CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

# UM ESTUDO COMPARATIVO PARA PREDIÇÃO DE CONSUMO DE FERTILIZANTES EM UM CENÁRIO DE SMALL DATA

Adalberto Mineiro de Andrade

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Pedro Henrique González Silva
Eduardo Soares Ogasawara

Rio de Janeiro,

Janeiro de 2021

# UM ESTUDO COMPARATIVO PARA PREDIÇÃO DE CONSUMO DE FERTILIZANTES EM UM CENÁRIO DE SMALL DATA

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Adalberto Mineiro de Andrade

Aprovada por:

---

Presidente, Prof. D.Sc. Pedro Henrique González Silva (CEFET/RJ) (orientador)

---

Professor D.Sc. Eduardo Soares Ogasawara (CEFET/RJ) (coorientador)

---

Professor D.Sc. Eduardo Bezerra da Silva (CEFET/RJ)

---

Professor D.Sc. Cristina Gomes De Souza (CEFET/RJ)

---

Professor D.Sc. Igor Machado Coelho (UFF)

Rio de Janeiro,

Janeiro de 2021

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A553 Andrade, Adalberto Mineiro de
Um estudo comparativo para predição de consumo de fertilizantes em um cenário de small data / Adalberto Mineiro de Andrade – 2020.
32f : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2020.
Bibliografia : f. 29-32
Orientador: Pedro Henrique González Silva
Coorientador: Eduardo Soares Ogasawara

1. Consumo de fertilizantes. 2. Análise de séries temporais. 3. Aprendizado de máquina – Previsao – Pré-processamento de dados.I. Silva, Pedro Henrique González (Orient.). II. Ogasawara, Eduardo Soares (Coorient.). III. Título.

CDD 631.8

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

# DEDICATÓRIA

Este trabalho é dedicado à minha mãe, Laurice M. de Andrade (in memorian), uma pessoa humilde, sábia, de valor inestimável, que sempre me apoiou em todos momentos de minha vida. Sua lembrança foi a fonte de minha força para persistir e superar todos os obstáculos dessa minha jornada.

# AGRADECIMENTOS

Ao longo destes três anos tive a oportunidade de conviver no ambiente acadêmico do Programa de Pós-Graduação em Ciência da Computação (PPCIC) do CEFET-RJ, onde recebi um aprendizado de excelência com professores altamente capacitados, encontrei um quadro de funcionários competentes, fiz colegas e amigos de curso. Por serem muitas pessoas, talvez seja difícil agradecer a todos que de alguma forma contribuíram direta ou indiretamente para a minha formação, mas deixo aqui minha tentativa de ser justo ao máximo.

Gostaria de expressar minha profunda gratidão ao Prof. D.Sc. Pedro Henrique González, meu orientador, e ao Prof. D.Sc. Eduardo Ogasawara, meu co-orientador. Ambos com muita dedicação, valiosa orientação, paciência, conselhos e com rica experiência acadêmica tornaram essa dissertação possível.

Agradeço a coordenação do PPCIC, onde os ensinamentos únicos e desafios constantes em cada disciplina, propiciaram um conhecimento amplo e completo ao curso de Mestrado em Ciência da Computação, além da alta qualidade no aprendizado importante e pioneiro que é a ciência de dados.

Agradeço ao dedicado e motivador corpo docente do CEFET-RJ.

Ao corpo discente acho que preciso agradecer aos meus contemporâneos de curso, pois de alguma forma aprendia sempre, seja com suas pesquisas, com as apresentações, com as ideias, com os trabalhos, em parcerias de estudo, em colaboração em artigos, pelas trocas de experiências acadêmicas, pelos desafios vividos juntos em sala de aula, enfim agradeço a todos, mas em especial agradeço ao Flávio Carvalho pela extrema boa vontade em ajudar, pela paciência, pelo alto nível da colaboração no artigo aceito no BreSci, pela alta capacidade de comprometimento nas tarefas e pela amizade. Também preciso citar o Antônio Castro pelo trabalho de disciplina desenvolvido em dupla, pela parceria em estudos e pela amizade. E por falta de espaço não cito outros colegas de curso.

Por último, e não menos importante minha família, agradeço a minha esposa Conceição por ter compreendido e acompanhado minha evolução em meu processo de formação acadêmica, meu filho Hugo e minha filha Natália.

# RESUMO

UM ESTUDO COMPARATIVO PARA PREDIÇÃO DE CONSUMO DE FERTILIZANTES EM UM CENÁRIO DE SMALL DATA

Adalberto Mineiro de Andrade

Orientadores:
Pedro Henrique González Silva
Eduardo Soares Ogasawara

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Os fertilizantes têm recebido crescente atenção do agronegócio, indústria, empresários, governos e entidades de pesquisa em todo o mundo. Como insumo crítico para a cadeia produtiva de alimentos e insumos orgânicos para outros setores, é importante prever o consumo de fertilizantes, para que o aumento de sua produção possa ser feito adequadamente planejado, sem comprometer o meio ambiente. Esta previsão apoia a tomada de decisões e o planejamento, particularmente para atividades agrícolas, fortemente dependentes do uso de fertilizantes. Tendo em vista os elementos citados, esta pesquisa tem como foco comparar abordagens analíticas de dados para melhorar as previsões do consumo de fertilizantes sob diferentes horizontes de passos à frente. Para tanto, exploramos maneiras de otimizar a construção de modelo considerando diferentes abordagens (ou seja, combinações de pares entre pré-processamento de dados e métodos de aprendizado de máquina). Avaliamos essas abordagens em um conjunto reduzido de observações, correspondentes aos quatro principais fertilizantes usados nos dez principais países que os consomem. Os resultados obtidos mostraram que o uso das ferramentas analíticas propostas pode ser uma maneira promissora de obtermos previsões para planejar demandas futuras.

Palavras-chave:
consumo de fertilizantes; análise de dados; previsão de séries temporais; aprendizado de máquina; pré-processamento de dados

Rio de Janeiro,

Janeiro de 2021

# ABSTRACT

UM ESTUDO COMPARATIVO PARA PREDIÇÃO DE CONSUMO DE FERTILIZANTES EM UM CENÁRIO DE SMALL DATA

Adalberto Mineiro de Andrade

Advisors:
Pedro Henrique González Silva
Eduardo Soares Ogasawara

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Fertilizer has received increasing attention from the agribusiness industry, entrepreneurs, governments, and research entities around the world. As critical input for the production chain of food and organic inputs for other sectors, it is important to predict fertilizer consumption, so the increase in its production could be adequately planned without compromising the environment. It supports decision-making and planning, particularly to agricultural activities, which are strongly dependent on the use of fertilizers. Due that, this research focuses on comparing data analytical approaches to improve predictions of fertilizer consumption under different horizons of steps forward. To do this, We explored ways to optimize the model construction considering different approaches (i.e., pair combinations between data preprocessing and machine learning methods). We evaluate these approaches in a reduced observations set, corresponding to the four main fertilizers of the top ten countries that demand them. The obtained results showed that using the proposed analytic tools can be a promising way to get predictions to plan for future demands.

Key-words:
fertilizer consumption; data analytics; time series prediction; machine learning; data preprocessing

Rio de Janeiro,

Janeiro de 2021

## Sumário

## Lista de Figuras

## Lista de Tabelas

## Lista de Abreviações

| | | |
|---|---|---|
| AN | Adaptive Normalization | 10 |
| AR | *Autoregressive Model* | 9 |
| ARIMA | *AutoRegressive Integrated Moving Average* | 9 |
| EELM | *Ensemble Extreme Learning Machines* | 11, 13 |
| ELM | *Extreme Learning Machine* | 11, 12 |
| EMLP | *Ensemble Multilayer Perceptron* | 11, 13 |
| IFA | *International Fertilizer Association* | 15 |
| MA | *Moving Average* | 9 |
| MLP | *Multilayer Perceptron* | 11 |
| MSE | *Mean Squared Error* | 14 |
| ONU | Organização Das Nações Unidas | 1 |
| RF | *Random Regression Forest* | 13 |
| SLFNS | *SINGLE-HIDDEN-LAYER FEEDFORWARD NETWORKS* | 11, 12 |
| SMAPE | *Symmetric Mean Absolute Percentage Error* | 14 |
| SVM | *Support Vector Machines* | 13 |

# Capítulo I Introdução

De acordo com a Organização das Nações Unidas (ONU) [UN, 2019], a população mundial atingirá 9,8 bilhões de pessoas até 2050. Portanto, a produção de alimentos terá que aumentar quase 50% acima do nível atual para acompanhar essa demanda (FAO, 2019), o que é um desafio considerável. Fazê-lo de uma forma que não comprometa a integridade ambiental é um desafio ainda mais significativo. O consenso é que atender a essa demanda na produção agrícola é essencial para a estabilidade, equidade política e social global [Tilman et al., 2002].

Paralelamente, a agricultura terá que adaptar-se, buscar e expandir a adoção de práticas agrícolas sustentáveis para mitigar os efeitos das mudanças climáticas. Em números a agricultura contribui com cerca de 14% de todas as emissões de gases de efeito estufa que levam ao aquecimento global [FAO, 2019] e ocupa cerca de 11% (1,5 bilhões de hectares) da superfície terrestre global (13,4 bilhões de hectares) [FAO, 2019].

Neste cenário, embora iniciativas na utilização de robôs, sensores de temperatura, sensores de umidade, imagens aéreas, radiofrequência, *drones* e GPS contribuam para melhorar as atividades agrícolas [Kirkpatrick, 2019], produzir mais com menos e preservar o meio ambiente é um desafio vital para o futuro do planeta. Melhorias substanciais na eficiência do uso e conservação dos recursos naturais devem ser alcançadas globalmente para atender ao crescimento e mudança na demanda de alimentos, ao mesmo tempo em que se evita a degradação ambiental.

Atualmente, os fertilizantes são inseridos como insumos essenciais neste desafio global. O uso adequado de fertilizantes desempenha um papel vital nesta tarefa [FAO, 2019]. O uso de fertilizantes é a chave não apenas para alcançar a segurança alimentar no mundo [Stewart and Roberts, 2012], como também é fundamental para a economia mundial. E quanto maior for a necessidade de produção de alimentos, maior será a quantidade de fertilizantes necessária. A partir disto, as consequências ambientais da escala da produção de fertilizantes precisam também ser consideradas.

Consideremos, por exemplo, o impacto da produção de fertilizantes fosfatados [Attallah et al., 2019]. Este processo consiste em várias etapas, e cada etapa impacta o meio ambiente no seu caminho. Além disso, causam emissões atmosféricas, erosão e assoreamento. A primeira etapa, a supressão da vegetação, pode levar à perda da biodiversidade e de resíduos vegetais (como folhas, galhos). A segunda etapa, remoção do solo orgânico e limpeza de resíduos, pode gerar terrenos estéreis e mudança de paisagem. A terceira, perfuração e desmontagem do minério (com explosivos

ou escavação mecânica), pode causar interferência na dinâmica das águas superficiais e na poluição subterrânea da água. A quarta etapa, após a tomada do material para a planta, a desativação do empreendimento - fechamento/recuperação, pode gerar interferência na dinâmica da água e nos impactos socioeconômicos. Todos esses impactos são graves, e o descuido pode até mesmo levar a um desastre natural.

Além do fertilizante fosfato, todos os outros fertilizantes geram seu impacto sobre o meio ambiente. Dados os desafios ambientais que surgem na produção de fertilizantes, é essencial prever com precisão a demanda quantitativa de consumo dos principais fertilizantes para que o aumento da produção possa ser feito de forma que o impacto ao meio ambiente aconteça de forma controlada.

