



UNIVERSIDADE FEDERAL DA BAHIA - UFBA  
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA - IME  
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA  
DISSERTAÇÃO DE MESTRADO



ANÁLISE E PREVISÃO DE MATERIAIS  
PARTICULADOS EM MINAS GERAIS  
UTILIZANDO MODELOS ESTATÍSTICOS, REDES  
NEURAIS E ABORDAGENS HÍBRIDAS

KIM LEONE SOUZA DA SILVA

Salvador-Bahia

ANÁLISE E PREVISÃO DE MATERIAIS  
PARTICULADOS EM MINAS GERAIS  
UTILIZANDO MODELOS ESTATÍSTICOS, REDES  
NEURAIS E ABORDAGENS HÍBRIDAS

KIM LEONE SOUZA DA SILVA

Dissertação de Mestrado apresentada ao  
Colegiado da Pós-Graduação em Matemática da  
Universidade Federal da Bahia como requisito  
parcial para obtenção do título de Mestre em  
Matemática. Área de Concentração: Estatística

**Orientador: Prof. Dr. Paulo Jorge Canas  
Rodrigues**

Salvador-Bahia

2024

Ficha catalográfica elaborada pela Biblioteca Universitária de  
Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

S586 Silva, Kim Leone Souza da

Análise e Previsão de Materiais Particulados em Minas Gerais Utilizando Modelos Estatísticos, Redes Neurais e Abordagens Híbridas/ Kim Leone Souza da Silva. – Salvador, 2024.

182 f.

Orientador: Prof. Dr. Paulo Jorge Canas Rodrigues

Dissertação (Mestrado) – Universidade Federal da Bahia. Instituto de Matemática, 2024.

1. Poluição. 2. Material Particulado. 3. Previsões de Séries Temporais. 4. Estatística. I. Rodrigues, Paulo Jorge Canas. II. Universidade Federal da Bahia. III. Título.

CDU: 519.2:614.71

# **Análise e previsão de materiais particulados em Minas Gerais utilizando modelos estatísticos, redes neurais e abordagens híbridas.**

Kim Leone Souza da Silva

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática.

## **Banca examinadora**

---

Prof<sup>o</sup> Dr<sup>o</sup> Paulo Jorge Canas Rodrigues (orientador - UFBA)

---

Prof<sup>o</sup>. Dr<sup>o</sup>. Valderio Anselmo Reisen (Interno - UFBA)

---

Prof<sup>a</sup>. Dra. Gecynalda Soares da Silva Gomes (Externa)



# Agradecimentos

Agradeço à minha família, amigos, colegas e professores que foram parte essencial e muito significativa nesta jornada.

Ao meu orientador, Paulo Canas, pela orientação e incentivo.

Aos membros da banca examinadora, pela disponibilidade.

À CAPES pelo apoio financeiro.

*“Models are, for the most part, caricatures of reality, but if they are good, like good caricatures, they portray, though perhaps in a disturbed manner, some features of the real world.”*

(Mark Kac)

# Resumo

O impacto humano no planeta é evidente, e a necessidade de mitigar esses efeitos torna-se cada vez mais urgente. A poluição atmosférica, por exemplo, está intrinsecamente ligada ao meio ambiente e à saúde humana. Materiais particulados com diâmetros inferiores a 10 e 2,5 micrômetros, chamados de  $PM_{10}$  e  $PM_{2,5}$ , respectivamente, podem impactar gravemente a saúde humana devido à sua capacidade de penetrar no sistema respiratório. Este estudo foca na previsão das concentrações de  $PM_{10}$  e  $PM_{2,5}$  em estações da rede de monitoramento do estado brasileiro de Minas Gerais. Na busca por melhores desempenhos preditivos, foram desenvolvidos modelos híbridos que combinam abordagens paramétricas, não paramétricas e redes neurais na previsão de materiais particulados. O desempenho desses modelos é avaliado em diferentes horizontes de previsão. Os resultados indicam que os modelos híbridos propostos apresentam desempenho competitivo em diversos horizontes de tempo para ambos os poluentes, quando comparados ao uso dos modelos de forma isolada. Essas descobertas ressaltam a importância de combinar diferentes abordagens de modelagem, aproveitando as forças de cada método para capturar padrões complexos e melhorar a precisão das previsões. Além disso, a integração de técnicas paramétricas, não paramétricas e redes neurais permite uma modelagem mais robusta e flexível, capaz de se adaptar a diferentes padrões temporais e características dos dados. Os métodos e insights obtidos neste estudo podem ser aplicados a pesquisas sobre outros poluentes em diferentes localidades, contribuindo, assim, para uma compreensão mais ampla da qualidade do ar em contextos variados e apoiando políticas públicas mais eficazes na gestão da poluição atmosférica.

**Palavras-chave:** Poluição, Material Particulado, Previsões de Séries Temporais, Estatística

# Abstract

The human impact on the planet is evident, and the need to mitigate these effects is becoming increasingly urgent. Atmospheric pollution, for example, is intrinsically linked to the environment and human health. Particulate matter with diameters smaller than 10 and 2.5 micrometers, called  $PM_{10}$  and  $PM_{2.5}$ , respectively, can severely impact human health due to their ability to penetrate the respiratory system. This study focuses on predicting the concentrations of  $PM_{10}$  and  $PM_{2.5}$  at monitoring stations in the Brazilian state of Minas Gerais. In the pursuit of better predictive performance, hybrid models were developed that combine parametric, non-parametric, and neural network approaches for particulate matter forecasting. The performance of these models is evaluated across different forecasting horizons. The results indicate that the proposed hybrid models exhibit competitive performance at various time horizons for both pollutants when compared to the use of the models in isolation. These findings highlight the importance of combining different modeling approaches, leveraging the strengths of each method to capture complex patterns and improve the accuracy of predictions. Furthermore, the integration of parametric, non-parametric, and neural network techniques allows for a more robust and flexible model, capable of adapting to different temporal patterns and data characteristics. The methods and insights obtained in this study can be applied to research on other pollutants in different locations, thus contributing to a broader understanding of air quality in various contexts and supporting more effective public policies in air pollution management.

**Keywords:** Pollution, Particulate Matter, Time Series Forecasts, Statistics

# Sumário

<b>Lista de Figuras</b>	<b>11</b>
<b>Lista de Tabelas</b>	<b>12</b>
<b>1 INTRODUÇÃO</b>	<b>1</b>
1.1 Revisão da Literatura . . . . .	2
1.2 Apresentação dos Capítulos . . . . .	3
<b>2 SÉRIES TEMPORAIS</b>	<b>5</b>
2.1 Noções Básicas . . . . .	5
2.2 Modelos SARIMA . . . . .	6
2.3 Modelos ARFIMA . . . . .	9
2.4 Análise Espectral Singular (SSA) . . . . .	10
2.5 Long Short-Term Memory (LSTM) . . . . .	12
2.6 Modelos Híbridos . . . . .	15
2.7 Validação . . . . .	16
<b>3 RESULTADOS E DISCUSSÃO</b>	<b>18</b>
3.1 Ajuste e Previsão de Séries Temporais em dados de $PM_{10}$ . . . . .	19
3.1.1 Análise Descritiva . . . . .	19
3.1.2 Ajuste dos Modelos para $PM_{10}$ . . . . .	26
3.1.3 Previsão de $PM_{10}$ . . . . .	27
3.2 Modelos Híbridos para Séries com Memória Longa em dados de $PM_{2,5}$ . . . . .	32
3.2.1 Análise Descritiva . . . . .	32
3.2.2 Ajuste e Previsão dos Modelos para $PM_{2,5}$ . . . . .	35
<b>4 CONCLUSÃO</b>	<b>38</b>
4.1 Resultados Alcançados . . . . .	39
4.2 Trabalhos Futuros . . . . .	40
<b>Referências Bibliográficas</b>	<b>41</b>

# Lista de Figuras

2.1	Estrutura de uma LSTM. Fonte: [Olah 2015]	13
2.2	Esquema da atualização do estado da célula de uma LSTM. Fonte: [Olah 2015]	13
2.3	Esquema da atualização do estado oculto da célula de uma LSTM. Fonte: [Olah 2015]	14
2.4	Diagrama ilustrativo das abordagens híbridas para os dados de $PM_{10}$ e $PM_{2,5}$ .	16
3.1	O painel esquerdo mostra um mapa da América do Sul, enquanto o painel direito exibe o mapa que inclui o estado brasileiro de Minas Gerais com marcadores indicando a localização das estações de monitoramento.	19
3.2	Séries temporais das concentrações médias semanais de $PM_{10}$ para todas as 29 estações de monitoramento.	20
3.3	Boxplots das concentrações de $PM_{10}$ (em $\mu g/m^3$ ) por hora do dia para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de $PM_{10}$ observadas entre meia-noite e 1h).	22
3.4	Boxplots das concentrações de $PM_{10}$ (em $\mu g/m^3$ ) por dia do mês para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de $PM_{10}$ observadas no primeiro dia do mês).	23
3.5	Boxplots das concentrações de $PM_{10}$ (em $\mu g/m^3$ ) por mês do ano para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de $PM_{10}$ observadas em janeiro).	24
3.6	Boxplots das concentrações de $PM_{10}$ (em $\mu g/m^3$ ) por ano para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de $PM_{10}$ observadas em 2015).	25

3.7	Boxplots do RMSE e SMAPE para o ajuste e a previsão dos modelos, considerando as 29 estações de monitoramento. Os quatro modelos analisados são: SARIMA, LSTM, SSA e o modelo híbrido que combina SARIMA e LSTM. . . . .	31
3.8	Séries temporais das concentrações de $PM_{2,5}$ nas seis estações de monitoramento consideradas após imputação de dados ausentes. . . . .	32
3.9	Boxplots da concentração de $PM_{2,5}$ (em $\mu m^2$ ) por hora em cada estação. . .	33
3.10	Boxplots da concentração de $PM_{2,5}$ (em $\mu m^2$ ) por dia em cada estação. . .	33
3.11	Boxplots da concentração de $PM_{2,5}$ (em $\mu m^2$ ) por mês em cada estação. . .	34
3.12	Boxplots da concentração de $PM_{2,5}$ (em $\mu m^2$ ) por ano em cada estação. . .	34
3.13	Autocorrelação das concentrações de $PM_{2,5}$ em cada estação. . . . .	35

# Lista de Tabelas

3.1	Informações detalhadas para cada estação de monitoramento, incluindo código, nome da estação, empresa responsável pela estação de monitoramento, longitude, latitude e taxa de valores ausentes. . . . .	19
3.2	Medidas de acurácia (RMSE e SMAPE) para avaliação do ajuste dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM.	26
3.3	Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão uma hora à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM. . . . .	28
3.4	Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão 24 horas (um dia) à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM. . . . .	29
3.5	Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão 168 horas (uma semana) à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM. . . . .	30
3.6	Avaliação do RMSE do ajuste e das previsões para os modelos SARIMA, ARFIMA, LSTM e Modelos Híbridos ARFIMA-LSTM e SSA-ARFIMA-LSTM em horizontes de 1, 7 e 30 dias . . . . .	36
3.7	Avaliação do SMAPE do ajuste e das previsões para os modelos SARIMA, ARFIMA, LSTM e Modelos Híbridos ARFIMA-LSTM e SSA-ARFIMA-LSTM em horizontes de 1, 7 e 30 Dias . . . . .	37

# Capítulo 1

## INTRODUÇÃO

O impacto humano no planeta é evidente, e a necessidade de mitigar esses efeitos torna-se cada vez mais urgente. A poluição atmosférica, por exemplo, está intrinsecamente ligada ao meio ambiente e à saúde humana. Um dos fatores críticos que afetam a saúde é o tamanho das partículas presentes na atmosfera, uma vez que sua permanência e capacidade de penetração no sistema respiratório variam de acordo com essas dimensões. Nesse contexto, o monitoramento dos chamados materiais particulados, especialmente aqueles com tamanho inferior a 10 ou 2,5 micrômetros ( $PM_{10}$  e  $PM_{2,5}$ ), é de suma importância, pois o impacto sobre a saúde está diretamente relacionado ao tamanho dessas partículas.

Neste sentido, o presente trabalho tem como objetivo analisar as concentrações de material particulado em municípios do estado de Minas Gerais, Brasil. Para isso, as concentrações desse poluente são monitoradas ao longo do dia, gerando dados que refletem os níveis de poluição a cada hora. Esses dados são organizados em séries temporais, permitindo uma compreensão mais aprofundada das variações das concentrações de material particulado ao longo do tempo. Essa análise é fundamental para identificar padrões, tendências e possíveis fontes de poluição, contribuindo assim para a formulação de políticas públicas voltadas à melhoria da qualidade do ar e à proteção da saúde da população.

No estudo das séries temporais, diversas técnicas são utilizadas para a realização de previsões. Este trabalho investiga e compara diferentes abordagens de modelagem, visando identificar os métodos mais eficazes para prever os níveis de material particulado em diferentes localizações. A pesquisa adota uma abordagem abrangente, incorporando modelos paramétricos, não-paramétricos e redes neurais recorrentes. Além disso, propõe a utilização de modelos híbridos que combinam essas abordagens, buscando aprimorar a performance em comparação com a aplicação isolada de cada técnica. Ao longo do estudo, realizamos comparações entre as diferentes metodologias em variados horizontes de previsão. Essa análise proporciona insights valiosos sobre a eficácia preditiva de cada

modelo, permitindo identificar quais abordagens se destacam em contextos específicos. O objetivo é contribuir para o avanço das técnicas de previsão, oferecendo uma compreensão mais profunda das dinâmicas temporais dos dados analisados.

Para alcançar esses objetivos, foram analisadas as concentrações de  $PM_{10}$  e  $PM_{2,5}$  em alguns municípios de Minas Gerais, utilizando dados da GESAR - Gerência de Monitoramento da Qualidade do Ar e Emissões. Esses dados, que estão disponíveis publicamente conforme a Lei nº 12.527, de 18 de novembro de 2011, regulam o acesso à informação e garantem a transparência dos dados de monitoramento contínuo da qualidade do ar das estações que compõem a rede de monitoramento automático do estado. Com essas informações, foi possível avaliar a emissão de poluentes na região.

## 1.1 Revisão da Literatura

Baseados na média anual de  $PM_{10}$ , a Organização Mundial da Saúde (OMS) classificou Ahvaz no Irã como a cidade mais poluída do mundo com  $372 \mu g/m^3$  [Maleki et al. 2016]. No Brasil, por meio da Resolução Conama nº 005/1989, temos o Programa Nacional de Controle da Qualidade do Ar (Pronar), que é uma tentativa de construção das bases para uma política nacional de proteção da qualidade do ar. No entanto, ainda que o Pronar seja o início para uma política nacional de qualidade do ar, ele possui grande fragilidade jurídica já que a base legal que o sustenta é infralegal. Algumas das dificuldades enfrentadas são: a atualização periódica dos padrões nacionais, carências financeiras e de recursos humanos na sustentação de uma rede nacional de monitoramento, além da ausência de um abrangente inventário nacional de emissões. Para além disso, ainda existe uma clara assimetria entre as regiões do país e a maior parte dos instrumentos de gestão da qualidade do ar estão localizados no sudeste do Brasil [Artaxo 2021].

Na Europa o projeto Apheis desenvolveu diretrizes para análise e coleta de dados sobre a qualidade do ar e os impactos na saúde pública. Apheis é um sistema de vigilância cujo intuito é fornecer aos tomadores de decisão europeus, nacionais, regionais e locais, profissionais de saúde ambiental e ao público em geral informações atualizadas e fáceis de usar sobre poluição e saúde pública. O estudo apresentou o impacto na saúde em 19 cidades de países da Europa Oriental e Ocidental. Os resultados apontam que a redução à exposição de longo prazo de  $PM_{10}$  em  $5 \mu g/m^3$  poderia evitar cerca de 3300 a 7700 mortes anuais prematuras, 500 e 1000 das quais estão associadas à exposição de curto prazo. Apheis mostrou que na Europa urbana a poluição do ar atual têm um impacto não desprezível na saúde pública e que mesmo em cidades com baixa poluição do ar as medidas de prevenção podem reduzir danos [Medina et al. 2004].

