Pós-Graduação em Ciência da Computação

RICARDO TAVARES ANTUNES DE OLIVEIRA

O Impacto do Número de Preditores no Desempenho de Comitês Baseados em Cópulas

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife

2019

RICARDO TAVARES ANTUNES DE OLIVEIRA

O Impacto do Número de Preditores no Desempenho de Comitês Baseados em Cópulas

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Inteligência Artificial.

Orientador: Prof. Dr. Adriano Lorena Inácio de Oliveira

Coorientador: Prof. Dr. Tiago Alessandro Espínola Ferreira

Recife

2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

O48i Oliveira, Ricardo Tavares Antunes de
O impacto do número de preditores no desempenho de comitês baseados em cópulas / Ricardo Tavares Antunes de Oliveira. – 2019.
125 f.: il., fig., tab.

Orientador: Adriano Lorena Inácio de Oliveira.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e anexos.

1. Inteligência artificial. 2. Combinado de modelos. 3. Previsão de séries temporais. I. Oliveira, Adriano Lorena Inácio de (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2019-086

RICARDO TAVARES ANTUNES DE OLIVEIRA

O Impacto do Número de Preditores no Desempenho de Comitês Baseados em Cópulas

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovada em: 22 de Fevereiro de 2019.

Orientador: Prof. Dr. Adriano Lorena Inácio de Oliveira

BANCA EXAMINADORA

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Universidade Federal de Pernambuco

Prof. Dr. Cleber Zanchettin
Universidade Federal de Pernambuco

Prof. Dr. Germano C. Vasconcelos
Universidade Federal de Pernambuco

Prof. Dr. Paulo Renato Alves Firmino
Universidade Federal do Cariri

Prof. Dr. Jarley Palmeira Nóbrega
MCT/CETENE

AGRADECIMENTOS

Primeiramente e acima de tudo, agradeço a DEUS por estar sempre ao meu lado nos momentos bons e ruins, por me proteger, abençoar, por iluminar minhas trajetórias. Agradeço ao SENHOR por mais essa vitória em minha vida.

À minha querida mãe Maria Tavares, por estar ao meu lado me apoiando durante todo esse longo percurso dos meus estudos, dando força e me encorajando, pois sem seu apoio nada disso seria possível.

Ao meu querido pai José Antunes, vulgo "Tetê", por sempre me incentivar a trilhar este caminho, pelo apoio para que eu pudesse um dia chegar a conquistar esse título. O apoio e incentivo do meu pai foi crucial para mais essa conquista.

À toda a minha família, pela torcida em favor do meu sucesso, não só profissional mas também pessoal. Em especial a memória da minha amada avó Maria de Lourdes que apesar de não está mais entre nós, sempre foi amável, carinhosa e a todo o momento torceu por mim, para que eu pudesse conquistar meus objetivos. As minhas tias Eliene, Conceição, Eliana e tio Edivaldo por sempre estarem na torcida por mim.

Agradeço imensamente ao meu orientador Prof. Adriano Lorena, por me acolher como seu orientando, pela paciência, por confiar em mim, por sempre está disponível e principalmente pelos ensinamentos que foram fundamentais no decorrer deste trabalho.

Ao Prof. Tiago Alessandro, ser extraordinário, de qualidades ímpares, que sempre será uma referência para mim tanto como pesquisador quanto ser humano. Meus agradecimentos por confiar e acreditar em mim desde da época do mestrado. Por todos os ensinamentos, pela paciência e amizade. Apesar dos seus curtos momentos de tempo, sempre contribuiu não apenas com este trabalho, mas também com meu crescimento pessoal como cidadão e profissional.

Ao sempre disponível Prof. Paulo Firmino, ser fantástico e pesquisador exemplar, que me orienta desde o mestrado. Na época não sabia e hoje contínuo sem saber expressar e descrever meus sinceros agradecimentos por todos os ensinamentos, pela amizade e por todas as horas que se dedicou para me orientar nesse trabalho.

Eu sinceramente tive a sorte e o imenso privilégio de ser orientado por esses três pesquisadores formidáveis. Que desde de o início fizeram a diferença, meu eterno agradecimento.

Aos amigos, em especial aqueles que acompanharam de perto as dificuldades e realizações nesse tempo: Allisson Dantas, Mariane Ocanha, Anselmo Lacerda, Filippo Régis e Thaíze Fernandes. Muito obrigado por existirem.

A todos os meus colegas de trabalho do Instituto Federal de Mato Grosso do Sul (IFMS) pelo apoio e compreensão. E principalmente, pelas horas de capacitação disponibilizadas pelo IFMS para execução deste trabalho e o término dos meus estudos, meu muito obrigado.

Finalmente, meus agradecimentos aos professores e colegas do Centro de Informática da Universidade Federal de Pernambuco pelo ambiente acolhedor e produtivo, pelo apoio financeiro para participação de eventos, por disponibilizar infraestrutura fantástica para poder desenvolver a pesquisa.

