



UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA APLICADA

Um Processo Baseado em Séries Temporais para Predição da Resistência a Antimicrobianos

Anderson Rodrigues Cavalcante

Recife - PE, Agosto 2025

Anderson Rodrigues Cavalcante

Um Processo Baseado em Séries Temporais para Predição da Resistência a Antimicrobianos

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

Universidade Federal Rural de Pernambuco – UFPE

Departamento de Estatística e Informática

Programa de Pós-Graduação em Informática Aplicada

Orientador: Prof. Dr. Victor Wanderley Costa de Medeiros

Coorientador: Prof. Dr. Rodrigo Gabriel Ferreira Soares

Recife - PE

Agosto 2025

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Gerada automaticamente, mediante os dados fornecidos
pelo(a) autor(a)

C377p Cavalcante, Anderson Rodrigues.
Um processo baseado em séries temporais para
predição da resistência a antimicrobianos /
Anderson Rodrigues Cavalcante. – Recife, 2025.
85 f.; il.

Orientador(a): Victor Wanderley Costa de
Medeiros.
Co-orientador(a): Rodrigo Gabriel Ferreira
Soares.

Dissertação (Mestrado) – Universidade Federal
Rural de Pernambuco, Programa de Pós-Graduação
em Informática Aplicada, Recife, BR-PE, 2024.

Inclui referências e apêndice(s).

1. ARIMA. 2. Autorregressão. 3. Média móvel. 4.
Séries temporais 5. AMR. I. Medeiros, Victor
Wanderley Costa de, orient. II. Soares, Rodrigo
Gabriel Ferreira, coorient. III. Título

CDD 004

Anderson Rodrigues Cavalcante

Um Processo Baseado em Séries Temporais para Predição da Resistência a Antimicrobianos

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

Aprovada em Recife - PE, Agosto 26, 2025:

Prof. Dr. Victor Wanderley Costa de Medeiros (Orientador)
Universidade Federal Rural de Pernambuco

Prof. Dr. Tiago Alessandro Espínola Ferreira
Universidade Federal Rural de Pernambuco

Dra. Izolda Maria Fernandes de Moura
Universidade Federal de Pernambuco

Recife - PE
Agosto 2025

*Dedico ao Senhor Jesus, a minha esposa Mylena, a meus pais Naildo e Tânia e a meu
irmão Gustavo.*

Agradecimentos

A meu professor e orientador Victor Medeiros que sempre me incentivou e me ajudou no mundo da pesquisa acadêmica e toda equipe de enfermagem e biomedicina do Hospital das Clínicas PE representados por Carol Medeiros, Andreza, Shirley, Izolda e Danylo.

Ao instituto SiDi, sempre fomentando a pesquisa em ciência e tecnologia, me permitindo durante todo o curso, dedicar horas semanais a minha pesquisa e aulas.

Agradeço a meus pais, Tânia e Naildo, que sacrificaram muito para me criarem com muita educação, também a meu irmão Gustavo, meu amigo durante a vida.

À minha mulher Mylena que sempre me ajudou a superar meus obstáculos e ser uma esposa incrível.

Por último e mais importante, agradeço a meu Senhor e Salvador Jesus Cristo, que em sua grande misericórdia olhou para a humanidade e decidiu salvá-la e nunca me deixou de lado.

“Entrega o teu caminho ao Senhor; confia nele, e ele o mais fará.”
(Salmos 37:5)

Resumo

A Resistência Antimicrobiana (AMR) representa uma das maiores ameaças à saúde pública global, sendo responsável por milhões de mortes anualmente. Os métodos diagnósticos tradicionais como o antibiograma, embora precisos, apresentam um tempo de resposta elevado, de 24 a 48 horas, e um custo considerável, o que pode atrasar o início do tratamento adequado e impactar negativamente os desfechos clínicos. Diante desse cenário, a busca por métodos preditivos e de baixo custo é fundamental para auxiliar decisões hospitalares e fortalecer a vigilância microbiológica. Esta dissertação propõe um processo para utilização do Modelo Autoregressivo Integrado de Média Móvel (ARIMA) como ferramenta de predição da evolução de resistência bacteriana em um ambiente hospitalar. O objetivo é validar o processo de análise de séries temporais como uma ferramenta robusta para antecipar tendências de resistência, oferecendo suporte à decisão clínica e contribuindo para a otimização dos Programas de Gerenciamento de Antibióticos (PGA). Para isso, foi pré-processada e construída uma base de dados a partir de 19 mil registros de antibiogramas do Hospital das Clínicas de Pernambuco, abrangendo o período de 2016 a 2024. Devido a lacunas e ausência de informações nos dados, as taxas de resistência foram agregadas em intervalos mensais. Esse foi um dos motivos de abordagens de aprendizado de máquina terem sido descartadas para análise. Foram modeladas as séries temporais dos patógenos mais relevantes do hospital - *Klebsiella*, *Escherichia coli* e *Pseudomonas* - bem como os antibióticos específicos para cada um deles. O desempenho do ARIMA foi comparado com abordagens de regressão (Logit Binomial, Linear e Log-Linear Poisson). Os resultados indicaram uma superioridade geral do ARIMA. No entanto, a alta ocorrência de valores extremos trouxe um maior grau de complexidade para predição dos modelos propostos. Ainda assim, conclui-se que o ARIMA é uma ferramenta poderosa para vigilância preditiva de AMR, especialmente em séries com forte componente temporal.

Palavras-chave: ARIMA, autorregressão, média móvel, séries temporais, AMR, resistência bacteriana.

Abstract

Antimicrobial Resistance (AMR) represents one of the greatest threats to global public health, being responsible for millions of deaths annually. Although accurate, traditional diagnostic methods, such as the antibiogram, present a high response time of 24 to 48 hours and considerable cost, which can delay the initiation of appropriate treatment and negatively impact clinical outcomes. In this scenario, searching for predictive and low-cost methods is essential to support hospital decision-making and strengthen microbiological surveillance. This dissertation proposes a process for using the Autoregressive Integrated Moving Average (ARIMA) model as a predictive tool to evolve bacterial resistance in a hospital environment. The goal is to validate the time series analysis process as a robust tool to anticipate resistance trends, provide clinical decision support, and contribute to optimizing Antimicrobial Stewardship Programs (ASP). For this purpose, a database was pre-processed and constructed from 19,000 antibiogram records from the Hospital das Clínicas of Pernambuco, covering the period from 2016 to 2024. Due to data gaps and missing information, resistance rates were aggregated into monthly intervals, which is also why we discarded machine learning approaches for analysis. Time series were modeled for the hospital's most relevant pathogens — *Klebsiella*, *Escherichia coli* and *Pseudomonas* — as well as for the specific antibiotics used against them. We compared the performance of ARIMA with regression approaches (Binomial Logit, Linear, and Log-Linear Poisson). The results indicated an overall superiority of ARIMA. However, the high occurrence of extreme values introduced greater complexity to the prediction of the proposed models. Nevertheless, it is concluded that ARIMA is a powerful tool for predictive AMR surveillance, especially in series with a strong temporal component.

Keywords: ARIMA, autoregression, moving average, time series, AMR, bacterial resistance.

Lista de ilustrações

Figura 1 – Mortes causadas por AMR até 2050. Fonte: (UNEP, 2023).	3
Figura 2 – Como acontece a resistência bacteriana. Fonte: Elaborada pelo autor.	11
Figura 3 – Série temporal de um Lote e o seu Rendimento. Fonte: (BOX et al., 2016)	14
Figura 4 – Fluxograma de execução do trabalho. Fonte: Elaborada pelo autor	17
Figura 5 – Formato da coluna Bactéria e Antibióticos - Dataset Mastertools 284. Fonte: Elaborada pelo autor.	19
Figura 6 – Formato da coluna Bactéria e Antibióticos - Dataset Mastertools 762. Fonte: Elaborada pelo autor.	20
Figura 7 – A estrutura usada para a taxonomia de um microorganismo do pacote AMR. O ID consiste em iniciais separadas por underscore que são, respectivamente: o reino do microorganismo; o gênero; espécie; e subespécie. Fonte: (BERENDS et al., 2022).	22
Figura 8 – Série temporal Klebsiella-Cefepima. Fonte: Elaborada pelo autor.	24
Figura 9 – Série temporal da Klebsiella-Gentamicina. Fonte: Elaborada pelo autor.	25
Figura 10 – Série temporal Klebsiella-Meropenem. Fonte: Elaborada pelo autor.	25
Figura 11 – ACF e PACF da taxa de resistência da <i>Escherichia coli</i> em amostra de sangue com o antibiótico Amicacina entre 2015–2019. Fonte: (LOBO ASHA KAMATH, 2021).	32
Figura 12 – Modelo ARIMA(2,1,1) constatado como subajustado para a série temporal da Klebsiella-Cefepima. Fonte: Elaborada pelo autor	34
Figura 13 – Resíduos do modelo ARIMA(2,1,1) constatado como ruído branco, ou seja, ideal para a série temporal da <i>Escherichia coli</i> -Ampicilina. Fonte: Elaborada pelo autor	34
Figura 14 – Em (A) ACF e PACF do par <i>Escherichia coli</i> -Ampicilina. Em (B) o par <i>Escherichia coli</i> - Gentamicina. Fonte: Elaborada pelo autor.	36
Figura 15 – ACF e PACF do par <i>Pseudomonas</i> -Amicacina. Fonte: Elaborada pelo autor.	36
Figura 16 – Previsão da resistência da Klebsiella quando ministrado Cefepima utilizando ARIMA(2,1,1). Fonte: Elaborada pelo autor.	40
Figura 17 – Previsão da resistência da Klebsiella quando ministrado Cefepima utilizando Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.	40
Figura 18 – Gráfico da ACF e PACF para a Klebsiella-Cefepima. Fonte: Elaborada pelo autor.	42
Figura 19 – Previsão da resistência da Klebsiella-Cefepima utilizando o ARIMA(2,1,1). Fonte: Elaborada pelo autor.	42

Figura 20 – Gráfico da previsão utilizando Regressão Linear para o par Klebsiella-Gentamicina. Fonte: Elaborada pelo autor.	43
Figura 21 – Previsão da resistência da Escherichia-coli quando ministrado Ampicilina utilizando ARIMA(2,1,1). Fonte: Elaborada pelo autor.	44
Figura 22 – Previsão da resistência da Escherichia-coli quando administrada a Ampicilina utilizando a Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.	45
Figura 23 – Previsão da resistência da Escherichia-coli quando ministrado Ceftriaxona utilizando a Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.	45
Figura 24 – Previsão da resistência da Escherichia-coli para a Ampicilina utilizando ARIMA(3,1,1) na série quinzenal. Fonte: Elaborada pelo autor.	46
Figura 25 – Previsão da resistência da Escherichia-coli para a Ampicilina utilizando Regressão Linear na série quinzenal. Fonte: Elaborada pelo autor.	47
Figura 26 – Previsão da resistência da Escherichia-coli para a Ceftriaxona utilizando ARIMA(2,1,1) na série quinzenal. Fonte: Elaborada pelo autor.	47
Figura 27 – Previsão da resistência da Pseudomonas quando ministrado Amicacina utilizando ARIMA(1,1,1). Fonte: Elaborada pelo autor.	48
Figura 28 – Previsão da resistência da Pseudomonas quando ministrado Amicacina utilizando ARIMA(3,1,1). Fonte: Elaborada pelo autor.	49
Figura 29 – Previsão da resistência da Pseudomonas quando ministrado Amicacina utilizando Logit Binomial da biblioteca AMR for R. Fonte: Elaborada pelo autor.	49
Figura 30 – Previsão da resistência da Pseudomonas-Amicacina utilizando ARMA(2,0,2) na série quinzenal. Fonte: Elaborada pelo autor.	50
Figura 31 – Previsão da resistência da Pseudomonas-Meropenem utilizando Regressão Logit Binomial na série quinzenal. Fonte: Elaborada pelo autor.	51
Figura 32 – Previsão da resistência da Pseudomonas-Meropenem utilizando o ARIMA(2,1,1) na série quinzenal. Fonte: Elaborada pelo autor.	51

Lista de tabelas

Tabela 1 – Detalhamento dos Trabalhos Relacionados.	8
Tabela 2 – Culturas de bactérias coletadas de maio de 2016 a dezembro de 2024. .	23
Tabela 3 – Pares Antibiótico-Bactéria utilizados para análise.	24
Tabela 4 – Modelos para cada antibiótico administrado na Klebsiella.	35
Tabela 5 – Modelos para cada antibiótico administrado na Escherichia-coli.	35
Tabela 6 – Modelos para cada antibiótico administrado na Pseudomonas.	35
Tabela 7 – Intervalos analisados com seus respectivos bimestres limites de treino e teste.	37
Tabela 8 – Comparação dos resultados da previsão para Cefepima e Gentamicina ministrados na Klebsiella da análise mensal.	41
Tabela 9 – Comparação de desempenho dos modelos para a previsão quinzenal da resistência da Klebsiella à Cefepima e Gentamicina	41
Tabela 10 – Comparação dos resultados da previsão da Ampicilina, Ceftriaxona e Gentamicina para a Escherichia coli usando ARIMA e modelos de Regressão da biblioteca AMR For R.	44
Tabela 11 – Comparação dos resultados da previsão da Escherichia coli utilizando dados de resistência quinzenal.	46
Tabela 12 – Comparação dos resultados da previsão Pseudomonas implementando modelos ARIMA e a biblioteca AMR For R utilizando modelos de Regressão.	48
Tabela 13 – Tabela de resultados da Pseudomonas com a série temporal quinzenal.	50

Lista de abreviaturas e siglas

ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
AR	Auto-regressive
AMR	Resistência Antimicrobiana
ARIMA	Modelo Autoregressivo Integrado de Média Móvel
ATB	Antibiótico
BIC	Bayesian Information Criterion
BRCAT	Comitê Brasileiro de Testes de Sensibilidade aos Antimicrobianos
EUCAT	Comitê Europeu para o Estudo da Suscetibilidade Antimicrobiana
FS	Feature Selection
I	Indeterminado
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
MA	Moving Average
MIC	Minimum Inhibitory Concentration
NT	Não Testado
PGA	Programa de Gerenciamento de Antibióticos
R	Resistente
RL	Regression Logistic
S	Sensível
SSWM	Strong Selection Weak Mutation
UTI	Unidade de Terapia Intensiva
WHO	World Health Organization

Sumário

1	INTRODUÇÃO	1
1.1	Motivação	1
1.2	Objetivos	3
1.3	Organização do trabalho	4
2	TRABALHOS RELACIONADOS	5
3	REFERENCIAL TEÓRICO	9
3.1	Antibióticos	9
3.2	Bactérias	10
3.3	Resistência Antimicrobiana	10
3.3.1	Classificação de Ambler	12
3.4	Séries Temporais	12
3.5	Análise Estatística	13
3.6	Caminhada Aleatória	14
4	MATERIAIS E MÉTODOS	17
4.1	Fluxo do Trabalho	17
4.2	Base de Dados	18
4.2.1	Pacote R para resistência antimicrobiana	18
4.2.2	Pré-Processamento	19
4.2.3	Integração das bases de dados	20
4.3	Seleção de Bactérias e Antibióticos	23
4.4	ARIMA	25
4.4.1	Estimativa dos Coeficientes ϕ e θ	27
4.4.2	Estacionariedade de Séries Temporais	28
4.4.3	Escolha de Parâmetros	30
4.4.3.1	Teste Dickey-Fuller	30
4.4.3.2	Teste KPSS	31
4.4.3.3	Autocorrelação e Autocorrelação Parcial	31
4.4.4	Validação dos Parâmetros	31
4.4.4.1	AIC/BIC	32
4.4.4.2	Ljung-Box	33
4.4.5	Treino e Teste	33
4.5	Métricas de Desempenho	37

5	RESULTADOS E DISCUSSÃO	39
5.1	Análise Preditiva da Klebsiella	39
5.1.1	Agregação Quinzenal para a Klebsiella	41
5.2	Análise Preditiva da Escherichia coli	43
5.2.1	Agregação Quinzenal para a Escherichia coli	46
5.3	Análise Preditiva da Pseudomonas	47
5.3.1	Agregação Quinzenal para a Pseudomonas	49
6	CONCLUSÕES	53
6.1	Dificuldades na Previsão e a Volatilidade das Séries	54
6.2	Contribuições	55
6.3	Trabalhos Futuros	56
	REFERÊNCIAS	59
	Apêndices	63

1 Introdução

Os antibióticos revolucionaram o sistema de saúde no combate a infecções bacterianas desde a criação do primeiro (Penicilina) por Alexander Fleming ([TAN; TATSUMURA, 2015](#)), que lhe rendeu o Nobel de Medicina/Fisiologia em 1945. Com o advento desse medicamento, tratamentos mal administrados e mutações genéticas naturais fizeram com que os microorganismos passassem a desenvolver resistência a muitos antibióticos.

Segundo a *World Health Organization* (WHO) quando um patógeno se torna resistente a um antimicrobiano significa dizer que o tratamento com este medicamento se tornou ineficiente e é muito difícil ou até mesmo impossível a infecção ser curada. Isso é um processo natural onde os patógenos passam por mutações genéticas, o que dificulta o tratamento ([WHO, 2023](#)).

Resistência antimicrobiana (AMR) é um dos maiores problemas de saúde pública global atualmente tratados. É estimado que mundialmente a AMR foi diretamente responsável por 1,27 milhão de mortes e contribuiu indiretamente para 4,95 milhões de mortes em 2019 ([MURRAY, 2022](#)). Segundo o Institute for Health Metrics and Evaluation ([IHME, 2023](#)), 3,83 milhões de mortes foram relacionadas a infecções bacterianas na região africana da WHO, com a AMR sendo responsável por mais de 25% dessa mortalidade. Além do custo de milhões de vidas, os gastos trazidos por este desafio chegam a cifras significativas. O Banco Mundial estima que até 2050 a AMR pode resultar em 1 trilhão de dólares em custos adicionais com saúde, e perdas de US\$ 1 trilhão a US\$ 3,4 trilhões no produto interno bruto (PIB) por ano até 2030.

1.1 Motivação

O resultado de um antibiograma leva em torno de 24 a 48 horas. O resultado irá mostrar se o microrganismo é Resistente ou Sensível ao ATB. Em alguns casos o resultado pode ser Indeterminado. Quanto mais rápido o resultado for dado, mais rápido o paciente será encaminhado para o tratamento adequado. Além disso, os procedimentos analíticos mais comumente utilizados para detecção de AMR, como o teste de Concentração Inibitória Mínima (MIC), possuem alto custo ([KIM et al., 2023](#)).

Dependendo do custo computacional, métodos estatísticos podem produzir estimativas preditivas mais rápidas, indicando se o patógeno se tornará resistente ou se está respondendo positivamente ao tratamento, ou seja, sensível ao antimicrobiano.

A velocidade no diagnóstico pode servir de guia para direcionar a equipe hospitalar a implementar tratamentos mais consonantes e projetar PGA (Programa de Gerenciamento

de Antibióticos) de maneira mais rápida e mais congruente ao estado mais recente dos pacientes. Dessa forma, se um algoritmo, em um baixo tempo e com alta acurácia, prevê que uma bactéria se tornará resistente, os profissionais poderão alterar a abordagem no tratamento de forma que a resistência não aconteça.

Essa efetividade somada a um menor custo trazido pelo uso da estatística, podem gerar diversos benefícios, como a preservação de milhões vidas humanas (MURRAY, 2022; IHME, 2023). O método ARIMA, por exemplo, já foi empregado com sucesso para detecção de resistência antimicrobiana em fazendas de suínos (KIM et al., 2023), sendo assim, útil também no setor pecuário.

Considerando-se o custo que o tratamento de infecções bacterianas trazem, mitigar os impactos de microrganismos resistentes podem trazer importantes implicações econômicas. Estimativas do Banco Mundial apontam que a AMR pode custar 1 trilhão de dólares ao redor do mundo até 2030 e perdas de até 3,4 trilhões de dólares por ano até 2030.

Devido a melhora do sistema de saúde global do último século o uso de antibióticos aumentou (MOHAMMED et al., 2018) e com isso uma tendência crescente de seu uso indevido. Isso é causado por fatores como a falta de educação sobre o uso de antibióticos, má regulamentação de medicamentos prescritos e vendas sem receita (LOOD ROLF, 2017).

Um estudo sobre o consumo de ATBs mostrou uma crescente de mais de 70 milhões de unidades no período de 2000 a 2010, o que representou um aumento de 35% em relação à década anterior (BOECKEL et al., 2015).

Não por acaso, de acordo o relatório da ONU (UNEP, 2023) intitulado “Preparando-se para Super Bactérias”, resistência antimicrobiana é, junto com o câncer, de longe a maior causa de milhões de mortes em relação a outras doenças e infecções, como mostrado na Figura 1.

Sabendo da importância de iniciativas voltadas para o estudo de resistência antimicrobiana, o BRCAST (Comitê Brasileiro de Testes de Sensibilidade aos Antimicrobianos) 2024 determina, baseado também no EUCAST (Comitê Europeu para o Estudo da Suscetibilidade Antimicrobiana) 2024, três categorias de sensibilidade (BRCAST, 2024):

- S - Sensível, dose padrão: Um microrganismo é categorizado como Sensível, quando há uma alta probabilidade de sucesso terapêutico utilizando o regime de dose padrão do agente.
- I - Sensível, aumentando exposição: Um microrganismo é categorizado como Sensível, aumentando a exposição, quando há alta probabilidade de sucesso terapêutico devido ao aumento da exposição do antimicrobiano ao microrganismo.
- R - Resistente: um microrganismo é categorizado como Resistente quando o antimicrobiano não consegue pausar sua multiplicação ou eliminá-lo.

