Imputação de valores ausentes em séries temporais utilizando o pacote imputeTS

Pedro de Brito Neto

7 de dezembro de 2023

1 Resumo

A falta de dados é um problema comum para os tipos de dados, não diferente para dados de séries temporais. Métodos convencionais como imputação e média e moda não costumam trazer bons resultados. Métodos mais robustos normalmente são mais adequados para tratar dados faltantes no contexto de séries temporais. Além de ser necessário conhecer os dados e os métodos mais adequados para cada cenário, é importante também conhecer ferramentas para a implementação desses métodos. No software R diversos pacotre

2 Introdução

Um conjunto de dados é uma coleção organizada de informações ou observações, geralmente apresentações de forma estruturada, que podem ser evidenciadas para obter compreensões ou entendimentos, fazer inferências ou tomar decisões. Conjuntos de dados podem ser obtidos por diferentes métodos, como, por exemplo, realizar pesquisas, entrevistas ou experimentos para coletar dados específicos, usar sensores ou dispositivos de medição para coletar informações em tempo real, dados disponíveis em portais governamentais, organizações públicas ou outras fontes de dados abertas, etc.

Analisar um conjunto de dados, seja por qual for o objetivo ou método de análise utilizado, conta com uma etapa importante: o pré-processamento dos dados. Está etapa consiste em manipular, limpar e organizar os dados brutos encontrados para torná-los mais adequados e eficientes para análise. O tratamento de valores ausentes está incluso na estapa de pré-processamento.

Valores ausentes referem-se à ausência de dados em determinadas posições ou variáveis dentro de um conjunto de dados. Em outras palavras, são espaços em branco ou informações não registradas em locais onde deveria haver dados.

O motivo da ausência de dados são vários. Em uma pesquisa de opinião por exmeplo, os dados ausêntes podem ocorrer devido a erros humanos durante a coleta dos dados, onde os respondentes se recusam a responder algumas perguntas ou não completam toda a pesquisa.

Dados ausentes por falhas técnicas também podem ocorrer. Suponhamos que um sensor de temperatura foi implantado em vários locais para monitorar as variações diárias. Durante um período específico, uma falha técnica ocorreu em um dos sensores, resultando na ausência de dados de temperatura para esse local.

Valores ausentes são imprevisíveis e provavelmente inevitáveis. No contexto de séries temporais, dados de séries temporais são também passíveis de valores faltantes. Uma série temporal, em resumo, é um conjunto de dados que são coletados ou registrados em intervalos regulares ao longo do tempo. Esses dados são organizados em uma sequência cronológica, onde cada ponto de dados está associado a um momento específico. Alguns exemplos de séries temporais são: registros de temperatura, preços diários de ações, vendas diárias ou mensais de produtos. Por conta dessa sequência cronológica, alguns cuidados devem ser tomados para a imputação de dados em séries temporais, como: séries temporais frequentemente exibem padrões sazonais e tendências ao longo do tempo. Os métodos de imputação precisam ser capazes de capturar e preservar esses padrões temporais.

A presença de dados faltantes pode seguir padrões diferentes, e esses padrões são frequentemente classificados em três categorias principais:

- **Ausente Completamente Aleatoriamente (MCAR – Missing Completely At Random):** significa que a ocorrência dos dados faltantes é completamente plausível, não relacionada aos valores observados ou não observados. Em outras palavras, a probabilidade de um dado estar faltando é a mesma para todos os pontos da série temporal, independentemente dos valores observados. Esse é um cenário mais simples e, muitas vezes, mais fácil de lidar estatisticamente;

- **Desaparecido Aleatoriamente (MAR - Missing At Random):** indica que a probabilidade de um dado estar faltando pode depender de outros valores observados, mas não depende do próprio valor que está faltando. Embora haja uma dependência, essa dependência é conhecida e pode ser modelada, facilitando a imputação ou preenchimento dos dados faltantes;

- **Ausente Não Aleatoriamente (MNAR - Not Missing At Random):** Refere-se a uma situação em que a probabilidade de um dado estar faltando depende do valor que está faltando. Isso torna a imputação mais solicitada, pois a dependência está relacionada à informação não observada. Lidar com dados não faltantes pode envolver estratégias mais complexas.

