Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Maike Willian Martins dos Santos

Análise de Séries Temporais para Previsão de Máquinas em Serviços de DaaS

Curitiba
2024

Maike Willian Martins dos Santos

Análise de Séries Temporais para Previsão de Máquinas em Serviços de DaaS

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Wagner Hugo Bonat

Curitiba
2024

Análise de Séries Temporais para Previsão de Máquinas em Serviços de DaaS

Time Series Analysis for Forecasting Machines in DaaS Services

Maike Willian Martins dos Santos[1], Maike W. Martins dos Santos[2]

[1]Aluno do programa de Especialização em Data Science & Big Data
[2]Professor do Departamento de Estatística - DEST/UFPR, wbonat@gmail.com

Este trabalho tem como fundamento principal o estudo da evolução da quantidade de máquinas ligadas em dois ambientes de Data as a Service (DaaS) a partir de tendências e melhoria nos data centers. Os dados utilizados foram extraídos do Citrix, uma plataforma de gerenciamento de máquinas DaaS, para analisar a quantidade de máquinas ativas hora a hora. Para a revisão dos dados, foi feita a aplicação do modelo de Decomposição Sazonal de médias via suavização exponencial para a captação de padrões de cada dia da semana. O trabalho se encerrou com uma série de análises de resíduo que mostrou significativas autocorrelações e foram necessários modelos mais complexos para a obtenção de uma previsão mais acurada. A conclusão a que se chegou por meio do referido experimento foi a de que é, sim, possível melhorar a infraestrutura de data centers ao entender melhor como as máquinas são utilizadas e aloca melhor os recursos, conseguindo reduzir custos. Além disso, foram realizados testes de validação do modelo escolhido, incluindo AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) e o teste de Ljung-Box, os quais evidenciaram a presença de padrões não aleatórios nos resíduos, destacando a importância de incorporar a sazonalidade e outros fatores temporais nos modelos preditivos.
Palavras-chave: Séries temporais, previsão, DaaS, máquinas, data centers, Citrix

This work is primarily based on the study of the evolution of the number of active machines in two Data as a Service (DaaS) environments, focusing on trends and improvements in data centers. The data used were extracted from Citrix, a DaaS management platform, to analyze the number of active machines on an hourly basis. For data review, the Seasonal Decomposition model using exponential smoothing was applied to capture the patterns for each day of the week. The study concluded with a series of residual analyses that showed significant autocorrelations, indicating that more complex models were necessary to achieve more accurate predictions. The conclusion reached through this experiment was that it is indeed possible to improve data center infrastructure by better understanding how the machines are used and by allocating resources more efficiently, thereby reducing costs. Additionally, validation tests of the chosen model were conducted, including AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and the Ljung-Box test, which revealed the presence of non-random patterns in the residuals, highlighting the importance of incorporating seasonality and other temporal factors into predictive models.
Keywords: Time series, forecasting, DaaS, machines, data centers, Citrix

1. Introdução

O presente trabalho se destina a estudar a evolução do número de máquinas conectadas em um ambiente de Data as a Service (DaaS). Data as a Service (DaaS) é um modelo de serviço baseado em nuvem que oferece acesso sob demanda a recursos de armazenamento, processamento e gerenciamento de computadores [1]. Ele permite que as organizações gerenciem e analisem dados de forma eficiente, sem a necessidade de manter sua própria infraestrutura de dados. Com este propósito, obtivemos os dados do Citrix, que é um gerenciador das máquinas virtuais, o que nos forneceu toda a informação detalhada e respectiva evolução delas ao longo do tempo.

Nos ambientes de DaaS, o conhecimento da utilização das máquinas torna-se fundamental, a fim de assegurar o uso mais eficiente dos recursos disponíveis, tornando assim a redução dos custos oportuna. Por conseguinte, a quantidade de máquinas ligadas tem

comportamento variável ao longo do dia e da semana, o que é gerado por vários fatores como por exemplo os horários do trabalho, picos de utilizações e manutenção, entre outros. Para a apreensão da natureza desses padrões foram utilizados os modelos de decomposição sazonal e alisamento exponencial, possibilitando a identificação e a análise das componentes sazonais diárias e semanais presentes nos dados.

