Series Temporais

Fundamentos de previsão e análise

Prof. Dr. Rafael Traldi Moura

Prof. Dr. Marlon Sproesser Mathias

Objetivos de hoje

O objetivo dessa aula é apresentar algumas ferramentas para realizar previsões e analisar séries temporais. Essas ferramentas são:

- Métodos simples de previsão;
- Formas de simplificar a tarefa de previsão usando transformações e ajustes dos dados;
- Métodos para avaliar a qualidade de previsões e desempenho de um modelo;
- Técnicas para calcular intervalos de confiança;
- Técnica para identificar anomalia do tipo "outliner".

Objetivos de hoje

Os métodos simples para realizar previsões são:

1. Método base;
2. Método base para dados com sazonalidade;
3. Método de tendência;
4. Média móvel;
5. Média móvel ponderada.

Modelo base

Vamos começar usando um modelo baseado em uma hipótese ingênua: "o valor da série no instante de tempo seguinte é igual ao valor atual".

Esse modelo é definido matematicamente por:

$$\hat{y}_t = y_{t-1}$$

onde y_{t-1} é a amostra no instante de tempo $t - T_a$, T_a é o período de amostragem (intervalo de tempo entre a coleta de duas amostras) e $\hat{y}_t$ é o valor previsto para a série no instante de tempo t.

Modelo base para séries com sazonalidade

Existe um método similar ao método base, que é útil para séries com sazonalidade bem definida.

Nesse caso a previsão é definida como sendo igual ao último valor observado da mesma temporada anterior, por exemplo, para uma série que apresenta sazonalidade anual usa-se o valor da amostra do mesmo mês do ano anterior.

Nesse método, a previsão para o próximo instante de tempo é dada por:

$$\hat{y}_t = y_{t-s}$$

na qual $\hat{y}_t$ é o valor previsto para a série no instante de tempo t ,onde y_{t-s} é a amostra no instante de tempo equivalente da temporada anterior e s é o número de amostras de cada temporada.

Modelo de tendência simples

Uma variação do método base é introduzir na previsão a tendência dos dados. Essa tendência é calculada considerando que a mudança ao longo do tempo é definida como a variação entre a última amostra e a amostra m instantes de tempo atrás.

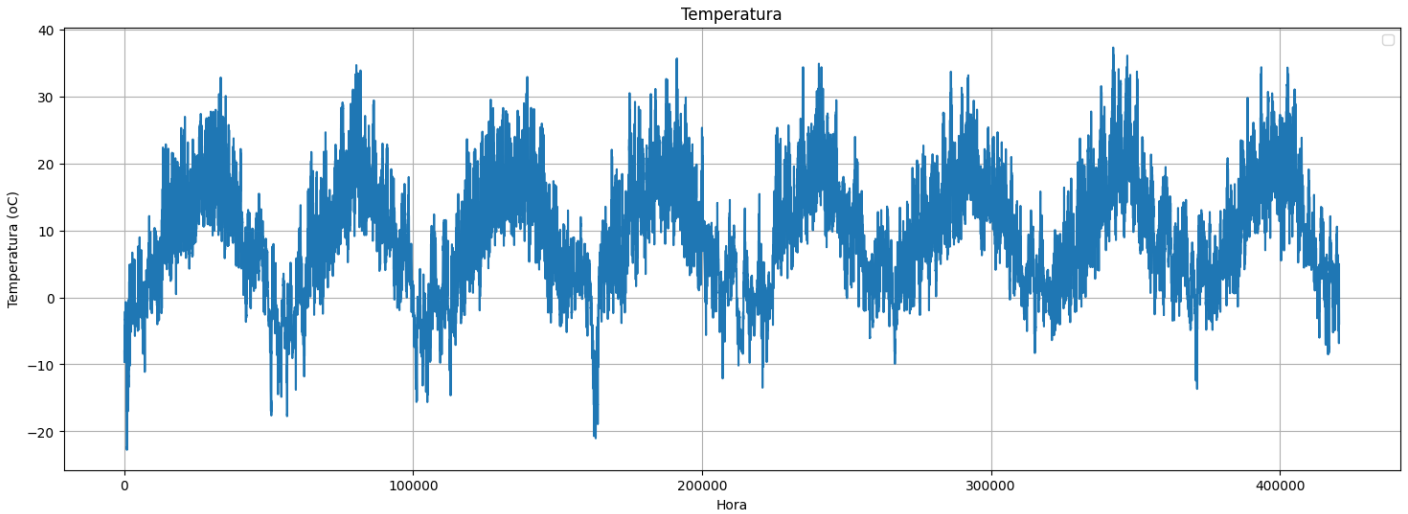
Nesse caso, a previsão para o tempo t, é dada por:

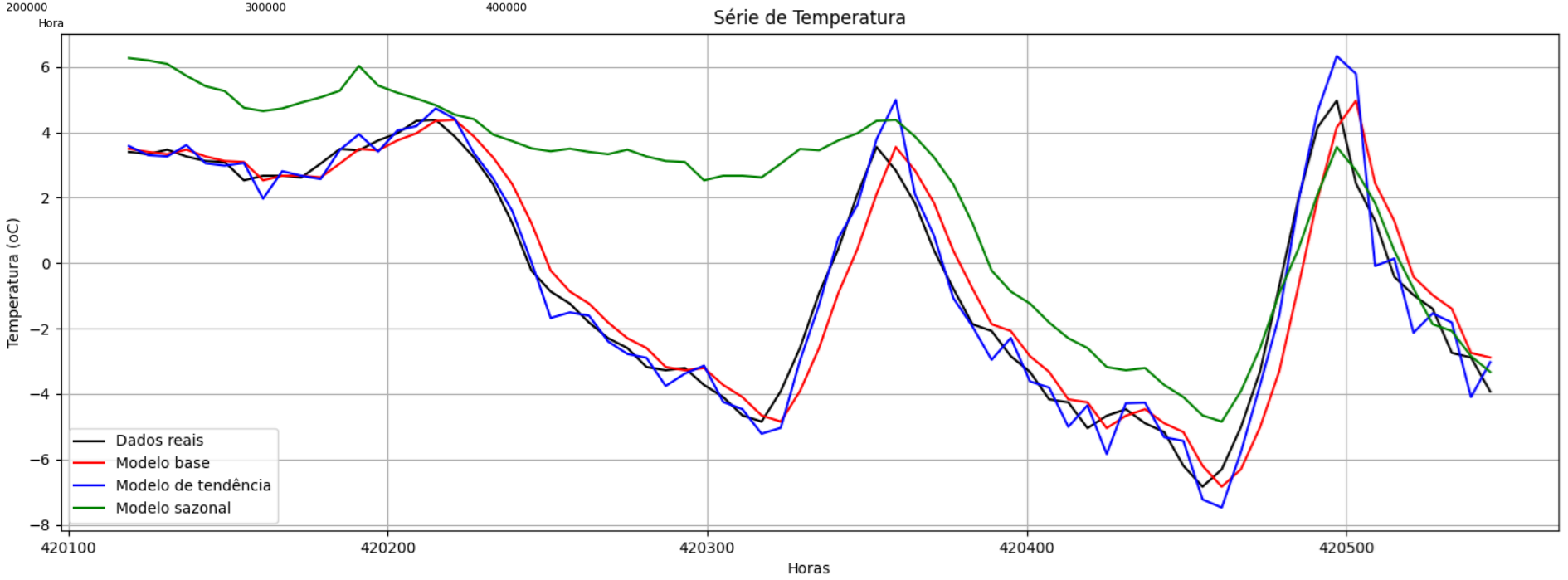
$$\hat{y}_t = y_{t-1} + \frac{(y_{t-1} - y_{t-m})}{m-1}$$

