Forecasting of Espírito Santo State ICMS Revenue through Cascading Feature Selection and Machine Learning Techniques

Previsão de receitas de ICMS do estado do Espírito Santo através de Seleção de Características em Cascata e técnicas de Aprendizado de Máquina

Marcelo Magalhães do Carmo[1], **Francisco de Assis Boldt**[2], **Karin Satie Komati**[2]

[1] Programa de Pós-graduação em Ciência de Dados com Big Data

[2]Programa de Pós-graduação em Computação Aplicada (PPComp)
Campus Serra do Instituto Federal do Espírito Santo (Ifes)
Serra – ES – Brasil

marcelo.mmcarmo@gmail.com, {franciscoa,kkomati}@ifes.edu.br

Abstract. *Two methods to forecast the ICMS tax (Imposto sobre Circulação de Mercadorias e Serviços) income in the state of Espírito Santo (ES) are proposed. The first uses an artificial neural network with a cascade feature selection, and the second uses a combination of statistical methods. The proposed machine learning method outperformed the univariate neural network with a sliding window six months ahead for epochs in 2018. The combined statistical methods outperformed all the others in the 120 epochs tested.*

Resumo. *Dois métodos para prever a arrecadação do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) no estado do Espírito Santo (ES) são propostos. O primeiro usa uma rede neural artificial com uma seleção de características em cascata, e o segundo usa uma combinação de métodos estatísticos. O método de aprendizagem de máquina proposto superou a rede neural univariada de referencia, com previsões em uma janela deslizante de seis meses a frente para épocas de teste em 2018. O combinado de métodos estatísticos superou na média todos os demais nas 120 épocas testadas.*

1. Introdução

Problemas de predição em séries temporais são desafiadores e tem sido estudados pela comunidade acadêmica. Revisitar técnicas de predição e adaptá-las a problemas concretos é sempre uma abordagem de interesse de pesquisa [Castanho 2011, de Azevedo et al. 2017, Makridakis et al. 2018b, Gonçalves and Rosa 2018, Domingos Jr et al. 2018]. Além disso, existem as *surpresas* que, apesar de serem fortemente estudadas pelo mercado e por pesquisadores [Caruso 2019], são dificilmente previsíveis pela sua natureza. Eventos de caráter caótico ou de padrão extremamente complexo dificilmente serão previstos pelas técnicas tradicionais, mas certamente dão informações no presente que afetarão as arrecadações futura de impostos, e por consequência podem apoiar os modelos de predição em

antecipar a ciência dos seus efeitos. O ambiente ao redor das variáveis econômicas não é estacionário ao longo de grandes períodos de tempo.

Este trabalho propõe uma aplicação na área de gestão pública, mais especificamente no estudo da arrecadação do ICMS do estado do Espírito Santo (ICMS-ES). A última revisão dos métodos e modelos de previsão do ICMS-ES publicada pelo Instituto Jones dos Santos Neves (IJSN) data de 2010 [Ribeiro 2010], e a pesquisa acadêmica mais similar à esta proposta e mais recente é datada de 2011 [Castanho 2011]. Os estudos apontam que o total de importações deve corresponder em torno de 30% da arrecadação direta do ICMS do estado, e o restante provem das atividades econômica industrial, comercial e serviços [Castanho 2011]. A necessidade de melhorar a previsão de receita nas contas públicas é contínua, e ainda mais relevante no momento atual de volatilidade econômica. Assim, em continuidade e extensão aos trabalhos de Castanho (2011) e do IJSN (2010), este tem por objetivo explorar variáveis econômicas relacionadas ao ES, como exportações e importações, e realizar comparações contra modelos estatísticos de referência [Hyndman and Athanasopoulos 2018], de forma a aprimorar a qualidade da previsão de arrecadação do ICMS-ES.

A hipótese deste trabalho foi que, pelo fato do ES ter grandes empresas exportadoras e atividades portuárias que movimentam sua economia, os dados de exportação e importação do MDIC (Ministério da economia, indústria, comércio exterior e serviços) tem representatividade para apoiar na explicação, direta ou indiretamente, do aumento ou diminuição da arrecadação do ICMS-ES, com defasagem no tempo. Uma das dificuldades está em saber quais as melhores variáveis, dentre centenas, que explicam a variabilidade. Para tal, será adotada a abordagem utilizada em seleção de variáveis explicativas baseada ao artigo de Assis Boldt et al. (2017) adaptada para regressão.

O trabalho está organizado da seguinte forma: a seção 2 apresenta os métodos escolhidos de aprendizado de máquina e as métricas de comparação de resultados; a seção 3 apresenta a análise exploratória e técnica de seleção em cascata de variáveis explicativas; a seção 4 apresenta os experimentos realizados e os resultados obtidos; a seção 5 apresenta as conclusões e considerações finais deste trabalho.

2. Referencial teórico

Esta seção apresenta as técnicas de aprendizado de máquina usadas e as métricas de comparação de resultados. Competições internacionais de referência em previsão para séries temporais foram consultadas para escolha destes métodos e métricas.