Devido à importância de minimizar os impactos ambientais, na presente pesquisa temos como objetivo:

- geral: a previsão do consumo dos quatro principais fertilizantes utilizados (Nitrogênio (N), Fósforo ($P_2O_5$), Potássio ($K_2O$) e NPK (NPK)) nos dez principais países (Brasil, Canadá, China, Estados Unidos, França, Índia, Indonésia, Paquistão, Rússia e Turquia) que os demandam.

- específico: a realização de uma avaliação completa de abordagens analíticas de dados (um par de combinações de pré-processamento de dados e métodos de aprendizado de máquinas). Avaliamos estas abordagens em 40 séries temporais diferentes de fertilizantes, que correspondem ao conjunto de dados descrito no objetivo geral.

Considerando as explicações fornecidas, o cenário da presente pesquisa poderia ser representado conforme Figura I.1.

Esta pesquisa contribui de três maneiras principais. Primeiro, estabelece um *workflow* que realiza a análise de dados de previsão do consumo de fertilizantes sob diferentes horizontes de previsão passos à frente e replicação durante a fase de validação-treinamento. Em segundo lugar, fornece uma análise mais profunda usando uma ampla gama de abordagens de análise de dados. Em terceiro, apresenta as tendências do consumo mundial de fertilizantes para os próximos anos.

Além deste capítulo introdutório, esta pesquisa está organizada em mais cinco capítulos. No capítulo II descrevemos os principais trabalhos relacionados ao estudo desta dissertação. O capítulo III descreve os conhecimentos prévios do consumo de fertilizantes e abordagens de análise de dados para predição de séries temporais. O capítulo IV detalha a metodologia desenvolvida nesta pesquisa. O capítulo V apresenta a avaliação experimental e discussão conduzida nesta pesquisa. Por fim, o capítulo VI conclui e apresenta a continuidade da pesquisa em possível trabalho futuro.

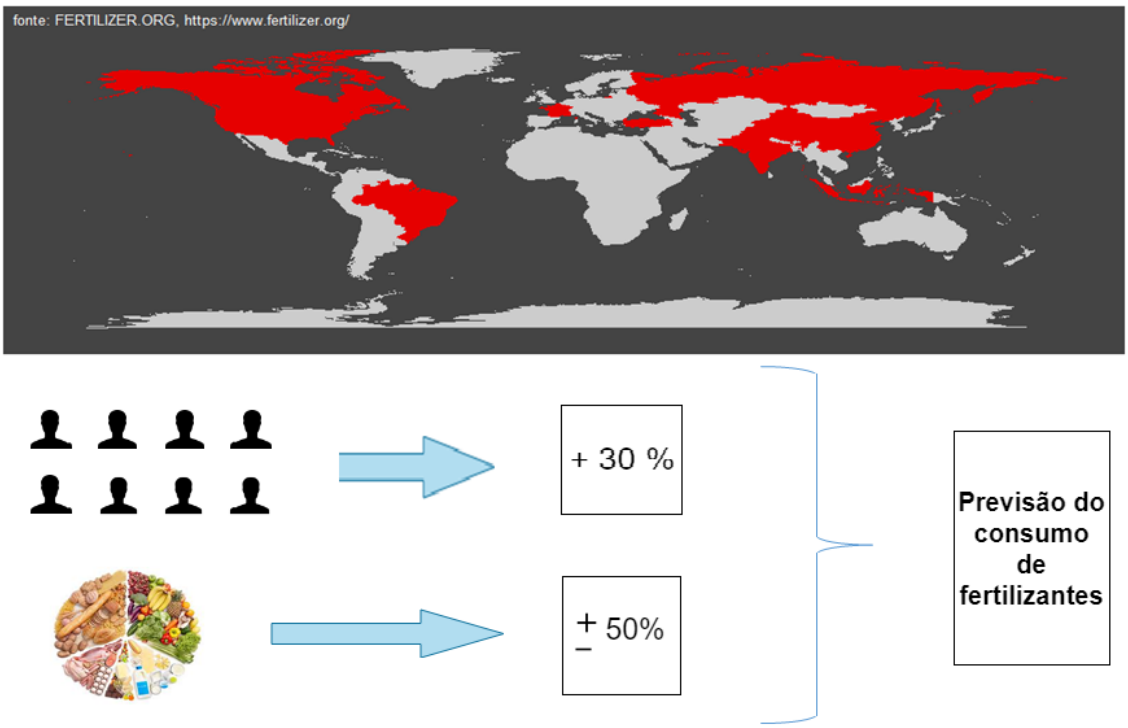



Figura I.1: *C*enário da presente pesquisa

# Capítulo II Revisão da Literatura

Ao longo dos anos, tem sido coletados diversos dados oriundos de diversas fontes. A extração do conhecimento vem ganhando cada vez mais força em todas as áreas do conhecimento. Especificamente na área de fertilizantes, valiosas informações sobre consumo dos principais fertilizantes podem ser fornecidas a sociedade e aos tomadores de decisões. Considerando a necessidade de produção de quase 50% a mais de alimentos em relação ao nível atual para suprir o aumento populacional até 2050 [UN, 2019], a tarefa de previsão do consumo de fertilizantes tem ganho importância.

Além dos citados fatos, fertilizantes são considerados insumos agronômicos relevantes para melhoria da produção agrícola e sua aplicação está associada a um consequente crescimento econômico gerado pela expansão agrícola dos países McArthur and McCord [2017]. Em função destes fatores, muita pesquisa tem sido feita na comunidade científica sobre fertilizantes. Nesta pesquisa, estamos particularmente interessados em avaliar artigos que abordem a previsão do consumo de fertilizantes. Para encontrar esses artigos, um mapeamento sistemático da base indexada SCOPUS a partir de *string* ("predict" OR "forecast") AND "fertilizer" AND ("consumption" OR "demand") foi realizado. Somente artigos de periódicos ou de conferências foram considerados. Além disso, somente os artigos escritos em inglês foram estudados. A consulta retornou 370 artigos do período de 1972 a 2019. A consulta foi executada em 2 de outubro de 2019. Todos os resumos foram analisados e 79 artigos foram selecionados para leitura posterior com base em sua relevância.

A maioria dos artigos encontrados na literatura está relacionada à agricultura e analisa fertilizantes no contexto de otimização de seu uso nas lavouras. Nesse caso, são feitos estudos com base nas propriedades das espécies cultivadas e no solo em que são plantadas. Esses estudos são utilizados como base para a identificação e quantificação da demanda de fertilizantes. Outros estudos conduzem análises socioeconômicas a partir das tendências de consumo de fertilizantes. As análises deste tipo não foram selecionadas para discussão nesta seção, pois elas estão associadas a outras áreas de domínio preocupadas com os efeitos dos valores previstos, e não sobre como prever.

No final, foram selecionados para discussão um conjunto de artigos que abordavam a previsão do consumo de fertilizantes. Buscamos trabalhar com propostas como esta pesquisa, mais relacionada à previsão do consumo de fertilizantes em territórios maiores, como países ou até o mundo, em detrimento de lavouras de espécies específicas em ambientes controlados. Resumimos, imediatamente a seguir, a principal ideia de cada um dos nove principais estudos selecionados que estão

relacionados a presente pesquisa, bem como visualizamos na Tabela II.1 características importantes de cada um destes estudos.

Styhr Petersen [1977] apresenta um método de previsão que prevê o consumo do fertilizante nitrogênio na Dinamarca com dados da agricultura daquele país.

Deadman and Ghatak [1979] aborda uma gama mais extensa de fertilizantes em suas análises e apresentam uma pesquisa que realiza uma previsão de longo prazo da produção e consumo de fertilizantes no mundo.

No estudo de Gilland [1993], é analisado as tendências na relação entre a produção de cereais e o consumo de fertilizantes nitrogenados até o ano de 2030. Conclui-se que a duplicação do consumo de nitrogênio químico durante esse período é necessária para manter o consumo de nitrogênio na produção de cereais no mundo nos níveis atuais.

A pesquisa de Howarth et al. [2002] consiste em utilizar dados derivadas das agricultura dos Estados Unidos de 1961 até 1997 para realizar a previsão do fertilizante Nitrogênio até 2030.

Em sua pesquisa, Dobermann and Cassman [2005] considera as diferenças de regiões, países e culturas para realizar uma tendência em escala global para atender o consumo do fertilizante nitrogênio.

Zhang and Zhang [2007] apresentam uma pesquisa cujo objetivo é prever o consumo de fertilizantes em todo o mundo, a fim de fornecer informações para a tomada de decisão sobre a produção de fertilizantes e avaliação do impacto ambiental. Nesse estudo verificou-se que o consumo de fertilizantes era dependente da população humana e o aumento do consumo de fertilizantes foi principalmente resultante da expansão da população humana.

Tenkorang and Lowenberg-Deboer [2009] realiza uma pesquisa que trabalha na previsão de longo prazo dos principais fertilizantes em nove regiões globais.

Ogasawara et al. [2013] apresentam um método de previsão para vinte anos à frente do consumo do Brasil dos fertilizantes NPK, enxofre, rocha fosfática, potássio e nitrogênio baseado no modelo ARIMA e no modelo de função logística. Para esta tarefa de previsão, essa pesquisa trabalha com três variáveis: crescimento do PIB, crescimento populacional e consumo dos fertilizantes. Foram utilizados nesta pesquisa três cenários de crescimento da economia: um pessimista, um conservador e um otimista, no qual cada um possui uma taxa de crescimento própria.

O estudo de Pires et al. [2015] investiga o relacionamento entre a evolução da produção de cereal e o uso do fertilizante nitrogênio no Brasil relacionado com o uso eficiente do Nitrogênio e emissão de gás estufa. Ao final deste estudo é realizada a previsão a longo prazo do fertilizante nitrogênio no Brasil.

Finalmente, é possível observar que a maioria dos artigos estudados utiliza modelos causais para prever os dados de consumo de fertilizantes. Os modelos causais associam uma ou mais variáveis

de interesse com a geração de uma curva característica de seu relacionamento. Modelos causais, tais como, modelos de regressão linear/polinomial, às vezes são preferíveis quando projeções de variáveis relacionadas estão disponíveis.

Neste levantamento realizado, observa-se que o número de publicações abordando a análise de séries temporais para previsão do consumo de fertilizantes ainda é escasso. Existe uma lacuna potencial de pesquisa para estudo, como estudos que utilizam uma gama mais ampla de pré-processamento de dados, métodos de aprendizado de máquina e um número maior de países e tipos de fertilizantes.

Tabela II.1: Trabalhos relacionados sobre predição do consumo de fertilizantes

| Artigo | Região | Fertilizante | Domínio | Metódo |
|---|---|---|---|---|
| Styhr Petersen [1977] | Dinamarca | N | Agricultura | Regressão |
| Deadman and Ghatak [1979] | Mundo | N, P, K | Agricultura | Regressão |
| Gilland [1993] | Mundo | N | Agricultura | Regressão |
| Howarth et al. [2002] | EUA | N | Ambiental | Regressão |
| Dobermann and Cassman [2005] | Mundo | N | Agricultura | Regressão |
| Zhang and Zhang [2007] | Mundo | N, P, K | Ambiental | Regressão |
| Tenkorang and Lowenberg-Deboer [2009] | Mundo | N, P, K | Agricultura | Regressão |
| Ogasawara et al. [2013] | Brasil | NPK, S, N, P, K | Agricultura | ARIMA |
| Pires et al. [2015] | Brasil | N | Agricultura | Regressão |

# Capítulo III Fundamentação Teórica

Neste capítulo, vários conceitos necessários para entender e discutir este trabalho são apresentados. Para facilitar o entendimento desses conceitos, este capítulo está dividido em seis seções. A seção III.1 apresenta as ideias fundamentais relacionadas a fertilizantes e quais fertilizantes são considerados nesta pesquisa. Na seção III.2, os conceitos de séries temporais e como eles podem ser utilizados para representar um evento distribuído no tempo é discutido. Na sequência a seção III.3 apresenta as principais características do modelo arima. Depois disso, a seção III.4 apresenta várias técnicas de pré-processamento de dados. Na seção III.5 seis técnicas de aprendizado de máquina para séries temporais são apresentadas. Por fim, na seção III.6, uma métrica é apresentada para avaliar o desempenho dos métodos utilizados durante a avaliação experimental.