na qual $\hat{y}_t$ é o valor previsto para a série no instante de tempo t, y_{t-1} é a amostra no instante de tempo $t - T_a$, y_{t-m} é a amostra m instantes anteriores.

Isso equivale a traçar uma linha entre a primeira e a m-ésima amostra anterior e extrapolá-la para o futuro.

Exemplo: série de temperatura





Modelo de tendência simples

Uma variação do método base é introduzir na previsão a tendência dos dados. Essa tendência é calculada considerando que a mudança ao longo do tempo é definida como a variação entre a última amostra e a amostra m instantes de tempo atrás.

Nesse caso, a previsão para o tempo t, é dada por:

$$\hat{y}_t = y_{t-1} + \frac{(y_{t-1} - y_{t-m})}{m-1}$$

na qual $\hat{y}_t$ é o valor previsto para a série no instante de tempo t, y_{t-1} é a amostra no instante de tempo $t - T_a$, y_{t-m} é a amostra m instantes anteriores.

Isso equivale a traçar uma linha entre a primeira e a m-ésima amostra anterior e extrapolá-la para o futuro.

Transformações e pré-processamento

Ajustar e pré-processar os dados de uma série temporal muitas vezes torna a tarefa de **previsão/análise mais simples**. Existem muitas forma de pré-processar os dados de uma série temporal e a melhor forma depende de cada série.

Vamos ver quatro tipos de ajustes e pré-processamento:

1. Ajustes de calendário;
2. Ajustes de população;
3. Ajustes de inflação;
4. Transformações matemáticas.

O objetivo de realizar ajustes e transformações nos dados é simplificar os padrões presentes, removendo fontes conhecidas de variação ou tornando o padrão mais consistente em todo o conjunto de dados → **padrões mais simples geralmente levam a previsões mais precisas**.

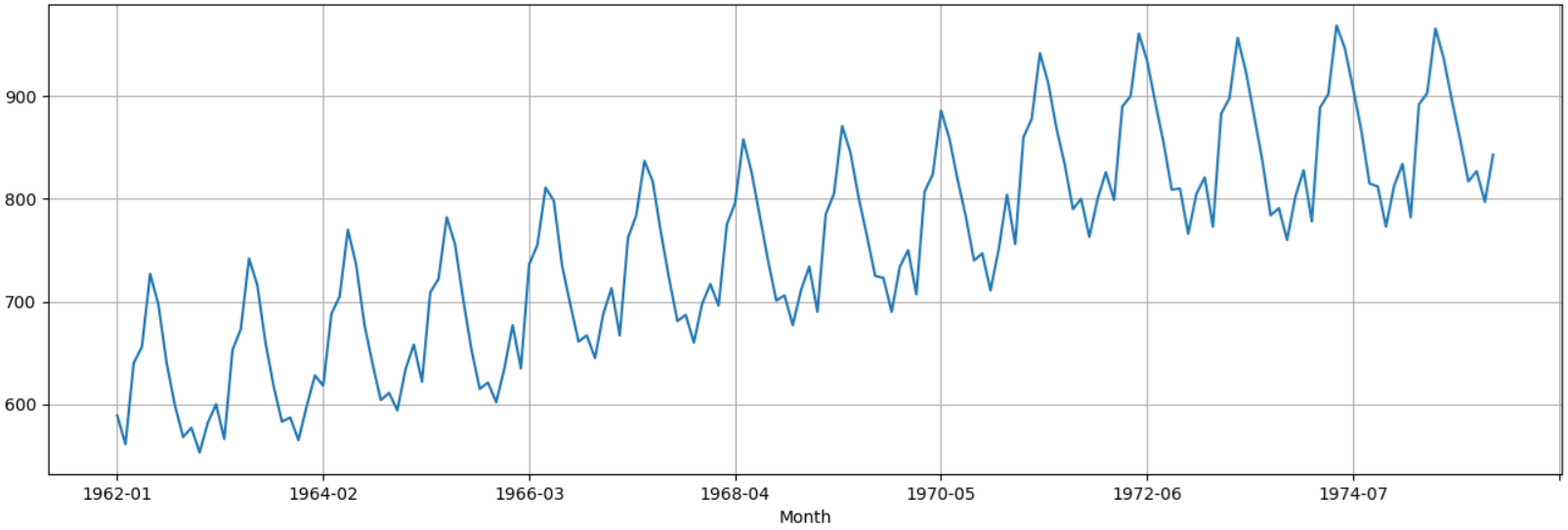
Ajustes do calendário

Algumas das variações observadas nos dados sazonais podem ser devidas a simples efeitos de calendário. Nesses casos, geralmente é muito mais fácil **remover a variação** antes de ajustar um modelo de previsão.

