essa validação é bastante utilizada quando os dados são muito raros, onde existem apenas dezenas de amostras de dados disponíveis (GUTTAG, 2013).

2.3 Trabalhos Relacionados

Abaixo segue uma síntese dos trabalhos relacionados à este. Esta seleção apresenta trabalhos que fazem uso de técnicas de interpolação espacial para diferentes tipos de dados, tanto utilizando técnicas com base em geoestatística, técnicas não-geoestatísticas e técnicas de aprendizagem de máquina.

Em (LI; HEAP, 2014) é feita uma revisão dos estudos relacionados a interpolação espacial de dados ambientais, fornecendo uma visão geral e classificação de 25 métodos baseados em suas características. A classificação é apresentada na forma de uma árvore de decisão que permite selecionar os métodos mais adequados com base na natureza dos dados sob estudo, tipo da estimativa e características de cada método. O trabalho também indica uma lista de pacotes de software comumente utilizados para interpolação espacial.

O trabalho de Xavier, King e Scanlon (2016) fez uma comparação de métodos de interpolação para dados de precipitação e evapotranspiração provenientes de estações meteorológicas do território brasileiro no período de 1980 a 2013. Foram realizados experimentos com seis métodos de interpolação distintos: Média dentro da área de 0,25º × 0,25º (MÉDIA), Interpolação Natural (NATURAL), Spline de Superfície Fina (THINPLATE), IDW, Ponderação pela Distância Angular (ADW) e OK de ponto comum (OPK). O trabalho empregou a validação cruzada *leave-one-out* e computou diversas métricas de erro para avaliação do desempenho de cada modelo. Além disso, foi criado um ranking para classificação das técnicas, o qual apontou que o IDW e ADW foram os melhores métodos de interpolação para as duas variáveis sob estudo. Usando os modelos com melhores resultados do ranking, o autor construiu e disponibilizou uma rede 0.25º x 0.25º de dados diários e mensais para precipitação e evapotranspiração para o Brasil inteiro de 1980 a 2013.

No estudo de Li et al. (2011), foram comparados o desempenho de 23 métodos, utilizando dados de amostras de conteúdo de lama da margem sudoeste do litoral da Austrália. Esse artigo se destaca por ter sido pioneiro ao empregar técnicas de aprendizagem de máquina para a interpolação espacial. Os métodos utilizados foram divididos pelo autor com as seguintes categorias: métodos de interpolação espacial não-geoestatísticos, como o IDW; métodos geoestatísticos como a Krigagem e suas variações; métodos estatísticos baseado em regressão, como a Regressão Linear e

o Modelo Linear Generalizado; métodos de aprendizagem de máquina como a árvore de regressão e a RF. Além disso, o trabalho apresentou resultados para combinações destes métodos, utilizando um dos métodos para a tarefa de interpolação principal e outro para construção de um modelo dos erros do método principal. Os resultados dos experimentos mostraram que a RF, bem como combinações desta com outros métodos, forneceram as estimativas mais precisas.

O trabalho de Appelhans et al. (2015) utilizou 14 algoritmos de aprendizagem de máquina para a previsão de padrões espaciais de temperatura, com dados provenientes das encostas do sul do Monte Kilimanjaro na Tanzânia. Utilizando a validação cruzada foi constatado que, no geral, os modelos baseados em árvores de regressão obtiveram uma melhor performance do que os modelos de regressão linear e não linear.

Em Silva et al. (2019) o autor faz uma execução de múltiplas simulações e desenvolve um programa para acesso e visualização de uma base de dados georreferenciados. Para escolha do melhor método de interpolação a ser utilizado, foram feitos experimentos com o IDW, *Shepard Modified*, *Natural Neighbour*, *Nearest Neighbour*, *Radial Basis Function*, *Kernel smoothing*, Krigagem e *Trend Surface Analysis*. As métricas de avaliação utilizadas foram Erro Médio (EM), MAE, Erro Quadrático Médio(MSE), RMSE, Eficiência do Modelo(EFM), D de Willmott(D), Coeficiente de Determinação(R^2). Os dados utilizados foram provenientes de agrupamento de estações meteorológicas da base do Instituto de Tecnologia de Pernambuco (ITEP) e INMET, somando um total de 329 estações meteorológicas, do interior e contorno do estado de Pernambuco, no período de 1950 a 2012. Os experimentos apresentaram que a técnica com os melhores resultados na precisão de suas interpolações foi o *Trend Surface Analysis*.

O trabalho de Júnior (2018) faz uma avaliação de precisão e desempenho computacional de algoritmos para interpolação de dados de evapotranspiração. Utilizando técnicas convencionais como o IDW e o OK, e técnicas de aprendizagem de máquina, como o RF e a árvore de regressão, o trabalho avalia a qualidade da interpolação da evapotranspiração diária em todo o Nordeste brasileiro do mês de janeiro de 2017. Os experimentos do autor mostraram que, para os dados em estudo, os algoritmos clássicos como o IDW e OK obtiveram um resultado inferior no aspecto do desempenho computacional comparado aos algoritmos de aprendizagem de máquina, em especifico a Árvore Aleatória. Em termos de precisão, os algoritmos de aprendizagem, em destaque a RF, apresentaram os melhores resultados.

3 Metodologia

Esta seção descreve como foram realizados os experimentos para avaliação da interpolação dos dados de precipitação. São apresentadas também justificativas da utilização dos métodos de interpolação e seus parâmetros; como foram feitos os recortes temporal e espacial da base de dados utilizada; e a forma do tratamento dos dados e execução dos experimentos.

