Leonardo Meireles - NºUSP: 4182085, Antonio Moreira - NºUSP: 9779242

Análise de Séries Temporais Utilizando Modelos Lineares Dinâmicos

Brasil

2020, v-1

Leonardo Meireles - NºUSP: 4182085, Antonio Moreira - NºUSP: 9779242

Análise de Séries Temporais Utilizando Modelos Lineares Dinâmicos

Projeto final da matéria Séries Temporais e Aprendizado Dinâmico apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP em que o objetivo é familiarizar os alunos com modelos lineares dinâmicos e suas aplicações

Universidade de São Paulo - USP

Instituto de Ciências Matemáticas e de Computação

Programa de Graduação

Brasil

2020, v-1

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

FAC	Função de Autocorrelação
FACP	Função de Autocorrelação Parcial
ST	Série Temporal
DLM	Dynamic Linear Model
MSE	Mean Squared Error
MAE	Mean Absolute Error

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence
δ	Letra grega minúscula delta

Sumário

Introdução

Este documento tem como objetivo detalhar as atividades realizadas do projeto final da matéria *SME0808 - Séries Temporais e Aprendizado Dinâmico*, cuja principal tarefa analisar uma ou mais séries temporais via modelos lineares dinâmicos (*DLM - Dynamic Linear Model*).

A ideia principal introduzida com estes novos modelos é que conforme o tempo corre a informação de certa forma "envelhece", isto é, ao longo do tempo novas tendências podem surgir nos dados aumentando a incerteza nos mesmos e nosso modelo deve ser capaz de capturar tal efeito.

Assim, para capturar esses efeitos temporais é necessário introduzir elementos capazes de representar a evolução temporal do modelo ao longo do tempo, assim, a ideia é introduzir um novo modelo parametrizado por θ_t (o estado do processo no tempo t), que irá caracterizar tal evolução temporal.

Ademais, uma análise descritiva da série será apresentada, tal análise visa entender o comportamento da série, identificar tendências e possíveis sazonalidades. Finalmente após entender os padrões da série, um modelo linear dinâmico será proposto a fim de modelar e representar a série em questão.

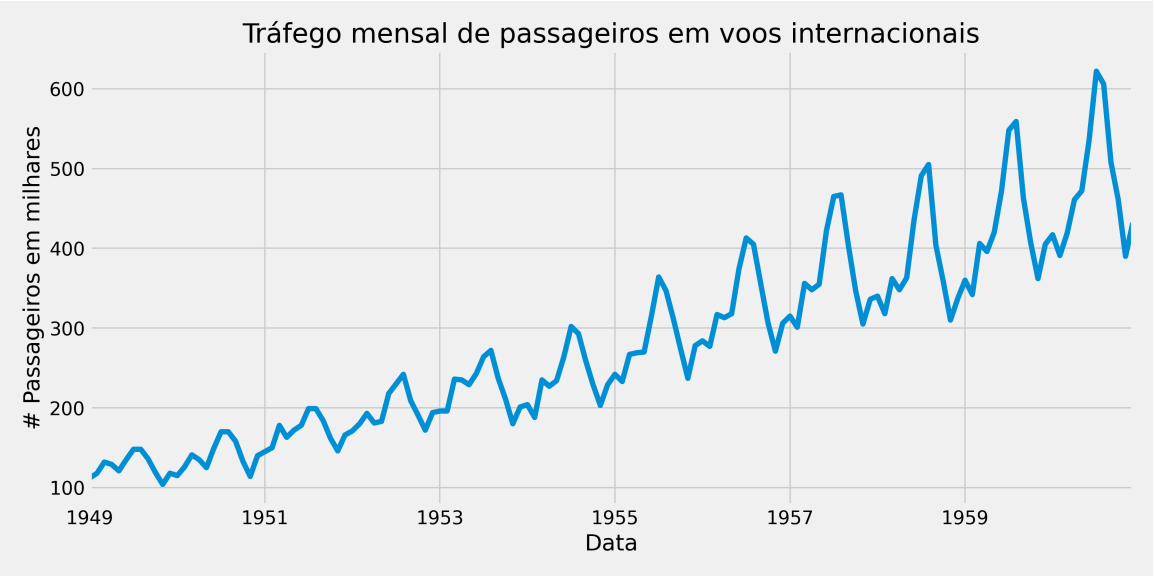
1 Análise descritiva da série temporal

Este capitulo tem como objetivo apresentar a série temporal de estudo, assim como uma breve análise descritiva das características da mesma.