Existem diferentes métodos para introdudiz esses dados, onde, a depender de vários fatores como, por exemplo o problema, o tipo de dados faltante, o tamanhho da base, a quantidade de valores faltantes e até mesmo aa análise que esta sendo realizada, métodos convencionais podem ser úteis ou apenas os mais robustos.

Métodos convencionais são os primeiros métodos que surgiram no estudo de tratamento de valores ausentes. São métodos simples, mas que podem ser problemáticos, visco que resolvem um problema, mas podem criar outro. Preencher os valores faltantes com a média, mediana ou moda dos valores coletados são exemplos de métodos simples.

Tendo em vista que o tratamento dos valores faltantes será fetio na etapa de pré-processamento, é importante indentificar os possíveis problemas e encontrar o melhor método para imputar esses dados. Além disso, conhecer ferramentas que podem ajudar nesse processo torna-se igualmente útil.

Neste trabalho iremos fornecer uma revisão de alguns estudos estudos relacionados aos principais métodos ou técnicas de tratamentos de valores ausentes. Além disso será apresentado o pacote **imputeTS**, especializado em imputação de séries temporais univariadas.

3 Métodos

Existem vários métodos, técnicas e abordagens propostas para o tratamento de valores faltantes. Métodos básicos como, imputação por média, moda, zero, ignorar ou ecxluir são propostos em (Little & Rubin, 2019). Esses métodos são extremamente simples, mas eficazes apenas em cenários específicos, como baixo percentual de falta. Já em casos onde se tem um alto percentual de falta, o resultado da análise será afetado ou até mesmo podera gerar resultados tendenciosos (Aydilek & Arslan, 2013). A qualidade da análise é influenciada pela qualidade dos dados, portanto, em determinados casos, é necessário possui os dados completos, tornando essencial a estimação de valores ausentes.

Nesta seção será apresentado uma breve explicação sobre os principais métodos e abordagens para a estimação de valores faltantes.

3.0.1 Métodos Convencionais

Esses métodos são realmente simples, alguns utilizam princípios estatísticos como base.

- **Ignorando:** O primeiro método apresentando em (Aydilek & Arslan, 2013) é ignorar o valor faltante, é descrita como a maneira mais simples de lidar com dados faltantes. Consiste em ignorar completamente os valores faltantes à medida que a análise prossegue. Entretanto, pode ser um método muito arriscado, principalmente se a porcentagem de valores faltantes for muito alta;

- **Exclusão:** Este método consiste em simplismente excluir a observação ausente para continuar a análise.

- **Imputação por média ou moda:** Este métdo consiste em imputar os valores faltantes com a média ou a moda, pode ser mais eficiente que os métodos anteriores, tendo em vista que resolve o

problema dos valores faltantes e o número de dados permanece o mesmo. Uma desvatagem é o viés causado por tantos valores de dados com valores semelhantes, além de outros motivos.

3.0.2 Procedimentos de Imputação

Estes métodos tratam do problema de valores ausentes substituindo cada um dos valores por alguns valores específico, eles valores variam de acordo com o método.

- **Imputação de Hot e Cold Deck:** A abordagem de Hot Deck envolve a substituição de valores ausentes por observações existentes no mesmo conjunto de dados. Essencialmente, um "deck"de valores disponíveis é reservado, e os valores ausentes são preenchidos com valores retirados aleatoriamente desse deck. Esta abordagem é particularmente útil quando há uma estrutura temporal nos dados ou quando existe uma relação lógica entre os valores ausentes e outros observados na amostra (Andridge & Little, 2010). Ou seja, pode funcionar bem no contexto de séries temporais. Diferente de Hot Deck, a imputação de Cold Deck consiste na substituição de valores ausentes por valores provenientes de uma fonte externa ou de um conjunto de dados de referência ou valores predefinidos, muitas vezes determinados antes da análise dos dados. Os valores a serem imputados são retirados de um conjunto fixo de dados, independentemente da estrutura ou padrões nos dados originais. Essa abordagem é útil quando se busca consistência nos resultados ou quando há justificativa teórica ou prática para substituir os valores ausentes por um conjunto específico de números.

- **Imputação Múltipla:** A Imputação Múltipla é uma técnica estatística utilizada para enfrentar o desafio dos dados ausentes em conjuntos de dados. Quando há valores faltantes, a Imputação Múltipla não se contenta com uma única estimativa, mas cria múltiplos conjuntos de dados completos, cada um com diferentes imputações para os valores ausentes. Essa abordagem é valiosa porque reconhece a incerteza associada à imputação, gerando estimativas mais robustas e confiáveis. O processo inicia com a identificação das variáveis que contêm valores ausentes. Em seguida, modelos estatísticos são desenvolvidos para compreender a relação entre as variáveis com dados ausentes e aquelas que são observadas. A partir desses modelos, diversos conjuntos de dados completos são gerados, representando diferentes possíveis realizações dos dados ausentes. Cada conjunto de dados imputado é então submetido à análise estatística de interesse. Esta etapa é crucial, pois permite explorar a variabilidade introduzida pelas diferentes imputações. Os resultados dessas análises são, posteriormente, combinados utilizando métodos estatísticos de regra de combinação. As vantagens dessa abordagem incluem a consideração da incerteza associada à imputação, a preservação da variabilidade subjacente nos dados e a capacidade de mitigar o viés introduzido pela exclusão de casos com dados ausentes. No entanto, há desvantagens a serem consideradas. A complexidade computacional pode ser um obstáculo, pois a geração de múltiplos conjuntos de dados e a combinação de resultados podem exigir recursos computacionais significativos. Em (Hopke et al., 2001) é feito uma aplicação e discussão deste método. A imputação múltipla em séries temporais é uma abordagem poderosa, mas também é computacionalmente intensiva e requer um entendimento sólido da estrutura temporal dos dados.

- **Interpolação linear:** Se os intervalos da série temporal forem regulares, mas alguns valores simplesmente não estiverem presentes, os valores faltantes poderão ser estimados usando interpolação linear (Shao et al., 2014). O funcionamento da interpolação linear é fundamentado na ideia de que a mudança entre dois pontos consecutivos é uniforme. A técnica conecta esses pontos com uma reta e utiliza essa linha para estimar valores desconhecidos dentro desse intervalo. O cálculo do valor interpolado é feito utilizando a equação da reta que liga os pontos conhecidos. Esta abordagem é direta e eficiente, sendo particularmente útil em conjuntos de dados menores. Suas vantagens incluem a simplicidade e a rapidez de implementação, tornando-a uma escolha prática para situações em que se deseja uma estimativa rápida. Entretanto, a interpolação linear apresenta limitações importantes. Ela assume uma relação linear entre pontos adjacentes, o que pode não representar adequadamente padrões mais complexos nos dados. Além disso, é sensível a outliers, pois uma única observação atípica pode influenciar significativamente a reta de interpolação. Em resumo, A imputação por interpolação linear pode ser útil em casos simples e quando a linearidade entre os pontos é uma suposição razoável. No entanto, em situações mais complexas ou quando a série temporal possui padrões não lineares.