2.1. Métodos

Os modelos escolhidos neste trabalho foram obtidos de algoritmos clássicos, utilizados nas competições M3 [Makridakis and Hibon 2000], M4 [Makridakis et al. 2018a] e literatura científica [Hyndman and Athanasopoulos 2018]. Estes modelos possuem código aberto e público, podendo ser utilizados tanto para reproduzir o resultado das competições citadas quanto para comparações de novos modelos com outros conjuntos de séries temporais a serem previstas. Os métodos clássicos de previsão escolhidos para comparações foram o ingênuo com ajuste de sazonalidade (**naive2**), ingênuo com ajustes de sazonalidade e adição da tendência linear (**snaivedrift**), Holt clássico (**holt**) [Hyndman and Athanasopoulos 2018, Makridakis et al. 2018a], Theta clássico

[Assimakopoulos and Nikolopoulos 2000] (**ThetaClassic**) e método Box–Jenkins *Auto-Regressive Integrated Moving Average* [Hyndman and Athanasopoulos 2018] (**arima**). Além destes, um *multilayer perceptron* (**MLP**) com uma camada escondida foi incluso, seguindo código de referência da competição M4 [Makridakis et al. 2018a]. A mesma técnica de decomposição clássica multiplicativa ($y_t = T_t \times S_t \times R_t$) [Hyndman and Athanasopoulos 2018] foi utilizada para obtenção e extração da componente de sazonalidade em todos os métodos.

Dois novos modelos foram criados, um sendo uma combinação dos métodos clássicos **Theta**, **Holt**, **ARIMA**, aqui denominado por **comb_wavg**, e o outro é uma variação de MLP, aqui denominado por **mlp_cfs**. O método **comb_wavg** é apresentado no Algoritmo 1 e está sendo proposto a partir de uma média ponderada dos modelos clássicos. Os pesos foram escolhidos por análises empíricas em várias rodadas de experimentos. O mesmo método de remoção da sazonalidade utilizada nos modelos clássicos foi replicada neste modelo combinado. A definição dos parâmetros ARMA(p,q) do modelo ARIMA foi obtida automaticamente por procedimento *stepwise* na minimização do Critério de Informação de Akaike (AIC), conforme proposto no artigo [Hyndman et al. 2007], e a definição do parâmetro I(d) foi obtida por teste de estacionariedade de Dickey-Fuller.

Algoritmo 1: comb_wavg - Modelo combinado Theta, Holt e ARIMA

```
R_comb_wavg <- function(des_input,SIout, fh){
  # Codigo R para media ponderada de metodos classicos
  # des_input: ts removida sazonalidade
  # SIout: coponente multiplicativa da sazonalidade
  # fh : forecast horizon
  f1 <- forecast(auto.arima(des_input, max.p = 2*fh, max.q = 2*fh,
                            start.p = 1, start.q = 1,d = 1),
                            h=6)$mean*SIout # ARIMA
  f2 <- holt(des_input, h=fh, damped=F)$mean*SIout # Holt
  f3 <- Theta.classic(input=des_input, fh=fh)$mean*SIout # Theta
  f4 <- (0.5*f3+0.3*f2+0.2*f1)# Weighted average
  return(f4) }
```

O método **mlp_cfs** é apresentado no Algoritmo 2, sendo este um MLP com duas camadas escondidas, utilizando como entrada as variáveis explicativas escolhidas no processo seleção em cascata. A segunda camada foi inclusa para melhor capturar as relações não lineares complexas entre as variáveis explicativas. Por padrão, o modelo irá prever $\hat{y}_{t+1}$ e realizar as próximas previsões por iterações, conforme demostrada maior eficiência e menor complexidade em modelos MLP no artigo [Makridakis et al. 2018b]. Todas as camadas do modelo proposto são densamente conectadas, com padrão de (n, 2n, n/2, 1) onde n é o número de variáveis de entrada no modelo. Para redução de *overfiting* nos treinos, uma camada reguladora de *dropout* [Srivastava et al. 2014] foi inserida antes da camada de saída do modelo com taxa de 0.2, escolhida por experimentos. O método otimizador utilizado foi o Adam [Kingma and Ba 2014] com parâmetro de taxa de aprendizado de 0.001. A validação cruzada utilizou a razão 90%/10% para treino/validação, sem aleatoriedade e sempre validando contra os dados mais recentes na série.

Algoritmo 2: mlp_cfs - Modelo MLP proposto

```
def create_model_mlp_cfs(n=100,odim=1):
    # n_steps: numero de variaveis escolhidas pela selecao em cascata
    # odim: horizonte de previsao. por padrao a saida sera y_hat(t+1)
    model = Sequential()
    model.add(Dense(n, activation='linear', input_dim=n, name='Input'))
    model.add(Dense(n*2, activation='relu', name='NonLinearCorr_Dense1'))
    model.add(Dense(n//2, activation='relu', name='DimReduction_Dense2'))
    model.add(Dropout(rate=0.2))
    model.add(Dense(odim, activation='linear', name='forecast_output'))
    model.compile(optimizer='Adam', loss='mean_squared_error',
                  metrics = ['mean_squared_error'])
    return model
```

2.2. Métricas

As métricas utilizadas para comparação dos resultados seguiram equações e códigos similares aos da competição M4, conforme explicação detalhada no artigo [Makridakis et al. 2018a].