## III.1 Fertilizantes

O termo fertilizante vem da palavra latina *fertilis*, que significa frutificação. Um fertilizante é um nutriente para planta, servindo quase como um suplemento vitamínico essencial para o solo. Vários materiais podem servir como fontes de nutrientes para as plantas. Dentro disto, existe uma classificação que os distingue: orgânicos, mineral (sintético ou inorgânico) e biofertilizantes (FAO, 2019). Esta pesquisa estuda estritamente fertilizantes minerais, os quais são os mais comuns e utilizados na agricultura.

Por sua vez, os fertilizantes minerais são classificados como fertilizantes diretos (também conhecidos como macronutrientes primários), que contêm um dos três principais nutrientes Nitrogênio (N), Fósforo ($P_2O_5$) ou Potássio ($K_2O$) em sua composição, ou fertilizantes compostos/complexos, que contêm mais de um dos macronutrientes primários. O fertilizante complexo mais amplamente utilizado é o NPK, cujo conteúdo é escrito na sequência N, $P_2O_5$ e $K_2O$ [FAO, 2019].

A fertilização do solo desempenha um papel de destaque na atividade agrícola, uma vez que é o principal responsável pelos ganhos de produtividade das lavouras [IFA, 2019]. Inclusive, no contexto da contribuição dos fertilizantes na produção de alimentos, estima-se que os fertilizantes são responsáveis por 40 a 60 por cento da produção global de alimentos [Roberts et al., 2009]. Ainda no contexto de fertilizantes, é comumente utilizado o termo "consumo aparente"como uma medida de consumo, e que corresponde ao valor da produção doméstica mais o valor da quantidade

importada menos o valor da quantidade exportada [de Planta Ciência e Tecnologia, 2018].

### III.2 Séries temporais

Uma série temporal é qualquer sequência de observações de um fenômeno ao longo do tempo. Portanto, podemos dizer que uma série temporal $t$ é uma sequência $< t_1, t_2, t_3, \cdots, t_n >$, onde $t_1$ é a primeira observação e $t_p$ é a observação mais recente. O comprimento da série temporal é representado por $|t| = p$. Geralmente, a evolução dos valores em uma série temporal não é uniforme, razão pela qual a maioria dos autores adota a tendência, a sazonalidade e aleatoriedade como componentes das séries temporais [Box et al., 2015a].

De acordo com esses componentes, a literatura relata a existência de vários modelos de previsão de séries temporais, a maioria dos quais assume que as séries temporais são estacionárias [Gujarati, 2002a]. Nas séries temporais estacionárias, as propriedades estatísticas média, variância e covariância apresentam valores constantes ao longo do tempo [Shumway and Stoffer, 2017]. No entanto, na prática, pode-se observar que essas propriedades não são constantes em muitas aplicações reais, como séries temporais envolvendo fenômenos socioeconômicos [Tsay, 2010], onde frequentemente encontramos séries temporais não-estacionárias.

Uma vez que a série temporal é representada como um conjunto de observações, os eventos de previsão do futuro podem ser descritos principalmente em duas etapas. A primeira é pré-processar os dados de entrada. A segunda é usar ferramentas de aprendizado de máquina para prever eventos futuros com base no conjunto de observações fornecidas como entrada.

### III.3 Modelo ARIMA

Nos dias de hoje, a previsão é uma ferramenta valiosa na definição das estratégias. Uma previsão com valores futuros precisos garante uma tomada de decisão mais ponderada e justificada. Previsão de cura da doença de um paciente pela evolução de seu histórico clínico, previsão do crescimento populacional de um país, previsão do crescimento do PIB de um país, previsão do movimento de ações na bolsa de valores, previsão do consumo de energia, previsão do consumo de água, previsão da oscilação da vendas de um produto, previsão do estoque de um produto, inúmeras são as demandas. Motivos pelos quais fazem a previsão uma tarefa de crescente importância em vários campos do conhecimento e tem despertado o interesse de pesquisadores.

Em aplicações do mundo real, as diversas propriedades e complexidades dos dados impedem um processo de previsão trivial que seja possível produzir resultados confiáveis. Modelos matemáticos que desenvolvem a capacidade de treinar e representar a diversidade dos dados são necessários em cenários do mundo real.

*Autoregressive Model* (ar) foi o primeiro modelo a ser formulado na década de 1930. O modelo AR faz com que a série temporal seja regredida em seus próprios dados passados (AR($p$)). Anos depois surge o *Moving Average* (ma). O modelo MA indica que o erro de previsão é uma combinação linear dos erros respectivos anteriores (MA($q$)).

Box et al. [2015b], na década de 1970, com seu livro "*Times Series Analysis: Forecasting and Control*", criam o Modelo *AutoRegressive Integrated Moving Average* (arima) (p, d, q). O Modelo arima é derivado de uma composição do Modelo Autoregressivo (ar) e do Modelo de Médias Móveis (ma) (respectivamente representados por p e q) com um processo de diferenciação adicional (representado por d), desenvolvidos para tratar séries temporais não-estacionárias [Gujarati, 2002b]. O modelo formal arima (p, d, q) é definido na Equação III.1.

$$\phi_q(B)(1-B)^d x_t = \theta_q(B)a_t. \tag{III.1}$$

Na presente pesquisa, utilizamos na configuração o auto arima. A vantagem desta opção é de que o modelo arima realiza as tarefas de ajuste do modelo, da previsão e do cálculo do erro de previsão, todos de forma automaticamente. O arima é um modelo consolidado e tradicionalmente utilizado nas tarefas de previsão em séries temporais, motivo pelo qual ele é utilizado como *baseline* e seus resultados de previsão são comparados com outros métodos de previsão na presente pesquisa.

### III.4 Pré-processamento de Dados

O pré-processamento de dados é uma etapa importante durante a análise de dados. Um pré-processamento de dados adequadamente aplicado em séries temporais não-estacionárias pode levar a significativas melhoras nas previsões [Salles et al., 2019]. Nesta pesquisa, exploramos três abordagens gerais: (i) diferenciação; (ii) janela deslizante com min-max; (iii) normalização adaptativa. Entretanto, antes da explicação destas transformações, o conceito geral de janela deslizante é explicado, pois ele é usado para introduzir métodos de aprendizado de máquina.

Uma subsequência é uma amostra de séries temporais. A subsequência $i$-th de uma série temporal de tamanho $p$ para uma série temporal $t$ é representada como $seq_{p,i}(t)$ e corresponde à sequência de valores ordenados $< (t_i, t_{i+1}, \ldots, t_{i+p-1}) >$, onde $|seq_{p,i}(t)| = p$ e $1 \leq i \leq |t| - p$. Dado $A = sw_p(t)$, $\forall a_i \in A$, $a_i = seq_{p,i}(t)$. Vale ressaltar que as janelas deslizantes organizam as colunas da matriz $A$, de modo que a coluna $j$-th corresponda a uma referência de atraso para a série temporal original $t$ para os valores anteriores $p - j$ .

A transformação de diferenciação (diff) pode ser usada para eliminar tendências, usando o operador de retrocesso $B$ [Salles et al., 2019]. O diff pode ser definido por sua sequência, conforme descrito na Equação III.2, a qual define a estrutura da tendência para ser eliminada. Para $d = 1$, diff

elimina tendência linear, para $d = 2$ diff elimina tendência quadrática e assim por diante. Além do diff original, algumas variantes podem ser encontradas na literatura, como diferenciação fracionária (fdiff) e diferenciação sazonal (sdiff) [Salles et al., 2019]. Essa série temporal transformada é então fornecida como entrada para transformação da janela deslizante para treinamento adicional de aprendizado de máquina

$$\nabla^d = (1 - B)^d, B^k t_i = t_{i-k} \tag{III.2}$$

A janela deslizante com min-max (swmm) é o processo de transformar uma série temporal em uma janela deslizante de tamanho $p$ e, posteriormente, aplicar uma transformação para cada linha. Formalmente, dado $A = sw_p(t)$, $\forall a_i \in A$, $minmax(a_i)$ é descrito pela Equação III.3 [Han et al., 2011].

$$minmax(a_i) = \frac{a_{ij} - \min(a_i)}{\max(a_i) - \min(a_i)} \tag{III.3}$$

Normalização adaptativa (Adaptive Normalization (AN)) é uma técnica de normalização cujo processo pode ser dividido em três etapas. O primeiro passo é transformar uma série temporal não-estacionária em uma série estacionária, o que cria uma sequência de janelas deslizantes. O segundo passo consiste na remoção de *outliers*. Finalmente, na terceira etapa, a normalização min-max para toda a janela deslizante [Ogasawara et al., 2010].

Considerando uma janela deslizante $A$ de tamanho $p$ para uma série temporal $t$, tal que ($A = sw_p(t)$). O primeiro passo transforma a matriz $A$ em adaptativo normalizado $\hat{A}$. Para cada linha $a_i$ em $A$, uma média móvel é aplicada sobre $(a_{i_1}, \cdots, a_{i_p})$ conforme descrito na Equação III.4. A média móvel $ma$ traz inércia para a série temporal $t$. Quando a Equação III.4 é aplicada para todas as linhas $a_i \in A$, produz normalização adaptativa $\hat{A}$ [Ogasawara et al., 2010]. Após esta etapa, espera-se que $\hat{A}$ tenha média 1 e variância 2 $\sigma^2$.

$$\hat{a}_{i_j} = a_{i_j}/ma(a_{i_1}, \cdots, a_{i_p}), \forall j \in [1, p+1] \tag{III.4}$$

Na segunda etapa, os *outliers* são removidos. Todas as linhas nas quais um valor está fora de $[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$ para a distribuição de valores em $\hat{A}$ são removidos. Esse critério é o mesmo aplicado pela a análise de *box-plot* [Han et al., 2011]. Finalmente, na terceira etapa, uma transformação min-max é executada para toda a matriz $\hat{A}$, conforme descrito na Equação III.5.

$$minmax(\hat{A}) = \frac{\hat{a}_{ij} - \min(\hat{A})}{\max(\hat{A}) - \min(\hat{A})} \tag{III.5}$$

## III.5 Aprendizado de máquina para séries temporais

Observa-se que as séries temporais de fertilizantes compartilham algumas semelhanças (como demanda global por consumo de alimentos, minas disponíveis, pequena quantidade de dados disponíveis para ajustes de parâmetros) usando os mesmos princípios de aprendizagem de transferência homogênea [Weiss et al., 2016]. Tal característica no contexto de aprendizado de máquina implica no uso dos resultados de vários modelos para aproveitá-los no processo de aprendizagem. Na presente pesquisa, os resultados de seis diferentes modelos de aprendizado de máquina para a previsão do consumo de fertilizantes são investigados em 40 séries temporais.

