Relatório de progresso

Iniciação Científica

Rafael Jordane de Souza Oliveira

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Breve introdução

Ao iniciarmos as produções da pesquisa, foi feita uma introdução ao estudo de séries temporais e aos modelos de ajuste ARIMA. A intenção do trabalho realizado é pesquisar, testar e comparar diferentes metodologias para a imputação de dados faltantes em séries temporais. A ideia é que, ao encontrarmos o método mais eficiênte, esse método possa ser aplicado em um modelo de séries temporais que relaciona dados de poluição atmosférica no município de VItória/ES e mortes por doenças respiratórias agudas, também no município. Com melhor ajuste e refinamento da série, é possível que a relação entre a poluição do ar e as mortes fiquem ainda mais evidentes, possibilitando que órgãos publicos e a população como um todo tenho mais conhecimento do impacto dessa poluição, e possam posteriormente desenvolver medidas mais diretas para a mitigação de tais danos.

Os Ajustes

As séries temporais foram geradas aleatoriamente à partir do modelo de ajuste ARIMA (*autoregressive integrated moving average*ados), seguindo os seguintes ajustes: AR(1) ($\phi = 0,4$), AR(1)($\phi = 0,6$), AR(2) ($\phi_1 = 0.2, \phi_2 = 0,4$), AR(2) ($\phi_1 = 0.4, \phi_2 = 0,5$) e ARMA(1,1) ($\phi = 0.2, \theta = 0,4$).

Depois de geradas, as mesmas series foram replicadas para os diferentes métodos de imputação dos dados faltantes.

Imputações com base em observações próximas

(Explicação)
Para os modelos LOCF (Last Observation Carried Forward) e NOCB (Next Observation Carried Backward), após a imputação inicial, foi realizada uma imputação adicional utilizando a média da série para preencher os dados faltantes remanescentes. Essa etapa foi necessária devido à possibilidade de ausência de dados nas primeiras ou últimas observações da série, que não são contempladas pelos métodos LOCF e NOCB.

Interpolação

Explicação breve (vou precisar de ajuda)

Resultados

A forma para a medição do impacto do método de imputação dos dados faltantes **(citar método utilizado)** em cada ajuste foi a geração de séries duplicadas. Após remoção aleatória de dados em diferentes porcentagens, sendo elas 5%, 10%, 20% e 40% essas lacunas foram preenchidas com algum método de imputação de dados, e depois disso, foi calculado a raiz do erro quadrático médio (RMSE) entre essa série que passou por modificação e a série que se mateve original. Os resultados foram registrados em tabelas para cada método de imputação e para dois tamanhos de amostra, 100 observações e 1000 observações. Esses resultados, que se encontram à seguir, foram separados em subceções para cada tamanho de amostra. Dessa forma, podemos comparar o efeito de cada método de imputação de dados para amostras de cada tamanho separadamente, analisando o impacto em relação a quantidade de dados faltantes gerados e o ajuste da série.

Amostras de 100 observações:

Nessa subseção se encontram os resultados obtidos através do processo de imputação de dados faltantes à partir de diferentes métodos em modelos de séries temporais, de diferentes ajustes, todos com uma amostra de 100 observações.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.3892315	0.3148000	0.2144880	0.1182557	0.2564963
10	0.4468177	0.3444869	0.3684174	0.4033684	0.3813407

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
20	0.5851220	0.4943355	0.5647307	0.6563425	0.4924350
40	0.6918194	0.7612301	0.7140997	0.9830260	0.8438136

Tabela 1: Resultados dos cálculos do RMSE com dados faltantes foram substituidos pela média (n=100).

A Tabela 1 revela que, à medida que o percentual de dados faltantes nas séries ajustadas aumenta, o valor do RMSE também tende a crescer. Isso indica que o método de imputação resulta em séries cada vez mais distantes da série original à medida que o número de dados ausentes aumenta.