RESUMO

A combinação de preditores através de comitês (*ensembles*) tem atraído pesquisadores de diversas áreas, principalmente por sua acurácia e eficiência em termos estatísticos. Assim, *ensembles* têm superado os respectivos resultados apresentados por modelos individuais. Desta forma, as combinações lineares como média simples, média ponderada, mediana e moda têm sido alternativas clássicas de agregação de previsões apresentadas na literatura de combinação de preditores. Neste trabalho, será adotado um *ensemble* baseado no formalismo de cópulas para combinar uma grande quantidade de modelos individuais de previsão. A literatura em geral, apresenta métodos combinando tipicamente dois, três, cinquenta ou até cem modelos individuais. No entanto, na literatura não existe ou são raros os trabalhos que estudam a quantidade ideal de modelos individuais que devem ser utilizados na combinação, tão pouco um estudo para avaliar a correlação entre o número de modelos e o erro cometido. Assim, existem pressupostos na literatura que quanto mais modelos individuais são agregados à combinação, melhores serão os resultados alcançados. Neste sentido, este trabalho visa por meio de cópulas apresentar um estudo envolvendo a combinação de diversos modelos individuais e analisar o desempenho do *ensemble* à medida que aumenta a quantidade destes modelos na combinação. O *ensemble* adotado provê uma estrutura de combinação que envolve (i) a geração dos preditores individuais, (ii) modelagem dos erros de predição e (iii) a combinação dos modelos individuais via cópulas. Neste trabalho, o desempenho do *ensemble* baseado em cópulas é avaliado considerando séries temporais financeiras, bem como fenômenos meteorológicos, demográficos e hidrológicos. Análises com testes estatísticos foram utilizadas para avaliar se o número de modelos individuais influencia na qualidade das previsões combinadas. A busca pela quantidade ótima de modelos individuais e comparações entre o *ensemble* baseado em cópulas e os bem conhecidos métodos clássicos de combinação linear: média simples, média ponderada, moda e mediana são apresentados. Os testes estatísticos mostram que o número de modelos individuais influencia na qualidade das previsões combinadas para o *ensemble* baseado em cópulas a um nível de significância de 5%, por outro lado os métodos clássicos não passaram nos testes de hipóteses para todas as séries. Os resultados alcançados pelo *ensemble* baseado em cópulas mostram sua superioridade quando comparado com os métodos clássicos de combinação linear.

Palavras-chave: Combinação de modelos. Previsão de séries temporais. Cópulas. Máquina de aprendizagem extrema.

ABSTRACT

The combination of predictors through ensembles has attracted researchers from diverses areas, mainly due to thein statistical efficiency. Thus, ensembles have been superior to the respective single models in statistical terms. In this way, the linear combinations as simple average, weighted average, median and mode have been classical alternatives for prediction aggregation presented in the literature. In this work, it is presented a ensemble based on the copulas to combine a large number of single models. The literature in general presents methods that typically involve two, three, fifty or even one hundred single models. However, the literature does not exist or are rare the studies that study the ideal amount of individual models that should be used in the combination, as well as a study to evaluate the correlation between the number of models and the error committed. However, there is evidence that the more single models are aggregated in the combination, better the results achieved. In this sense, this work aims through copulas to present a study involving the combination of several single models and to analyze the performance of ensemble as the number of single models in the combination increases. The ensemble proposed provides a combination structure involving (i) generation of sigle predictors, (ii) modeling of prediction errors, and (iii) combining individual models via copula. In this work, the performance of the ensemble based on copulas is evaluated considering financial time series as well as series of natural phenomena. Statistical tests show that the number of individual models influences the quality of the predictions for copulas-based ensemble at a significance level of 5%, therwise the classical methods did not pass the hypothesis tests for all the series. The results achieved by the proposed ensemble show superiority when compared to the linear combination method.

Keywords: Ensembles. Time series forecasting. Copulas. Extreme learning machine.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARIMA	*Autoregressive Integrated Moving Average* - Modelo Auto-Regressivo Integrado de Média Móveis
CB	*Copula-Based ensemble* - Método proposto baseado em cópulas
DJ	Série financeira Dow Jones Industrial Average
ELM	*Extreme Learning Machine* - Máquinas de Aprendizagem Extrema
FDA	Função de Distribuição Acumulada
IFM	*Inference Function for Margins* - Inferência de marginais
LL	Modelo Log-Linear
LI	Modelo Linear
MAE	*Mean Absolute Error* - Erro absoluto médio
ME	*Median* - Método mediana
MMV	Método de Máxima Verossimilhança
MO	*Mode* - Método moda
MSE	*Mean Squared Error* - Erro quadrático médio
MV	Máxima Verossimilhança
ND	Série financeira Nasdaq Index
PM	Série de precipitação em Melbourne
PO	Série meteorológica de precipitação em Oldman
QB	Série demográfica da cidade de Quebec
RF	Série hidrológica do Rio Fisher
RMSE	*Root Mean Square Error* - Raiz do erro quadrático médio
RNA	Redes Neurais Artificiais
RO	Série hidrológica do rio Oldman
RS	Série hidrológica do rio Saugeen

SA	*Simple Average* - Método média simples
SLFN	*Single-hidden layer feedforward neural network* - Rede neural com uma única camada escondida
SP	Série financeira S&P500
THEILU	Coeficiente U de Theil
TO	Série meteorológica da temperatura de Oldman
WA	*Weighted Average* - Método média ponderada

LISTA DE SÍMBOLOS

a	Estimativas dos parâmetros do modelo de regressão linear
A, A_x e A_y	Variáveis de estudo
A_{MO}	Amplitude da classe utilizada na moda de Czuber
α_i	Vetor de parâmetros das distribuições marginais dos resíduos
$\hat{\alpha}_i$	Vetor de parâmetros estimados
b	Estimativas dos parâmetros do modelo de regressão linear
β	Parâmetro de dependência da cópula Normal
B	Matriz de bias da RNA via ELM
$C(\cdot)$	Função de distribuição acumulada multivariada
$c(\cdot)$	Função de Distribuição de Probabilidade da Cópula
d	Diferença entre postos
$\dagger$	Indica matriz transposta
db	Base de dados normalizada
$\mathbf{E}$	Vetor de erros
e	Erro
ε	Erro aleatório
$f(\cdot)$	representação de uma função
F_{MO}	Frequência absoluta da classe modal
F_{MO-1}	Frequência anterior da classe modal
F_{MO+1}	Frequência posterior da classe modal
$G(\cdot)$	Função de distribuição acumulada univariada
$\mathbf{H}$	Matriz resultante da saída da função de ativação da RNA
$H(\cdot)$	Função de distribuição acumulada multivariada