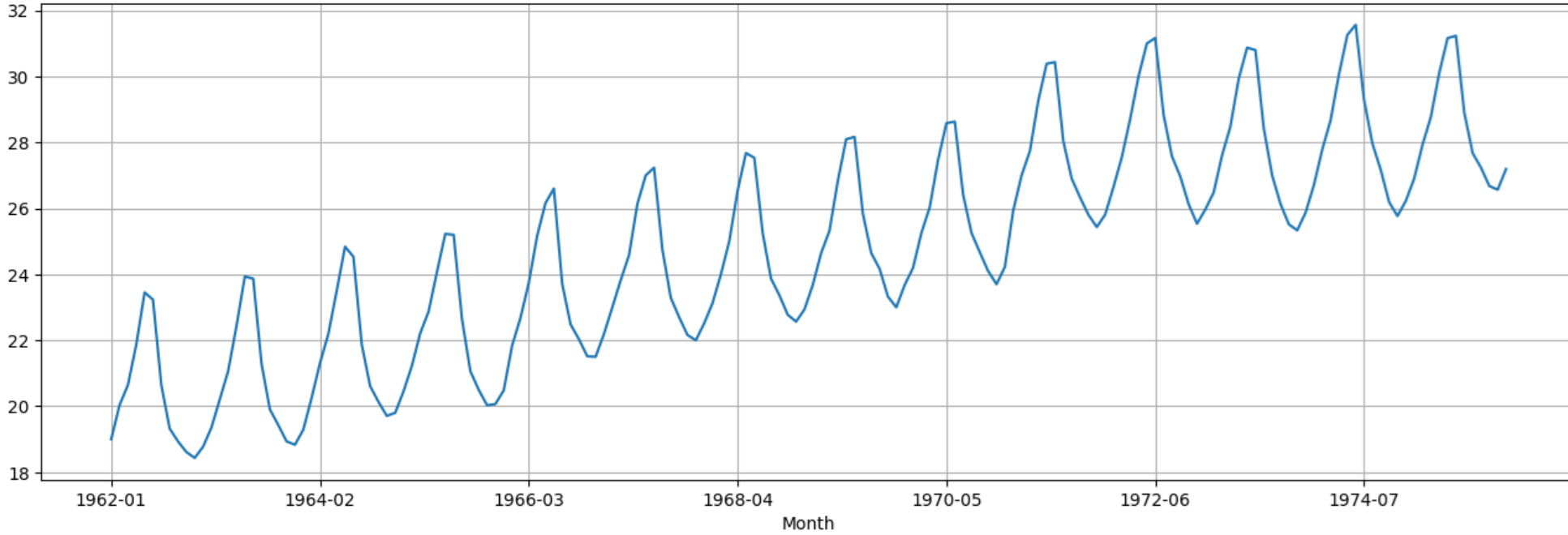
Por exemplo, se você estiver estudando a **produção mensal** de uma fábrica, ou o total mensal de vendas de uma loja, haverá variação entre os meses simplesmente por causa dos **diferentes números de dias em cada mês**, além da **variação sazonal ao longo do ano**.

No caso de uma série que representa a produção mensal de uma fábrica, basta ajustar a produção pela média da produção da fábrica em cada dia do mês, dividindo a produção mensal pelo número de dias do mês. Ao olhar para a produção média diária em vez da produção mensal total, remove-se efetivamente a variação devido aos diferentes números de dias de cada mês.

Ajustes do calendário – produção de leite



Produção e Produção dividida por dias úteis por mês



Ajustes de população

Quaisquer dados que **dependem de mudanças populacionais** podem ser ajustados para fornecer dados per capita → ou seja, considerando os **dados por pessoa** (ou por mil pessoas, ou por milhão de pessoas) em vez do total.

Por exemplo, se os dados forem o número de leitos hospitalares em uma determinada região ao longo do tempo, os resultados serão muito mais fáceis de interpretar se forem removidos os efeitos das mudanças populacionais considerando o número de leitos por mil pessoas → dessa forma é possível ver se houve aumentos reais no número de leitos, ou se os aumentos se devem ao aumento da população. É possível que o número total de leitos aumente, mas o número de leitos por mil pessoas diminua. Isso ocorre quando a população está aumentando mais rapidamente do que o número de leitos hospitalares.

Ajustes de inflação

Recomenda-se que dados relacionados ao valor do dinheiro sejam ajustados antes da modelagem.

Por exemplo, o custo médio de uma casa nova terá aumentado nas últimas décadas devido à inflação. Uma casa de 2 milhões este ano não é o mesmo que uma casa de 2 milhões vinte anos atrás.

As séries temporais financeiras geralmente são ajustadas para que todos os valores sejam expressos em valores do dinheiro em um determinado momento. Por exemplo, os dados do preço da casa podem ser declarados em reais do ano 2020.

Para fazer esses ajustes, é necessário obter e usar algum índice de inflação.

Transformações matemáticas

Temos

- a transformação logarítmica: $w_t = log(y_t)$

usada para séries temporais com variação exponencial;

- Valor de retorno: $r_t = log(p_t/p_{t-1})$

usada para séries, por exemplo, de preço. Valor positivo, aumento, negativo, redução.

- Transformações de potência: $w_t = \frac{y_t^{\lambda}}{\lambda}$

usada para diminuir variações nos dados que variam em função da sua magnitude, de forma a fazer com que as variações se tornem da mesma ordem de grandeza em qualquer instante de tempo da série.

Além de outras como transformada de Fourier, Transformada Cosseno, etc.

Resíduo

Os ***resíduos*** em um modelo de série temporal são as diferenças entre os valores reais e os valores previstos pelo modelo, ou seja:

$$e_t = y_t - \hat{y}_t$$

Na qual e_t é o resíduo no instante de tempo t, y_t e $\hat{y}_t$ são respectivamente o valor real e o valor previsto da série pelo modelo no instante de tempo t.

Os resíduos são úteis para verificar se um modelo capturou adequadamente as informações nos dados.

Um bom método de previsão deve produzir resíduos com as seguintes propriedades:

- Os resíduos não são correlacionados → se houver correlações entre os resíduos, então, há informações deixadas nos resíduos que devem ser usadas no cálculo das previsões.
- Os resíduos têm média zero → se os resíduos tiverem uma média diferente de zero, as previsões são tendenciosas.

Resíduo

Se qualquer uma dessas propriedades não for satisfeita, o método de previsão pode ser modificado para fornecer melhores previsões.

Por exemplo, ajustar um modelo para eliminar viés é fácil: se os resíduos tiverem média e, então, basta adicionar essa média aos valores previstos e o problema de viés é resolvido. Corrigir o problema de correlação é mais difícil.