Por outro lado, o impacto do aumento da complexidade do ajuste não segue um padrão tão claro. Embora o RMSE tenha aumentado em algumas situações e diminuído em outras, é possível notar uma leve tendência de aumento nos valores do RMSE em todas as porcentagens à medida que os modelos ajustados de séries temporais se tornam mais complexos. Essa variação sugere que a complexidade do modelo pode influenciar a precisão das previsões, mas não de maneira uniforme.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.1804544	0.3164807	0.2147843	0.1137387	0.2444201
10	0.1979142	0.3498192	0.3683235	0.4034666	0.3811761
20	0.4049888	0.5031434	0.5647040	0.6577882	0.4927660
40	0.6466893	0.7632819	0.7210497	0.9913392	0.8409461

Tabela 2: Resultados dos cálculos do RMSE com dados faltantes foram substituidos pela mediana (n=100).

A Tabela 2, assim como observado anteriormente na Tabela 1, demonstra que, à medida que o percentual de dados faltantes nas séries ajustadas aumenta, o valor do RMSE tende a crescer. Isso evidencia que o método de imputação gera séries progressivamente mais distantes da série original à medida que a quantidade de dados ausentes se eleva.

Por outro lado, o impacto do aumento na complexidade do ajuste não apresenta um padrão bem definido. Embora em algumas situações o RMSE tenha aumentado e, em outras, diminuído, percebe-se uma leve tendência de crescimento nos valores de RMSE em todas as porcentagens à medida que os modelos ajustados de séries temporais se tornam mais complexos. Essa oscilação sugere que a complexidade do modelo pode afetar a precisão das previsões, mas de maneira não uniforme.

Não foi observada uma mudança clara nos resultados que indique influência da substituição do método de imputação de dados da média para a mediana.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2811430	0.3922194	0.1801501	0.1923050	0.3409044
10	0.3189514	0.4801306	0.1981648	0.4239092	0.3932324
20	0.5615312	0.6840932	0.4562173	0.7876901	0.5994673
40	0.7089940	0.8710560	0.7441356	1.0883977	0.8507686

Tabela 3: Resultados dos cálculos do RMSE com dados faltantes foram substituidos atravéz do método LOCF (n = 100)

Na Tabela 3, foi identificado o mesmo padrão de aumento da distância entre as séries com imputações e as séries originais à medida que a porcentagem de dados faltantes aumentava. Observou-se um comportamento semelhante ao dos demais métodos de imputação em relação ao aumento da complexidade dos ajustes. No entanto, destaca-se que, no ajuste AR(2) ($\phi_1 = 0.2, \phi_2 = 0.4$), o método LOCF apresentou desempenho significativamente melhor em comparação com os outros ajustes.

Em relação à imputação pela média e pela mediana, o método LOCF resultou em um aumento do RMSE na maioria das combinações de porcentagens de dados faltantes e ajustes. A única exceção foi no ajuste AR(2) ($\phi_1 = 0.2, \phi_2 = 0.4$), onde o LOCF teve desempenho superior em todas as porcentagens de dados faltantes, exceto em 40%, onde o RMSE foi semelhante ao dos demais métodos. Além disso, observou-se que, com o LOCF, os valores do RMSE tendiam a crescer de forma mais acentuada conforme aumentavam as porcentagens de dados faltantes.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2811430	0.3922194	0.1801501	0.1923050	0.3409044
10	0.3189514	0.4801306	0.1981648	0.4239092	0.3932324
20	0.5615312	0.6840932	0.4562173	0.7876901	0.5994673
40	0.7089940	0.8710560	0.7441356	1.0883977	0.8507686

Tabela 4: Resultados dos cálculos do RMSE com dados faltantes substituidos atravéz do método NOCB (n = 100)

Na Tabela 4 são apresentados os valores do RMSE obtidos com a substituição dos dados faltantes pelo método NOCB. O comportamento em relação ao aumento da porcentagem de dados faltantes segue o padrão observado em outros métodos de imputação. No entanto, em relação à complexidade dos modelos, o NOCB gerou resultados distintos.