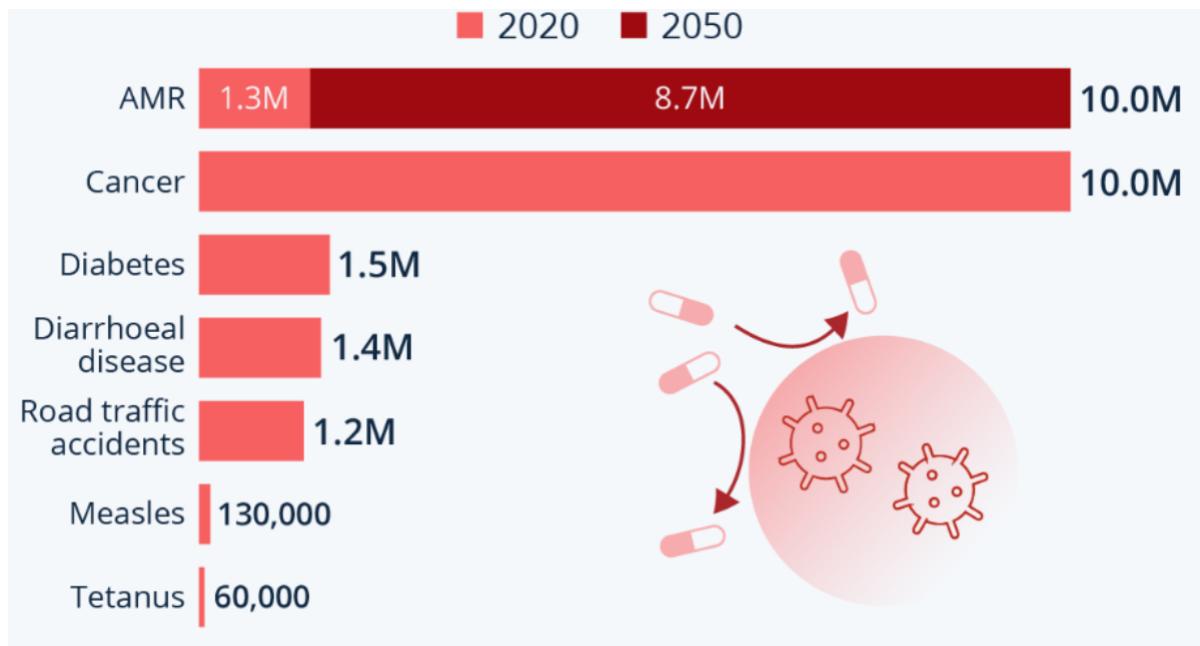


Figura 1 – Mortes causadas por AMR até 2050. Fonte: (UNEP, 2023).

1.2 Objetivos

O objetivo principal desta dissertação é propor e validar um processo baseado em séries temporais para a predição da resistência a antimicrobianos em ambiente hospitalar. Esse processo abrange desde o pré-processamento dos dados e seleção das técnicas estatísticas até a avaliação comparativa dos modelos, com ênfase no uso do modelo Autoregressivo Integrado de Média Móvel (ARIMA). Pretende-se, assim, demonstrar a viabilidade da análise de séries temporais como uma ferramenta para antecipar tendências de resistência bacteriana, contribuindo para o apoio à decisão clínica e para a otimização de estratégias de controle de infecções.

E como objetivos específicos, pretende-se:

- Predizer a evolução futura da resistência bacteriana utilizando modelos lineares estatísticos;
- Organizar uma base de dados que será utilizada neste trabalho e em estudos futuros;
- Oferecer suporte à decisão clínica e contribuir para políticas de controle de infecção bacteriana.
- Avaliar e comparar o desempenho dos modelos preditivos estatísticos de séries temporais;

1.3 Organização do trabalho

Este trabalho está organizado em seis capítulos, sendo o primeiro de introdução e o segundo sobre os trabalhos relacionados. O Capítulo 3 apresenta o referencial teórico sobre aprendizado de máquina, sendo supervisionado e não supervisionado e análise estatística de séries temporais. O Capítulo 4 descreve os materiais e métodos utilizados no desenvolvimento do projeto, incluindo a base de dados utilizada, seu pré-processamento, as considerações sobre as séries temporais e suas limitações, os algoritmos testados e as métricas utilizadas para a avaliação de resultados. O Capítulo 5 apresenta e discute os resultados obtidos. No Capítulo 6 têm-se as conclusões e considerações finais sobre este trabalho, e também são apresentadas ideias para trabalhos futuros.

2 Trabalhos Relacionados

A Tabela 1 traz as características de cada trabalho relacionado, permitindo avaliar de forma mais objetiva as diferenças e similaridades. Algo fundamental a ser mencionado é a natureza dos dados utilizados. Os trabalhos que utilizam aprendizado de máquina e aprendizado profundo possuem dados ricos em atributos, onde cada registro representa um paciente ou exame com múltiplas variáveis clínicas e demográficas. Por outro lado, esta dissertação se alinha intimamente com os trabalhos de (KIM et al., 2023) e (BINDEL; ROLAND, 2025), que utiliza uma base de dados de série temporal agregada e univariada submetendo-as ao ARIMA e suas variantes. Dessa forma, os dados direcionaram qual tipo de abordagem seria seguida.

Modelos preditivos têm grande potencial para antecipar casos de resistência antimicrobiana e direcionar esforços para outros tratamentos de forma que a bactéria se mantenha ainda sensível aos medicamentos.

Esse potencial foi demonstrado em (PUURONEN SEPPÖ, 2008) que, por meio de *Ensemble Learning* com seleção de características (*Feature Selection*) locais para prever a sensibilidade de um patógeno com base em dados sobre o antibiótico, sobre o patógeno isolado e as características demográficas e clínicas do paciente. Esse método fez uso de janelas deslizantes com tamanho de 3 meses e deslocamento de 1 mês. Foram testadas janelas sobrepostas e não sobrepostas. Vários classificadores foram criados para execução ao passo que as janelas iam se deslocando.

O uso de *Feature Selection* com variáveis clínicas (dias na UTI, departamento, dias internado antes dos testes) e demográficas (sexo, idade) foi utilizado com *Naive Bayes*. Segundo (PUURONEN SEPPÖ, 2008), produziu precisões médias maiores em comparação com os cenários sem FS. O mesmo foi observado para Decision Trees e KNN, métodos de Aprendizado de Máquina utilizados neste mesmo trabalho para predição.

Willmann (2013) combinou o ARIMA com um método chamado Função de Transferência Multivariada e explorou a relação entre duas ou mais séries temporais. Um de seus resultados foi uma redução anual de 4% no uso de Meropenem ao longo de 3 anos, de 2009 a 2011. Essa pequena redução produziu uma atenuação de 62,5% na tendência crescente de *Pseudomonas aeruginosa* multirresistente, mostrando como a previsão de AMR pode dar suporte ao PGA.

O uso de Regressão Logística, KNN, Árvores de Decisão, *Random Forest* e Perceptron Multicamadas sobre variáveis demográficas, variáveis relacionadas a culturas de bactérias (bactéria, amostra), e variáveis relacionadas ao antibiograma (antibiótico, resultado, data dos resultados, entre outras), que segundo (AGUERO et al., 2019), detectou o

surgimento de resistência para avançar medidas preventivas, isolamentos e reduzir taxas de AMR. Foram usadas duas abordagens de FS: Uma baseada em conhecimento de equipe médica e outra por informações mútuas. Os melhores resultados (acurácia, especificidade, sensibilidade e f1-score) de cada família de antibiótico na maioria dos casos vieram da *Feature Selection* por informações mútuas.

Outra abordagem estatística bastante utilizada é a Regressão Logística, que em (YELIN et al., 2019) foi utilizada para construir um modelo preditivo. Além da RL, *Gradient Boosting* também foi utilizado, demonstrado nos resultados fortes associações de AMR com informações demográficas (idade, sexo, origem clínica, comorbidades, entre outras) dos pacientes, cultura de urina e histórico de compra de medicamentos.

Lewin-Epstein 2021 utilizou este dados de antibiogramas dos pacientes, ou seja, registros eletrônicos pré-processados, para prever resistência bacteriana utilizando regressão logística, redes neurais e árvores de gradiente aumentado. Seu foco foi prever a resistência aos cinco antibióticos mais comumente testados quanto à resistência no conjunto de dados: ceftazidima, gentamicina, imipenem, ofloxacina e sulfamethoxazole-trimethoprim, tendo em seus resultados uma área abaixo da curva ROC entre 0,8 e 0,88.

Kim 2023 comparou ARMA, ARIMA e SARIMA para prever a porcentagem trimestral de resistência bacteriana. A proporção entre registros sensíveis e resistentes foi calculada e esses três métodos foram utilizados para prever a resistência do trimestre seguinte, obtendo uma tendência futura prevista em que o RMSE foi calculado para todos os algoritmos e produzindo um erro baixo em alguns casos como para a bactéria *Pasteurella multocida*, *Streptococcus suis* e *Bordetella bronchiseptic*.

A comparação entre algoritmos de aprendizado de máquina e algoritmos estatísticos também pode ser feita para escolha do modelo que melhor se encaixa na realidade dos dados estudados. Segundo (SAKAGIANNI et al., 2023), prever AMR de forma rápida e eficaz é crucial para o sucesso do tratamento clínico. O estudo avaliou cinco modelos de aprendizado de máquina e regressão para prever a suscetibilidade antimicrobiana e direcionar a equipe médica para o melhor tratamento. Os modelos utilizados foram: J48; Random Forest; Regressão Logística, KNN e Perceptron Multi-Camadas. O modelo KNN obteve os melhores resultados, com F1 Score de 0,724 e uma área abaixo da ROC de 0,758.

Também é importante considerar a limitação dos dados em relação a quantidade (número de exames) e qualidade (dados importantes ausentes da série, como idade, sexo e comorbidades, lacunas temporais como dias ou meses sem exames). Como mencionado anteriormente, (KIM et al., 2023) organizou a série temporal trimestralmente para superar a falta de dados em alguns meses, conseguindo extrair de maneira mais confiável a porcentagem de resistência.

Além disso, os níveis de resistência mudam em cada hospital. Uma bactéria pode

ser mais resistente a determinado antibiótico em um hospital e menos em outro, porque o universo do hospital varia. Dessa forma, o tipo de série temporal também varia, sendo necessário também analisar seus variados componentes, como a aleatoriedade, pois isso vai direcionar o estudo.

Análises levando em consideração uma série de caminhada aleatória foram feitas em (BINDEL; ROLAND, 2025), onde foi utilizado modelos ARIMA para previsão de longo prazo em resistência a alguns antimicrobianos, entre eles amoxicilina, clindamicina, azitromicina, nitrofurantoína, ciprofloxacina, penicilina e doxiciclina.

Foi também necessário um modelo de caminhada aleatória, $ARIMA(0,1,0)$, que não possui os componentes autorregressivo e de média móvel, apenas a diferenciação de primeira ordem. Em séries de caminhada aleatória, valores futuros dependem do valor atual e de um termo de erro.

A Tabela 1 traz as características de cada trabalho mencionado, permitindo avaliar de forma mais objetiva as diferenças e similaridades. Algo fundamental a ser mencionado é a natureza dos dados utilizados. Os trabalhos que utilizam aprendizado de máquina e aprendizado profundo possuem dados ricos em atributos, onde cada registro representa um paciente ou exame com múltiplas variáveis clínicas e demográficas. Por outro lado, esta dissertação se alinha intimamente com os trabalhos de (KIM et al., 2023) e (BINDEL; ROLAND, 2025), que utiliza uma base de dados de série temporal agregada e univariada submetendo-as ao ARIMA e suas variantes.

Tabela 1 – Detalhamento dos Trabalhos Relacionados.

Trabalho	Base de Dados	Técnicas Utilizadas	Tarefa
(PURONEN SEPO, 2008)	Variáveis clínicas e demográficas	Ensemble Learning	Predição
(WILLMANN et al., 2013)	Variáveis clínicas, demográficas e antibiogramas	ARIMA	Predição
(AGUIERO et al., 2019)	Variáveis clínicas, demográficas e antibiogramas	KNN, RL, RF, Perceptron Multilayer	Predição e Classificação
(YELIN et al., 2019)	Variáveis clínicas e demográficas	RL, Gradient Boosting	Predição
(LEWIN-EPSTEIN et al., 2021)	Variáveis clínicas, demográficas e antibiogramas	RL, NN, GradientBoosting Tree	Predição
(KIM et al., 2023)	Taxas de resistência e MIC	ARMA, ARIMA, SARIMA	Predição
(SAKAGIANNI et al., 2023)	Variáveis clínicas, demográficas e antibiogramas	J48, RF, RL, KNN, Perceptron Multilayer	Predição
(BINDEL; ROLAND, 2025)	Consumo diário de antibióticos	ARIMA	Predição

3 Referencial Teórico

Neste capítulo serão apresentados os conceitos fundamentais que sustentam esta dissertação. A Seção 3.1 aborda os antibióticos e suas classificações. A Seção 3.2 detalha as características das bactérias, com foco naquelas relevantes para a ação dos antimicrobianos. A Seção 3.3 explora a Resistência Antimicrobiana (AMR), seus mecanismos e seu impacto global. As Seções 3.4, 3.5 e 3.6 estabelecem a base estatística do trabalho, introduzindo os conceitos de Séries Temporais, Análise Estatística aplicada à AMR e o modelo de Caminhada Aleatória como um caso particular do ARIMA.

3.1 Antibióticos

Os antibióticos representam uma das mais importantes descobertas da medicina no século XX, sendo agentes terapêuticos capazes de combater infecções bacterianas. A sua ação principal foca na destruição ou inibição do crescimento da célula bacteriana, podendo ser classificados em duas categorias principais: bactericidas e bacteriostáticos. Os agentes bactericidas são aqueles que causam a morte do patógeno, enquanto os bacteriostáticos apenas impedem sua multiplicação, permitindo que o sistema imunológico do hospedeiro elimine a infecção.

Para quantificar a eficácia de um antibiótico, são utilizadas duas métricas principais: a Concentração Inibitória Mínima (MIC) e a Concentração Bactericida Mínima (CBM). O MIC é definido como a menor concentração de um antibiótico que inibe o crescimento visível de uma bactéria após 24 horas de incubação ([PATEL et al., 2025](#)). Já o CBM corresponde à menor concentração capaz de reduzir a densidade bacteriana em 99,9% (ou 1000 vezes) no mesmo período ([PANKEY; SABATH, 2004](#)).

Embora seu papel na saúde humana seja fundamental, o uso inadequado de antibióticos pode levar não apenas ao desenvolvimento de resistência antimicrobiana, mas também a reações adversas nos pacientes. Estima-se que um em cada cinco pacientes hospitalizados desenvolva algum efeito colateral relacionado ao uso de antibióticos, o que reforça a necessidade de uma prescrição criteriosa e monitorada ([TAMMA et al., 2017](#)). Tais reações podem afetar múltiplos sistemas, incluindo renal, dermatológico, cardíaco, hematológico e neurológico. Por essa razão, a administração de altas doses na tentativa de eliminar rapidamente um patógeno deve ser evitada, buscando sempre um equilíbrio entre eficácia e segurança.

3.2 Bactérias

As bactérias são microrganismos unicelulares e procariontes ([NHGRI, 2022](#)), o que significa que seu material genético não está contido em um núcleo delimitado por membrana. Elas são os principais representantes do reino Monera e exibem uma vasta diversidade metabólica e estrutural.

Embora muitas espécies sejam conhecidas por causar doenças, a maioria das bactérias não é patogênica para os seres humanos. Pelo contrário, muitas desempenham papéis vitais, como as que compõem a microbiota do sistema digestivo, auxiliando na absorção de nutrientes, e as que habitam a pele, protegendo o sistema imunológico contra a invasão de bactérias prejudiciais ([CLINIC, 2022](#)).

Uma das classificações mais importantes na bacteriologia é baseada na estrutura da parede celular, através da coloração de Gram ([KRISTENSEN et al., 2023](#)).

- Bactérias Gram-positivas: Possuem uma espessa camada de peptidoglicano em sua parede celular, que retém o corante violeta de cristal, conferindo-lhes uma coloração roxa;
- Bactérias Gram-negativas: Possuem uma parede celular mais complexa, com uma fina camada de peptidoglicano e uma membrana externa adicional, que impede a retenção do corante inicial, fazendo com que se coram de vermelho ou rosa após a aplicação de um contra-corante.

Essa diferença estrutural é crucial, pois a membrana externa das bactérias Gram-negativas atua como uma barreira seletiva, dificultando a penetração de muitos antibióticos e tornando-as naturalmente mais resistentes a certas classes de medicamentos.

3.3 Resistência Antimicrobiana

Existem dois tipos de resistência bacteriana segundo ([REYGAERT, 2018](#)):

- Resistência Natural - Pode ser Induzida ou Intrínseca. Na induzida, os genes ocorrem naturalmente na bactéria, mas só são expressos após a exposição a um antibiótico. A resistência intrínseca é quando uma espécie bacteriana já possui resistência e é independente da exposição prévia a antibióticos. Não está relacionada à transferência horizontal de genes.
- Resistência Adquirida - É quando uma bactéria que, anteriormente à sua mutação genética, era sensível a um antibiótico passa a não sentir mais efeitos desse mesmo

medicamento. Ela pode ser adquirida por meio de mutações genéticas espontâneas ou transferência de genes de um microrganismo para outro.

A Figura 2 ilustra esse processo de seleção e multiplicação de bactérias resistentes:

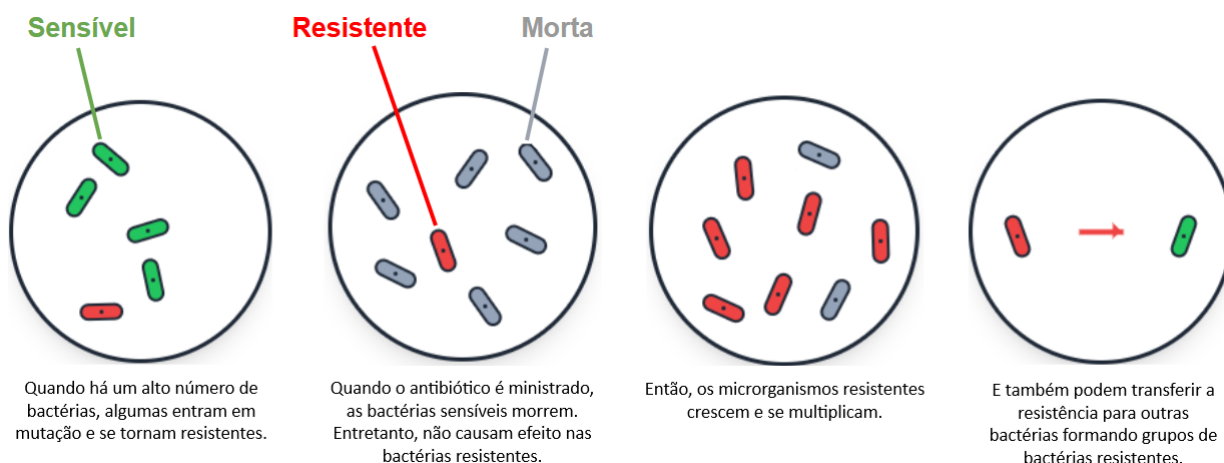


Figura 2 – Como acontece a resistência bacteriana. Fonte: Elaborada pelo autor.

Na resistência adquirida o microorganismo sofre mutações que o torna inviável de ser combatido através do mesmo antibiótico utilizado e poucos são os que terão algum efeito positivo contra o patógeno. Essas mutações acontecem por meio de alterações no nucleotídeo, moléculas orgânicas que formam o DNA (ácido desoxirribonucleico), e são classificadas de acordo com sua alteração.

A resistência adquirida pode ser gerada de duas formas (ZHENG et al., 2020):

- Com substituição de um nucleotídeo: Envolve a substituição de uma única base do DNA, fazendo com que, na replicação, o material genético seja disseminado já com uma alteração permanente e com chances de produzir resistência.
- Deleção ou adição de nucleotídeo: durante a replicação do DNA, um nucleotídeo é totalmente removido ou incluído.

Além disso, existem definições que caracterizam os níveis de resistência (MAGIO-RAKOS et al., 2012):

- MDR: Os microrganismos multidroga resistentes são aqueles que possuem resistência a mais de um agente antimicrobiano. Existem algumas controvérsias sobre esses microrganismos e outras definições, como a resistência a três ou mais classes de antimicrobianos;

- XDR: Os Extremamente Resistentes (ou extensivamente resistentes a drogas) são organismos resistentes não apenas a múltiplos agentes antimicrobianos, mas também a alta probabilidade de resistência a quase todos os ATBs aprovados.
- PDR: Patógenos classificados como Pan-droga resistente (Pan vem de todos) são aqueles que possuem resistência a todas as classes de antimicrobianos testados rotineiramente.

3.3.1 Classificação de Ambler

Os antibióticos Beta-Lactâmicos constituem um grande arsenal antimicrobiano. Em bactérias Gram-negativas enzimas Beta-Lactamases são armas de defesa das bactérias contra esses medicamentos.

Proposta por Richard P. Ambler, a classificação de Ambler é um sistema de classificação para as enzimas beta-lactamases, dividindo-as em quatro classes: A, B, C e D.

Segundo (BUSH K., 2010), enquanto que as classes A, C e D quebram o antibiótico utilizando parte de sua estrutura denominada serina, a classe B utiliza um íon de zinco para executar essa hidrólise do Beta-Lactâmico.

As carbapenemases são tipos de Beta-Lactamases com grande relevância para o Hospital das Clínicas de Pernambuco. Resistentes aos carbapenêmicos (antibióticos de grande espectro) e MDR, esses microrganismos representam grande alerta no contexto hospitalar.

3.4 Séries Temporais

Uma série temporal é um conjunto de observações z registradas sobre algum índice cronológico, em tempos t regulares regulares ou não.

A análise de séries temporais é um campo da estatística que lida com dados coletados e ordenados ao longo do tempo, geralmente em intervalos equidistantes. Sua aplicação é vasta e, em saúde pública, constitui uma ferramenta poderosa para o monitoramento de doenças, a avaliação de intervenções e a previsão de tendências futuras, auxiliando na tomada de decisões e no planejamento de recursos.

Uma característica inerente a este tipo de dado é a dependência temporal (BOX et al., 2016), onde o valor de uma observação em um determinado momento está correlacionado com os valores de observações passadas. Isso significa que o passado da série contém informações valiosas que podem ser usadas para compreender e prever seu comportamento futuro. A Figura 3 exibe um exemplo gráfico de uma série temporal.

Um conceito central na análise de séries temporais é a estacionariedade. Uma série é considerada estacionária quando suas propriedades estatísticas, como média e variância, permanecem constantes ao longo do tempo. A maioria dos modelos estatísticos, incluindo o ARIMA, pressupõe que a série seja estacionária ou que possa ser transformada em uma. Séries não estacionárias, que exibem tendências ou mudanças na volatilidade, precisam ser tratadas, geralmente através de um processo de diferenciação, para que os modelos possam ser aplicados de forma confiável.

Há ainda dois conceitos fundamentais dentro do universo das séries temporais - as univariadas e as multivariadas.

As séries temporais univariadas são aquelas que observam uma única variável registrada em um intervalo de tempo. As séries temporais modeladas nesta pesquisa são um exemplo univariado, em que se observa a evolução da resistência dos pares Bactéria-Antibiótico ao longo do tempo. A variável a ser observada é apenas a proporção de resistência mensal.

Por outro lado, uma série temporal multivariada consiste em observações de duas ou mais variáveis simultaneamente ao longo do tempo. Este tipo de análise é mais complexa pois tenta prever os valores futuros de mais de uma variável. Um exemplo de uma análise multivariada é além de prever a proporção de resistência de um antibiótico, fosse também analisado o consumo mensal deste antibiótico para a evolução dessa resistência, não só para visualizar como a resistência evolui no futuro, mas também para observar o impacto do consumo mensal do antibiótico influencia nesse valor de resistência mensal. Esses dois cenários combinados formam uma série temporal multivariada.

3.5 Análise Estatística

As abordagens estatísticas são amplamente utilizadas para resolver problemas de regressão e classificação. Em muitos casos, a análise de tendência e sazonalidade são úteis para detecção de padrões dentro de séries temporais. Martínez 2020 utilizou séries temporais mensais para analisar tendências, sazonalidade e a associação entre o uso comunitário de antibióticos para detectar resistência antimicrobiana (AMR) em *Escherichia coli* e *Klebsiella pneumoniae* em infecções do trato urinário.

Técnicas como Regressão Logística e as variações ARMA, ARIMA e SARIMA podem realizar predições que são úteis para direcionar a equipe médica ao mostrar em uma curta janela futura o caminho que o tratamento está seguindo, se a resistência de determinado par bactéria-antibiótico vai aumentar ou diminuir nos próximos meses, alertando para um possível aumento da resistência e ajudando no tratamento não apenas daquele paciente, mas também evitando surto no ambiente hospitalar.

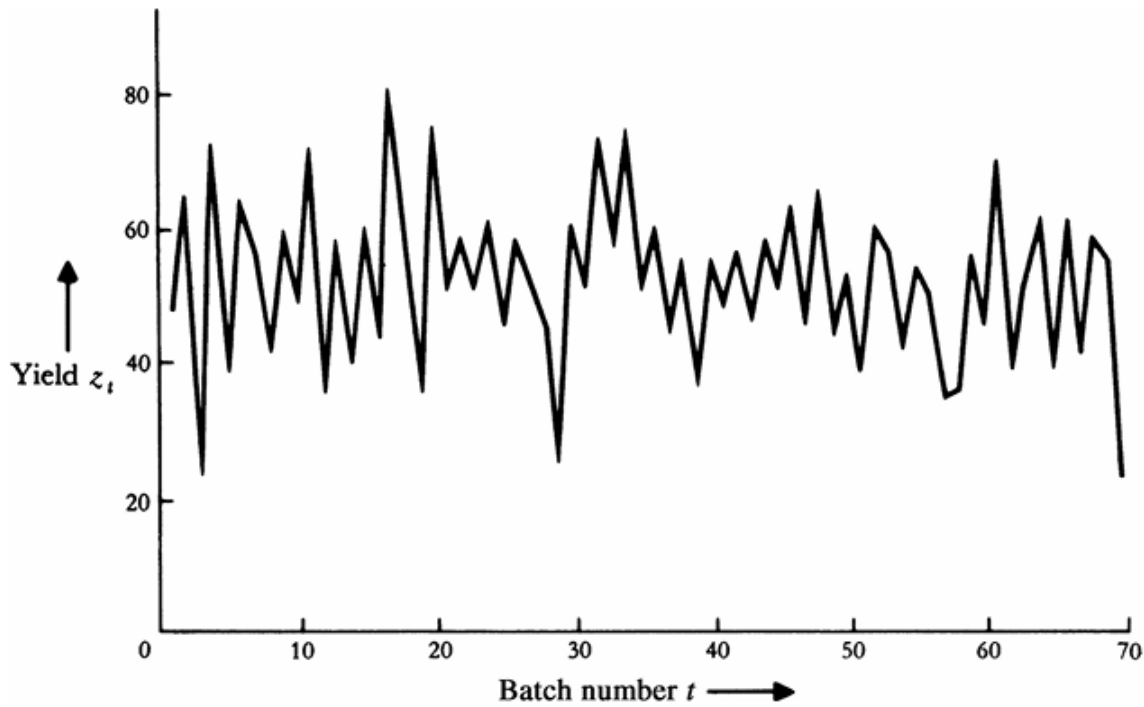


Figura 3 – Série temporal de um Lote e o seu Rendimento. Fonte: (BOX et al., 2016)

Em Kim 2023 ARIMA e suas variações foram usadas para previsão de resistência antimicrobiana em fazendas de suínos sobre duplas de bactéria-antibiótico. Tais previsões foram úteis para antecipar aumento na resistência bacteriana, uma vez que os resultados mostram se o uso de tal antibiótico está sendo eficaz uma vez que foi observado se a porcentagem de resistência futura aumentará ou diminuirá.

Por conta da versatilidade em operar vários componentes de séries temporais, Bindel e Roland 2025 utilizaram ARIMA para previsão de resistência bacteriana. Entre os modelos utilizados, o ARIMA (0,1,0), foi o mais adequado para capturar tendências em todos os medicamentos antibacterianos em algumas séries temporais de caminhada aleatória. As demais séries temporais utilizaram modelos com outros parâmetros colhidos por meio da submissão de funções de Autocorrelação e Autocorrelação Parcial.

3.6 Caminhada Aleatória

Um modelo de Caminhada Aleatória é um dos processos estocásticos mais fundamentais na análise de séries temporais. Ele descreve uma trajetória que consiste em uma sucessão de passos aleatórios. Matematicamente, um modelo de caminhada aleatória simples é definido pela equação:

$$Z_t = Z_{t-1} + a_t \quad (3.1)$$

onde Z_t é o valor da série no tempo t , Z_{t-1} é o valor no período imediatamente anterior, e a_t é um termo de erro aleatório (ruído branco) com média zero e variância constante. Isso significa que a melhor previsão para o valor de amanhã (Z_{t+1}) é simplesmente o valor de hoje (Z_t).

A principal característica de uma série de caminhada aleatória é que ela é não estacionária. No entanto, a sua primeira diferença, $\nabla Z_t = Z_t - Z_{t-1} = a_t$, é, por definição, um processo de ruído branco, que é estacionário.

Essa propriedade estabelece uma conexão direta com os modelos ARIMA. Um processo de caminhada aleatória é, na verdade, um caso especial de um modelo ARIMA, especificamente um modelo ARIMA(0,1,0):

- $p=0$: Não há componente autorregressivo (a série não depende de seus lags passados, apenas do valor mais recente);
- $d=1$: A série precisa de uma diferenciação para se tornar estacionária;
- $q=0$: Não há componente de média móvel (o erro não depende de erros passados).

Em estudos de previsão, o modelo de caminhada aleatória é frequentemente utilizado como um baseline ou modelo "ingênuo" de referência. Se um modelo mais complexo não consegue gerar previsões significativamente melhores do que as de um simples ARIMA(0,1,0), sua utilidade e complexidade adicional podem ser questionadas, como observado por (BINDEL; ROLAND, 2025) em sua análise.

4 Materiais e métodos

Este capítulo apresentará os materiais e métodos empregados no desenvolvimento do trabalho, isto é, o conjunto de dados, os métodos selecionados e as métricas adotadas para avaliação dos experimentos.

4.1 Fluxo do Trabalho

Primeiramente, o problema foi definido a partir das necessidades apontadas pelos especialistas do Hospital das Clínicas de Pernambuco, sendo estabelecido que seria adotada uma abordagem voltada à previsão da resistência antimicrobiana (AMR).

As etapas da metodologia empregada no trabalho estão representadas na Figura 4.

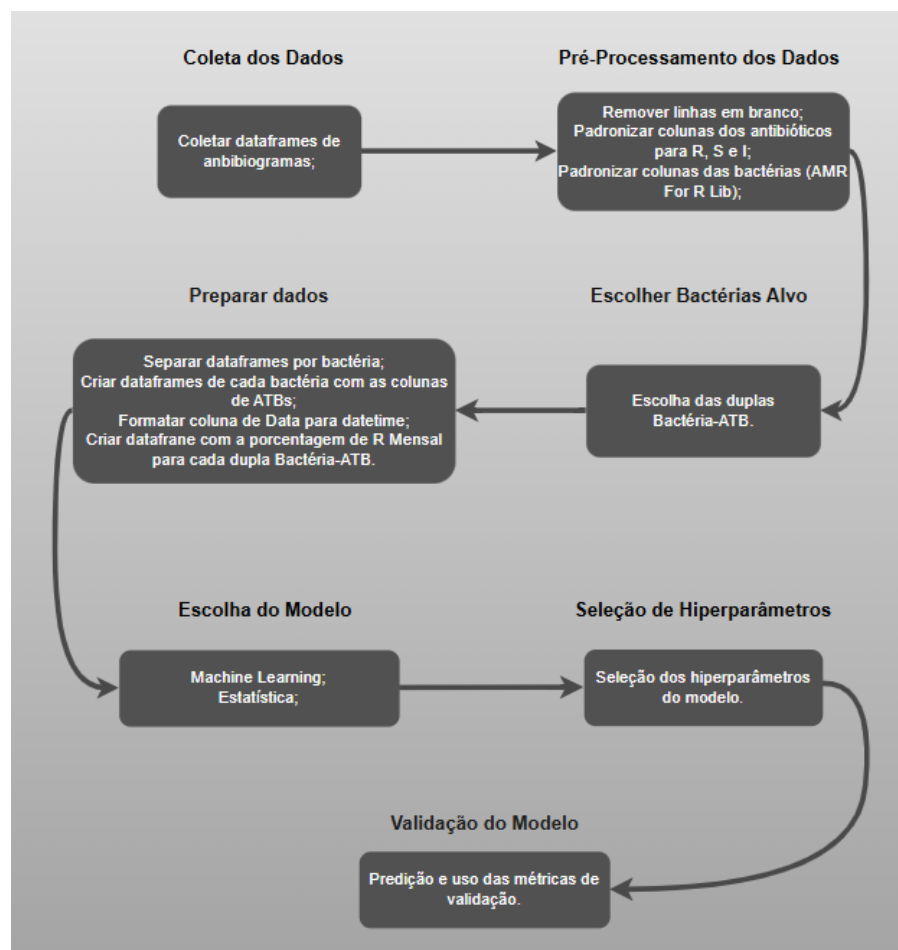


Figura 4 – Fluxograma de execução do trabalho. Fonte: Elaborada pelo autor

4.2 Base de Dados

Foi utilizado um conjunto de dados do HC-UFPE dos anos de 2013 à 2024, após pré-processamento. Esses dados são relativos a exames de antibiogramas realizados em pacientes do hospital, possuindo as seguintes informações:

- Variáveis clínicas - Origem clínica, data de internação, setor, código da clínica, leito, código da especialidade e exame.
- Variáveis relacionadas a culturas de bactérias - Bactéria, amostra;
- Variáveis relacionadas ao antibiograma - Antibiótico, MIC, resultado e data de internação.

Estes antibiogramas estão divididos em duas bases de dados principais oriundas de duas máquinas com sistemas distintos que fazem parte do laboratório de análises clínicas do hospital e que estão disponíveis para a Unidade de Vigilância em Saúde, responsável pela vigilância microbiológica do HC-UFPE. Inicialmente, antes do pré-processamento, ambas as bases possuíam juntas aproximadamente 19 mil linhas, incluindo as linhas em branco. No entanto, as duas bases apresentavam falta de padronização e não estavam estruturadas, contendo campos de preenchimento aberto inseridos manualmente e um grande número de dados faltantes, o que gerou considerável dificuldade para a organização e consolidação das informações.

Considerando nosso objetivo de contribuir para políticas eficazes de controle de infecção bacteriana, esta pesquisa propõe uma padronização inicial dos registros, estabelecendo a base para futuras análises e intervenções estratégicas.

4.2.1 Pacote R para resistência antimicrobiana

O *AMR Package For R* ([BERENDS et al., 2022](#)) é um pacote gratuito e de código aberto escrito em R que simplifica o processamento de diversas informações para pesquisas relacionadas à resistência antimicrobiana. Entre elas, a padronização de taxonomia de bactérias e fungos (reino, família, espécie), padronização de antibióticos, base de dados de antibiograma sintético, relação de 50 mil registros de resistência intrínseca, entre outras funcionalidades.

O Pacote R foi base para a publicação do artigo científico ([BERENDS et al., 2022](#)) e duas teses de doutorado: ([BERENDS, 2021](#)) e ([LUZ, 2021](#)). Além das funções de pré-processamento e base de dados, este pacote dispõe também de métodos para predição de resistência antimicrobiana baseado na regressão logística.

4.2.2 Pré-Processamento

As duas bases de dados utilizadas foram provenientes de duas máquinas distintas e não apresentam a mesma organização dos dados. Os antibiogramas coletados abarcam o período de 2013 até o final de 2024. Os dois datasets possuem mais de 9 mil linhas cada e estão organizados de maneira distinta, além disso, aproximadamente metade das linhas não puderam ser aproveitadas por estarem mal estruturadas, com dados faltantes ou impossíveis de serem interpretadas com confiabilidade, especialmente nos campos abertos.

Para se ter uma ideia da diferença entre estes conjuntos de dados podemos observar as colunas de cada um dos dois. O primeiro conjunto de dados (Master 762) possui as colunas descritas no [Apêndice I](#), enquanto que o segundo (Master 284) possui as colunas descritas no [Apêndice II](#).

Apesar de ambos os conjuntos de dados possuírem as primeiras colunas semelhantes (colunas em negrito), as colunas dos antibióticos e dos microrganismos (coluna BACTERIA é equivalente a MICROORGANISMO 1 e 2) têm uma apresentação distinta nos dois. Enquanto que no segundo as informações de antibiótico e perfil de resistência estão dispostos de maneira direta, onde o antibiótico é a coluna e o perfil de resistência R, S, I ou NT, como apresentado na Figura 5. No primeiro conjunto de dados as informações relativas aos micro-organismos e antibióticos estão dispostos como apresentado na Figura 6.

O	P	Q	R	S	T	U	V
BACTERIA	AMICACINA	AMPICILINA	AMPICILINA_S	AZTREONAM	CEFAZOLINA	CEFEPIMA	CEFTAZIDIMA
Staphylococcus aureus							
Proteus mirabilis	S	R	R		R	S	
Morganella morganii	S	R	R		R	S	
Klebsiella pneumoniae	S	R	R		R	S	
Klebsiella pneumoniae	I	R	R		R	S	
Micrococcus sp							
Acinetobacter baumannii	R		R			R	R
Escherichia coli	S	R	S		S	S	
Staphylococcus aureus							
Staphylococcus coagulase negativa		R					
Klebsiella pneumoniae	S	R	R		R	R	
Enterobacter cloacae	S	R	R		R	S	
Micrococcus							
Enterobacter cloacae (200.000 UFC/mL)	S	R	R		R	R	
Micrococcus							
Escherichia coli	Sensível	Resistente	Resistente		Resistente	Resistente	
Escherichia coli Ac. 1.000.000ufc/ml	S	R	S		R	S	
Escherichia coli Ac. 1.000.000ufc/ml	S	R	R		R		

Figura 5 – Formato da coluna Bactéria e Antibióticos - Dataset Mastertools 284. Fonte: Elaborada pelo autor.

Nessa organização dos dados as colunas "Droga" correspondem aos antibióticos; SIR são as colunas referentes aos perfis de resistência R, S ou I; CIM é a Concentração Inibitória Mínima, o valor mínimo de antibiótico administrado para que o tratamento seja eficaz, não prejudicando o paciente.

O	P	Q	R	S	T	U	V	W	X
MICROORGANISMO1	DROGA11	CIM11	SIR11	DROGA12	CIM12	SIR12	DROGA13	CIM13	SIR13
Pseudomonas aeruginosa (>1.000.000 UFC/mL)				Amicacina	<=8	S	Aztreonam	S	R
Staphylococcus epidermidis				Clindamicina	>4	R	Daptomicina	R	S
Klebsiella pneumoniae ssp ozae	Ampicilina	>16	R	Amicacina		16 S	Ampicilina-Sulbactam	S	R
Enterococcus faecalis	Ampicilina		1 S	Daptomicina		4 S	Estreptomicina-Sin	S	S
Escherichia coli (>1.000.000 UI)	Ampicilina	>16	R	Amicacina	<=8	S	Ampicilina-Sulbactam	S	R
Klebsiella pneumoniae ssp pneu	Ampicilina	>16	R	Amicacina	<=8	S	Ampicilina-Sulbactam	S	R
	Ampicilina	>16	R	Amicacina	<=8	S	Ampicilina-sulbactam	S	R

Figura 6 – Formato da coluna Bactéria e Antibióticos - Dataset Mastertools 762. Fonte: Elaborada pelo autor.

A estrutura de organização do segundo conjunto de dados (Master 284) mostrou-se mais adequada para a realização deste trabalho e, portanto, foi necessária a construção de *scripts* para reestruturar o primeiro conjunto (Master 762), permitindo a integração de ambos em um único conjunto.

Um terceiro conjunto de dados foi incorporado ao conjunto de dados integrado, e refere-se a antibiogramas realizados a partir da metade de 2022 e finalizando em dezembro de 2024. Apesar do curto intervalo de tempo, este conjunto possui uma grande quantidade de dados, sendo de 20 mil linhas a quantidade total de registros. Este número de culturas se deu principalmente pelo aumento progressivo dos pacientes regulados de outros serviços pela Secretaria de Saúde. Sendo esse aumento de 30% dos leitos na metade de 2023 e 50% na metade de 2024. Esse aumento de pacientes regulados tem repercussão direta na solicitação de culturas de vigilância.

A atualização no parque de máquinas e a alteração na rotina de análise das culturas tornaram os resultados mais céleres e detalhados. As análises que anteriormente não eram realizadas nos finais de semana e feriados passaram a ser realizadas diariamente e os microrganismos passaram a ser identificados até a espécie como também pela produção de carbapenemases. Tais mudanças refletiram no maior interesse dos profissionais em realizar a solicitação das culturas para melhor guiar a prescrição de antibióticos.

4.2.3 Integração das bases de dados

A melhor estrutura para os dados serem trabalhados nos modelos propostos foi o formato utilizado pelo dataset Mastertools 284 (Figura 5). Considerando isso, precisou-se de um esforço para converter o formato do Mastertools 762 (Figura 6) no 284.

A primeira etapa do pré-processamento dos dados envolveu a exclusão de todas as linhas em branco. Em seguida, precisou-se padronizar os nomes dos antibióticos. Para isso, foi utilizada a função `as.ab()` do Pacote R para AMR. São aceitas entradas como nome oficial ou subtexto, código ATC e ID do composto.

No Código 4.1 é apresentado um exemplo do uso da função `as.ab()`.

Código 4.1 – Conversão dos nomes de antibióticos para o código padrão

```
R> as.ab("Amoxicillin")
      Class <ab>
      [1] AMX

R> as.ab(c("Amoxil", "dispermox", "amox", "J01CA04"))
      Class <ab>
      [1] AMX AMX AMX AMX
```

A próxima etapa foi a padronização do nome dos microrganismos por conta do alto número de entradas com sintaxe incorreta ou falta de padronização.

Pelo fato de existirem bactérias com nomes similares, o método `as.mo()` da biblioteca AMR possui um método para calcular uma pontuação e classificar a saída com o maior grau de similaridade com a entrada. Esse método utiliza a Equação 4.1 para determinar esta similaridade.

$$m_{(x,n)} = \frac{l_n - 0.5 * \min(l_n, lev(x, n))}{l_n * p_n * k_n} \quad (4.1)$$

Onde:

- n é a taxonomia (gênero, espécie e subespécie);
- l_n é o tamanho de n ;
- lev é a função da distância de Levenshtein (1966), que contém qualquer inserção, deleção e substituição como 1 que é a mudança necessária para mudar x dentro de n ;
- p_n é o patógeno humano prevalente ao grupo de n , conforme descrito abaixo;
- k_n é a taxonomia do reino de n , definido como Bacteria = 1, Fungi = 2, Protozoa = 3, Archaea = 4, others = 5.;

A função gera a taxonomia do microorganismo de acordo com a estrutura observada na Figura 7. Alguns exemplos de execução da função podem ser observados no Código 4.2.

Código 4.2 – Conversão dos nomes dos microorganismos para a taxonomia padrão

```
R> as.mo("K. pneumoniae") - B_KLBSL_PNMN
R> as.mo("KLEPNE") - B_KLBSL_PNMN
R> as.mo("KLPN") - B_KLBSL_PNMN
```

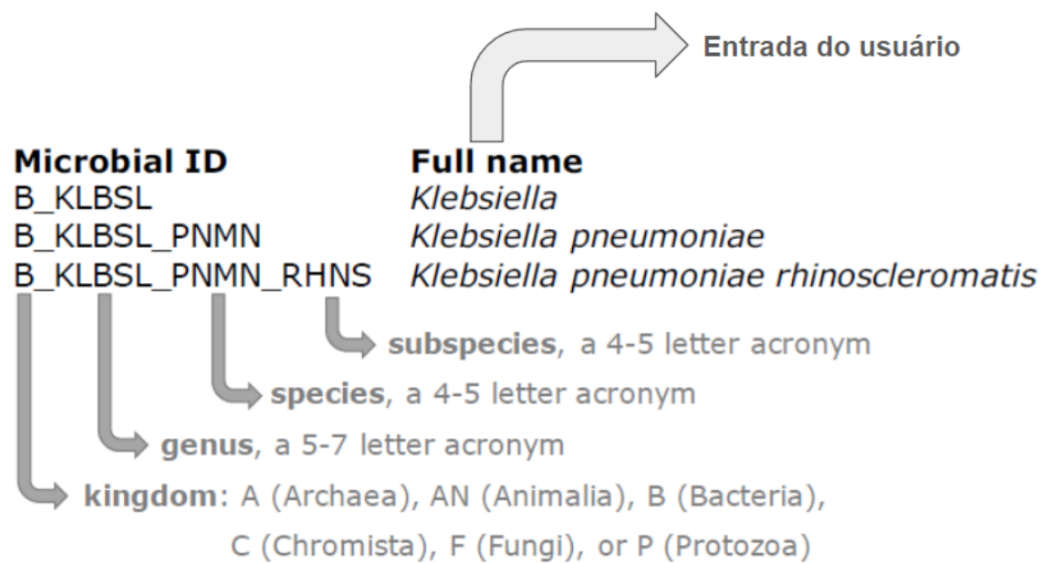


Figura 7 – A estrutura usada para a taxonomia de um microorganismo do pacote AMR. O ID consiste em iniciais separadas por underscore que são, respectivamente: o reino do microorganismo; o gênero; espécie; e subespécie. Fonte: (BERENDS et al., 2022).