O erro médio absoluto percentual simétrico (do inglês *symmetric Mean Absolute Percentage Error* - sMAPE) foi a métrica escolhida para *benchmarks*. Na equação 1 do sMAPE, y_t é o valor real observado no período t, e $\hat{y}_t$ é o valor previsto pelo modelo no mesmo período. O erro médio absoluto escalado (do inglês *Mean Absolute Scaled Error* - MASE) foi a segunda métrica escolhida de *benchmark*, sendo comumente utilizada para medir desempenho de modelos contra séries de escalas diferentes. Neste trabalho, a escala do ICMS-ES entre as épocas é semelhante, mas não exatamente igual se olharmos um horizonte de 10 anos (120 épocas) de previsão para próximos 6 meses. Na equação 2 do MASE, y_t é o valor real observado, $\hat{y}_t$ é o valor previsto pelo modelo e f é a sazonalidade. O MASE, por definição, também é uma comparação direta com o método "ingênuo" (*naive*), onde a previsão para $\hat{y}_{t+1} = y_t$.

As demais métricas utilizadas de desempenho foram a raiz quadrática do erro médio (do inglês root mean square error - RMSE) e o erro percentual médio absoluto (do inglês *Mean Absolute Percentage Error* - MAPE). O RMSE (equação 3) dá a ideia da escala de grandeza do erro absoluto em R$, e o MAPE (equação 4) é utilizado em vários trabalhos similares da área de ciências econômicas, mas não utilizado nas competições M3 e M4 pelos problemas de singularidade quando y_t tende a zero. Pressupõe-se que na previsão da série ICMS-ES não haverá este problema de singularidade.

(1) $$sMAPE = \left(\sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \right) \frac{100}{n}$$

(2) $$MASE = \sum_{t=1}^{n} \left(\frac{|y_t - \hat{y}_t|}{\frac{1}{n-f} \sum_{t=f+1}^{n} |y_t - y_{t-1}|} \right) \frac{1}{n}$$

(3) $$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$$

(4) $$MAPE = \frac{100}{n} \left(\sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right)$$

3. Análise exploratória e seleção de variáveis

Análise exploratória é uma etapa importante em qualquer projeto de modelagem. Para este trabalho a análise foi dividida em 2 partes, sendo a primeira para série histórica do ICMS-ES e a segunda para demais séries explicativas de importação e exportação utilizando ferramentas de aprendizado de máquina como apoio.

3.1. Fonte de dados

Para este trabalho, foram utilizados somente dados públicos disponíveis nas fontes oficiais de divulgação do governo:

- SEFAZ-ES[1]: 17 anos de séries mensais de ICMS consolidado e arrecadado, em reais (R$), para o estado do Espírito Santo.
- MDIC[2]: 20 anos de séries mensais de exportação e importação em dólares, decompostas por **357** categorias de Pauta de Produtos Importados (PPI) e **454** categorias de Pauta de Produtos Exportados (PPE), segundo metodologia de categorização do MDIC. Exemplo de categorias PPI e PPE: ***'PPI_TECIDOS DE MALHA'*** e ***'PPE_PRODUTOS LAMINADOS PLANOS DE FERRO OU ACOS'***.

Para a tarefa de captura, extração, transformação e modelagem dos dados, foram usadas as ferramentas: *Microsoft Power Query/Power BI* e *Python/Jupyter Notebook*, que auxiliaram na visualização dos dados para extração em *datasets* simplificados.

A base de dados completa do MDIC é particularmente grande, com 4,58GB de informações. Neste caso, a estratégia foi filtrar os dados contextualizados ao estado do ES e exportar as séries temporais sob o domínio de interesse. Os filtros utilizados foram dados de importação e exportação agrupados por PPI e PPE relativos ao estado do Espírito Santo, dados de exportação relativos a outros estados mas que foram escoados por Unidades da Receita Federal (URF) situada no Espírito Santo e valores entre as datas de 2000 e 2018.

3.2. Análise da série do ICMS-ES

A série do ICMS-ES possui valores de janeiro de 2002 até dezembro de 2018, em períodos mensais com 204 amostras. Pelo método de Dickey-Fuller[3] (p-value 0.497) temos indicativo de não estacionariedade. Esta não estacionariedade é comum em séries econômicas e precisa ser tratada antes da entrada nos métodos de previsão [Hyndman and Athanasopoulos 2018].

O primeiro gráfico da Figura 1 apresenta a série temporal do ICMS-ES. Decompondo a série pelo método multiplicativo[4] ($y_t = T_t \times S_t \times R_t$) avalia-se que há uma tendência (T_t) apresentada na Figura 1(b), uma sazonalidade (S_t visível em 12 meses) na Figura 1(c) e um resíduo (R_t) na Figura 1(d). Tanto a série residual 1(d) quanto a diferença mês a mês da série (Figura 2) apresentam indicativos de estacionariedade ($p-value < 0.05$).