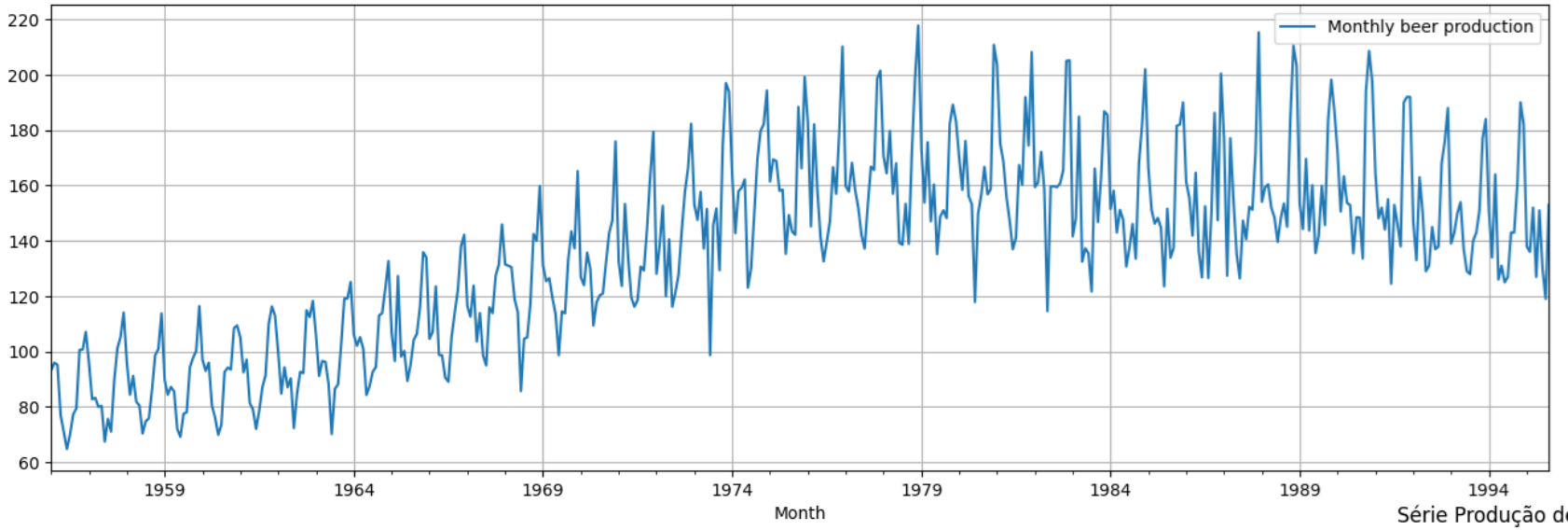
Além dessas propriedades essenciais, é útil (mas não necessário) que os resíduos também tenham as duas propriedades a seguir:

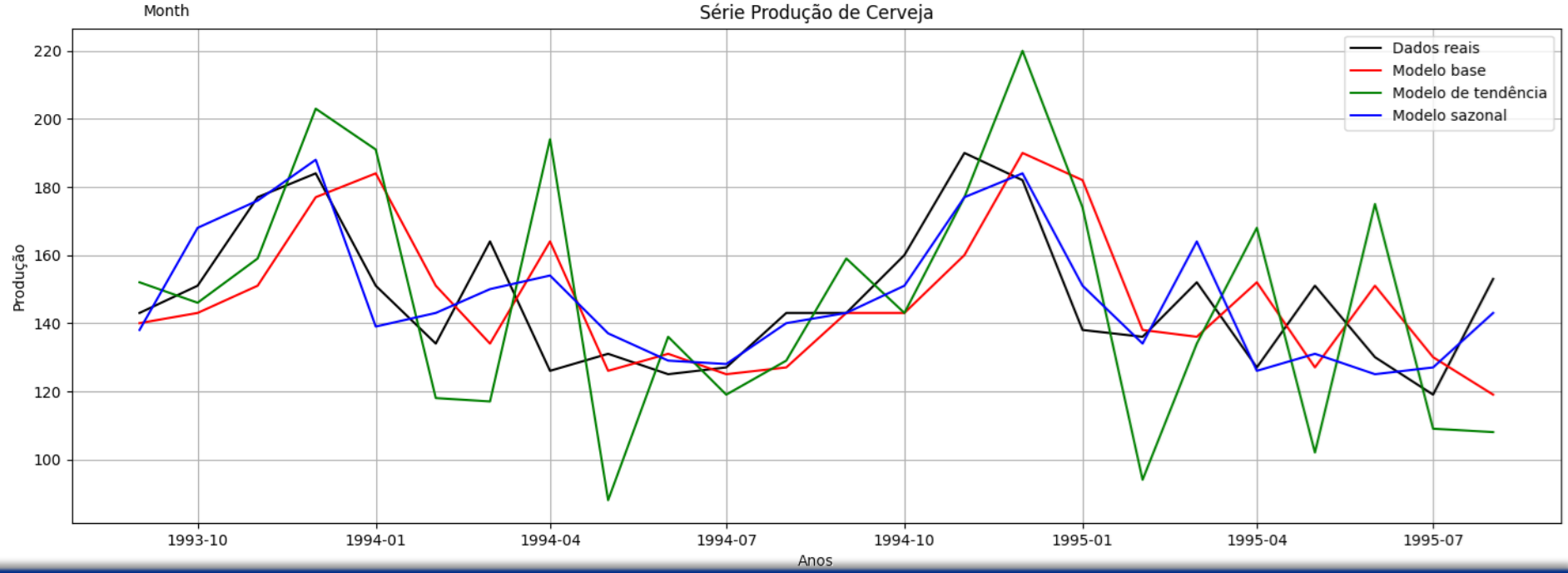
- Os resíduos têm variância constante.
- Os resíduos possuem distribuição normal.

No entanto, um método de previsão que não satisfaz essas propriedades, em alguns casos, não pode ser melhorado. Às vezes, aplicar uma transformação de potência pode ajudar com essas propriedades, mas geralmente há pouco que você pode fazer para garantir que os resíduos tenham variância constante e uma distribuição normal.

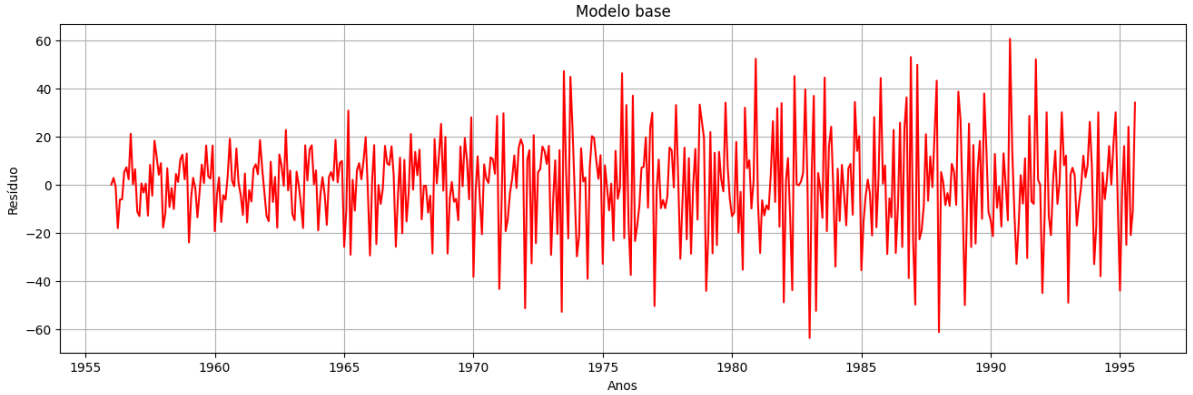
Se as previsões forem boas, os resíduos devem ter uma distribuição de probabilidade gaussiana e a sua auto-correlação deve ser zero para qualquer instante de tempo.