1.1 Descrição da série

A série temporal utilizada neste projeto consiste do tráfego mensal de passageiros em voos internacionais entre 1949 e 1960, este clássico *dataset* foi retirado do livro Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1976) *Time Series Analysis, Forecasting and Control*. Third Edition. Holden-Day. Series G. A escolha deste dado deve-se ao fato de apresentar nitidamente características interessantes a serem exploradas na presente análise, como sazonalidade e tendência. A representação gráfica da série temporal, exibindo a número de passageiros (em milhares) em função do tempo, pode ser vista na Figura 1.

Figura 1 – Série temporal que será utilizada no estudo



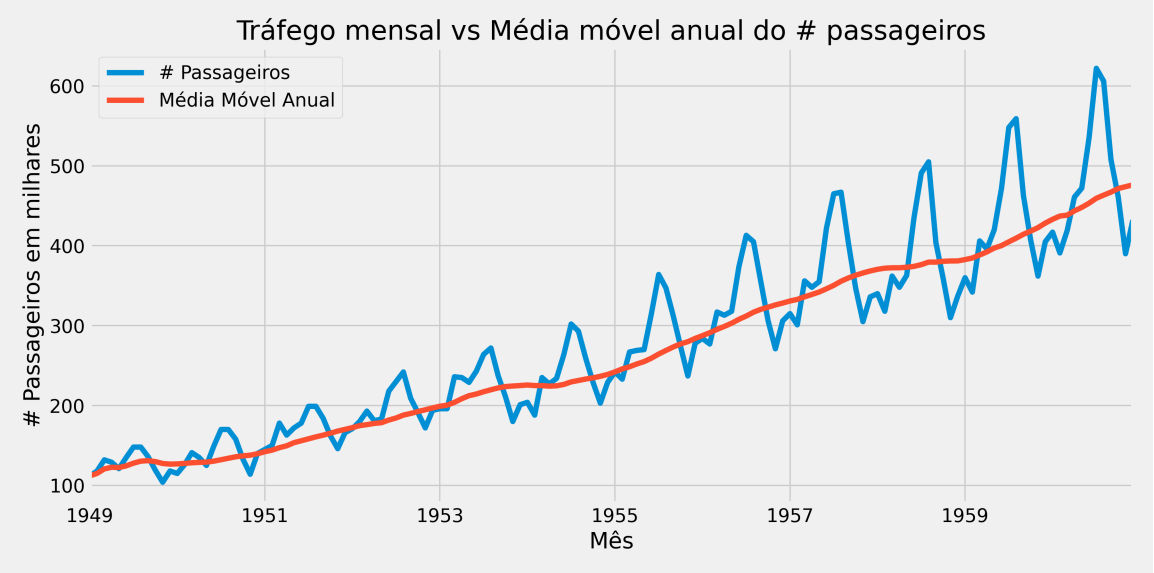
Como estamos tratando de uma série temporal em tempo discreto, para efeito de notação, denotaremos tal processo estocástico como $\{y_t, t \in [0, 143] \subset N\}$.

1.2 Estacionariedade e Tendência

Ao analisar séries temporais devemos checar se a mesma demonstra um comportamento estacionário ao longo do tempo. Da definição de estacionariedade fraca, podemos checar se a média dos valores da série ao longo do tempo se mantém constante, o que não ocorre na série em questão, basta observarmos a Figura 2 onde notamos um compor-

tamento de crescimento da média móvel anual (curva em vermelho) em sobreposição a série original (curva em azul). Este comportamento também nos fornece um indicativo gráfico de tendência positiva de crescimento que, por conseguinte, no presente trabalho a modelagem buscará exprimir como um comportamento linear ao longo do tempo.

Figura 2 – Série temporal e sua média móvel anual



1.3 Sazonalidade

Para verificar se a série possui algum comportamento sazonal podemos olhar como sua função de autocorrelação amostral $\widehat{\rho}$ se comporta, um indicativo de sazonalidade é caso ela possua valores altos em intervalos fixos de tempo. A FAC observada na Figura 3 apresenta tal comportamento descrito anteriormente com um intervalo de 12 meses.