A execução desses experimentos tem o intuito de efetuar uma análise da precisão dos métodos na estimativa espacial do valor de precipitação pluviométrica, com o objetivo de conhecer qual técnica de interpolação terá melhor resultado, no contexto dos dados de precipitação diários correspondentes a estações meteorológicas no estado de Pernambuco, no período de 2013 até 2018.

Os métodos utilizados para essa avaliação foram o IDW, OK, RF e RFsp. O IDW foi escolhido por ser bastante utilizado nos experimentos que fazem interpolação utilizando dados ambientais (LI et al., 2011), incluindo dados de precipitação pluviométrica (XAVIER; KING; SCANLON, 2016; SILVA et al., 2019). Os experimentos extensivos de Xavier, King e Scanlon (2016) para o todo Brasil durante mais de 30 anos, demonstram que o IDW obtém os melhores resultados. Assim como o IDW, a escolha do OK foi feita por ser um algoritmo comumente utilizado nos trabalhos relacionados, e também possuir um resultado aproximado ao do IDW.

O trabalho (LI et al., 2011) foi o pioneiro na utilização de técnicas de aprendizagem de máquina para interpolação espacial, demonstrando a viabilidade na utilização destas técnicas para este propósito pelo fato da RF ter obtido bons resultados nos experimentos, tanto individualmente quanto em combinação com o OK e IDW. Dos métodos de aprendizagem de máquina, a RF foi escolhida dado estes resultados e a sua utilização em outros trabalhos similares (JúNIOR, 2018; HENGL et al., 2018).

Os dados utilizados nos experimentos são provenientes das estações meteorológicas convencionais do sistema HidroWeb[1], mantido pela Agência Nacional de Águas (ANA). O recorte espacial utilizado é referente às estações pertencentes exclusivamente ao estado de Pernambuco, totalizando 46 estações distribuídas pelo estado. No recorte temporal, foi utilizado o período compreendido de 1º de janeiro de 2013 à 31 de dezembro de 2018. A Figura 2 mostra a disposição espacial das estações meteorológicas empregadas.

As estações convencionais da ANA disponibilizam dados diários do acúmulo de precipitação, portanto, cada estação contida no experimento tem um registro para cada

[1] https://www.snirh.gov.br/hidroweb/publico/apresentacao.jsf

dia do ano. A base do ANA foi escolhida devido ao fato de ter apresentado uma maior cobertura de sensores, como demonstrado em (XAVIER; KING; SCANLON, 2016). O recorte temporal foi escolhido para utilização de uma base mais atual para esse tipo de experimento.

Figura 2 – Localização das estações meteorológicas

Os registros diários de precipitações referentes ao estado de Pernambuco no período de 2013 até 2018 somam um total de 99138 registros. Com esses registros foram feitos alguns tratamentos para remoção de registros faltantes ou inconsistentes. Inicialmente foram removidos 1924 registros nulos, que são os registros com valores que não são numéricos devido, provavelmente, à falhas na leitura. Em seguida, foram removidos 2 registros contendo a mesma localização e os mesmos valores de precipitação, deixando apenas um único registro.

Em relação a qualidade dos registros de precipitação, foram considerados apenas os registros de precipitação maiores ou iguais a 0 e menores que 450 mm, conforme o controle feito no trabalho de Xavier, King e Scanlon (2016). Esta análise, em nossa base, indicou que todos os 97212 registros encontram-se neste intervalo.

A Tabela 1 apresenta estatísticas do valor do acúmulo anual da precipitação de cada estação por ano. Pode-se perceber que algumas estações meteorológicas apresentam um acúmulo de precipitação anual bastante baixo, tal como a estação da Barragem de Glória do Goita que obteve apenas 32,3 mm em 2016, apenas 10% do primeiro quartil deste ano (321,6 mm). Efeito similar ocorreu em 2017 com a estação de Belém de São Francisco.

Os resultados do desvio padrão para todos os anos demonstra que existe uma grande variação na quantidade precipitada em Pernambuco (o coeficiente de variação é de mais de 0,5 em todos os casos). Este resultado é corroborado ao observar-se que a maior parte dos dados de precipitação acumulada (terceiro quartil) tem valor muito abaixo da máxima (sendo a máxima até aproximadamente 3 vezes o valor do terceiro quartil). Estes resultados indicam a alta variabilidade do dado o que pode impactar nos resultados dos métodos de interpolação.

Ano	Média	Desvio Padrão	Mínima	Máxima	Q1	Q2	Q3
2013	703,3	570,7	157,6	2131,1	316,0	445,1	864,9
2014	735,9	545,4	143,4	2349,6	369,8	581,6	833,5
2015	628,3	520,1	175,9	2171,1	271,2	442,7	825,9
2016	604,9	437,0	32,3	1785,5	321,6	421,6	838,6
2017	734,6	579,3	61,9	2405,3	282,4	578,9	988,1
2018	666,9	353,5	304,5	1748,8	437,5	555,2	745,2

Tabela 1 – Sumário da precipitação anual (em mm/ano) das estações meteorológicas de 2013 a 2018

Os parâmetros dos modelos utilizados neste trabalho foram baseados em experimentos observados nos trabalhos relacionados. Para o IDW foi definido o valor de potência igual a 2, devido aos bons resultados obtidos nos trabalhos de Xavier, King e Scanlon (2016) e Silva et al. (2019). Sobre os algoritmos de aprendizagem de máquina, RF e RFsp, foi definido apenas o valor da quantidade de árvores. No caso foram usados dois valores: 500 árvores referente aos experimentos executados em (JúNIOR, 2018); e 2000 árvores referente ao trabalho de (LI et al., 2011).