Estudos anteriores sugerem uma correlação do ICMS-ES com a série defasada no tempo para o câmbio [Braatz and da Rocha Gonçalves 2018], proveniente da dependência econômica do estado nas atividades relacionadas a exportações e importações (sistemafindes.org.br).

3.3. Análise da série do MDIC - variação mensal de exportação e importação

Para as séries históricas dos grupos de produto de exportação e importação, existe uma distribuição próxima de exponencial, entre a variação mensal dos principais grupos de

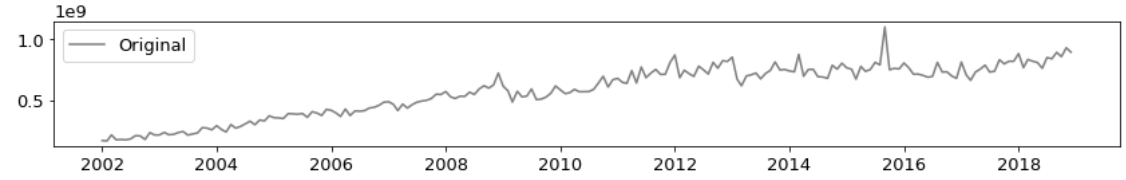
[1] internet.sefaz.es.gov.br

[2] www.mdic.gov.br

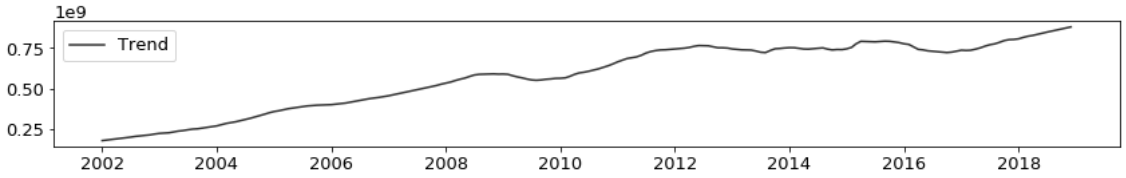
[3] statsmodels.tsa.stattools.adfuller

[4] statsmodels.tsa.seasonal.seasonal_decompose

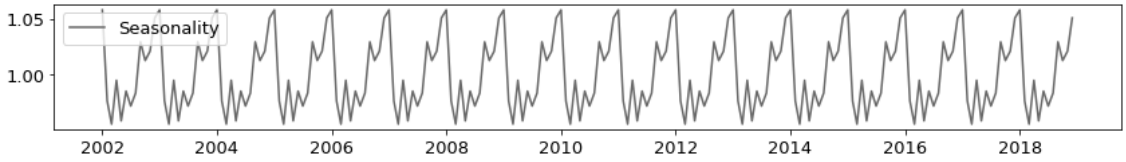
(a) Série mensal do ICMS - Original



(b) Decomposição do ICMS - Tendência



(c) Decomposição do ICMS - Sazonalidade



(d) Decomposição do ICMS - Resíduo

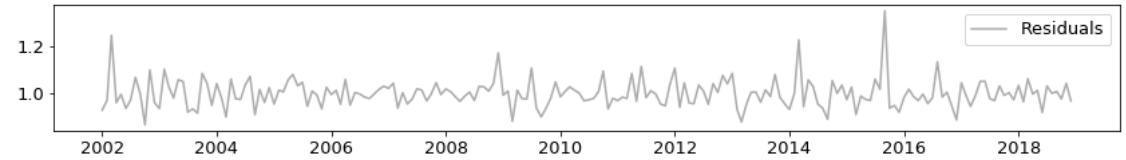


Figura 1. Decomposição ICMS-ES - tendência(b), sazonalidade(c) e resíduo(d)

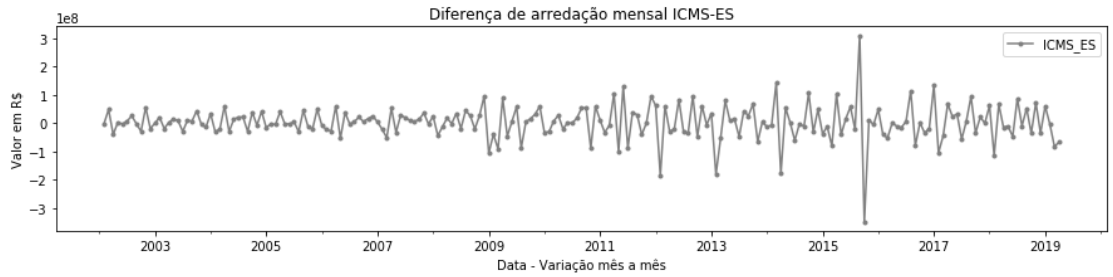


Figura 2. Variação mensal - ICMS-ES

produtos. Aproximadamente 20% dos itens PPI e PPE correspondem a aproximados 80% dos valores de importações e exportações, conforme experimentos preliminares, tendo maior destaque os itens de *commodities*. Não foi escopo deste trabalho avaliar a causalidade econômica destas variáveis, mas sim o seu potencial de auxílio na previsão do ICMS-ES.