- **Interpolação spline:** Neste método, os valores ausentes são substituídos com base em uma interpolação spline dos valores disponíveis. A interpolação spline emprega polinômios por partes para aproximar os dados, capturando padrões não lineares (Shao et al., 2014). Este método é adequado para dados de séries temporais, mas pressupõe uma certa suavidade nos dados. O funcionamento da interpolação spline envolve a criação de polinômios separados para cada segmento entre dois pontos adjacentes do conjunto de dados. Esses polinômios são escolhidos de modo que, além de passar pelos pontos conhecidos, sejam suaves, evitando oscilações bruscas e proporcionando uma transição contínua entre os segmentos. Essa técnica é particularmente útil quando a relação entre os pontos no conjunto de dados não é linear. A interpolação spline oferece maior flexibilidade para se adaptar a padrões mais complexos. Existem diferentes tipos de splines, sendo o spline cúbico natural um dos mais comuns, onde a segunda derivada nos extremos é fixada como zero para garantir suavidade nas extremidades. s vantagens da interpolação spline incluem a capacidade de lidar com padrões mais complexos nos dados, oferecendo uma representação suave e flexível entre os pontos conhecidos. Além disso, a interpolação spline é menos sensível a outliers em comparação com métodos mais simples. No entanto, a interpolação spline também apresenta desvantagens. A principal delas é a possibilidade de introduzir oscilações indesejadas, especialmente em conjuntos de dados pequenos. Além disso, a complexidade computacional pode ser maior em comparação com métodos mais simples, tornando-a menos prática em situações onde a eficiência computacional é crucial.

- **Última observação realizada (LOCF - Last Observation Carried Forward) e próxima observação realizada para trás (NOCB - Next Observation Carried Backward):**Esses métodos substituem os valores ausentes pelo valor observado imediatamente anterior (LOCF) ou pelo valor observado subsequente (NOCB). Eles são potencialmente úteis para dados de séries temporais, mas podem introduzir viés se os dados não forem estacionários. Esses métodos devem ser usados com muito cuidado no contexto de séries temporais, pois assumem que dados não mudam significativamente entre observações consecutivas.

3.0.3 Filtro de Kalman:

A imputação em séries temporais usando o Filtro de Kalman é um método avançado que busca preencher valores ausentes em dados temporais de maneira mais robusta, levando em consideração a dinâmica temporal do processo. Este método é particularmente útil quando há incertezas e variações nos dados de séries temporais.

O Filtro de Kalman é um algoritmo recursivo que opera em duas fases principais: predição e correção. Na fase de predição, o filtro utiliza um modelo dinâmico do sistema para estimar o próximo estado com base nas informações disponíveis até o momento. Esta estimativa leva em conta a dinâmica temporal do processo, bem como qualquer ruído ou incerteza associado. Em seguida, na fase de correção, o filtro ajusta a estimativa com base na nova observação disponível, ponderando a informação observada com a previsão feita na etapa de predição. A incerteza associada tanto à previsão quanto à observação é cuidadosamente considerada durante esse processo.

Ao lidar com imputação em séries temporais, o Filtro de Kalman utiliza os pontos observados adjacentes para prever os valores ausentes e atualiza essa previsão conforme novas observações são disponibilizadas. A capacidade do filtro de adaptar-se dinamicamente às mudanças no sistema e incorporar a incerteza faz com que seja particularmente eficaz em situações onde a relação entre as observações pode ser modelada de forma dinâmica.

A imputação por Filtro de Kalman em séries temporais é uma abordagem avançada, útil em cenários onde métodos mais simples, como a interpolação linear, podem não ser adequados. No entanto, sua implementação requer um entendimento sólido da dinâmica temporal do processo e pode ser computacionalmente intensiva. Em casos de séries temporais mais simples, pode ser mais apropriado considerar métodos mais diretos, enquanto o Filtro de Kalman brilha em cenários mais complexos com dinâmicas não triviais e incertezas. Em (Harvey, 1990), o método é discutido, assim como sua aplicação em séries temporais. Este método de imputação é muito eficiente.

