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	0.4699637	0.5463677	0.5634988	0.7271051	0.5849156
10	0.5854688	0.7402376	0.7618335	0.9470237	0.7286870

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
20	0.7614553	0.9848464	0.9750527	1.2051350	0.9203268
40	1.0173459	1.3679721	1.2347454	1.5877383	1.2440685

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	0.4604036	0.5351261	0.5898960	0.7468039	0.5758298
10	0.5581029	0.7094723	0.7776561	0.9797072	0.7048509
20	0.7183403	0.9987362	0.9429366	1.1886295	0.8812158
40	0.9560941	1.4249596	1.1549140	1.5077228	1.2053637

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	0.3140226	0.4194391	0.4438822	0.3603400	0.3393137
10	0.4796570	0.7703249	0.5623183	0.4375707	0.5698656
20	0.6342560	1.0623778	0.7659245	0.8086849	0.7281756
40	1.0707756	1.3236721	1.1740775	1.3237643	1.3039913

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.

Em relação à imputação pela média e pela mediana, o método LOCF resultou em uma diminuição do RMSE em quase todas as combinações de porcentagens de dados faltantes e ajustes. A melhora foi mais significativa sobretudo nos ajustes de maior complexidade como AR (2)($\phi = 0.4, \phi = 0.5$) e ARMA (1,1)($\phi = 0.2, \theta = 0.4$).

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.3763733	0.2504095	0.3903083	0.3874545	0.3905667
10	0.4101555	0.3797480	0.5530478	0.5201887	0.6065266
20	0.6844843	0.8640071	0.8300174	0.8866716	0.8431176
40	0.9237649	1.1103111	1.1763707	1.3939464	1.1590771

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.

Comparando com os métodos de imputação pela média e mediana, houve uma redução do RMSE em todas as combinações de porcentagens de dados faltantes e ajustes. Novamente

destacou-se a melhora nos ajustes de maior complexidade, principalmente o ajuste AR $(2)(\phi = 0.4, \phi = 0.5)$. Em relação ao método LOCF, houve melhora significativa na performance no ajuste AR $(1)(\phi = 0.6)$ enquanto nos demais ajustes houve melhora sutil ou não houve melhora

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.2907758	0.3032111	0.4362034	0.4420587	0.2434931
10	0.3773786	0.5495050	0.5741398	0.4064126	0.5729463
20	0.5431922	0.7903430	0.7631143	0.6609947	0.7020057
40	1.1191562	1.1651119	1.4545111	1.3649269	1.4805275

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.

De maneira geral, a aplicação da Interpolação Cúbica demonstrou melhorias significativas nos valores de RMSE em comparação com a imputação por média e mediana. Em relação ao método NOCB, houve diminuição do RMSE em alguns ajustes mas aumento em outros. Destaca-se que na porcentagem de 20% de dados faltantes, o método da Interpolação Cúbica performou melhor que todos os demais métodos.

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	0.5503887	0.6179312	0.5257339	0.9687625	0.6099121
10	0.6904739	0.7889848	0.6838812	1.2559460	0.7866031
20	0.8762627	1.0198500	0.8935735	1.6019046	1.0042092
40	1.1656180	1.3764251	1.2170042	2.0331025	1.3143559

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 houve aumento significativo do RMSE em todas os ajustes com excessão de AR(2) com $\phi = 0.2$ e $\phi = 0.4$, que apresentou singela melhora. 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 e os testes feitos com 100 observações. Esse padrão sugere que a aumento do número de observações pode influenciar negativamente a precisão das previsões quando utilizado esse método 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	0.5403800	0.6003763	0.5006211	0.9777281	0.6031007
10	0.6637104	0.7522562	0.6511108	1.2533568	0.7676380
20	0.8329958	0.9716486	0.8319952	1.5644571	0.9639041
40	1.0810908	1.2869749	1.1213610	1.9354181	1.2535745

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 6 também 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	0.4356524	0.3550447	0.4000016	0.4012626	0.4025512
10	0.6369169	0.4984366	0.5297828	0.5852661	0.5648038
20	0.7996506	0.7477458	0.7420153	1.1147183	0.8145155
40	1.1725513	1.1087192	1.0384939	1.6771536	1.1573435

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

Aqui está o seu texto revisado para maior clareza e fluidez:

A Tabela 8 revelou 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. Observou-se uma melhora no desempenho em relação aos métodos de imputação por média e mediana em todas as observações.