Um estudo na região metropolitana de Lima, a capital do Peru, propôs uma

visualização espaço-temporal para análise dos níveis de  $PM_{10}$ . As maiores concentrações de  $PM_{10}$  foram registradas próximas a morros e estradas de tráfego intenso e ruas não pavimentadas [Encalada-Malca et al. 2021]. Em Berlim, 50% do  $PM_{10}$  é causado pelo transporte de longo alcance [Lenschow et al. 2001]. No Brasil, na região metropolitana de São Paulo (MASP), 40% da emissão de  $PM_{10}$  são provenientes de fontes móveis. Além disso, ozônio e  $PM_{10}$  são os poluentes com maior impacto na qualidade do ar na MASP. Um estudo, realizado nos túneis Jânio Quadros e Maria Maluf de São Paulo, indica que a emissão de veículos pesados a diesel é a maior fonte de material particulado fino  $PM_{2,5}$  [Sánchez-Ccoyllo et al. 2009].

Pesquisas na área de saúde sugerem que a exposição contínua a partículas tem diversos impactos na saúde dos indivíduos expostos [Mukherjee e Agrawal 2017, Sacks et al. 2011, Bae e Hong 2018, Thompson 2018]. Em Itabira (MG), o aumento de  $10 \mu g/m^3$  de  $PM_{10}$  foi associado ao aumento de doenças respiratórias no pronto-socorro. Os resultados do estudo mostraram que um aumento nos níveis de  $PM_{10}$  têm grande impacto na população exposta [Braga et al. 2007].

Um estudo realizado na Região da Grande Vitória (RGV), Espírito Santo, Brasil, utilizou o modelo SARIMAX (modelo Autorregressivo Integrado de Médias Móveis Sazonal com Regressores Exógenos) para modelar e prever as concentrações de  $PM_{10}$ . Neste estudo, a velocidade do vento e a precipitação pluvial foram significativas e ajudaram a melhorar o ajuste do modelo [Pinto, Reisen e Monte 2018]. Outro estudo, no município de Canoas, Rio Grande do Sul, mostrou que o modelo ARMAX, com a inclusão das variáveis exógenas Monóxido de Carbono (CO) e Dióxido de Enxofre (SO<sub>2</sub>), obteve melhor desempenho comparado aos modelos ARIMA, Alisamento Exponencial Simples, Biparamétrico e Holt-Winters, para a previsão de  $PM_{10}$  [Tibulo et al. 2021].

A aplicação de redes neurais na previsão de material particulado tem se mostrado promissora ao analisar padrões complexos e não lineares nos dados ambientais [Li, Hua e Wu 2020, Memarianfard e Hatami 2017]. Diversos estudos mostram que a combinação de modelos ARIMA (modelo Autorregressivo Integrado de Médias Móveis) com redes neurais pode ser útil para aumentar a precisão das previsões de materiais particulados [Li, Xu e Cheng 2022, Díaz-Robles et al. 2008, Samia, Kaouther e Abdelwahed 2012].

## 1.2 Apresentação dos Capítulos

Este trabalho está organizado em quatro capítulos. No Capítulo 1, é apresentada a introdução, que inclui uma breve revisão da literatura. O Capítulo 2 descreve detalhadamente a metodologia utilizada para a análise das séries temporais. Nesta seção, são explicados os métodos estatísticos e os modelos empregados. No Capítulo 3, são

apresentados os resultados da modelagem dos materiais particulados. Esta seção inclui a análise dos dados, a aplicação dos modelos selecionados e a interpretação dos resultados obtidos. Finalmente, no Capítulo 4, são apresentadas as conclusões do estudo. Esta seção sintetiza os principais achados da pesquisa e propõe possíveis direções para investigações subsequentes.

# Capítulo 2

## SÉRIES TEMPORAIS

### 2.1 Noções Básicas

Ao analisar fenômenos que ocorrem ao longo do tempo, é apropriado considerar correlações entre observações. Isso dificulta o uso de métodos estatísticos convencionais, que assumem independência e distribuição idêntica das observações. Em casos onde essas premissas não se aplicam, a abordagem mais apropriada são as técnicas de análise de séries temporais.

Uma série temporal é caracterizada por um conjunto de observações que são sistematicamente geradas em intervalos regulares ao longo de um período. Quando os valores futuros de uma série temporal são precisamente determinados matematicamente por uma função, ela é classificada como uma série temporal determinística. No entanto, na maioria das séries temporais, os valores futuros só podem ser descritos por meio de uma distribuição probabilística, sendo assim denominadas séries temporais não-determinísticas. Nesse cenário, a série pode ser concebida como a trajetória de um processo estocástico.

Seja  $\tau$  um conjunto arbitrário, segundo [Morettin 1999, Morettin e Toloí 2018], um processo estocástico é uma família  $\{X(t), t \in \tau\}$ . Assim, um processo estocástico é uma família de variáveis aleatórias supostamente definidas no mesmo espaço de probabilidades. Para cada  $t \in \tau$ ,  $X(t)$  é uma variável aleatória definida sobre  $\Omega$ . Portanto, na verdade,  $X(t)$  é uma função de dois argumentos:  $X(t, \omega)$ , onde  $t \in \tau$  e  $\omega \in \Omega$ . Assim, para cada  $t \in \tau$ , temos uma variável aleatória  $X(t, \omega)$ , assumindo a existência de uma função de densidade de probabilidades  $f_X(x)$ . Por outro lado, para cada  $\omega \in \Omega$ , fixo, obtemos uma série temporal, ou seja, uma realização do processo.

A coleção de valores  $\{X(t), t \in \tau\}$  é referida como o *espaço de estados*  $S$  de um processo estocástico, onde  $X(t)$  representa os estados do processo. Se o conjunto  $\tau$  for finito ou enumerável, o processo é chamado de *discreto em parâmetro*. Quando  $\tau$  pertence ao conjunto dos números reais, o processo é classificado como *contínuo em*

*parâmetro*. O espaço de estados pode ser discreto ou contínuo. No primeiro caso,  $X(t)$  pode representar uma contagem, como o número de chamadas em uma central telefônica durante um período de duas horas. No segundo caso,  $X(t)$  denota uma medida que varia continuamente, como a temperatura, por exemplo.

Neste trabalho, concentraremos nossa análise em um processo estocástico que possui um parâmetro discreto e um espaço de estados contínuo. Isso implica que a coleção de valores  $X(t), t \in \tau$  será estudada considerando um conjunto  $\tau$  finito (no conjunto dos inteiros), caracterizando o processo como *discreto em parâmetro*. Em outras palavras, estaremos investigando situações onde as observações são coletadas em intervalos distintos ou em pontos específicos no tempo.

Ao mesmo tempo, o espaço de estados  $S$  associado a esse processo será considerado contínuo, o que significa que as variáveis  $X(t)$  assumirão valores em um conjunto contínuo, em oposição a uma contagem discreta. Este contexto é especialmente relevante quando estamos interessados em fenômenos que variam de maneira suave e contínua ao longo do tempo, como, por exemplo, a emissão de poluentes na atmosfera.

Nas situações em que se deseja empregar modelos para a análise de séries temporais, é imperativo introduzir conceitos que permitam a análise de determinadas classes de processos estocásticos. Um desses conceitos é a estacionariedade. Um processo estocástico é considerado estacionário se ele evolui ao longo do tempo de tal maneira que a escolha da origem dos tempos não tem relevância. Em outras palavras, as características de  $X(t - \tau)$ , para todo  $\tau$ , são idênticas às de  $X(t)$ . Teoricamente, existem dois tipos de estacionariedade: a fraca (ou ampla, ou de segunda ordem) e a estrita (ou forte).

Um processo estocástico  $X = \{X(t), t \in \tau\}$  é estritamente estacionário se todas as distribuições unidimensionais são invariantes sob translação no tempo. Por outro lado, um processo estocástico é fracamente estacionário se e somente se:

- $E(X(t)) = \mu(t) = \mu$ , constantes, para todo  $t \in \tau$ ;
- $E\{X^2(t)\} < \infty$ , para todo  $t \in \tau$ ;
- $\gamma(t_1, t_2) = Cov\{X(t_1), X(t_2)\}$  é uma função de  $|t_1 - t_2|$ .

## 2.2 Modelos SARIMA

### Modelos Autorregressivos

Modelos autorregressivos são caracterizados pelo fato de uma variável ser originada a partir de um processo estocástico estacionário, que incorpora os valores anteriores dessa mesma variável. As previsões correspondem a uma combinação linear dos valores

anteriores da variável. A expressão matemática que descreve um processo autorregressivo é a seguinte:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t, \quad (2.1)$$

onde,  $E(\epsilon_t) = 0$ ,  $Var(\epsilon_t) = \sigma^2$ ,  $cov(\epsilon_t \cdot \epsilon_{t-s}) = 0, \forall j \neq s$  e  $\mu$  é um parâmetro determinando o nível da série.

O operador defasagem, denotado por  $B$  e definido como  $BZ_t = Z_{t-1}$ , pode ser utilizado para a reescrita da equação (2.1):

$$\begin{aligned} Y_t &= \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t \\ Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_p Y_{t-p} &= \mu + \epsilon_t \\ Y_t - \phi_1 B Y_t - \phi_2 B^2 Y_t - \dots - \phi_p B^p Y_t &= \mu + \epsilon_t \\ Y_t (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) &= \mu + \epsilon_t \end{aligned}$$

Finalmente, escrevendo de forma compacta:

$$Y_t \phi(B) = \mu + \epsilon_t \quad (2.2)$$

Desse modo, na equação (2.2),  $\phi(B)$  é o operador autorregressivo de ordem  $p$  que representa um polinômio de grau  $p$  em  $B$ .

## Modelos de Médias Móveis

$Y_t$  é gerada por um processo Médias Móveis (MA) se o valor corrente é dado por uma média ponderado dos choques aleatórios correntes e passados. Esses modelos, por sua própria natureza, exibem estacionariedade fraca, uma vez que são formados por uma combinação linear de ruídos brancos ( $\epsilon_t$  tem média zero, variância constante e não-autocorrelacionado).

De forma geral, podemos escrever um modelo de médias móveis da seguinte forma:

$$Y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (2.3)$$

Reescrevendo a equação 2.3 utilizando o operador de defasagem  $B$ :

$$\begin{aligned}
Y_t - \mu &= \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \\
Y_t - \mu &= \epsilon_t - \theta_1 B \epsilon_t - \dots - \theta_q B^q \epsilon_t \\
Y_t - \mu &= \epsilon_t (1 - \theta_1 B - \dots - \theta_q B^q) \\
Y_t - \mu &= \theta(B) \epsilon_t
\end{aligned}$$

Chamando  $Y_t - \mu = \tilde{Y}_t$  temos:

$$\tilde{Y}_t = \theta(B) \epsilon_t,$$

onde  $\theta(B)$  é o operador de médias móveis de ordem  $q$ .

## Modelos ARMA

Em muitas séries temporais observadas na prática, o uso de modelos autorregressivos ou de médias móveis pode apresentar desafios, como a necessidade de lidar com um grande número de parâmetros, entre outros problemas. Nesses casos, a aplicação de modelos autorregressivos de médias móveis (ARMA), que combinam ambos os modelos, pode ser uma alternativa mais adequada.

Assim, temos os modelos ARMA(p,q) que podem ser escritos como:

$$\tilde{Y}_t = \phi_1 \tilde{Y}_{t-1} + \dots + \phi_p \tilde{Y}_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q}. \quad (2.4)$$

Utilizando os operadores autorregressivos e de médias móveis em 2.4 obtemos a forma compacta da equação:  $\phi(B) \tilde{Y}_t = \theta(B) \epsilon_t$

## Modelos ARIMA

Os modelos apresentados anteriormente são apropriados para séries estacionárias. Na prática, muitas séries apresentam comportamento não estacionário. No entanto, muitas dessas séries se tornam estacionárias após serem diferenciadas. Por exemplo, embora  $Z_t$  possa ser não estacionária, a aplicação da diferenciação pode torná-la estacionária.

$$W_t = Z_t - Z_{t-1} = (1 - B)Z_t = \Delta Z_t$$

Desse modo, séries  $Z_t$  que se tornam estacionárias após a aplicação de um número finito de diferenças são chamadas de não estacionárias homogêneas ou integradas de ordem

d. Seja  $Z_t$  uma série não estacionária homogênea. Podemos considerar que  $Z_t$  segue um modelo autorregressivo integrado de médias móveis, ou modelo ARIMA.

A equação 2.5 representa um modelo de ordem  $(p,d,q)$  escrita como ARIMA $(p,d,q)$ , onde  $p$  e  $q$  são as ordens de  $\phi(B)$  e  $\theta(B)$ , respectivamente.

$$\phi(B)\Delta^d Z_t = \theta(B)\epsilon_t \quad (2.5)$$

## Modelos SARIMA

O modelo *SARIMA* (modelo Autorregressivo Integrado de Médias Móveis Sazonal) está entre os métodos mais amplamente utilizados para previsão de séries temporais. O modelo *SARIMA* (*Sazonal ARIMA*) estende o *ARIMA* adicionando uma combinação linear de valores sazonais e/ou erros de previsão.

Seja  $Y_t(t = 1,2,3,\dots)$  um processo linear representado pela equação (2.6):

$$(1 - B)^d(1 - B^S)^D \Phi(B^S)\phi(B)Y_t = \Theta(B^S)\theta(B)\epsilon_t, \quad (2.6)$$

em que  $S$  é o período sazonal do processo,  $B$  é o operador defasagem,  $\epsilon_t$  é o ruído branco (RB) definido como uma sequência de variáveis aleatórias não correlacionadas com média zero e variância constante ao longo do tempo,  $\epsilon_t \sim RB(0, \sigma_\epsilon^2)$ . Além disso,  $d$  e  $D$  são números inteiros não negativos e expressam o número de diferenças simples e sazonais, respectivamente, aplicadas sobre o processo  $Y_t$  [Pinto, Reisen e Monte 2018].

Assim, a equação (2.6) é denominado modelo autorregressivo integrado e de médias móveis sazonal multiplicativo e pode ser representado por (2.7)

$$SARIMA(p, d, q)(P, D, Q)s. \quad (2.7)$$

## 2.3 Modelos ARFIMA

Nos processos *ARMA* $(p, q)$ , o coeficiente de autocorrelação  $\rho_j$  tende a diminuir rapidamente em direção a zero. Essa propriedade caracteriza os processos *ARMA* $(p, q)$  como sendo de “memória curta”. Em contrapartida, um processo que exhibe “memória longa” é aquele em que a função de autocorrelação decai de maneira mais gradual em direção a zero.

A presença do fenômeno de memória longa é identificada em diversas áreas, incluindo estudos de hidrologia, pesquisa climática, análises econômicas e financeiras [Graves et al. 2017]. Para lidar com séries temporais que exibem memória longa, um dos modelos frequentemente empregados é o modelo *ARFIMA* (modelo de Médias Móveis

AutoRegressivo com Integração Fracionária), que representa uma generalização do modelo *ARIMA*.

Defina-se, o operador de diferença fracionária, para todo número real  $d > -1$ , como

$$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k. \quad (2.8)$$

$Y_t$  pode ser caracterizado como um processo *ARFIMA*( $p, d, q$ ), onde  $d$  pertence ao intervalo  $(-\frac{1}{2}, \frac{1}{2})$ . Isso ocorre quando  $Y_t$  é estacionário e atende à seguinte equação:

$$\phi(B)(1 - B)^d Y_t = \theta(B)a_t, \quad (2.9)$$

onde  $a_t$  representa o ruído branco, enquanto  $\phi(B)$  e  $\theta(B)$  são polinômios em  $B$  de graus  $p$  e  $q$ , respectivamente.

A escolha dessa família de processos para a modelagem de séries temporais com comportamento de memória longa é justificada pelo fato de que o efeito do parâmetro  $d$  em observações distantes diminui de forma hiperbólica à medida que a distância aumenta. Em contraste, os efeitos dos parâmetros  $\phi$  e  $\theta$  diminuem exponencialmente. Portanto, a seleção de  $d$  visa explicar a estrutura de correlação de ordens altas na série, enquanto os parâmetros  $\phi$  e  $\theta$  explicam a estrutura de correlação de ordens baixas.

## 2.4 Análise Espetral Singular (SSA)

A Análise Espetral Singular (SSA, do inglês “Singular Spectrum Analysis”) é uma técnica poderosa utilizada para análise de séries temporais e processamento de sinais [Golyandina, Nekrutkin e Zhigljavsky 2001, Golyandina, Korobeynikov e Zhigljavsky 2018, Hassani e Mahmoudvand 2018, Rodrigues e Mahmoudvand 2018]. A SSA envolve a decomposição de uma série temporal em seus componentes, identificação de padrões subjacentes e realização de previsões com base neles. Os passos principais em SSA são os seguintes:

1. Incorporação: Construir uma matriz de trajetórias  $\mathbf{X}$  a partir dos dados da série temporal  $x_1, \dots, x_N$  incorporando a série temporal em uma matriz. Seleciona-se o tamanho da janela  $L$  onde  $(1 < L < N)$ . A matriz de trajetórias  $X$  terá dimensões  $L \times K$ , com  $K = N - L + 1$ .