3.1 Impute TS

O pacote **imputeTS**(Moritz & Bartz-Beielstein, 2017) pode ser encontrado no CRAN do software R. É um pacote dedicado exclusivamente à imputação univariada de séries temporais e inclui diversos algoritmos. O pacote permite a imputação de valores faltantes por diversos métodos como interpolação, mediana, moda, Kalman, entre outros. O desempenho da imputação é sempre muito dependente das características da série temporal. Em alguns casos, imputação com valores médios pode funcinar bem. Em séries temporais com forte sazonalidade, o método Kalman costuma se sair melhor. O pacote **imputeTS** disponibiliza conjuntos de dados disponíveis em uma versão com dados faltantes e uma versão completa. isso permite comparar o desempenho dos métodos de imputação. Utilizaremos o pacote **imputeTS** para dar mostrar exemplos de como realizar imputação de dados faltantes no R. Além disso, será possível também comparar o desempenho de alguns métodos.

4 Análises

Para as análises, iremos trabalhar com a série temporal "AirPassengers". A série tem origem em (Box et al., 2015), sendo comumente usada como exemplo na literatura de análise de sérires temporais. No pacote **imputeTS**, a série foi dividida em duas, sendo elas:

- **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.

Iremos aplicar três diferentes formas de imputação na base de dados com NA e comparar com a base original.

4.0.1 Imputação Por Média

A imputação por média é feita por meio da função **na.mean**. Essa função possui outros argumentos, tornando possível a imputação pela moda e pela mediana.

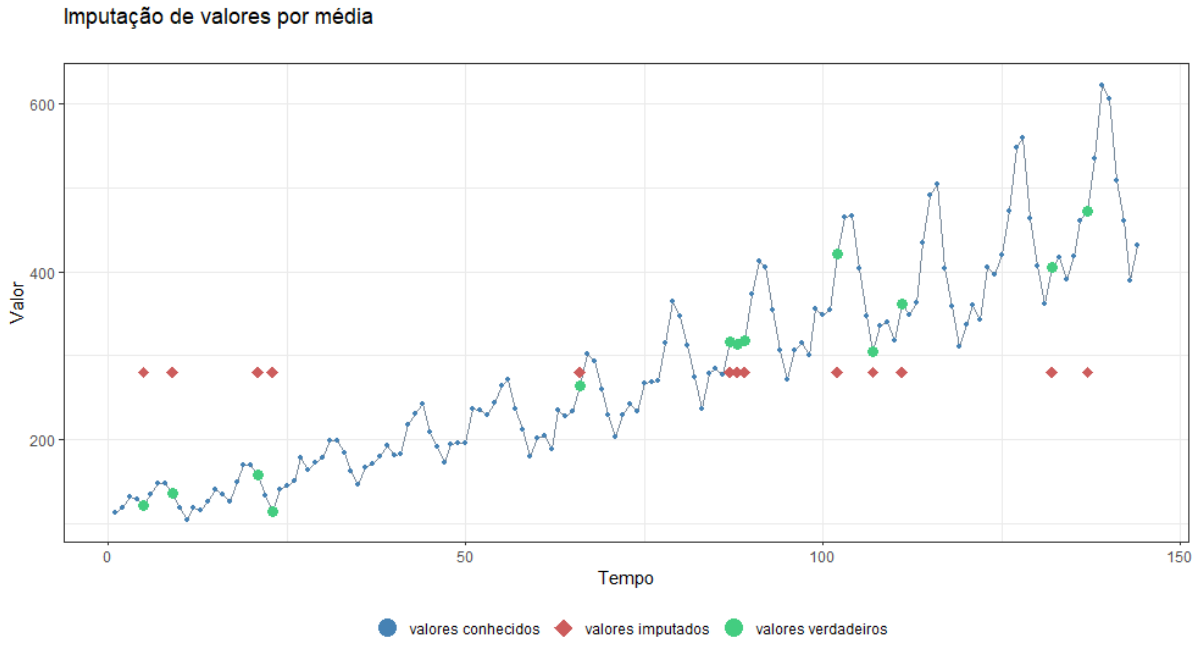

Figura 1: Imputação de valores faltantes por média.