Infelizmente na base de dados, por conta de algum problema na coleta, teve alguns dados missing. Para o tratamento dos dados missing, empregamos o valor médio centrado na hora e no dia da semana, considerado que a respectiva distribuição das amostras por tempos refletia adequadamente as medianas calculadas. Desta forma, nas análises de séries temporais, a utilização dos métodos de imputação dos missing values varia devido à necessidade de manter a preservação da estrutura temporal dos dados no sentido dos preservar os modelos preditivos.

Foi realizada também a análise dos resíduos para a avaliação da qualidade dos modelos ajustados, e foi observada a existência de correlação estatisticamente significativa entre os próprios resíduos, o que indicava a necessidade de emprego dos modelos mais complexos no sentido de captar a estrutura temporal dos respectivos dados. Posteriormente, foram realizados os testes de Ljung-Box, ao qual foi verificado a presença dos padrões não aleatórios dos resíduos, e consequentemente a necessidade de considerar a sazonalidade na escolha dos modelos preditivos a serem utilizados.

O trabalho foi dividido em três seções principais. Na seção 2, discutimos brevemente os métodos de decomposição sazonal e alisamento exponencial, bem como os modelos selecionados para o estudo. Na seção 3, apresentamos o delineamento dos experimentos criados para o estudo e a descrição do processo de modelagem e avaliação dos modelos. Por fim, na seção 4, discutimos os resultados obtidos e suas implicações para a otimização de data centers em ambientes DaaS.

Para identificar o melhor modelo, foi utilizado uma abordagem sistemática que envolveu a análise de diferentes modelos e a avaliação de seus resíduos. Os modelos foram ajustados e validados utilizando a metodologia de validação cruzada, e os resíduos foram analisados para detectar autocorrelações e padrões não aleatórios. Os resultados dos testes de Ljung-Box nos permitiram avaliar a adequação dos modelos e determinar quais eram mais eficazes na captura da estrutura temporal dos dados.

Através desta análise, buscamos fornecer insights valiosos para a otimização dos data centers, melhorando a eficiência operacional e reduzindo custos. Este estudo contribui para o entendimento dos padrões de uso de máquinas em ambientes DaaS e oferece uma base sólida para futuras pesquisas e aplicações práticas.

2. Discussão

Neste estudo, apresentamos os resultados da análise de séries temporais dos dados da quantidade de máquinas ligadas em um ambiente DaaS. Os dados consistem em números inteiros que representam a quantidade de máquinas ligadas para cada data e hora. Os dados a serem analisados foram retirados do gerenciador da Citrix.

Para obter conclusões confiáveis, adotamos diferentes abordagens de decomposição sazonal para capturar tendências diárias e semanais. Foi utilizada a média para substituir os dados ausentes, na qual foi calculada em função da hora do dia e do dia da semana, permitindo-nos imputar valores precisos em conformidade com os padrões temporários. A análise de resíduos mostrou a presença de autocorrelações significativas, indicando que são necessários modelos mais complexos para descrever adequadamente a estrutura temporal dos dados. Os testes de Ljung-Box confirmaram que os fatores temporais não podem ser negligenciados ao ajustar um modelo.

A análise de resíduos revelou a presença de autocorrelações significativas, indicando que modelos mais complexos podem ser necessários para capturar adequadamente a estrutura temporal dos dados. Os testes de Ljung-Box reforçaram a necessidade de considerar fatores temporais ao ajustar os modelos.

Para concluir, poderíamos modelar adequadamente as tendências mais importantes apesar das mudanças de uso serem bastante complexas. No entanto, a autocorrelação dos resíduos sugere que a modelagem da série temporal poderia ser ainda mais precisa usando um modelo que inclua algumas componentes não lineares ou tecnologias avançadas de ajuste de série temporal. Ao mesmo tempo, o estudo atual fornece a base para otimizar a operação de data centers em ambientes DaaS, fornecendo uma série de recomendações valiosas para um melhor recurso-alvo e operação de centros de dados de nível intermediário. As áreas de investigação futura podem incluir o ajuste de métodos

mais avançados como redes neuronais ou aprendizado profundo.