Quanto à aplicação do método nos modelos com n=100, verificou-se uma redução do RMSE em alguns casos, enquanto em outros houve aumento. Isso sugere que o impacto do número de observações nas simulações que utilizam o método de imputação LOCF não segue um padrão linear.

Porcentagem	AR_0.4	AR_0.6	AR_0.2_0.4	AR_0.4_0.5	ARMA_0.4_0.2
5	0.4732984	0.4307927	0.3478948	0.3949038	0.4300506
10	0.6006231	0.5609582	0.4793141	0.5845785	0.5643438
20	0.8476235	0.7573554	0.6947071	1.0831990	0.8346765
40	1.1122577	1.1788858	0.9692013	1.5163241	1.1931220

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 resultado observado utilizando o método LOCF. Esse padrão pode ser observado tanto na comparação com os métodos de média e mediana quanto se comparado o uso do método com 100 e 1000 observações.

O destaque desse método quando comparado diretamente com o LOFC são os valores do RMSE dos ajustes AR(2)($\phi_1 = 0.4, \phi_2 = 0.5$) e AR(2)($\phi_1 = 0.2, \phi_2 = 0.4$), que apresentaram diminuição do valor da métrica em todas 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	0.2907758	0.3032111	0.4362034	0.4420587	0.2434931
10	0.3773786	0.5495050	0.5741398	0.4064126	0.5729463
20	0.5431922	0.7903430	0.7631143	0.6609947	0.7020057
40	1.1191562	1.1651119	1.4545111	1.3649269	1.4805275

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

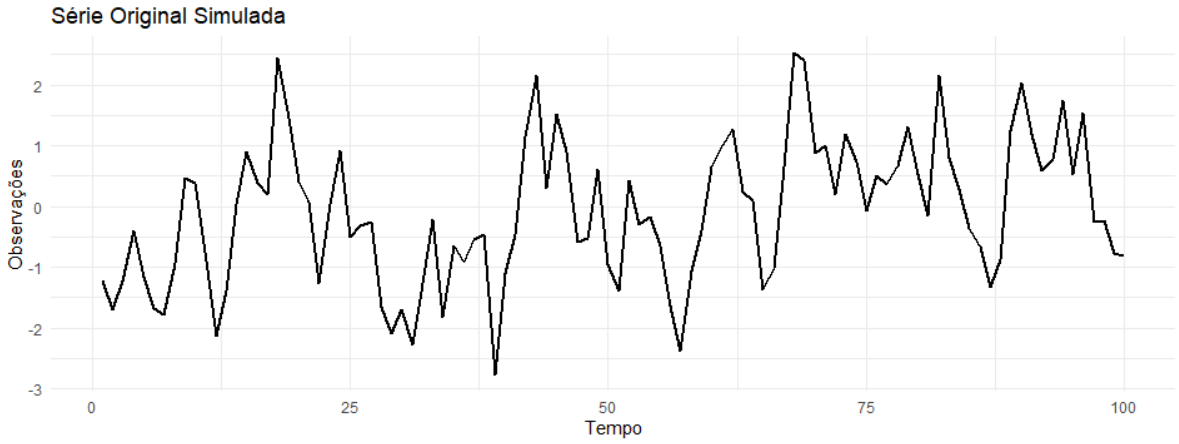
O comportamento do RMSE para o método de imputação por Interpolação Cúbica com 1000 observações seguiu o padrão observado nos outros métodos, confirmando que as considerações sobre o aumento do RMSE conforme cresce a porcentagem de dados faltantes e a complexidade dos modelos permanecem válidas.