Para a OK utilizou-se o modelo esférico, levando em consideração sua utilização em (MARCUZZO; ANDRADE; MELO, 2011) e (XAVIER; KING; SCANLON, 2016). Para o ajuste do semivariograma foi escolhido o variograma mensal, que utiliza a média dos registros diários de cada estação para cada mês, ou seja, foi construído um único variograma para cada mês utilizando-o em todos os dias referentes ao respectivo mês. Essa escolha foi feita considerando o desempenho desta técnica no trabalho do Xavier, King e Scanlon (2016). A Figura 3 mostra o variograma mensal de Junho de 2018.

Para a execução dos experimentos com a RFsp, foi construído uma matriz de distâncias euclidianas para todas as 46 estações pertencentes ao estado de Pernambuco. Nesta matriz, cada linha é um vetor que indica a distância para um estação especifica. Essa matriz foi salva em um arquivo no formato Rdata, pertencente ao ambiente R, para que posteriormente seus valores fossem resgatados em cada interação do experimento. Desta forma é possível reduzir o tempo de execução do experimento, já que os valores de distância estão pré-computados.

Para comparar a precisão dos métodos de interpolação foi utilizada a validação cruzada *leave-one-out* (XAVIER; KING; SCANLON, 2016). Para cada dia e para cada método de interpolação, remove-se uma estação que servirá para teste e os registros das estações restantes são utilizados para estimar o modelo. Isso é repetido para cada estação presente no respectivo dia.

Como em (XAVIER; KING; SCANLON, 2016) e (JúNIOR, 2018) as seguintes

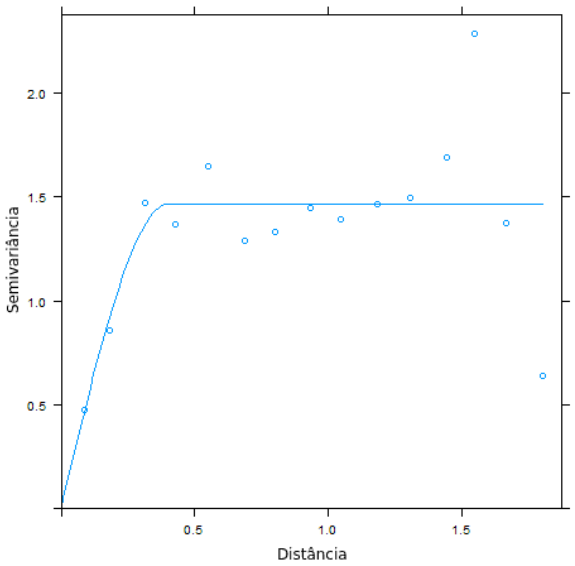


Figura 3 – Variograma Esférico Mensal - Junho de 2018

métricas de avaliação foram utilizadas: coeficiente de correlação (*R* descrito na Equação 3.1); viés (*BIAS* descrito na Equação 3.2); raiz do erro médio quadrático (*Root Mean Square Error - RMSE*, descrito na Equação 3.3) e erro médio absoluto (*Mean Absolute Error - MAE*, descrito na Equação 3.4). Nestas equações, X e Y são, respectivamente, o valor da precipitação observada e interpolada no dia em uma estação i, n é o número de estações com dados válidos no dia, $\overline{X}$ é o valor médio espacial da precipitação das estações em um dia e $\overline{Y}$ é o valor médio da precipitação interpolada.

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}\sqrt{(X_i - \overline{X})^2}\sqrt{(Y_i - \overline{Y})^2}} \tag{3.1}$$

$$BIAS = \overline{Y} - \overline{X} \tag{3.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}} \tag{3.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{3.4}$$

Essas métricas foram utilizadas para medir o desempenho dos modelos na validação cruzada *leave-one-out* e cada uma captura um aspecto distinto da precisão das técnicas. O coeficiente de correlação é uma medida de similaridade, onde o grau de dependência linear das variáveis é medido, variando entre -1 a 1. Já o *BIAS* indica se as estimativas interpoladas tendem a serem menores ou maiores que os dados observados. No caso o valor ideal de*BIAS* é 0. O *MAE* mede a precisão da interpolação,

quando os dados observados e estimados são semelhantes, quanto mais próximo de 0 indica que a interpolação foi precisa. Semelhante ao *MAE*, o *RMSE* também mede erro, porém, calcula o quadrado do desvio entre os valores observados e estimados sendo, portanto, mais sensível a erros maiores.

O script de execução do experimento foi construído na linguagem de programação R, utilizando o ambiente integrado de desenvolvimento (IDE) Rstudio[2]. A biblioteca utilizada para a implementação do IDW e OK foi a gstat[3](LI; HEAP, 2014), e para a RF e RFsp foi utilizado a biblioteca ranger[4] empregada em (HENGL et al., 2018). O computador utilizado para execução dos experimentos foi um Intel Core i5-6300U 2.40GHz com 8GB de memoria RAM.

[2] https://www.rstudio.com/
[3] https://www.rdocumentation.org/packages/gstat/versions/2.0-2
[4] https://www.rdocumentation.org/packages/ranger/versions/0.11.2/topics/ranger

4 Resultados e Discussões

Neste capítulo é apresentado a análise dos resultados obtidos pela execução dos experimentos. Os resultados gerais de cada método são apresentados na Tabela 2, enquanto as Figuras 4, 5, 6 e 7 mostram os valores das métricas obtidos para cada dia em cada ano. Por fim, a Figura 8 apresenta os mapas de interpolação para cada método analisado para o dia 01/06/2017.

Nos resultados globais, pode-se observar que os métodos apresentaram resultados bastante aproximados. Contudo, pode-se dizer que as métricas mostraram um desempenho ligeiramente superior para a técnica convencional IDW que apresentou melhores valores para as métricas R, RMSE e MAE. Seguindo, os melhores métodos em sequência foram RF, OK e, por último, RFsp.