3. Materiais e métodos

A base de dados utilizada para o estudo foi extraída do Citrix Manager,um componente crucial em ambientes DaaS. O Citrix é uma plataforma que permite a entrega segura de aplicações e desktops virtuais, proporcionando aos usuários acesso remoto a suas máquinas virtuais [2], foi utilizando a API do Citrix para a extração e Engenharia de Dados, cujos auxiliaram no trabalho apresentado. Esta funcionalidade é essencial para a gestão eficiente de recursos em data centers, otimizando a utilização de máquinas e garantindo a disponibilidade conforme a demanda dos usuários.

Para este estudo, foi realizada uma engenharia de dados utilizando a API do Citrix, como mostrado na pela Figura 1, permitindo a coleta precisa dos dados de uso de máquinas virtuais ao longo do tempo. A API do Citrix foi utilizada para extrair informações detalhadas sobre o número de máquinas ligadas em diferentes horários e dias da semana. Os dados coletados são números inteiros que representam a quantidade de máquinas ativas em cada ponto temporal, essenciais para a análise de tendências e sazonalidades.

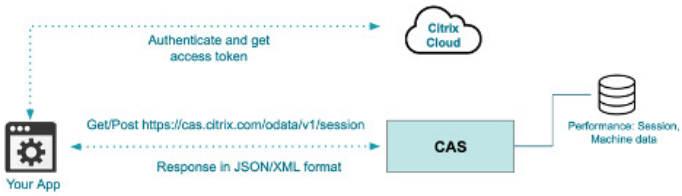

Figura 1: Fluxograma de extração dos dados.

Foi estruturada uma tabela em uma banco de dados SQL com a coleta dos dados. Utilizando a linguagem Python para coletar os dados foi feito um script para coletar os dados diariamente e salvando na tabela, como mostra a figura 2.

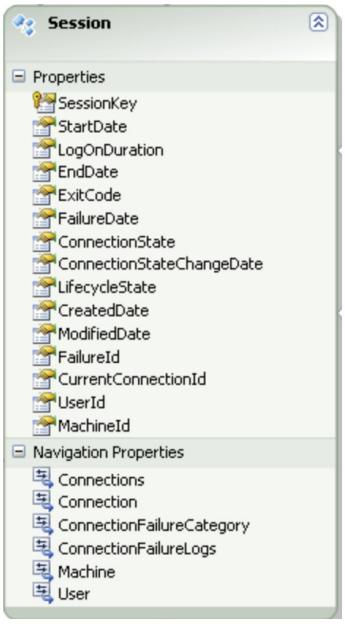

Figura 2: Tabela da extração dos dados do Citrix.

A fim de salvar o detalhe do dado e tratar os pontos faltantes, foi utilizada a média para preencher os pontos faltantes através da hora do dia e dia da semana, preservando esta imputação exata à capacidade do modelo da tendência temporal original dos dados.

Os modelos que foram usados para essa análise incluem técnicas de decomposição sazonal e alisamento exponencial que ajudam a capturar os padrões diários e semanais abaixo:

- **Modelo stlm (Seasonal and Trend decomposition using Loess):** Aplica decomposição de tendência e sazonalidade usando Loess [3].
- **Modelo ets (Exponential Smoothing State Space Model):** Utiliza alisamento exponencial para prever séries temporais com componentes de erro, tendência e sazonalidade [3].

Utilizamos esses modelos para verificar a sazonalidade ao longo do tempo juntando os modelos de predições abaixo:

- **Modelo auto.arima (AutoRegressive Integrated Moving Average):** Ajusta automaticamente um modelo ARIMA para capturar padrões e autocorrelações em séries temporais [4] .
- **TBATS:** É uma modelagem de série temporal que lida com múltiplas sazonalidades e transformações de Box-Cox, ARMA, tendência de nível, entre outros [5].

Com base nesses modelos, foi possível analisar os dados e capturar padrões e tendências anuais. Especi-

ficamente, a abordagem de decomposição permitiu separar os três componentes de tendência, sazonalidade e ruído, enquanto o alisamento exponencial suavizou as flutuações aleatórias dos dados.