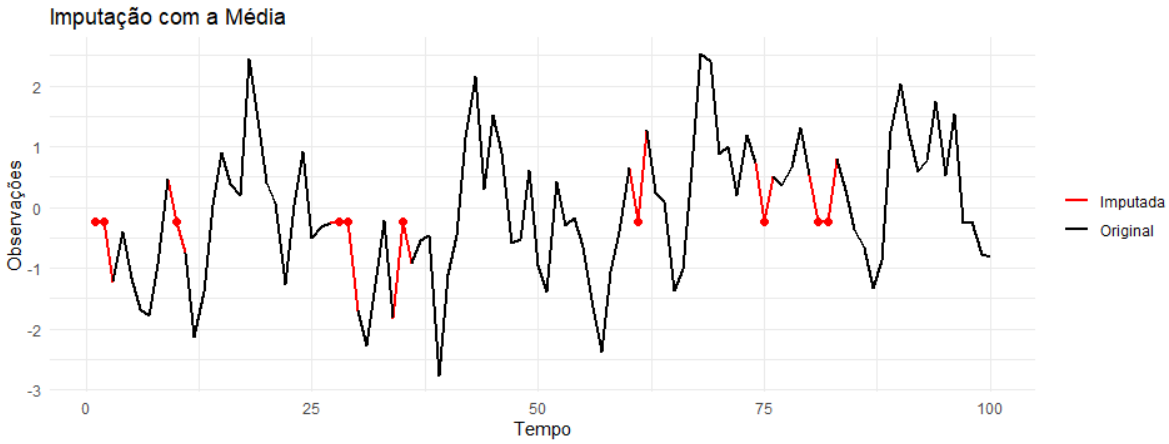
Comparado com os demais métodos de imputação para $n = 1000$, a Interpolação Cúbica demonstrou desempenho no geral superior. Quando posto em comparação com as imputações pela média e mediana, o desempenho foi melhor em todos os cenários. Já quando comparado com os métodos LOCF e NOCB, houve uma melhora em alguns ajustes, mas piora no desempenho dos ajustes AR(2)($\phi_1 = 0.4, \phi_2 = 0.5$) e AR(2)($\phi_1 = 0.2, \phi_2 = 0.4$), apresentando RMSE c maior em todas as proporções de dados faltantespara os modelos. Por fim, quando comparamos com o método aplicado em ajustes com 100 observações, foi observado uma melhor do desempenho em alguns ajustes e piora em outros de forma não linear.

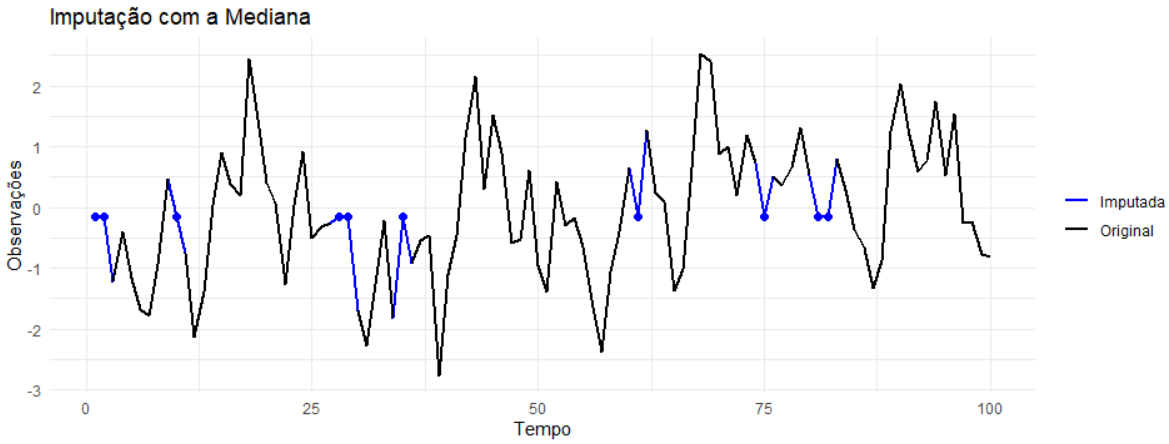
Embora o método seja superior ao de imputação por média e mediana, sua eficácia tem vantagens e desvantagens em relação aos métodos LOCF e NOCB.