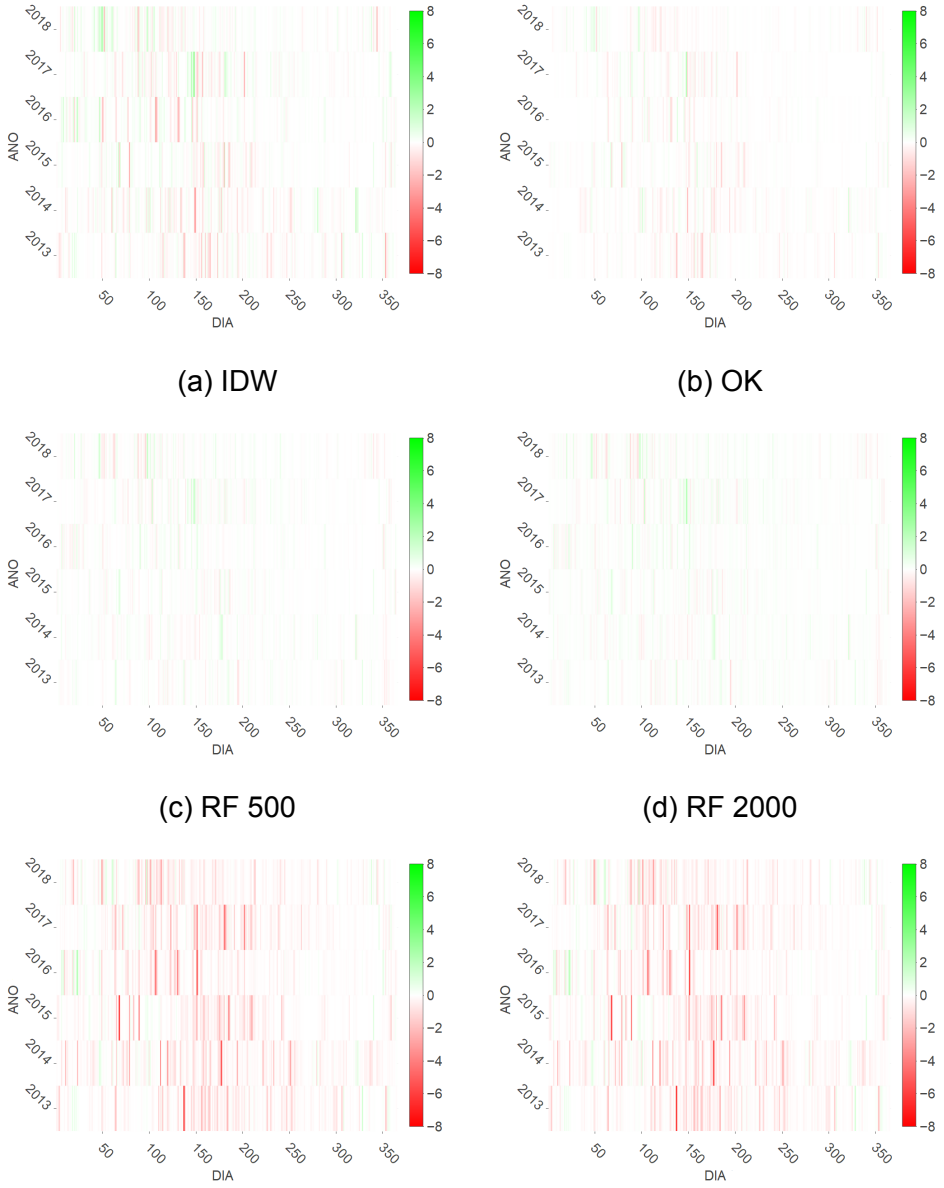
Outra observação que se pode fazer a partir destes resultados é que o incremento no parâmetro número de árvores (de 500 para 2000) nos métodos RF e RFsp não influenciaram os resultados. Podendo o parâmetro 500 ser usado neste caso, sem maiores perdas.

Algoritmo	R	BIAS	RMSE	MAE
IDW	0,6652	-0,0202	5,2630	1,8058
OK	0,6352	-0,0270	5,4733	1,9590
RF 500	0,6380	0,0129	5,4233	1,8778
RF 2000	0,6383	0,0133	5,4215	1,8771
RFsp 500	0,6388	-0,2259	5,4614	1,9904
RFsp 2000	0,6383	-0,2255	5,4636	1.9910

Tabela 2 – Resultado das métricas dos experimentos

Na métrica BIAS, o RFsp obteve o menor valor. Pode-se ver que os valores estimados pela RFsp para ambos parâmetros de quantidade de árvores tenderam a ser menores que os valores observados. A Figura 4 reforça esta conclusão da métrica geral ao ilustrar o BIAS das estimativas diárias para cada ano. Ainda sobre o BIAS, pode-se observar um contraste no resultado em relação à RF. Já que a RF obteve o melhor resultado de BIAS, i.e, valor de BIAS geral mais próximo de 0 quando comparado às outras técnicas.