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_k \\ x_2 & x_3 & \dots & x_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \dots & x_N \end{bmatrix}$$

onde  $x_i$  representa os dados da série temporal e  $N$  é o comprimento da série temporal.

2. Decomposição em Valores Singulares (SVD): Realizar SVD na matriz de trajetórias  $\mathbf{X}$  para a decompor em valores singulares e vetores singulares.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

onde  $\mathbf{U}$  e  $\mathbf{V}$  são matrizes ortogonais contendo os vetores singulares esquerdo e direito, respectivamente, e  $\mathbf{\Sigma}$  é uma matriz diagonal contendo os valores singulares de  $X$ .

3. Agrupamento: Agrupar os valores singulares e vetores singulares correspondentes em componentes com base em seus comportamentos. O passo de agrupamento envolve dividir as matrizes de posto um obtidas no passo anterior em  $m$  subconjuntos disjuntos  $I_1, \dots, I_m$  e consolidar as matrizes dentro de cada grupo. Em nossa aplicação específica, vamos focar no caso onde  $m = 2$ , representando dois grupos. Os conjuntos  $I_1 = 1, \dots, r$  e  $I_2 = r + 1, \dots, L$  são atribuídos aos componentes de sinal e ruído, respectivamente, para ajudar a identificar padrões e tendências nos dados.

A escolha de  $r$ , que determina a quantidade de componentes retidos no conjunto  $I_1$ , pode ser feita com base na análise da contribuição acumulada dos valores singulares para a variância total da série temporal. Tipicamente,  $r$  é escolhido de forma que a soma dos primeiros  $r$  valores singulares capture uma proporção significativa da variância, enquanto os valores singulares restantes são associados ao ruído.

4. Reconstrução: Reconstruir a série temporal original usando os componentes selecionados para capturar os padrões desejados.

$$\hat{\mathbf{X}} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$$

onde  $\mathbf{U}_r$ ,  $\mathbf{\Sigma}_r$ , e  $\mathbf{V}_r$  são as matrizes truncadas contendo os primeiros  $r$  valores e vetores singulares, aqueles associados ao sinal da série temporal.

5. Reconstrução da Série Livre de Ruído: Após a seleção dos componentes associados ao sinal, a série temporal é reconstruída com a remoção do ruído. Para isso, aplica-se

a técnica de diagonalização por médias na matriz de trajetórias reconstruída  $\widehat{\mathbf{X}}$ . Esse processo consiste em construir uma nova matriz de Hankel  $\widetilde{\mathbf{X}}$ , onde cada elemento  $\tilde{x}_{ij}$  é a média dos elementos da matriz reconstruída  $\widehat{\mathbf{X}}$  ao longo das diagonais.

A média ao longo da diagonal de cada elemento  $\tilde{x}_{ij}$  é dada por:

$$\tilde{x}_{ij} = \frac{1}{\#(A_s)} \sum_{(l,k) \in A_s} \widehat{x}_{lk}$$

onde  $A_s = \{(l, k) : l + k = s, 1 \leq l \leq L, 1 \leq k \leq K\}$  representa o conjunto de elementos ao longo da diagonal  $s$ , e  $\#(A_s)$  é o número de elementos em  $A_s$ .

A série temporal reconstruída  $\tilde{y}_i$  é obtida a partir da matriz  $\widetilde{\mathbf{X}}$  como:

$$\tilde{y}_i = \begin{cases} \tilde{x}_{i1} & \text{para } i = 1, \dots, L, \\ \tilde{x}_{Lj} & \text{para } i = L + 1, \dots, N, \end{cases}$$

onde  $j = i - L + 1$ . A série  $\tilde{y}_i$  representa a versão filtrada da série temporal original, reduzindo o impacto do ruído. Essa série reconstruída pode ser utilizada para previsões e outras análises subsequentes.

SSA é especialmente útil para analisar séries temporais com características não lineares e não estacionárias, pois pode separar efetivamente diferentes padrões e tendências nos dados. Tem aplicações em diversos campos, como ciência ambiental, finanças e processamento de sinais, onde compreender padrões subjacentes e fazer previsões precisas são cruciais [Vautard e Ghil 1989, Malamiri et al. 2020, Hassani e Thomakos 2010].

## 2.5 Long Short-Term Memory (LSTM)

As Redes Neurais Recorrentes (RNN, do inglês “Recurrent Neural Network”) constituem uma categoria de redes neurais artificiais onde as conexões entre as unidades formam um ciclo direcionado. Elas representam modelos de redes neurais com potencial substancial em diversas aplicações de aprendizado de máquina [Song et al. 2020, Sahoo et al. 2019]. Entre as variantes de RNNs, as redes neurais recorrentes de Longo Curto Prazo (LSTM, do inglês “Long Short-Term Memory”) se destacam por sua notável capacidade de capturar dependências de longo prazo de forma eficaz.

A Figura 2.1 mostra a estrutura de uma rede LSTM. Nesta estrutura, cada linha transporta um vetor inteiro, desde a saída de um nó até as entradas de outros. Os círculos rosa simbolizam operações específicas, como a soma de vetores, enquanto as caixas amarelas representam camadas de redes neurais que foram treinadas. As linhas que se

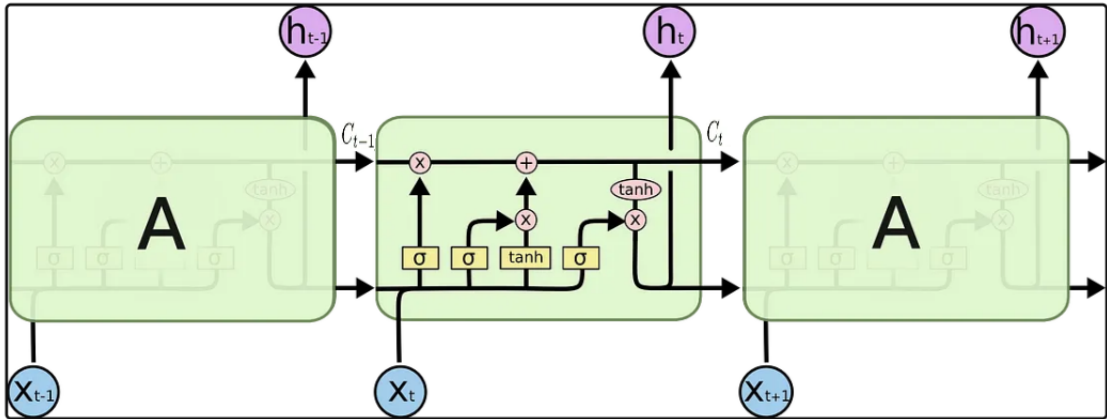


Figura 2.1: Estrutura de uma LSTM. Fonte: [Olah 2015]

juntas indicam uma concatenação, enquanto uma linha que se divide significa que seu conteúdo está sendo replicado, com as cópias enviadas para diferentes destinos.

Uma das principais características das redes LSTM é o estado da célula, representado pela linha horizontal que percorre a parte superior do diagrama na Figura 2.1. Esse estado funciona como uma correia transportadora, atravessando toda a sequência com apenas algumas interações lineares ao longo do caminho, como podemos ver na Figura 2.2. Essas interações irão definir quanto das novas informações serão esquecidas (com  $f_t$  sendo a porta de esquecimento) e quanto delas serão passadas a diante ( $i_t$  determina quanto da nova informação armazenar).

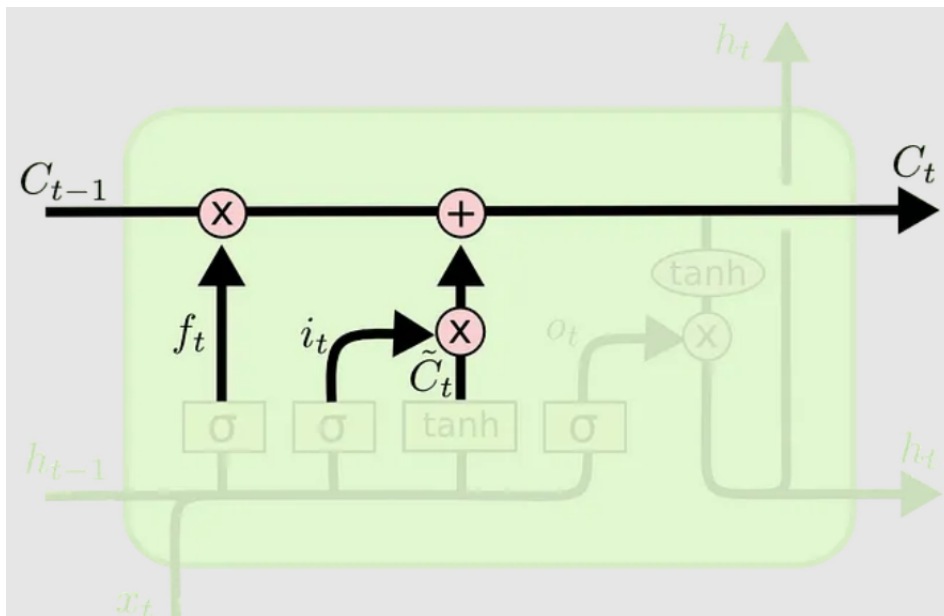


Figura 2.2: Esquema da atualização do estado da célula de uma LSTM. Fonte: [Olah 2015]

Por fim, é necessário determinar o que será gerado como saída. Essa saída será derivada do estado da célula, mas em uma versão filtrada. Primeiro, aplicamos uma

camada sigmoide que define quais partes do estado da célula serão utilizadas. Em seguida, o estado da célula passa por uma função tanh (que comprime os valores entre -1 e 1) e multiplicamos esse resultado pela saída da sigmoide, garantindo que apenas as partes selecionadas sejam produzidas. A Figura 2.3 detalha o processo descrito.

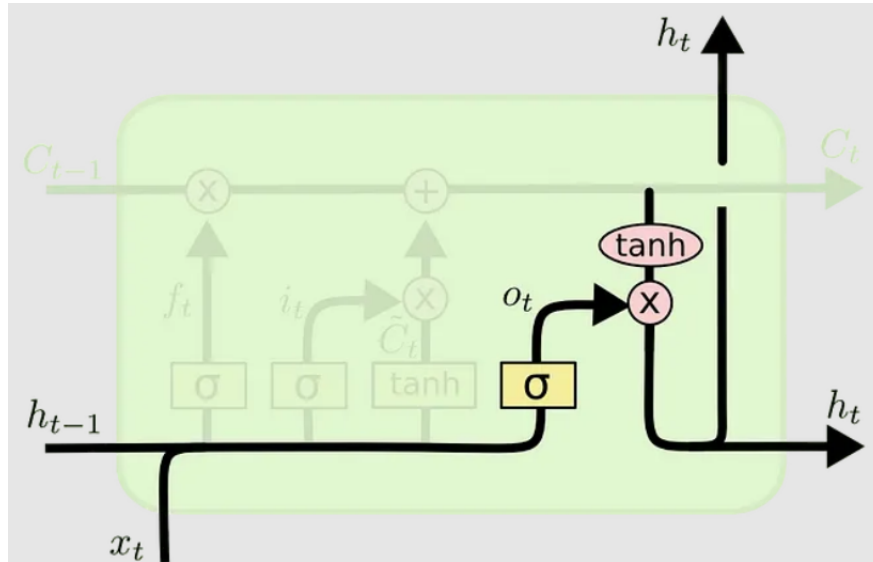


Figura 2.3: Esquema da atualização do estado oculto da célula de uma LSTM.

Fonte: [Olah 2015]

Em outras palavras, dentro de um neurônio (célula) LSTM, uma sequência de entrada  $x_t$  é processada, e cada porta utiliza unidades de ativação para regular seu uso. O parâmetro  $C_t$  denota o estado da célula no tempo  $t$ , encapsulando informações acumuladas até aquele passo temporal. A porta de esquecimento, denotada como  $f_t$ , decide quais informações descartar da célula. A porta de entrada, rotulada como  $i_t$ , supervisiona a atualização do estado da célula com valores de entrada, enquanto a porta de saída, referida como  $O_t$ , determina a saída com base nos dados de entrada e na memória da célula. As saídas dessas portas são então combinadas usando operações de concatenação, multiplicação ou adição. As equações abaixo ilustram a funcionalidade de uma rede LSTM. [Santos e Spancerski 2021].

1. Porta de Entrada ( $i_t$ ): Determina quanto de nova informação armazenar no estado da célula.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

2. Porta de Esquecimento ( $f_t$ ): Controla quanto do estado anterior da célula deve ser esquecido.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

3. Atualização do Estado da Célula ( $\tilde{C}_t$ ): Calcula o novo estado candidato da célula.

$$\tilde{C}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

4. Atualização do Estado da Célula ( $C_t$ ): Atualiza o estado da célula com base nas portas e no estado anterior da célula.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

5. Porta de Saída ( $o_t$ ): Determina o novo estado oculto com base no estado atualizado da célula.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

6. Atualização do Estado Oculto ( $h_t$ ): Calcula o novo estado oculto usando a porta de saída e o estado atualizado da célula.

$$h_t = o_t \cdot \tanh(C_t)$$

Nas equações acima,  $W$  representa as matrizes de peso,  $b$  é o vetor de viés,  $\sigma$  denota a função de ativação sigmoide, e  $\tanh$  é a função tangente hiperbólica. Os mecanismos de portões da rede LSTM permitem regular o fluxo de informações, tornando-a eficaz para lidar com dependências de longo prazo em dados sequenciais.

## 2.6 Modelos Híbridos

Técnicas híbridas na previsão de séries temporais fazem a ponte entre modelos estatísticos clássicos e metodologias de aprendizado de máquina de ponta, harmonizando interpretabilidade e flexibilidade. Ao fundir um espectro de técnicas como SARIMA, SSA e LSTM, esses modelos híbridos buscam elevar a precisão preditiva em uma ampla gama de domínios. Combinando as forças das metodologias estatísticas e de aprendizado de máquina, os modelos híbridos oferecem uma representação abrangente e precisa das dinâmicas dos dados de séries temporais. A introdução do primeiro modelo híbrido por Zhang em 2003, que mesclava ARIMA e redes neurais feedforward, foi um marco significativo [Zhang 2003]. Estudos subsequentes têm explorado uma ampla gama de abordagens híbridas, refletindo a evolução contínua na previsão de séries temporais [Kazienko, Lughofer e Trawinski 2015, López-Gonzales et al. 2023, Iftikhar et al. 2023, Sulandari et al. 2020, Sulandari et al. 2020, Ismail, Shabri e Samsudin 2011, Xu et al. 2022, Silva et al. 2022, Cordova et al. 2021]. Este campo dinâmico continua a testemunhar o desen-

volvimento de métodos e técnicas inovadoras, impulsionados pela busca de modelos mais precisos e robustos para previsão de dados de séries temporais.

Neste estudo, foram desenvolvidos modelos híbrido que combinam as técnicas SARIMA, ARFIMA, SSA e LSTM. A metodologia compreendeu várias etapas: inicialmente, um modelo SARIMA foi empregado para prever as concentrações de  $PM_{10}$  em cada estação. Subsequentemente, a rede neural LSTM foi aplicada aos resíduos do modelo SARIMA. Finalmente, as previsões foram derivadas combinando os resultados de ambos os modelos.

Para o  $PM_{2,5}$  foram propostos dois modelos híbridos. O primeiro modelo híbrido combina os modelos ARFIMA e LSTM. Inicialmente o modelo ARFIMA é utilizado para prever as concentrações de  $PM_{2,5}$  e, em seguida, os resíduos obtidos do modelo ARFIMA são ajustados pelo modelo LSTM. Por fim, realiza-se a soma das previsões do modelo ARFIMA com as previsões do modelo LSTM. O segundo modelo híbrido considera os modelos SSA, ARFIMA e LSTM. Primeiramente, utiliza-se o modelo SSA para decompor as séries temporais nas componentes de sinal e ruído. Em seguida, o sinal é modelado com o modelo ARFIMA e o ruído é modelado pelo modelo LSTM. As previsões são realizadas para ambas as componentes e, posteriormente, somam-se as duas previsões.

A Figura 2.4 exibe diagramas que ilustram as propostas de modelos híbridos aplicados aos dados de  $PM_{10}$  e  $PM_{2,5}$ .