Além dá análise de FAC, também podemos utilizar a primeira diferença como técnica de detecção de sazonalidade, com tal técnica as novas observações passam a ser derivadas da seguinte maneira $y'_t = y_t - y_{t-1}$. A nova série diferenciada amortece os impactos da tendência facilitando a análise gráfica da sazonalidade vide Figura 4 em que também é observada uma sazonalidade apontada pela ocorrência anual dos três picos nos meses de dezembro, março e julho.

Figura 3 – Gráficos da FAC e FACP com 26 lags mensais

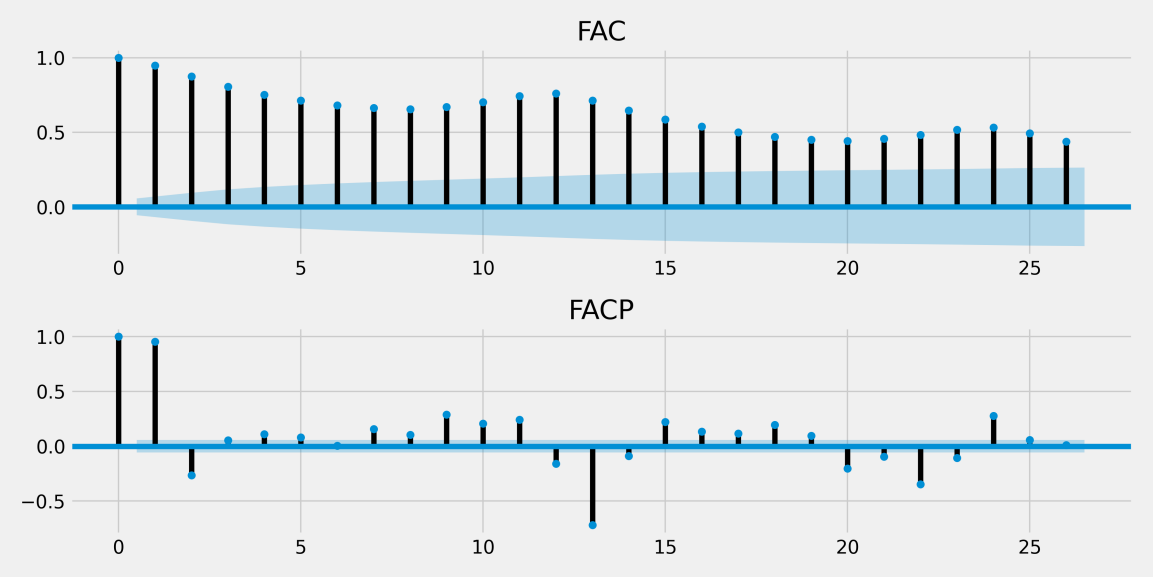
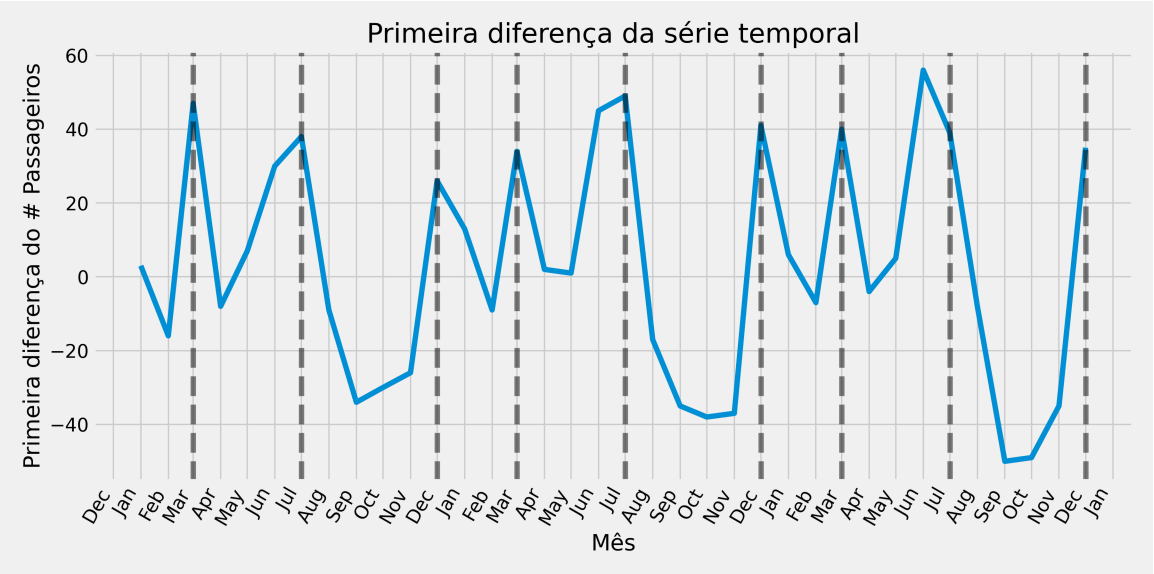