Na Figura 1 tempos o gráfico da imputação por média. De vermelho temos os valores imputados, de ver os valores verdadeiros e de azul os valores conhecidos. Podemos observar que de fato a imputação pela média é um método bem simples, não acompanha nem um pouco a tendência da série. Sendo necessário escolher esse método com muita cautela.

4.0.2 Imputação por Interpolação

A imputação por média é feita por meio da função **na.iterpolaion**. Essa função possui outros argumentos, tornando possível a imputação pela utilizando spline, stine ou linear. Neste caso, estamos usando interpolação linear

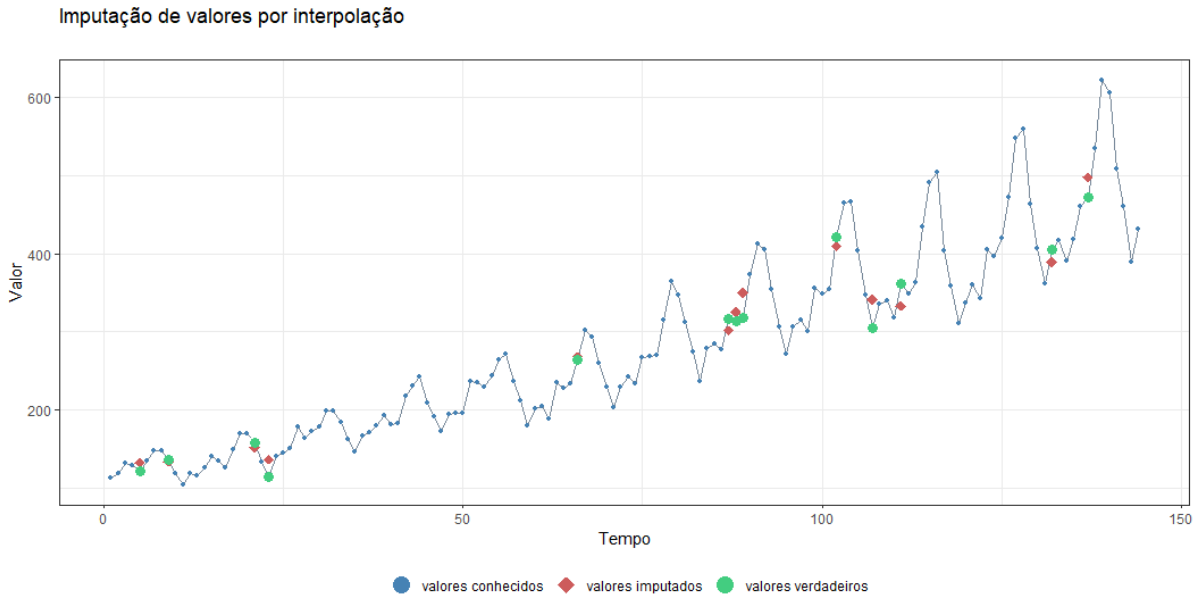

Figura 2: Imputação de valores faltantes por interpolação.

a Figura 2 tempos o gráfico da imputação por interpolação. De vermelho temos os valores imputados, de ver os valores verdadeiros e de azul os valores conhecidos. Podemos observar que para está série o método parece funcionar bem, pegando a tendência da séria em grande parte das vezes.

4.0.3 Imputação Por Kalman

A imputação por média é feita por meio da função **na.kalman**. Para séries muito grandes e com muitos valores faltantes, esse método pode ser computacionalmente demorado.