Esta seção está organizada em cinco subseções que descrevem os seis modelos de aprendizado de máquina que são comumente utilizados em previsões de séries temporais e que são utilizados na presente pesquisa. A subseção III.5.1 descreve o modelo de aprendizado de máquina *Multilayer Perceptron*. A subseção III.5.2 faz uma descrição do modelo de aprendizado de máquina *Extreme Learning Machines*. A subseção III.5.3 apresenta o modelo *Random Regression Forest*. O modelo *Support Vector Machines* é apresentado na subseção III.5.4. Por último, a subseção III.5.5 apresenta o modelo comitê *Ensemble Extreme Learning Machines* (eelm) e o modelo comitê *Ensemble Multilayer Perceptron* (emlp) implementados na presente pesquisa

### III.5.1 *Multilayer Perceptron*

*Multilayer Perceptron* (mlp) é um tipo de rede neural *Feedforward*. Uma rede mlp tradicional possui pelo menos três tipo de camada: uma de entrada, uma oculta e uma de saída. Sendo uma rede neural artificial, cada nó corresponde a um neurônio, que por sua vez terá uma função de ativação. Uma rede mlp utiliza o algoritmo *backpropagation* para treinar seu modelo. *Backpropagation* é um método iterativo e recursivo para calcular as atualizações de pesos, onde o valor retornado na saída é comparado com o valor desejado. Esta é uma medida de erro. Em seguida, calcula o erro associado a cada neurônio da camada anterior. Este processo é repetido até que a camada de entrada seja atingida. Como o erro é propagado para trás (no sentido saída para os atributos de entrada), através da rede para ajustar os pesos e vieses, essa abordagem é conhecida como *backpropagation* [Lewis, 2017]. Este processo de treinamento de modelo também é conhecido como aprendizado supervisionado. A Figura III.1 ilustra o modelo MLP.

### III.5.2 *Extreme Learning Machine*

*Extreme Learning Machine* (elm) é uma rede *Feedforward* cuja arquitetura apresenta apenas uma única camada oculta, mais conhecida como *SINGLE-HIDDEN-LAYER FEEDFORWARD NETWORKS* (SLFNS), que utiliza uma função aditiva conhecida como *radial basis function (RBF)*

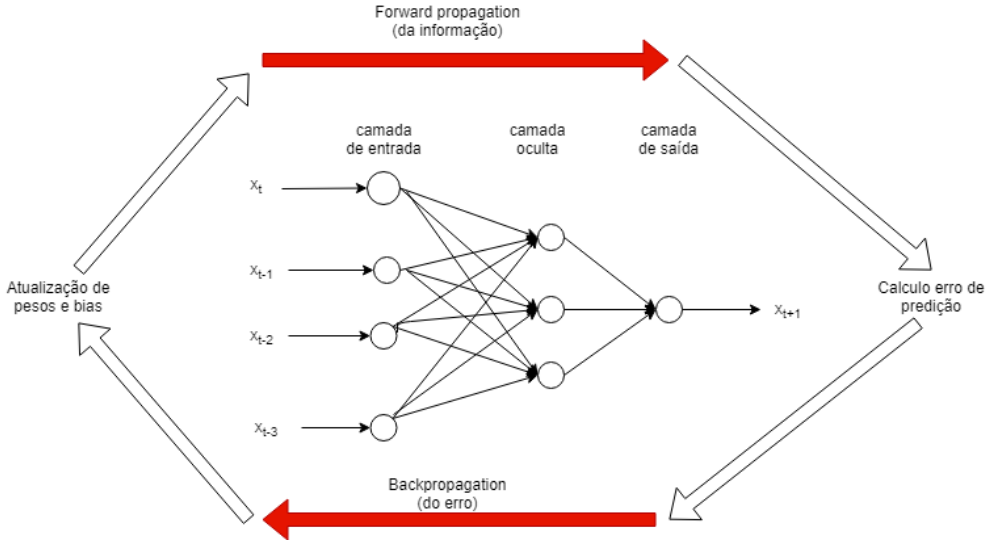



Figura III.1: Ciclo de aprendizado supervisionado utilizando modelo MLP. Adaptado de Lewis [2017]

nos nós ocultos [Huang et al., 2006]. Na rede elm os pesos sinápticos entre as camadas de entrada e oculta são escolhidos aleatoriamente, não necessitando de treinamento, os pesos são mantidos com seus valores fixos até o final do algoritmo. Apenas os pesos dos neurônios entre a camada oculta e saída são analiticamente calculados. Tais características, conferem ao elm um treinamento extremamente rápido, boa generalização, aproximação universal e capacidade de classificação [Tang et al., 2015].

Na teoria do elm, as SLFNS possuem L nós ocultos que podem ser representados pela Equação III.6, onde $a_i \in R^d$ , $b_i$ , $\beta_i \in R$. Por sua vez, a função de ativação do enésimo nó oculto é denotada por $G_i$, $a_i$ é o vetor de peso de entrada conectando a camada de entrada a enésima camada oculta, $b_i$ é o peso de polarização da enésima camada oculta e $\beta_i$ é o peso de saída. Enquanto $X$ é o vetor de entrada. Para nós aditivos com função de ativação $g$, $G_i$ é definido conforme Equação III.7.

$$f_L(X) = \sum_{i=1}^{L} G_i(x, a_i, b_i).\beta_i \tag{III.6}$$

$$G_i(x, a_i, b_i) = g(a_i.x + b_i) \tag{III.7}$$

E para nós de *radial basis function (RBF)* com função de ativação $g$,$G_i$ é definido como na Equação III.8.

$$G_i(x, a_i, b_i) = g(b_i, ||x - a_i||) \tag{III.8}$$

### III.5.3 *Random Regression Forest*

O *Random Regression Forest* (rf) é um algoritmo de aprendizado supervisionado que usa o método de aprendizagem de comitê para regressão. A ideia por trás de rf consiste em construir várias árvores de decisão no momento do treinamento e gerar a previsão média das árvores individuais [Palmer et al., 2007]. A ideia do rf é representada pela figura III.2.

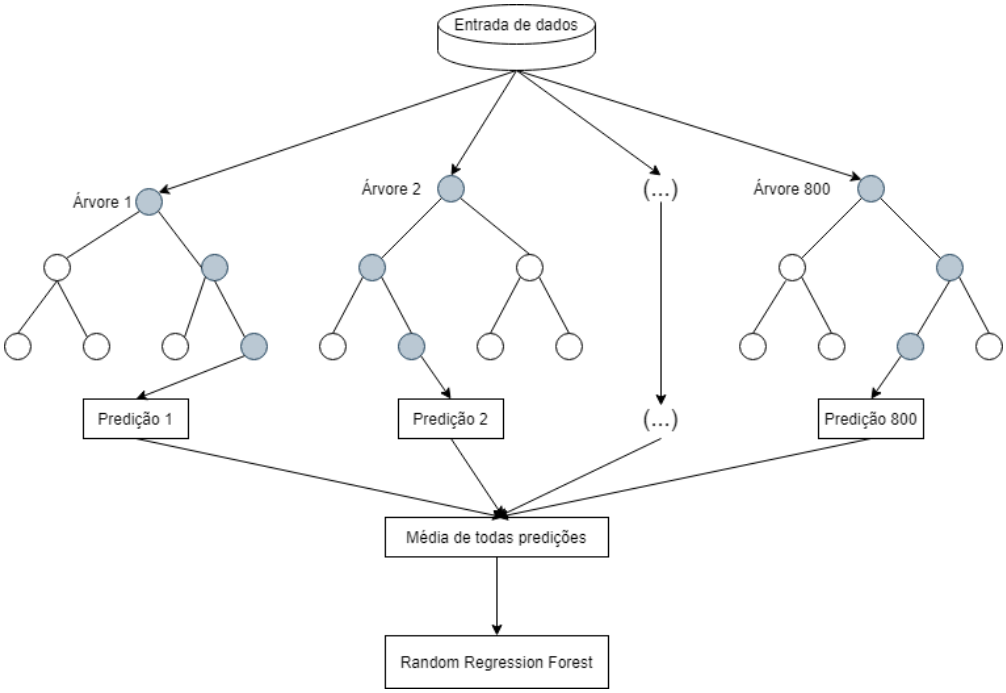



Figura III.2: Arquitetura do modelo *Random Regression Forest*, adaptado de Palmer et al. [2007]

### III.5.4 *Support Vector Machines*

Embora geralmente seja usado para classificação, o *Support Vector Machines* (svm) pode ser usado como um método de regressão. No caso de regressão, o svm tenta ajustar o erro dentro de um certo limite, em vez de minimizar a taxa de erro [Sapankevych and Sankar, 2009].

### III.5.5 Comitê

Na presente pesquisa o modelo comitê é implementada através do modelo emlp e do modelo eelm. Modelos comitê são modelos compostos por conjuntos de modelos de aprendizado de máquina. A ideia dos modelos comitê é retornar um consenso entre os vários modelos membros. Existem duas maneiras principais de criar modelos comitê. A primeira maneira de compor os comitês é treinar o mesmo tipo de rede em diferentes conjuntos de dados ou subconjuntos. Uma segunda maneira de configurar o comitê é usar diferentes tipos de técnicas de previsão e treiná-las no mesmo conjunto de dados. Neste trabalho, foi decidido usar a primeira maneira com o método de aprendizagem elm e mlp [Zhang and Berardi, 2001]. A ideia do comitê é representada pela figura III.3.

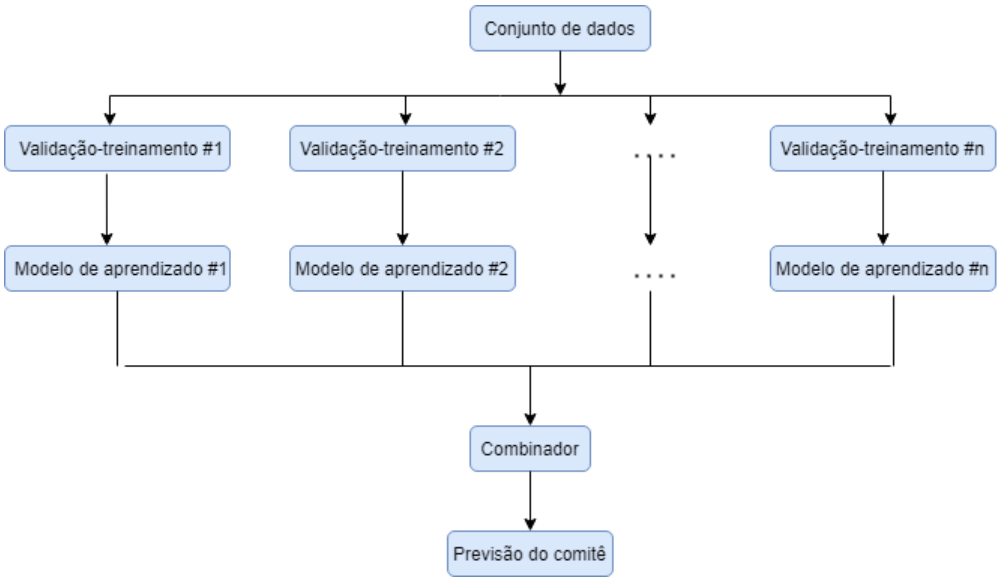



Figura III.3: Arquitetura do modelo comitê, adaptado de Zhang and Berardi [2001]

### III.6 Desempenho de previsão

A avaliação do desempenho de um modelo preditivo pode ser feita de várias maneiras. A mais comum faz uso do *Mean Squared Error* (MSE). O MSE é definido como a média das diferenças ao quadrado entre o valor previsto e o valor real. Formalmente, é definida pela Equação III.9 onde $\hat{t}_i$ é o valor previsto pelo modelo treinado e $t_i$ é o valor real.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2 \tag{III.9}$$

O *Symmetric Mean Absolute Percentage Error* (SMAPE) é uma medida de acurácia com base em erros de porcentagem (ou relativos). Geralmente é definido como apresentado na Equação III.10 [Crone et al., 2011], onde a diferença absoluta entre o valor real $t_i$ e o valor predito $\hat{t}_i$ são divididos pela metade da soma dos valores absolutos do valor real $t_i$ e do valor previsto $\hat{t}_i$.

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2|\hat{t}_i - t_i|}{\left(|t_i| + |\hat{t}_i|\right)}. \tag{III.10}$$

# Capítulo IV Metodologia

A metodologia aplicada neste trabalho é definida de acordo com o *workflow* representado na Figura IV.1. Ele pode ser resumido nas seguintes etapas de análises de dados: (i) seleção e limpeza de dados; (ii) partição de séries temporais em validação-treinamento e conjuntos de teste; (iii) otimização da construção do modelo; (iv) avaliação de modelo. Essas etapas são descritas em detalhes nas seguintes seções.