$\mathbf{I}$	Matriz inversa da Matriz $\mathbf{H}$
$\mathbb{I}$	Matriz identidade
k	Número de modelos individuais
k_f	Dimensionalidade do espaço de medição
L	Número de modelos individuais
λ	Parâmetro de suavização da cópula de Cacoullos
LI_{MO}	Limite inferior da classe modal
m	Tamanho do conjunto de teste
M^{-1}	Matriz inversa de covariância
$\overline{\text{MAE}}$	Média do MAE
$\widehat{\overline{\text{MAE}}}$	Estimativa do $\overline{\text{MAE}}$
$max(\cdot)$	Função retorna o valor máximo
$min(\cdot)$	Função retorna o valor mínimo
$\overline{\text{MSE}}$	Média do MSE
$\widehat{\overline{\text{MSE}}}$	Estimativa do $\overline{\text{MSE}}$
μ	Média
n	Quantidade de observações da série temporal
N	tamanho da amostra para calcular a média das métricas
n_f	Número total de padrões de treinamento
NC	Número de Classes utilizada na moda de Czuber
ω_i	Peso do método WA
p	Quantidade de características de entrada da RNA via ELM
$P(\cdot)$	Função de distribuição acumulada univariada
$p(\cdots)$	Função de densidade de probabilidade multivariada
$\Phi(\cdot)$	Função inversa da distribuição acumulada multivariada normal
$\varphi^{-1}(\cdot)$	Função inversa da distribuição acumulada univariada normal

$\varphi(\cdot)$	Função de ativação da RNA
r_s	Coeficiente de correlação de Spearman
$mathbbR$	Conjunto dos números reais
ρ	Coeficiente de correlação linear
$\overline{\text{RMSE}}$	Média do RMSE
σ	Desvio padrão
t	Valor do teste de hipótese da distribuição t-student
$\mathbf{T}$	Padrão de entrada da cópula de Cacoullos
τ	Coeficiente de correlação de Kendall
$\overline{\text{THEILU}}$	Média de Theil'U
θ	Parâmetro da cópula Gumbel-Hougaard e Clayton
U_t	Varável aleatória da série temporal no instante t
u_t	Observação da série temporal no instante t
$\hat{u}_t$	Estimativa da série temporal no instante t
v	Instância da distribuição de probabilidade marginal univariada
W	Matriz de pesos da RNA
X	Variáveis de estudo
x_i	Instancia da i-ésima variável X
Y	Variáveis de estudo
$\hat{y}$	Estimativa via regressão linear
y_i	Instancia da i-ésima variável Y

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo serão apresentados os principais conceitos de *ensembles* baseado em cópulas, voltados para o problema de combinação de modelos de previsão de séries temporais. Também serão apresentados métodos alternativos de *ensembles* presentes na literatura. Além disso, serão introduzidos os objetivos, justificativa, as questões de pesquisa e, por fim, a estrutura da tese.

1.1 Considerações Iniciais

Ensembles para previsão de séries temporais têm produzido resultados convincentes, como em Kourentzes, Barrow e Petropoulos (2019), Sobhani et al. (2019), Oliveira et al. (2018), Oliveira et al. (2017), Oliveira et al. (2016), Oliveira (2014), Firmino, Neto e Ferreira (2014), Lux e Morales-Arias (2010), Menezes, Bunn e Taylor (2000). Estatisticamente, as previsões combinadas obtidas pelo *ensemble* são superiores quando comparadas com as previsões dos modelos individuais em termos de acurácia e eficiência. Essa afirmação é suportada por Sammut e Webb (2011), Amendola e Storti (2008), Clemen (1989), Jeong e Kim (2009), Dell'Aquila e Ronchetti (2006), Wallis (2011). *Ensembles* são usualmente aplicados em vários problemas presentes na literatura, como: reconhecimento de padrões por Chen, Dantcheva e Ross (2016), classificação de padrões em Sesmero et al. (2015), Omari e Figueiras-Vidal (2015) e previsão de séries temporais por Bone e Cardot (2008), por exemplo.

Entre as alternativas de *ensembles*, o método de média simples é uma das formas mais triviais de combinar modelos apresentada na literatura (veja, Kourentzes, Barrow e Crone (2014), Krishnamurti et al. (2000)). Neste sentido, o método média simples assume que os modelos individuais contribuem igualmente no processo de combinação, isto é, um modelo com pouca acurácia contribui na combinação igual aos modelos muito acurados. Krishnamurti et al. (2000) comenta que a agregação via média simples, envolvendo esses modelos ruins, degradam (pioram) os resultados globais à medida que mais destes modelos são incluídos.

Na literatura podemos encontrar outros métodos de *ensembles*, que podem ser mais sofisticados que a média simples, como por exemplo, o *ensemble* que combina os modelos individuais através da média ponderada (ARARIPE, 2008; KRISHNAMURTI et al., 2000), onde cada modelo contribui com a combinação de acordo com sua eficiência estatística, tal que o melhor modelo terá a maior contribuição (maior peso) no processo de combinação, enquanto o pior modelo terá a menor contribuição (menor peso). Assim, teoricamente,

este tipo de combinação tende a ser mais eficiente e acurado estatisticamente[1] em relação aos modelos via média simples. Existem também *ensembles* baseados em outros métodos de combinação como mediana e moda (KOURENTZES; BARROW; CRONE, 2014), por exemplo. Contudo, como a média simples, moda e mediana são basicamente medidas de tendência central, a qualidade da estimativa final pode ser muito prejudicada quando os modelos individuais envolvidos têm uma variância grande ou *outliers* que são tidas na estatística como algum valor atípico, isto é, uma observação que apresenta um grande afastamento das demais observações da série, ou que é inconsistente e não faz sentido.

Trabalhos como Oliveira et al. (2017), Oliveira et al. (2016), Assis (2016), Oliveira (2014), Oliveira et al. (2013) têm estudado cópulas para combinar modelos de previsão de séries temporais. Esta abordagem pode ser usada para construir modelos combinados (*ensembles*) para predição de séries temporais através do método de máxima verossimilhança ou mínima variância. Estas combinações são tipicamente realizadas através de funções não lineares. Cópulas são funções que combinam duas ou mais distribuições de probabilidade marginal univariada para construir uma distribuição de probabilidade multivariada, incorporando as interdependências entre estas distribuições univariadas (NELSEN, 2006). Na teoria, combinação de modelos individuais através de cópulas são ainda mais eficazes em problemas de previsão de séries temporais aonde os modelos subestimam ou superestimam a série temporal do evento observado, como apresentado por Oliveira (2014). Além disso, cópulas também têm sido aplicadas em aprendizagem de máquina como Elidan (2014), bem como problema de classificação apresentado por Salinas-Gutiérrez et al. (2010).