Analisando por experimentos as séries decompostas em PPI e PPE, percebe-se a existência de itens com maior correlação (tanto positiva quanto negativa) do que as séries sumarizadas (soma dos valores PPI e PPE). Por estes motivos e pela provável não linearidade da correlação, uma técnica de seleção em cascata dará apoio na escolha dos

prováveis melhores conjuntos, incluindo as regressões das séries explicativas em até 3 períodos no passado (do inglês *lags*).

3.4. Escolha das variáveis de importação e exportação

A proposição do processo de seleção progressiva em cascata, onde cada passo de seleção simples é seguido de um passo mais complexo computacional foi adaptada de Assis Boldt et al. (2017), e se mostrou um eficiente método para o problema apresentado neste trabalho. O fluxo proposto está descrito na Figura 3, sendo as caixas brancas contribuições de novas variáveis, as caixas cinza os passos de eliminações e as caixas centrais os acumuladores dos descartes, onde a escala de cinza simboliza o aumento da complexidade na escolha. As variáveis descartadas são armazenadas e acumuladas em uma soma para serem utilizadas na avaliação final.

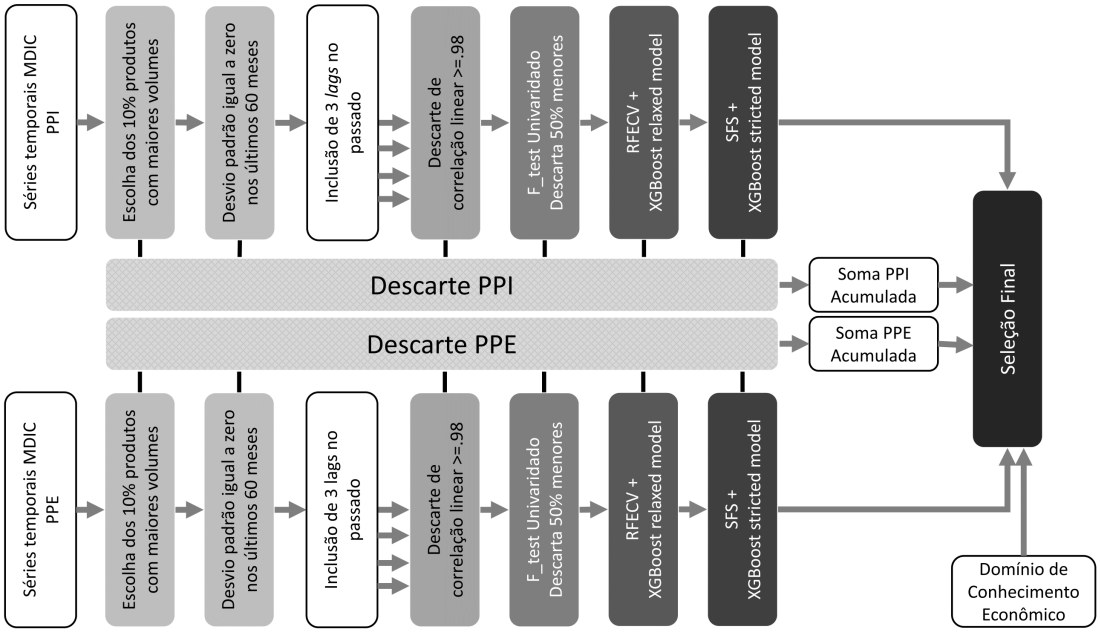


Figura 3. Fluxo de seleção em cascata de variáveis explicativas

Os primeiros passos do fluxo precisam ser executados por métodos simples e pouco restritivos, como a eliminação das variáveis menos representativas em volumes e variáveis monotônicas (desvio padrão zero) nos ultimos 60 meses. Para os demais passos, um cálculo de coeficiente de correlação de *Pearson* com altíssimo valor de corte (> 0.98 para eliminação de redundâncias) é executado e também cálculos estatísticos simples univariados, com teste F para regressão contra a diferença mês-a-mês da série do ICMS-ES. Os métodos foram conservadores, com intuito de eliminar somente as variáveis certamente menos relevantes.

As fases subsequentes de seleção são de avaliação computacional dos resultados por validação cruzada. A primeira é a eliminação recursiva de características por validação cruzada, utilizando o pacote do *scikit-learn RFECV* [Pedregosa et al. 2011]. A segunda, e computacionalmente mais complexa, é a seleção progressiva sequencial (do inglês *Sequential Forward Selection* - SFS) de características [de Assis Boldt et al. 2017],

utilizando o pacote *mlxtend SFS*[5] [Raschka 2018].