A similaridade dos nomes dos microrganismos pôde ser superada através do método `as.mo()`. Um exemplo prático pode ser observado com a seguinte entrada:

A execução de `as.mo("E. coli")` retornará o ID referente à bactéria *Escherichia coli* ($m = 0.688$) que é uma bactéria mais comum em humanos, e não *Entamoeba coli* ($m = 0.381$) que é bem menos presente em seres humanos, mesmo que a entrada a princípio dificulte a diferenciação e que *Entamoeba* seja o primeiro em ordem alfabética.

Para testar a capacidade da biblioteca em determinar corretamente a taxonomia, fizemos 10 testes para cada bactéria. No Código 4.3 temos os testes para a *Escherichia coli* e a pudemos observar a mesma saída para todos. O mesmo foi feito para a *Klebsiella* e *Pseudomonas*, bem como para obter as siglas dos antibióticos e também padronizá-los.

Código 4.3 – Testes de Taxonomia da *Escherichia coli*

```
R> as.mo("Escherichia_coli") - B_ESCHR_COLI
R> as.mo("E_coli") - B_ESCHR_COLI
R> as.mo("E_coli") - B_ESCHR_COLI
R> as.mo("Eschericia_coli") - B_ESCHR_COLI
R> as.mo("Escherichia_colli") - B_ESCHR_COLI
R> as.mo("escherichia_coli") - B_ESCHR_COLI
R> as.mo("e-coli") - B_ESCHR_COLI
R> as.mo("Esche_col") - B_ESCHR_COLI
R> as.mo("Echherichia_coli") - B_ESCHR_COLI
R> as.mo("Escherichia-coli") - B_ESCHR_COLI
```

4.3 Seleção de Bactérias e Antibióticos

Após padronização dos dados, uma última etapa planejada inicialmente foi a correção de lacunas temporais que a série apresentava e que influencia na escolha dos parâmetros do ARIMA e, conseqüentemente, na predição.

O objetivo inicial seria prever a porcentagem de resistência semanalmente. Porém, devido a lacunas nos dados, muitas semanas, e também quinzenas, não tinham registros de resistência ou tinham apenas um registro que foi resistente, dois cenários indesejados se apresentaram pois resultam em 0% de resistência em algumas semanas e 100% de resistência em outras. Esses cenários extremos podem enviesar a predição.

Para superar essas limitações, uma solução foi utilizar proporções de AMR mensais para cada um dos microrganismos estudados, semelhante a (KIM et al., 2023) que utilizou dados trimestrais pelos mesmos motivos.

Além das lacunas temporais nos dados, um outro problema era a quantidade de culturas disponíveis. Uma vez que os dados foram agrupados em meses, o volume diminuiu bastante, comprometendo a significância estatística nas amostras de cada par ATB-Bactéria para estudo. Mesmo utilizando os microrganismos mais numerosos em exames, a quantidade de culturas foi muito inferior a encontrada nos trabalhos relacionados.

Isso pôde ser minimizado consultando a equipe médica especialista do Hospital das Clínicas que sugeriu agrupar por gênero dois dos três microrganismos de maior significância do hospital, sendo eles os patógenos do gênero *Pseudomonas* e *Klebsiella*. O resultado desse agrupamento pode ser visualizado na Tabela 2. A *Escherichia coli* não sofreu esse agrupamento com outros do seu gênero por causa do princípio ativo dos antibióticos e de não fazer sentido para o universo microbiológico do hospital.

Tabela 2 – Culturas de bactérias coletadas de maio de 2016 a dezembro de 2024.

Bactéria	Quantidade
Escherichia coli	4949
Klebsiella	4358
Pseudomonas	2300

Uma vez que a análise requer filtragem dada por pares específicos de Bactéria-Antibiótico, a quantidade de registros disponíveis para cada série temporal diminuiu consideravelmente. Diante deste desafio, através de interação com os especialistas do Hospital das Clínicas, foram realizados dois agrupamentos a nível taxonômico superior à

espécie, neste caso o gênero. Desta forma, a agregação foi feita sobre os gêneros *Klebsiella* e *Pseudomonas* para aumentar o número de observações e tornar a série temporal mais robusta. Sendo assim, foi importante ouvir os especialistas em infectologia do HC-UFPE a respeito do conjunto de antibióticos mais adequados para o tratamento de uma determinada bactéria. A Tabela 3 traz a recomendação dos especialistas para o estudo das três bactérias alvo com seus respectivos antibióticos.

Tabela 3 – Pares Antibiótico-Bactéria utilizados para análise.

Bactéria	Antibióticos
Escherichia coli	Ampicilina, Ceftriaxona, Gentamicina
Klebsiella	Cefepima, Gentamicina, Meropenem
Pseudomonas	Amicacina, Meropenem

O estudo foi conduzido em um período de tempo de 2016 a 2024. A seguir, as Figuras 8, 9 e 10 apresentam três das séries temporais de pares bactéria-antibiótico seguindo uma agregação mensal da resistência bacteriana. A proporção de resistência mensal foi calculada como demonstrado na Equação 4.2:

$$ProporçãoR = \frac{R}{R + S + I}$$

(4.2)

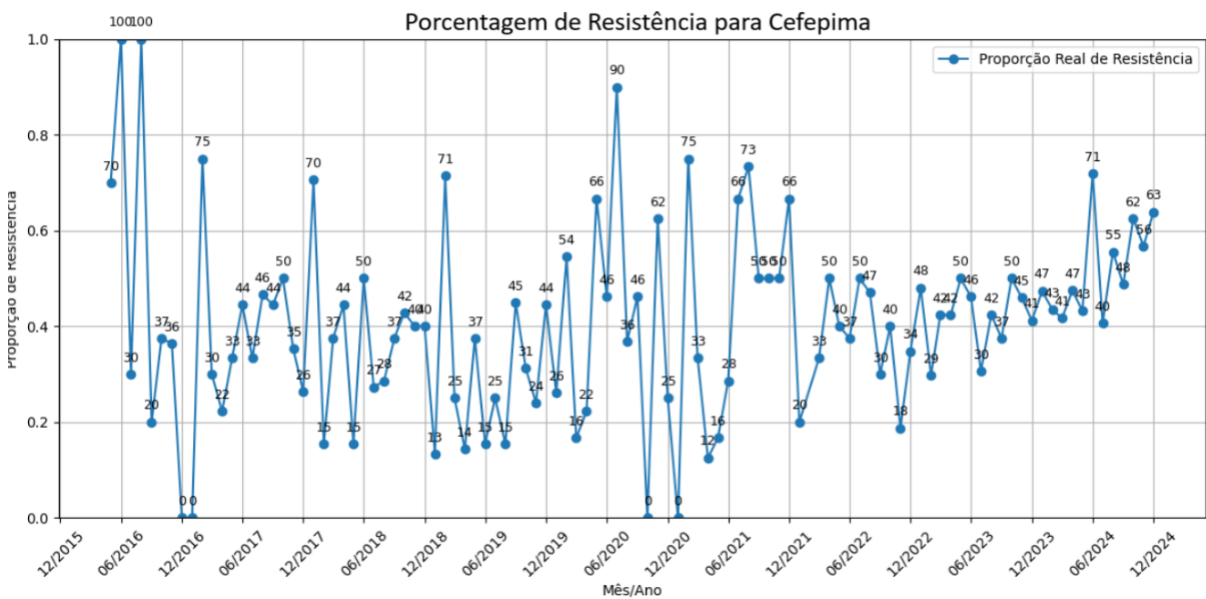


Figura 8 – Série temporal Klebsiella-Cefepima. Fonte: Elaborada pelo autor.

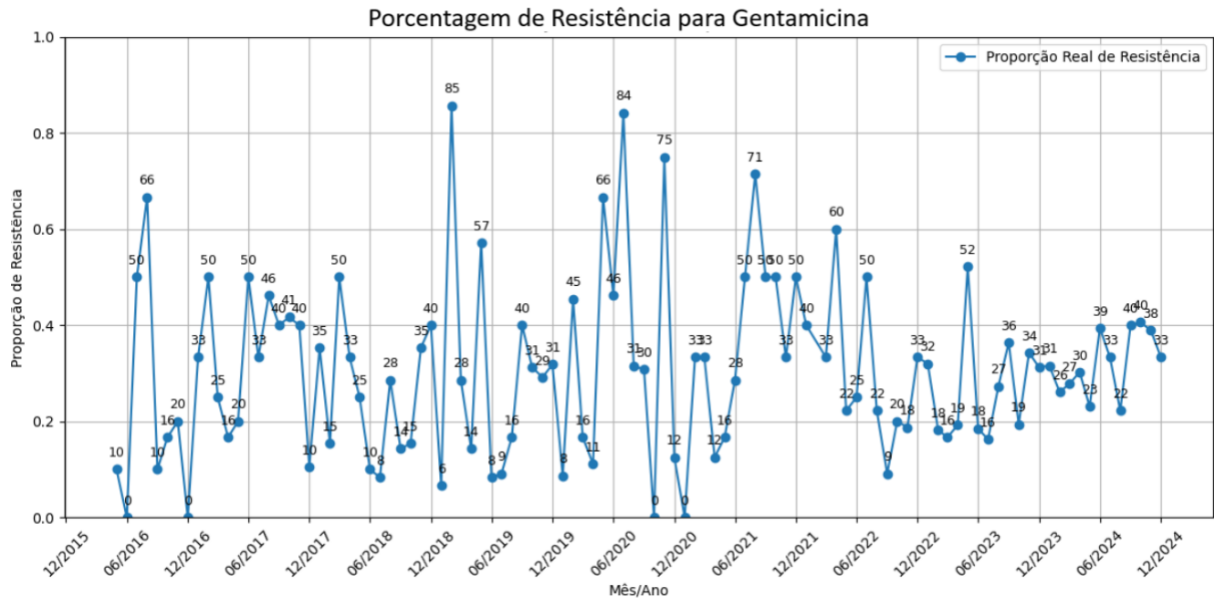


Figura 9 – Série temporal da Klebsiella-Gentamicina. Fonte: Elaborada pelo autor.

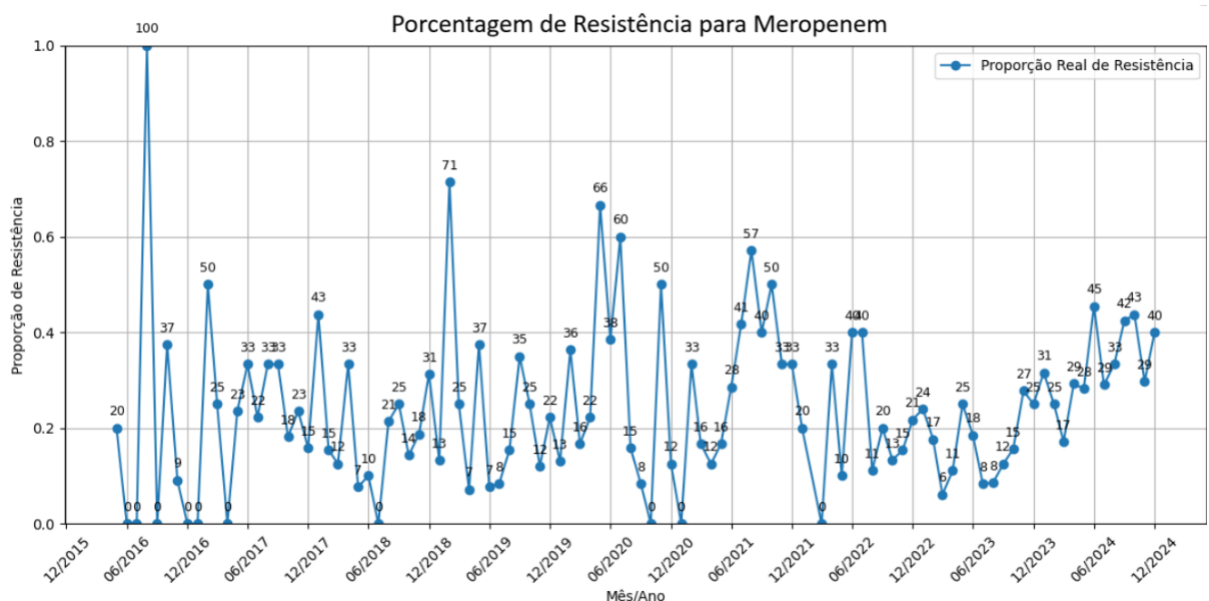


Figura 10 – Série temporal Klebsiella-Meropenem. Fonte: Elaborada pelo autor.

4.4 ARIMA

Esta seção apresenta uma explicação teórica do modelo autoregressivo integrado de média móvel. O problema atual é definido como um problema de regressão, em que a previsão será dada a partir de uma série temporal de valores reais e a saída serão valores reais. O ARIMA é uma extensão do método Auto Regressivo (AR) e de Média Móvel (MA), com um adicional de que pode ser utilizado em cenários que não apresentam estacionariedade, através de seu componente Integrado (I). Por isso, é de suma importância avaliar se os componentes de uma série são invariantes no tempo, pois a premissa de

estacionariedade é fundamental para a modelagem, como mostrado em (STRAHLBERG, 2021).

Matematicamente um modelo ARIMA pode ser descrito da seguinte forma (4.3):

$$\phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \quad (4.3)$$

- ϕ - Componente Autoregressivo (de ordem p);
- B - Operador de defasagem;
- ∇ - Operador de diferenciação aplicado d vezes;
- θ - Operador de média móvel (de ordem q);
- z_t - Série temporal original no tempo t ;
- α_t - Termo de erro (ruído branco no tempo t);
- θ_0 - Constante de deriva (*drift*).

A capacidade preditiva do modelo ARIMA reside na sua arquitetura híbrida, que combina dois componentes estatísticos fundamentais: o componente Autorregressivo (AR) e o de Média Móvel (MA). Durante a etapa de treinamento, o algoritmo utiliza os dados históricos para estimar os coeficientes em que esses dois componentes se baseiam, criando um modelo que aprende tanto com os valores passados da própria série quanto com seus próprios erros de previsão passados.

O componente Autorregressivo (AR), cuja ordem é definida pelo parâmetro p , tem como fundamento a ideia de que o valor atual de uma série temporal pode ser modelado como uma função linear de seus valores passados (lags). A teoria para este componente, detalhada em (BOX et al., 2016), modela a série estacionária após a diferenciação como:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t \quad (4.4)$$

A ordem p define exatamente quantos períodos anteriores são utilizados nesse cálculo. Por exemplo, um modelo com $p=3$ observará os três últimos valores da série para entender seu comportamento, sendo este a tendência de subida ou de queda dos valores. Em essência, este parâmetro observa a tendência natural. Se a resistência microbiana apresentou uma tendência de aumento nos últimos meses, o componente AR tentará projetar essa continuidade. Os parâmetros estimados para esta parte do modelo são os coeficientes autorregressivos, denotados por *phi* ($\phi_1, \phi_2, \dots, \phi_p$) na Equação(4.4). Cada coeficiente ϕ representa o peso ou a força da influência de um valor passado específico na previsão do valor atual.

Por outro lado, o componente de Média Móvel (MA), cuja ordem é definida pelo parâmetro q , executa um mecanismo de autocorreção aprimorado. Em vez de olhar para os valores passados da série, este componente foca na memória dos erros de previsão recentes, pressupondo que o valor atual é influenciado pelos erros de previsão passados. Nesta formulação, $\tilde{z}_t$ representa a série temporal Z_t centralizada, ou seja, $Z_t - \bar{Z}$ ((BOX et al., 2016) Capítulo 3), onde $\bar{Z}$ é a média da série. Sua fórmula é:

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (4.5)$$

O processo de ajuste do ARIMA envolve a previsão dos próprios valores conhecidos da série submetida como conjunto de treino. Ele os prevê e consegue calcular o erro (α_t) em cada ponto do tempo com base nos valores reais e os utiliza para ajustar a previsão atual. Este componente é crucial para modelar choques ou eventos inesperados que não são explicados apenas pela tendência da série. Os parâmetros estimados são seções de choque do processo para a absorção de eventos aleatórios, denotados por theta ($\theta_1, \theta_2, \dots, \theta_q$) na Equação (4.5).

A força do modelo ARIMA está na combinação de ambos os conceitos AR e MA para uma série estacionária. A equação completa que unifica os dois componentes pode ser escrita da seguinte forma após a série diferenciada:

$$z'_t = c + \phi_1 z'_{t-1} + \dots + \phi_p z'_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (4.6)$$

4.4.1 Estimativa dos Coeficientes ϕ e θ

Ambos os parâmetros ϕ e θ são fundamentais para a construção do modelo. Eles não são calculados por uma fórmula direta, mas sim estimados a partir dos dados de treino através de um processo de otimização numérica. O Capítulo 7 do livro (BOX et al., 2016) é inteiramente dedicado a este processo de estimação de coeficientes. A meta é encontrar a combinação que resulta no modelo que melhor se ajusta aos dados observados. Para isso, duas abordagens teóricas são centrais: a de Mínimos Quadrados e a de Máxima Verossimilhança.

A abordagem de Mínimos Quadrados (Least Squares) é encontrar os valores de ambos os coeficientes que minimizam a soma dos quadrados dos resíduos. Os resíduos (a_t) são os erros de previsão de um passo à frente que o modelo teria cometido dentro do conjunto de treino. A função a ser minimizada, chamada de Função de Soma dos Quadrados, é formalmente definida como:

$$S(\phi, \theta) = \sum a_t^2 \quad (4.7)$$

A função de Verossimilhança (Likelihood Function), por sua vez, assume que os erros do modelo (a_t) seguem uma distribuição de probabilidade (normal). Ela calcula a probabilidade de ter observado exatamente os dados que foram coletados, dado um conjunto de parâmetros. O método de Máxima Verossimilhança busca encontrar os valores dos coeficientes que maximizam essa probabilidade.

Em um modelo ARIMA, uma vez que os erros são assumidos como normalmente distribuídos, maximizar a função de verossimilhança ou minimizar a soma dos quadrados dos resíduos são abordagens equivalentes.

4.4.2 Estacionariedade de Séries Temporais

Um dos pressupostos fundamentais para a estimação confiável dos componentes Autorregressivo (AR) e de Média Móvel (MA) de um modelo ARIMA é que a série temporal analisada seja estacionária. Uma série é considerada estacionária quando suas propriedades estatísticas, notadamente a média e a variância, são constantes ao longo do tempo. Contudo, a maioria das séries temporais observadas no mundo real, como as de resistência microbiana, são não-estacionárias, frequentemente exibindo tendências claras de aumento ou diminuição, ou mudanças em sua volatilidade.

A modelagem de uma série não-estacionária com um modelo que assume estacionariedade pode levar a parâmetros enviesados, intervalos de confiança não confiáveis e, em última análise, a previsões de baixa acurácia. Para contornar essa questão, o modelo ARIMA incorpora o componente I (Integrado), que corresponde à aplicação da diferenciação para transformar a série original não-estacionária em uma nova série estacionária.

A ordem do componente integrado, denotada pelo parâmetro d , indica quantas vezes a operação de diferenciação precisa ser aplicada. Para a maioria das séries, até duas diferenciações ($d=1$) é suficiente. Esta operação consiste em subtrair de cada observação o valor da observação anterior, criando uma nova série que representa a variação entre os períodos. Matematicamente, a série original z_t é transformada na série diferenciada w_t é dada por (4.8):

$$w_t = (1 - B)^d z_t = \nabla^d z_t \quad (4.8)$$

Substituindo na equação geral (4.3) a diferenciação:

$$\phi(B)w_t = \theta_0 + \theta(B)a_t \quad (4.9)$$

$\phi(B)$ é o operador autorregressivo, enquanto que $\theta(B)$ é o operador de média móvel. Eles usam o operador de defasagem B , mas aplicado ao tempo. Podemos decompor $\phi(B)$

e $\theta(B)$ respectivamente como:

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) \quad (4.10)$$

$$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) \quad (4.11)$$

Juntando ambos os componentes na fórmula geral temos:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)w_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)a_t \quad (4.12)$$

$$w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} - \dots = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \dots \quad (4.13)$$

O que depois de diferenciada a série nos leva de volta para a Equação (4.6) do modelo ARMA:

$$w_t = \theta_0 + \phi_1 w_{t-1} + \phi_2 w_{t-2} + \dots + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \dots \quad (4.14)$$

É crucial aplicar a diferenciação apenas o número de vezes estritamente necessário para alcançar a estacionariedade. A aplicação de diferenciações adicionais a uma série que já é estacionária, um processo conhecido como sobrediferenciação, é contraproducente e prejudicial ao modelo por duas razões principais:

- **Introdução de Correlação Artificial:** A sobrediferenciação introduz uma estrutura de dependência serial que não existe nos dados originais. Por exemplo, se a primeira diferença de uma série (w_t) já resulta em ruído branco (um processo perfeitamente estacionário e aleatório), diferenciá-la novamente criará um processo de Média Móvel de ordem 1 (MA(1)), tornando a estrutura dos dados artificialmente mais complexa.
- **Aumento da Complexidade do Modelo:** Consequentemente, o modelo final para a série original se torna desnecessariamente mais complexo (ex: um ARIMA(0,2,1) em vez de um ARIMA(0,1,0)). Modelos mais complexos são mais difíceis de estimar e, seguindo o princípio da parcimônia, devem ser evitados se um modelo mais simples for adequado. A sobrediferenciação pode também levar a problemas de não-invertibilidade no componente MA, o que causa instabilidade e viés na estimação dos parâmetros.

Para evitar a armadilha da sobrediferenciação e determinar empiricamente a ordem correta de d , são empregados testes estatísticos formais. O Teste de Dickey-Fuller

Aumentado (ZHENG et al., 2020) e o Teste Kwiatkowski-Phillips-Schmidt-Shin (HERNANDEZ et al., 2021) são utilizados em conjunto para fornecer uma avaliação robusta da estacionariedade da série, garantindo que a diferenciação seja aplicada de forma necessária e suficiente.

4.4.3 Escolha de Parâmetros

Como explicado da seção anterior, resumidamente os parâmetros do ARIMA seguem as definições:

- p : Ordem da Auto Regressão;
- d : Quantas vezes a série temporal precisa ser diferenciada para se tornar estacionária;
- q : Ordem da Média Móvel.

A estimativa de parâmetros é crucial para que o algoritmo possa ser executado corretamente. (BOX et al., 2016) traz o conceito de parcimônia, defendendo que um modelo precisa ter o menor número possível de parâmetros para representar os dados. Um modelo com parâmetros em excesso pode realizar previsões de baixa qualidade e menos confiável. Existem algumas técnicas para que essa estimativa seja a mais precisa o possível.