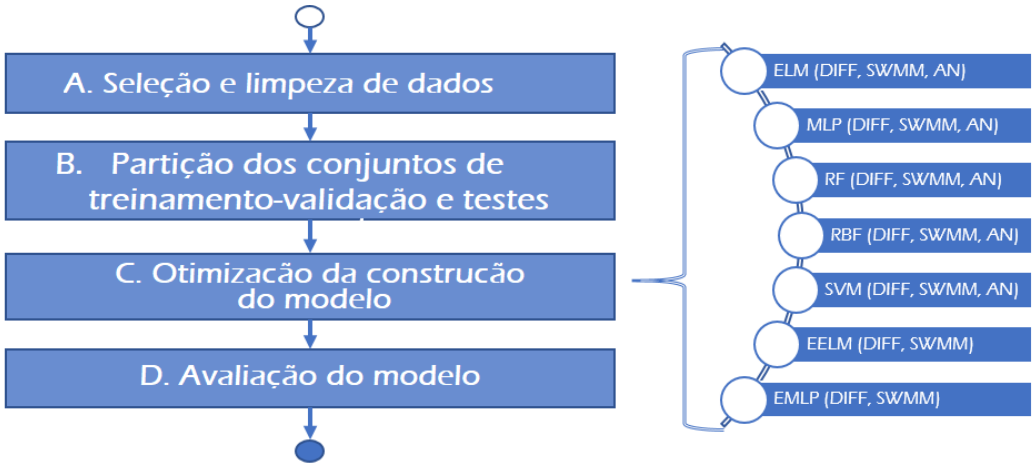



Figura IV.1: *Workflow* de análise de dados aplicado na metodologia

## IV.1 Seleção e limpeza de dados

O banco de dados utilizado na presente pesquisa é composto por dados públicos da *International Fertilizer Association* (IFA) disponível em `http://www.fertilizer.org`. Escolhemos o consumo para o banco de dados e *Grand Total* para nitrogênio (N), fosfato ($P_2O_5$), potássio ($K_2O$) e NPK (nitrogênio-fosfato-potássio) para fertilizantes, para os dez maiores países consumidores de fertilizantes do mundo: Brasil, Canadá, China, Estados Unidos, França, Índia, Indonésia, Paquistão, Rússia e Turquia, durante o período de 1961 a 2016.

O consumo de fertilizantes para cada país foi organizado como séries temporais anuais. Na formação destes conjuntos de dados a maioria das séries temporais foi composta de 56 observações. Nas séries temporais para o consumo de fertilizantes do Paquistão tiveram alguns valores faltantes, enquanto que nas séries temporais para o consumo de fertilizantes da Rússia apenas existiam dados disponíveis a partir de 1990. Por esse motivo, valores faltantes no início da série temporal foram

descartados, enquanto os valores faltantes no meio da série temporal foram interpolados.

### IV.2 Partição de conjuntos de validação-treinamento e teste

Considere uma série temporal $t_i$ com $n$ observações. Considere também o objetivo de prever $k$ passos à frente de observações. Devido a ordenação dos dados, a série temporal é particionada de $t_1$ a $t_{n-k}$ para validação-treinamento e de $t_{n-k+1}$ a $t_n$ para teste.

Devido à necessidade de otimização dos hiperparâmetros, o conjunto de validação-treinamento é particionado em dois subconjuntos. Um para o treinamento real e o outro para validação dos modelos desenvolvidos durante a otimização dos hiperparâmetros. Essa divisão é replicada $r$ vezes. Assim, o treinamento real do modelo usa observações de $t_1$ a $t_{n-2k-p}$ e o de validação usa observações de $t_{n-2k+1-p}$ a $t_{n-k-p}$, para cada replicação definida por $p$, tal que $p \in \{0, \cdots, r-1\}$. A razão para o processo de replicação é permitir que a otimização dos hiperparâmetros possa obter modelos estáveis, ou seja, modelos consistentemente mais precisos do que outros sob o conjunto de validação $r$.

### IV.3 Otimização da construção do modelo

A otimização da construção do modelo visa identificar a melhor abordagem (i.e., um par de pré-processamento de dados e métodos de aprendizado de máquina) para a previsão do consumo de fertilizantes. O processo consiste em usar *grid search* para refinar os hiperparâmetros [Thornton et al., 2013].

A tabela IV.1 descreve as dimensões do espaço dos hiperparâmetros que consideramos como opções. A opção de modelo de nome pré-processamento de dados avaliado neste trabalho teve como valores candidatos a diferenciação (diff), janela deslizante com normalização min-max (swmm), normalização adaptativa (an). Os métodos de aprendizado de máquina foram *Extreme Learning Machine* (elm), *Multilayer perceptron* (mlp), *Random Regression Forest* (rf), *Support Vector Machine* (svm), *Ensemble Extreme Learning Machine* (eelm), *Ensemble Multilayer Perceptron* (emlp).

Tabela IV.1: (C) Otimização da construção do modelo

| **Opções de modelos** | **Valores dos candidatos** |
|---|---|
| Pré-processamento de dados | {diff, swmm, an} |
| Aprendizado de máquina | {elm, mlp, rf, svm, eelm, emlp} |
| Estrutura interna | [3..10] |
| Número de entradas | [3..10] |

A opção do modelo estrutura interna possui valores que variam de acordo com o tipo de aprendizado de máquina: (i) elm e mlp (o número de nós ocultos); (ii) rf (o número de árvores de decisão); (iii) svm (o tipo de núcleos utilizados: linear, polinomial, base radial e sigmóide); (iv)

eelm (o número de eelm interno para o conjunto) e (v) emlp (o número de emlp interno para o conjunto). Por último define-se o número de entradas, que varia de 3 à 10.

Das possíveis opções de modelo descritas na Tabela IV.1, cada série temporal é treinada por 1152 modelos diferentes. Este valor é multiplicado pelo número de replicações ($r$) durante a otimização da construção do modelo.

Esta pesquisa analisa duas maneiras diferentes de otimizar construções de modelos. A primeira, chamada `opt`, é método proposto nesta pesquisa, que aplica uma técnica de otimização de busca dos melhores modelos de previsão em cada série temporal, escolhendo o modelo que minimiza o MSE ou SMAPE de cada abordagem. A segunda identifica a abordagem que é mais consistente para prever todas as séries temporais. Cada modelo ainda é ajustado de acordo com seus hiperparâmetros (número de entradas e estrutura interna). No entanto, uma única abordagem é adotada para todas as séries temporais. A abordagem que fornece o limite superior de erro mais baixo do intervalo de confiança para SMAPE é escolhida.

### IV.4 Avaliação do modelo

Uma vez que o modelo foi obtido a partir da otimização dos hiperparâmetros, ele é usado para prever observações para o conjunto de testes. Para isso, é feito um treinamento adicional usando todo o conjunto de validação-treinamento para prever observações $k$ passos à frente. A previsão é comparada com o conjunto de testes. As medidas de erro são coletadas usando tanto MSE quanto SMAPE. Além disso, a previsão de cada modelo é comparada com a previsão usando o modelo ARIMA ajustado com o algoritmo Hyndman-Khandakar [Hyndman and Khandakar, 2008]. Este processo avalia a qualidade do modelo de previsão [Salles et al., 2015].

# Capítulo V Resultados e discussão

Este capítulo apresenta os resultados obtidos e discute seu impacto. Para facilitar o entendimento deste capítulo, ele foi dividido em seis seções. A primeira seção apresenta a configuração experimental. A segunda apresenta a análise da otimização da construção do modelo. Em seguida, as Seções V.3, V.4 e V.5 apresentam discussões sobre o desempenho das abordagens propostas. A Seção V.6 apresenta as previsões para o NPK de oito anos passos à frente, considerando 2008 como linha de base. Em seguida, é apresentada tendências para os anos de 2017 a 2024, considerando 2016 como linha de base.

## V.1 Configuração Experimental

A avaliação experimental implementada na presente pesquisa consistiu em prever $k$ passos à frente de observações para previsão do consumo dos quatro principais fertilizantes consumidos globalmente (nitrogênio (N), fosfato ($P_2O_5$), potássio ($K_2O$) e NPK (nitrogênio-fosfato-potássio)) nos dez maiores países consumidores destes fertilizantes no período de 1961 até 2016 (Brasil, Canadá, China, Estados Unidos, França, Índia, Indonésia, Paquistão, Rússia e Turquia). Isso corresponde a 40 séries temporais anuais. Durante a avaliação experimental, variamos $k$ de 1 a 8. Para fins de comparação, o período de treinamento para todas as avaliações foi definido de 1961 a 2008, e os testes foram definidos de 2009 a 2016. Quando $k = 1$, o objetivo era prever o consumo de fertilizantes em 2009, enquanto que quando $k = 8$, a meta era prever de 2009 a 2016 de uma só vez. Além disso, a replicação ($r$) variou de 1 a 5.

Considerando o cenário de *small data* das séries temporais de fertilizantes anteriormente descrita, após execução de experimentos, a tarefa de previsão não se beneficiava de múltiplas camadas ocultas, motivo pelo qual é utilizado uma única camada oculta para cada método de aprendizado de máquina usado na presente pesquisa. Entretanto, existem outros parâmetros relevantes nos métodos de aprendizado de máquina, cujos valores influenciam na construção dos modelos. A literatura relata a existência de algumas abordagens que solucionam a escolha destes valores. Na presente pesquisa, para projetar pesquisa sistemática para testar diferentes configurações de aprendizado de máquina é utilizado a técnica de *grid search*, que executa uma combinação automatizada na utilizando do intervalo de valores para o número de neurônios por camada oculta e o

intervalo de valores das entradas de atraso para cada método de aprendizado de máquina e técnica de pré-processamento de dados conforme definições da tabela V.1.

Tabela V.1: Valores dos principais atributos de aprendizado de máquina utilizados

| **Entradas de atraso** | **Neurônios por camada oculta** | **Aprend máquina** | **Pré-Proc** |
|---|---|---|---|
| 3..9 | 2..10 | emlp | swmm |
| 3..9 | 2..10 | emlp | diff |
| 3..9 | 2..10 | eelm | swmm |
| 3..9 | 2..10 | eelm | diff |
| 3..9 | 1..10 | mlp | an |
| 3..9 | 1..10 | mlp | swmm |
| 3..9 | 1..10 | mlp | diff |
| 3..9 | 1..10 | elm | an |
| 3..9 | 1..10 | elm | swmm |
| 3..9 | 1..10 | elm | diff |
| 3..9 | 1..10 | svm | an |
| 3..9 | 1..10 | svm | swmm |
| 3..9 | 1..10 | svm | diff |
| 3..9 | 1..10 | rf | an |
| 3..9 | 1..10 | rf | swmm |
| 3..9 | 1..10 | rf | diff |

Por fim, a função de ativação restringe a amplitude da saída de um neurônio a um valor finito. Este valor pode variar conforme o tipo de função de ativação que se estiver utilizando. Nos experimentos de previsão das séries temporais de fertilizantes, é utilizada a função de ativação hiperbólica tangente nos modelos de aprendizado de máquina elm, eelm, mlp e emlp, cuja definição podemos visualizar na Figura V.1 .

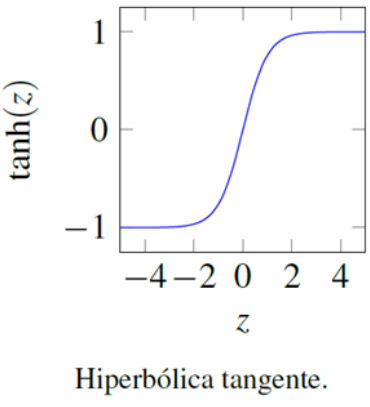



Figura V.1: Função de ativação dos métodos de aprendizado de máquina utilizados.

Toda avaliação experimental foi desenvolvida em R [Shumway and Stoffer, 2017]. O número de modelos criados durante a otimização da construção do modelo foi de 225.738 para todas as séries temporais. Para agilizar o processo de otimização, esses modelos foram criados em paralelo usando *Sparklyr* [Venkataraman et al., 2016] em uma estação de trabalho com 16 núcleos e 128 GB de RAM.

## V.2 Análise da otimização da construção do modelo

A Figura V.2 apresenta o desempenho geral de cada abordagem em todas as séries temporais considerando todos os cenários de previsão de passos à frente e valores de replicação durante validação-treinamento. Cada barra corresponde ao SMAPE mediano, cercado pelo seu intervalo de confiança para todos os cenários de previsão usando essa abordagem. Quanto menor o valor do SMAPE, melhor é a previsão. A primeira barra corresponde ao ARIMA, sendo este método utilizado como linha de base para comparação com outras abordagens analíticas de dados. A barra de nome `opt` é a forma proposta neste trabalho, enquanto que as outras barras correspondem a otimização dos hiperparâmetros com o modelo fixado.

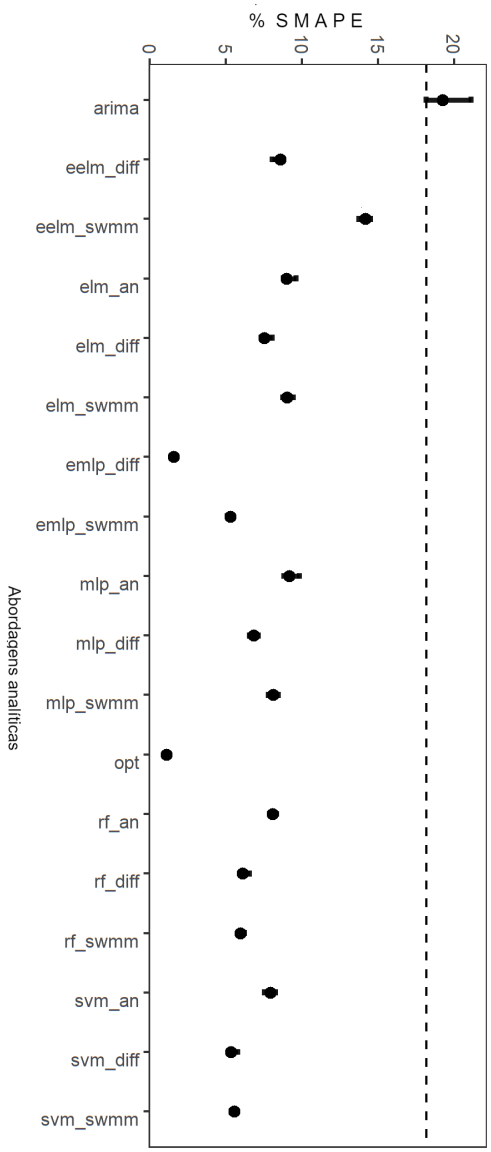



Figura V.2: Erro geral do SMAPE (em porcentagem) do consumo de fertilizantes em todas as abordagens durante a validação-treinamento.

De acordo com os resultados da Figura V.2, `opt` superou todos os métodos durante a fase de

validação-treinamento. Esse resultado é esperado devido ao seu processo de otimização intrínseco. Além disso, excluindo `opt`, considerando o limite superior do intervalo de confiança do SMAPE para cada abordagem, o `emlp_diff` foi o método mais consistente durante a fase de validação para todas as séries temporais. Devido a isso, foi o método selecionado para comparação com `arima` e `opt`. Por fim, todas as abordagens tiveram um desempenho significativamente melhor que o `arima`.

### V.3 Desempenho geral das previsões das principais abordagens

Considerando todos os cenários de avaliação experimental (todos os cenários do conjunto de testes), a Figura V.3 mostra a mediana das diferenças percentuais entre o SMAPE do `arima` e o SMAPE de cada abordagem cercado por seu intervalo de confiança. Isso significa que observações maiores ou menores que zero, respectivamente, correspondem ao caso em que a abordagem forneceu desempenho de previsão melhor ou pior do que `arima`. A maioria das abordagens era melhor que a `arima`, incluindo `emlp_diff`.

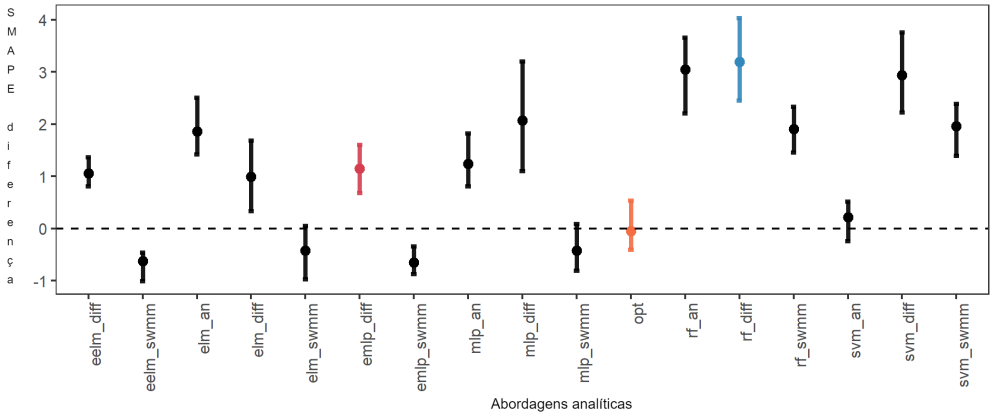



Figura V.3: A diferença geral (em porcentagem) de erros SMAPE das principais abordagens em comparação com `arima` durante o teste

Embora `opt` tenha tido um desempenho superior durante a fase de validação-treinamento, isso não o levou a previsões significativamente melhores durante a fase de teste. Outro resultado interessante é que a escolha da abordagem mais consistente durante a validação-treinamento (ou seja, a abordagem `emlp_diff`) levou a previsões melhores do que o `arima` durante o teste. Isso corrobora com iniciativas de aprendizagem de transferência homogênea [Weiss et al., 2016].

Além disso, o `rf_diff` ficou em torno de 3% melhor que o `arima` durante o teste. No entanto, ele não foi escolhido durante a validação-treinamento. Este resultado abre espaço para estudarmos maneiras de selecionarmos melhores abordagens durante a validação-treinamento. Assim, `rf_diff` também é apresentado durante comparações para fins de análises *what-if* (Kegel et al., 2017).

A Figura V.4 apresenta a porcentagem de vezes que cada abordagem superou o `arima` durante o teste. Novamente, o `emlp_diff` foi mais eficiente que o `arima`, mas algumas outras abordagens

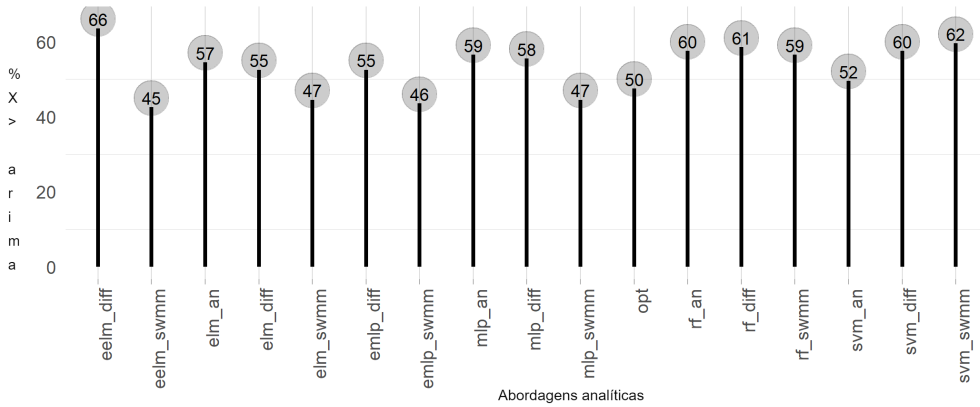



Figura V.4: A porcentagem de vezes que cada abordagem superou `arima` durante o teste

foram ainda melhores, como `eelm_diff` e `rf_diff`. Além disso, `opt` só foi melhor do que `arima` na metade dos casos. Esse resultado indica que procurar por abordagens mais consistentes em relação a todas as séries temporais é melhor do que otimizar cada série temporal.

## V.4 Desempenho das abordagens em relação a países e fertilizantes

Para aprofundar a análise, apresentamos o desempenho relativo de previsões das principais abordagens em comparação com o `arima` para cada país (Figura V.5) e cada fertilizante (Figura V.6). Valores significativamente maiores que zero fornecem previsões melhores do que `arima`.

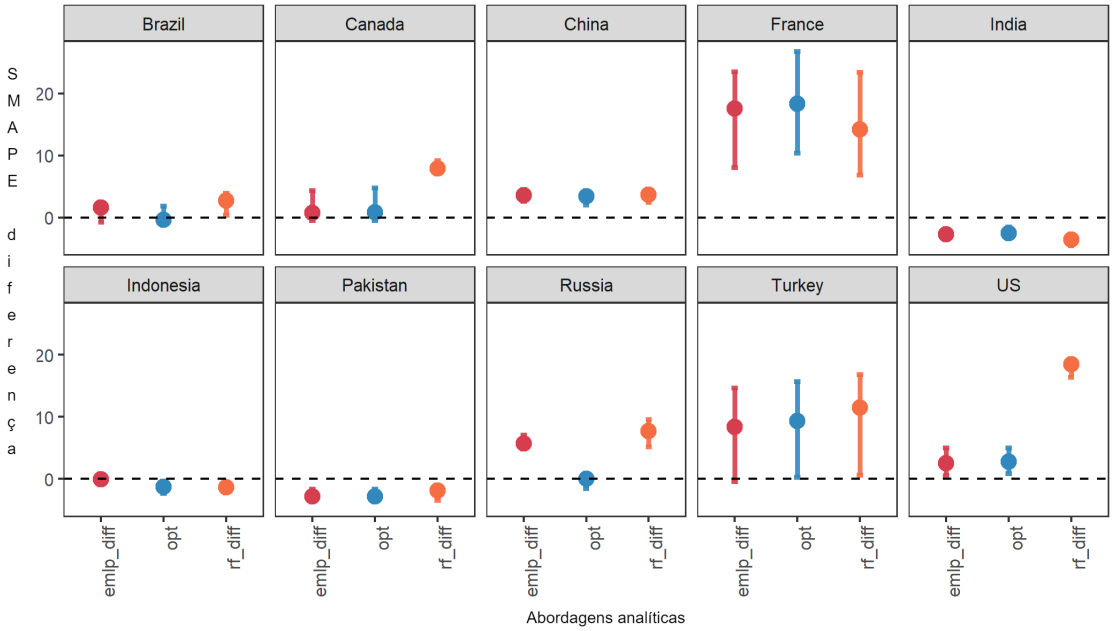



Figura V.5: A diferença (em porcentagem) de erros de abordagem SMAPE em comparação com `arima` durante o teste para cada país

Com relação à análise de cada país, é possível observar que `rf_diff` obteve um desempenho nas previsões superior à maioria das outras abordagens. No entanto, ele falhou em fornecer previsões