Para a métrica R, no geral, para todos os métodos utilizados tivemos uma correlação positiva de proporção mediana. Apesar de todos os valores estarem bem próximos uns dos outros, o IDW mostrou ter uma maior correlação positiva comparado aos demais métodos, com valor geral de 0,6652. Em destaque na Figura 5, no período de



(a) IDW (b) OK

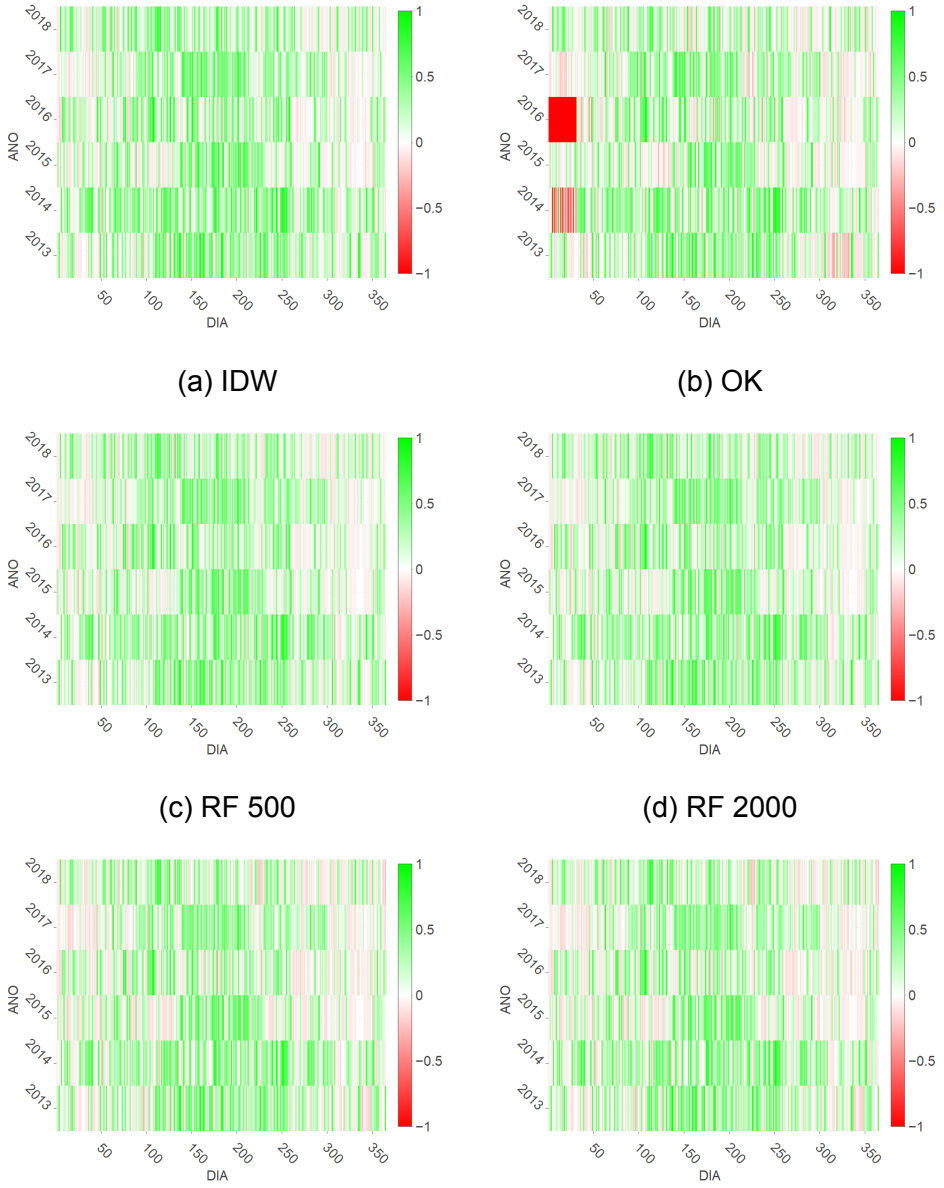
(c) RF 500 (d) RF 2000

(e) RFsp 500 (f) RFsp 2000

Figura 4 – Métrica BIAS de avaliação para os métodos.

2014 e 2015, as estimativas no início do ano para o OK obtiveram uma alta correlação negativa. Como consequência, isto teve um impacto na métrica geral, fazendo com que o método OK tivesse a menor correlação frente aos resultados de todos os outros métodos.

Em relação à métrica RMSE, os resultado obtidos pela maioria dos métodos mostraram-se bem próximos uns dos outros, como mostra a Tabela 2. Este resultado é corroborado pela Figura 6, que mostra que uma grande proximidade entre os métodos. Ainda assim, deve-se destacar que o IDW obteve uma precisão ligeiramente melhor em suas estimativas para essa métrica, tendo um valor de 5,2630 mm/dia.