4.4.3.1 Teste Dickey-Fuller

A primeira escolha é a da diferenciação (parâmetro d). É necessário saber quantas vezes a série temporal deve ser diferenciada para ser de fato executada pelo ARIMA. Isso é feito através do teste de Dickey-Fuller Aumentado (ZHENG et al., 2020).

O teste ADF é uma categoria de testes chamada de Teste de Raiz Unitária, onde caso exista uma raiz unitária para a série temporal em questão, esse conjunto de dados é não estacionário. Esses possuem duas hipóteses:

- Hipótese Nula: A série temporal possui raiz unitária (não é estacionária);
- Hipótese Alternativa: A série temporal não possui raiz unitária (é estacionária).

Para isso, existe um valor de probabilidade chamado p -value que é obtido no teste ADF e determina se a hipótese nula pode ser aceita ou rejeitada. O p -value para que a hipótese nula seja rejeitada deve ser menor do que 0,05 (para um nível de confiança de 95%) e considerar a série estacionária.

4.4.3.2 Teste KPSS

Uma vez que a ausência de raiz unitária não é uma prova completa de não estacionariedade, mas uma forte evidência, uma outra forma de atestar melhor a hipótese nula, segundo (HERNANDEZ et al., 2021), é o teste Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

O KPSS é um método útil para a análise de séries temporais para determinar se existe estacionariedade, sendo semelhante ao teste ADF, mas com hipóteses inversas. O teste KPSS inicia presumindo que a série temporal é estacionária e busca evidências para rejeitá-la através da ausência de uma raiz unitária:

- Hipótese Nula: A série temporal é estacionária e não possui raiz unitária;
- Hipótese Alternativa: A série temporal não é estacionária e possui raiz unitária. Se o valor de probabilidade (p -value) for maior do que o nível de significância 0,05, significa que a série é estacionária.

4.4.3.3 Autocorrelação e Autocorrelação Parcial

Os parâmetros p e q , que condizem respectivamente à Auto Regressão e Média Móvel, segundo (LOBO ASHA KAMATH, 2021), podem ser identificados calculando as funções de Autocorrelação Parcial (PACF) e Autocorrelação (ACF), usados para selecionar um atraso máximo dos termos AR e MA, respectivamente.

A Figura 11 representa os gráficos produzidos pelo ACF e PACF do antibiótico Amicacina e a bactéria *Escherichia coli* que ajudam a identificar o número de atrasos necessários para definir os possíveis valores de p e q .

Considerando o cenário em que tanto informações importantes dos pacientes como sexo, idade e comorbidades, como informações clínicas como o MIC, não estão disponíveis na base de dados, e devido a baixa quantidade de dados, para um estudo com técnicas mais robustas, que também séries temporais mostraram ser alternativas na previsão de resistência microbiana, como em (GUO et al., 2019) e (WANG et al., 2018), e que o ARIMA é uma técnica amplamente utilizada no estudo de séries temporais e previsão AMR (LOBO ASHA KAMATH, 2021), sua escolha foi se justificando ao longo da pesquisa ao passo que esses motivos foram sendo descobertos.

4.4.4 Validação dos Parâmetros

Essa etapa é importante para levantar a certeza de que a escolha dos parâmetros está correta e o modelo não está subajustado ou sobreajustado.

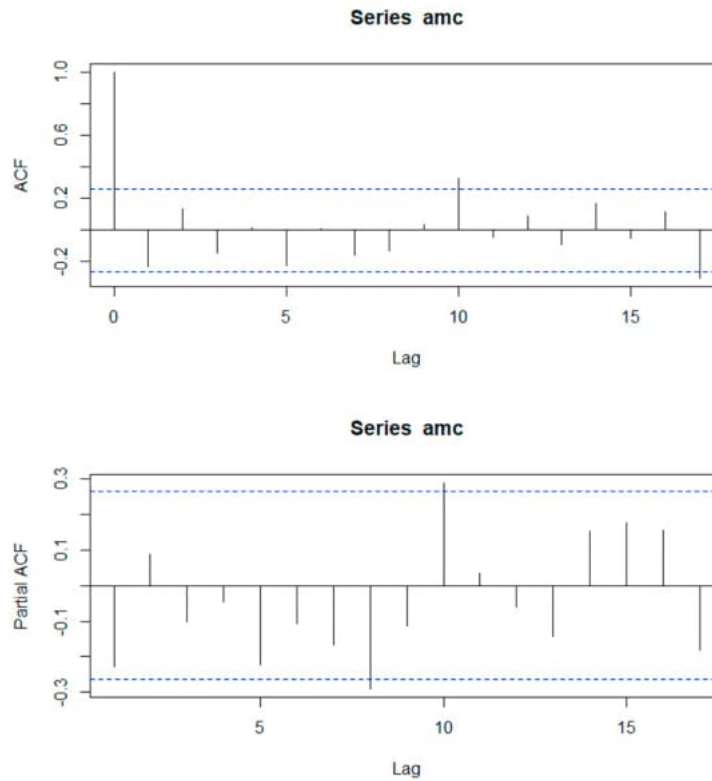


Figura 11 – ACF e PACF da taxa de resistência da *Escherichia coli* em amostra de sangue com o antibiótico Amicacina entre 2015–2019. Fonte: (LOBO ASHA KAMATH, 2021).

4.4.4.1 AIC/BIC

O Critério de Informação Akaike (AIC) é um critério que serve para penalizar modelos que possam estar sobreajustado, seja a quantidade de parâmetros usados ou o valor deles. O AIC testa a qualidade do ajuste do modelo com simplicidade, se baseando no conceito de entropia equilibrando complexidade do modelo com o desempenho nos dados. Ele premia modelos que explicam melhor os dados, mesmo tendo mais parâmetros ou mais valores. Seu cálculo pode ser observado na Equação (4.15):

$$AIC = 2 * k - 2 * \ln(L) \quad (4.15)$$

O Critério de Informação Bayesiana é semelhante mas usa um cálculo diferente. Ele pune mais complexidade e premia os que conseguem explicar os dados de maneira menos complexa, mas de forma equilibrada. Para que modelos mais complexos sejam preferidos, eles precisam fazer muita diferença no resultado final. Para obter um modelo mais complexo, é necessário encontrar um equilíbrio entre os parâmetros. A Equação (4.16) demonstra como o BIC pode ser calculado.

$$BIC = k * \ln(n) - 2 * \ln(L) \quad (4.16)$$

Onde:

- k - Quantidade de parâmetros do modelo;
- L - Função de probabilidade;
- n - Tamanho do conjunto de dados.

Esses critérios são amplamente utilizados para avaliar a qualidade do ajuste, como mostrado em (ZHANG, 2023). Assim, em alguns casos quando os gráficos de ACF e PACF podem mostrar lags muito próximo da área de significância, esses critérios podem auxiliar na escolha. As Tabelas Tabela 5 e 6 mostram alguns casos em que foi preciso utilizar esses critérios considerando os possíveis valores de P e Q apresentados nas Figuras 14 e 15.

4.4.4.2 Ljung-Box

O teste Ljung-Box, segundo (NASON, 2014), serve para validar os resíduos de um modelo. Em nosso caso o utilizamos para validar se nossos modelos ARIMA são adequados para nossos dados. Esse teste possui duas hipóteses:

- H_0 - Hipótese nula. Os resíduos não tem autocorrelação, ou seja, são independentes e distribuídos de forma idêntica. Pode-se dizer também que estes resíduos são aleatórios, implicando que o modelo é adequado aos dados. Se a hipótese nula for rejeitada (H_0 falso), é um indício de que o modelo possa estar subajustado. H_0 verdadeiro indica que todos os resíduos são ruído branco.
- H_1 - Os resíduos não são independentes e não distribuídos de forma idêntica. Ao fazer o teste, pode existir algum valor de autocorrelação $< 0,05$, e sugere que o modelo não capturou em toda a estrutura dos dados e os erros ainda contém informações que poderiam ser modeladas.

A Figura 12 é um exemplo de um modelo subajustado e a Figura 13 de um modelo onde os resíduos são ruído branco. Todas as análises estão discriminadas nas Tabela 4, 5 e 6.

4.4.5 Treino e Teste

Para avaliar a capacidade do modelo ARIMA, o conjunto de dados foi dividido em duas partes distintas: conjunto de treino e conjunto de teste. Esta abordagem é fundamental

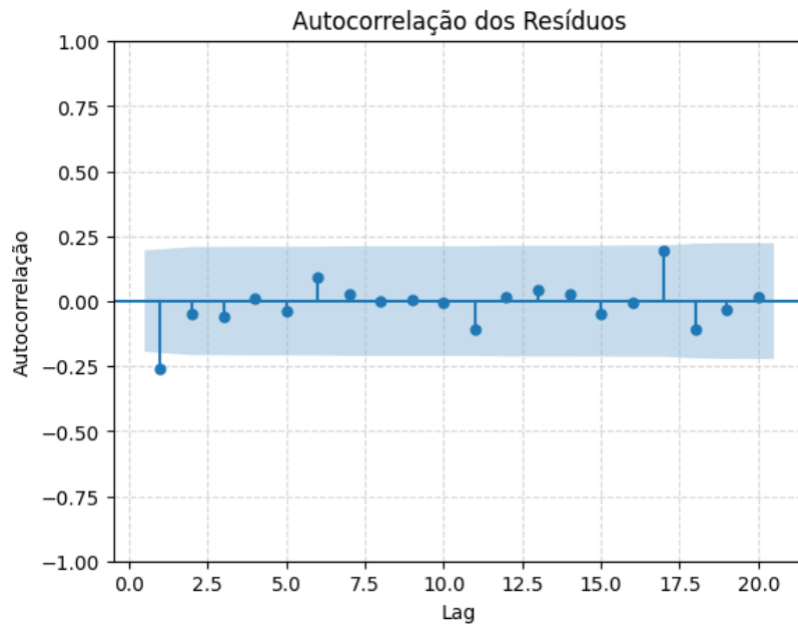


Figura 12 – Modelo ARIMA(2,1,1) constatado como subajustado para a série temporal da *Klebsiella*-Cefepima. Fonte: Elaborada pelo autor

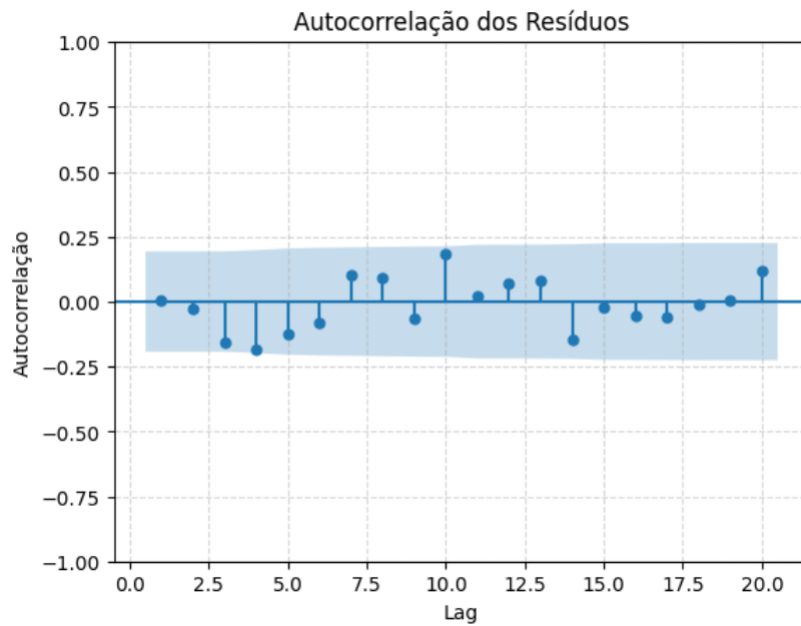


Figura 13 – Resíduos do modelo ARIMA(2,1,1) constatado como ruído branco, ou seja, ideal para a série temporal da *Escherichia coli*-Ampicilina. Fonte: Elaborada pelo autor

na modelagem de séries temporais, pois simula um cenário real onde o modelo deve fazer previsões para o futuro utilizando apenas as informações disponíveis do passado.

Diferentemente de outros problemas de aprendizado de máquina onde a divisão pode ser aleatória, em séries temporais a separação deve ser estritamente cronológica.

Tabela 4 – Modelos para cada antibiótico administrado na Klebsiella.

Antibiótico	Modelo	Ljung-Box
Cefepima	ARIMA(2,1,1)	Modelo subajustado.
Gentamicina	ARIMA(2,1,1)	Os resíduos do modelo são ruído branco.
Meropenem	ARIMA(2,1,1)	Os resíduos do modelo são ruído branco.

Tabela 5 – Modelos para cada antibiótico administrado na Escherichia-coli.

Antibiótico	Modelo	Ljung-Box	AIC/BIC
Ampicilina	ARIMA(2,1,1)	Os resíduos do modelo são ruído branco.	262.10 / 272.60
	ARIMA(1,1,1)	Os resíduos do modelo são ruído branco.	272.64 / 280.52
Ceftriaxona	ARIMA(2,1,1)	Os resíduos do modelo são ruído branco.	N/A
Gentamicina	ARIMA(2,1,1)	N/A	266.41 / 276.87
	ARIMA(3,1,1)	Os resíduos do modelo são ruído.	260.07 / 273.15

Tabela 6 – Modelos para cada antibiótico administrado na Pseudomonas.

Antibiótico	Modelo	Ljung-Box	AIC/BIC
Amicacina	ARIMA(3,1,1)	Os resíduos do modelo são ruído branco.	272.51 / 285.59
	ARIMA(1,1,1)	Os resíduos do modelo são ruído branco.	281.90 / 289.74
Meropenem	ARIMA(3,1,1)	Os resíduos do modelo são ruído branco.	N/A

Após pré-processamento, o conjunto de treino foi composto pelos dados históricos de maio de 2016 até agosto de 2024. Os dados anteriores a maio de 2016 foram desconsiderados por que o ano de 2015até abril-2016 não tinha nenhuma informação. Este conjunto de dados foi utilizado para todas as etapas de desenvolvimento do modelo, o que inclui:

- A análise exploratória da série temporal;
- A identificação dos parâmetros ótimos do modelo ARIMA (p, d, q), conforme o processo detalhado na seção 4.4.1 e ilustrado no fluxograma da Figura 3;
- A estimação dos coeficientes do modelo.

O conjunto de teste foi formado pelos quatro meses subsequentes, de setembro de 2024 a dezembro de 2024. Este conjunto foi mantido completamente "cego" durante a fase de treino e ajuste de parâmetros. Sua única finalidade é servir como uma amostra de dados futuros e não vistos para validar o desempenho do modelo finalizado. A previsão gerada pelo modelo para este período foi então comparada com os valores reais observados, permitindo o cálculo das métricas de acurácia (MAE, MAPE e RMSE) e uma avaliação fidedigna da sua performance em um contexto prático.

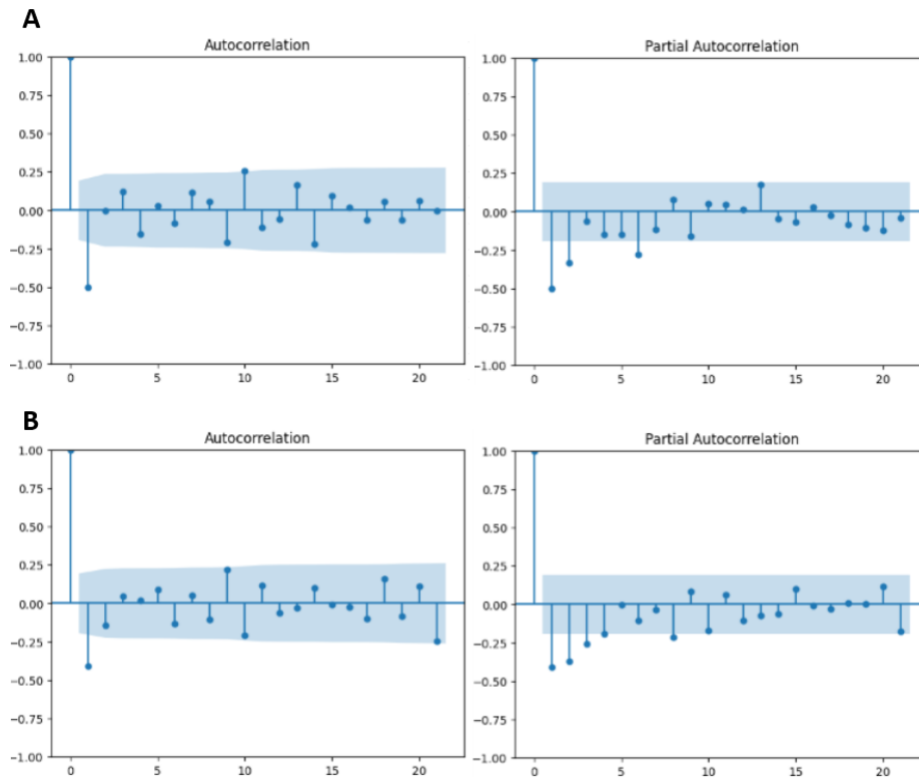


Figura 14 – Em (A) ACF e PACF do par *Escherichia coli*-Ampicilina. Em (B) o par *Escherichia coli* - Gentamicina. Fonte: Elaborada pelo autor.

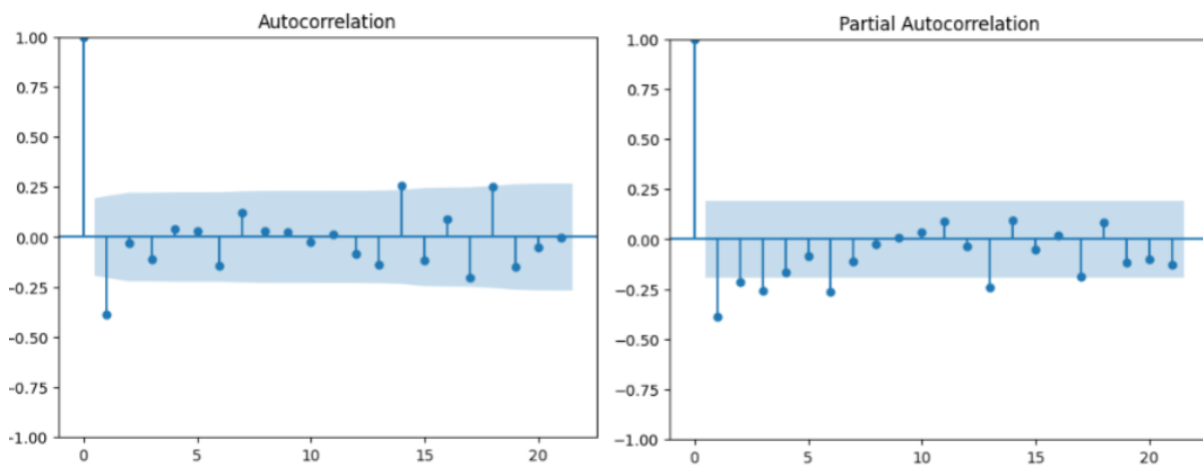


Figura 15 – ACF e PACF do par *Pseudomonas*-Amicacina. Fonte: Elaborada pelo autor.

Na Seção 4.4.1 mostramos como foi feita a busca pelos parâmetros para o ARIMA ser executado simulando um caso real:

1. O conjunto de treino foi de maio de 2016 até agosto de 2024;
2. Este mesmo conjunto foi utilizado para verificar os diferentes parâmetros do ARIMA obtidos na Seção 4.4.1 e que estão ilustrados no fluxograma da Figura 4;

3. Uma vez validados todos os parâmetros, estes foram utilizados para a previsão. Perceba que os quatro últimos meses não são conhecidos pelas ACF e PACF.
4. Na agregação mensal dos dados, os quatro últimos meses foram previstos de uma única vez e comparados com os quatro últimos que já existem na série temporal e foram calculadas as métricas de validação. O mesmo aconteceu para a agregação quinzenal, para prever as quatro últimas quinzenas.

A Tabela 7 mostra as datas de limite para Treino e Teste. Em nosso caso, os quatro últimos bimestres são referentes a teste e validação e todo o período anterior à data de início de teste foi o conjunto de treino.

Tabela 7 – Intervalos analisados com seus respectivos bimestres limites de treino e teste.

Intervalo	Mês Limite de Treino	Meses de Previsão (Teste)
2016-2024	ago-2024.	set, out, nov, dez/2024

4.5 Métricas de Desempenho

Sabe-se que não é prudente utilizar apenas uma métrica para verificar o desempenho de algo tão complexo e dinâmico, como a resistência de um patógeno. Dessa forma, utilizamos as seguintes métricas de erro: MAE e MAPE utilizada por (BINDEL; ROLAND, 2025) e RMSE utilizada por (KIM et al., 2023).

O Erro Absoluto Médio (MAE), Equação (4.17), é uma média aritmética das diferenças entre os valores reais e os previstos. Ele é útil porque é uma maneira direta de nos mostrar, na mesma escala dos dados, o tamanho do erro médio que nosso modelo cometeu. Mesmo sendo de fácil interpretação, ele é sensível a outliers e erros maiores podem passar despercebidos. Por isso a importância de olharmos para outras métricas.

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (4.17)$$

Onde:

- n - Quantidade de observações;
- y - Valor real no ponto i;
- x - Valor previsto no ponto i.

O Erro Percentual Absoluto Médio (MAPE) é uma métrica que quantifica o erro da previsão como uma porcentagem dos valores observados, fornecendo uma medida de erro relativo. Seu cálculo é dado pela seguinte fórmula (4.18):

$$MAPE = \sum_{i=1}^n \left| \frac{y_i - x_i}{y_i} \right| \times 100 \quad (4.18)$$

Ele é calculado pela média dos erros percentuais absolutos individuais de cada ponto. O MAPE é uma métrica útil para comparar o desempenho de modelos com diferentes séries temporais ainda que essas tenham escalas diferentes, isso porque ele é uma medida expressa em termos percentuais.

A Raiz do Erro Quadrado Médio (RMSE) calcula a magnitude média dos erros entre os valores previstos e os valores observados penalizando erros maiores. O RMSE deriva do MSE (Erro Quadrado Médio) com uma diferença de calcular a raiz quadrada do resultado final, convertendo a métrica para a mesma unidade da variável original, facilitando a interpretação direta. Seu resultado é dado pela seguinte fórmula (4.19):

$$RMSE = \sqrt{\sum_{i=1}^n \left(\frac{(y_i - x_i)}{n} \right)^2} \quad (4.19)$$

5 Resultados e discussão

Neste capítulo serão apresentados os resultados das análises comparativas entre os diferentes modelos do ARIMA e séries temporais das duplas ATB-Bactéria, bem como a utilização da biblioteca AMR For R, que faz uso de três modelos (Logit Binomial, Log Linear Poisson, e Linear), para comparação com nossos resultados. A estrutura deste capítulo foi organizada com base nos microrganismos avaliados: *Klebsiella*, *Escherichia coli* e *Pseudomonas*.

Enquanto que o modelo de regressão Logit Binomial, segundo Christensen 1997, o objetivo é prever uma variável de saída que é uma probabilidade ou proporção, a regressão Linear, conforme mostrado por (DISKIN, 1970), é focada em prever uma variável contínua e sem limites. Já o modelo de Poisson, mostrado por (DOBSON; BARNETT, 2001) é projetado para dados de contagem (o número de ocorrências de um evento), prevendo o número de vezes que um evento ocorrerá em um intervalo.

Para cada par bactéria-antibiótico, serão apresentados os resultados dos modelos ARIMA ajustados, seguidos da avaliação de seu desempenho preditivo por meio das métricas de erro MAE, RMSE e MAPE. Também foi feito o uso do método de predição de resistência microbiana utilizando o método da biblioteca AMR for R com o objetivo de compará-lo com nosso método. Por fim, uma discussão geral será conduzida para sintetizar os achados, contextualizá-los com aos trabalhos relacionados e discutir as implicações práticas e as limitações do estudo.

Para todos os casos, após a análise de estacionariedade, das funções de ACF e PACF e verificação dos critérios AIC/BIC, bem como o teste Ljung-Box, diferentes modelos do ARIMA foram selecionados, em alguns casos mais de um por dupla ATB-Bactéria. Todos os esses testes para seleção de parâmetros foram discutidos na Seção 4.

5.1 Análise Preditiva da Klebsiella

A Figura 16 apresenta a série histórica de 2016 a 2024 para a dupla *Klebsiella*-Cefepima utilizando ARIMA(2,1,1) para os quatro últimos meses de 2024. No gráfico, a previsão é demarcada após a linha vertical pontilhada. Percebe-se que a série é marcada por uma alta volatilidade, atingindo em alguns meses picos de 100% de resistência e vales de 0%. Todos os pré-processamentos e transformações foram demonstrados no Capítulo 4. Os resultados, bem como uma comparação com a previsão realizada na biblioteca AMR For R, estão apresentados na Tabela 8.

A previsão gerada pelas duas abordagens aponta para uma estabilização em torno

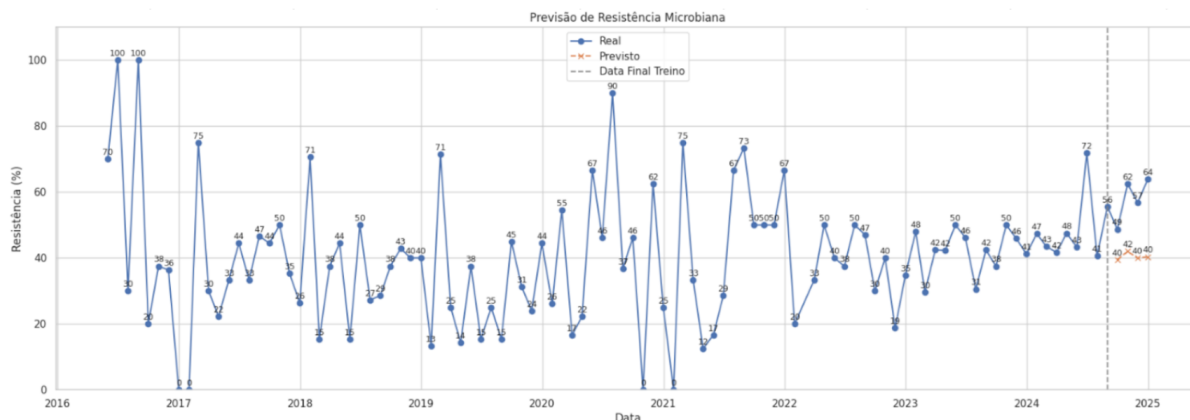


Figura 16 – Previsão da resistência da Klebsiella quando ministrado Cefepima utilizando ARIMA(2,1,1). Fonte: Elaborada pelo autor.

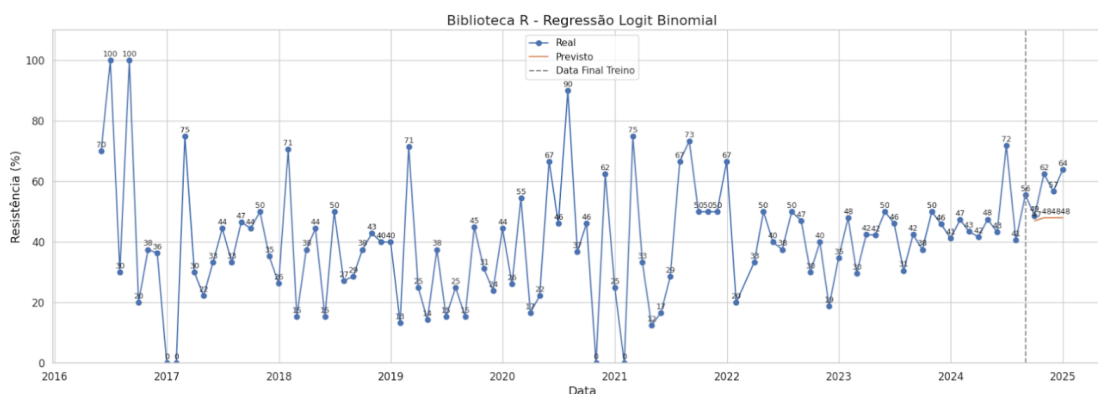


Figura 17 – Previsão da resistência da Klebsiella quando ministrado Cefepima utilizando Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.

de 41-45%, seguindo uma tendência real observada nos últimos meses, porém, sem capturar os picos reais e a tendência de crescimento apresentados nos meses previstos.

A Figura 17 traz a previsão dos últimos 4 meses de 2024 utilizando o modelo de regressão Logit Binomial da biblioteca AMR For R. A Tabela 8 mostra que o modelo da biblioteca AMR foi ligeiramente melhor, prevendo uma porcentagem de 44% e 45%, enquanto que o modelo ARIMA(2,1,1) ficou mais distante dos valores reais prevendo valores que variam entre 40% e 42%.

O teste Ljung-Box para o modelo ARIMA(2,1,1) para a Klebsiella-Cefepima foi constatado como sub ajustado, mesmo que os gráficos de ACF/PACF tenham indicado um $p = 2$ e $q = 1$. Não tivemos opções para suposição de um novo modelo, por isso utilizamos o que foi encontrado pelas funções e testes citados. Como mencionado no parágrafo anterior, os resultados do modelo ARIMA mostraram uma inferioridade grande em relação à regressão, muito provavelmente por conta desse subajuste que os testes Ljung-Box constataram.

Tabela 8 – Comparação dos resultados da previsão para Cefepima e Gentamicina ministrados na Klebsiella da análise mensal.

Antibiótico	Abordagem	MAE	MAPE	RMSE
Cefepima	ARIMA(2,1,1)	17.5	26	19.3
	Logit Binomial	10.21	16.64	11.93
	Log-Linear Poisson	42.03	74.49	42.45
	Linear	13.46	22.39	14.65
Gentamicina	ARIMA(2,1,1)	6.0	15.48	7.24
	Logit Binomial	11.28	30.42	11.75
	Log-Linear Poisson	61.78	163.32	61.85
	Linear	7.25	22.75	8.79

O destaque do ARIMA ficou para a Gentamicina, onde as abordagens regressivas não conseguiram superar o modelo autorregressivo e de média móvel. Mesmo que por pouco o ARIMA(2,1,1) superou a Regressão Linear, conseguindo capturar melhor padrões temporais e estimar os 4 meses futuros.

5.1.1 Agregação Quinzenal para a Klebsiella

Seguindo a mesma abordagem do processo proposto, esta seção se dedica à análise com uma agregação quinzenal. O volume de dados consistentemente maior no período de 2022 à 2024 viabilizou a agregação quinzenal. Esse volume maior de dados não apenas enriquece a análise ao dobrar o número de observações, mas também mitiga um dos desafios da agregação mensal: a ocorrência de picos e vales extremos. Como resultado, as séries temporais quinzenais apresentaram um comportamento menos instável.

Após executado todo o processo descrito nas seções anteriores, para o par Klebsiella-Cefepima o modelo ARIMA(2,1,1) foi o encontrado e comparado com os mesmos modelos de regressão da biblioteca AMR For R (BERENDS et al., 2022). A Tabela 9 resume o desempenho de cada modelo.

Tabela 9 – Comparação de desempenho dos modelos para a previsão quinzenal da resistência da Klebsiella à Cefepima e Gentamicina

Antibiótico	Abordagem	MAE	MAPE	RMSE
Cefepima	ARIMA(2,1,1)	12.12	21.61	16.92
	Logit Binomial	12.34	21.93	16.97
	Log-Linear Poisson	45.53	70.03	48.53
	Linear	13.11	22.31	17.24
Gentamicina	ARMA(1,0,1)	16.88	53.36	17.87
	Logit Binomial	15.25	49.08	16.24
	Log-Linear Poisson	28.05	73.65	32.47
	Linear	15.50	46.63	16.84

A análise quinzenal para a Cefepima revela um cenário de alta competitividade entre as abordagens. O modelo ARIMA(2,1,1) foi escolhido conforme mostrado na Figura 18 e o parâmetro q passou pelos testes ADF e KPSS mostrando necessário uma diferenciação. Este modelo obteve uma leve vantagem, apresentando os menores valores em todas as métricas de erro (MAE de 12.12, RMSE de 16.92 e MAPE de 21.61%). Contudo, o desempenho do modelo de Regressão Logit Binomial foi semelhante, com um MAE de 12.34, RMSE de 16.97 e MAPE de 21.93%. Isso indica que os dois modelos conseguiram modelar a série de maneira similar para atingir resultados próximos.

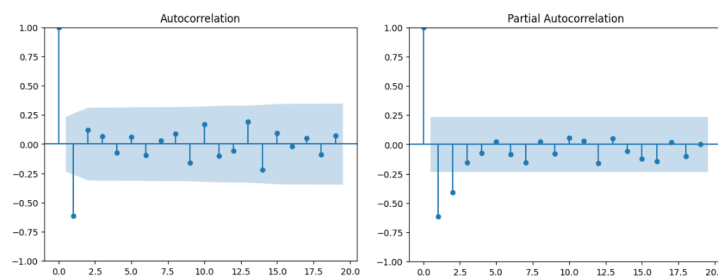


Figura 18 – Gráfico da ACF e PACF para a Klebsiella-Cefepima. Fonte: Elaborada pelo autor.

Ao contrário da análise mensal, em que a Regressão Logit Binomial superou o modelo ARIMA, na previsão quinzenal o desempenho foi à favor de nossa abordagem, mesmo que sutil. Visualmente, o gráfico da Figura 19 demonstra que, embora a volatilidade da série quinzenal seja um desafio, o modelo ARIMA conseguiu reproduzir com exatidão as duas últimas quinzenas previstas.

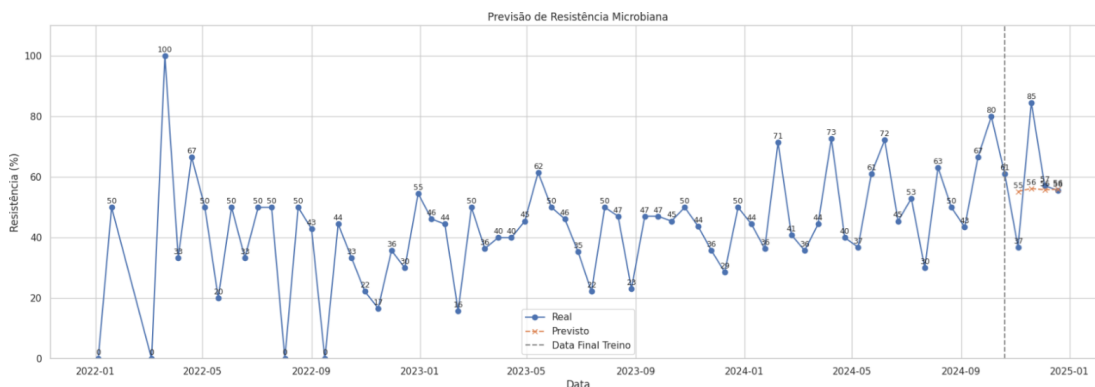


Figura 19 – Previsão da resistência da Klebsiella-Cefepima utilizando o ARIMA(2,1,1). Fonte: Elaborada pelo autor.

Já quando analisamos o par Klebsiella-Gentamicina temos uma inversão de desempenho em relação à análise mensal (Seção 5.1). Enquanto na agregação mensal o modelo ARIMA(2,1,1) havia se mostrado superior, na análise quinzenal, os modelos de Regressão Linear e Logit Binomial superaram o modelo dessa vez ARMA, uma vez que não possui o componente Integrado.

Essa mudança no desempenho sugere que a estrutura da série temporal quinzenal para este par específico pode ser dominada por uma tendência mais simples ou por flutuações que não são bem capturadas pela estrutura de autocorrelação do modelo ARMA(1,0,1) ajustado. O modelo ARMA, neste caso, apresentou os maiores erros absolutos e quadráticos, indicando uma dificuldade em se adaptar à dinâmica mais ruidosa da série quinzenal.

Visualmente, os gráficos confirmam a alta volatilidade da série, com picos e vales acentuados. Todos os modelos, incluindo os de regressão, (Figura 20 - Regressão Linear), geraram previsões mais horizontalizadas, prevendo uma tendência mais estável.

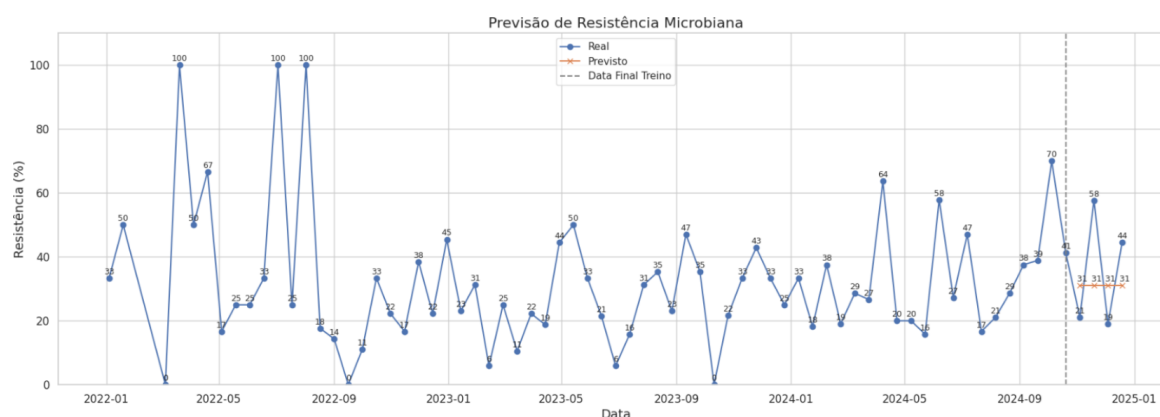


Figura 20 – Gráfico da previsão utilizando Regressão Linear para o par Klebsiella-Gentamicina. Fonte: Elaborada pelo autor.

Assim como nas análises anteriores, o modelo de Regressão Log-Linear de Poisson demonstrou ser inadequado, apresentando erros significativamente mais altos em todas as métricas. A consistência deste resultado ruim em diferentes séries e agregações reforça a conclusão de que o modelo de Poisson não é uma abordagem viável para este tipo de dado.

5.2 Análise Preditiva da *Escherichia coli*

Para a dupla Ampicilina-*Escherichia coli* foram testados cinco modelos distintos. A Tabela 10 apresenta resultados competitivos, com destaque para o modelo de Regressão Logit Binomial.

O modelo de Regressão Logit Binomial demonstrou uma superioridade em todas as três métricas de erro. O MAE de 6.04 foi aproximadamente 18% melhor em termos de erro absoluto do que o ARIMA (2,1,1). O gráfico de ambas as abordagens pode ser vista nas Figuras 21 e 22.

A proximidade entre o MAE (6.04) e RMSE (6.42) nos apresenta uma distribuição de erros mais consistente, com menor impacto de outliers. O fato do RMSE penalizar erros maiores, se seu resultado está muito próximo ao MAE (que é uma métrica de

Tabela 10 – Comparação dos resultados da previsão da Ampicilina, Ceftriaxona e Gentamicina para a Escherichia coli usando ARIMA e modelos de Regressão da biblioteca AMR For R.

Antibiótico	Abordagem	MAE	MAPE	RMSE
Ampicilina	ARIMA(2,1,1)	7.35	8	8.27
	ARIMA(1,1,1)	9.89	11	10.58
	Logit Binomial	6.04	6.68	6.42
	Log-Linear Poisson	10.2	11.57	10.9
	Linear	7.95	9.07	8.89
Ceftriaxona	ARIMA(2,1,1)	13.22	35	14.4
	ARIMA(1,1,1)	13.24	35	14.46
	Logit Binomial	17.34	47.11	18.2
	Log-Linear Poisson	64.15	185.79	64.39
	Linear	17.34	47.11	18.2
Gentamicina	ARIMA(3,1,1)	7.24	29	8.2
	ARIMA(2,1,1)	7.09	30.32	8.28
	ARIMA(1,1,1)	6.39	28.33	7.08
	Logit Binomial	7.06	33.23	7.33
	Log-Linear Poisson	77.7	398.21	78.02
	Linear	6.81	29.25	7.69

média), isso mostra nenhum erro grave na previsão. Isso também vale para as previsões do ARIMA(2,1,1), ainda que seus resultados tenham sido inferiores à Regressão Logit-Binomial.

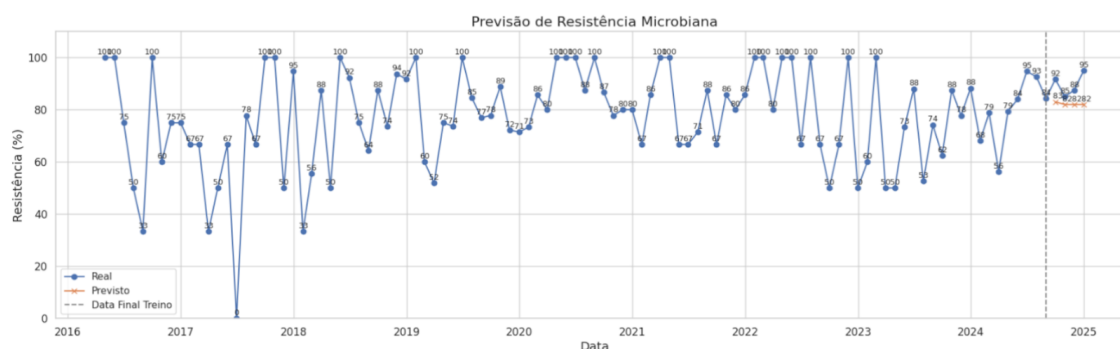


Figura 21 – Previsão da resistência da Escherichia-coli quando ministrado Ampicilina utilizando ARIMA(2,1,1). Fonte: Elaborada pelo autor.

Para a série da Ceftriaxona, Figura 23, os resultados dos dois modelos ARIMA foram superiores a todos os modelos de regressão. O ARIMA(2,1,1) obteve um MAE de 13.22 e foi mais de 4 pontos percentuais inferior ao dos modelos Logit Binomial e Linear. O desempenho do modelo Log-Linear Poisson foi extremamente baixo, com um MAPE de 185.79%, indicando que este modelo é inadequado para este tipo de dados.

Diferente da Ampicilina, a série da Ceftriaxona, embora volátil, opera em níveis

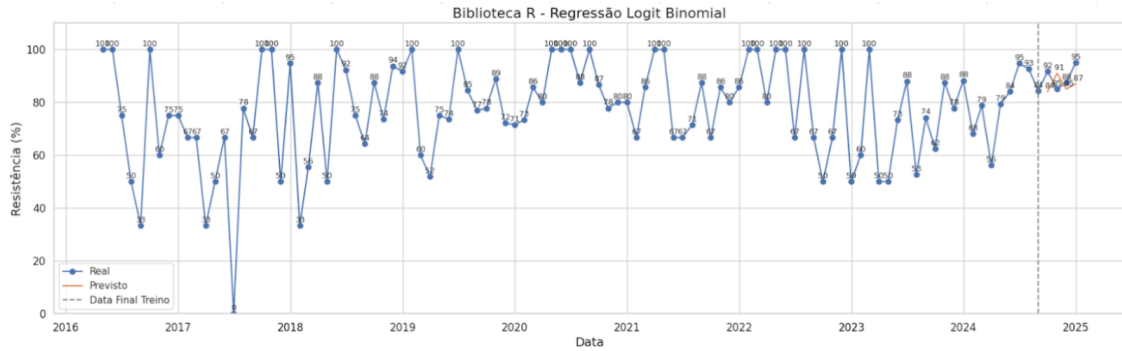


Figura 22 – Previsão da resistência da *Escherichia coli* quando administrada a Ampicilina utilizando a Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.

de resistência mais intermediários. Neste cenário, a capacidade do ARIMA de modelar a dependência serial parece ser o fator mais crucial para erros menores, superando as abordagens de regressão que não incorporam essa estrutura temporal de forma explícita.

Os dois modelos ARIMA testados (2,1,1 e 1,1,1) tiveram desempenho virtualmente idêntico, sugerindo que o modelo mais simples, ARIMA(1,1,1), poderia ser escolhido com base no princípio da parcimônia. Os resultados são apresentados na Tabela 10.

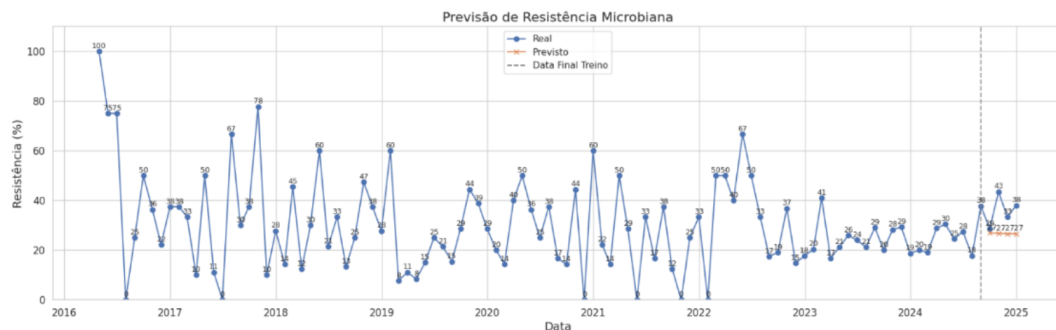


Figura 23 – Previsão da resistência da *Escherichia coli* quando ministrado Ceftriaxona utilizando a Regressão Logit Binomial da biblioteca AMR For R. Fonte: Elaborada pelo autor.

Já na série temporal da Gentamicina os resultados dos modelos ARIMA(1,1,1) e Regressão Linear foram muito semelhantes, com diferenças de erro de menos de uma unidade, sugerindo que ambos os modelos podem ser adequados para este caso, apesar do ARIMA ter resultados melhores.

Mais uma vez, o modelo Log-Linear Poisson falhou drasticamente, confirmando sua inadequação. A escolha do ARIMA(1,1,1) como o melhor modelo para Gentamicina demonstra que, por vezes, modelos de séries temporais mais simples podem ser mais adequados do que modelos mais complexos (como o ARIMA(3,1,1)), evitando o sobreajuste ao ruído da série de treino. Todo o comparativo pode ser observado na Tabela 10.

5.2.1 Agregação Quinzenal para a Escherichia coli

Para o par Escherichia coli-Ampicilina, a análise quinzenal foi conduzida com um modelo ARIMA(3,1,1). A Tabela 11 apresenta o desempenho de cada modelo para a previsão das quatro últimas quinzenas de 2024.

Tabela 11 – Comparação dos resultados da previsão da Escherichia coli utilizando dados de resistência quinzenal.

Antibiótico	Abordagem	MAE	MAPE	RMSE
Ampicilina	ARIMA(3,1,1)	5.75	5.95	7.44
	Logit Binomial	5.74	6.06	6.63
	Log-Linear Poisson	80.80	87.52	80.96
	Linear	4.25	4.62	4.91
Ceftriaxona	ARIMA(2,1,1)	16.23	66.88	20.60
	Logit Binomial	16.5	86.82	20.98
	Log-Linear Poisson	29.91	66.71	36.37
	Linear	17	88.20	21.1

Para esta análise, a Ampicilina apresenta um resultado diferente com comparação com a Ceftriaxona. Para a Ampicilina, as regressões Logit Binomial e Linear foram superiores ao ARIMA(3,1,1), porém muito similar ao que aconteceu na análise mensal, onde os modelos ARIMA foram inferiores aos regressivos com um detalhe importante: os erros foram muito baixos e resultados bastante próximos. O modelo Linear se destacou para a agregação quinzenal, podendo sugerir que a projeção de uma tendência linear sobre os dados de resistência foi mais eficaz do que a abordagem autorregressiva.

A visualização gráfica da previsão do ARIMA(3,1,1) e da Regressão Linear pode ser visualizada na Figura 24 e Figura 25 respectivamente.

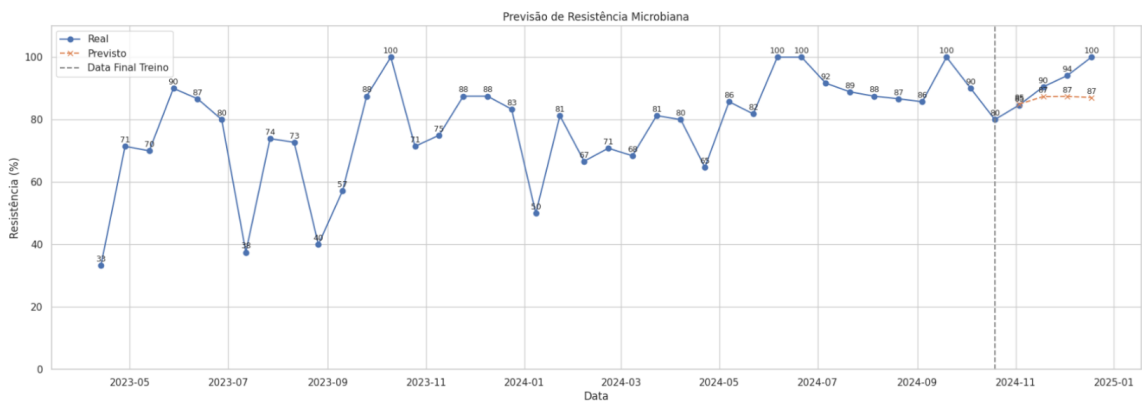


Figura 24 – Previsão da resistência da Escherichia-coli para a Ampicilina utilizando ARIMA(3,1,1) na série quinzenal. Fonte: Elaborada pelo autor.

Com relação a Ceftriaxona, a superioridade do ARIMA se manteve assim como no teste de agregação mensal. Pode-se observar pela Figura 26 que as últimas quinzenas

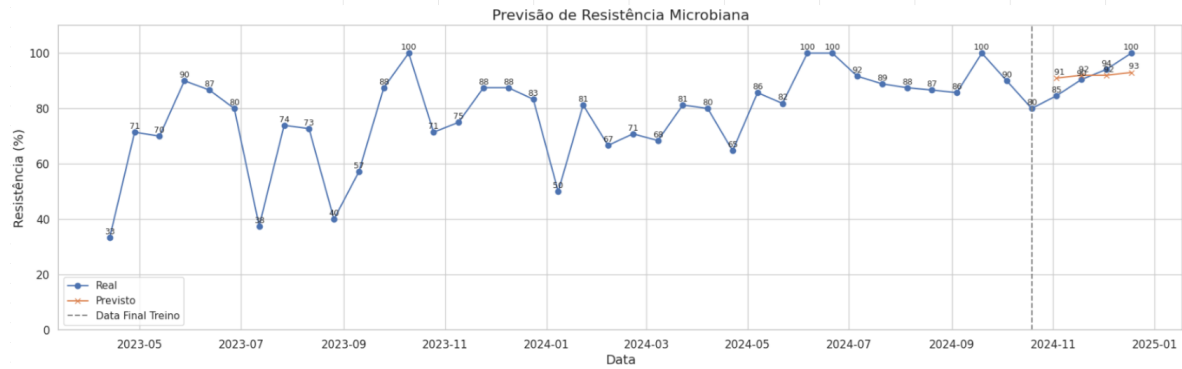


Figura 25 – Previsão da resistência da Escherichia coli para a Ampicilina utilizando Regressão Linear na série quinzenal. Fonte: Elaborada pelo autor.

previstas possuem bastante variação, mesmo assim, o modelo ARIMA(2,1,1) conseguiu capturar a tendência de alta da resistência, podendo assim servir como um alerta para a equipe médica.

Essa presença ruidosa com picos pode explicar o valor do MAPE para as regressões Logit-Binomial e Linear, ambos atingindo mais de 80%, como pode ser visto na Tabela 11.

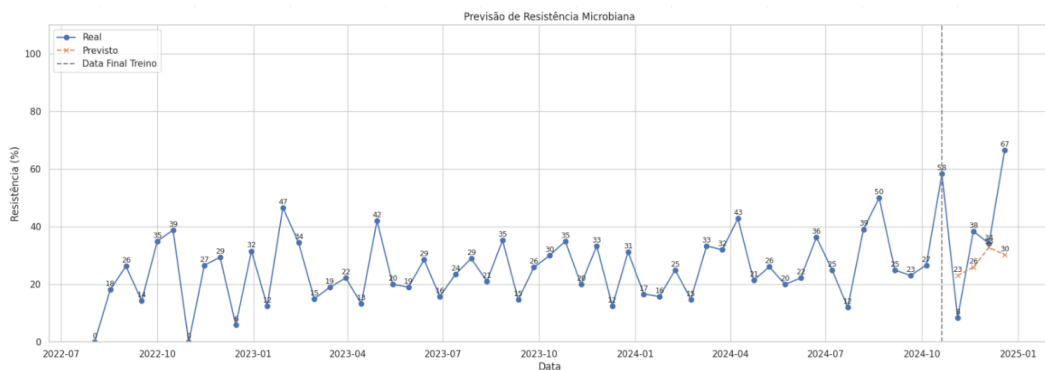


Figura 26 – Previsão da resistência da Escherichia coli para a Ceftriaxona utilizando ARIMA(2,1,1) na série quinzenal. Fonte: Elaborada pelo autor.

5.3 Análise Preditiva da Pseudomonas

Para a série temporal de resistência à Amicacina, os testes de ACF, PACF, AIC/-BIC e Ljung-Box mostraram dois possíveis modelos: ARIMA(3,1,1) e ARIMA(1,1,1). Constatamos que o segundo modelo apresentou um MAPE (48%) e RMSE (8.37) inferior em comparação com o ARIMA(3,1,1) (64% e 8.78), conforme apresentado na Tabela 12.

Ao comparar as duas abordagens fica evidente a superioridade dos modelos ARIMA. Ambos ARIMA(3,1,1) e ARIMA(1,1,1) obtiveram um MAE de 6.69 e 5.47 respectivamente, o que é menor do que todos os MAE dos modelos de Regressão da biblioteca AMR For R. o MAE da Regressão Logit Binomial (Regressão Logística com Distribuição Binomial) foi

de 9.07, sendo em torno de 26% maior em comparação ao modelo ARIMA com menor performance. A Tabela 12 apresenta os resultados consolidados. As Figuras 27, 28 e 29 mostram graficamente a previsão em comparação com a série histórica entre os modelos ARIMA e a Regressão Logit Binomial.

A discrepância fica maior ao olharmos a Regressão Log-Linear com distribuição de Poisson e o MAPE de 586.56, se mostrando inadequada para a previsão de AMR para este tipo de série temporal. Fica evidente que o ARIMA conseguiu modelar melhor características temporais (auto regressão e tendência).

Tabela 12 – Comparação dos resultados da previsão *Pseudomonas* implementando modelos ARIMA e a biblioteca AMR For R utilizando modelos de Regressão.

Antibiótico	Abordagem	MAE	MAPE	RMSE
Amicacina	ARIMA(3,1,1)	6.69	64	8.78
	ARIMA(1,1,1)	5.47	59.78	6.77
	Logit Binomial	9.07	50.62	10.97
	Log-Linear Poisson	52.11	586.56	64.49
	Linear	13.77	92	14.98
Meropenem	ARIMA(3,1,1)	10.58	23.03	12.05
	Logit Binomial	27.4	61.73	28.63
	Log-Linear Poisson	39.33	88.63	41.53
	Linear	39.15	88	40.98

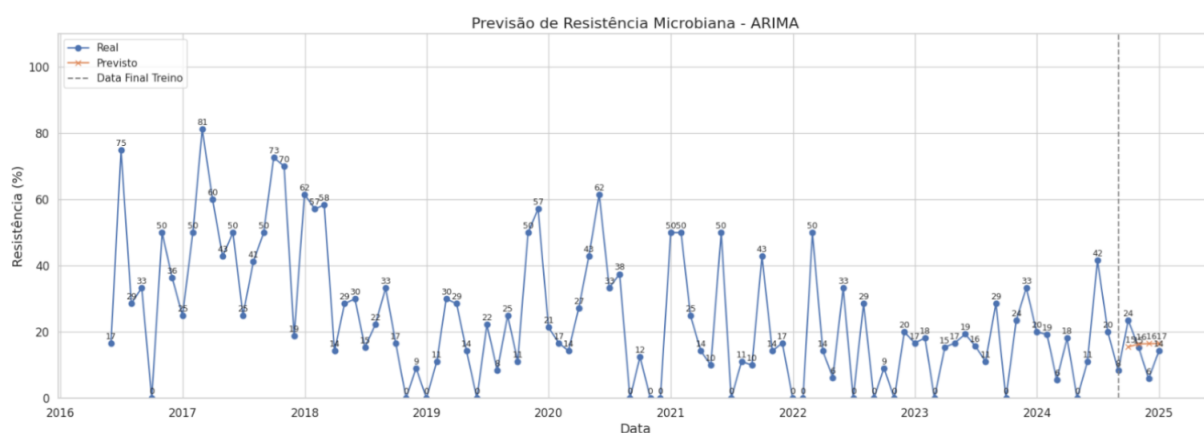


Figura 27 – Previsão da resistência da *Pseudomonas* quando ministrado Amicacina utilizando ARIMA(1,1,1). Fonte: Elaborada pelo autor.

A discrepância se mostrou maior ainda quando comparadas as performances da previsão na série temporal do Meropenem. O modelo ARIMA(3,1,1) apresentou um MAPE 38,7% menor em comparação com o melhor desempenho entre os modelos de regressão da biblioteca AMR For R, sendo este também o Logit-Binomial.

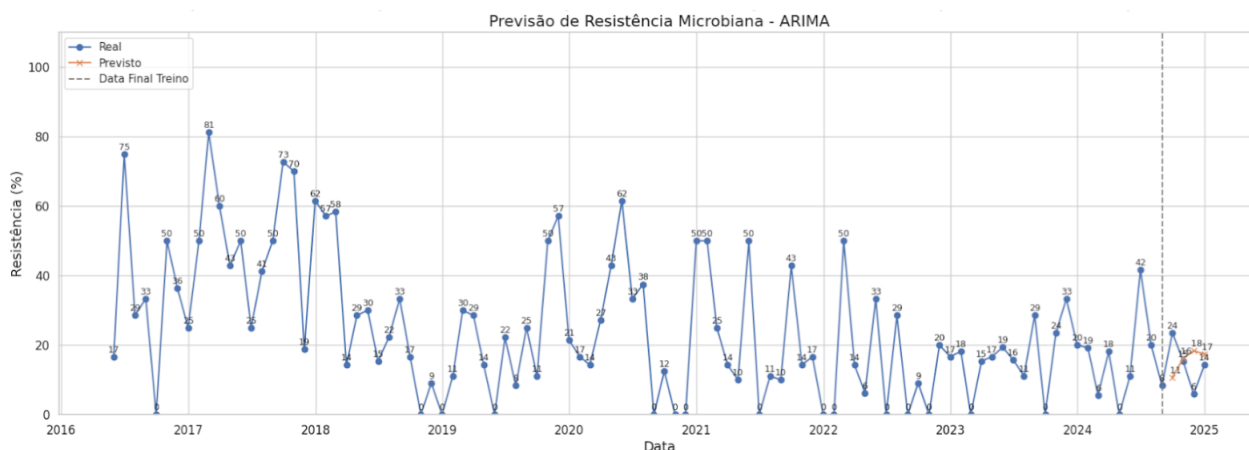


Figura 28 – Previsão da resistência da *Pseudomonas* quando ministrado Amicacina utilizando ARIMA(3,1,1). Fonte: Elaborada pelo autor.

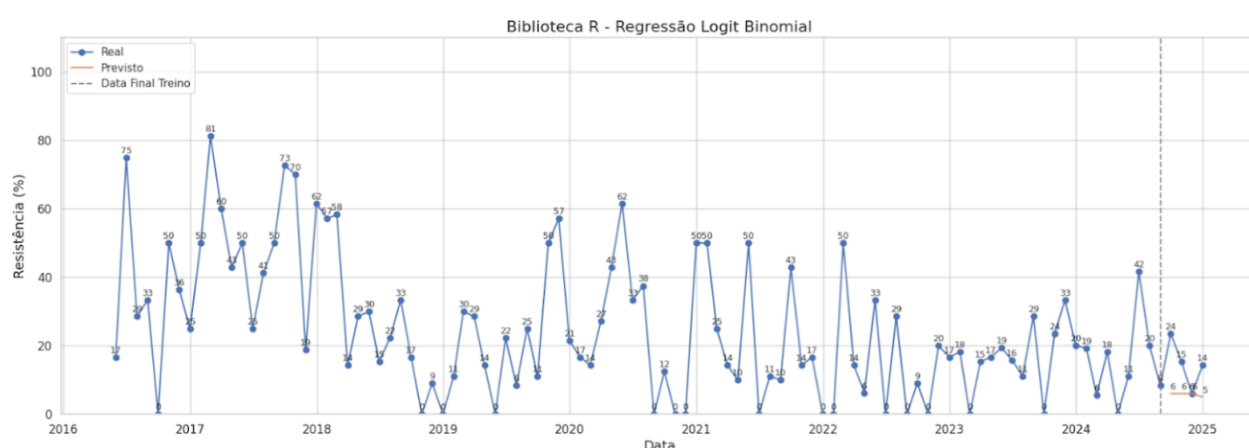


Figura 29 – Previsão da resistência da *Pseudomonas* quando ministrado Amicacina utilizando Logit Binomial da biblioteca AMR for R. Fonte: Elaborada pelo autor.

Analisando as métricas de erro da Tabela 12 fica evidente que os modelos de Regressão não conseguiram captar as tendências temporais tão bem quanto o ARIMA que, mesmo com valores consideráveis, cometeu erros bem menores.

5.3.1 Agregação Quinzenal para a *Pseudomonas*

A análise quinzenal para *Pseudomonas*-Amicacina reforça uma tendência observada ao longo deste trabalho: a superioridade do modelo ARIMA em métricas de erro absoluto. O ARMA(2,0,2) se destacou com o menor MAE (10.84) e RMSE (14.84). Isso indica que, em média, suas previsões foram as mais próximas dos valores reais. Este resultado está alinhado com a análise mensal (Seção 5.3), onde o ARIMA também superou os modelos de regressão de forma expressiva. A consistência do desempenho superior do ARIMA/ARMA em ambos os intervalos temporais sugere que a estrutura de autocorrelação desta série específica é um componente forte e persistente, que é mais bem capturado pela metodologia

(BOX et al., 2016). Os resultados estão consolidados na Tabela 13.

Visualmente, o gráfico da Figura 30, mostram que o ARMA tentou acompanhar a tendência de queda e subida no período de teste. Em contrapartida, as regressões produziram previsões mais estáveis e com menor variação, resultando em erros absolutos maiores. No geral, a análise para a Amicacina consolida o modelo ARMA/ARIMA como a abordagem mais robusta e acurada para a previsão da resistência do gênero *Pseudomonas* no contexto estudado.

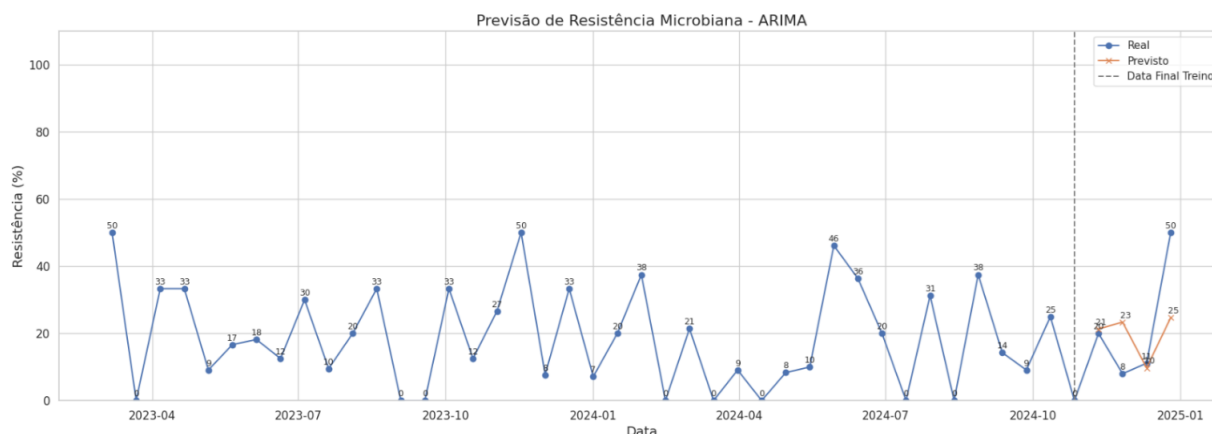


Figura 30 – Previsão da resistência da *Pseudomonas*-Amicacina utilizando ARMA(2,0,2) na série quinzenal. Fonte: Elaborada pelo autor.

Tabela 13 – Tabela de resultados da *Pseudomonas* com a série temporal quinzenal.

Antibiótico	Abordagem	MAE	MAPE	RMSE
Amicacina	ARMA(2,0,2)	10.84	65.46	14.84
	Logit Binomial	12.47	57.5	17.26
	Log-Linear Poisson	20.27	85.75	26.2
	Linear	12.72	61.62	17.41
Meropenem	ARIMA(2,2,1)	23.86	35.41	30.09
	Logit Binomial	20.4	24.98	31.72
	Log-Linear Poisson	53.9	89.69	58.96
	Linear	24.4	32.67	34.43

A análise do Meropenem quinzenal demonstra uma inversão no histórico dos resultados que até então favoreceram a performance do ARIMA para a *Pseudomonas*, incluindo o Meropenem com os intervalos mensais.

Os resultados podem ser vistos na Tabela 13, onde demonstram que o modelo de Regressão Logit Binomial emergiu como a abordagem mais acurada, alcançando o menor MAE (20.41) e MAPE (24.98%), apesar do ARIMA ter competido em MAE(23.86) e RMSE(30.09).

É possível que a estrutura de autocorrelação capturada na série mensal tenha se tornado menos predominante na série quinzenal. Um outro ponto interessante de ser observado na Figura 31 é que essa regressão conseguiu uma alta precisão nas duas primeiras quinzenas, isto porque na maioria das previsões os modelos regressivos preveram resistências mais horizontalizadas. Em contraste, o ARIMA(2,1,1), observado pela Figura 32, tentou reproduzir uma previsão ruidosa, com grandes picos e quedas assim como na série histórica, porém, sem capturar a tendência de uma alta íngreme que a série desenvolveu.

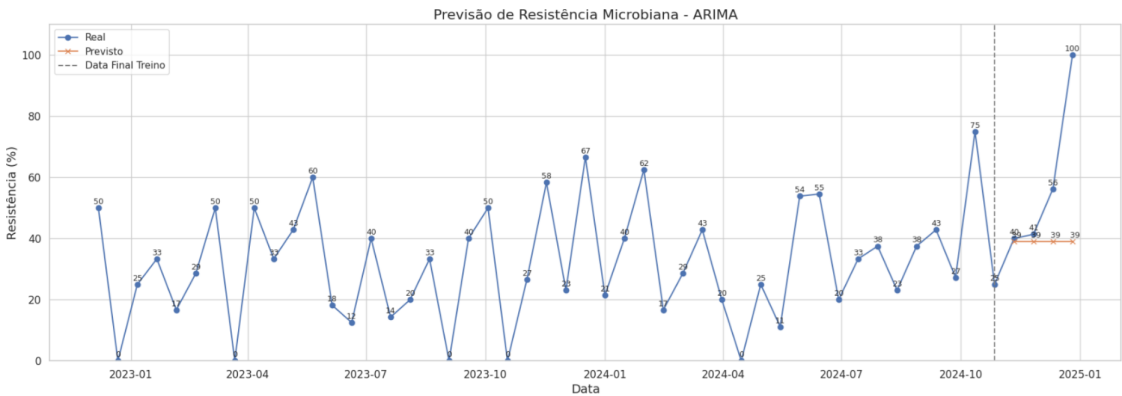


Figura 31 – Previsão da resistência da *Pseudomonas*-Meropenem utilizando Regressão Logit Binomial na série quinzenal. Fonte: Elaborada pelo autor.