O modelo de regressão não linear utilizado em ambas foi o *Gradient Boost Regression Tree* (GBRT), conforme referência em problema similar de seleção de características em séries temporais [Sala et al. 2018]. A técnica *XGBoost*[6] [Chen and Guestrin 2016] foi utilizada por ser bastante flexível em parâmetros, eficiente em desempenho, explicativa na relevância das variáveis (pré-requisito do RFECV) e por suportar tanto paralelismo quanto rotinas otimizadas para GPU. O modelo *XGBRegressor* foi utilizado tanto no RFECV quanto no SFS, mas de forma relaxada na primeira rodada do RFECV e um pouco mais restrito na segunda rodada do SFS.

Para relaxamento, os parâmetros *alpha* e *subsample* do método *XGBRegressor* foram alterados de forma a permitir a admissão de mais variáveis pouco influentes, eliminando recursivamente somente aquelas que realmente tiveram a pior contribuição no modelo (alpha=0.1, subsample=0.8). Para a rodada do SFS os parâmetros foram mais restritivos (alpha=0.01, subsample=0.9), procurando as melhores combinações possíveis de variáveis por seleção sequencial. Estes parâmetros foram escolhidos por experimentos. Em ambos os passos de RFECV e SFS, a validação cruzada utilizada foi adequada para séries temporais.

Todas as variáveis descartadas voltam consolidadas novamente em uma soma para seleção final. Este procedimento garante que nenhuma informação é perdida no processo, levando separadamente as variáveis explicativas que possuem maior potencial de apoio na previsão de variações do ICMS-ES por um método não linear. Após todo o processo de seleção, uma avaliação manual deve ser realizada por especialistas tanto nas séries ICMS-ES quanto nas séries selecionadas de PPI e PPE. Também são incluídas as variáveis de auto-regressão do ICMS-ES para o modelo final, assim como quaisquer outras variáveis explicativas que um economista julgar necessário.

Neste trabalho a seleção em cascata foi executada considerando séries até ano de 2017, deixando os períodos de 2018 para validação final. Somente foram inclusas, ao final da seleção automática, as variáveis de auto-regressão do ICMS-ES (12 *lags* pela sazonalidade identificada na análise exploratória da série). Por fim, foram escolhidas 41 variáveis para utilização no modelo MLP *mlp_cfs*.

4. *Experimentos e Resultados*

Todos os modelos foram executados em um horizonte de previsão de seis meses a frente, calculando a média dos resultados. Os comparativos foram executados contra um conjunto de 120 épocas simuladas do ICMS-ES em janelas deslizantes iniciando em 2008-06 até 2018-06, conforme Figura 4. Para treino/validação foi utilizada a série de y_1 até y_t, onde t é o fim do período da janela deslizante e y_{t+1} o início dos testes para previsão dos próximos seis meses ($\hat{y}_{t+1} \ldots \hat{y}_{t+6}$).

Rotinas R e Python foram chamadas simultaneamente utilizando a biblioteca *rpy2* [Gautier 2018]. Os códigos das técnicas de aprendizado de máquina foram baseados na publicação do github[7] da competição M4, sendo adap-

[5] rasbt.github.io/mlxtend

[6] xgboost.readthedocs.io

[7] github.com/M4Competition

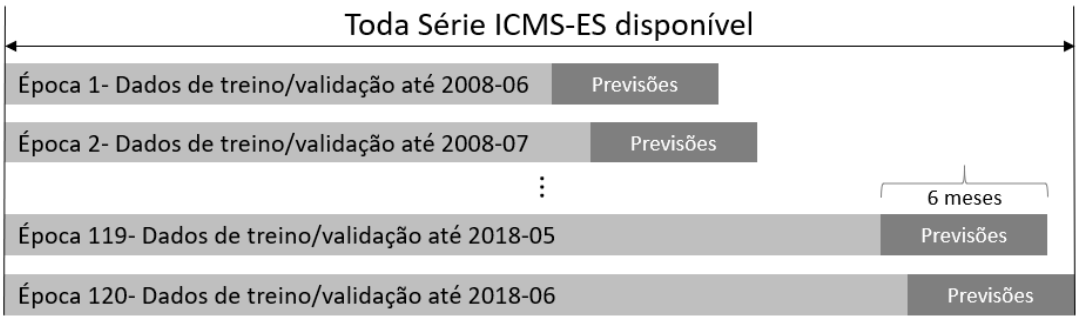


Figura 4. Janela deslizante para simulação das épocas de previsão no tempo

tados e atualizados para uso. Todos os modelos inclusos utilizaram as bibliotecas R *forecast* [Hyndman and Athanasopoulos 2018], bibliotecas Python *scikit-learn* [Pedregosa et al. 2011] e pacote *Keras/TensorFlow* [Chollet et al. 2015]

Todos os experimentos foram executados utilizando sempre o mesmo computador, com a configuração CPU i5 4210U, 8GB RAM, GPU NVidia GT740M 2GB RAM. A GPU foi utilizada sempre que disponível suporte a CUDA[8]. O tempo total de execução dos experimentos com todos os 8 modelos escolhidos para todas as 120 épocas foi de 1.843,55 segundos.

A Tabela 1 apresenta a visão geral dos resultados obtidos, ordenada por **sMAPE**, e em negrito estão os melhores valores, sendo separado na Figura 5 os resultados sMAPE para todos os modelos ao longo das épocas testadas entre 2015 e 2018. O desvio padrão do sMAPE (σ sMAPE) está presente na Tabela 1 para análise da variação do erro.