(a) IDW (b) OK

(c) RF 500 (d) RF 2000

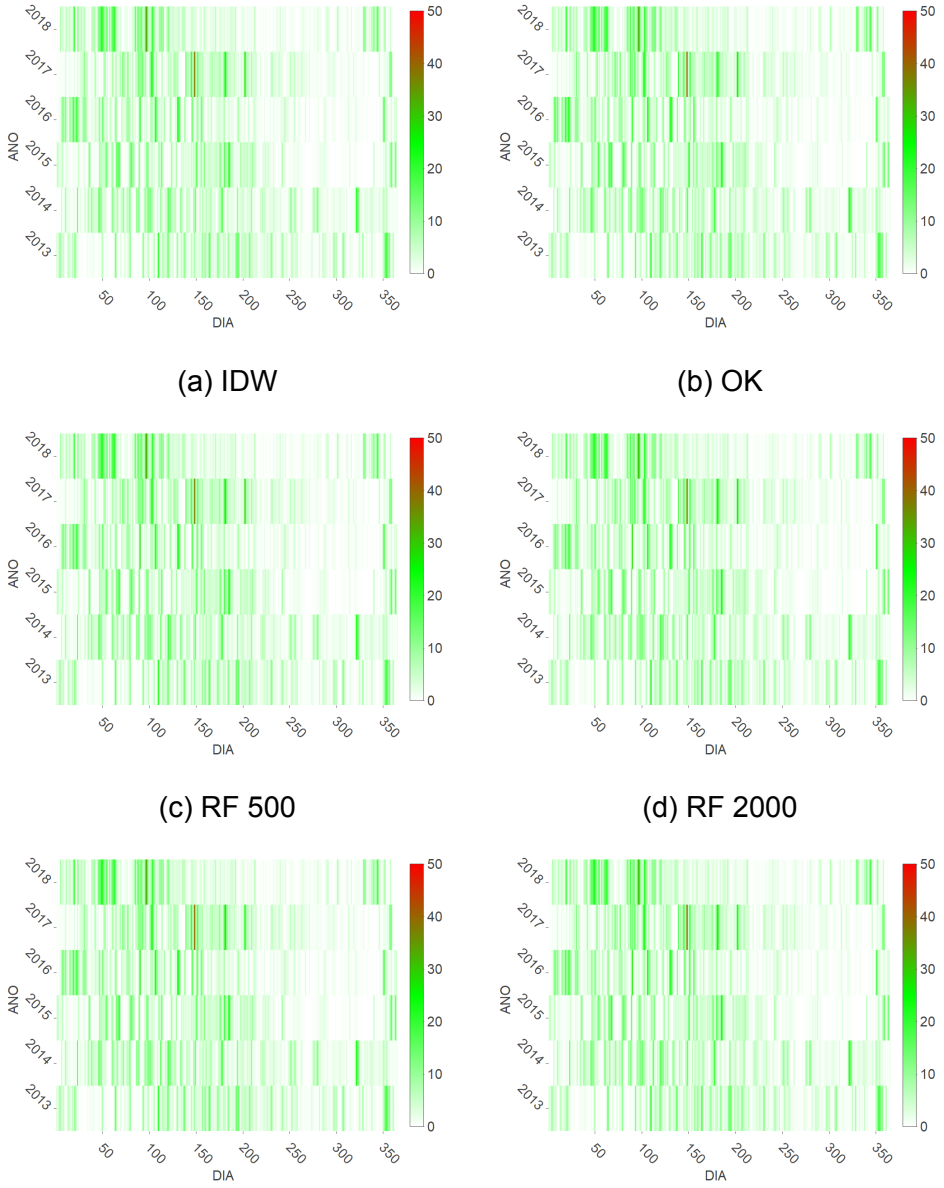
(e) RFsp 500 (f) RFsp 2000

Figura 5 – Métrica R de avaliação para os métodos.

O IDW também tem um resultado um pouco melhor para a métrica MAE (1,8058 mm/dia). Mas a análise dos dados diários, na Figura 7, indica que há pouquíssima diferença entre os métodos estudados.

Na Figura 8 podemos ver o mapa de interpolação do acumulo de precipitação no dia 01/05/2016 para todas as técnicas analisadas, para gerar o grid a ser interpolado no mapa utilizamos o método raster[1] com o parâmetro res = 1000. Os pontos no mapa representam as localidades das estações e as cores indicam a chuva em mm/dia conforme a escala de cores ao lado do gráfico.