A figura 3 mostra como se comporta a distribuição nos diferentes DataCenter ao longo 5 meses. E na figura 4 mostra a evolução semanal de acesso às máquinas por semana.

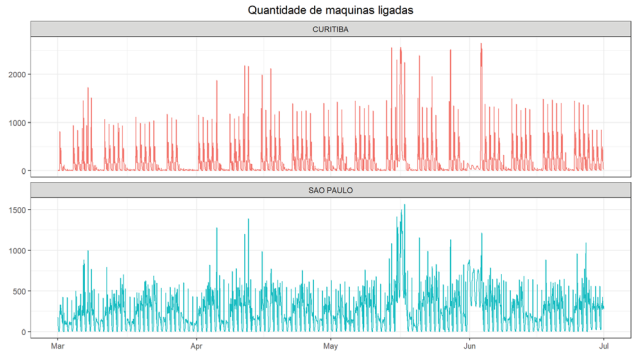

Figura 3: Distribuição da probabilidade predita via logit na amostra de validação.

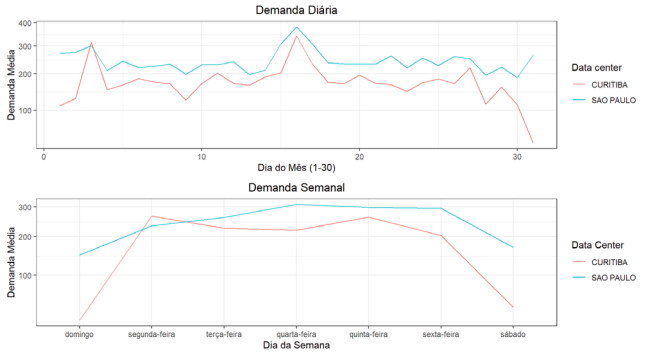

Figura 4: Medidas de validação

Para uma análise prévia dos DataCenter foi feito uma análise de decomposição STL. As duas imagens, como mostrado abaixo na figura 5, fornecem uma decomposição STL (Seasonal-Trend decomposition using Loess) das séries temporais referentes ao número de máquinas ligadas em dois data centers, um em Curitiba e outro em São Paulo. A decomposição STL separa a série temporal em três componentes principais: tendência (trend), sazonalidade (seasonal) e resíduos (remainder), além dos dados originais (data).

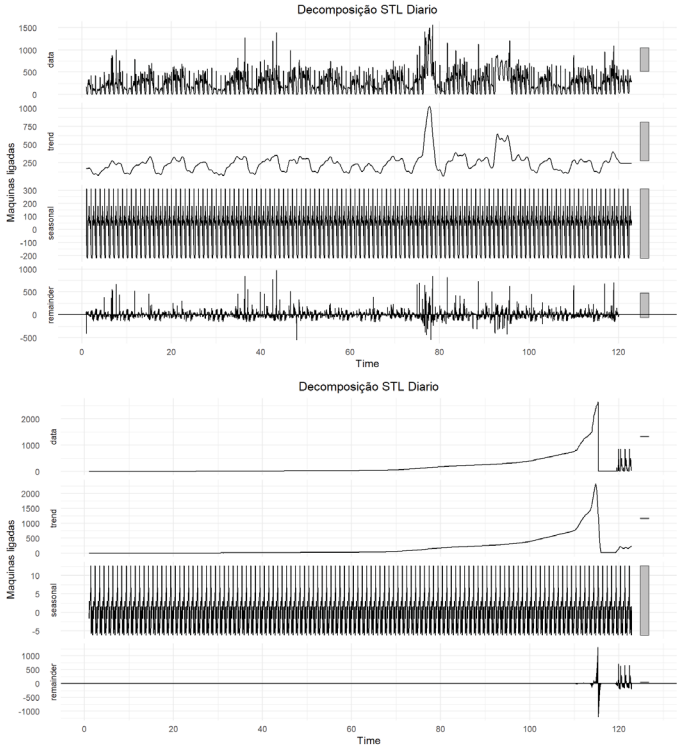

Figura 5: Medidas de validação dissidentes

A decomposição STL para ambos os datacenters mostram que, enquanto Curitiba tem uma tendência mais clara e sazonalidades bem definida, São Paulo é mais complexa para a tendência e sazonalidade, o que pode indicar uma demanda mais dinâmica e possivelmente menos previsível, ou diferenças, talvez, na natureza do trabalho que é processado em cada data Center.

Além disso, foram realizados testes de Ljung-Box para avaliar a presença de autocorrelação nos resíduos dos modelos ajustados e feita uma Análise dos Resíduos para verificar o comportamento do modelo escolhido. Esses testes são cruciais para identificar se os modelos capturam adequadamente a estrutura temporal dos dados ou se há necessidade de modelos mais complexos. E para explicar os resultados dos modelos utilizamos essas métricas:

- **Erro Teste (MAE - Mean Absolute Error):** Mede a precisão do modelo, sendo a média dos valores absolutos dos erros (diferença entre valor observado e previsto). Valores menores indicam melhor desempenho [6].
- **AIC (Akaike Information Criterion):** Critério de informação usado para comparar a qualidade dos modelos ajustados, penalizando modelos com mais parâmetros. Valores menores indicam modelos melhores [6].

- **BIC (Bayesian Information Criterion):** Similar ao AIC, mas com penalização mais forte para o número de parâmetros. Valores menores também indicam modelos melhores [6].

Por fim, para atingir os objetivos desejados utilizamos a linguagem *R* como ferramenta de análise e os pacotes *forecast* [7] , *timetk* [8], *padr* [9], e *ggplot2*, que contam com a implementação dos algoritmos e avaliações que foram necessárias para o estudo.

4. Resultados

A modelagem foi realizada utilizando diferentes métodos de séries temporais, incluindo STLM (Seasonal and Trend Decomposition using Loess), ETS (Exponential Smoothing State Space Model), auto.arima (Auto-Regressive Integrated Moving Average) e TBATS. Cada modelo foi avaliado com base em métricas como Erro Médio Absoluto (MAE), Critério de Informação de Akaike (AIC) e Critério de Informação Bayesiano (BIC), a fim de selecionar o modelo com melhor desempenho.

Conforme os resultados apresentados na tabela 6, o modelo STL aplicado às séries temporais com múltiplas sazonalidades (msts) foi identificado como melhor modelo com o menor erro de teste (MAE) para o data center de Curitiba. O modelo foi capaz de capturar detalhes dos padrões sazonais complexos porque o modelo é robusto e foi capaz de detectar a variabilidade dos dados de uso das máquinas virtuais. Assim sendo, o modelo STL foi capaz de detectar a sazonalidade e tendência ao longo do tempo para um melhor plano de infraestrutura de datacenter. A capacidade do modelo de identificar padrões sazonais e prever ações futuras da melhor maneira é muito significativa, porque as demandas dos usuários são muito variáveis, além do fato de que as demandas se alteram diariamente e semanalmente.

Datacenter	Modelos	Sazonalidades	Erro Teste (MAE)	AIC	BIC
CURITIBA	stlm	msts	77.08929	15119.323	15137.091
CURITIBA	tbats	msts	87.13095	17297.911	17368.983
CURITIBA	tbats	ts	99.57738	17459.597	17518.824
CURITIBA	stlm	ts	115.00595	17008.268	17026.035
CURITIBA	ets	ts	116.90476	17638.436	17798.347
CURITIBA	auto.arima	ts	185.20238	4162.602	4209.913
SAO PAULO	tbats	ts	71.58333	16611.205	16664.508
SAO PAULO	tbats	msts	73.01786	16570.799	16635.948
SAO PAULO	stlm	msts	75.64286	14782.281	14800.049
SAO PAULO	stlm	ts	139.43452	15947.214	15964.982
SAO PAULO	ets	ts	141.64286	16672.727	16832.638
SAO PAULO	auto.arima	ts	151.04167	3329.320	3388.458

Figura 6: Modelos testados com os testes repectivos

A análise dos resíduos realizada de acordo com as figuras 7 e 8 mostra que, embora o modelo STLM tenha identificado de maneira robusta os padrões sazonais, existem sinais de autocorrelação nos resíduos, conforme demonstrado nos gráficos de autocorrelação. Embora seja necessário melhorar os modelos para infraestrutura de datacenter, a análise de Ljung-Box mostrou que existem evidências de autocorrelação dos resíduos para os dois locais estudados. No caso de Curitiba, o valor foi de 1056,8 com 336 graus de liberdade e um erro de valor . Além disso, o teste complementar Box-Ljung apresentou um valor de 24,38 com um grau de liberdade. Assim, um erro de valor de maiores para os dois testes indica a presença de autocorrelação nos resíduos.

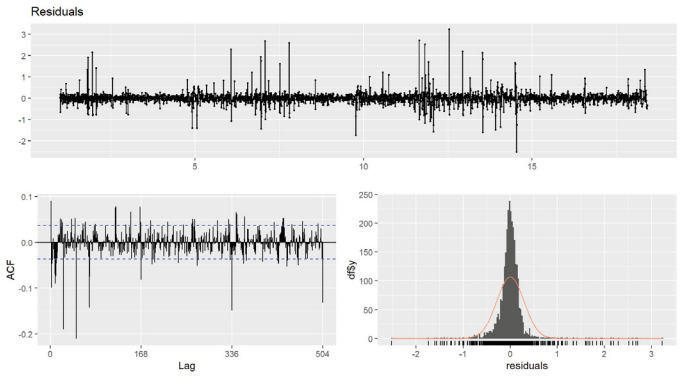

Figura 7: Análise de resíduos de Curitiba

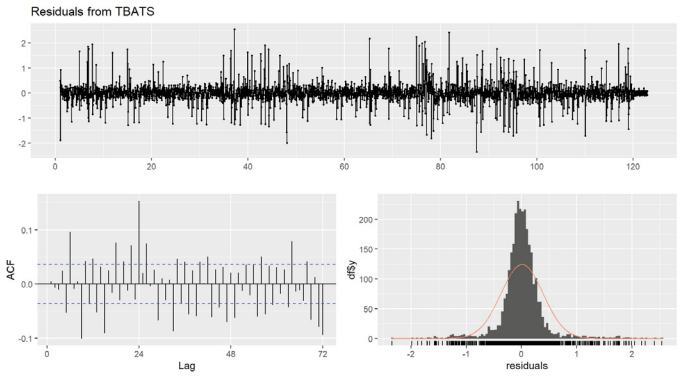

Figura 8: Análise de resíduos de São Paulo

O teste de Ljung-Box, cujos resultados estão detalhados na Figura 9, indicou a presença de autocorrelação significativa nos resíduos para ambos os locais analisados. Para Curitiba, o valor de Q* foi de 1056.8 com 336 graus de liberdade e um p-valor inferior a 2.2e-16, indicando forte evidência de autocorrelação nos resíduos. Além disso, o teste complementar Box-Ljung apresentou um valor de X-squared de 24.38 com 1 grau

de liberdade e p-valor de 7.907e-07, confirmando a necessidade de melhorias no modelo para este data center.

Em São Paulo, o Ljung-Box mostrou um Q* de 381.06 com 48 graus de liberdade e p-valor também inferior a 2.2e-16. Contudo, o teste Box-Ljung para o primeiro lag apresentou um X-squared de 0.066932 com p-valor de 0.7959, sugerindo que não há autocorrelação significativa para este lag específico. Isso implica que, embora o modelo tenha capturado bem a estrutura dos dados, ainda há espaço para ajustes, especialmente em Curitiba, onde a autocorrelação residual é mais evidente.

```
Análise de Resíduos - CURITIBA

        Ljung-Box test

data:  Residuals
Q* = 1056.8, df = 336, p-value < 2.2e-16

Model df: 0.   Total lags used: 336

        Box-Ljung test

data:  residuals(.x)
X-squared = 24.38, df = 1, p-value = 7.907e-07
```

```
Análise de Resíduos - SAO PAULO

        Ljung-Box test

data:  Residuals from TBATS
Q* = 381.06, df = 48, p-value < 2.2e-16

Model df: 0.   Total lags used: 48

        Box-Ljung test

data:  residuals(.x)
X-squared = 0.066932, df = 1, p-value = 0.7959
```

Figura 9: Teste Ljug-box

A Figura 10 apresenta as previsões obtidas com o modelo STLM e permite a comparação com os dados reais. Embora o modelo pareça ter uma boa capacidade de previsão, as diferenças entre as séries indicam que a autocorrelação residual pode ser ainda reduzida, e ajustes no modelo podem ser realizados.

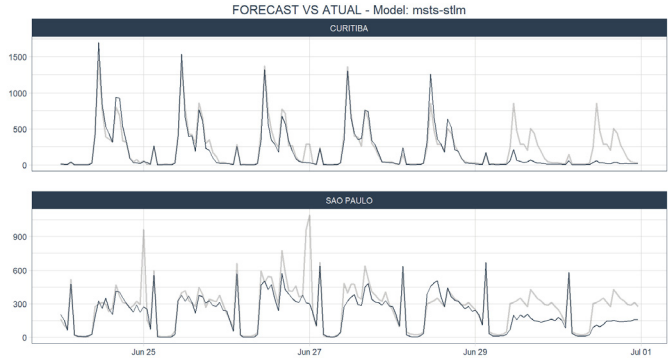

Figura 10: Comparação: Real vs previsto modelo msts-stlm

Por fim, as Figuras 11 e 12 apresentam as previsões para os próximos oito dias. A linha vermelha representa os valores reais, enquanto a linha azul indica as previsões do modelo. A segunda figura oferece um zoom na cauda da série temporal, facilitando a análise das previsões mais recentes. Estes resultados sugerem que, apesar do desempenho geral sólido, há oportunidades para refinar ainda mais as previsões, especialmente no que diz respeito à precisão e à redução da autocorrelação dos resíduos.

Em resumo, os modelos forneceram uma base sólida para fornecer previsões de clima em ambientes DaaS. No entanto, as comparações com os dados reais e previsões futuras indicam que ajustes específicos são possíveis, a fim de melhorar as previsões e reduzir a autocorrelação residual. Isso pode envolver a adoção de modelos mais avançados ou ajustes específicos para capturar a dinâmica total dos dados.

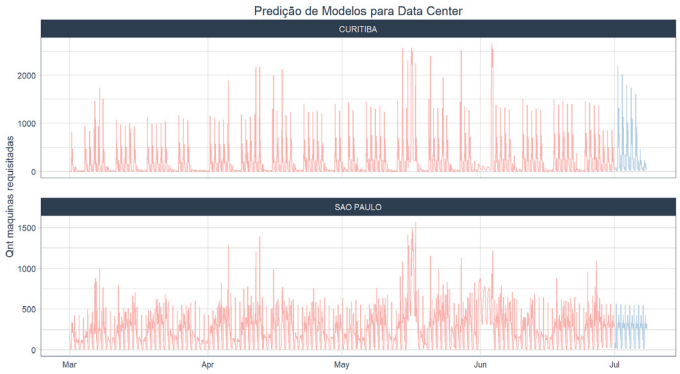

Figura 11: Modelo Escolhido

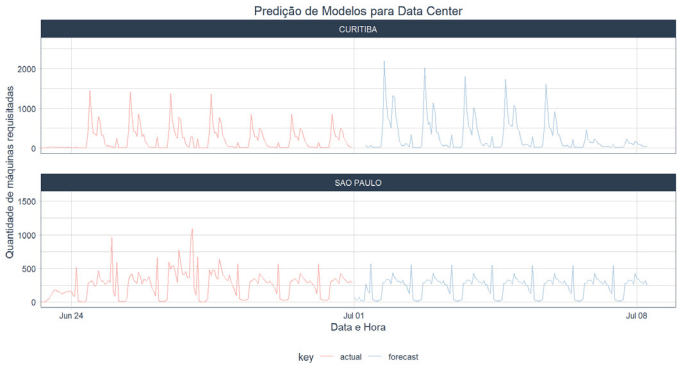

Figura 12: Modelo Escolhido - Zoom

5. Conclusão

Neste estudo, foram analisados os dados do uso de máquinas virtuais em um ambiente DaaS, que foram extraídos da API do Citrix para identificar padrões sazonais e prever a demanda futura. Usamos diferentes métodos da série do tempo, tal como STLM, ETS, auto.arima, e TBATS e avaliamos cada modelo usando MAE, AIC, BIC. Como resultado, o modelo de STLM pela série do tempo com sazonalidade múltipla foi o melhor em termos de erro de teste para o data center de Curitiba — isso representa saber mais preciso sobre o padrão sazonal e a tendência de longo prazo