As cópulas captam as dependências entre as distribuições de probabilidade marginal univariada, tal que a distribuição multivariada tem as principais informações sobre cada distribuição marginal univariada, e assim, a distribuição multivariada obtém as melhores características de cada distribuição univariada. Este trabalho visa estudar a combinação de modelos individuais aplicada em problemas de previsão de séries temporais por meio de cópulas. Mais especificamente, almeja-se estudar as vantagens e desvantagens em realizar combinações de modelos individuais. Além disso, a literatura de séries temporais parece não enfatizar estudos sobre a quantidade ideal de modelos individuais a serem utilizados no processo de combinação (NELSEN, 2006).

A literatura de cópulas para combinação de modelos individuais de previsão de séries temporais ainda não está consolidada, isto é, muitas questões sobre cópulas ainda precisam ser discutidas, como por exemplo, em quais casos pode ser mais adequado aplicar uma determinada família de cópulas. Neste sentido, requer-se que seja investigada a

[1] O conceito de acurácia (o resultado aproxima-se do alvo) e eficiência (resultados dispersos) abordado neste trabalho, trata-se da ideia de que, a acurácia indica quão próximo o resultado está do valor verdadeiro, enquanto que a eficiência indica o quão próximos os resultados estão uns dos outros. Mais especificamente, o pesquisador pode aferir a acurácia dos resultados por meio da média dos erros e a eficiência pode ser aferida através da variância dos erros (veja mais em Chapra e Canale (2016)).

acurácia e eficiência dos muitos tipos de cópulas como Clayton (WANG et al., 2010), Frank (SINGH; ZHANG, 2007), Gumbel-Hougaard (HOUGAARD, 1986), Normal (NELSEN, 2006), Cacoullos (OLIVEIRA et al., 2018) e outras, para avaliar a qualidade e capacidade de combinação para cada uma delas. E assim, explicar quais cópulas são mais indicadas para determinados fenômenos. Em Oliveira et al. (2017) é apresentado um método de combinação baseado em cópula para combinar apenas dois modelos individuais. Neste trabalho, foram realizados vários experimentos envolvendo 1027 séries temporais sintéticas. Os modelos individuais também foram obtidos de maneira sintética. Os resultados apontaram que as cópulas de Gumbel-Hougaard e Normal são mais adequadas em determinadas situações.

Um aspecto que pode apresentar grande importância quando se trata de combinação de modelos individuais de previsão é a distribuição de probabilidade do erro e sua estrutura de dependência. Estes aspectos são tratados em detalhes no trabalho de Assis (2016), o qual apresenta um método de combinação baseado em cópulas para agregar modelos individuais de diferentes metodologias para séries temporais do mundo real como de crescimento de peixes e séries financeiras. Assim, o método visa ajustar corretamente as distribuições de probabilidade e a estrutura de dependência de cada preditor individualmente. Os resultados mostram que apenas um entre 36 modelos individuais apresentou uma distribuição marginal diferente da normal, aonde a autora relata diante dos achados que assumir a normalidade das distribuições pode não ser um problema.

Os métodos utilizados no processo de combinação são essencialmente críticos para obtenção de melhores previsões. São várias as aplicações para combinação de preditores de séries temporais com Redes Neurais Artificiais (RNA) que têm sido usualmente apresentados na literatura de *ensembles* como por Firmino, Neto e Ferreira (2014), Kourentzes, Barrow e Crone (2014), Oliveira et al. (2013), Barrow, Crone e Kourentzes (2010), Bone e Cardot (2008), Zhang (2007).

No presente trabalho, será proposto um *ensemble* baseado em cópulas para combinar diversas redes neurais treinadas por meio do algoritmo de máquina de aprendizagem extrema. O desempenho deste *ensemble* será investigado junto com os métodos média simples (*simple average* - SA), média ponderada (*weighted average* - WA), moda (*mode* - MO) e mediana (*median* - ME). Um conjunto com diferentes séries temporais do mundo real foram utilizadas para avaliar os métodos, sendo propostos fenômenos financeiros, meteorológicos, demográficos e hidrológicos. Este trabalho se dedicará principalmente em avaliar a influência e impacto da quantidade de modelos individuais de previsão de séries temporais no processo de combinação do *ensemble* proposto.

O método baseado em cópulas, proposto por Oliveira et al. (2017), apresenta uma estrutura de combinação através da cópula normal para dois modelos individuais. Apesar da estrutura ser capaz de agregar mais que dois modelos, esta utiliza uma matriz inversa

convencional que possui alguns problemas se for adotado vários modelos na combinação. O método proposto neste trabalho trata o problema de calcular a matriz inversa através da pseudo-inversa generalizada de Moore-Penrose (veja Penrose (1955), Albert (1972) e Ben-Israel (2001)) que torna possível realizar combinações envolvendo mil modelos ou mais, além de possuir uma estrutura que visa combinar exclusivamente RNAs treinadas por meio do algoritmo de máquina de aprendizagem extrema.

1.2 Questões de Pesquisa

Neste trabalho, deseja-se apresentar resultados que possam responder as seguintes questões:

i À medida que mais modelos individuais são incorporados à combinação, é possível produzir previsões mais acuradas?

ii Existe uma quantidade específica de modelos individuais que possa ser incorporada na combinação, de modo que, garanta previsões estatisticamente eficientes e acuradas?

iii Combinar diversos modelos individuais de previsão de séries temporais, através de cópulas, conduz a melhores resultados comparados com outras abordagens presentes na literatura?

1.3 Objetivos Gerais

Estudar combinações de previsões de séries temporais e seu comportamento ao longo de várias combinações. Para isso, este trabalho adota um *ensemble* baseado no formalismo de cópulas para combinar diversos modelos individuais de previsão.

1.4 Objetivos Específicos

Mais especificamente, deseja-se:

i Analisar o grau de associação entre a relação erro $\times$ quantidade de modelos individuais. Assim, avaliar se existe uma associação negativa entre erro e a quantidade de preditores;

ii Analisar o comportamento da curva do erro do *ensemble* baseado em cópula, a medida que mais modelos individuais são incorporados à combinação;

iii Comparar o *ensemble* proposto com alternativas previamente estabelecidas na literatura.

1.5 Justificativa

Na literatura de previsão de séries temporais muitos trabalhos têm investigado diversos métodos para se conhecer melhor sobre um determinado fenômeno observado. Esta literatura se inicia com trabalhos que propõem a previsão da série temporal como, por exemplo, Ferreira, Vasconcelos e Adeodato (2008), com a finalidade de produzir uma estimativa mais próxima possível do valor real da série. Contudo, pesquisadores têm mostrado que a qualidade das previsões pode ser substancialmente melhorada pelas combinações de modelos individuais. Diversos autores como Amendola e Storti (2008), Lux e Morales-Arias (2010), Kim e Kim (1997) dão suporte a esta afirmação. Existem ainda, pesquisadores como Inoue e Kilian (2006) que têm estudado a seleção de modelos individuais para realizar combinações. Assim, antes de ocorrer o processo de combinação, os modelos são selecionados com base em sua acurácia e eficiência estatística. Desta forma, a ideia é selecionar apenas os modelos que irão contribuir para obter o melhor resultado de combinação. A literatura atual apresenta diversos meios de combinação, sendo por exemplo: média simples, moda, mediana, média ponderada, entre outros (KOURENTZES; BARROW; CRONE, 2014; KUNCHEVA, 2014). Kuncheva (2014) apresenta diversos métodos para combinação de modelos individuais, bem como usa pequenos experimentos envolvendo combinações com, no máximo, cinquenta modelos individuais direcionados para problemas de classificação. Heeswijk et al. (2009) apresentam o método de média ponderada para combinar cem modelos individuais de previsão de séries temporais. Contudo, estes trabalhos têm estudado o método de combinação, negligenciando o comportamento de seus métodos, à medida que mais modelos são incluídos na combinação. Cabrera (2006) apresenta um estudo similar a este trabalho, em que, o autor apresenta um *ensemble* de larga escala para combinar classificadores com duas classes baseados em três diferentes tipos de distribuições de probabilidade.

1.6 Delimitações da Tese

Este trabalho apresenta comparações entre o *ensemble* baseado em cópulas e os clássicos *ensembles* de combinação linear, que serão utilizados porque tais modelos são simples de serem implementados, possuem baixo custo computacional, além de serem largamente citados na literatura como mostra Sobhani et al. (2019), Oliveira et al. (2018), Kourentzes, Barrow e Petropoulos (2019), Barrow e Kourentzes (2016). Metodologias atuais e complexas foram negligenciadas neste primeiro momento, dado que este é um estudo inicial e visa primeiramente ser comparado com as técnicas clássicas de combinação de modelos individuais, assim como trabalhos futuros é proposto a comparação com metodologias alternativas.

Foram utilizadas exclusivamente as redes neurais artificiais neste trabalho pelo

fato destas serem capazes de realizarem seus treinamentos em tempo computacional consideravelmente baixo como exposto por Huang, Zhu e Siew (2006), outra qualidade das redes neurais é a capacidade de apresentar generalizações acuradas, bem como ser possível gerar redes diversificadas.

Os experimentos que serão apresentados não passaram por nenhuma análise de desempenho voltado para avaliar o custo computacional dos métodos, todavia estas avaliações não serão realizadas porque os *ensembles* expostos executam as combinações em tempo computacional baixo, isto é, o tempo utilizado para combinar geralmente é de alguns segundos para séries pequenas com até 2 mil pontos ou menos que dez minutos para séries maiores com até 25 mil observações.

1.7 Estrutura da Tese

A estrutura desta Tese é composta por 6 capítulos, descritos a seguir:

Capítulo 2 - Revisão da Literatura: neste capítulo são definidas algumas características das séries temporais. São apresentadas técnicas para medir o desempenho de modelos individuais. Os conceitos de correlação e regressão linear são abordados. A partir destas definições são expostos os modelos de combinação mais encontrados na literatura, além de abordar sobre redes neurais artificiais (RNA) uma das técnicas mais populares para a previsão de séries temporais. Por fim, as definições do formalismo matemático de cópulas são apresentadas.

Capítulo 3 - Método Proposto: neste capítulo é apresentada uma metodologia de *ensemble* a ser aplicada nos problemas de previsão de séries temporais. A metodologia é composta pelo formalismo matemático de cópulas que realiza a combinação de inúmeras RNAs em um procedimento de combinação que avalia individualmente cada modelo e atribui um grau de importância a cada RNA para determinar um comportamento mais acurado e eficientes das previsões agregadas.

Capítulo 4 - Metodologia dos Experimentos: este capítulo descreve o procedimento utilizado para realizar os experimentos deste trabalho. Desta forma, são expostas as séries temporais adotadas nos experimentos. Neste sentido, foram selecionadas diferente tipos de fenômenos de séries temporais com intuito de captar em detalhes o comportamento do método proposto para diversas situações. Além disso, é definido um conjunto de medidas para avaliar o desempenho das previsões do método proposto. Por fim, as definições dos experimentos para explicar a relação de causa e efeito entre a quantidade de RNAs e o desempenho do método proposto são apresentados por meio da analise de regressão linear.

Capítulo 5 - Análise e Resultados: nete capítulo são expostos os resultados dos experimentos realizados. O método proposto passa por uma conjunto de séries temporais reais com diversos tipos de fenômenos. Os resultados são analisados por meio de medidas de desempenho e gráficos que ilustram o comportamento do método proposto para cada um dos fenômenos. Também são apresentados os resultados dos experimentos para explicar a relação de causa e efeito entre a quantidade de modelos individuais e o desempenho do método proposto. Por fim, é apresentada uma comparação entre o método proposto e outras metodologias estabelecidas na literatura.

Capítulo 6 - Conclusões: neste capítulo são apresentadas as conclusões do trabalho. Discussões sobre o método proposto, contribuições e limitações do trabalho são realizadas. Posteriormente, propostas de novos métodos e, possíveis, extensões sobre o tema são apresentados.

2 REVISÃO DA LITERATURA

Neste capítulo são apresentados alguns conceitos básicos e definições quanto ao problema de combinação de previsões de séries temporais. São apresentadas as principais métricas para avaliar a qualidade das previsões. Conceitos de correlação e análise de regressão linear simples são abordados. As principais técnicas encontradas na literatura sobre combinação de previsões são expostas. Além disso, os conceitos de redes neurais artificiais e o algoritmo de aprendizagem extrema são apresentados. Finalmente, o formalismo de cópulas é mostrado em detalhes.

2.1 Séries Temporais

Segundo Box, Jenkins e Reinsel (1994) uma série temporal é tida como uma sequência de observações ordenadas ao longo do tempo. De modo geral, uma série temporal é resultado de uma sequência de observações medidas a partir de um determinado fenômeno ao longo de um espaço de tempo, o qual pode ser contínuo ou discreto. Fenômenos como o índice de inflação anual do Brasil, a quantidade de chuva mensal de uma região, o preço de fechamento diário de uma ação no pregão da bolsa de valores e a quantidade de acessos diários em um site da internet, entre outros. Assim, uma série temporal pode ser representada por $U_t = \{u_t \in \mathbb{R} | t = 1, 2, \cdots, n\}, n \in \mathbb{N}$ onde t é o índice cronológico e n é a quantidade total de observações da amostra, tal que $n \geq 1$. Assim, o objetivo é prever a próxima observação da série u_t, ou seja, estimar u_{t+1}.

As séries temporais podem apresentar características particulares, isto é, as séries podem mostrar um padrão comportamental. A Figura 1 [1] apresenta uma série temporal do quantitativo de passagens internacionais emitidas por companhias aéreas no período de Janeiro de 1949 até Dezembro de 1960. Nesta série temporal é fácil notar que existe uma tendência, ou seja, a série tem aumentado o quantitativo de passagens emitidas ao longo dos anos. Assim, é notável o crescimento das emissões de passagens quando se observa o período entre 1949 e 1950 com 1950 e 1951, bem como ocorre com os períodos seguintes até 1960. Esta série temporal também apresenta a característica de sazonalidade, que pode ser definida como um padrão que se repete periodicamente ao longo da série temporal (BOX; JENKINS; REINSEL, 1994).

A Figura 2 [2] apresenta outra série temporal com outra caraterística importante a ser

[1] Esta série temporal foi obtida por meio do site Datamarket disponível em: <https://datamarket.com/data/set/22u3/international-airline-passengers-monthly-totals-in-thousands-jan-49-dec-60#!ds=22u3&display=line> e publicada por Box e Jenkins (1976).

[2] Esta série temporal foi obtida por meio do site Datamarket disponível em: <https://datamarket.com/

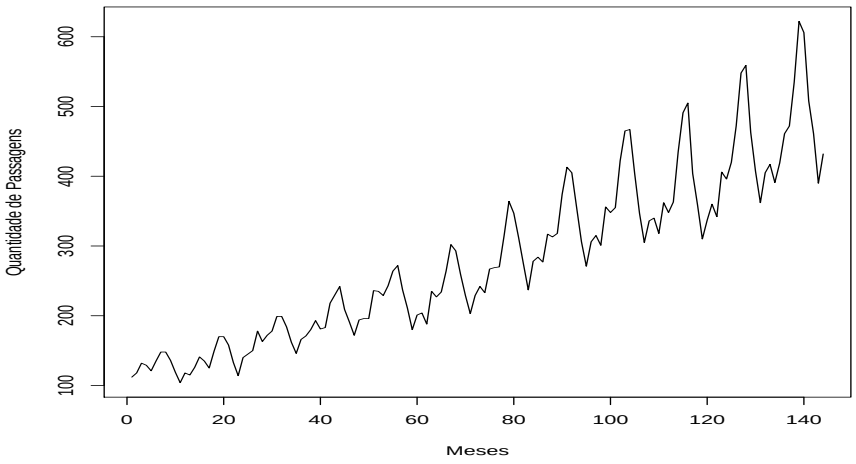


Figura 1 – Quantidade de passagens de companhias aéreas internacionais: Total de passagens mensais em milhares entre Janeiro de 1949 e Dezembro de 1960 (Fonte: Box e Jenkins (1976)).

observada que trata-se da estacionariedade. Assim, uma série temporal é dita estacionária em sua média quando as observações da série se mantém ao redor da média ao longo do tempo e a média não está em função do tempo, ou seja, não se altera ao longo do tempo. Como pode ser notado na figura, o número médio de nascimentos diários em Quebec é de aproximadamente 250, e esta média se mantém ao passar do tempo. A série temporal também pode ser estacionária em relação a variância, assim é possível observar que a variância da série não se altera ao longo do tempo (FILHO; PESSOA, 2015).

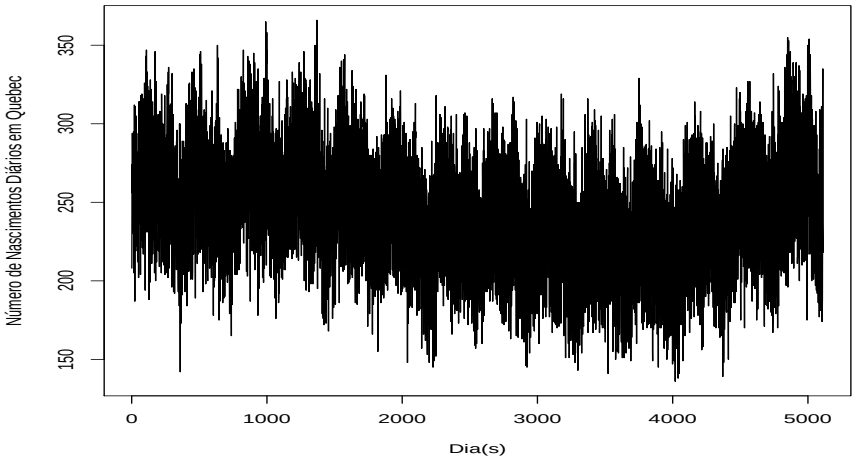


Figura 2 – Número de nascimentos diários em Quebec ocorridos entre 1 de Janeiro de 1977 à 31 de Dezembro de 1990 (Fonte: Hipel e McLeod (1994)).

data/set/235j/number-of-daily-births-in-quebec-jan-01-1977-to-dec-31-1990#!ds=235j&display=line> e publicada por Hipel e McLeod (1994).