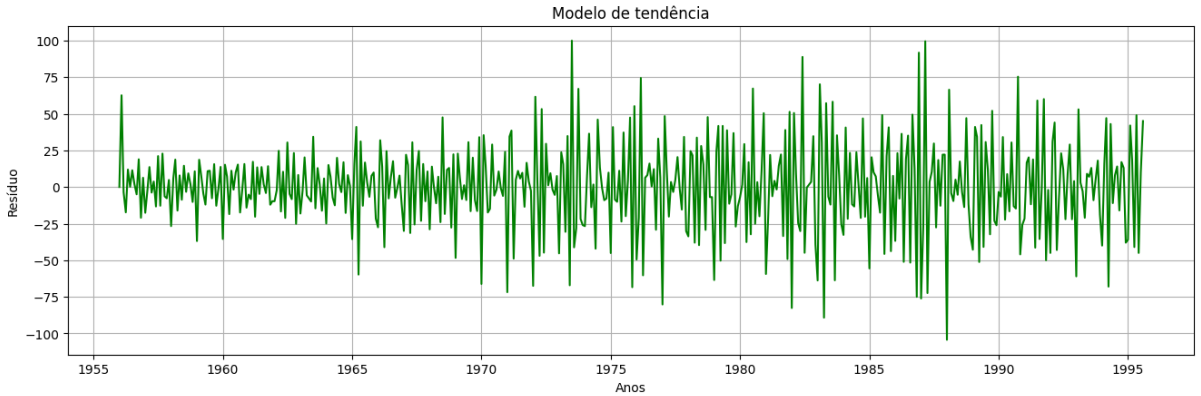
Resíduo – Produção de cerveja na Austrália