Figura 4 – Primeira diferença da série entre os anos 1954 e 1956



2 Metodologia

Neste capítulo buscamos introduzir o leitor a alguns conceitos essenciais para o entendimento da escolha do modelo utilizado.

2.1 Considerações Iniciais

Diversas técnicas de modelagem de séries temporais são baseadas no conceito de suavização do dado (*smoothing*), isto é, decompor a série como a soma de duas componentes, uma componente "suavizada" e outra componente que contém as características que a componente suavizada não pode descrever. Este procedimento é similar ao conceito de "sinal mais ruído" em processamento de sinais. Duas maneiras para suavizarmos uma série temporal consistem em:

- Ajustar uma regressão linear para descrever a componente de tendência ou, de maneira mais geral, ajustar uma regressão polinomial. Suavizar por uma regressão polinomial consiste, basicamente, em identificar (estimar) os parâmetros $\beta_i, i = 0, ..., p$ do modelo, onde ϵ_t é um erro aleatório:

$$y_t = \left(\sum_{i=o}^{p} \beta_i \cdot t^i \right) + \epsilon_t, \epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$$

- De modo similar, uma regressão harmônica nos permite descrever os ciclos da série temporal, ou seja, descrever a componente sazonal. Então se desejarmos remover as componentes periódicas com frequências $w_1, ..., w_p$, precisamos estimar $a_1, b_1, ..., a_p, b_p$ do seguinte modelo, novamente com um erro ϵ_t aleatório:

$$y_t = \left(\sum_{j=1}^{p} a_j cos(2\pi w_j t) + b_j sin(2\pi w_j t) \right) + \epsilon_t, \epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$$

Podemos notar que independentemente da abordagem que decidamos seguir é necessária a estimação de certos parâmetros, que aqui denotaremos pelo vetor $\theta = (\theta_0, ..., \theta_p)$, de tal forma que nossa série temporal $\{y_t\}$ dependa de tais parâmetros. Quando visualizamos desta forma o nos deparamos com um problema de **inferência Bayesiana**. Neste contexto, pelo Teorema de Bayes, podemos obter a distribuição a *posteriori* $p(\theta|y_t)$ da utilizando:

1. **Função de verossimilhança**: $p(y_t|\theta)$
2. **Distribuição a *priori***: $p(\theta)$
3. **Função de densidade preditiva**: $p(y_t) = \int p(\theta)p(y_t|\theta)\ \mathrm{d}\theta$

Tomando o parâmetro θ como um vetor desconhecido, a distribuição a *priori* indica a incerteza do parâmetro e é construída analisando e quantificando informações históricas da série. Para obter mais informações sobre o parâmetro θ devemos observar os dados, a distribuição conjunta do mesmo como função do vetor de parâmetros é a função de verossimilhança. Para obtermos a a distribuição a *posteriori*, a incerteza sobre θ é atualizada à luz da nova informação e pode ser calculada a partir do Teorema de Bayes:

$$p(\theta|y_t) = \frac{p(\theta)p(y_t|\theta)}{p(y_t)}$$

2.2 Justificativa do modelo e processo de treinamento

A motivação para a escolha de um modelo linear dinâmico (**DLM**) para representar a série temporal deve-se ao fato do mesmo surgir da formulação dos espaços de estados e como uma solução natural para estruturar séries temporais com componentes não-estacionárias. A estrutura geral para a classe de **modelos lineares dinâmicos para séries temporais univariadas com observações igualmente espaçadas** pode ser descrita pelas equações:

$$y_t = F_t'\theta_t + \nu_t$$
$$\theta_t = G_t\theta_{t-1} + w_t$$

Onde,

- θ_t: é o vetor de estados no tempo t.
- F_t: é um vetor p-dimensional de constantes conhecidas ou regresssores no tempo t.
- ν_t: é o erro de observação no tempo t. $\nu_t \sim \mathcal{N}(0, v_t)$.
- G_t: é uma matriz bloco-diagonal conhecida como matriz de evolução no tempo t.
- w_t: é o erro de evolução no tempo t. $w_t \sim \mathcal{N}(0, W_t)$.

A partir desta estrutura geral podemos modelar a série de muitas maneiras, neste trabalho faremos uso do **aprendizado sequencial**. Aqui F_t e G_t não variam ao longo do tempo. Além disso, para permitir uma maior variabilidade na transição dos estados foi utilizado um fator de desconto δ no lugar de fixarmos uma matriz de variâncias e covariâncias para os erros, neste caso:

$$W_t = P_t \left(\frac{1-\delta}{\delta} \right) \quad \forall \delta \in (0, 1]$$
$$P_t = GC_tG'$$

A incerteza do modelo, conforme a variabilidade no fator de desconto é quantificada como proporção de P_t, que por sua vez pode ser interpretada como a matria de variância a priori em um modelo sem erro de evolução.

A compreensão da escolha do modelo é simples: *por serem lineares, tais modelos são aditivos tornando possível a superposição de modelos. De modo prático, é possível tratar a sazonalidade e a tendência de modo conjunto.*

Neste trabalho, como temos dados mensais, i.e., com periodicidade $p = 12$, modelamos a série fazendo a superposição de um modelo polinomial de de primeiro grau com um modelo sazonal harmônico.

Deste modo, o modelo final é ($\alpha_{t,m}$ é devido a paridade de periodicidade):

$$y_t = \mu_t + \left(\sum_{j=1}^{m-1} \alpha_{t,j} \right) + \alpha_{t,m} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, v_t)$$

$$\mu_t = \mu - t - 1 + w_t, w_t \sim \mathcal{N}(0, P_t \left(\frac{1-\delta}{\delta} \right))$$

As componentes principais do modelo podem ser vistas como, onde cada $G_i, i \in [1, m-1]$ é uma matriz harmônica:

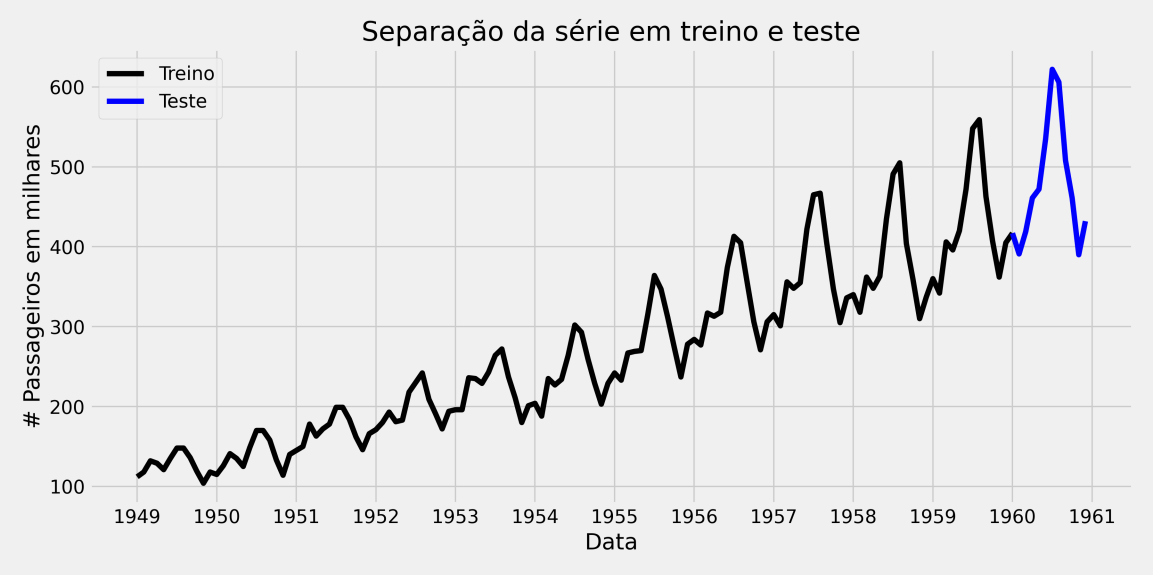
$$F = (1, 1, 0, 1, ..., 0, 1)$$
$$G = bloco\_diagonal(1, G_1, ..., G_{m-1}, -1)$$

Como dito anteriormente, o modelo consiste na concatenação do modelo linear com o o modelo harmônico, em seguida o algoritmo realiza o processo de análise retrospectiva da série (*Smooth*).