Comparando com os métodos de imputação pela média e mediana, houve uma redução do RMSE nos testes com 40% de dados faltantes na maioria das porcentagens, exceto no modelo ARMA (1,1)($\phi = 0.2, \theta = 0.4$), onde os resultados foram praticamente iguais. Em relação ao método LOCF, os resultados para variaram entre rasoavelmente melhores, rasoavelmente piores e semelhantes. No geral, não é possível afirmar que o método NOCB se destaca como superior aos demais métodos de imputação.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.3387105	0.2130691	0.2876818	0.1881545	0.1185403
10	0.3665967	0.3135915	0.2649126	0.3714423	0.4110724
20	0.6342673	0.6562248	0.5024862	0.5985662	0.7142722
40	0.7497217	0.9526146	1.0898719	1.5165599	0.8538800

Tabela 5: Resultados dos cálculos do RMSE com dados faltantes substituidos atravéz do método de Interpolação Cúbica (slice)

A Tabela 5 revela que o padrão de crescimento do RMSE acompanha o aumento da complexidade dos modelos e da porcentagem de dados faltantes, mantendo uma tendência semelhante aos demais métodos de imputação. No caso da Interpolação, os maiores valores de RMSE foram registrados nas proporções mais altas de dados ausentes (20% e 40%), sugerindo uma redução na confiabilidade do método quando há uma quantidade excessiva de valores faltantes. Em contrapartida, nas porcentagens menores (5% e 10%), observou-se maior estabilidade nos RMSE, com desempenho ocasionalmente superior ao de outros métodos, embora sem um padrão consistente.

De maneira geral, a aplicação da Interpolação Cúbica não demonstrou melhorias significativas nos valores de RMSE em comparação com as outras abordagens de imputação analisadas para $n = 100$.

Amostras de 1000 observações:

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2357325	0.2518541	0.2570215	0.4021781	0.3428851
10	0.3449642	0.3665185	0.3797436	0.5752837	0.4279775
20	0.4934541	0.4960622	0.5387605	0.8067815	0.5667309
40	0.6981557	0.7561148	0.7177946	1.2912187	0.7537121

Tabela 6: Resultados dos cálculos do RMSE com dados faltantes foram substituidos pela média (n=1000).

A Tabela 6 mostra que, com o aumento do percentual de dados faltantes nas séries ajustadas, os valores da raiz do erro quadrático médio (RMSE) também se elevam. Esse comportamento sugere que o método de imputação gera séries progressivamente mais distantes da série original à medida que cresce a quantidade de dados ausentes.

Diferentemente dos resultados obtidos na imputação pela média com 100 observações, nesta análise o aumento da complexidade do ajuste ocasiona valores mais divergentes. Destaca-se que, no modelo mais elaborado (AR(2) com $\phi = 0.4$ e $\phi = 0.5$), houve um crescimento expressivo no RMSE em relação aos demais modelos em todas as porcentagens avaliadas. Esse padrão sugere que a complexidade do modelo pode influenciar negativamente a precisão das previsões.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2351958	0.2514991	0.2551379	0.4030686	0.3428318
10	0.3450636	0.3688178	0.3784964	0.5708639	0.4279534
20	0.4933475	0.4978918	0.5391817	0.8006419	0.5666016
40	0.6971285	0.7563042	0.7171511	1.2834627	0.7534568

Tabela 7: Resultados dos cálculos do RMSE com dados faltantes foram substituidos pela mediana (n=1000)

Na Tabela 7 é possível observar novamente que não há diferença clara entre os métodos de imputação pela média e pela mediana. As considerações feitas à respeito da Tabela 56também são válidas para a Tabela 7.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2874326	0.2906962	0.2160859	0.3765961	0.2264049
10	0.4236037	0.4299546	0.3358650	0.4602698	0.4112478
20	0.6006022	0.6048240	0.4751475	0.7056996	0.6160392
40	0.9239159	0.9284677	0.7727094	1.1556546	1.0360355