2.2 Métricas de Desempenho

Conforme mencionado no Capítulo 1, são diversos os trabalhos apresentados na literatura de previsão de séries temporais, que propõem metodologias para obter estimativas de um determinado fenômeno (série temporal). A principal motivação destes trabalhos consiste em produzir estimavas mais próximas possíveis dos valores reais observados no fenômeno. Assim, quando as estimativas tem o centro de gravidade o valor real da série temporal implica dizer que a metodologia realizou estimativas sem viés, ou seja, sem erro médio associado. Na literatura de séries temporais, uma das formas mais adotadas para se calcular o desempenho das metodologias é através do erro quadrático médio (*Mean Squared Error* - MSE) que calcula a média dos erros quadráticos, essa métrica normalmente apresenta valores altos quando comparado com outras métricas, mas o MSE é largamente utilizado como medida de qualidade das estimativas pela sua capacidade de avaliar tanto acurácia quanto a eficiência do modelo. O MSE pode ser escrito da seguinte forma:

$$\text{MSE} = \frac{1}{M}\sum_{t=1}^{M}(\hat{u}_t - u_t)^2; \tag{2.1}$$

em que M é o tamanho do conjunto de teste para a série temporal u_t e a estimativa $\hat{u}_t$. Outra medida comum utilizada para medir o desempenho dos modelos individuais é o erro absoluto médio (*Mean Absolute Error* - MAE) que ao contrário do MSE não utiliza o erro quadrático mas o erro absoluto, essa métrica tem pouca sensibilidade a variância do erro e apresenta valores pequenos comparado com o MSE. O MAE pode ser obtido através da seguinte fórmula:

$$\text{MAE} = \frac{1}{M}\sum_{t=1}^{M}|\hat{u}_t - u_t|. \tag{2.2}$$

O coeficiente U de Theil pode ser visto na obra de Theil (1965), Theil (1966) *apud* Bliemel (1973) que descreve a diferença entre as duas estatísticas propostas por Theil em *Economic Policy and Forecast* e *Applied Economic Forecasting*. Tal métrica é denominada neste trabalho como THEILU, podendo seu coeficiente estar no intervalo entre 0 e 1. Para THEILU=0 indica uma previsão perfeita ($u_t = \hat{u}_t$, para todo t), e caso contrário existe uma desigualdade entre os valores reais e os previstos. Assim, THEILU=1 indica que pelo menos uma das variáveis possui valor zero ($u_t = 0$ ou $\hat{u}_t = 0$ para todo t), ou ainda existe uma proporcionalidade negativa (BLIEMEL, 1973). O coeficiente U de Theil é dado por:

$$\text{THEILU} = \frac{\left[\frac{1}{M}\sum_{t=1}^{M}(u_t - \hat{u}_t)^2\right]^{\frac{1}{2}}}{\left[\frac{1}{M}\sum_{t=1}^{M}u_t^2\right]^{\frac{1}{2}} + \left[\frac{1}{M}\sum_{t=1}^{M}\hat{u}_t^2\right]^{\frac{1}{2}}}; \tag{2.3}$$

e a raiz do MSE conhecida como *Root Mean Square Error* - RMSE que mostra-se mais sensível a variância dos erros do modelo comparado com o MAE, dado que os erros são elevados ao quadrado é apenas por último calculada a raiz quadrada, isso indica que o RMSE pode ser melhor aplicado quando os erros grandes são indesejados. Logo o RMSE pode ser dado por:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{t=1}^{M}(\hat{u}_t - u_t)^2}. \tag{2.4}$$

2.3 Correlação

Uma maneira de avaliar se a quantidade dos modelos individuais influenciam no desempenho do método proposto é através do coeficiente de correlação. Este coeficiente é uma medida de associação entre duas variáveis quantitativas. Esta medida foi proposta inicialmente por Karl Pearson em 1896. Por isso, tem sido largamente conhecida por coeficiente de correlação de Pearson. Atualmente, já existe também o coeficiente de Spearman e Kendall (CALLEGARI-JACQUES, 2009).

O coeficiente de correlação de Pearson é dado por:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.5}$$

tal que, n é o tamanho da amostra, x_i e y_i são as observações no índice i, enquanto que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (média da variável X) e $\bar{y}$ segue a mesma ideia.

O coeficiente de correlação pode assumir infinitos valores (conjunto dos reais) entre -1 e +1, onde valores negativos indicam uma correlação do tipo negativa (inversa) e valores positivos ocorrem quando a correlação for positiva. Assim, quando duas variáveis X e Y estão positivamente correlacionadas (isto é, são diretamente proporcionais) quando X cresce, os valores de Y tendem a crescer juntos, ou seja, ambas as variáveis variam no mesmo sentido. Por outro lado, quando a correlação é negativa (isto é, são inversamente proporcionais), quando X cresce, os valores de Y tendem a diminuir (BARBETTA; REIS; BORNIA, 2010; CALLEGARI-JACQUES, 2009).

