

Análise de Séries Temporais dos Casos de Dengue em Sergipe

Mateus Ramos Silva

January 21, 2025

Abstract

Este estudo tem como objetivo a análise de séries temporais dos casos de dengue em Sergipe, com o intuito de identificar padrões sazonais e prever os casos futuros da doença. Os dados foram coletados do sistema DATASUS, com dois pacotes distintos: o primeiro abrangendo o período de 2007 a 2013, e o segundo de 2014 a 2024. Os dados foram combinados e transformados em uma série temporal no R, a partir da qual foi realizada a decomposição para identificar a tendência, a sazonalidade e os erros.

Foi realizado o teste de sazonalidade utilizando o teste de Dickey-Fuller, seguido da aplicação de um modelo ARIMA para modelar a série. Após a análise, os dados foram suavizados utilizando média móvel, e o modelo foi reavaliado. O modelo final apresentou um MAPE (Erro Percentual Absoluto Médio) de 16%, indicando que as previsões com o modelo ajustado possuem um bom desempenho, com um erro médio de 16% em relação aos valores reais.

1 Introdução

A dengue é uma das doenças virais mais prevalentes no Brasil, com surtos recorrentes que afetam principalmente as regiões tropicais. Em Sergipe, os casos de dengue têm mostrado variação ao longo dos anos, com picos de incidência em determinados períodos. Com o objetivo de entender melhor a dinâmica dessa doença, este estudo realiza uma análise de séries temporais utilizando dados dos casos de dengue em Sergipe entre 2007 e 2024.

A série temporal foi construída a partir de dois pacotes de dados: um referente ao período de 2007 a 2013, e o outro de 2014 a 2024. Estes dados foram combinados e transformados para serem analisados no software

R, permitindo a aplicação de técnicas estatísticas para detectar tendências, sazonalidade e comportamento da doença ao longo do tempo. Para modelar a série temporal, foram aplicados testes de estacionariedade, decomposição de séries e o modelo ARIMA, com o objetivo de prever os casos futuros e avaliar a eficácia das previsões.

Além disso, o processo de suavização dos dados foi realizado para minimizar o impacto de variações extremas, proporcionando uma análise mais precisa das tendências subjacentes. O desempenho do modelo foi avaliado utilizando o MAPE, com um resultado de 16%, indicando que o modelo é eficaz para prever os casos de dengue em Sergipe.

2 Metodologia

A metodologia adotada para a análise dos casos de dengue em Sergipe seguiu uma abordagem estruturada, com o objetivo de entender as tendências e padrões sazonais da doença. A seguir, descrevem-se as etapas do processo.

2.1 Coleta e Preparação dos Dados

Os dados utilizados nesta análise foram coletados do sistema DATASUS, que fornece informações sobre casos de dengue registrados em diversas regiões do Brasil. Foram utilizados dois pacotes de dados: o primeiro, referente ao período de 2007 a 2013, e o segundo, de 2014 a 2024. Os dados originais estavam organizados em formato mensal, com informações sobre o número de casos registrados a cada mês.

Após a coleta, os dados foram combinados em um único conjunto de dados para abranger o período de 2007 a 2024. A combinação foi realizada utilizando o software R, onde os dados de ambos os pacotes foram unidos e transformados para o formato adequado para análise de séries temporais. Para isso, foi utilizado o comando `pivot_longer` do pacote `tidyr`, e as variáveis de mês e ano foram transformadas para uma coluna de data.

2.2 Análise de Séries Temporais

Com os dados combinados, a primeira etapa foi a transformação dos dados em uma série temporal. Para isso, utilizou-se a função `ts` do R, com frequência mensal (12) e início em janeiro de 2007. Essa série temporal foi a base para todas as análises subsequentes.

A série temporal foi, então, decomposta utilizando a função `decompose` para identificar as três componentes principais: tendência, sazonalidade e

erro. Essa decomposição forneceu uma visão mais clara dos padrões subjacentes aos casos de dengue.