Tabela 8: Resultados dos cálculos do RMSE com dados faltantes foram substituidos atravéz do método LOCF (n = 1000)

A Tabela 8 apresentou um padrão de comportamento do RMSE semelhante ao observado na comparação entre o método LOCF e as imputações por média e mediana com 100 observações. O aumento geral do RMSE, previamente identificado nas simulações com $n = 1000$, também foi constatado nesta análise.

Destaca-se uma redução significativa do RMSE no ajuste AR(2) ($\phi_1 = 0.2, \phi_2 = 0.4$), em linha com o que já havia sido observado na simulação com $n = 100$. No entanto, nesta simulação, essa redução também foi registrada no ajuste AR(2) ($\phi_1 = 0.4, \phi_2 = 0.5$), que apresentou RMSE inferior ao das imputações por média e mediana em todas as porcentagens avaliadas. Apesar disso, este ajuste teve desempenho inferior aos demais da mesma simulação, refletindo o impacto negativo de seu alto grau de complexidade.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2761077	0.2049902	0.2559661	0.2664016	0.2836093
10	0.3888442	0.3330139	0.3855374	0.3872747	0.4356191
20	0.5867092	0.5400304	0.5359320	0.6513307	0.6138997
40	0.9128088	0.8432034	0.7856316	1.0116014	0.9838729

Tabela 9: Resultados dos cálculos do RMSE com dados faltantes foram substituidos atravéz do método NOCB (n = 1000)

O comportamento dos resultados do RMSE utilizando o método de imputação NOCB para 1000 observações, representado na Tabela 9, é semelhante ao dos demais métodos no que se refere à reação da métrica ao aumento da porcentagem de dados faltantes. O destaque desse método está na "estabilidade" observada nos valores do RMSE entre os diferentes tipos de ajustes. Diferentemente do observado em outras simulações, especialmente nas realizadas com 100 observações, o método NOCB para $n = 1000$ apresentou valores de RMSE muito

próximos entre os ajustes. Uma variação mais acentuada foi percebida apenas para 40% de dados faltantes.

Em comparação ao método de imputação pela média, o NOCB apresentou desempenho superior nos modelos AR(1)($\phi = 0.4$) e ARMA(1,1)($\phi = 0.2, \theta = 0.4$), mas foi consideravelmente inferior no modelo AR(2)($\phi_1 = 0.4, \phi_2 = 0.5$). Já em relação ao método LOCF, o método NOCB apresentou resultados semelhantes no modelo ARMA(1,1)($\phi = 0.2, \theta = 0.4$) e ligeiramente melhores no modelo AR(2)($\phi_1 = 0.2, \phi_2 = 0.4$, porém teve desempenho inferior em todos os outros modelos.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.3387105	0.2130691	0.2876818	0.1881545	0.1185403
10	0.3665967	0.3135915	0.2649126	0.3714423	0.4110724
20	0.6342673	0.6562248	0.5024862	0.5985662	0.7142722
40	0.7497217	0.9526146	1.0898719	1.5165599	0.8538800

Tabela 10: Resultados dos cálculos do RMSE com dados faltantes foram substituidos atravéz do método NOCB (n = 1000)