Para comparar os resultados da seleção em cascata, o modelo *mlp_bench* univariado da competição M4 e o modelo desenvolvido neste trabalho (*mlp_cfs*) incluindo as variáveis explicativas PPI e PPE foram selecionados. O modelo *mlp_bench* possui arquitetura de camadas (12, 6, 1) com 241 parâmetros para treino, enquanto o modelo *mlp_cfs* possui arquitetura (41, 82, 20, 1) com 6.847 parâmetros para treino. Os resultados estão dispostos na Tabela 2 e Figura 6.

Tabela 1. Média dos resultados das 120 épocas de previsão para os próximos 6 meses (entre 2008-07 até 2018-06)

Modelos	***sMAPE***	*σ sMAPE*	*MASE*	*RMSE*	*MAPE*	*Tempo comp.(s)*
comb_wavg	**6,832**	3,62	**0,752**	**5,98E+07**	6,93	1,849
ThetaClassic	6,865	**3,41**	0,754	6,01E+07	**6,91**	0,012
holt	6,907	3,79	0,760	6,03E+07	7,03	0,011
arima	7,119	3,98	0,787	6,19E+07	7,32	1,836
snaivedrift	7,745	5,24	0,863	6,68E+07	7,95	0,012
naive2	7,875	4,76	0,867	6,76E+07	7,91	0,008
mlp_cfs	9,492	3,70	1,034	8,13E+07	9,43	7,457
mlp_bench	9,717	3,83	1,086	8,64E+07	10,00	0,058

[8]developer.nvidia.com/cuda-toolkit

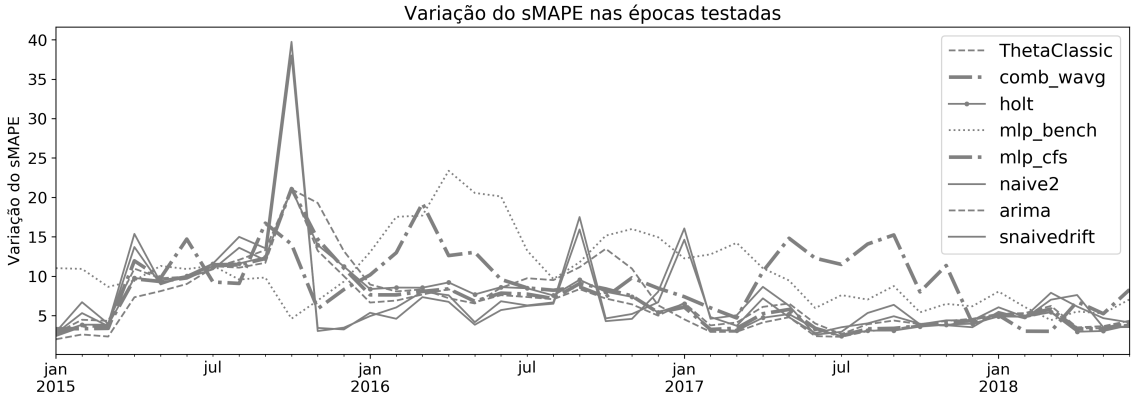


Figura 5. Variação do sMAPE nas épocas previstas entre 2015 até 2018

Tabela 2. Comparativo somente dos modelos MLP - Média dos resultados de previsão dentro do ano de 2018

Modelos	***sMAPE***	*σ sMAPE*	*MASE*	*RMSE*	*MAPE*	*Tempo comp.(s)*
mlp_cfs	**5,231**	**1,31**	**0,634**	**5,38E+07**	**5,15**	10,584
mlp_bench	6,076	2,08	0,750	6,86E+07	6,26	0,067