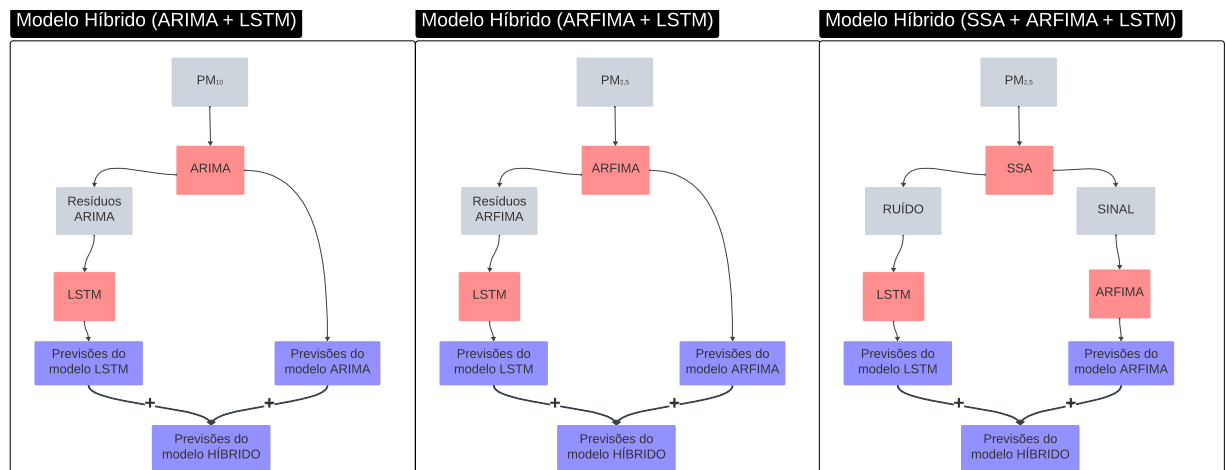


Figura 2.4: Diagrama ilustrativo das abordagens híbridas para os dados de  $PM_{10}$  e  $PM_{2,5}$ .

## 2.7 Validação

Para avaliar o desempenho dos modelos, utilizamos o Root Mean Square Error (RMSE) como métrica de erro, definido pela equação

$$\mathbf{RMSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (2.10)$$

onde  $n$  é o número de previsões realizadas,  $y_i$  representa a observação da série temporal no instante  $i$ , e  $\hat{y}_i$  é o valor previsto pelo modelo para o mesmo instante  $i$ .

Além disso, o erro percentual médio absoluto simétrico (SMAPE) foi empregado como métrica para avaliação do modelo, conforme dado pela equação (2.11). O SMAPE fornece um valor entre 0 e 2, onde valores mais baixos indicam um melhor ajuste do modelo.

$$\mathbf{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{m_t}, \text{ com } m_t = \frac{|\hat{y}_i| + |y_i|}{2}. \quad (2.11)$$

As previsões foram realizadas para uma hora à frente, 24 horas (um dia) à frente e 168 horas (uma semana) à frente, para o  $\text{PM}_{10}$ . Os horizontes de tempo para previsões do  $\text{PM}_{2,5}$  foram de um dia, sete dias (uma semana) e 30 dias (um mês). Para calcular os erros, empregamos o método de validação cruzada. Isso envolveu a segregação das últimas 24 observações para teste enquanto as observações restantes eram utilizadas para treinar os modelos. Posteriormente, utilizando os modelos treinados, geramos previsões para os horizontes temporais mencionados anteriormente. Essa abordagem nos permite comparar o desempenho da previsão de cada modelo em várias séries temporais.

## Capítulo 3

# RESULTADOS E DISCUSSÃO

Os dados utilizados provêm da Fundação Estadual do Meio Ambiente (FEAM), disponível no link: <http://www.feam.br/qualidade-do-ar/dados>. A FEAM oferece informações provenientes do monitoramento contínuo da qualidade do ar para diversos municípios no Estado de Minas Gerais. Cada um desses municípios está equipado com uma ou mais estações de monitoramento.

A definição dos poluentes a serem monitorados em cada estação é estabelecida no contexto do processo de licenciamento ambiental, atualmente determinados após a avaliação do Estudo de Dispersão Atmosférica (EDA). Em geral, os poluentes requisitados correspondem àqueles que possuem critérios de qualidade especificados na Resolução CONAMA nº 491/2018.

Neste trabalho, os poluentes analisados são materiais particulados de tamanho inferior a 10 e 2,5 micrometros ( $PM_{10}$  e  $PM_{2,5}$ ). Os dados de  $PM_{10}$  foram coletados no período de janeiro de 2015 a dezembro de 2019 e correspondem a 42.824 observações horárias do poluente em 29 estações de monitoramento. Já os dados de  $PM_{2,5}$  foram coletados entre janeiro de 2016 e dezembro de 2022. Com um número de estações analisadas reduzido para seis porque a medição de  $PM_{2,5}$  foi realizada em um número menor de estações, e o período analisado também limita as estações ativas na coleta desse poluente. Os dados disponíveis são horários e correspondem a um total de 61.368 observações. No entanto, para a realização das previsões, foi utilizada a média diária de  $PM_{2,5}$ , resultando em uma base de dados com 2.557 observações.

Os dados possuem valores ausentes que foram imputados utilizando a função `na.kalman` do pacote `imputeTS` [Moritz e Bartz-Beielstein 2017] no software R [R Core Team 2024]. A Tabela 3.1 contém informações sobre as estações, incluindo percentual de valores faltantes e localização geográfica.

Tabela 3.1: Informações detalhadas para cada estação de monitoramento, incluindo código, nome da estação, empresa responsável pela estação de monitoramento, longitude, latitude e taxa de valores ausentes.

Nome	Nome da estação	Empresa Responsável	Longitude	Latitude	Valores Faltantes PM <sub>10</sub> (%)	Valores Faltantes PM <sub>2.5</sub> (%)
Betim1	Alterosa	Refinaria Gabriel Passos	44d09'39.06" W	19d56'51.18" S	16,73	11,79
Betim2	Centro Administrativo Betim	Refinaria Gabriel Passos	44d12'26.00" W	19d58'07.98" S	8,77	15,29
Betim3	Petrovale	Refinaria Gabriel Passos	44d06'40.05" W	19d59'39.93" S	12,16	22,95
BH1	Centro Av. do Contorno	Refinaria Gabriel Passos	43d56'08.87" W	19d54'48.02" S	4,38	11,18
BH2	Delegacia Amazonas	V&M do Brasil S.A	43d59'41.50" W	19d56'14.90" S	22,50	-
Brumadinho2	Comunidade do Feijão	Vale S.A	44d06'32.19" W	20d08'09.13" S	24,92	-
C. Fabriciano	Senac - (Vale do Aço)	Aperam Inox S.A	42d37'38.00" W	19d31'49.00" S	11,19	-
Contagem	Cidade Industrial	V&M do Brasil S.A	44d01'54.27" W	19d57'38.07" S	27,09	-
Ibirité1	Cascata	Refinaria Gabriel Passos	44d05'10.73" W	19d59'15.31" S	10,19	20,91
Ibirité2	Piratininga (Ibitermo)	Refinaria Gabriel Passos	44d03'34.89" W	20d00'14.27" S	4,71	11,69
Ipatinga1	Bom Retiro	Usiminas	42d33'25.28" W	19d30'34.55" S	8,77	-
Ipatinga2	Caruru	Usiminas	42d31'43.46" W	19d29'28.92" S	9,29	-
Ipatinga3	Cidade Nobre	Usiminas	42d33'36.74" W	19d27'40.22" S	7,45	-
Itabira2	Félix (EAMA 31)	Vale S.A	43d14'14.68" W	19d39'15.00" S	19,09	-
Itabira3	Major Lage (EAMA 21)	Vale S.A	43d14'13.02" W	19d38'08.09" S	5,60	-
Itabira4	Panorama (EAMA 41)	Vale S.A	43d13'19.10" W	19d38'04.10" S	8,85	-
Itabira5	Pará (EAMA 11)	Vale S.A	43d13'50.85" W	19d37'10.00" S	19,24	-
Paracatu1	Clube da União	Kinross Gold Corporation	46d53'10.86" W	17d12'56.88" S	22,63	-
Paracatu2	Copasa	Kinross Gold Corporation	46d52'27.36" W	17d12'19.74" S	24,67	-
Paracatu3	Lagoa Trindade Rodrigues	Kinross Gold Corporation	46d49'58.50" W	17d09'01.47" S	16,42	-
Paracatu4	São Domingos	Kinross Gold Corporation	46d52'27.36" W	17d11'53.82" S	17,46	-
Paracatu5	Sérgio Ulhoa	Kinross Gold Corporation	46d52'30.60" W	17d13'28.44" S	12,62	-
S.J.daLapa1	Célia (Vespasiano)	Belocal	43d58'07.83" W	19d42'45.01" S	20,41	-
S.J.daLapa2	Centro (Delegacia)	Belocal	43d57'36.70" W	19d42'00.54" S	17,03	-
S.J.daLapa3	Jardim Encantado	ICAL	43d56'17.33" W	19d41'49.21" S	21,14	-
S.J.daLapa4	Escola Municipal Filinha Gama	ICAL	43d57'50.73" W	19d42'47.45" S	33,71	-
Timóteo1	Cecília Meireles	Aperam Inox S.A	42d39'23.00" W	19d32'47.00" S	25,33	-
Timóteo2	Escola Sementinha	Aperam Inox S.A	42d40'16.00" W	19d32'12.00" S	9,08	-
Timóteo3	Hospital Vital Brazil	Aperam Inox S.A	42d38'33.00" W	19d32'59.00" S	10,17	-

### 3.1 Ajuste e Previsão de Séries Temporais em dados de PM<sub>10</sub>

#### 3.1.1 Análise Descritiva

A Figura 3.1 ilustra a distribuição geográfica desses municípios, destacando uma concentração de estações em regiões com maior densidade populacional [Silva et al. 2023].

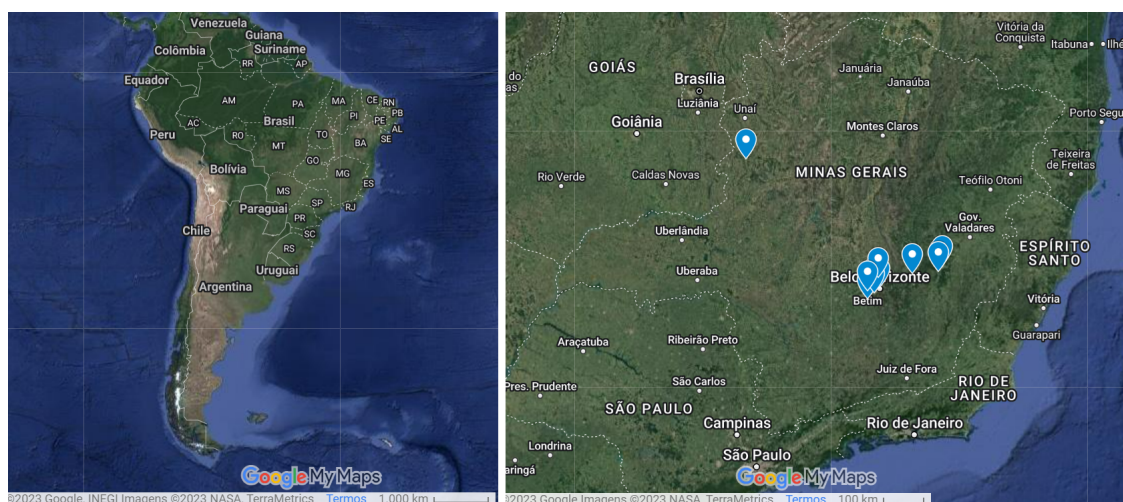


Figura 3.1: O painel esquerdo mostra um mapa da América do Sul, enquanto o painel direito exhibe o mapa que inclui o estado brasileiro de Minas Gerais com marcadores indicando a localização das estações de monitoramento.

Foi feita, inicialmente, a análise exploratória dos dados, onde podemos visualizar

gráficos a cerca do comportamento das concentrações de  $PM_{10}$ .

A Figura 3.2 apresenta os valores médios semanais de concentração de  $PM_{10}$  em todas as 29 estações de monitoramento após imputação dos valores faltantes. Foram utilizadas as concentrações médias pois facilita a visualização da série.

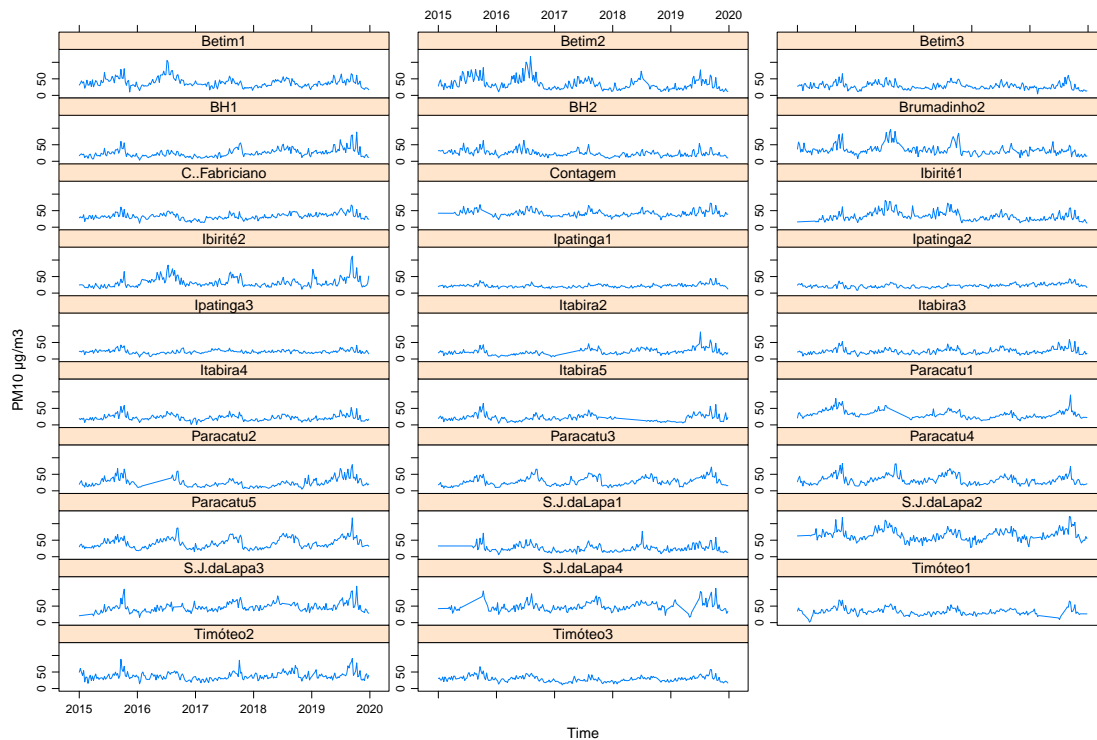


Figura 3.2: Séries temporais das concentrações médias semanais de  $PM_{10}$  para todas as 29 estações de monitoramento.

Nas Figuras 3.3-3.6 são apresentados os boxplots para os dados de  $PM_{10}$ . Na Figura 3.3, as observações são agrupadas considerando as horas do dia. Na Figura 3.4, considera-se os dias; na Figura 3.5, os meses; e, por fim, na Figura 3.6, os anos. Dessa forma, podemos analisar as variações sazonais considerando as horas, os dias, os meses e os anos. Essas análises são fundamentais para entender os padrões de poluição do ar e suas possíveis causas.

A análise dos gráficos de boxplot revela padrões gerais nas concentrações de  $PM_{10}$  ao longo do tempo, com variações sazonais observáveis. No entanto, é importante destacar que esses comportamentos podem variar entre as 29 estações de monitoramento. Mesmo assim, algumas análises gerais são apresentadas a seguir. Nos gráficos de concentrações horárias, observa-se um aumento nas concentrações de  $PM_{10}$  no período da manhã e no fim da tarde, indicando picos de poluição nesses horários. Já nos gráficos diários, a sazonalidade é menos pronunciada, mas alguns padrões de variação ao longo da semana podem ser observados, sem uma tendência acentuada. Nos gráficos mensais, a maioria das

estações apresenta um aumento nas concentrações entre julho e dezembro, possivelmente associado a fatores climáticos ou a outras atividades que se intensificam nesse período. Nos gráficos anuais, não parece haver sazonalidade ao longo dos anos, com as concentrações apresentando variações mais homogêneas. Vale frisar que essas observações são gerais e podem não refletir o comportamento específico de todas as 29 estações, visto que diferenças locais, como a topografia, atividades econômicas e fontes de poluição, podem influenciar a dinâmica de cada estação de monitoramento. Para verificar se essas tendências são aplicáveis a uma estação específica, é recomendável consultar o boxplot individual de cada estação.