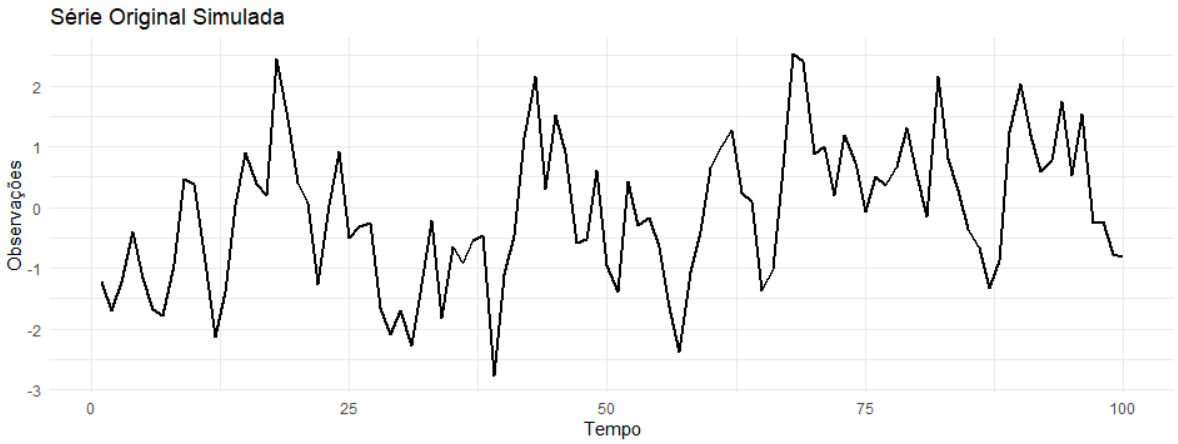
O comportamento do RMSE para o método de imputação por Interpolação Cúbica com 1000 observações seguiu o padrão observado nos outros métodos, confirmando que as considerações sobre o aumento do RMSE conforme cresce a porcentagem de dados faltantes e a complexidade dos modelos permanecem válidas.

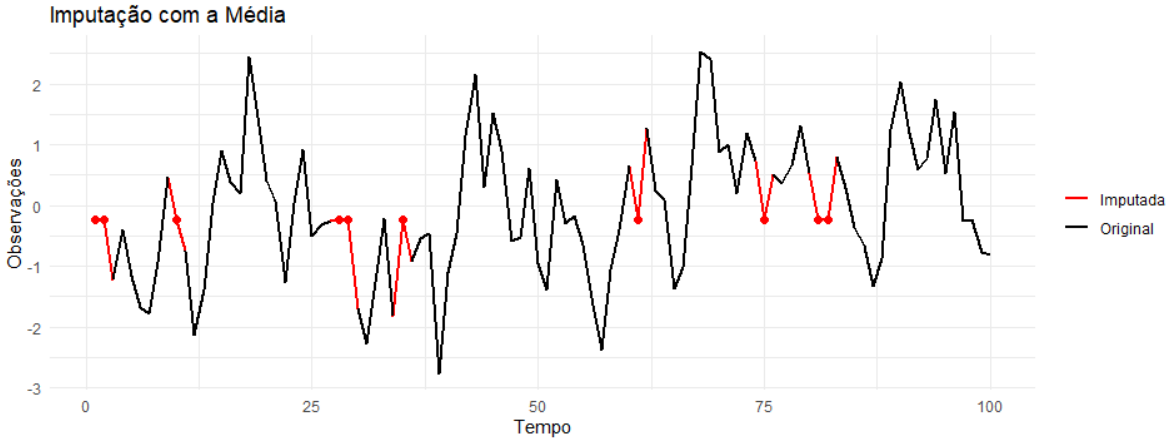
Comparado com os demais métodos de imputação para $n = 1000$, a Interpolação Cúbica demonstrou desempenho notavelmente inferior, apresentando RMSE consistentemente maior em todas as proporções de dados faltantes para mais de um modelo, quando comparado com as abordagens de imputação pela média, LOCF e NOCB. O padrão identificado na Tabela 5, que mostrava maior estabilidade nas menores porcentagens de dados faltantes (5% e 10%), também foi observado com $n = 1000$; no entanto, os resultados da Interpolação Cúbica apresentaram variações mesmo nessas condições.