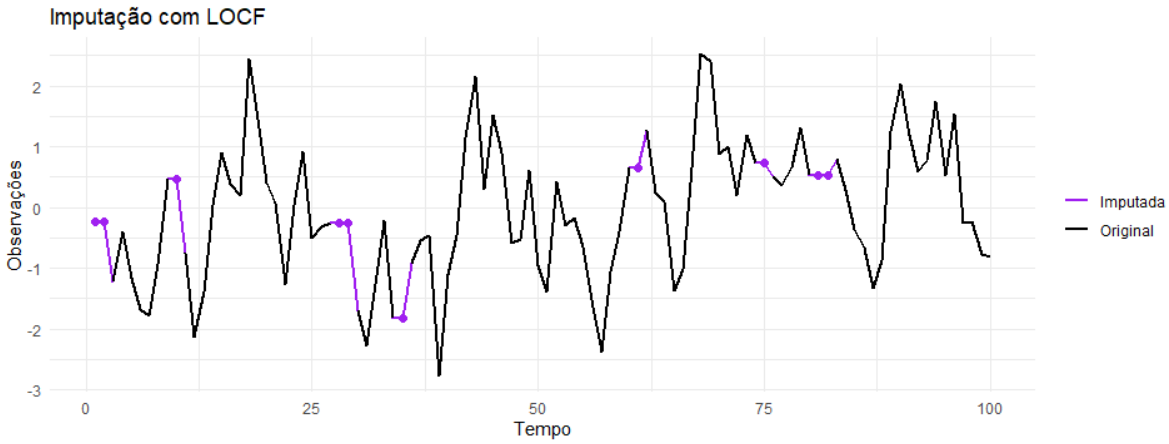
Visualização gráfica das imputações

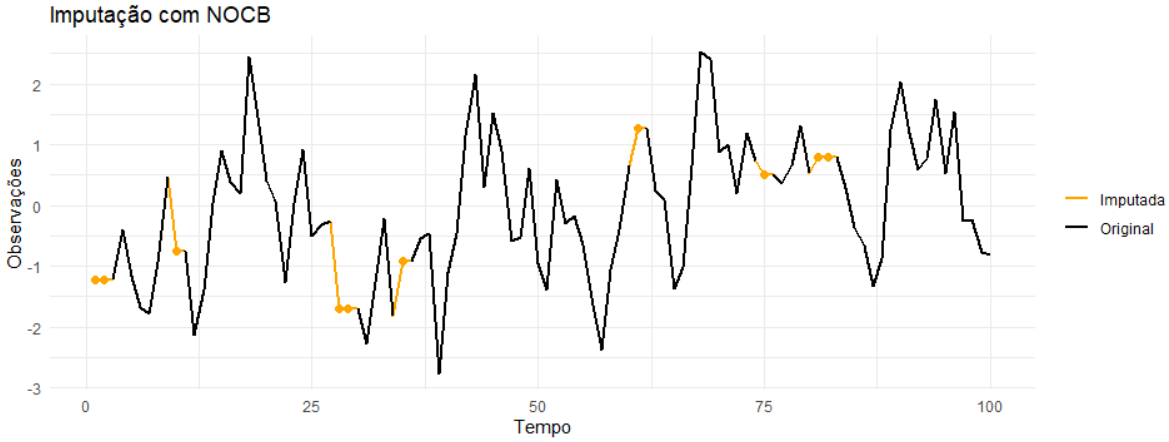
Para visualização da imputação dos dados faltantes na série foram gerados gráficos das séries temporais ARMA(1,1)($\phi = 0.2, \theta = 0, 4$) com 10% de dados faltantes, onde os trechos coloridos representam os dados imputados. Por vias de comparação, o gráfico da simulação sem a retirada dos dados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

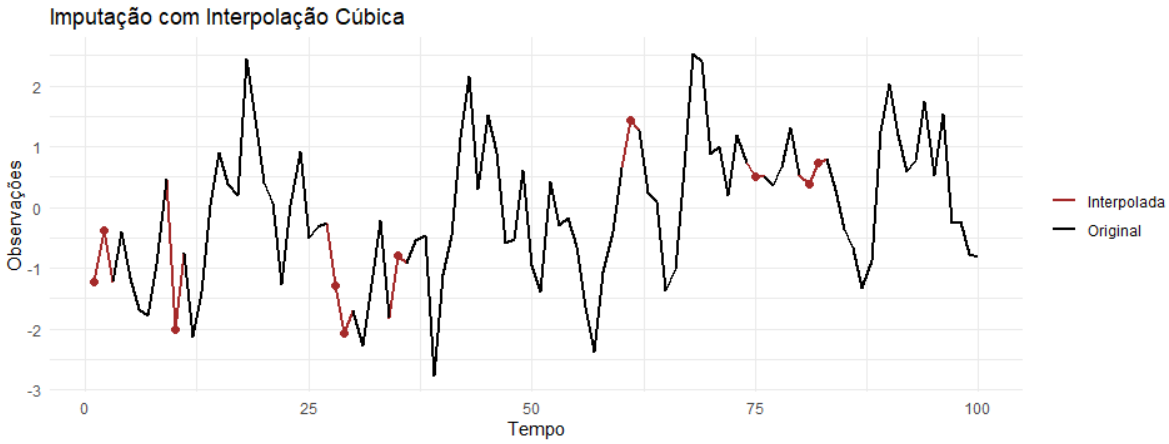



Imputação com a Média
Observações
Tempo
Imputada
Original


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
0
25
50
75
100
2
1
0
-1
-2
-3
Interpolada
Original