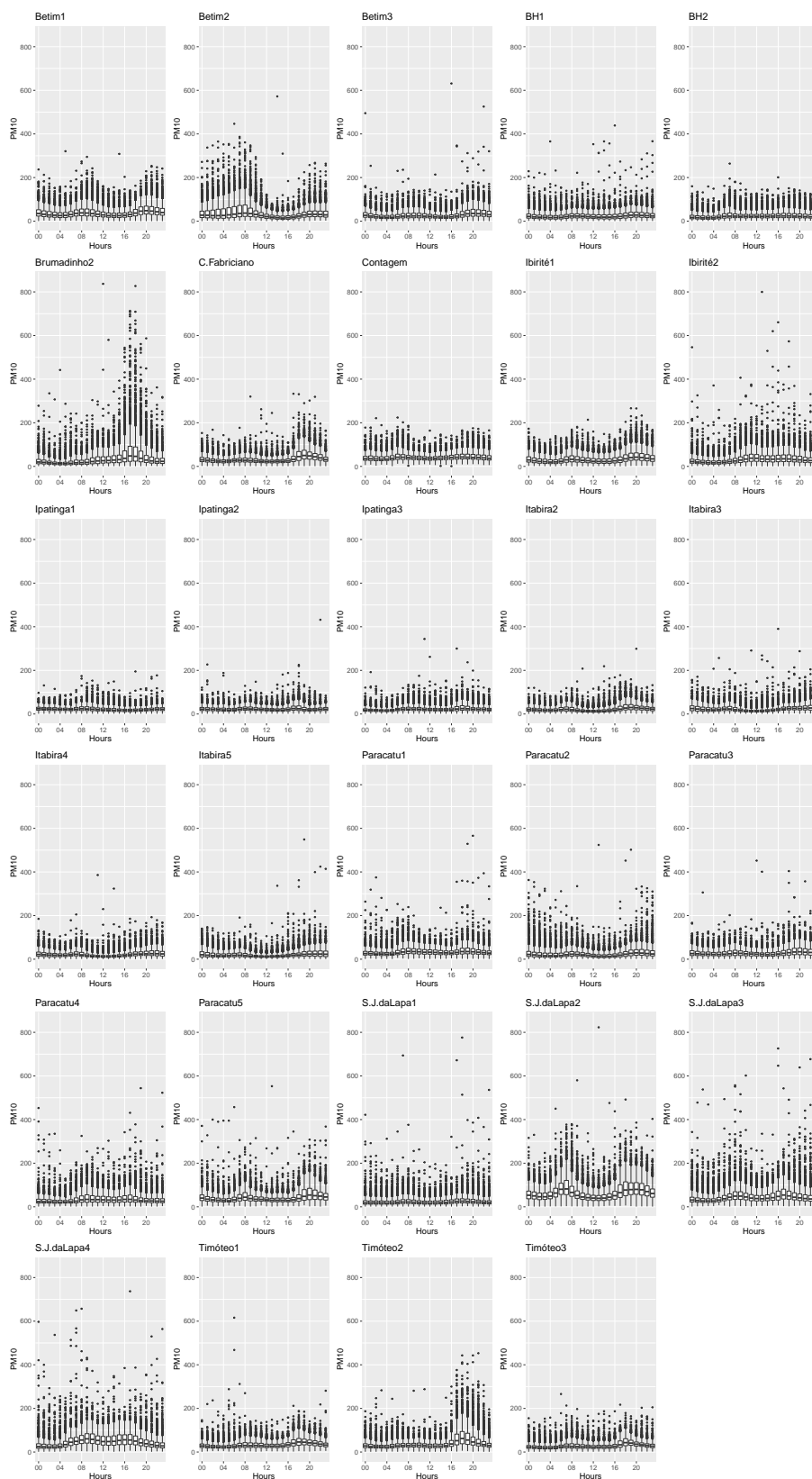


Figura 3.3: Boxplots das concentrações de  $PM_{10}$  (em  $\mu g/m^3$ ) por hora do dia para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de  $PM_{10}$  observadas entre meia-noite e 1h).

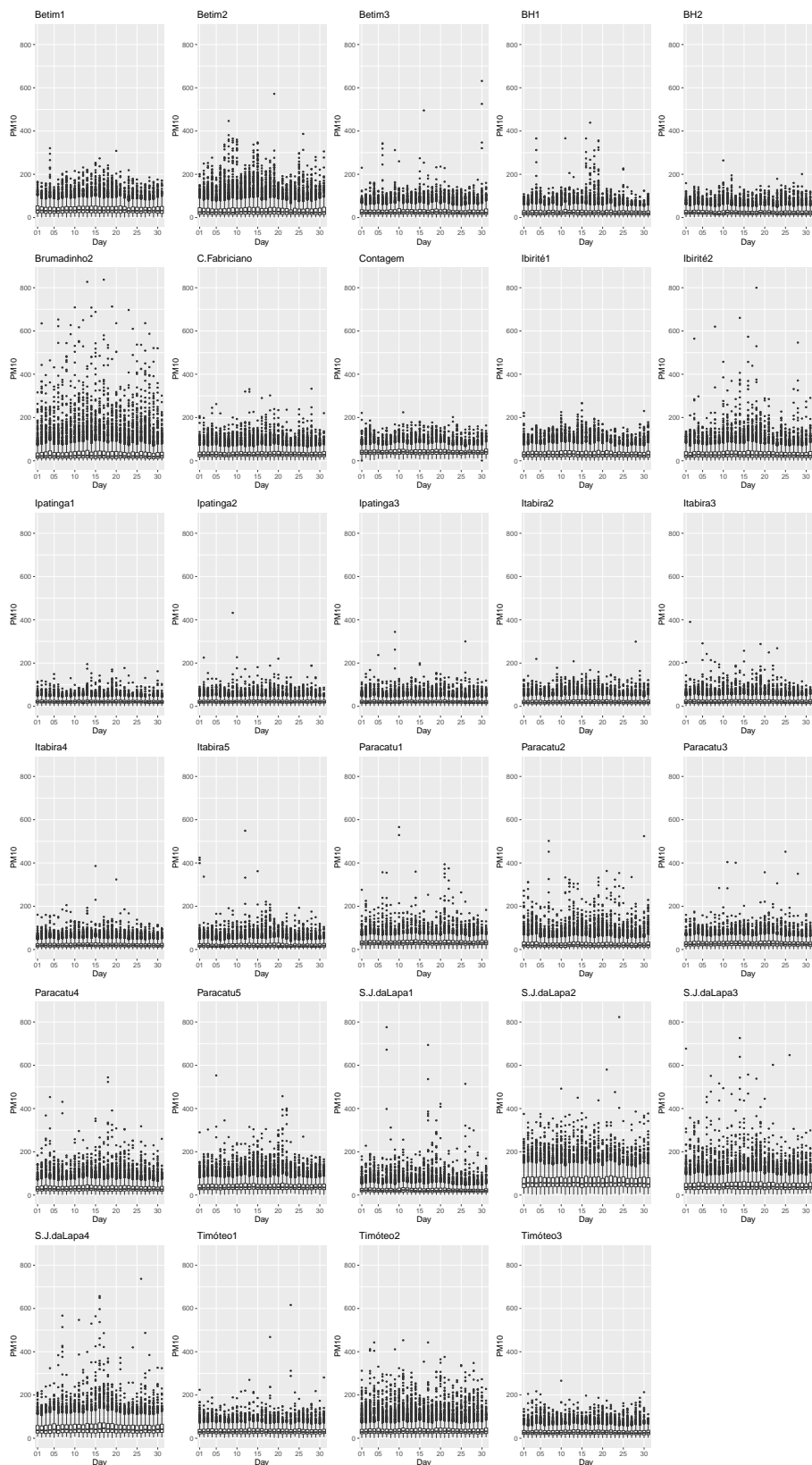


Figura 3.4: Boxplots das concentrações de  $PM_{10}$  (em  $\mu g/m^3$ ) por dia do mês para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de  $PM_{10}$  observadas no primeiro dia do mês).

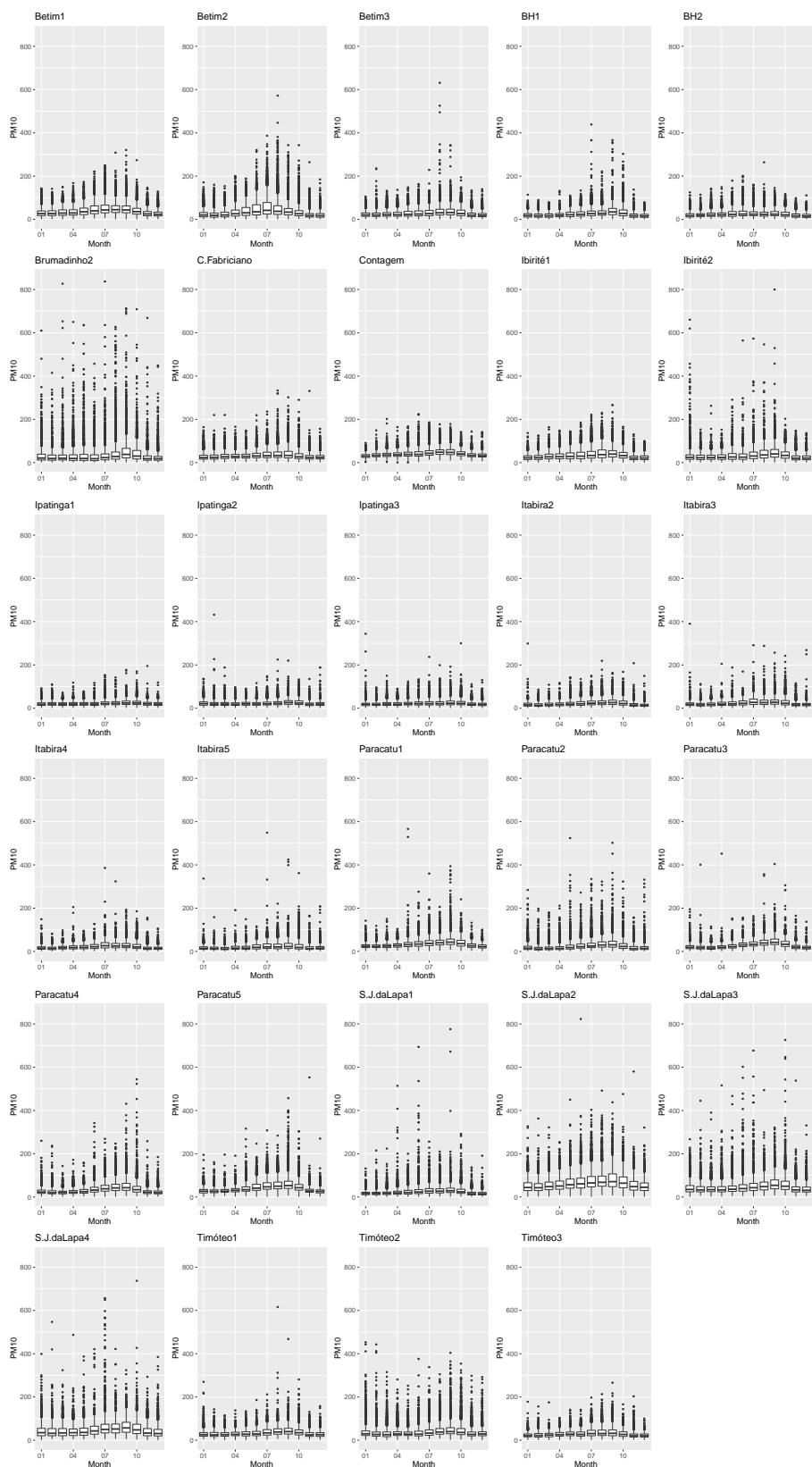


Figura 3.5: Boxplots das concentrações de  $PM_{10}$  (em  $\mu g/m^3$ ) por mês do ano para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de  $PM_{10}$  observadas em janeiro).

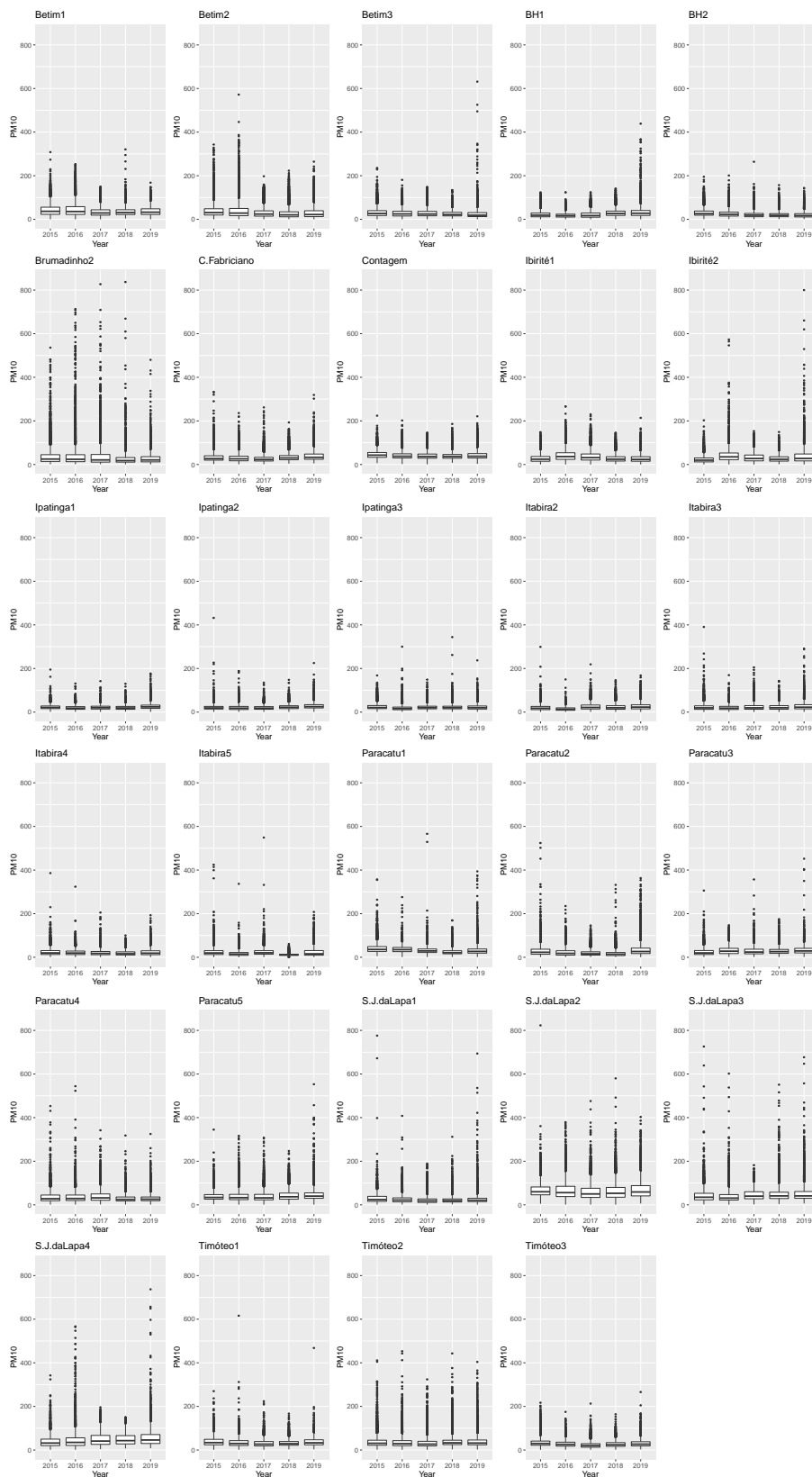


Figura 3.6: Boxplots das concentrações de  $PM_{10}$  (em  $\mu g/m^3$ ) por ano para todas as 29 estações de monitoramento (por exemplo, o primeiro boxplot em cada gráfico é baseado em todas as concentrações de  $PM_{10}$  observadas em 2015).

### 3.1.2 Ajuste dos Modelos para PM<sub>10</sub>

Foram testados quatro modelos para avaliar a concentração de PM<sub>10</sub>: ARIMA, LSTM, SSA e um modelo híbrido proposto, que combina ARIMA e LSTM. O ajuste dos modelos foi avaliado com base nos resultados apresentados na Tabela 3.2.