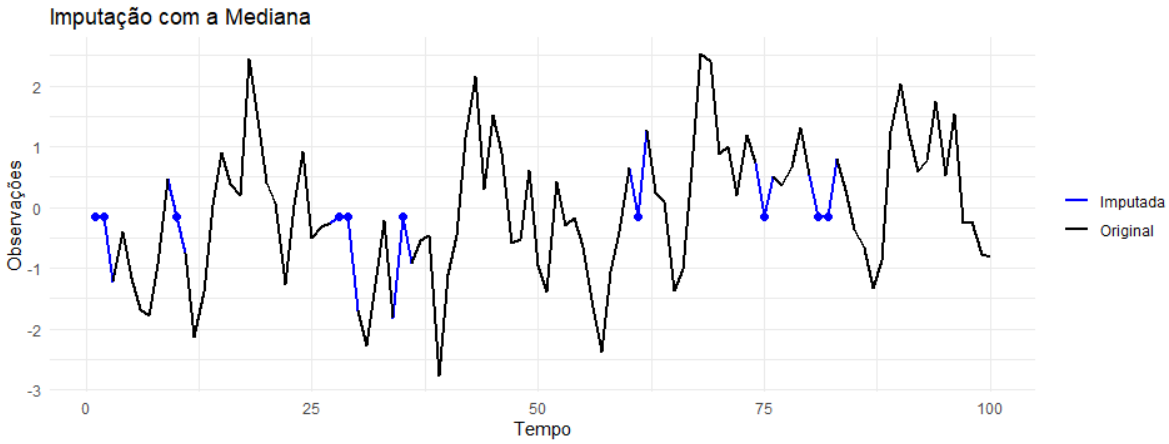
No geral, a Interpolação Cúbica não conseguiu produzir resultados que se aproximassem satisfatoriamente dos modelos originais das séries geradas, podendo indicar limitações na sua aplicabilidade em cenários com grandes volumes de dados e altos níveis de ausência.