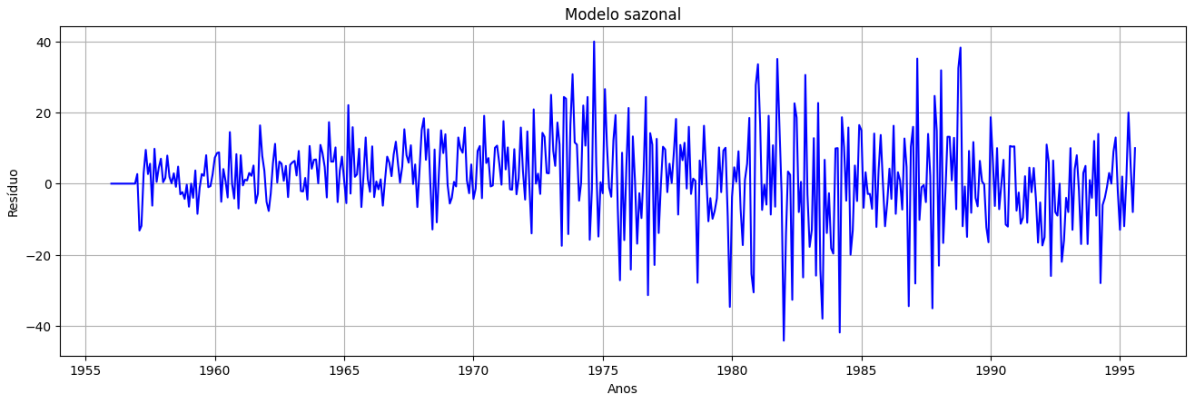


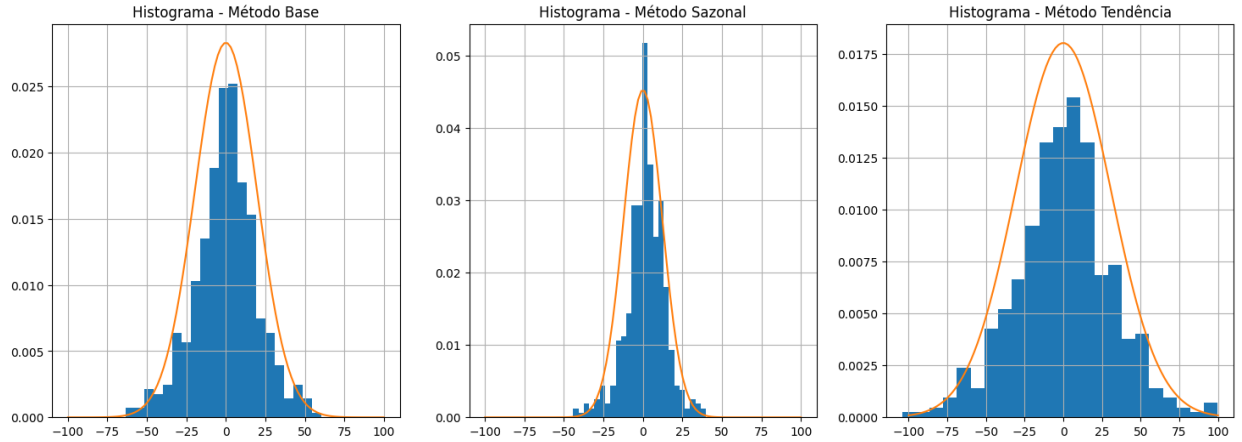


Resíduo







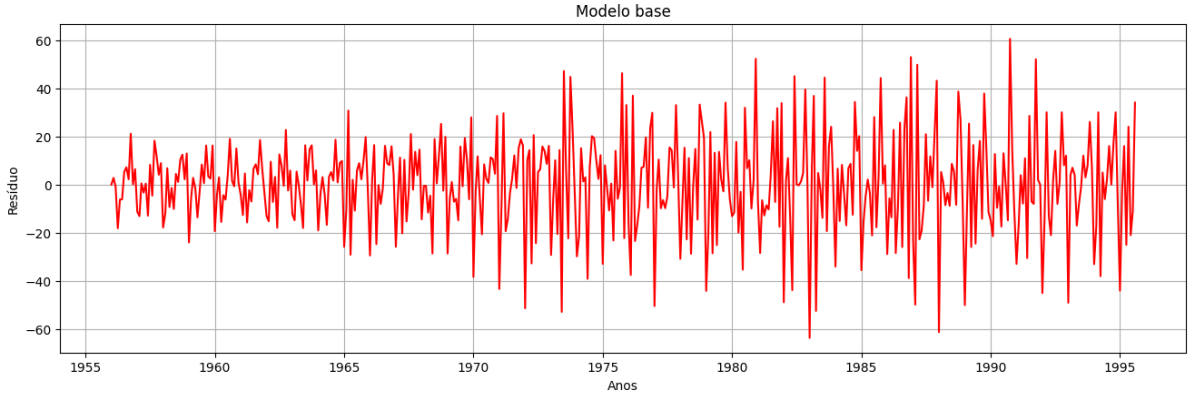


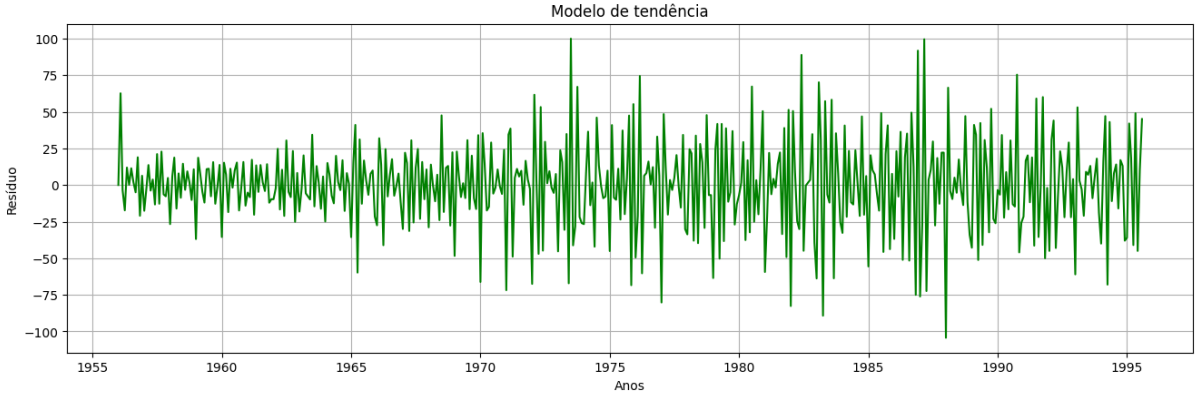

```
Desvio padrão - método base: 19.612241466978855
Desvio padrão - método sazonal: 12.257327552470413
Desvio padrão - método tendencia: 30.824947466418553
```

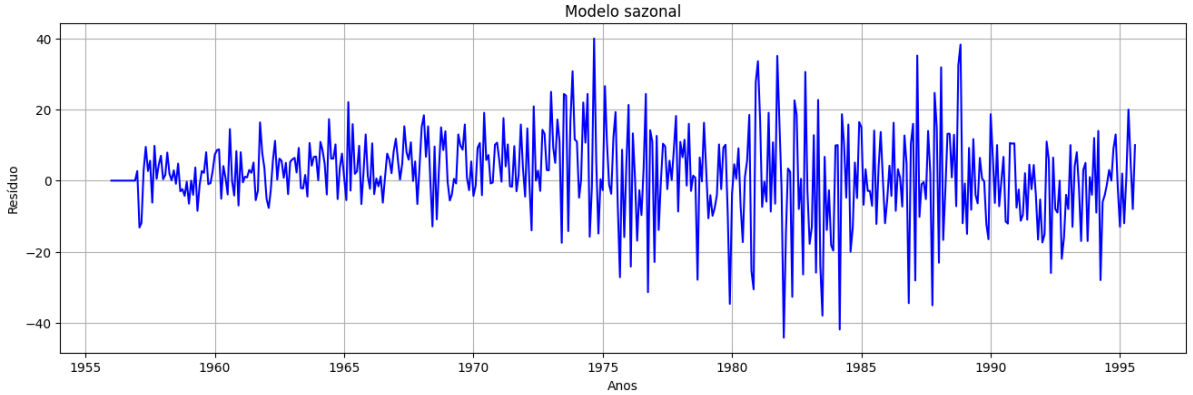
As médias dos resíduos são próximas de zero;

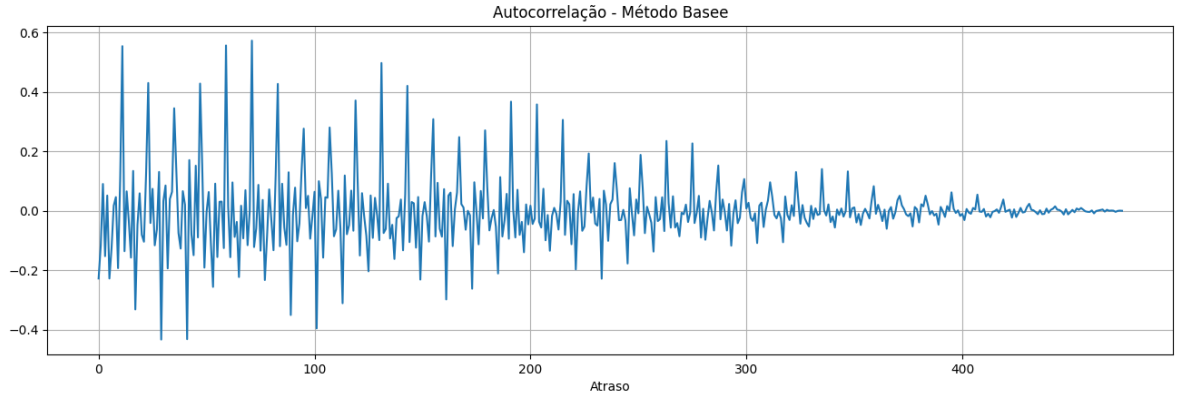
O seus valores máximos são próximos dos valores das respectivas distribuições normais.

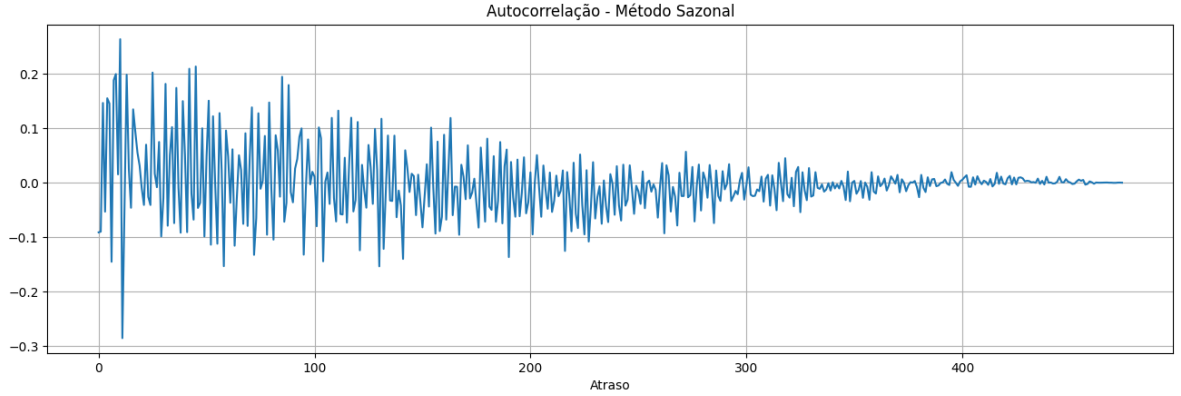
Resíduo

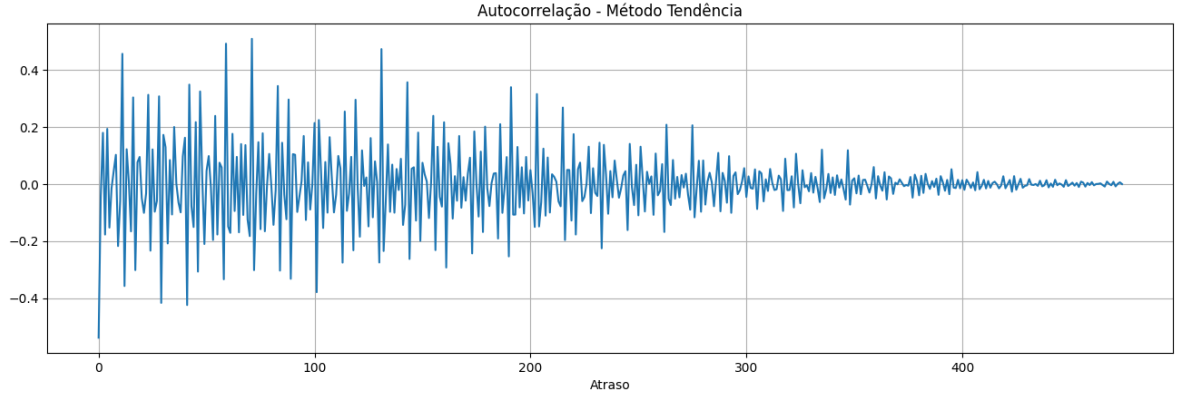

Modelo base
Resíduo
Anos


Modelo de tendência
Resíduo
Anos


Modelo sazonal
Resíduo
Anos


Autocorrelação - Método Basee
Atraso


Autocorrelação - Método Sazonal
Atraso


Autocorrelação - Método Tendência
Atraso

Métricas para avaliar previsões e modelos

R squared ($R2$)

Coeficiente de determinação (em econometria, pode ser interpretado como sendo a porcentagem da variância explicada pelo modelo):

$$R^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

onde yi é um dado da série, $\hat{y}_i$ é o valor previsto para esse dado, $\bar{y}$é o valor médio dos dados reais da série e N é o número de dados (amostras) da série.

Função em Python: `sklearn.metrics.r2_score`

Erro absoluto médio (MAE)

É uma das métricas mais usadas por ser a mais fácil de ser interpretada:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

onde y_i é um dado da série, $\hat{y}_i$ é o valor previsto para esse dado, e N é o número de dados (amostras) da série.

Função em Python: `sklearn.metrics.mean_absolute_error`

Erro absoluto médio porcentual (MAEP)

É o mesmo que o MAE, mas é calculado como sendo uma porcentagem, o que é mais conveniente quando se quer analisar a qualidade de um resultado:

$$MAEP = \frac{100}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i}$$

onde y_i é um dado da série, $\hat{y}_i$ é o valor previsto para esse dado, e N é o número de dados (amostras) da série.

Erro absoluto médio percentual modificado (MAEPM)

Para eliminar o problema de divisão por zero pode-se usar uma modificação do MAEP.

$$MAEPM = \frac{100}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{\bar{y}}$$

onde yi é um dado da série, $\hat{y}_i$ é o valor previsto para esse dado, $\bar{y}$é o valor médio dos dados reais da série e N é o número de dados (amostras) da série.

Modelo de média móvel

O modelo de média móvel é um dos mais usados para séries temporais em razão da sua simplicidade aliado à sua capacidade de fornecer parâmetros importantes para análise de uma série temporal.

O modelo de média móvel assume que o valor futuro da variável depende da média de seus valores anteriores.

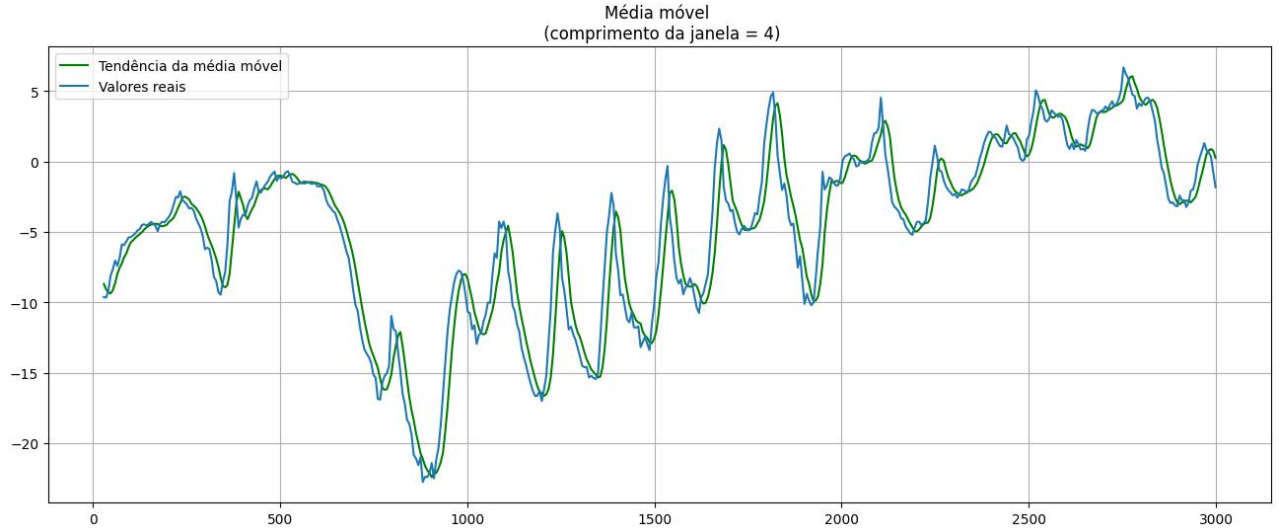
O modelo de média móvel fornece uma previsão para a próxima amostra tendo as amostras anteriores, usando a seguinte equação:

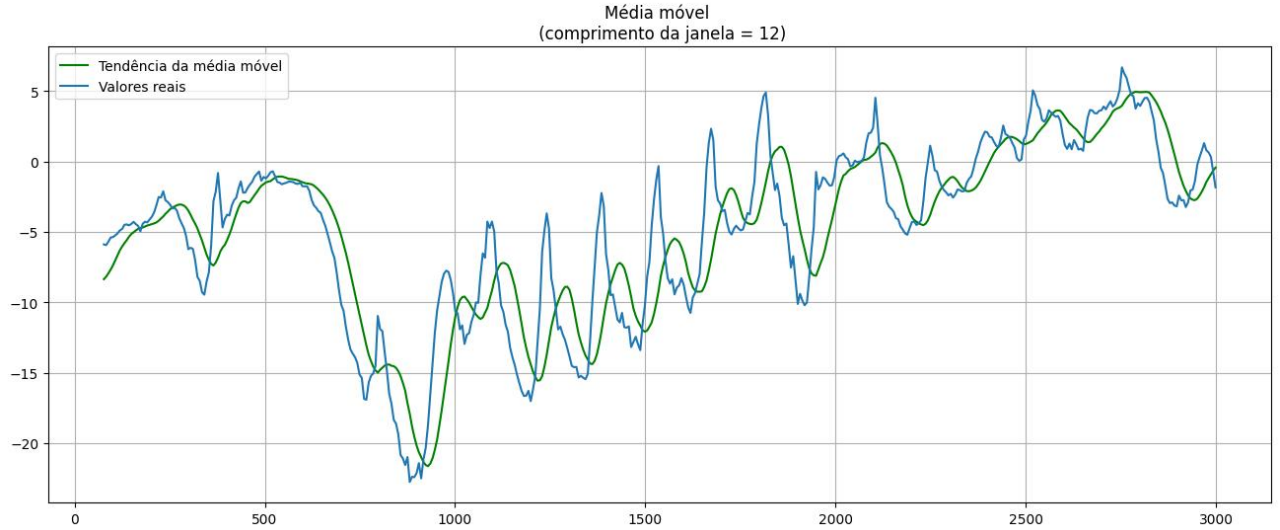
$$\hat{y}_t = \frac{1}{m} \sum_{i=1}^{m} y_{t-i}$$

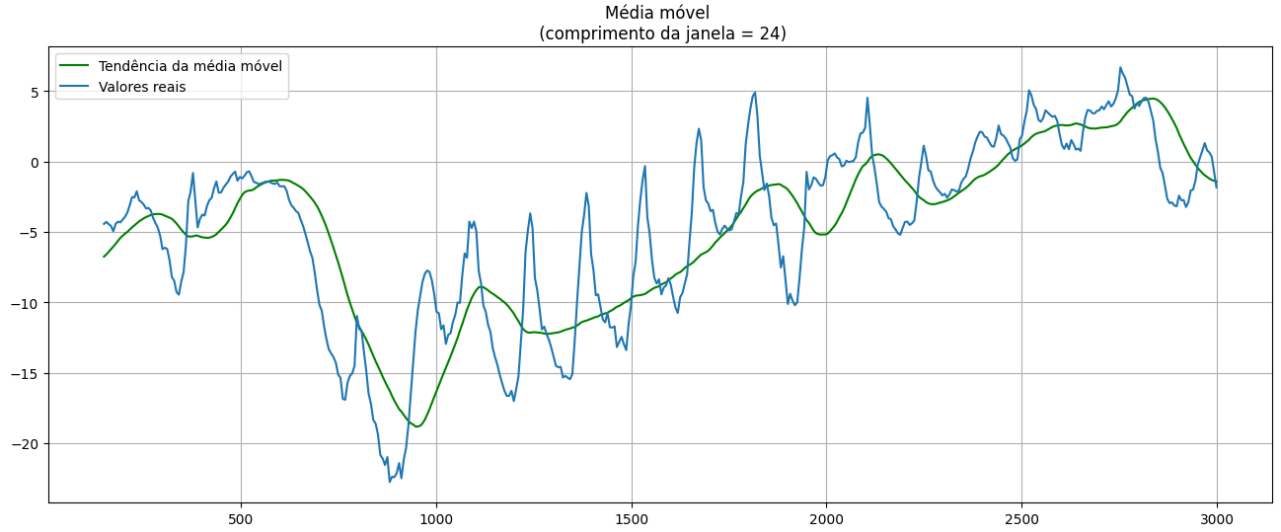
onde m é o número de amostras anteriores usadas para calcular a média. O valor de m é função do comportamento da série.

Observe que esse modelo pode ser visto como sendo uma extensão do método base.

Modelo de média móvel







Erro absoluto médio percentual modificado (MAEPM)

Um intervalo de confiança fornece um intervalo dentro do qual espera-se que a previsão tenha uma probabilidade de acerto determinada. Supondo que os erros de previsão tenham uma distribuição de probabilidade normal, então, um intervalo de confiança de uma dada % é definido por:

$$\hat{y}_i \pm c\sigma$$

O valor dos intervalos de confiança expressam a incerteza nas previsões. Se produzirmos apenas previsões, não há como dizer quão precisas elas são. No entanto, se também produzirmos intervalos de confiança, fica claro quanta incerteza está associada a cada previsão.

Porcentagem	Multiplicador (c)
50	0,67
60	0,84
70	1,04
80	1,28
90	1,64
95	1,96
99	2,54

Detecção de anomalias

Usando esses métodos simples de previsão, incluindo os métodos de média móvel, é possível criar um sistema de detecção de anomalias da forma de variações bruscas nos dados ("outliners").

Para detectar "outliners" basta identificar os dados reais que ficam foram do intervalo de confiança das previsões.

Exemplo de intervalo de confiança e detecção de anomalias