Utilizamos as métricas Erro Quadrático Médio (RMSE) e Erro Percentual Médio Absoluto Simétrico (SMAPE) para avaliar a precisão dos ajustes realizados pelos modelos. A Tabela 3.2 apresenta os valores dessas métricas para cada modelo avaliado. De maneira geral, os resultados indicam que o modelo SARIMA oferece o melhor ajuste dentre os modelos testados. Este desempenho superior do SARIMA reflete-se em menores valores de RMSE e SMAPE, evidenciando sua capacidade de capturar melhor as características dos dados.

Tabela 3.2: Medidas de acurácia (RMSE e SMAPE) para avaliação do ajuste dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM.

	SARIMA		LSTM		SSA		Híbrido	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Betim1	13,88	0,27	15,75	0,33	22,25	0,43	13,88	0,27
Betim2	12,92	0,29	13,70	0,31	28,34	0,55	12,93	0,30
Betim3	11,60	0,28	12,20	0,29	18,59	0,44	11,58	0,28
BH1	8,63	0,26	9,87	0,30	17,85	0,48	8,79	0,27
BH2	8,69	0,22	9,53	0,25	14,89	0,42	8,68	0,22
Brumadinho2	29,40	0,39	30,44	0,37	38,33	0,62	29,43	0,39
C..Fabriciano	13,53	0,26	14,59	0,30	18,75	0,39	13,50	0,26
Contagem	10,50	0,15	11,16	0,17	15,45	0,25	10,48	0,15
Ibirite1	9,96	0,25	11,96	0,32	20,19	0,55	9,92	0,26
Ibirite2	14,15	0,30	15,14	0,31	22,76	0,47	14,23	0,30
Ipatinga1	7,68	0,25	7,91	0,27	10,72	0,37	7,68	0,25
Ipatinga2	9,45	0,28	10,97	0,32	11,98	0,37	9,81	0,29
Ipatinga3	9,18	0,26	10,39	0,30	11,95	0,37	9,16	0,26
Itabira2	7,52	0,20	9,13	0,26	12,95	0,42	7,55	0,22
Itabira3	8,74	0,23	9,46	0,26	13,32	0,39	8,90	0,23
Itabira4	8,35	0,24	9,82	0,35	13,58	0,44	8,41	0,25
Itabira5	8,42	0,20	8,74	0,22	14,56	0,40	8,41	0,20
Paracatu1	12,00	0,21	18,52	0,42	15,71	0,29	12,04	0,21
Paracatu2	11,50	0,24	12,17	0,26	19,30	0,45	11,66	0,28
Paracatu3	11,35	0,25	17,05	0,51	14,71	0,37	11,50	0,26
Paracatu4	15,85	0,26	22,06	0,56	20,15	0,38	16,29	0,28
Paracatu5	15,73	0,22	19,91	0,30	21,17	0,33	15,91	0,23
S.J.daLapa1	12,70	0,22	17,92	0,69	18,36	0,39	12,86	0,24
S.J.daLapa2	26,78	0,26	33,78	0,38	33,26	0,37	26,78	0,26
S.J.daLapa3	23,81	0,26	31,64	0,60	29,03	0,39	24,64	0,30
S.J.daLapa4	23,02	0,25	24,89	0,36	28,95	0,38	22,99	0,25
Timoteo1	11,43	0,19	14,72	0,33	15,55	0,34	11,92	0,24
Timoteo2	19,00	0,27	20,82	0,33	26,12	0,41	18,99	0,28
Timoteo3	10,61	0,24	12,11	0,28	14,34	0,36	10,55	0,24

### 3.1.3 Previsão de $PM_{10}$

As previsões foram feitas para três horizontes de tempo distintos: 1 hora à frente, 24 horas (1 dia) à frente e 168 horas (1 semana) à frente. As métricas RMSE e SMAPE foram calculadas utilizando validação cruzada, com as 24 observações mais recentes reservadas para teste, e o restante da série foi utilizado para treinar os modelos. Esse processo foi repetido 24 vezes, e as métricas foram computadas comparando os valores previstos e reais.

O desempenho dos modelos analisados variou conforme os diferentes horizontes de previsão, como mostrado nas Tabelas 3.3-3.5. Essas tabelas apresentam as medidas de precisão (RMSE e SMAPE) para avaliar as previsões feitas uma hora adiante, 24 horas (um dia) adiante e 168 horas (uma semana) adiante, respectivamente, para os modelos SARIMA, LSTM, SSA e híbrido (SARIMA e LSTM). A Figura 3.7 exibe os boxplots de RMSE e SMAPE, respectivamente, demonstrando o ajuste do modelo e o desempenho da previsão nas 29 estações de monitoramento para os quatro modelos: SARIMA, LSTM, SSA e o híbrido (SARIMA e LSTM).

Para todos os três horizontes de previsão, os modelos SARIMA e híbrido surgem como os de melhor desempenho, com o modelo híbrido apresentando um desempenho geral ligeiramente melhor. Para previsões de uma hora e um dia adiante, o LSTM superou seus concorrentes para várias estações de monitoramento.

Tabela 3.3: Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão uma hora à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM.

	SARIMA		LSTM		SSA		Híbrido	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Betim1	14,87	0,35	6,16	0,32	29,26	0,76	15,11	0,35
Betim2	4,73	0,33	4,02	0,30	8,89	0,64	4,78	0,34
Betim3	5,98	0,30	4,43	0,24	9,71	0,59	5,87	0,29
BH1	5,74	0,36	7,01	0,51	18,79	0,84	6,16	0,40
BH2	7,56	0,32	2,73	0,27	8,78	0,43	7,58	0,32
Brumadinho2	3,84	0,27	8,32	0,47	9,20	0,88	3,37	0,26
C..Fabriciano	5,39	0,14	12,44	0,43	11,81	0,25	6,27	0,17
Contagem	0,30	0,01	3,02	0,09	13,85	0,36	0,23	0,01
Ibirite1	9,85	0,44	3,71	0,29	22,72	0,68	9,85	0,44
Ibirite2	22,07	0,64	14,12	0,23	50,39	1,12	22,33	0,65
Ipatinga1	5,96	0,23	6,15	0,46	11,02	0,41	5,94	0,23
Ipatinga2	15,09	0,26	9,23	0,36	13,15	0,23	15,70	0,30
Ipatinga3	12,89	0,25	6,28	0,30	19,95	0,39	12,95	0,25
Itabira2	4,01	0,22	5,67	0,26	8,51	0,35	4,02	0,24
Itabira3	2,83	0,15	3,06	0,14	7,44	0,36	2,41	0,13
Itabira4	4,15	0,29	12,56	0,45	7,72	0,55	4,08	0,26
Itabira5	1,51	0,07	9,13	0,25	6,58	0,25	1,14	0,06
Paracatu1	10,21	0,27	8,55	0,37	17,73	0,55	10,55	0,28
Paracatu2	7,65	0,37	5,99	0,22	30,03	1,04	8,01	0,44
Paracatu3	5,36	0,45	9,90	0,59	15,85	0,95	5,19	0,37
Paracatu4	6,65	0,40	11,83	0,62	10,82	0,54	7,46	0,48
Paracatu5	11,82	0,22	6,29	0,20	17,47	0,35	12,87	0,24
S.J.daLapa1	19,89	0,16	8,27	0,72	10,23	0,52	3,49	0,18
S.J.daLapa2	12,15	0,39	16,49	0,30	28,70	0,56	20,37	0,39
S.J.daLapa3	12,15	0,28	13,50	0,51	16,44	0,40	13,48	0,29
S.J.daLapa4	14,82	0,41	14,47	0,36	18,28	0,44	14,90	0,41
Timoteo1	0,01	~0,00	6,56	0,28	8,12	0,33	3,28	0,13
Timoteo2	11,18	0,21	19,87	0,41	21,35	0,41	10,75	0,21
Timoteo3	11,30	0,31	11,17	0,41	11,48	0,33	11,49	0,32

Tabela 3.4: Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão 24 horas (um dia) à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM.

	SARIMA		LSTM		SSA		Híbrido	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Betim1	36,02	0,62	10,25	0,46	35,43	0,72	36,46	0,62
Betim2	6,80	0,26	4,79	0,41	9,40	0,63	6,88	0,26
Betim3	7,65	0,40	13,42	0,67	11,54	0,60	7,62	0,40
BH1	7,39	0,58	15,86	0,86	22,04	1,09	6,37	0,52
BH2	7,65	0,62	8,84	0,63	6,41	0,54	7,51	0,61
Brumadinho2	10,79	0,42	11,39	0,64	18,13	0,78	10,57	0,35
C..Fabriciano	5,87	0,17	14,78	0,46	13,15	0,31	5,84	0,16
Contagem	0,78	0,02	5,75	0,17	15,30	0,39	0,80	0,02
Ibirite1	28,39	0,66	8,98	0,56	33,01	0,73	28,40	0,66
Ibirite2	43,28	0,79	32,87	0,67	34,71	1,10	42,24	0,80
Ipatinga1	7,83	0,35	11,25	0,69	10,58	0,45	7,98	0,35
Ipatinga2	11,02	0,33	11,09	0,54	11,80	0,37	11,68	0,32
Ipatinga3	18,08	0,41	10,09	0,52	20,83	0,53	18,20	0,41
Itabira2	7,54	0,54	8,81	0,33	9,73	0,53	7,12	0,52
Itabira3	9,87	0,60	3,54	0,16	7,25	0,46	8,90	0,56
Itabira4	8,77	0,70	17,90	0,50	5,77	0,46	8,34	0,67
Itabira5	8,74	0,39	16,88	0,45	7,94	0,33	8,39	0,38
Paracatu1	10,55	0,27	6,80	0,28	13,88	0,41	11,31	0,31
Paracatu2	14,63	0,44	18,23	0,64	25,40	0,88	15,21	0,44
Paracatu3	6,62	0,41	14,97	0,71	12,38	0,66	7,23	0,43
Paracatu4	8,84	0,36	12,06	0,43	8,42	0,33	11,96	0,57
Paracatu5	9,65	0,18	7,77	0,24	12,92	0,27	11,25	0,24
S.J.daLapa1	6,08	0,39	13,67	0,70	11,79	0,63	5,72	0,39
S.J.daLapa2	24,18	0,38	12,21	0,21	21,77	0,27	24,63	0,38
S.J.daLapa3	10,54	0,31	18,53	0,49	18,04	0,49	8,79	0,22
S.J.daLapa4	17,68	0,57	30,30	0,63	19,21	0,54	18,01	0,58
Timoteo1	0,02	~0,00	6,51	0,25	8,25	0,33	3,28	0,13
Timoteo2	12,92	0,23	21,10	0,36	15,21	0,31	12,67	0,23
Timoteo3	8,14	0,27	10,84	0,52	7,52	0,26	8,32	0,28

Tabela 3.5: Medidas de acurácia (RMSE e SMAPE) para avaliação da previsão 168 horas (uma semana) à frente dos modelos SARIMA, LSTM, SSA e do híbrido que considera SARIMA e LSTM.

	SARIMA		LSTM		SSA		Híbrido	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Betim1	6,01	0,30	16,04	0,66	13,70	0,56	5,68	0,29
Betim2	5,60	0,43	6,01	0,45	7,19	0,59	6,01	0,45
Betim3	4,69	0,25	16,51	0,77	7,98	0,44	4,63	0,24
BH1	7,47	0,53	19,26	0,94	20,39	0,97	6,97	0,50
BH2	6,76	0,53	12,10	0,77	4,97	0,41	6,65	0,52
Brumadinho2	9,41	0,53	10,00	0,55	14,71	0,66	9,75	0,56
C..Fabriciano	14,65	0,46	17,45	0,58	17,67	0,54	14,75	0,46
Contagem	0,78	0,02	5,43	0,16	13,85	0,36	0,77	0,02
Ibirite1	9,82	0,59	13,67	0,77	5,24	0,40	9,79	0,59
Ibirite2	47,98	0,89	34,03	0,68	45,33	0,94	48,13	0,92
Ipatinga1	10,27	0,66	12,46	0,75	17,82	0,92	10,81	0,68
Ipatinga2	11,63	0,55	11,48	0,54	16,30	0,71	10,40	0,48
Ipatinga3	16,77	0,74	9,75	0,51	7,68	0,43	16,51	0,73
Itabira2	9,19	0,39	8,68	0,42	11,06	0,49	9,46	0,39
Itabira3	3,64	0,16	3,07	0,14	13,15	0,75	5,28	0,25
Itabira4	20,04	0,56	19,58	0,65	17,75	0,43	20,29	0,54
Itabira5	16,70	0,66	17,73	0,47	13,91	0,55	16,60	0,65
Paracatu1	7,29	0,30	7,41	0,27	15,66	0,51	7,72	0,31
Paracatu2	11,41	0,45	19,16	0,66	25,92	0,77	12,56	0,52
Paracatu3	9,57	0,43	16,35	0,74	11,60	0,57	10,56	0,49
Paracatu4	9,94	0,42	14,15	0,49	8,16	0,33	12,60	0,61
Paracatu5	3,33	0,07	4,81	0,13	7,54	0,22	4,52	0,14
S.J.daLapa1	6,62	0,42	13,48	0,70	8,32	0,51	6,42	0,37
S.J.daLapa2	17,40	0,31	16,91	0,31	24,63	0,43	19,78	0,36
S.J.daLapa3	8,99	0,26	20,87	0,54	13,38	0,35	7,06	0,19
S.J.daLapa4	5,22	0,14	36,41	0,71	9,73	0,24	5,42	0,15
Timoteo1	0,06	~0,00	2,49	0,09	8,92	0,37	3,35	0,13
Timoteo2	29,10	0,44	27,88	0,53	20,61	0,43	28,48	0,46
Timoteo3	13,53	0,62	13,36	0,60	15,00	0,67	13,47	0,60

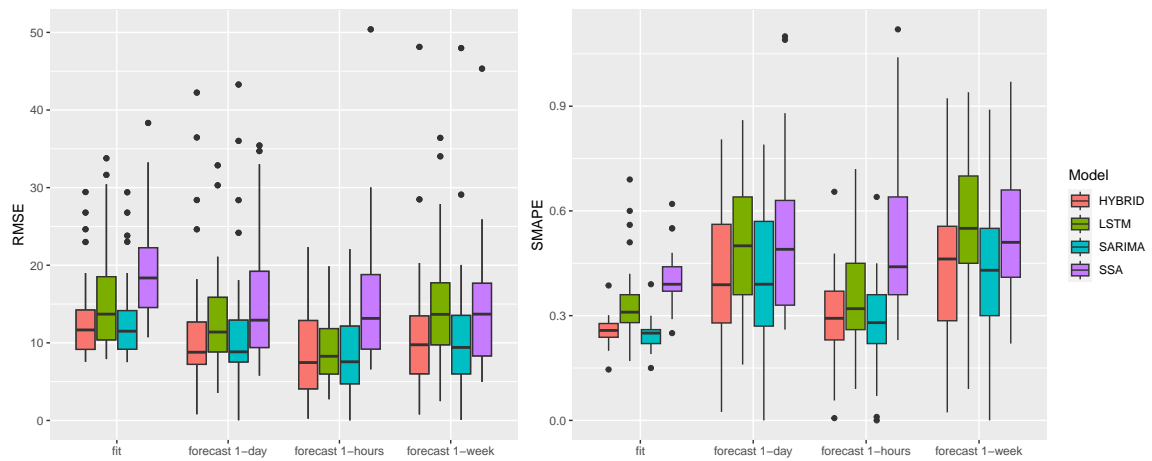


Figura 3.7: Boxplots do RMSE e SMAPE para o ajuste e a previsão dos modelos, considerando as 29 estações de monitoramento. Os quatro modelos analisados são: SARIMA, LSTM, SSA e o modelo híbrido que combina SARIMA e LSTM.

## 3.2 Modelos Híbridos para Séries com Memória Longa em dados de $PM_{2,5}$

### 3.2.1 Análise Descritiva

A Figura 3.8 apresenta as concentrações de  $PM_{2,5}$  nas seis estações de monitoramento consideradas após imputação dos valores faltantes.

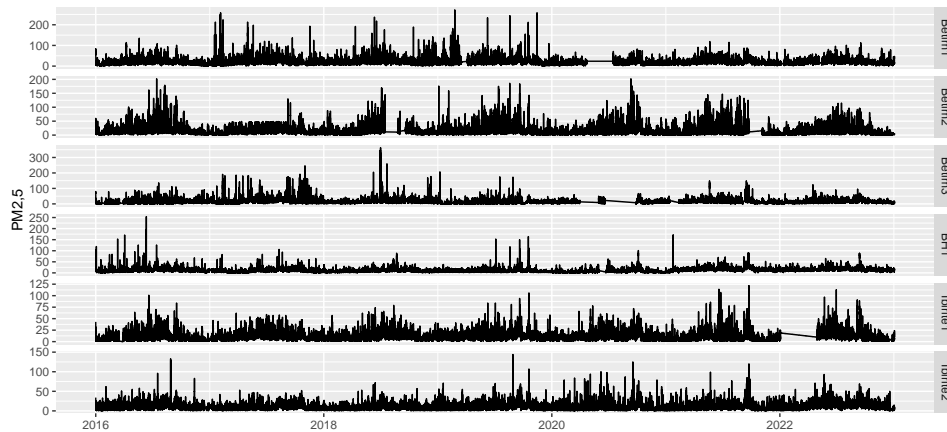


Figura 3.8: Séries temporais das concentrações de  $PM_{2,5}$  nas seis estações de monitoramento consideradas após imputação de dados ausentes.

Nas Figuras 3.9-3.12 são apresentados os boxplots das concentrações de  $PM_{2,5}$ . Na Figura 3.9, as medições estão agrupadas de acordo com as horas do dia. A Figura 3.10 ilustra as observações organizadas por dias. Já na Figura 3.11, as medições são agrupadas por meses e na Figura 3.12, os dados estão organizados por anos. Assim, podemos examinar as variações sazonais nas concentrações de  $PM_{2,5}$  considerando diferentes escalas temporais: horas, dias, meses e anos. Essa abordagem nos permite identificar padrões e tendências ao longo do tempo, proporcionando uma visão detalhada das flutuações sazonais do poluente.

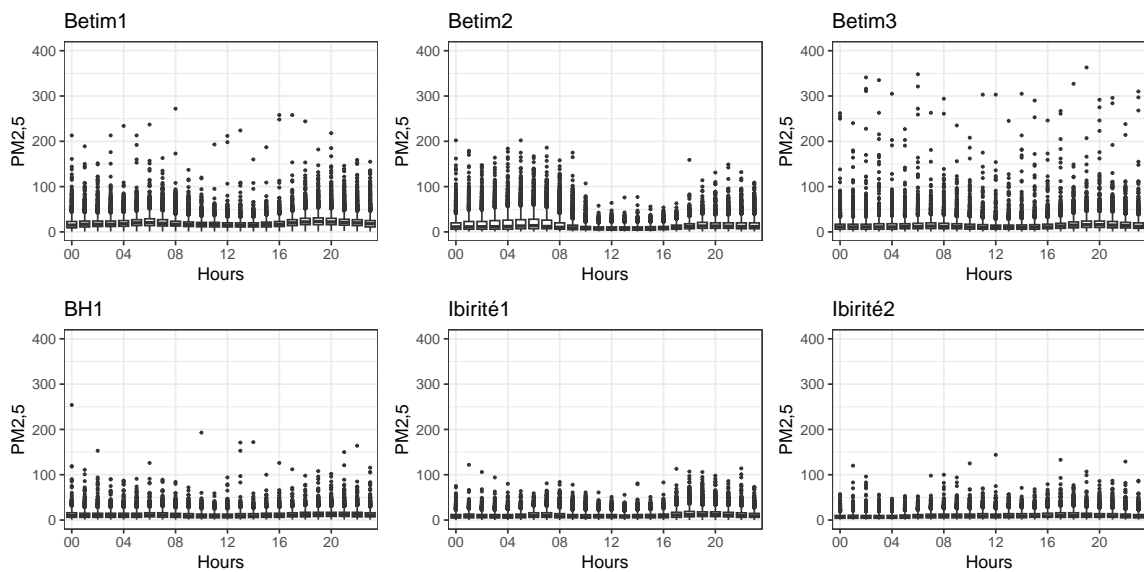


Figura 3.9: Boxplots da concentração de PM<sub>2,5</sub> (em  $\mu m^2$ ) por hora em cada estação.

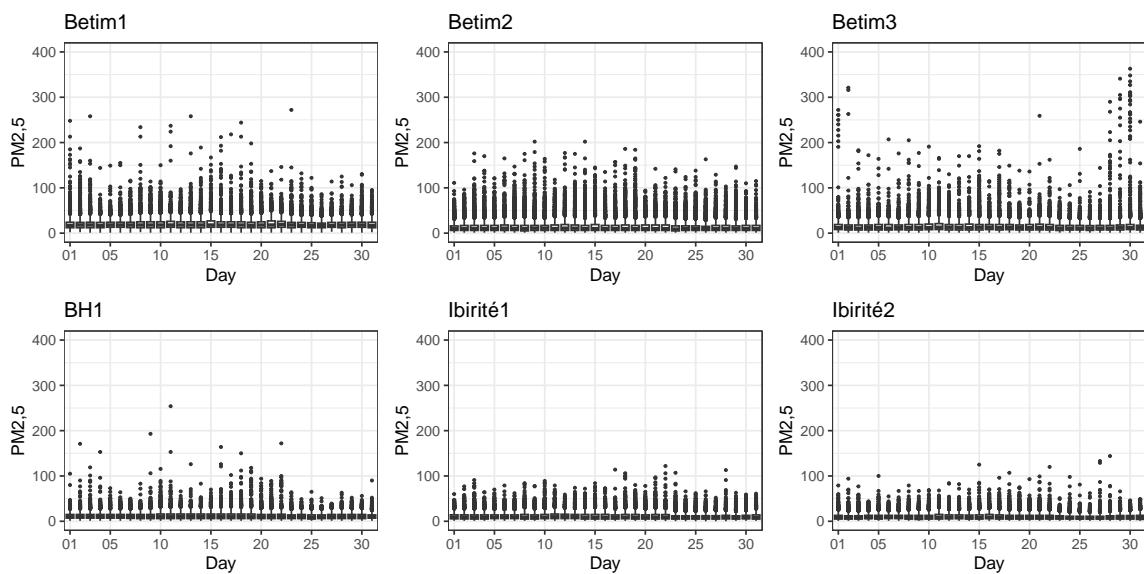


Figura 3.10: Boxplots da concentração de PM<sub>2,5</sub> (em  $\mu m^2$ ) por dia em cada estação.

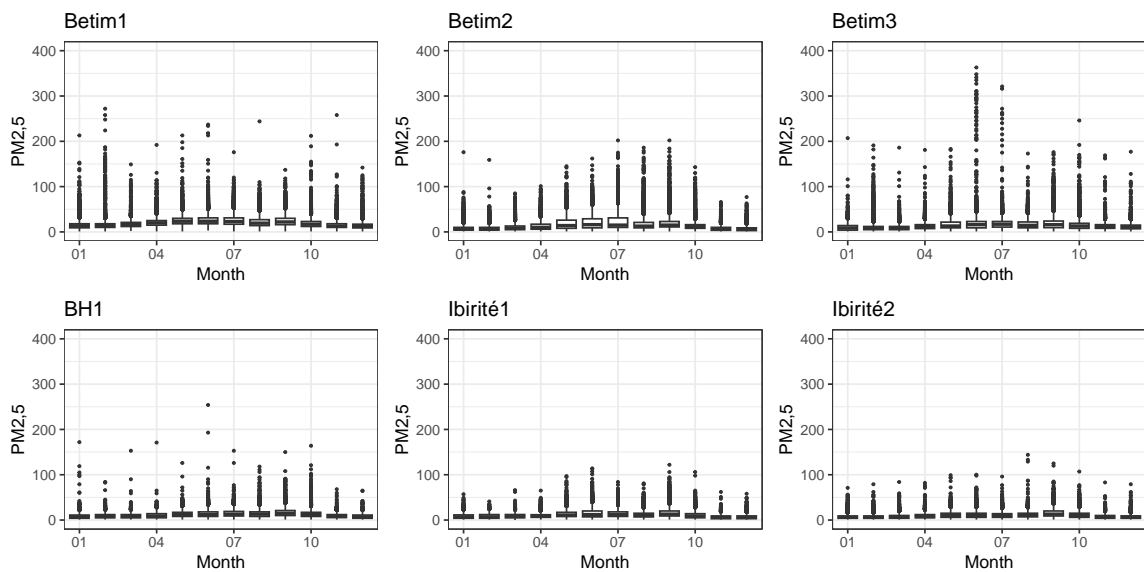


Figura 3.11: Boxplots da concentração de  $\text{PM}_{2,5}$  (em  $\mu\text{m}^2$ ) por mês em cada estação.

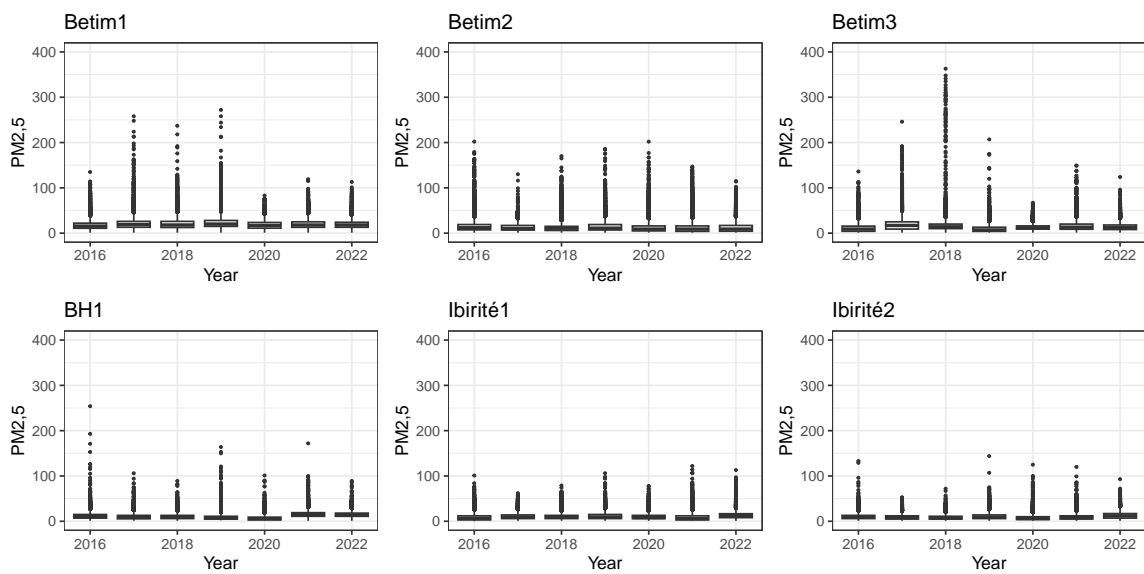


Figura 3.12: Boxplots da concentração de  $\text{PM}_{2,5}$  (em  $\mu\text{m}^2$ ) por ano em cada estação.

### 3.2.2 Ajuste e Previsão dos Modelos para $PM_{2,5}$

Foram testados cinco modelos para realizar as previsões de  $PM_{2,5}$ : SARIMA, ARFIMA, LSTM e duas propostas de modelos híbridos, uma que combina o modelo ARFIMA e LSTM, e outro que combina os modelos SSA, ARFIMA e LSTM. A escolha do modelo ARFIMA deve-se ao fato das séries apresentarem um comportamento característico de séries com memória longa, conforme observado na Figura 3.13, que exibe uma autocorrelação significativa em longos períodos.

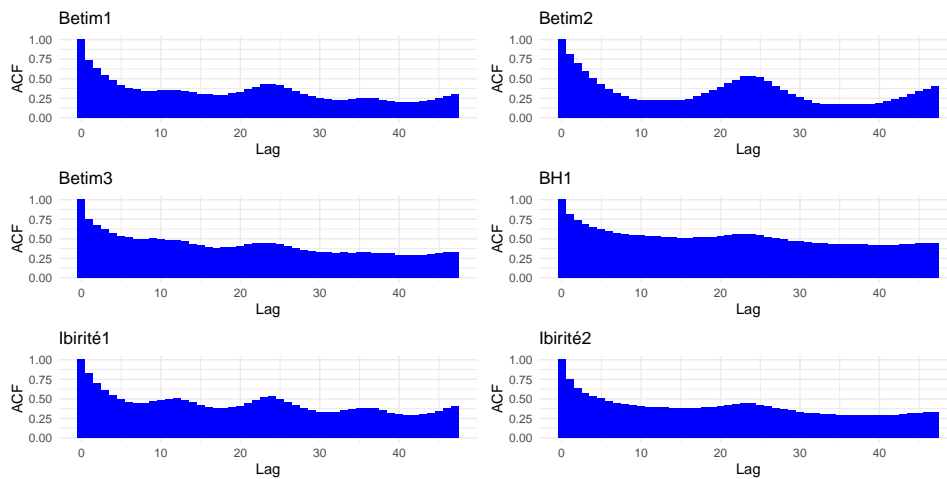


Figura 3.13: Autocorrelação das concentrações de  $PM_{2,5}$  em cada estação.

As Tabelas 3.6 e 3.7 apresentam os valores das métricas de erro para cada modelo avaliado, considerando o ajuste dos modelos e as previsões em três horizontes de tempo: um dia à frente, sete dias à frente (uma semana) e 30 dias à frente (um mês). Podemos observar que o modelo LSTM apresenta o melhor ajuste dentre todos os modelos em todas as estações analisadas. No entanto, ao avaliar as métricas de previsão, o modelo LSTM não obteve as melhores previsões.

De maneira geral, os resultados indicam que os modelos SARIMA e ARFIMA são superiores para previsões de um dia à frente, com uma leve vantagem para o modelo SARIMA. Para previsões de sete dias à frente, os modelos SARIMA, ARFIMA e o modelo Híbrido, que considera os modelos SSA, ARFIMA e LSTM, mostraram-se eficazes. Nas previsões de 30 dias à frente, os modelos Híbridos apresentaram os melhores resultados, sendo o modelo Híbrido que combina ARFIMA e LSTM o que obteve as melhores previsões em uma maior quantidade de estações.

Esses achados destacam a importância de selecionar o modelo adequado para diferentes horizontes de previsão e sugerem que a combinação de modelos pode oferecer vantagens significativas em previsões de longo prazo.

Tabela 3.6: Avaliação do RMSE do ajuste e das previsões para os modelos SARIMA, ARFIMA, LSTM e Modelos Híbridos ARFIMA-LSTM e SSA-ARFIMA-LSTM em horizontes de 1, 7 e 30 dias

Ajustes					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	20,99	21,76	<b>6,09</b>	20,87	21,94
Betim2	15,56	15,42	<b>7,73</b>	15,08	10,63
Betim3	16,10	15,62	<b>8,15</b>	16,35	9,38
BH1	12,38	12,90	<b>4,52</b>	11,75	10,96
Ibirité1	11,18	11,45	<b>4,01</b>	10,98	9,07
Ibirité2	10,58	10,58	<b>3,70</b>	10,33	21,04

Previsão - um dia à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	<b>2,61</b>	2,64	3,87	2,68	3,94
Betim2	<b>1,64</b>	1,73	10,39	2,17	3,97
Betim3	2,13	2,41	6,03	2,24	<b>1,31</b>
BH1	<b>1,40</b>	1,41	2,82	1,69	1,57
Ibirité1	<b>1,31</b>	1,33	3,34	1,40	1,80
Ibirité2	1,39	1,56	3,53	1,76	<b>0,96</b>

Previsão - sete dias à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	<b>3,88</b>	4,47	6,12	4,91	3,91
Betim2	<b>3,51</b>	4,60	8,85	4,75	6,94
Betim3	<b>4,40</b>	6,49	14,83	7,57	5,06
BH1	<b>1,71</b>	2,32	2,99	2,81	1,87
Ibirité1	<b>2,29</b>	2,50	3,72	2,52	2,68
Ibirité2	3,27	4,62	3,83	4,75	<b>1,64</b>

Previsão - 30 dias à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	4,55	5,89	6,38	4,38	<b>4,19</b>
Betim2	5,83	7,74	12,00	<b>4,62</b>	8,75
Betim3	<b>4,95</b>	9,11	8,24	6,77	8,38
BH1	<b>1,83</b>	2,25	2,76	2,43	3,41
Ibirité1	3,88	5,09	6,32	<b>2,42</b>	4,06
Ibirité2	5,40	7,95	5,07	<b>4,64</b>	4,76

Tabela 3.7: Avaliação do SMAPE do ajuste e das previsões para os modelos SARIMA, ARFIMA, LSTM e Modelos Híbridos ARFIMA-LSTM e SSA-ARFIMA-LSTM em horizontes de 1, 7 e 30 Dias

Ajustes					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	1,70	1,66	<b>0,22</b>	1,69	1,77
Betim2	1,64	1,60	<b>0,47</b>	1,61	1,08
Betim3	1,66	1,60	<b>0,36</b>	1,68	0,48
BH1	1,68	1,63	<b>0,35</b>	1,62	1,55
Ibirité1	1,64	1,58	<b>0,32</b>	1,61	1,38
Ibirité2	1,70	1,65	<b>0,31</b>	1,67	2,00

Previsão - um dia à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	<b>0,14</b>	<b>0,14</b>	0,21	<b>0,14</b>	0,24
Betim2	<b>0,59</b>	0,62	1,45	0,70	1,00
Betim3	0,34	0,38	0,76	0,35	<b>0,22</b>
BH1	<b>0,09</b>	<b>0,09</b>	0,19	0,12	0,10
Ibirité1	<b>0,28</b>	<b>0,28</b>	0,55	0,29	0,33
Ibirité2	0,26	0,29	0,59	0,33	<b>0,20</b>

Previsão - sete dias à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	<b>0,22</b>	0,24	0,29	0,26	<b>0,22</b>
Betim2	<b>0,94</b>	1,07	1,35	1,09	1,27
Betim3	<b>0,59</b>	0,78	1,19	0,86	0,67
BH1	<b>0,12</b>	0,18	0,21	0,21	0,15
Ibirité1	<b>0,42</b>	0,46	0,59	0,45	0,46
Ibirité2	0,55	0,67	0,62	0,70	<b>0,28</b>

Previsão - 30 dias à frente					
	SARIMA	ARFIMA	LSTM	Híbrido (ARFIMA+LSTM)	Híbrido (SSA+ARFIMA+LSTM)
Betim1	<b>0,24</b>	0,29	0,32	<b>0,24</b>	<b>0,24</b>
Betim2	1,19	1,32	1,50	<b>1,07</b>	1,37
Betim3	<b>0,65</b>	0,95	0,91	0,80	0,90
BH1	<b>0,14</b>	0,16	0,20	0,19	0,23
Ibirité1	0,60	0,73	0,83	<b>0,44</b>	0,82
Ibirité2	0,76	0,96	0,74	<b>0,68</b>	0,70

# Capítulo 4

## CONCLUSÃO

Este estudo teve como objetivo utilizar modelos de séries temporais para prever as concentrações de  $PM_{10}$  e  $PM_{2,5}$  em estações de monitoramento no estado brasileiro de Minas Gerais, avaliando o ajuste do modelo e o desempenho da previsão por meio das métricas RMSE e SMAPE. Nossa análise abrangeu diferentes horizontes de previsão: uma hora, um dia, uma semana e um mês à frente.