2.3 Teste de Estacionariedade e Sazonalidade

Para verificar a estacionariedade da série temporal, foi realizado o teste de Dickey-Fuller. Este teste avalia se a série apresenta uma raiz unitária, ou seja, se a série temporal é estacionária ou se é necessário diferenciar os dados.

Em seguida, foi aplicado o teste de sazonalidade, considerando a natureza da série e a possibilidade de padrões repetitivos ao longo dos meses. O objetivo foi identificar se há variação sazonal no número de casos de dengue ao longo do ano, o que é esperado devido ao caráter epidêmico da doença, que tem maior incidência em períodos chuvosos.

2.4 Modelagem ARIMA

Após a análise da decomposição e sazonalidade, foi ajustado um modelo ARIMA (AutoRegressive Integrated Moving Average) à série temporal. O modelo ARIMA é uma técnica amplamente utilizada para prever séries temporais com base em suas próprias observações passadas. Utilizou-se a função `auto.arima` do pacote `forecast` do R, que seleciona automaticamente os melhores parâmetros do modelo.

O modelo ARIMA foi avaliado com base em diversas métricas, incluindo o AIC (Akaike Information Criterion) e o BIC (Bayesian Information Criterion), que ajudam a selecionar o modelo mais adequado para os dados.

2.5 Suavização dos Dados

Como a série temporal de casos de dengue apresentou picos significativos, foi realizada a suavização dos dados utilizando média móvel. Essa técnica ajuda a reduzir a variabilidade extrema e permite uma melhor visualização das tendências subjacentes.

Para suavizar os dados, foi utilizada a função `rollmean` do pacote `zoo`, com uma janela de 12 meses. O objetivo foi observar a tendência de longo prazo dos casos de dengue, minimizando o impacto das flutuações mensais.

2.6 Avaliação do Modelo

Após a aplicação do modelo ARIMA e da suavização, o desempenho do modelo foi avaliado utilizando o MAPE (Erro Percentual Absoluto Médio). O

MAPE é uma métrica comum para avaliar a precisão de modelos de previsão. Para esse estudo, o MAPE foi calculado para as previsões realizadas com os dados suavizados, e o valor obtido foi de 16%, indicando uma boa performance do modelo.

3 Resultados

3.1 Teste de Dickey-Fuller

O teste de Dickey-Fuller foi realizado para verificar a estacionariedade da série temporal dos casos de dengue em Sergipe. O valor do teste Dickey-Fuller foi de -5.6622, com um p-valor de 0.01. Como o p-valor é menor do que o nível de significância adotado (geralmente 0.05), podemos rejeitar a hipótese nula de que a série temporal possui uma raiz unitária. Portanto, concluímos que a série temporal é estacionária.

A estacionariedade é uma condição importante para a modelagem de séries temporais, pois séries não estacionárias podem levar a modelos de previsão imprecisos. Com a série temporal sendo estacionária, podemos avançar para a construção de modelos de previsão, como o ARIMA.

Saída do R:

Augmented Dickey-Fuller Test

```
data: ts_dengue
Dickey-Fuller = -5.6622, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

3.2 Modelo ARIMA

Foi ajustado um modelo ARIMA(3,0,2)(2,0,0)[12] aos dados da série temporal dos casos de dengue em Sergipe. O modelo contém os seguintes coeficientes:

- $AR(1) = 0.6131$, $AR(2) = -0.1007$, $AR(3) = -0.0233$
- $MA(1) = 0.2581$, $MA(2) = -0.1286$
- $SAR(1) = -0.0025$, $SAR(2) = -0.0722$
- Média = 353.6623

A variância dos erros é de 766.000, e o logaritmo da verossimilhança foi de -1766.15. O AIC (Critério de Informação de Akaike) é 3550.3, o AICc é 3551.18 e o BIC (Critério de Informação Bayesiano) é 3580.68.

As medidas de erro do conjunto de treinamento são:

- Erro Médio (ME) = 3.68
- Raiz do Erro Quadrático Médio (RMSE) = 858.85
- Erro Absoluto Médio (MAE) = 222.47
- Erro Percentual Médio (MPE) = -204.24
- Erro Percentual Absoluto Médio (MAPE) = 233.23
- Erro Absoluto Médio Escalonado (MASE) = 0.42
- Autocorrelação do primeiro erro (ACF1) = 0.004

Saída do R:

Series: ts_dengue

ARIMA(3,0,2)(2,0,0)[12] with non-zero mean

Coefficients:

	ar1	ar2	ar3	ma1	ma2	sar1	sar2	mean
	0.6131	-0.1007	-0.0233	0.2581	-0.1286	-0.0025	-0.0722	353.6623
s.e.	2.0257	0.9499	0.2250	2.0202	0.8568	0.0678	0.2040	121.8292

sigma^2 = 766000: log likelihood = -1766.15

AIC=3550.3 AICc=3551.18 BIC=3580.68

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	3.683449	858.8537	222.4657	-204.24	233.2273	0.4182495	0.004037816

O modelo ARIMA ajustado apresenta resultados muito elevados, indicando que o modelo não possui um nível de previsão adequado, o que pode ser devido a picos de casos; por esse motivo, foi feita uma suavização do modelo e refeito os testes.

3.3 Teste de Dickey-Fuller para a Série Suavizada

Para verificar a estacionariedade da série temporal suavizada dos casos de dengue em Sergipe, foi aplicado o teste de Dickey-Fuller aumentado (ADF). Os resultados obtidos são apresentados a seguir:

- Estatística de Dickey-Fuller: -4.9896
- Ordem de defasagem: 5
- Valor-p: 0.01

A hipótese alternativa do teste afirma que a série é estacionária. Como o valor-p ($p = 0.01$) é menor do que o nível de significância usual ($\alpha = 0.05$), rejeita-se a hipótese nula de não estacionariedade. Assim, conclui-se que a série temporal suavizada é estacionária.

Saída do R:

Augmented Dickey-Fuller Test

```
data: ts_dengue_suavizados
Dickey-Fuller = -4.9896, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

3.4 Ajuste do Modelo ARIMA na Série Temporal Suavizada

Após a suavização dos dados, foi ajustado um modelo ARIMA para capturar a dinâmica temporal dos casos de dengue em Sergipe. O modelo ajustado foi identificado como **ARIMA(1,1,1)(2,0,1)[12]**, indicando uma diferenciação simples ($d = 1$) para alcançar estacionariedade e sazonalidade modelada com parâmetros adicionais.

- **Coefficientes estimados:**
 - AR(1): 0.3659 ± 0.0883
 - MA(1): 0.4974 ± 0.0814
 - SAR(1): -0.4318 ± 0.2115
 - SAR(2): -0.1813 ± 0.2024
 - SMA(1): -0.5136 ± 0.2084
- Variância residual (σ^2): 5579

- Log-verossimilhança: -1173.21
- Critérios de informação:
 - AIC: 2358.41
 - AICc: 2358.84
 - BIC: 2378.32

Métricas de erro para o conjunto de treinamento:

- Erro médio (ME): -3.31
- Raiz do erro quadrático médio (RMSE): 73.59
- Erro absoluto médio (MAE): 24.57
- Erro percentual médio (MPE): -8.62%
- Erro percentual absoluto médio (MAPE): 16.20%
- Erro absoluto médio padronizado (MASE): 0.061
- Primeira autocorrelação dos resíduos (ACF1): 0.0034

Os valores de *AIC* e *BIC* indicam um bom ajuste do modelo ao conjunto de dados. Além disso, o MAPE de 16.20% mostra que o modelo possui uma precisão aceitável para previsões.