2.3 Preparação dos dados

Nenhum pré-processamento especial foi necessário pois não foi detectado *outliers* nos dados somente a separação dos dados em conjuntos de treino e teste representado na Figura 5.

Figura 5 – Separação dos dados em conjuntos de treino e teste



2.4 Bibliotecas utilizadas

O projeto foi desenvolvido em *Python3* utilizando o *Jupyter Notebook* e as bibliotecas utilizadas para desenvolvimento estão listadas abaixo:

- numpy: módulo para cálculo científico;
- pandas: biblioteca de análise exploratória de dados;
- pyDLM: biblioteca contendo a implementação do pacote dlm do *R* em *Python3*;
- matplotlib: módulo para criação de gráficos.

Os códigos e arquivos podem ser encontrados em repositório do projeto Github.

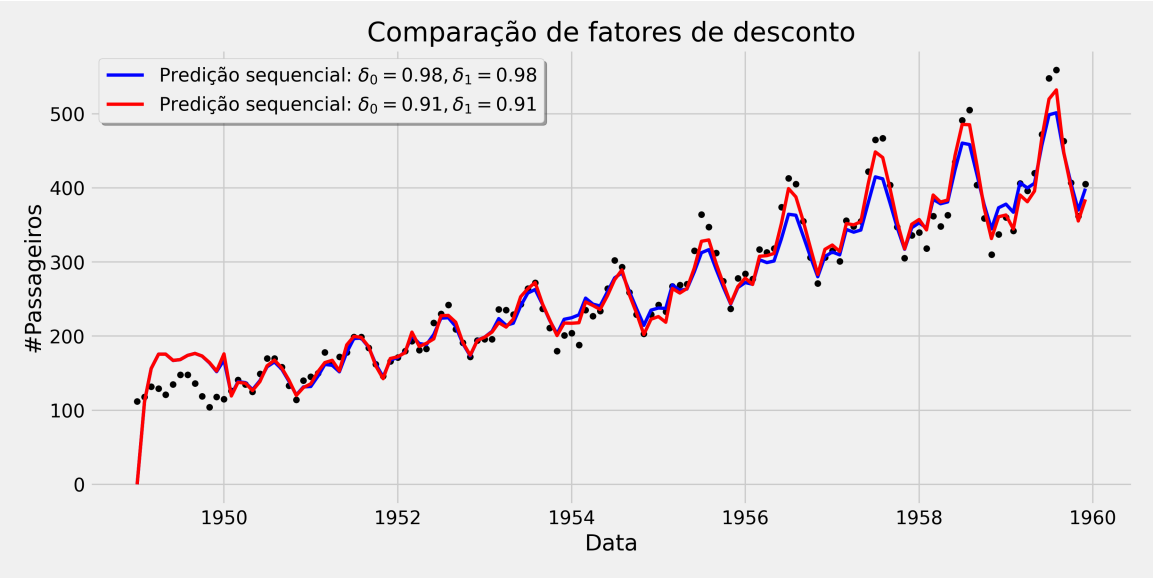
3 Resultados

Este capitulo tem como objetivo mostrar os resultados obtidos com a aplicação da metodologia discutida.

3.1 Fator de desconto

Para visualizar a influência dos fatores de desconto no modelo, simulações com diferentes valores foram realizadas, tal experimento é exemplificado na Figura 6. Nota-se pela figura que quanto menor o fator maior flexibilidade o modelo terá para capturar novos efeitos, entretanto, gera maior variabilidade consequentemente podendo gerar mais erros. Em suma, deve-se atentar para valores baixos pois estes podem causar o *overfitting* do modelo.

Figura 6 – Estimador da componente de tendencia do melhor modelo



3.2 Componentes

Após a escolha do fatores de desconto iguais a 0.91 para cada componente do modelo o modelo foi treinado gerando as estimações das componentes de tendência (Figura 7) e sazonalidade (Figura 8).