Considerando o ajuste dos modelos das séries temporais de  $PM_{10}$ , o modelo SARIMA apresentou o melhor desempenho em 17 estações, seguido de perto pelo modelo híbrido em 11 estações, enquanto o modelo LSTM demonstrou o melhor ajuste em apenas um caso. Para previsões de uma hora à frente, o modelo Híbrido se destacou em nove estações de monitoramento, com o modelo SARIMA alcançando um RMSE mais baixo em sete estações e o modelo LSTM em 13 estações. Analisando as previsões de um dia à frente, o modelo LSTM demonstrou desempenho superior em nove estações de monitoramento, enquanto o modelo SARIMA produziu um RMSE mais baixo em sete estações, o modelo SSA em cinco estações e o modelo híbrido obteve os melhores resultados em oito estações. Por fim, para previsões de uma semana à frente, o modelo LSTM mostrou o melhor desempenho em cinco estações de monitoramento, enquanto o modelo SARIMA entregou o menor RMSE em dez estações, o modelo SSA em sete estações e o modelo híbrido obteve os melhores resultados em sete estações.

Para o ajuste das concentrações de  $PM_{2,5}$ , o modelo LSTM apresentou os menores erros. Para as previsões de um dia à frente, o modelo Híbrido que combina ARFIMA e LSTM obteve a melhor previsão para a estação de Ibrité2, enquanto o modelo SARIMA se destacou nas demais estações. Esse mesmo padrão foi observado para previsões com horizonte de uma semana. Para previsões de um mês à frente, o modelo Híbrido que combina ARFIMA e LSTM obteve as melhores previsões em três estações, o modelo Híbrido que inclui SSA, ARFIMA e LSTM se destacou na estação de Betim1, e o modelo SARIMA obteve as melhores previsões em duas estações. Ou seja, os modelos híbridos

demonstram superioridade em horizontes de tempo mais longos e, embora não sejam os melhores para horizontes mais curtos, ainda se mostram competitivos.

A utilização de modelos híbridos apresenta vantagens em relação ao uso isolado das técnicas, uma vez que combina diferentes abordagens, cada uma com suas próprias fortalezas, para capturar características complementares dos dados. No caso deste estudo, a combinação de modelos como SARIMA ou ARFIMA e LSTM permitiu melhorar a precisão das previsões, visto que cada modelo aborda aspectos distintos da série temporal, como tendências, sazonalidade, memória longa e padrões não lineares. Ao ajustar os resíduos dos modelos tradicionais com redes neurais, como o LSTM, é possível reduzir erros que poderiam ser negligenciados por modelos isolados, resultando em previsões mais robustas e confiáveis.

Em conclusão, este estudo forneceu insights valiosos para prever as concentrações de  $PM_{10}$  e  $PM_{2,5}$  no estado brasileiro de Minas Gerais, utilizando diferentes classes de modelos de séries temporais, modelos paramétricos, não paramétricos, de aprendizado profundo e híbridos, fornecendo insights valiosos para formuladores de políticas ambientais e partes interessadas no desenvolvimento de estratégias eficazes para o gerenciamento da qualidade do ar. Os modelos e propostas utilizados podem ser aplicados para analisar e prever dados de outras localidades e até mesmo outros tipos de poluentes. Pesquisas futuras podem se basear nesses resultados explorando modelos híbridos adicionais e refinando ainda mais as técnicas de previsão para aumentar a precisão preditiva.

## 4.1 Resultados Alcançados

Os resultados deste trabalho foram amplamente disseminados na comunidade científica por meio de publicações e apresentações em eventos. Um dos principais resultados foi a publicação de um artigo na revista *Scientific Reports*. Além disso, outro artigo já foi submetido para publicação. Um terceiro manuscrito está em fase de escrita, com previsão de submissão em breve. O trabalho também foi apresentado em diversos eventos acadêmicos e conferências, incluindo o Congresso UFBA, o EBEST (Encontro Baiano de Estatística), o SINAPE (Simpósio Nacional de Probabilidade e Estatística) e o CSDS (Conference on Statistical and Data Science). Essas publicações e apresentações evidenciam a relevância e o impacto do estudo, consolidando sua contribuição para a área de previsão da qualidade do ar e modelos de séries temporais.

## 4.2 Trabalhos Futuros

Este estudo focou na previsão de concentrações de  $PM_{10}$  e  $PM_{2,5}$  utilizando apenas as séries temporais univariadas das concentrações dos poluentes. No entanto, um caminho promissor para trabalhos futuros é a inclusão de variáveis exógenas nos modelos. Fatores como condições meteorológicas (temperatura, umidade, velocidade do vento), dados de tráfego, e informações socioeconômicas podem influenciar as concentrações de poluentes e, ao serem incorporados aos modelos, podem aumentar significativamente a capacidade preditiva. Abordagens de séries temporais multivariadas ou LSTM multivariado, poderiam ser exploradas para integrar essas variáveis exógenas de forma mais eficiente, capturando as interdependências entre múltiplas séries temporais.

Outro aspecto relevante para estudos futuros é a realização de análises de intervenção. Avaliar o impacto de políticas públicas ou mudanças ambientais, como restrições ao tráfego de veículos ou a implementação de tecnologias mais limpas em áreas urbanas, pode fornecer insights valiosos sobre a eficácia dessas medidas na redução das concentrações de  $PM_{10}$  e  $PM_{2,5}$ . Métodos de análise de intervenção podem ser aplicados para detectar e quantificar mudanças abruptas nas séries temporais após essas intervenções, contribuindo para o desenvolvimento de políticas públicas mais informadas e eficazes no controle da poluição atmosférica.

# Referências Bibliográficas

ARTAXO, P. O estado da qualidade do ar no brasil. *Working Paper WRI Brasil*, p. 32, 2021.

BAE, S.; HONG, Y.-C. Health effects of particulate matter. *Journal of the Korean Medical Association*, Korean Medical Association, v. 61, n. 12, p. 749–755, 2018.

BRAGA, A. L. F. et al. Association between air pollution and respiratory and cardiovascular diseases in itabira, minas gerais state, brazil. *Cadernos de Saúde Pública*, SciELO Brasil, v. 23, p. S570–S578, 2007.

CORDOVA, C. H. et al. Air quality assessment and pollution forecasting using artificial neural networks in metropolitan lima-peru. *Scientific Reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 24232, 2021.

DÍAZ-ROBLES, L. A. et al. A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile. *Atmospheric Environment*, Elsevier, v. 42, n. 35, p. 8331–8340, 2008.

ENCALADA-MALCA, A. A. et al. A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima. *Atmosphere*, MDPI, v. 12, n. 5, p. 609, 2021.

GOLYANDINA, N.; KOROBAYNIKOV, A.; ZHIGLJAVSKY, A. *Singular spectrum analysis with R*. [S.l.]: Springer, 2018.

GOLYANDINA, N.; NEKRUTKIN, V.; ZHIGLJAVSKY, A. A. *Analysis of time series structure: SSA and related techniques*. [S.l.]: CRC press, 2001.

GRAVES, T. et al. A brief history of long memory: Hurst, mandelbrot and the road to arfima, 1951–1980. *Entropy*, MDPI, v. 19, n. 9, p. 437, 2017.

HASSANI, H.; MAHMOUDVAND, R. *Singular spectrum analysis: Using R*. [S.l.]: Springer, 2018.

HASSANI, H.; THOMAKOS, D. A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, International Press of Boston, v. 3, n. 3, p. 377–397, 2010.

IFTIKHAR, H. et al. Forecasting day-ahead brent crude oil prices using hybrid combinations of time series models. *Mathematics*, MDPI, v. 11, n. 16, p. 3548, 2023.

ISMAIL, S.; SHABRI, A.; SAMSUDIN, R. A hybrid model of self-organizing maps (som) and least square support vector machine (lssvm) for time-series forecasting. *Expert Systems with Applications*, Elsevier, v. 38, n. 8, p. 10574–10578, 2011.

- KAZIENKO, P.; LUGHOFER, E.; TRAWINSKI, B. Editorial on the special issue “hybrid and ensemble techniques in soft computing: recent advances and emerging trends”. *Soft computing*, Springer, v. 19, p. 3353–3355, 2015.
- LENSCHOW, P. et al. Some ideas about the sources of pm10. *Atmospheric environment*, Elsevier, v. 35, p. S23–S33, 2001.
- LI, J.; XU, G.; CHENG, X. Combining spatial pyramid pooling and long short-term memory network to predict pm2.5 concentration. *Atmospheric Pollution Research*, Elsevier, v. 13, n. 3, p. 101309, 2022.
- LI, T.; HUA, M.; WU, X. A hybrid cnn-lstm model for forecasting particulate matter (pm2.5). *Ieee Access*, IEEE, v. 8, p. 26933–26940, 2020.
- LÓPEZ-GONZALES, J. L. et al. Self-organizing topological multilayer perceptron: A hybrid method to improve the forecasting of extreme pollution values. *Stats*, MDPI, v. 6, n. 4, p. 1241–1259, 2023.
- MALAMIRI, H. R. G. et al. Comparison of harmonic analysis of time series (hants) and multi-singular spectrum analysis (m-ssa) in reconstruction of long-gap missing data in ndvi time series. *Remote Sensing*, MDPI, v. 12, n. 17, p. 2747, 2020.
- MALEKI, H. et al. Temporal profile of pm10 and associated health effects in one of the most polluted cities of the world (ahvaz, iran) between 2009 and 2014. *Aeolian research*, Elsevier, v. 22, p. 135–140, 2016.
- MEDINA, S. et al. Aphis: public health impact of pm10 in 19 european cities. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 58, n. 10, p. 831–836, 2004.
- MEMARIANFARD, M.; HATAMI, A. Artificial neural network forecast application for fine particulate matter concentration using meteorological data. *GLOBAL JOURNAL OF ENVIRONMENTAL SCIENCE AND MANAGEMENT*, 2017.
- MORETTIN, P. A. *Ondas e Ondaletas Vol. 23*. [S.l.]: Edusp, 1999.
- MORETTIN, P. A.; TOLOI, C. M. *Análise de Séries Temporais: Modelos lineares univariados*. [S.l.]: Editora Blucher, 2018.
- MORITZ, S.; BARTZ-BEIELSTEIN, T. imputeTS: Time Series Missing Value Imputation in R. *R Journal*, v. 9, n. 1, 2017.
- MUKHERJEE, A.; AGRAWAL, M. World air particulate matter: sources, distribution and health effects. *Environmental chemistry letters*, Springer, v. 15, p. 283–309, 2017.
- OLAH, C. *Understanding LSTM Networks*. 2015. Retrieved from <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Disponível em: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- PINTO, W. d. P.; REISEN, V. A.; MONTE, E. Z. Previsão da concentração de material particulado inalável, na região da grande vitória, es, brasil, utilizando o modelo sarimax. *Engenharia Sanitária e Ambiental*, SciELO Brasil, v. 23, p. 307–318, 2018.

- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. R package version 4.3.1. Disponível em: <https://www.R-project.org/>.
- RODRIGUES, P. C.; MAHMOUDVAND, R. The benefits of multivariate singular spectrum analysis over the univariate version. *Journal of the Franklin Institute*, Elsevier, v. 355, n. 1, p. 544–564, 2018.
- SACKS, J. D. et al. Particulate matter–induced health effects: who is susceptible? *Environmental health perspectives*, National Institute of Environmental Health Sciences, v. 119, n. 4, p. 446–454, 2011.
- SAHOO, B. B. et al. Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, Springer, v. 67, n. 5, p. 1471–1481, 2019.
- SAMIA, A.; KAOUTHER, N.; ABDELWAHED, T. A hybrid arima and artificial neural networks model to forecast air quality in urban areas: case of tunisia. *Advanced Materials Research*, Trans Tech Publ, v. 518, p. 2969–2979, 2012.
- SÁNCHEZ-CCOYLLO, O. R. et al. Vehicular particulate matter emissions in road tunnels in sao paulo, brazil. *Environmental monitoring and assessment*, Springer, v. 149, p. 241–249, 2009.
- SANTOS, J. A. A. dos; SPANCERSKI, J. S. Previsão da produtividade de arroz: uma aplicação de redes neurais recorrentes lstm. *Revista Cereus*, v. 13, n. 2, p. 163–175, 2021.
- SILVA, F. Leite Coelho da et al. Statistical and artificial neural networks models for electricity consumption forecasting in the brazilian industrial sector. *Energies*, MDPI, v. 15, n. 2, p. 588, 2022.
- SILVA, K. L. S. da et al. Spatio-temporal visualization and forecasting of pm 10 in the brazilian state of minas gerais. *Scientific Reports*, Nature Publishing Group UK London, v. 13, n. 1, p. 3269, 2023.
- SONG, X. et al. Time-series well performance prediction based on long short-term memory (lstm) neural network model. *Journal of Petroleum Science and Engineering*, Elsevier, v. 186, p. 106682, 2020.
- SULANDARI, W. et al. Indonesian electricity load forecasting using singular spectrum analysis, fuzzy systems and neural networks. *Energy*, Elsevier, v. 190, p. 116408, 2020.
- SULANDARI, W. et al. Ssa-based hybrid forecasting models and applications. *Bulletin of Electrical Engineering and Informatics*, v. 9, n. 5, p. 2178–2188, 2020.
- THOMPSON, J. E. Airborne particulate matter: human exposure and health effects. *Journal of occupational and environmental medicine*, LWW, v. 60, n. 5, p. 392–423, 2018.
- TIBULO, C. et al. Previsão da concentração de material particulado inalável, através de modelos estatísticos de séries temporais para o município de canoas, rio grande do sul. *Revista Thema*, v. 19, n. 1, p. 134–152, 2021.
- VAUTARD, R.; GHIL, M. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D: Nonlinear Phenomena*, Elsevier, v. 35, n. 3, p. 395–424, 1989.

XU, D. et al. Application of a hybrid arima-lstm model based on the spei for drought forecasting. *Environmental Science and Pollution Research*, Springer, v. 29, n. 3, p. 4128–4144, 2022.

ZHANG, G. P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, Elsevier, v. 50, p. 159–175, 2003.

Universidade Federal da Bahia - UFBA  
Instituto de Matemática / Programa de pós-graduação em Matemática

---

Av. Adhemar de Barros, s/n, Campus Universitário de Ondina, Salvador - BA

CEP: 40170-110

<<http://www.pgmat.ufba.br>>