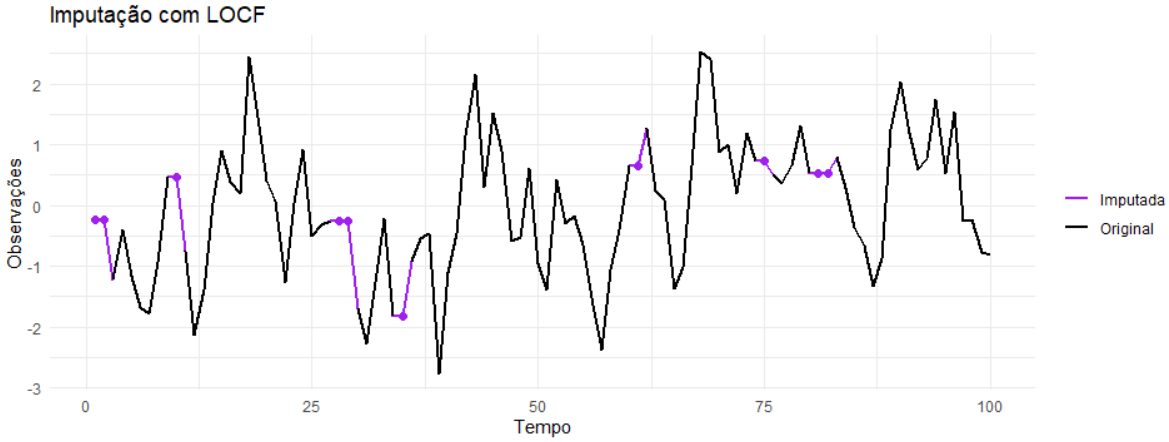
Visualização gráfica das imputações

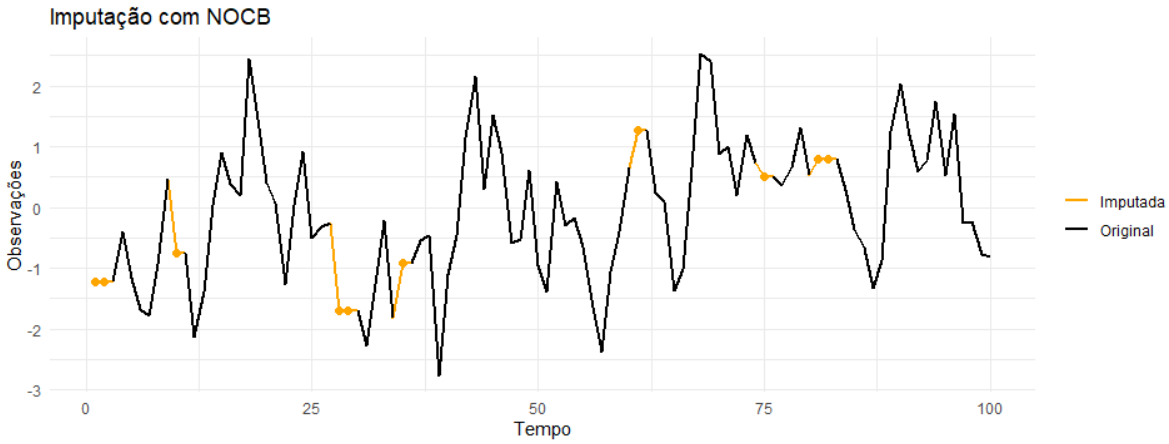
Para visualização da imputação dos dados faltantes na série foram gerados gráficos das séries temporais ARMA(1,1)($\phi = 0.2, \theta = 0,4$) com 10% de dados faltantes, onde os trechos coloridos representam os dados imputados. Por vias de comparação, o gráfico da simulação sem a retirada dos dados faltantes também foi gerada.

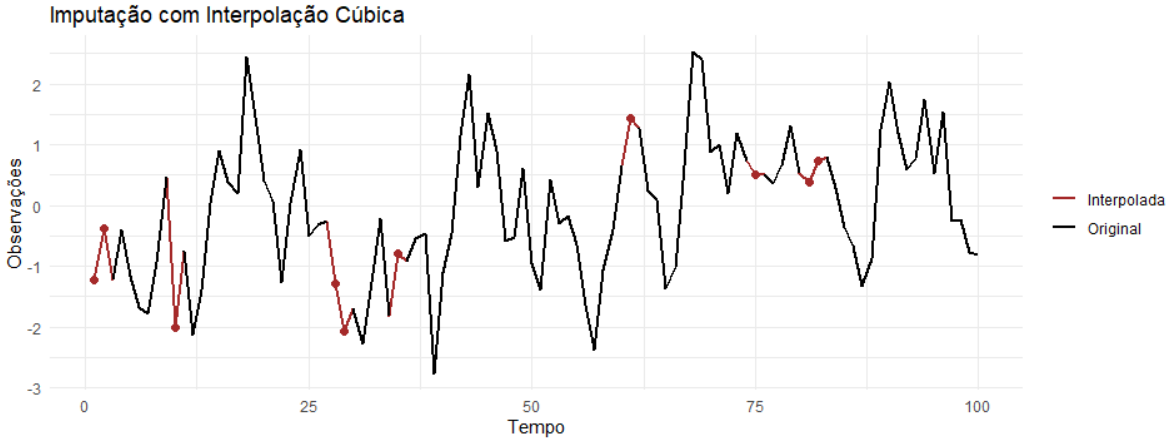

Série Original Simulada
Observações
2
1
0
-1
-2
-3
0
25
50
75
100
Tempo


Imputação com a Média
Observações
2
1
0
-1
-2
-3
0
25
50
75
100
Tempo
Imputada
Original


Imputação com a Mediana
Observações
Tempo
Imputada
Original


Imputação com LOCF
Observações
Tempo
Imputada
Original


Imputação com NOCB
Observações
Tempo
Imputada
Original


Imputação com Interpolação Cúbica
Observações
Tempo
Interpolada
Original