

Imputação de valores ausentes em séries temporais utilizando o pacote imputeTS

Pedro de Brito Neto

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

7 de dezembro de 2023

- 1 Introdução
- 2 Métodos
- 3 Aplicação
- 4 Referências

- Valores ausentes;
- Tipos: Ausente Completamente Aleatoriamente (MCAR – Missing Completely At Random),
Ausente Aleatoriamente (MAR – Missing At Random)
e Ausente Não Aleatoriamente (MNAR – Not Missing At Random);
- Dados faltantes em séries temporais;
- Revisão dos principais métodos de imputação e aplicações no R.

Métodos convencionais

- **Ignorando:** ignorar os valores faltantes ao longo da análise;
- **Exclusão:** excluir a observação ausente;
- **Imputação por média, moda ou mediana:** imputar os valores faltantes com a média, moda ou mediana.

Procedimentos de imputação

- **Imputação de Hot Deck:** Substitui valores ausentes por dados observados de unidades similares no mesmo conjunto de dados;
- **Imputação de Cold Deck:** Substitui valores ausentes por dados provenientes de fontes externas ou conjuntos de dados diferentes, geralmente assumindo que esses dados são comparáveis;
- **Imputação Múltipla:** cria várias estimativas para valores ausentes em um conjunto de dados, refletindo a incerteza associada à imputação.;
- **LOCF e NOCB:** LOCF preenche valores ausentes usando a última observação conhecida. NOCB utiliza a próxima observação conhecida para preencher valores ausentes, retrocedendo no tempo.

Procedimentos de imputação

- **Interpolação linear:** calcula valores intermediários com base em dois pontos adjacentes conhecidos, usando uma linha reta para estimar o valor ausente;
- **Interpolação spline:** preenche valores ausentes em um conjunto de dados criando uma curva suave (spline) que passa através dos pontos conhecidos. Splines são funções matemáticas definidas por segmentos em que cada segmento é um polinômio de baixo grau, em séries, os polinômios são de terceiro grau;
- **Filtro de Kalman:** É um algoritmo de estimativa recursiva que utiliza informações passadas e presentes para prever estados futuros em sistemas dinâmicos; o algoritmo é aplicado para prever e preencher valores ausentes em uma série temporal.

- O pacote **imputeTS** pode ser encontrado no CRAN do software R;
- Dedicado exclusivamente à imputação univariada de séries temporais e inclui diversos algoritmos;
- Algoritmos como: interpolação, mediana, moda, Kalman, entre outros;
- Disponibiliza conjuntos de dados disponíveis em uma versão com dados faltantes e uma versão completa.

- Trabalhamos com a série “AirPassengers” que foi dividada em duas;
- **tsAirgap**: Série temporal de passageiros aéreos mensais (com NAs). 144 linhas com 14 valores faltantes.
- **tsAirgapComplete**: Série temporal de passageiros aéreos mensais (completo). 144 linhas sem valores faltantes.

Imputação por média

A imputação por média é feita por meio da função **na.mean**.

Imputação de valores por média

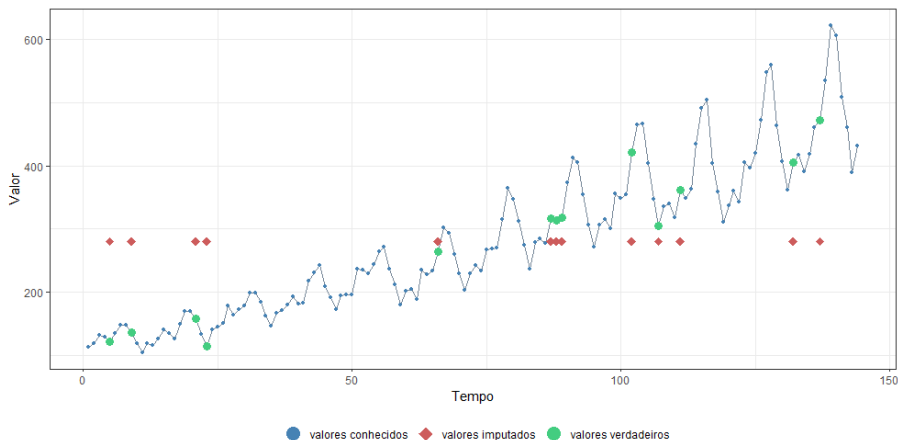


Figura: Imputação de valores faltantes por média.