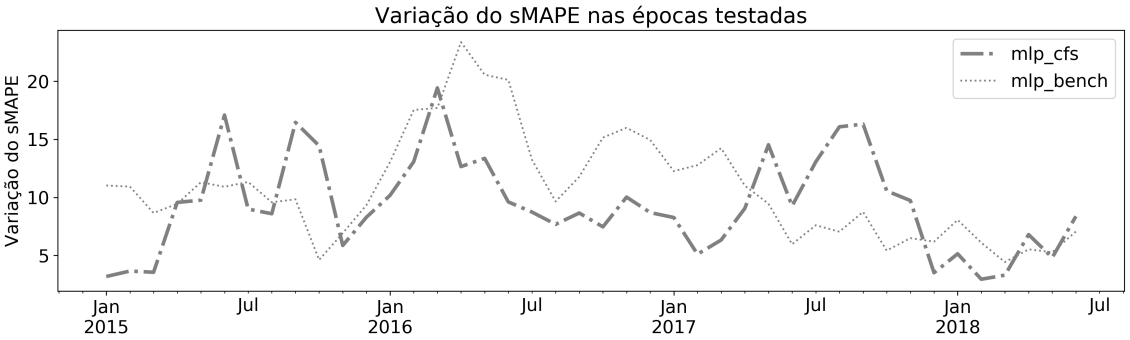


Figura 6. Comparativos entre modelos MLP - previsões entre 2015 até 2018

5. Conclusões e considerações finais

Este artigo apresenta uma proposta para previsão da série temporal do ICMS-ES e outra de seleção em cascata das principais séries explicativas de exportação e importação agrupadas por PPI e PPE, com finalidade de simplificar e automatizar o complexo trabalho de escolha e revisão dos modelos de previsão do ICMS-ES. Os testes sugerem que os resultados de *benchmark* nas séries sintéticas utilizadas nas competições M3 e M4 [Makridakis et al. 2018b, Makridakis et al. 2018a] são similares aos resultados gerados nas séries do ICMS-ES ao longo de 120 épocas (entre os anos de 2008 a 2018) para previsão dos próximos 6 meses de arrecadação. O método estatístico combinado proposto neste trabalho foi melhor em 0,5% na média do sMAPE em 120 épocas, apesar do maior desvio padrão comparado ao melhor método clássicos. Vemos ao longo do tempo que os melhores métodos variam por épocas, sugerindo revisões periódicas em momento de alta volatilidade econômica do estado e favorecendo, na média, modelos combinados.

Quanto a seleção de séries explicativas em cascata, o método proposto ultrapassou os resultados sMAPE médio do modelo MLP sugerido pela competição M4 contra o ICMS-ES em 2,4% ao longo das 120 épocas, e em 16,2% do sMAPE no período de previsões dentro de 2018. Estes resultados sugerem que as técnicas utilizadas de seleção em cascata de variáveis explicativas foram eficazes em aperfeiçoar a previsão da dinâmica econômica de um modelo univariado de referência. Trabalhos futuros podem explorar arquiteturas de redes neurais recorrentes, como RNN (do inglês Recurrent Neural Network) ou LSTM (do inglês Long Short-Term Memory), comparando-as aos resultados obtidos neste trabalho. Também podem se beneficiar deste trabalho quaisquer processos de previsão em outras séries econômicas correlacionadas as exportações e importações com poucas adaptações, como ICMS de outros estados.

Referências

[Assimakopoulos and Nikolopoulos 2000] Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4):521–530.

[Braatz and da Rocha Gonçalves 2018] Braatz, J. and da Rocha Gonçalves, R. (2018). Impactos regionais assimétricos da política cambial sobre a arrecadação do icms no brasil: uma abordagem com o método var. *Revista Estudo & Debate*, 25(3).

[Caruso 2019] Caruso, A. (2019). Macroeconomic news and market reaction: Surprise indexes meet nowcasting. *International Journal of Forecasting*.

[Castanho 2011] Castanho, B. J. d. S. (2011). Modelos para previsão de receitas tributárias: o icms do estado do espírito santo. Master's thesis, Universidade Federal do Espírito Santo.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

[Chollet et al. 2015] Chollet, F. et al. (2015). Keras. `https://keras.io`.

[de Assis Boldt et al. 2017] de Assis Boldt, F., Rauber, T. W., and Varejao, F. M. (2017). Cascade feature selection and elm for automatic fault diagnosis of the tennessee eastman process. *Neurocomputing*, 239:238–248.

[de Azevedo et al. 2017] de Azevedo, R. R., da Silva, J. M., and Gatsios, R. C. (2017). Análise crítica dos modelos de previsão de série temporal com base no icms estadual. *Revista de Gestão, Finanças e Contabilidade*, 7(1):164–184.

[Domingos Jr et al. 2018] Domingos Jr, S. d. O., de Holanda, L. G., and de Mattos Neto, P. S. (2018). Uma abordagem de combinação não-linear arima-svm para previsão de séries temporais.

[Gautier 2018] Gautier, L. (2018). rpy2—r in python. *Online https://bitbucket.org/rpy2/rpy2*.

[Gonçalves and Rosa 2018] Gonçalves, V. H. and Rosa, J. L. (2018). Forecasting economic time series using chaotic neural networks. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 823–834. SBC.

[Hyndman and Athanasopoulos 2018] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[Hyndman et al. 2007] Hyndman, R. J., Khandakar, Y., et al. (2007). *Automatic time series for forecasting: the forecast package for R*. Number 6/07. Monash University, Department of Econometrics and Business Statistics ....

[Kingma and Ba 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Makridakis and Hibon 2000] Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476.

[Makridakis et al. 2018a] Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018a). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808.

[Makridakis et al. 2018b] Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018b). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Raschka 2018] Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *J. Open Source Software*, 3(24):638.

[Ribeiro 2010] Ribeiro, L. (2010). Modelos mensal e trimestral para projeção de arrecadação do icms para o estado do espírito santo. *Instituto Jones dos Santos Neves. Texto para Discussão*, (10).

[Sala et al. 2018] Sala, D. A., Jalalvand, A., Van Yperen-De Deyne, A., and Mannens, E. (2018). Multivariate time series for data-driven endpoint prediction in the basic oxygen furnace. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1419–1426. IEEE.

[Srivastava et al. 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.