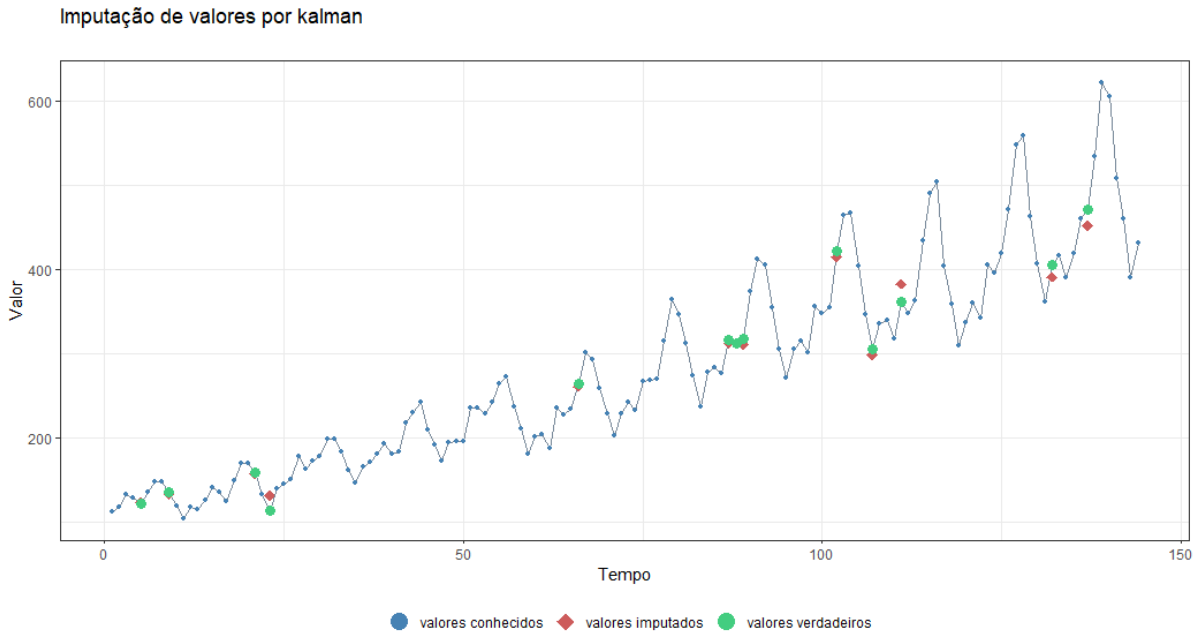

Figura 3: Imputação de valores faltantes por Kalman.

a Figura 3 tempos o gráfico da imputação por interpolação. De vermelho temos os valores imputados, de ver os valores verdadeiros e de azul os valores conhecidos. Podemos observar que os valores imputados estão bem próximos dos valores verdadeiros na maioria das vezes. Até mesmo nos pontos em que estão um pouco mais afastado, estão sempre pegando a tendência verdadeira da série.

5 Resultados

A partir dos gráficos é possível observar uma diferença na imputação dos valores faltantes por cada método. Fica claro que a imputação por média é o menos eficiente destes apresentados, o que já era de se esperar. O método de interpolação e o de Kalman apresentaram bons resultados. Apesar de não ter sido apresentado nenhuma métrica em relação ao desempenho dos métodos, visualmente, o método de Kalman consegue acompanhar melhor o desempenho da série.

6 Conclusões

Existem diversas maneiras de tratar valores ausentes. Escolher qual a melhor maneira depende de diversos fatores como, tamanho da base, quantidade de valores ausentes, tipo de dado, entre outros. Conhecer bem o problema e os dados são fatores determinantes para a escolha do método. Neste trabalho apresentamos uma visão geral de alguns dos métodos de imputação de dados mais conhecidos, além de alguns das vantagens e desvantagens de cada método, podendo ser levadas em consideração em pesquisas futuras.

Além do mais, ter conhecimento em diferentes ferramentas para a aplicação dos métodos também é importante, pois pode influenciar muito no tempo das análises. Apresentamos neste trabalho o pacote **imputeTS** que fornece uma coleção de algoritmos e ferramentes adptadas para essa tarefa. Ilustramos a facilidade de uso das funções, além comparar, de forma visual, o desempenho que alguns métdodos de imputação de dados.

Referências Bibliográficas

Andridge, R. R. & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Aydilek, 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.

Harvey, A. C. (1990). Forecasting, structural time series models and the kalman filter.

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 *2014 4th IEEE International Conference on Information Science and Technology*, pages 176–179. IEEE.