Saída do R:

```
Series: ts_dengue_suavizados
ARIMA(1,1,1)(2,0,1)[12]
```

Coefficients:

	ar1	ma1	sar1	sar2	sma1
	0.3659	0.4974	-0.4318	-0.1813	-0.5136
s.e.	0.0883	0.0814	0.2115	0.2024	0.2084

```
sigma^2 = 5579: log likelihood = -1173.21
AIC=2358.41 AICc=2358.84 BIC=2378.32
```

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-3.312467	73.59093	24.5657	-8.618777	16.19647	0.06092293	0.00339458

3.5 Comparação entre Modelos com Dados Não Suavizados e Suavizados

Os gráficos apresentados mostram as previsões baseadas nos modelos ARIMA para os casos de dengue em Sergipe antes e depois da suavização dos dados (Figura 1 e Figura 2). As principais diferenças entre os dois resultados podem ser interpretadas da seguinte forma:

- **Modelo com Dados Não Suavizados (Figura 1):**
 - O modelo $ARIMA(3,0,2)(2,0,0)[12]$ foi ajustado aos dados originais, que incluem o pico extremo de 2008.
 - Este pico influencia significativamente o ajuste e as previsões, resultando em um modelo menos sensível às tendências e padrões sazonais dos períodos mais recentes.
 - As previsões para 2025 apresentam intervalos de confiança amplos, indicando maior incerteza devido à alta variabilidade dos dados.
- **Modelo com Dados Suavizados (Figura 2):**
 - Após a suavização, o pico extremo de 2008 foi atenuado, e o modelo $ARIMA(1,1,1)(2,0,1)[12]$ foi ajustado aos dados resultantes.
 - A suavização possibilitou um ajuste mais adequado às tendências e padrões sazonais, especialmente nos anos mais recentes.
 - As previsões apresentam intervalos de confiança mais estreitos, indicando maior precisão e menor incerteza nas estimativas.
 - Este modelo reflete melhor a dinâmica subjacente dos casos de dengue ao longo do tempo, desconsiderando flutuações atípicas como o pico de 2008.

Forecasts from ARIMA(3,0,2)(2,0,0)[12] with non-zero m

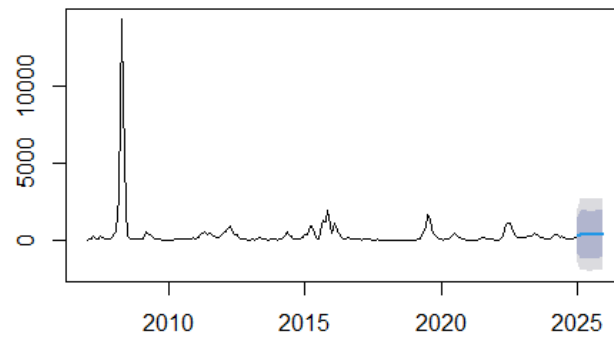


Figure 1: Previsões do modelo ARIMA com dados não suavizados.

Forecasts from ARIMA(1,1,1)(2,0,1)[12]

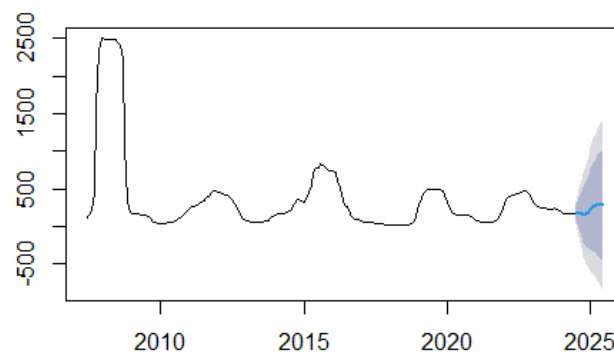


Figure 2: Previsões do modelo ARIMA com dados suavizados.

Interpretação Geral: A suavização dos dados permitiu uma análise mais robusta e previsões mais confiáveis, eliminando os efeitos de valores extremos. O modelo ajustado aos dados suavizados é mais apropriado para capturar tendências e padrões sazonais reais, com menor erro percentual médio absoluto (MAPE) e menor desvio.