[1] https://www.rdocumentation.org/packages/raster/versions/2.9-5/topics/raster



(a) IDW (b) OK

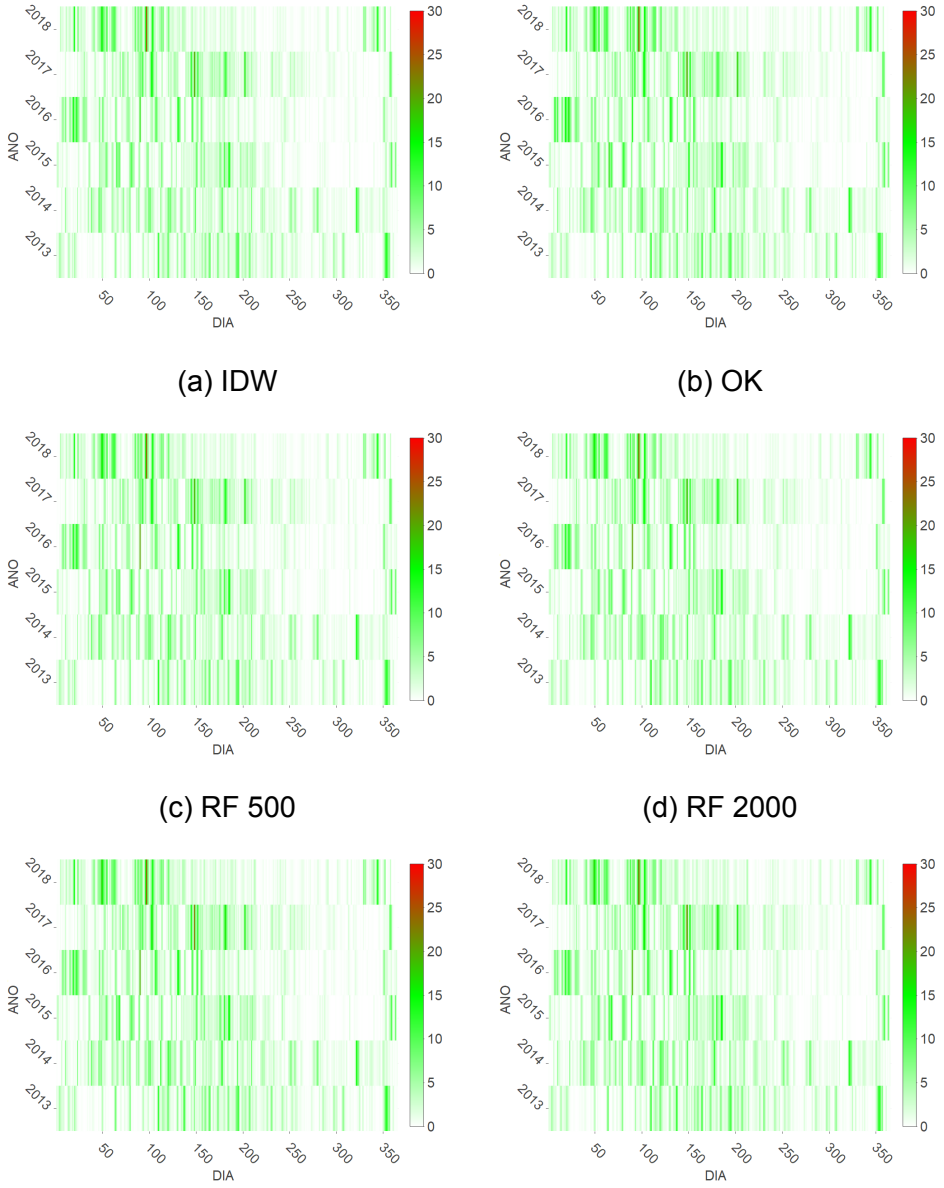
(c) RF 500 (d) RF 2000

(e) RFsp 500 (f) RFsp 2000

Figura 6 – Métrica RMSE de avaliação para os métodos.

Esse dia foi escolhido por ter uma boa distribuição visual da precipitação de chuva no estado. Para os mapas da RF e RFsp foram utilizados os modelos com quantidade de árvores igual à 500. Em aspectos visuais, pode-se observar que as técnicas IDW, OK e RFsp apresentam contornos mais suaves quando comparados à RF convencional. Isso ocorre porque as decisões tomadas nas árvores de decisão na RF fazem recortes lineares paralelos à latitude ou longitude causando os efeitos lineares característicos da RF.

Com esses resultados, podemos observar que assim como no trabalho do Xavier, King e Scanlon (2016), em nosso trabalho o IDW obteve os melhores resultados perante as outras técnicas. No âmbito geral a RF obteve o segundo melhor resultado,



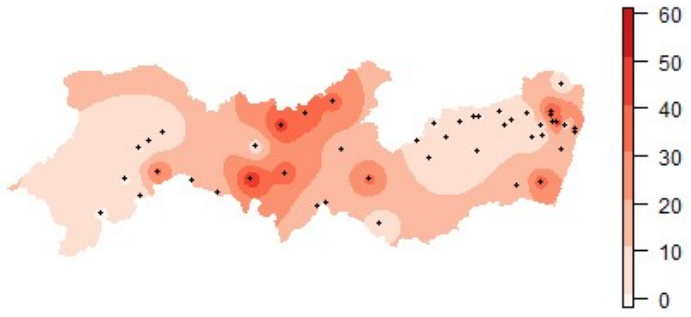
(a) IDW (b) OK

(c) RF 500 (d) RF 2000

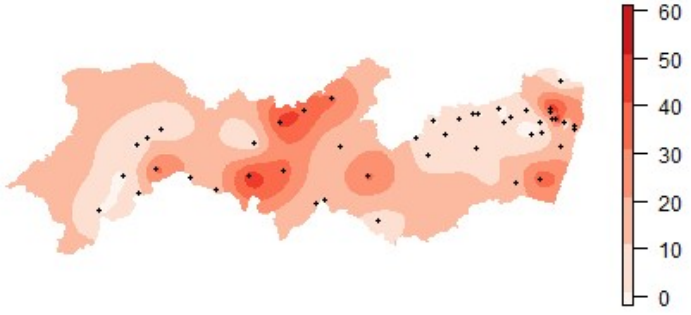
(e) RFsp 500 (f) RFsp 2000

Figura 7 – Métrica MAE de avaliação para os métodos.

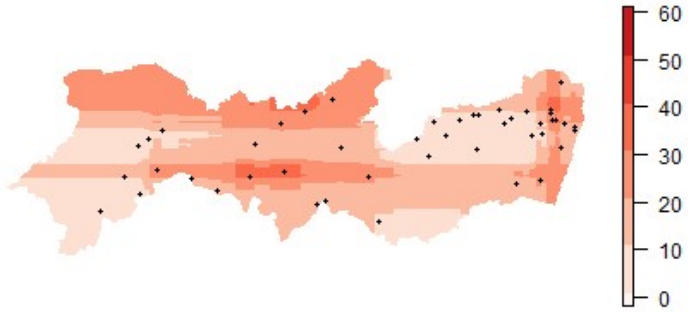
corroborando com o trabalho de Li et al. (2011). A configuração da RFsp não apresentou melhorias em termos de precisão comparada a RF, mas, assim como proposto em (HENGL et al., 2018), consegue aproximar-se dos resultados visuais de precipitação obtidos pelo IDW e OK e distanciando-se dos efeitos lineares da RF com coordenadas.



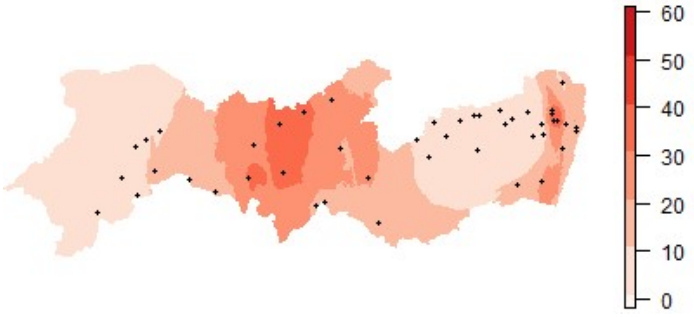
(a) IDW



(b) OK



(c) RF



(d) RFsp

Figura 8 – Mapa da precipitação de chuva interpolada para o dia 01/06/2017

5 Conclusão

Este trabalho investigou a capacidade de generalização dos métodos de interpolação convencionais (representados pelo IDW e o OK) para interpolação espacial de precipitação diária frente aos métodos de aprendizagem de máquina RF e sua modificação denominada RFsp. Esta avaliação de precisão dos métodos fez uso das métricas R, BIAS, RMSE, MAE, e empregou dados diários de precipitação pluviométrica no estado de Pernambuco, no período de 1º Janeiro de 2013 à 31 de Dezembro de 2018.