Imputação por Interpolação

A imputação por interpolação é feita por meio da função **na.interpolaion**.

Imputação de valores por interpolação

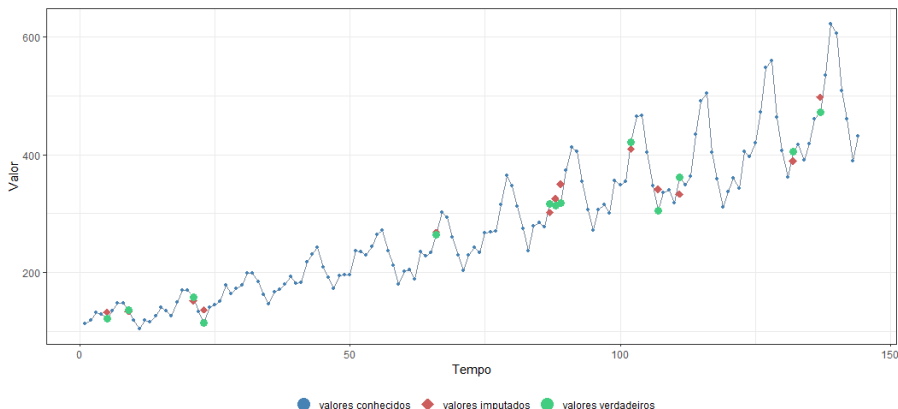


Figura: Imputação de valores faltantes por interpolação.

Imputação Por Kalman

A imputação por Kalman é feita por meio da função **na.kalman**

Imputação de valores por kalman

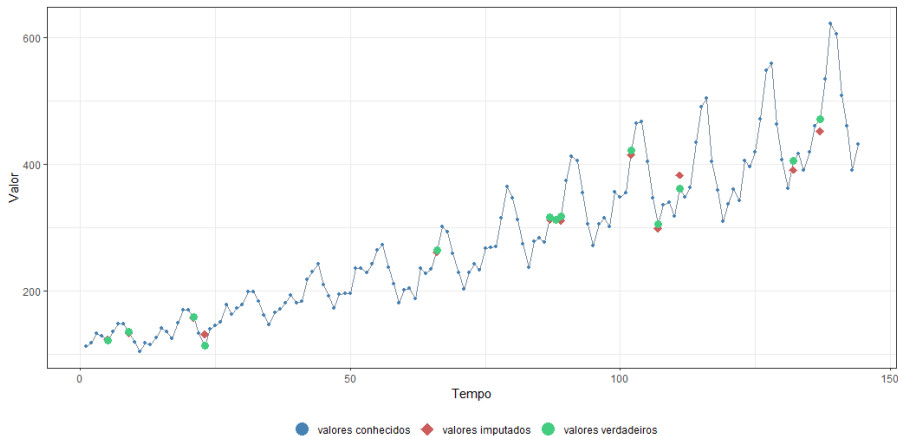


Figura: Imputação de valores faltantes por Kalman.

- Conhecer bem os dados, o problema e os objetivos da análise;
- Conhecer ferramentas para a aplicação dos métodos.

- Andridge, R. R. Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.
- Aydılek, I. B. Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233:25–35.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley Sons.
- Hopke, P. K., Liu, C., Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic. *Biometrics*, 57(1):22–33.
- Little, R. J. Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley Sons.
- Moritz, S. Bartz-Beielstein, T. (2017). imputets: time series missing value imputation in r. *R J.*, 9(1):207.
- Shao, C., Fang, F., Bai, F., Wang, B. (2014). An interpolation method combining snurbs with window interpolation adjustment. In

Obrigado!
Pedro de Brito Neto



UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO