Série dos Seminários de Acompanhamento à Pesquisa

DEI
DEPARTAMENTO DE ENGENHARIA INDUSTRIAL

Número 04 | 05 2021

Explorando métodos alternativos em previsões de séries temporais com bagging e clustering

Autor(es):

David Souza

PUC RIO

Série dos Seminários de Acompanhamento à Pesquisa

Número 04 | 05 2021

Explorando métodos alternativos em previsões de séries temporais com bagging e clustering

Autor(es):

David Souza

Orientador: Fernando Cyrino

CRÉDITOS:

SISTEMA MAXWELL / LAMBDA
https://www.maxwell.vrac.puc-rio.br/

Organizadores: Fernanda Baião / Soraida Aguilar

Layout da Capa: Aline Magalhães dos Santos

Agenda

1. Introdução
2. Método
3. Experimento
4. Conclusão

Resumo

Métodos de amortecimento exponencial são formulações versáteis para a previsão de séries temporais univariadas conhecidas desde a década de 1960. Modelos mais recentes tem feito uso do *bagging* para melhorar a qualidade das previsões com estes modelos. Um destes modelos, **BaggedETS**, desenvolvido em 2016, trouxe melhorias na qualidade de previsão e está disponível na biblioteca **forecast** para **R**. Uma proposta posterior, **BaggedClusterETS**, adicionou uma etapa de *clustering* e validação para tratar o efeito da covariância associada ao uso do *bagging*, resultando em ganhos adicionais de performance. Este artigo explora os efeitos do uso de quatro medidas de dissimilaridade ao gerar os *clusters*. Para testar as séries, 21 séries temporais da aviação civil e demanda energética foram empregadas.

Introdução

Contexto

- Séries temporais como uma coleção ordenadas de variáveis aleatórias, indexadas pelo tempo.
- Aplicações em economia, ciências sociais, medicina, e na indútria.
- Modelos de séries temporais como elemento-chave na tomada de decisão.
- Qualidade das previsões tem impacto nos custos operacional e de oportunidade do negócio.

.

Bagging

- Uso do *bagging* (ou *bootstrap aggregation*), reduz o erro de previsão ao agregar variações produzidas por *bootstrap* [Breiman, 1996].
- Para dados numéricos, a agregação é feita com a média ou mediana.
- Técnica com amplo uso em previsões de dados demográficos, séries financeiras, demanda de energia e aviação.

Modelos de amortecimento exponencial

- Desenvolvidos nas década de 1950 e 1960.
- Uso de médias ponderadas e hiper-parâmetros para o ajuste.
- O termo exponencial é referente ao decaimento dos pesos.
- Extensão com modelos de espaço de estado, permitindo intervalos de previsão: formulação ETS (***E**rror, **T**rend, **S**easonality*).

Combinando com o bagging

Boot.EXPOS proposto por Cordeiro e Neves [2009]: combina os modelos de amortecimento exponencial com o *sieve bootstrap* para realizar o *bagging*.

BaggedETS proposto por Bergmeir et al. [2016]: combina a transformação de Box-Cox, decomposição STL, *moving block bootstrap* e modelos ETS para realizar as previsões.

BaggedClusterETS proposto por Dantas e Oliveira [2018]: inclui etapas de validação, *clustering* e seleção para tratar o efeito da covariância que pode surgir ao empregar o *bagging*.

Motivação

Buscas nos agregadores científicos SCOPUS e *Web Of Science* não revelaram publicações que tenham trabalhado em extensões para o modelo de Dantas e Oliveira [2018].

Assim, há o interesse de explorar uma variação do modelo mencionado na construção dos *clusters* e o impacto resultante a qualidade das previsões.

Método

Métodos numéricos

Dada a maneira como as réplicas são geradas com o *bootstrap*, o comitê pode apresentar uma alta covariância, o que pode produzir efeitos no erro médio quadrático (*mean squared forecast error*, MSFE).

$$\mathrm{MSFE} = \mathrm{Var}(y_{t+1|t}) + \text{viés}(\hat{y}_{t+1|t})^2 + \mathrm{Var}(\hat{y}_{t+1|t}) \tag{1}$$

Métodos numéricos

$$\tilde{y}_{t+1|t} = \frac{1}{B}\sum_{i=1}^{B}\hat{y}^{*}_{(i)t+1|t} \tag{2}$$

$$\text{viés}(\tilde{y}_{t+1|t}) = \frac{1}{B}\sum_{i=1}^{B}\text{viés}(\hat{y}^{*}_{(i)t+1|t}) \tag{3}$$

$$\text{Var}(\tilde{y}_{t+1|t}) = \frac{1}{B}\sum_{i=1}^{B}\text{Var}(\hat{y}^{*}_{(i)t+1|t}) + \frac{1}{B^2}\sum_{i\neq i'}\text{Cov}[\hat{y}^{*}_{(i)t+1|t}, \hat{y}^{*}_{(i')t+1|t}] \tag{4}$$

Algoritmos

	BaggedETS	BaggedClusterETS
1		Cálculo do λ ótimo
2		Transformação de Box-Cox
3		Avaliação de sazonalidade para decomposição
4		STL (série sazonal); Loess (série não sazonal)
5		*Moving block bootstrap* aplicado ao resíduo
6		Construção de *B* réplicas
7		Reversão da transformação de Box-Cox
8	–	Validação e *ranking* com sMAPE
9	–	*Clustering*
10	–	Seleção das séries para cada *cluster*
11		Ajuste de modelos ETS
12		Previsão + Agregação

Tabela 1: Comparação entre **BaggedETS** e **BaggedClusterETS**

Clustering

Clustering é um método não-supervisionado de aprendizado de máquinas que busca identificar subconjuntos desconexos (*clusters*) em um conjunto de dados ou objetos, sem conhecimento prévio da composição do conjunto.

Para qualquer *cluster*, os elementos que o compõe apresentam alto grau de similaridade, sendo ao mesmo tempo altamente dissimilares quando comparados a elementos de outros *clusters* [Hennig e Meila, 2016; Aghabozorgi et al., 2015].

Clustering: tipos

Baseados em observações os dados brutos ou uma transformação adequada destes são usados para a criação dos *clusters*.

Baseados em atributos usa atributos extraídos dos dados, que podem ser extraídas dos domínios do tempo (correlação, autocorrelação, autocorrelação parcial), frequência (periodogramas, e ordenadas espectrais), ou podem ser baseados na decomposição por *wavelets*

Baseados em modelos assume que um conjunto de séries gerados por um mesmo modelo apresentam padrões semelhantes, e o agrupamento é realizado por meio de estimação de parâmetros ou ainda através dos resíduos de modelos ajustados.

Criando os clusters

Independente do tipo de agrupamento feito, é necessário que exista alguma formulação numérica que permita a mensuração da diferença entre os elementos. Com estas medidas é possível então construir uma matriz de dissimilaridade e assim agrupar os objetos [Montero e Vilar, 2014]. O **BaggedClusterETS** faz uso da distância euclidiana para avaliar a dissimilaridade entre duas séries. Esta matriz é então alimentada ao PAM para gerar os *clusters*. No entanto, considerando que o ruído presente nas séries pode causar interferência ao realizar agrupamentos com dados brutos [Caiado et al., 2016], é possível perguntar se diferentes medidas geram algum impacto na qualidade das previsões.

Medidas alternativas

Foram elencadas e testadas quatro medidas não-paramétricas, baseadas em atributos. São distâncias baseadas na correlação cruzada, transformada discreta de *wavelets*, e duas distâncias espectrais: uma baseada nos estimadores de mínimos quadrados do log-espectros das séries, e a outra baseada na razão da verossimilhança generalizada.

Assim, deseja-se verificar o impacto que estas quatro distâncias exercem na formaçãoo de clusters através do algoritmo PAM, e como consequência, na qualidade da previsão.

Avaliação

O algoritmo de clusterização original (PAM, *partitioning around medoids*) foi mantido. Para validar o número k ótimo de clusters, manteve-se o critério de silhueta. Para avaliar a qualidade das previsões foram usados o sMAPE e o MASE, como definidos nas equações 5 e 6.

Para este experimento, o cálculo do sMAPE foi feito em cima das previsões agregadas pela mediana.

$$\text{sMAPE} = \frac{200}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \tag{5}$$

$$\text{MASE}_S = \frac{y_t - \hat{y}_t}{\frac{1}{T-m} \sum_{t=m+1}^{T} |y_t - y_{t-m}|} \tag{6}$$

.

Experimento

Séries empregadas

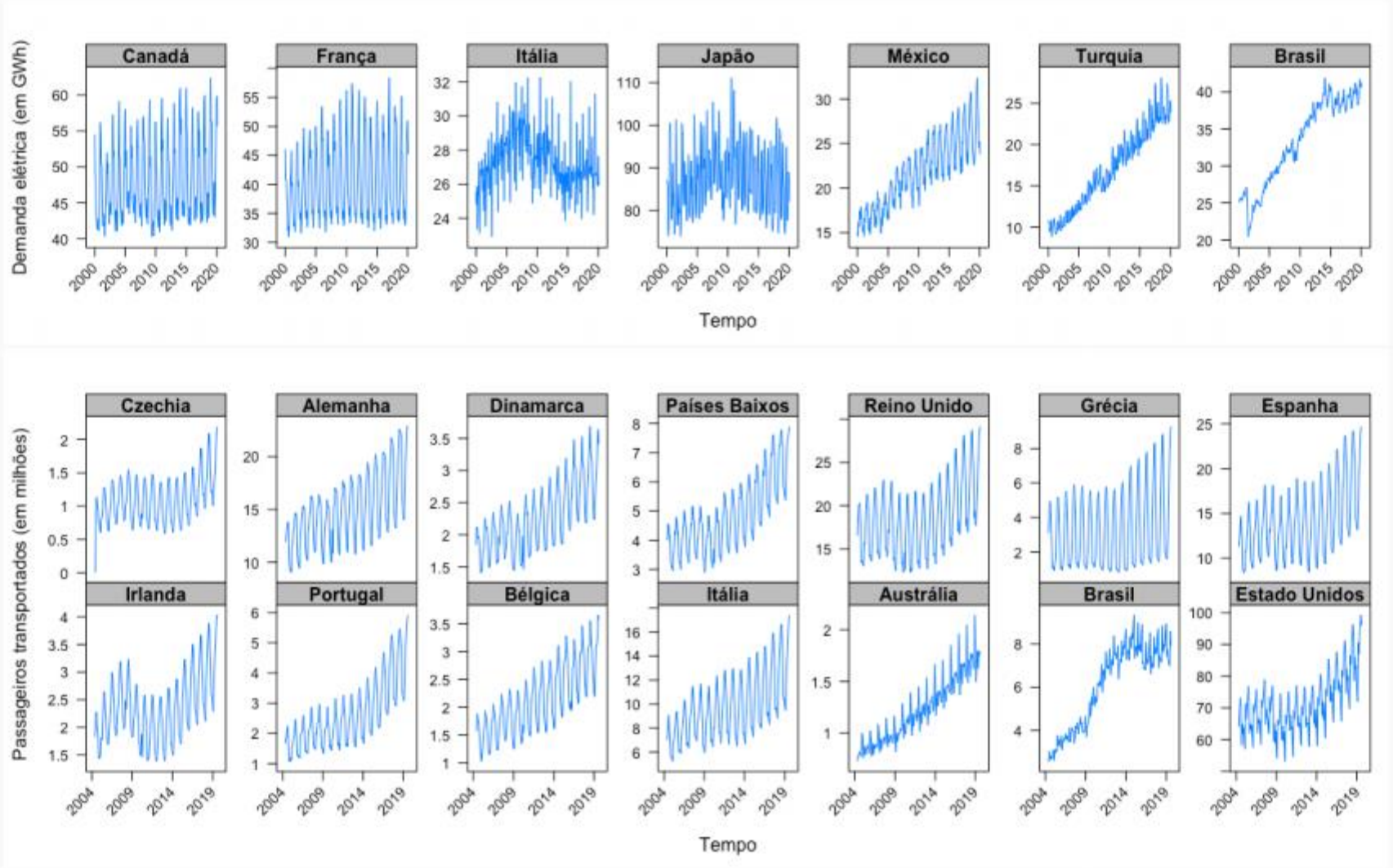

Figura 1: Séries de energia (topo) e de aviação (sopé) empregadas

Resultado: Energia

	sMAPE (%)					MASE				
	EUCL	COR	DWT	LLR	GLK	EUCL	COR	DWT	LLR	GLK
Canadá	0.948	0.949	**0.945**	0.945	0.950	**0.678**	0.680	0.678	0.678	0.679
França	**1.124**	1.124	1.136	1.133	1.140	0.529	0.529	0.525	**0.525**	0.529
Itália	1.501	1.499	**1.495**	1.513	1.499	1.074	1.072	**1.069**	1.079	1.070
Japão	1.289	1.280	1.289	**1.255**	1.280	0.613	0.611	0.616	**0.596**	0.614
México	3.309	3.309	3.303	3.292	**3.274**	2.239	2.239	2.239	2.232	**2.224**
Turquia	0.974	0.977	0.962	**0.958**	0.962	0.522	0.523	0.506	**0.505**	0.509
Brasil	**0.784**	0.787	0.786	0.784	0.810	0.440	**0.439**	0.439	0.444	0.450

Tabela 2: Tabela de erros: Séries de Energia

Resultado: Aviação

	sMAPE (%)					MASE				
	EUCL	COR	DWT	LLR	GLK	EUCL	COR	DWT	LLR	GLK
Rep. Checa	3.749	3.912	3.715	**3.650**	3.688	1.655	1.688	1.648	**1.437**	1.460
Alemanha	0.802	0.794	**0.792**	0.808	0.793	0.448	**0.433**	0.438	0.439	0.435
Dinamarca	1.340	1.352	1.316	1.305	**1.302**	0.666	0.651	**0.649**	0.655	0.659
Países Baixos	**1.522**	1.544	1.528	1.537	1.524	0.705	0.712	0.718	0.705	**0.701**
Reino Unido	0.862	**0.858**	0.888	0.863	0.867	0.497	0.489	0.491	0.484	**0.482**
Grécia	**1.306**	1.354	1.379	1.381	1.343	**0.561**	0.590	0.604	0.592	0.590
Espanha	**1.291**	1.304	1.347	1.329	1.343	**0.539**	0.540	0.549	0.544	0.546
Irlanda	1.199	1.147	1.176	**1.136**	1.143	0.404	**0.382**	0.390	0.383	0.390
Portugal	2.319	2.329	2.324	**2.312**	2.326	0.926	0.937	0.928	**0.923**	0.929
Bélgica	**1.264**	1.268	1.304	1.274	1.264	0.522	0.524	0.529	0.526	**0.521**
Itália	1.285	**1.195**	1.382	1.253	1.269	0.643	**0.604**	0.669	0.637	0.634
Austrália	1.183	1.185	**1.179**	1.187	1.179	**0.587**	0.590	0.589	0.589	0.589
Brasil	1.478	1.478	**1.417**	1.519	1.470	0.442	0.442	**0.430**	0.445	0.437
EUA	0.439	0.442	**0.425**	0.452	0.441	0.302	**0.298**	0.301	0.302	0.299

Tabela 3: Tabela de erros: Séries de Aviação

Discussão

Em uma comparação geral, considerando o sMAPE, os modelos com distâncias baseadas em atributos apresentaram um melhor desempenho para 15 das 21 séries. Ao considerar o MASE, o número de modelos com medidas baseadas em atributos vai de 15 para 17.

Avaliando ambas as métricas, há casos onde, embora a distância EUCL seja selecionada, uma ou mais medidas baseadas em atributos são bem próximas. Exemplos incluem, nas séries de energia, o sMAPE e o MASE para o Canadá, o sMAPE para o Brasil e o sMAPE da França; nas séries de aviação, tem-se o sMAPE da Bélgica.