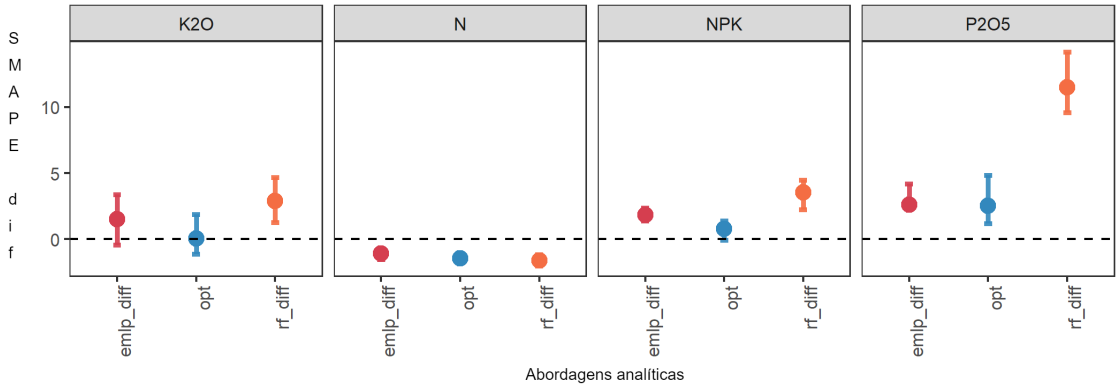
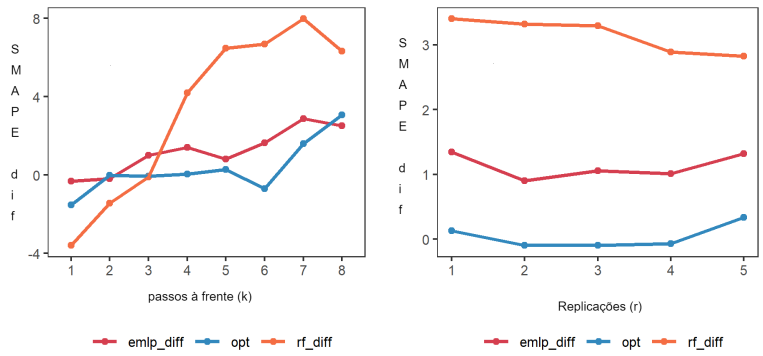



Figura V.6: A diferença (em porcentagem) de erros de SMAPE das abordagens em comparação com `arima` durante o teste para cada fertilizante

relativamente melhores do que `arima` para as séries temporais da Índia, Indonésia e Paquistão. Na maioria das vezes `emlp_diff` foi melhor que o `arima`, mas falhou em fornecer previsões melhores do que o `arima` na Índia e no Paquistão. Nas dez séries temporais, `opt` obteve um desempenho nas previsões de oitenta por cento pior do que as outras abordagens. Ele também falhou em fornecer previsões melhores do que a `arima` na Índia, Indonésia e Paquistão.

Quando se trata de diferentes tipos de fertilizantes, é possível observar que $P_2O_5$ foi o tipo de fertilizante no qual todas as abordagens levam a previsões significativamente melhor que `arima`. O N foi o tipo de fertilizante em que geralmente as abordagens não obtiveram previsões melhores do que o `arima`. Finalmente, para $K_2O$ e NPK, `rf_diff` teve desempenho significativamente melhor do que o `arima`.

## V.5 Desempenho de abordagens relacionadas à passo a frente e replicações




(a) (b)

Figura V.7: A influência de previsões passos à frente (a) e replicação (b) na diferença (em porcentagem) de erros do SMAPE entre as principais abordagens e `arima`

A Figura V.7.a apresenta a influência das previsões passos à frente em cada abordagem quando comparado ao `arima`. Valores acima de zero indicam desempenho de previsão melhor do que `arima`

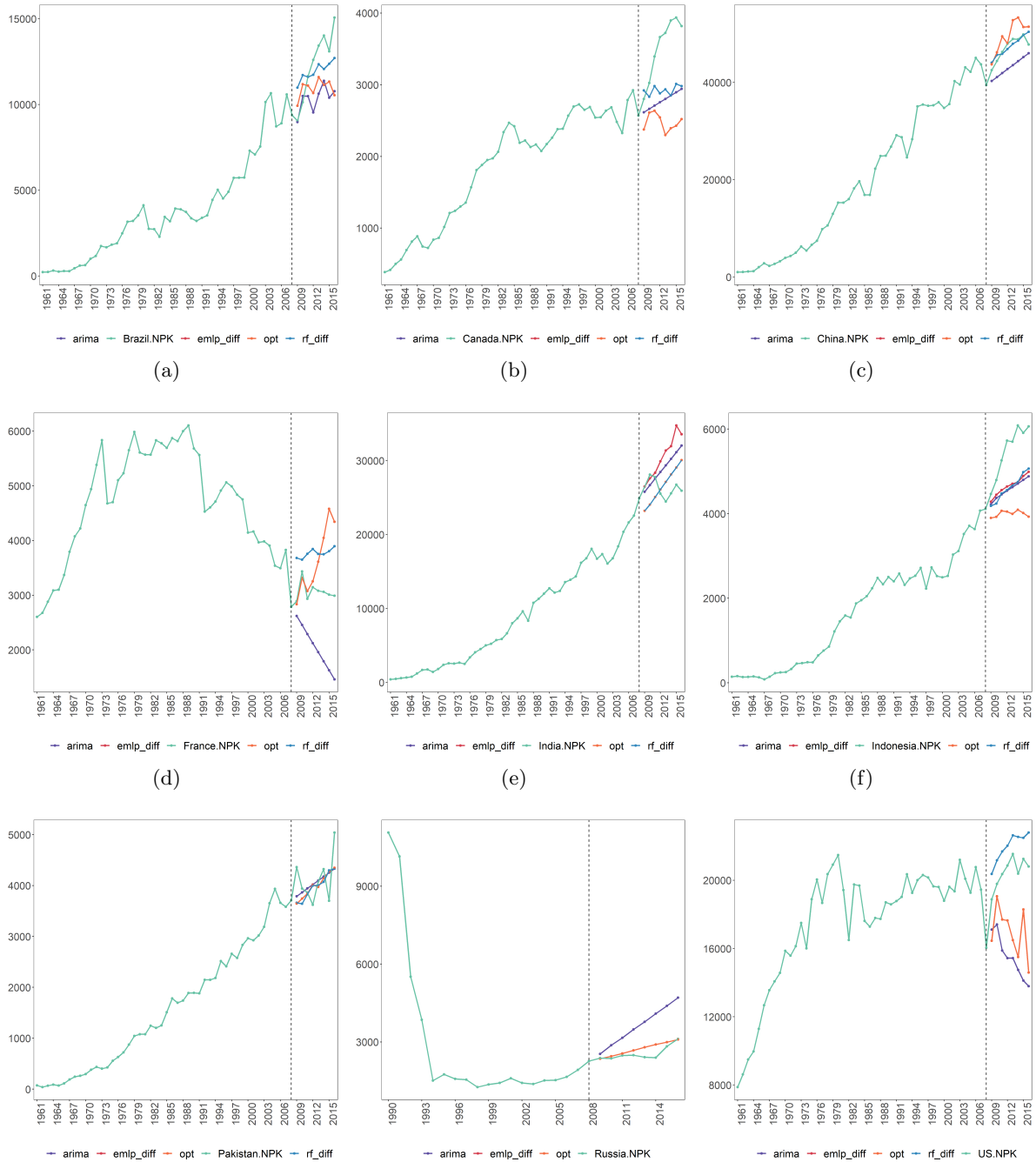



Figura V.8: Comparação de oito previsões de passos à frente (2009-2016 usando 2008 como linha base) para o consumo de NPK usando `arima`, `emlp_diff`, `opt`, `rf_diff`

(em porcentagem). Naturalmente, à medida que o valor do passo à frente avança, as previsões se tornam mais difíceis.

Todas as abordagens melhoraram seu desempenho relativo à medida que o número de passos à frente aumentam. A `opt` foi pior do que o `arima` entre 1 e 6 no cenário de previsão de passos à frente. Ele só forneceu previsões relativamente melhores entre 7 e 8 previsões de passos à frente. As previsões `emlp_diff` foram melhores do que `arima` de 3 a 8 previsões de passos à frente. Finalmente, as previsões de `rf_diff` foram melhores do que o `arima` de 4 a 8 previsões de passos à frente.

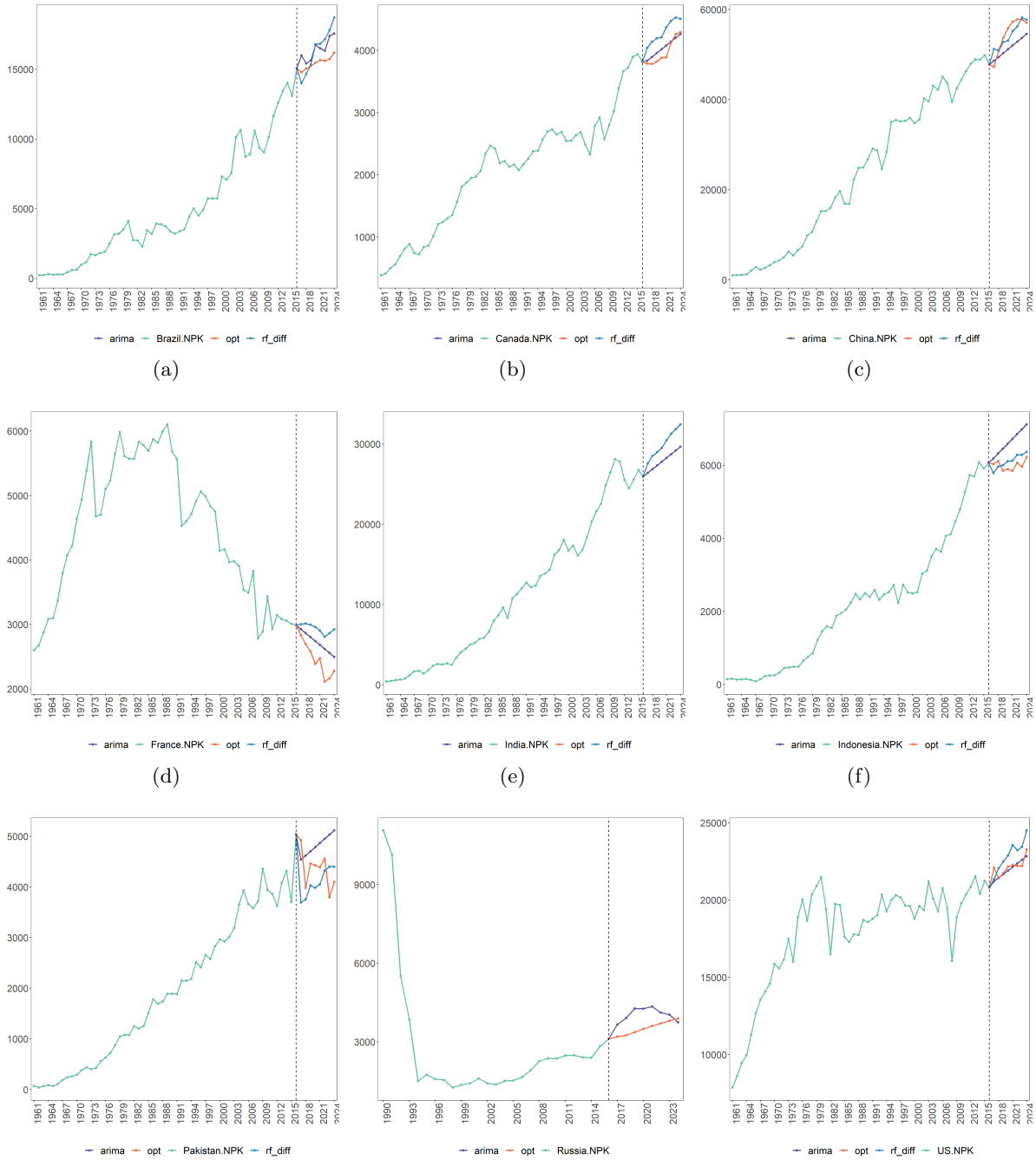



Figura V.9: Oito previsões de passos à frente (2017-2024 usando 2016 como linha base) para o consumo de NPK usando `arima`, `opt`, `rf_diff`

A Figura V.7.b apresenta como a replicação na validação-treinamento influencia a previsão do modelo durante o teste. Sem utilizar replicação levou `rf_diff` a obter um melhor desempenho. Por outro lado, uma replicação cinco vezes maior foi melhor para `emlp_diff` and `rf_diff`.