Figura 7 – Estimador da componente de tendencia do melhor modelo

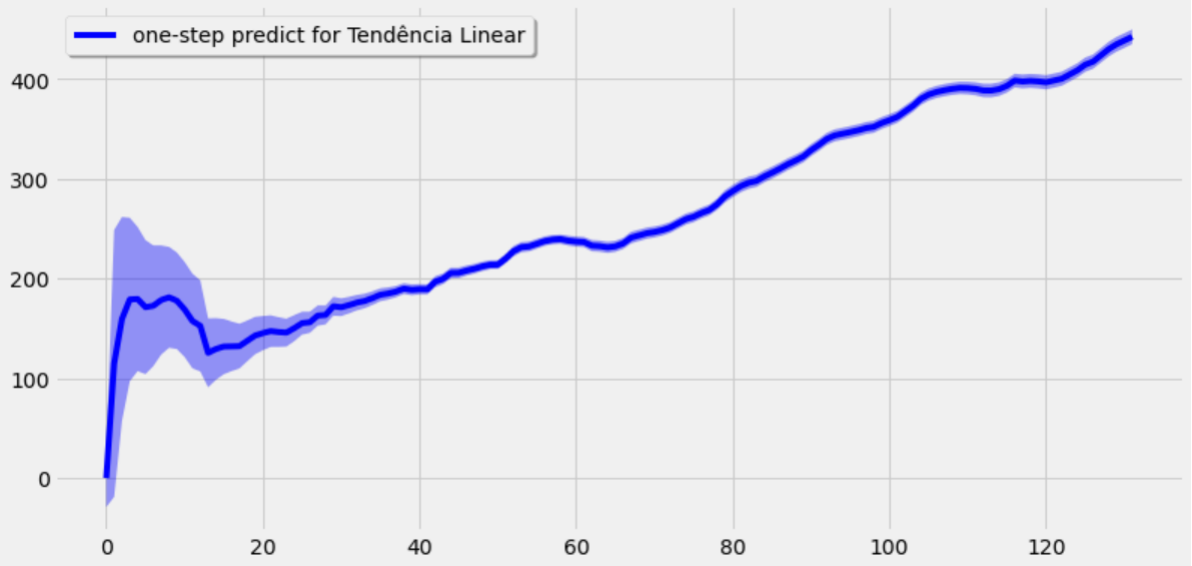
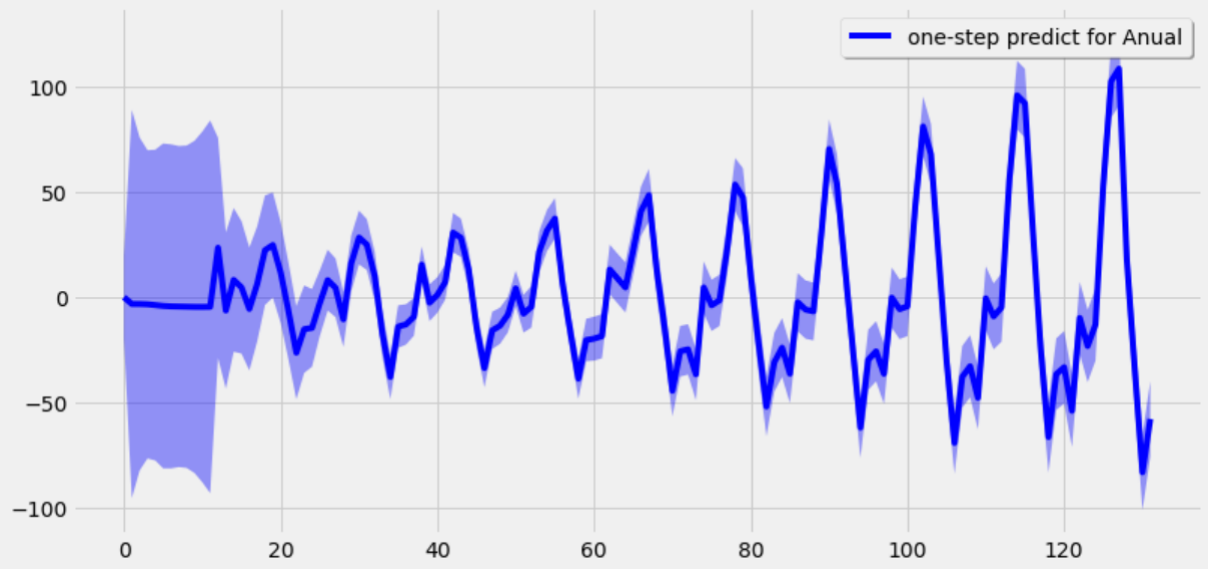