também é bastante bom quando se trata do planejamento de recurso em data centers. No entanto, a análise dos resíduos e os testes de Ljung-Box mostraram que há uma autocorrelação significativa nos resíduos, especialmente em Curitiba. Embora o modelo seja o melhor, os sinais significativos de autocorrelação sugerem que a modelagem e as previsões podem ser mais bem ajustadas para resolver os padrões não modelados completamente. Em São Paulo, a autocorrelação é menos significativa, no entanto, as melhorias ainda são possíveis. Em geral, os modelos que foram utilizados neste estudo fornecem bases fortes para a previsão da demanda em um ambiente DaaS. No entanto, os resultados sugerem que há espaço para aprimoramento, tanto na escolha do modelo quanto no ajuste de parâmetros para reduzir a autocorrelação residual e melhorar as previsões. Além disso, estudos futuros podem aplicar algoritmos de aprendizagem de máquina mais avançado ou analisar um conjunto de dados diferente para validar e, possivelmente, aprimorar nossas descobertas.

6. Agradecimentos

Gostaria de expressar meus agradecimentos a todos os docentes do curso de especialização em Data Science e Big Data que contribuíram para a realização do meu Trabalho, em especial ao meu orientador Wagner Bonat.

Referências

[1] R. Buyya, K. Ramamohanarao, C. Leckie, R. N. Calheiros, A. V. Dastjerdi and S. Versteeg, "Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions,"2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), Melbourne, VIC, Australia, 2015, pp. 75-84, doi: 10.1109/ICPADS.2015.18. keywords: Cloud computing;Business;Big data;Virtual machining;Computational modeling;Real-time systems,

[2] XenServer, Citrix. "Citrix XenServer™."

[3] Gupta, Hari & Gupta, Ashish & Biswas, Bhaskar & Dutta, Tanima. (2019). An Early Classification Approach for Multivariate Time Series of On-Vehicle Sensors in Transportation. IEEE Transactions on Intelligent Transportation Systems. 10.1109/TITS.2019.2957325.

[4] Yermal, Lakshmi & Balasubramanian, Palanisamy. (2017). Application of Auto ARIMA Model for Forecasting Returns on Minute Wise Amalgamated Data in NSE. 1-5. 10.1109/ICCIC.2017.8524232.

[5] MUNIM, Ziaul Haque. State-space TBATS model for container freight rate forecasting with improved accuracy. Maritime Transport Research, v. 3, p. 100057, 2022.

[6] KASSAMBARA, Alboukadel. Machine learning essentials: Practical guide in R. Sthda, 2018.

[7] Hyndman, R. J., et al. *forecast*: Forecasting functions for time series and linear models. R package version 8.23.0. Available online at: `https://pkg.robjhyndman.com/forecast/`..

[8] Timetk Package Documentation. Retrieved from `https://cran.r-project.org/web/packages/timetk/readme/README.html`

[9] Padr Package Documentation. Retrieved from `https://cran.r-project.org/web/packages/padr/padr.pdf`

[10] Forecasting in R. Retrieved from `https://rpubs.com/odenipinedo/forecasting-in-R`

[11] Introduction to Purrr in R. Retrieved from `https://rpubs.com/alfado/purrr`

[12] Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science Business Media. ISBN: 9783540719182.

[13] Livera, A. M. D., Hyndman, R. J., & Snyder, R. D. (2011). *Forecasting time series with complex seasonal patterns using exponential smoothing*. Journal of the American Statistical Association, 106(496), 1513-1527. DOI: 10.1198/jasa.2011.tm09771.