### V.6 Qualidade das previsões e tendências para NPK

Para avaliar intuitivamente a qualidade das previsões, a Figura V.8 apresenta previsões de passos à frente de NPK (de 2009 a 2016) usando 1961 a 2008 como validação-treinamento (com cinco repetições). O `rf_diff` apresentou melhores previsões no Brasil, Canadá, China, Índia e EUA. Enquanto `opt` apresentou melhores previsões na França, Índia e Rússia. Em todos os casos, o `arima` foi superado por pelo menos uma abordagem e, na maioria das vezes, por duas abordagens. Um aspecto positivo do `arima` foi sua capacidade de apresentar a tendência correta (aumento ou diminuição) do NPK em oito dos dez países.

Finalmente, a Figura V.9 apresenta a tendência para o período de 2017 a 2024 do consumo de NPK, onde é utilizado `arima`, `opt`, `rf_diff`, tendo como validação-treinamento a utilização referente ao período 1961 a 2016 das 40 diferentes séries temporais. Nesse cenário de previsão, quase todas as abordagens apresentaram tendências semelhantes. O consumo do Brasil, Canadá, China, Índia, Indonésia, Rússia e EUA é esperado um aumento. Na França, espera-se uma diminuição do consumo de fertilizantes. No Paquistão, o `arima` indica uma expectativa do aumento no consumo de NPK, enquanto outras abordagens indicaram preservação no nível de consumo.

# Capítulo VI Conclusões

No presente trabalho, abordagens de análise de dados foram avaliadas, a fim de melhorar as previsões do consumo de fertilizantes sob diferentes horizontes de passos à frente. Essa pesquisa explorou maneiras de otimizar a construção de modelo, levando em consideração diferentes abordagens, tais como combinações de pares entre pré-processamento de dados e métodos de aprendizado de máquina. Essas abordagens foram avaliadas usando 40 séries temporais diferentes de fertilizantes, correspondentes aos quatro principais fertilizantes dos dez principais países que os demandam.

Os experimentos dessa pesquisa foram divididos em duas fases: validação-treinamento e teste. Durante a fase de validação-treinamento, `opt` superou todos os métodos (batendo `arima` em quase 100% dos casos). Esse comportamento mudou durante a fase de teste, levando a resultados ligeiramente melhores que os obtidos por `arima` (50% dos casos). O segundo melhor método na fase de treinamento, `emlp_diff` selecionado por consistência (inspirada na aprendizagem de transferência homogênea), foi melhor na fase de teste, obtendo melhores desempenhos do que os `arima` em torno de 55% dos casos. Adicionalmente, `rf_diff` foi geralmente 3% melhor do que o `arima` durante a fase de teste. No entanto, ele não havia sido escolhido durante a fase de validação-treinamento. Tais resultados obtidos corroboraram com as hipóteses de que a otimização da construção do modelo precisa considerar a abordagem que é mais consistente em ter boas previsões para todas as diferentes séries temporais.

Em todos os experimentos, o modelo `arima` foi usado como *baseline* para a previsão do consumo de fertilizantes. O presente trabalho também apresenta a qualidade das previsões para o consumo de `NPK` dos dez principais países que os demandam. Também apresenta as tendências até 2024, usando observações de 1961 a 2016 como um conjunto de validação-treinamento. Nos principais resultados das tendências, o consumo do Brasil, Canadá, China, Índia, Indonésia, Rússia e EUA é esperado um aumento. Na França, espera-se uma diminuição do consumo de fertilizantes. No Paquistão o `arima` indica expectativa do aumento do consumno de NPK. Vale ressaltar que em todos os experimentos, nos deparamos com 56 observações em cada uma das 40 séries temporais, o que faz desse pequeno número de observações uma limitação natural que pode ter refletido nos resultados.

Como consequência dos estudos realizados no presente trabalho, ocorreu a divulgação dos resultados alcançados, bem como dos experimentos desenvolvidos na presente pesquisa. Durante o

período de estudos, o artigo intitulado "Uso de ciência de dados para predição do consumo de fertilizantes no Brasil" foi aceito e publicado nos anais no XIV BreSci – Brazilian e-Science Workshop 2020 (BreSci) do XL Congresso da Sociedade Brasileira de Computação (CSBC 2020).

Por fim, a análise dos resultados das previsões para países e tipos de fertilizantes mostrou que ainda há espaço para o desenvolvimento de formas mais avançadas de seleção de modelo, pois a abordagem proposta não foi consistente para alguns fertilizantes e países. Pesquisas futuras podem concentrar-se no desenvolvimento de métodos de otimização que projetem uma combinação automatizada para testar diferentes configurações de modelos de aprendizado de máquina, como por exemplo algoritmos genéticos e outras meta-heurísticas [Kerschke et al., 2018], para melhorar a seleção de modelos.

## Referências Bibliográficas

Nigel Lewis. *Neural networks for time series forecasting with R : an intuitive step by step blueprint for beginners*. AusCov, Place of publication not identified, 2017. ISBN 978-1544752952. , 11, 12

D.S. Palmer, N.M. O'Boyle, R.C. Glen, and J.B.O. Mitchell. Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47(1):150–158, 2007. , 13

G.P. Zhang and V.L. Berardi. Time series forecasting with neural network ensembles: An application for exchange rate prediction. *Journal of the Operational Research Society*, 52(6):652–664, 2001. , 13, 14

UN. United nations. Technical report, https://www.un.org/en/, May 2019. 1, 4

D. Tilman, K.G. Cassman, P.A. Matson, R. Naylor, and S. Polasky. Agricultural sustainability and intensive production practices. *Nature*, 418(6898):671–677, 2002. 1

FAO. Food and agriculture organization of the united nations. Technical report, http://www.fao.org, October 2019. 1, 7

Keith Kirkpatrick. Technologizing Agriculture. *Commun. ACM*, 62(2):14–16, January 2019. ISSN 0001-0782. 1

W.M. Stewart and T.L. Roberts. Food security and the role of fertilizer in supporting it. In *Procedia Engineering*, volume 46, pages 76–82, 2012. 1

MF Attallah, SS Metwally, SI Moussa, and Mohamed A Soliman. Environmental impact assessment of phosphate fertilizers and phosphogypsum waste: elemental and radiological effects. *Microchemical Journal*, 146:789–797, 2019. 1

John W McArthur and Gordon C McCord. Fertilizing growth: Agricultural inputs and their effects in economic development. *Journal of development economics*, 127:133–152, 2017. 4

H.J. Styhr Petersen. Forecasting Danish nitrogen fertilizer consumption. *Industrial Marketing Management*, 6(3):211–222, 1977. 5, 6

D. Deadman and S. Ghatak. Forecasting fertilizer consumption and production: Long- and short-run models. *World Development*, 7(11-12):1063–1072, 1979. 5, 6

B. Gilland. Cereals, nitrogen and population: an assessment of the global trends. *Endeavour*, 17 (2):84–88, 1993. 5, 6

R.W. Howarth, E.W. Boyer, W.J. Pabich, and J.N. Galloway. Nitrogen use in the United States from 1961-2000 and potential future trends. *Ambio*, 31(2):88–96, 2002. 5, 6

A. Dobermann and K.G. Cassman. Cereal area and nitrogen use efficiency are drivers of future nitrogen fertilizer consumption. *Science in China. Series C, Life sciences / Chinese Academy of Sciences*, 48 Spec No:745–758, 2005. 5, 6

W. Zhang and X. Zhang. A forecast analysis on fertilizers consumption worldwide. *Environmental Monitoring and Assessment*, 133(1-3):427–434, 2007. 5, 6

F. Tenkorang and J. Lowenberg-Deboer. Forecasting long-term global fertilizer demand. *Nutrient Cycling in Agroecosystems*, 83(3):233–247, 2009. 5, 6

E. Ogasawara, D. De Oliveira, F. Paschoal Jr., R. Castaneda, M. Amorim, R. Mauro, J. Soares, J. Quadros, and E. Bezerra. A forecasting method for fertilizers consumption in Brazil. *International Journal of Agricultural and Environmental Information Systems*, 4(2):23–36, 2013. 5, 6

M.V. Pires, D.A. Da Cunha, S. De Matos Carlos, and M.H. Costa. Nitrogen-use efficiency, nitrous oxide emissions, and cereal production in Brazil: Current trends and forecasts. *PLoS ONE*, 10 (8), 2015. 5, 6

IFA. International fertilizers association. Technical report, https://www.fertilizer.org, June 2019. 7

TL Roberts et al. The role of fertilizer in growing the world's food. *Better crops*, 93(2):12–15, 2009. 7

Nutrição de Planta Ciência e Tecnologia. Fertilizantes, 2018. 8

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, New Jersey, 5 edition edition, June 2015a. ISBN 978-1-118-67502-1. 8

Damodar Gujarati. *Basic Econometrics*. McGraw-Hill/Irwin, Boston; Montreal, 4 edition, March 2002a. ISBN 978-0-07-247852-5. 8

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, New York, NY, 4 edition, April 2017. ISBN 978-3-319-52451-1. 8, 19

Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, Cambridge, Mass, 3 edition, August 2010. ISBN 978-0-470-41435-4. 8

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015b. 9

Damodar Gujarati. Basic econometrics, 2002b. 9

R. Salles, K. Belloze, F. Porto, P.H. Gonzalez, and E. Ogasawara. Nonstationary time series transformation methods: An experimental review. *Knowledge-Based Systems*, 164:274–291, 2019. 9, 10

Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Haryana, India; Burlington, MA, 3 edition, July 2011. ISBN 978-0-12-381479-1. 10

E. Ogasawara, L.C. Martinez, D. De Oliveira, G. Zimbrão, G.L. Pappa, and M. Mattoso. Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *Proceedings of the International Joint Conference on Neural Networks*, 2010. 10

K. Weiss, T.M. Khoshgoftaar, and D.D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), 2016. 11, 21

Guang-Bin Huang, Lei Chen, Chee Kheong Siew, et al. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17(4):879–892, 2006. 12

Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. Extreme learning machine for multilayer perceptron. *IEEE transactions on neural networks and learning systems*, 27(4):809–821, 2015. 12

N. Sapankevych and R. Sankar. Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38, 2009. 13

S.F. Crone, M. Hibon, and K. Nikolopoulos. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, 27(3):635–660, 2011. 14

C. Thornton, F. Hutter, H.H. Hoos, and K. Leyton-Brown. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume Part F128815, pages 847–855, 2013. 16

R.J. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008. 17

Rebecca Salles, Eduardo Bezerra, Jorge Soares, and Eduardo Ogasawara. Evaluating Linear Models as a Baseline for Time Series Imputation. In *XXX Simpósio Brasileiro de Banco de Dados*, Petrópolis, RJ, October 2015. 17

S. Venkataraman, Z. Yang, D. Liu, E. Liang, H. Falaki, X. Meng, R. Xin, A. Ghodsi, M. Franklin, I. Stoica, and M. Zaharia. SparkR: Scaling R programs with spark. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, volume 26-June-2016, pages 1099–1104, 2016. 19

P. Kerschke, H.H. Hoos, F. Neumann, and H. Trautmann. Automated algorithm selection: Survey and perspectives. *Evolutionary Computation*, 27(1):3–45, 2018. 28