Figura 8 – Estimador da componente sazonal do melhor modelo



3.3 Avaliação do modelo

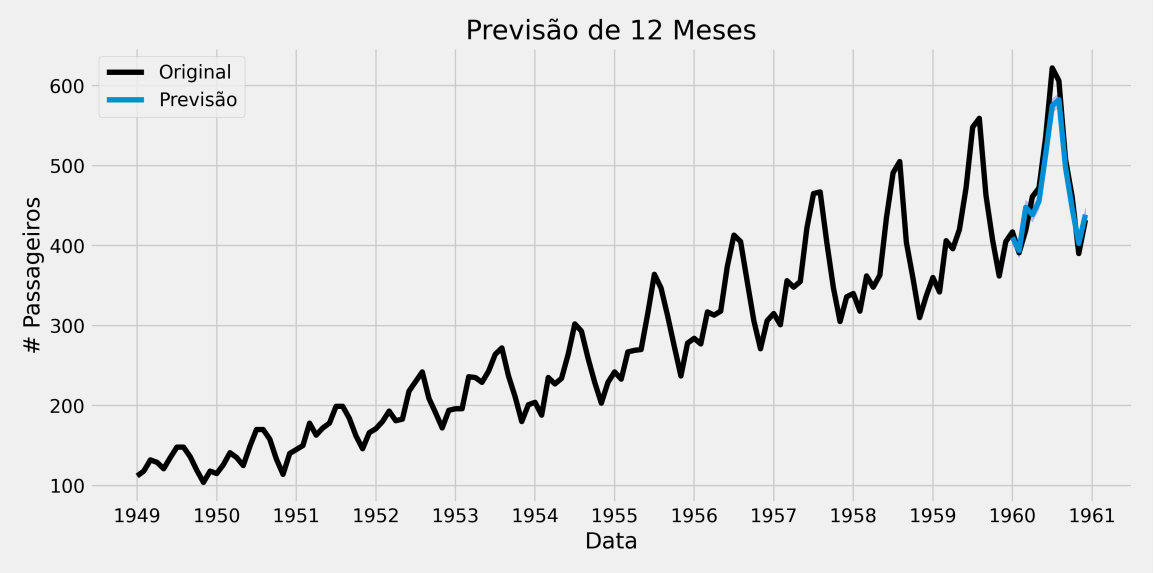
Para avaliar a performance do modelo métricas como MSE, MAE e R^2 foram calculadas entre os dados de teste e os dados originais, abaixo os valores das métricas do melhor modelo:

Tabela 1 – Métricas de avaliação.

Métrica	Resultado
R^2	0.92
MAE	17.34
MSE	410.40

A Figura 9 mostra a previsão em comparação com a série original, nota-se que o modelo foi capaz de capturar a tendência assim como também a sazonalidade anual da série.

Figura 9 – Previsão em comparação com série original



3.4 Custo computacional

De acordo com a biblioteca pyDLM a complexidade de previsão sequencial é $\mathcal{O}(n)$ fazendo com que o treinamento e previsão de n passos chegue a uma complexidade quadrática $\mathcal{O}(n^2)$.

4 Conclusão

Neste trabalho modelamos uma série temporal com clara tendência e sazonalidade usando um modelo linear dinâmico (DLM), que consiste na atualização progressiva dos parâmetros descritivos da série ao longo do tempo. O modelo foi criado a partir da concatenação de uma componente linear, para descrição da tendência, e de uma componente harmônica, para explica a sazonalidade, i.e., os possíveis ciclos. Os erros associados à atualização paramétrica foram descritos pelo fator de desconto, o qual variamos o valor para permitir uma melhor visualização do impacto deste parâmetro, chegando a conclusão que quanto menor este valor maior a possibilidade de variabilidade nos estados com a consequência de maior incerteza no modelo.

Como tratamos de uma série simples, a consequente também simples modelagem foi suficiente para descrever o comportamento da série, algo que provavelmente seria impossível para séries mais complexas. Para trabalhos futuros seria interessante tratar de novas séries permitindo maior flexibilidade na variação paramétrica, como por exemplo tirar a restrição de constância dos parâmetros F_t e G_t, com componentes regressoras.