Tempo de execução

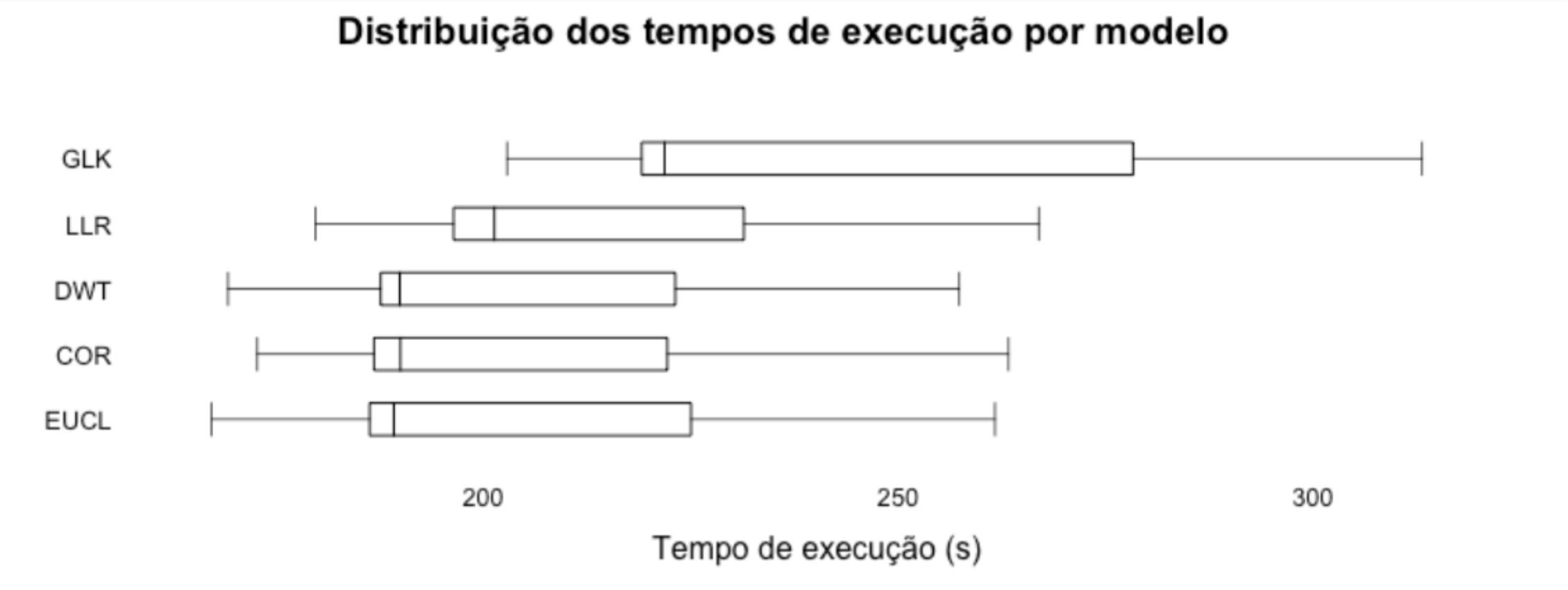

Figura 2: Distribuição dos tempos de execução por modelo

Conclusão

Conclusões

- Há ganhos de performance ao usar métricas baseadas em atributos. Mesmo que o ganho modesto, é necessário considerar a escala das aplicações.
- Mesmo com a troca de métricas, a escolha por uma medida baseada em atributos se mantém. É possível que estas ajudem a reduzir o efeito do ruído das séries ao calcular a dissimilaridade, mas não com força suficiente para agrupar réplicas geradas por *bootstrap*.
- Realizar mais testes com os modelos, usando um número maior de séries.

Referências

Aghabozorgi, S., Shirkhorshidi, A. S., e Wah, T. Y. (2015). Time-series clustering – a decade review. *Information Systems*, 53:16–38.

ANAC (2020). Relatório demanda e oferta do transporte aéreo. URL *https://www.anac.gov.br/assuntos/setor-regulado/empresas/envio-de-informacoes/relatorio-demanda-e-oferta-do-transporte-aereo-empresas-brasileiras*. Acesso em 02 jun. 2020.

Bergmeir, C., Hyndman, R. J., e Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and box–cox transformation. *International Journal of Forecasting*, 32(2):303–312.

BITRE (2020). International airline activity–time series. URL *https://www.bitre.gov.au/publications/ongoing/international_airline_activity-time_series*. Acesso em 02 jun. 2020.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw/Hill.

BTS (2020). Passengers, all carriers - all airports. URL *https://www.transtats.bts.gov/Data_Elements.aspx*. Acesso em 02 jun. 2020.

Caiado, J., Maharaj, E. A., e D'Urso, P. (2016). *Handbook of cluster analysis*, chapter 12. Time Series Clustering, p. 241–263. CRC Press.

Cordeiro, C. e Neves, M. M. (2009). Forecasting time series with BOOT.EXPOS procedure. *REVSTAT-Statistical Journal*, 7(2):135–149.

Dantas, T. M. e Oliveira, F. L. C. (2018). Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing. *International Journal of Forecasting*, 34(4):748–761.

Dantas, T. M., Oliveira, F. L. C., e Repolho, H. M. V. (2017). Air transportation demand forecast through bagging holt winters methods. *Journal of Air Transport Management*, 59:116–123.

de Oliveira, E. M. e Oliveira, F. L. C. (2018). Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. *Energy*, 144: 776–788.

Díaz, S. P. e Vilar, J. A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of classification*, 27(3):333–362.

Eurostat (2020). Transport database. URL *https://ec.europa.eu/eurostat/web/transport/data/database*. Acesso em 02 jun. 2020.

Hennig, C. e Meila, M. (2016). *Handbook of cluster analysis*, chapter 1. Cluster Analysis: An Overview, p. 1–20. CRC Press.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10.

Hyndman, R. J. e Athanasopoulos, G. (2019). *Forecasting: Principles and Practice*. OTexts, 3 edition.

Hyndman, R. J. e Khandakar, Y. (2008). Automatic time series for forecasting: the forecast package for r. *Journal of Statistical Software*, 27:1–22.

Liao, T. W. (2005). Clustering of time series data – a survey. *Pattern recognition*, 38(11): 1857–1874.

Maharaj, E. A., D'Urso, P., e Caiado, J. (2019). *Time series clustering and classification*. Chapman and Hall/CRC. ISBN (eBook): 978-0-429-05826-4.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529.

Montero, P. e Vilar, J. A. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43. URL ***http://www.jstatsoft.org/v62/i01/***.

Shumway, R. H. e Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer, 4 edition.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342.

Zhang, Y., Qu, H., Wang, W., e Zhao, J. (2020). A novel fuzzy time series forecasting model based on multiple linear regression and time series clustering. *Mathematical Problems in Engineering*, 2020.