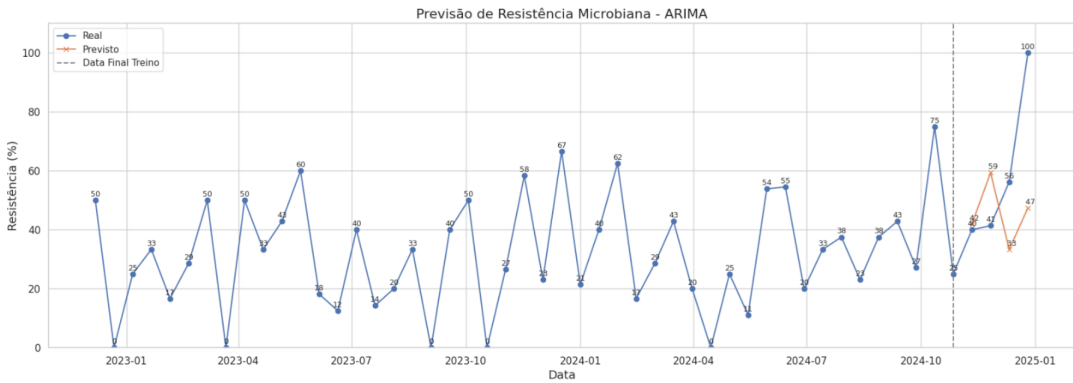


Figura 32 – Previsão da resistência da *Pseudomonas*-Meropenem utilizando o ARIMA(2,1,1) na série quinzenal. Fonte: Elaborada pelo autor.

6 Conclusões

Este trabalho se propôs a desenvolver e avaliar modelos estatísticos para a previsão de resistência bacteriana a antibióticos, utilizando dados do Hospital das Clínicas de Pernambuco. O objetivo principal foi empregar a metodologia de séries temporais, com foco no modelo ARIMA, como uma ferramenta de baixo custo e rápida execução para gerar previsões que possam auxiliar na tomada de decisão clínica e no controle de infecções. A metodologia envolveu um robusto processo de engenharia de dados, a modelagem estatística para diferentes pares de bactéria-antibiótico e a comparação de desempenho com outras abordagens de regressão.

Os resultados demonstraram que os modelos ARIMA/ARMA pode ser ferramentas viáveis e em muitos casos superiores à modelos de regressão para a previsão de taxas de resistência. Para os patógenos do gênero *Pseudomonas*, o ARIMA apresentou um desempenho notavelmente melhor que as abordagens de Regressão Linear, Logit Binomial e Log-Linear Poisson, indicando sua capacidade de capturar a dependência temporal intrínseca a essa série.

No caso da *Escherichia coli*, os resultados foram mistos: enquanto a abordagem de Regressão Logit Binomial errou menos na previsão para o par com Ampicilina, os modelos ARIMA foram mais eficazes para prever a resistência à Ceftriaxona. Este achado sugere que não existe uma única abordagem universalmente superior, e a escolha do modelo ideal pode depender das características específicas de cada série temporal.

Algo importante a ser mencionado é que para a série quinzenal *Escherichia coli*-Gentamicina os hiperparâmetros do modelo ARIMA foram $p = 0$, $d = 0$ e $q = 0$, ou seja, esta série se comporta como um ruído branco. No entanto, esta dissertação não pôde contemplar a execução de uma predição em uma série de ruído branco, sendo altamente desafiador por não haver padrões no passado que possam prever o futuro, fugindo dos objetivos deste trabalho.

Quando olhamos para a série do gênero *Pseudomonas*, a superioridade do ARIMA foi mais evidente na análise dos resultados da agregação mensal, onde os modelos apresentaram erros consistentemente inferiores aos de todos os modelos de regressão testados. Isso indica que a resistência para este patógeno possui dependência considerável de seus valores passados.

Quando a configuração da série muda, ou seja, tornando-se quinzenal, apesar do ARMA ter vencido para a Amicacina, o modelo Regressivo Logit Binomial superou o ARIMA(2,1,1) para o Meropenem.

É importante mencionar o baixo desempenho do modelo Log-Linear Poisson em quase todas as análises, se mostrando inadequado para este tipo de dados. Isso pode ser explicado por dois motivos principais descritos no Capítulo 9 de (DOBSON; BARNETT, 2001):

- Natureza da variável alvo (Proporção x Contagem) - O método de Poisson é feito para modelar dados de contagem, isto é, prever o número de vezes que um evento ocorrerá em um intervalo. Uma vez que nossos dados são resistência microbiana em Porcentagem, o modelo é inadequado para este caso;
- Premissa do método Poisson - Este tipo de distribuição define que a média dos dados deve ser igual à sua variância ($E[Y] = Var[Y]$). Os dados de proporção analisados neste trabalho são marcados por uma alta volatilidade, ocasionando a variância ser muito maior do que a sua média.

6.1 Dificuldades na Previsão e a Volatilidade das Séries

Um ponto central observado durante a análise foi a dificuldade de prever com baixos erros algumas séries temporais. A principal barreira está na volatilidade e na presença de valores extremos nos dados históricos, como observado na série *Klebsiella*-Cefepima, que em muitos meses apresentou picos de 100% e vales de 0%.

Essa volatilidade é um reflexo das limitações do conjunto de dados, que por sua vez é um reflexo do universo microbiológico somado à falta de insumos para a coleta no Hospital das Clínicas. A necessidade de agregar os dados mensalmente e quinzenalmente para contornar as lacunas temporais e a baixa quantidade de dados, em determinados períodos resultou em proporções de resistência extremas (0% ou 100%) quando poucos exames eram realizados.

Por mais que as tentativas de agregação mensal e quinzenal tenham sido exploradas, ainda não foram suficientes para resultar em séries temporais menos ruidosas.

Em muitos casos o ARIMA falhou em prever a magnitude exata dos eventos extremos, como demonstrado na previsão para *Klebsiella*-Cefepima, que se estabilizou em torno de 41% e 45% e não capturou picos reais, por outro lado, cumprindo nosso objetivo, conseguiu em muitos casos capturar a tendência geral de subida ou queda da resistência, como por exemplo nas agregações mensais da *Escherichia coli*-Ceftriaxona e *Pseudomonas*-Amicacina.

Uma análise mais aprofundada dos resultados mostra que a dinâmica da resistência antimicrobiana não é um fenômeno homogêneo, mas sim específico para cada patógeno.

Adicionalmente, a resistência bacteriana é um fenômeno complexo, influenciado por fatores externos não observados neste estudo, como mudanças em protocolos clínicos, idade, sexo e comorbidades dos pacientes, além do uso de múltiplos antibióticos. Uma série temporal univariada, por definição, modela o futuro com base apenas em seus próprios valores passados. Embora os efeitos desses fatores externos estejam implícitos na série, a informação sobre suas causas está ausente. Como o algoritmo não consegue avaliar o impacto isolado de cada variável, os eventos impulsionados por esses fatores são interpretados como 'choques' aleatórios, o que torna sua previsão um desafio maior.

6.2 Contribuições

O desenvolvimento dessa pesquisa resultou em contribuições significativas que vão desde a engenharia de dados até validações de algoritmos preditivos, criando um marco para futuras pesquisas pautadas na resistência microbiana no ambiente do Hospital das Clínicas de Pernambuco. As principais contribuições estão detalhadas a seguir.

- Engenharia de Dados e Geração de um Ativo para Pesquisas Futuras - Realizamos uma contribuição fundamental na consolidação dos dados brutos hospitalares de diferentes máquinas em uma única base de dados. Antes da formatação, os dados contavam como colunas heterogêneas, não padronização de nomes de microrganismos, de antibióticos e datas do exame. O resultado deste processo não foi apenas um requisito para a análise, mas a geração de um banco de dados mais coeso e padronizado, tornando-se valioso para pesquisas futuras evidenciando, inclusive, a importância da preocupação da qualidade da estruturação dos dados já no momento de sua coleta.
- Análise do modelo ARIMA como potencial solução para predição de resistência antimicrobiana superior à Regressão - Apesar de diversas outras técnicas estatísticas puderem ser aplicadas no contexto de predição de resistência microbiana, respeitamos as limitações dos dados e isso nos levou ao uso do ARIMA. Após as comparações entre nosso modelo e diferentes tipos de Regressão (Linear, Logit Binomial e Log Linear Poisson) os resultados mostraram uma superioridade da abordagem auto regressiva e de média móvel na maioria das duplas BACT-ATB e resultados muito semelhantes nos casos que a regressão se sobressaiu.
- Proposição de um novo processo para predição de resistência antimicrobiana através de séries temporais - Os dois tópicos anteriores contribuíram para a criação de um processo metodológico completo e replicável para análise de resistência microbiana. Desde a formatação dos dados utilizando uma biblioteca confiável [34], escolha do intervalo dos dados para superar limitações da série, testes de estacionariedade

(ADF, KPSS) e análises de ACF/PACF, utilização de critérios AIC/BIC e Teste Ljung-Box para identificação e validação do modelo mais adequado. Esse refinamento é essencial para entender os dados e escolha do melhor modelo. Além disso, a etapa final envolveu não apenas a predição, mas também a validação dos resultados e a comparação com uma outra técnica de predição utilizada pela biblioteca AMR For R.

Assim, nossa contribuição vai além de apenas uma previsão, oferecendo um acervo metodológico indo da entrega de uma base de dados refinada para análise, até a previsão validada, comparada e interpretada.

6.3 Trabalhos Futuros

Os resultados produzidos por essa dissertação estabelecem uma base sólida mas também, abrem a possibilidade para futuras abordagens promissoras. As propostas de trabalhos futuros se concentram em três importantes abordagens: Aprimoramento da qualidade e quantidade dos dados, exploração de outros métodos preditivos e contribuição para a comunidade científica e de código aberto.

O principal desafio encontrado neste trabalho, como discutido na Seção 4.3 foi a limitação dos dados, traduzida por muitas lacunas temporais, e também a quantidade restrita de observações resultadas não só por conta da agregação mensal, mas também pela baixa quantidade de culturas, mesmo para os microrganismos mais recorrentes do hospital. Adicionado a isto, uma informação muito importante é o MIC, que está presente em alguns registros mas não em outros, o que é de grande importância como demonstrado na Seção 3.1.

Dessa forma, propomos como trabalho futuro a incorporação de técnicas de Imputação de Dados, visando preservar informações valiosas que possam ter sido perdidas por causa da ausência em muitas culturas, como é o caso do MIC. Técnicas de Aumento de Dados *Data Augmentation* também são uma possível linha de pesquisa para preencher na série temporal lacunas diárias/semanais, e mesmo quinzenais, e ser criado um banco de dados mais numeroso que pode servir como base para técnicas preditivas que necessitam de uma quantidade muito maior de dados.

Uma vez que temos o ARIMA, e modelos de regressão, como técnicas validadas, e o trabalho de maior qualificação dos dados concluído, um próximo passo é a exploração de modelos mais complexos, tanto estatísticos (e.g. SARIMA) como abordagens baseadas em Inteligência Artificial.

Por fim, a formalização e incorporação do processo metodológico desenvolvido nessa dissertação como um novo módulo da biblioteca AMR For R ([BERENDS et al., 2022](#)) seria

uma grande contribuição para a comunidade científica. Como detalhado na seção 4.2.1, este pacote foi fundamental para a etapa de padronização dos dados. Essa contribuição representaria uma expansão significativa desta biblioteca, que atualmente conta com apenas uma técnica preditiva, indo desde a criação de funções para agregação temporal dos dados, identificando os hiperparâmetros do ARIMA até a geração das previsões e gráficos resultantes.

Referências

- AGUERO, S. et al. Machine learning techniques to identify antimicrobial resistance in the intensive care unit. *Entropy*, v. 21, p. 603, 2019. Citado 2 vezes nas páginas 5 e 8.
- BERENDS, M. S. *A New Instrument for Microbial Epidemiology: Empowering Antimicrobial Resistance Data Analysis*. [S.l.]: University of Groningen, 2021. Citado na página 18.
- BERENDS, M. S. et al. Amr: An r package for working with antimicrobial resistance data. *Journal of Statistical Software*, v. 104, p. 1–31, 2022. Citado 5 vezes nas páginas 13, 18, 22, 41 e 56.
- BINDEL, L. J.; ROLAND, S. Long-term forecast for antibacterial drug consumption in germany using arima models. *Naunyn-Schmiedeberg's Archives of Pharmacology*, v. 398, 2025. Citado 6 vezes nas páginas 5, 7, 8, 14, 15 e 37.
- BOECKEL, V. et al. Global trends in antimicrobial use in food animals. *Proc Natl Acad Sci USA*, p. 5649–5654, 2015. Citado na página 2.
- BOX, G. E. et al. *Time series analysis : forecasting and control*. [S.l.]: John Wiley Sons, Inc., Hoboken, New Jersey., 2016. Citado 7 vezes nas páginas 13, 12, 14, 26, 27, 30 e 50.
- BRCast. *Tabelas de pontos de corte para interpretação de CIMs e diâmetros de halos*. [S.l.], 2024. Citado na página 2.
- BUSH K., J. G. A. Updated functional classification of beta-lactamases. *Antimicrobial agents and chemotherapy*, v. 54, p. 969—976, 2010. Citado na página 12.
- CHRISTENSEN, R. *Logistic Regression, Logit Models, and Logistic Discrimination*. [S.l.]: Springer New York, 1997. 116–177 p. Citado na página 39.
- CLINIC, C. *Bacteria*. [S.l.], 2022. Citado na página 10.
- DISKIN, M. H. Definition and uses of the linear regression model. *Advancing Earth and Spaces Science*, 1970. Citado na página 39.
- DOBSON, A. J.; BARNETT, A. G. *An Introduction to Generalized Linear Models*. [S.l.]: Chapman Hall/CRC, 2001. Citado 2 vezes nas páginas 39 e 54.
- GUO, W. et al. Antimicrobial resistance surveillance and prediction of gram-negative bacteria based on antimicrobial consumption in a hospital setting: A 15-year retrospective study. *Medicine*, v. 21, p. 98, 2019. Citado na página 31.
- HERNANDEZ, B. et al. Resistance trend estimation using regression analysis to enhance antimicrobial surveillance: A multi-centre study in london 2009–2016. *Antibiotics (Basel)*, v. 10, p. 1267, 2021. Citado 2 vezes nas páginas 30 e 31.
- IHME. *Antimicrobial resistance leads to more deaths and illnesses in the WHO African region than anywhere else*. [S.l.], 2023. Citado 2 vezes nas páginas 1 e 2.

- KIM, J. et al. Predicting antimicrobial resistance of bacterial pathogens using time series analysis. *Frontiers in microbiology*, v. 14, 2023. Citado 9 vezes nas páginas 1, 2, 5, 6, 7, 8, 14, 23 e 37.
- KRISTENSEN, K. et al. Using image processing and automated classification models to classify microscopic gram stain images. *Computer Methods and Programs in Biomedicine Update*, 2023. Citado na página 10.
- LEVERNSHTEIN, V. I. Binary codes capable of correcting, deletions, insertions, and reversals. *Soviet Physics Doklady*, v. 10, p. 845–848, 1966. Citado na página 21.
- LEWIN-EPSTEIN, O. et al. Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records. *Clinical Infectious Diseases (Clin Infect Dis)*, v. 72, 2021. Citado 2 vezes nas páginas 6 e 8.
- LOBO ASHA KAMATH, V. K. E. J. Time series model for a proportion of antimicrobial resistance rate. *Clinical Epidemiology and Global Health*, v. 21, p. 101–290, 2021. Citado 3 vezes nas páginas 13, 31 e 32.
- LOOD ROLF, E. G. M. B. Revisiting antibiotic resistance spreading in wastewater treatment plants. *Frontiers in Microbiology*, v. 8, 2017. Citado na página 2.
- LUZ, C. F. *Data science for infection management antimicrobial stewardship*. [S.l.]: University of Groningen, 2021. Citado na página 18.
- MAGIORAKOS, P. et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clinical Microbiology and Infection (Clin Microbiol Infect)*, v. 18, p. 268—281, 2012. Citado na página 11.
- MARTÍNEZ, E. P. et al. Trends, seasonality and the association between outpatient antibiotic use and antimicrobial resistance among urinary bacteria in the netherlands. *Journal of Antimicrobial Chemotherapy*, v. 75, p. 2314—2325, 2020. Citado na página 13.
- MOHAMMED, I. et al. Antibiotics in wastewaters: a review with focus on oman. *Water Science*, v. 8, p. 199, 2018. Citado na página 2.
- MURRAY, C. J. L. e. a. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, v. 399, p. 629–655, 2022. Citado 2 vezes nas páginas 1 e 2.
- NASON, D. S. G. P. White noise testing using wavelets. *Stat Journal*, v. 3, p. 351–362, 2014. Citado na página 33.
- NHGRI. *Bacteria*. [S.l.], 2022. Citado na página 10.
- PANKEY, G. A.; SABATH, L. D. Clinical relevance of bacteriostatic versus bactericidal mechanisms of action in the treatment of gram-positive bacterial infections. *Clinical Infectious Diseases (Clin Infect Dis)*, v. 38, 2004. Citado na página 9.
- PATEL, P. et al. Antibiotics. *Treasure Island (FL), EUA*, 2025. Citado na página 9.

- PUURONEN SEPPÖ, P. M. T. A. Effectiveness of local feature selection in ensemble learning for prediction of antimicrobial resistance. *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2008. Citado 2 vezes nas páginas 5 e 8.
- REYGAERT, W. C. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiology*, v. 4, p. 482–501, 2018. Citado na página 10.
- SAKAGIANNI, A. et al. Using machine learning to predict antimicrobial resistance – a literature review. *Antibiotics (Basel)*, v. 12, 2023. Citado 2 vezes nas páginas 6 e 8.
- STRAHLBERG, D. Antibiotics resistance forecasting: A comparison of two time series forecast models. *SIAM Undergraduate Research Online*, v. 14, p. 383—399, 2021. Citado na página 26.
- TAMMA, P. D. et al. Association of adverse events with antibiotic use in hospitalized patients. *JAMA Internal Medicine (JAMA Intern Med)*, v. 177, 2017. Citado na página 9.
- TAN, S. Y.; TATSUMURA, Y. Alexander Fleming (1881-1955): Discoverer of penicillin. *Singapore Med Journal*, v. 50, p. 366–367, 2015. Citado na página 1.
- UNEP. *Bracing for Superbugs: Strengthening environmental action in the One Health response to antimicrobial resistance*. [S.l.], 2023. Citado 3 vezes nas páginas 13, 2 e 3.
- WANG, Y. et al. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep.*, v. 8, 2018. Citado na página 31.
- WHO. *Antimicrobial resistance*. [S.l.], 2023. Citado na página 1.
- WILLMANN, M. et al. Time series analysis as a tool to predict the impact of antimicrobial restriction in antibiotic stewardship programs using the example of multidrug-resistant *Pseudomonas aeruginosa*. *Antimicrobial Agents and Chemotherapy*, v. 57, 2013. Citado 2 vezes nas páginas 5 e 8.
- YELIN, I. et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature medicine*, v. 25, p. 1143–1152, 2019. Citado 2 vezes nas páginas 6 e 8.
- ZHANG, G. M. Y. Simulation of an adaptive model based on AIC and BIC ARIMA predictions. *Journal of Physics: Conference Series*, 2023. Citado na página 33.
- ZHENG, A. et al. An application of ARIMA model for predicting total health expenditure in China from 1978–2022. *Journal of Global Health (J Glob Health)*, v. 10, 2020. Citado 2 vezes nas páginas 11 e 30.

Apêndices

Apêndice I

Colunas do conjunto de dados Master 762.

DT_PEDIDO, NU_PEDIDO, CD_ATENDIMENTO, CD_PACIENTE, NM_PACIENTE, SETOR, CD_CLINICA, NM_CLINICA, DS_LEITO, CD_ESPECIALIDADE, NM_ESPECIALIDADE, RESULTADO, EXAME, MATERIAL, MICROORGANISMO1, AMOSTRA, DROGA11, CIM11, SIR11, DROGA12, CIM12, SIR12, DROGA13, CIM13, SIR13, DROGA14, CIM14, SIR14, DROGA15, CIM15, SIR15, DROGA16, CIM16, SIR16, DROGA17, CIM17, SIR17, DROGA18, CIM18, SIR18, DROGA19, CIM19, SIR19, DROGA110, CIM110, SIR110, DROGA111, CIM111, SIR111, DROGA112, CIM112, SIR112, DROGA113, CIM113, SIR113, DROGA114, CIM114, SIR114, DROGA115, CIM115, SIR115, DROGA116, CIM116, SIR116, DROGA117, CIM117, SIR117, DROGA118, CIM118, SIR118, DROGA119, CIM119, SIR119, DROGA120, CIM120, SIR120, MICROORGANISMO2, CIM21, SIR21, CIM22, SIR22, CIM23, SIR23, CIM24, SIR24, CIM25, SIR25, CIM26, SIR26, CIM27, SIR27, CIM28, SIR28, CIM29, SIR29, CIM210, SIR210, CIM211, SIR211, CIM212, SIR212, CIM213, SIR213, CIM214, SIR214, CIM215, SIR215, CIM216, SIR216, CIM217, SIR217, CIM218, SIR218, CIM219, SIR219, CIM1, SIR1, OBSERVACAO.

Apêndice II

Colunas do conjunto de dados Master 284.

DT_PEDIDO, NU_PEDIDO, CD_ATENDIMENTO, CD_PACIENTE, NM_PACIENTE, SETOR, CD_CLINICA, NM_CLINICA, DS_LEITO, CD_ESPECIALIDADE, NM_ESPECIALIDADE, RESULTADO, CD_MNEMONICO, MATERIAL, BACTERIA, AMOSTRA, AMICACINA, AMPICILINA, AMPICILINA_SULB, AZTREONAM, CEFAZOLINA, CEFEPIMA, CEFTAZIDIMA, CEFTRIAXONA, CEFUROXIMA, CEFOXITINA, CIPROFLOXACINA, CLINDAMICINA, DAPTOMICINA, ERITROMICINA, GENTAMICINA, IMIPENEM, LEVOFLOXACINA, LINEZOLIDA, MEROPENEM, MINOCICLINA, NITROFURANTOINA, OXACILINA, PENICILINA, PIPERACILINA, QUINUPRISTINA, RIFAMPICINA, TETRACICLINA, TOBRAMICINA, TRIMETOPRIM, VANCOMICINA, OBSERVACAO.