Os resultados das métricas de precisão indicaram que, de forma geral, os métodos se aproximaram bastante para os dados em estudo, com a técnica convencional IDW obtendo um desempenho ligeiramente melhor do que as demais, ao contrário da OK que no geral obteve a menor correlação positiva e a maior RMSE em relação às outras técnicas utilizadas. A RF ficou na segunda posição no resultado das suas métricas.

Para a RFsp, pode-se notar, que de fato, o uso das distâncias de *buffer* causa um efeito visual positivo na interpolação das predições, tornando as estimativas visualmente mais suaves que a RF e aproximando-as daquelas obtidas pelo IDW e OK.

5.1 Contribuições

Este trabalho, no melhor do conhecimento do autor, torna-se pioneiro pelo fato de obter resultados de um experimento utilizando métodos convencionais e de aprendizagem de máquina, especificamente por usar o recente método RFsp, que propõe melhorar as predições da RF para interpolação espacial usando informação sobre o referencial geográfico. Um detalhe importante é que, por ser um trabalho de conclusão de curso, o escopo é bastante reduzido, tanto no recorte espacial quando no temporal da base utilizada, então para solidificar esses resultados deve-se aumentar os intervalos de tempo e espaço dos dados.

5.2 Trabalhos futuros

Para trabalhos futuros espera-se conduzir novos experimentos utilizando dados provenientes de outras bases de estações meteorológicas, como o INMET e ITEP, incluindo também as estações automáticas, com um recorte espacial maior com todas as regiões do país em um recorte temporal maior. Em relação as interpolações , exe-

cutar os experimentos utilizando os acúmulos mensais e anuais. Pretende-se avaliar ainda o desempenho de outros algoritmos de aprendizagem de máquina tais como a Máquina de Vetores de Suporte (*Support Vector Machine* - SVM), utilizando também as distâncias de buffer como entradas para as predições. Também espera-se incluir o Modelo de Elevação Digital (*Digital Elevation Model* - DEM) para melhoria da RFsp tanto em generalização quanto em termos visuais. Pretende-se incluir também uma avaliação de desempenho dos algoritimos por tempo de processamento.

Referências

APPELHANS, T. et al. Evaluating machine learning approaches for the interpolation of monthly air temperature at mt. kilimanjaro, tanzania. *Spatial Statistics*, Elsevier, v. 14, p. 91–113, 2015. Citado na página 11.

BABAK, O.; DEUTSCH, C. V. Statistical approach to inverse distance interpolation. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 23, n. 5, p. 543–553, 2009. Citado na página 4.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 2.

CRESSIE, N. Spatial prediction and ordinary kriging. *Mathematical geology*, Springer, v. 20, n. 4, p. 405–421, 1988. Citado na página 2.

GUTTAG, J. V. *Introduction to computation and programming using Python*. [S.l.]: Mit Press, 2013. Citado 2 vezes nas páginas 9 e 10.

HENGL, T. et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, PeerJ Inc., v. 6, p. e5518, 2018. Citado 6 vezes nas páginas 2, 8, 9, 12, 16 e 21.

JúNIOR, J. C. d. S. Avaliação de técnicas para interpolação espacial de dados de evapotranspiração. 2018. Citado 5 vezes nas páginas 1, 7, 11, 12 e 14.

LI, J.; HEAP, A. D. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, Elsevier, v. 53, p. 173–189, 2014. Citado 3 vezes nas páginas 2, 10 e 16.

LI, J. et al. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, Elsevier, v. 26, n. 12, p. 1647–1659, 2011. Citado 4 vezes nas páginas 10, 12, 14 e 21.

LOH, W.-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 1, n. 1, p. 14–23, 2011. Citado na página 7.

LU, G. Y.; WONG, D. W. An adaptive inverse-distance weighting spatial interpolation technique. *Computers & geosciences*, Elsevier, v. 34, n. 9, p. 1044–1055, 2008. Citado 2 vezes nas páginas 2 e 4.

MARCUZZO, F. F. N.; ANDRADE, L. R.; MELO, D. C. d. R. Métodos de interpolação matemática no mapeamento de chuvas do estado do mato grosso. 2011. Citado 4 vezes nas páginas 4, 5, 6 e 14.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. *Encyclopedia of database systems*, Springer, p. 532–538, 2009. Citado na página 9.

RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, Elsevier, v. 71, p. 804–818, 2015. Citado na página 7.

SILVA, A. S. A. da et al. Comparison of interpolation methods for spatial distribution of monthly precipitation in the state of pernambuco, brazil. 2019. Citado 7 vezes nas páginas 1, 2, 4, 5, 11, 12 e 14.

VIOLA, M. R. et al. Métodos de interpolação espacial para o mapeamento da precipitação pluvial. *Revista Brasileira de Engenharia Agricola e Ambiental-Agriambi*, v. 14, n. 9, 2010. Citado 2 vezes nas páginas 1 e 4.

XAVIER, A. C.; KING, C. W.; SCANLON, B. R. Daily gridded meteorological variables in brazil (1980–2013). *International Journal of Climatology*, Wiley Online Library, v. 36, n. 6, p. 2644–2659, 2016. Citado 6 vezes nas páginas 1, 10, 12, 13, 14 e 20.