Um coeficiente de +1 indica que as duas variáveis são perfeitamente correlacionadas de maneria positiva, desta forma, quando o valor de uma variável cresce a outra tende a crescer proporcionalmente, assim os pontos formam uma linha reta inclinada crescente em um gráfico de dispersão. O coeficiente -1 indica uma perfeita correlação negativa, isso ocorre quando todos os pontos formam uma linha reta inclinada de forma decrescente. Por outro lado, quando não existe correlação linear, isto é, o coeficiente igual a 0, os pontos se distribuem em forma não linear (CALLEGARI-JACQUES, 2009). Vieira (2015) menciona que os coeficientes de correlação têm valores que podem variar conforme a área de atuação, como ocorre nas ciências físicas aonde os coeficientes são relativamente altos, enquanto

que nas ciências da saúde os valores são menores por causa da variabilidade dos fenômenos biológicos. Nas ciências do comportamento o autor relata que coeficientes iguais ou maiores que 0,70 são raríssimos.

2.3.1 Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman trata-se de uma estatística não-paramétrica padronizada que mede a grandeza de associação entre duas variáveis que não necessitam satisfazer a exigência de normalidade por meio de teste de hipótese, para avaliar se os dados seguem uma distribuição bivariada normal, como ocorre com o coeficiente de Pearson. Esta estatística surgiu em 1904 e apresenta-se na literatura como uma alternativa quando as variáveis não satisfazem as exigências para o teste do coeficiente de produto-momento de Pearson(CALLEGARI-JACQUES, 2009; FIELD, 2009).

Para calcular o coeficiente de correlação de Spearman (r_s) cada variável X e Y precisa ter um conjunto de amostras pareadas e com o mesmo tamanho, logo as variáveis x_i e y_i ($i = 1, \cdots, m$, onde $y_i \in Y$ e $x_i \in X$) são tidas como supostamente correlacionadas. Então, no primeiro momento, os valores das variáveis separadamente são ordenados (*rank*). Neste sentido, se as variáveis estiverem correlacionadas positivamente, os *ranks* baixos em uma delas serão de maneira geral acompanhados pelos *ranks* baixos na outra e vice-versa. Contudo, se não houver correlação, os *ranks* baixos em uma não irão influenciar nos *ranks* da outra, que poderão ser baixos, médios ou altos. Desta forma, a comparação entre os *ranks* de cada variável irá indicar o tipo de correlação existente entre elas (CALLEGARI-JACQUES, 2009). Calcula-se o coeficiente de Spearman pela seguinte equação (KUTNER et al., 2005):

$$r_s = \frac{\sum_{i=1}^{n}(R_{i,1} - \bar{R}_1)(R_{i,2} - \bar{R}_2)}{\left[\sum_{i=1}^{n}(R_{i,1} - \bar{R}_1)^2 \sum_{i=1}^{n}(R_{i,2} - \bar{R}_2)^2\right]^{\frac{1}{2}}} \tag{2.6}$$

onde n reflete o número de pares de valores. O *rank* de X é denotado por $R_{i,1}$ e similarmente o *rank* de Y é dado por $R_{i,2}$. O $\bar{R}_1$ é a média do *rank* $R_{i,1}$ e $\bar{R}_2$ é a média do *rank* $R_{i,2}$

Quando o tamanho da amostra for pequena e houver uma quantidade grande de *ranks* empatados, isto irá fazer com que após a ordenação haja muitos valores para X ou Y com o mesmo *rank*. Nessas situações deve ser adotado o τ de Kendall, outra estatística de correlação não-paramétrica, por esse motivo Field (2009) acredita que o coeficiente de Kendall seja mais preciso em relação ao coeficiente de Spearman, nesses casos.

2.3.2 Teste de Hipótese sobre a Correlação

Após o cálculo do coeficiente de correlação em uma amostra, resta ainda realizar o teste de hipótese para ter certeza de que a correlação encontrada não foi apenas casual, indicando simplesmente um erro de desvio da amostragem. O teste de hipótese supõe inicialmente

que a correlação entre as variáveis é nula. Assim, as hipóteses estatísticas são:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0, \end{aligned} \tag{2.7}$$

o nível de significância do teste é dado por α, enquanto que os graus de liberdade são dados por $gl = n - 2$ e o valor crítico do teste t pode ser calculado:

$$t = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}, \tag{2.8}$$

caso $|t| \leq t_{1-\alpha;gl}$ então não rejeita a hipótese H_0. O valor de $t_{1-\alpha;gl}$ pode ser obtido através da distribuição t quando a for utilizado o coeficiente de correlação de Pearson. Para os casos de Spearman usa-se uma abordagem não paramétrica (CALLEGARI-JACQUES, 2009).

2.4 Regressão Linear Simples

O termo regressão surgiu em 1886, quando Francis Galton publicou um artigo que apresentava um modelo matemático-estatístico para explicar o comportamento de uma variável através de outra. Neste sentido, o objetivo do trabalho era predizer a altura de um indivíduo através das alturas de seus pais. Este método foi aperfeiçoado e atualmente é conhecido por regressão linear e tem sido aplicado nas diversas áreas do conhecimento (BARBETTA; REIS; BORNIA, 2010; CALLEGARI-JACQUES, 2009).

A regressão linear é estudada quando existem razões para supor uma relação de causalidade entre duas variáveis quantitativas. Assim, a regressão estuda o relacionamento de uma variável Y chamada de variável resposta ou dependente, junto com a variável X, denominada como variável explicativa ou preditiva (BARBETTA; REIS; BORNIA, 2010; CALLEGARI-JACQUES, 2009). Desta forma, por meio da regressão linear podemos estimar o valor da variável resposta Y quando se manipula o valor da variável explicativa X. Assim, por exemplo, imaginando que exista uma correlação estatisticamente significativa entre as variáveis peso (Y) e altura (X), é possível estimar o peso de um individuo quando este tiver uma determinada altura.

A principal diferença entre os estudos com análise de regressão e correlação é porque na regressão, parte do pré-suposto que as variáveis Y e X possuem um efeito de causalidade. Assim, na análise de regressão é assumida teoricamente tal correlação com intuito de estudar a causa e efeito entre as variáveis, de tal maneira que as observações são pareadas $(x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i), \cdots, (x_m, y_m)$ (BARBETTA; REIS; BORNIA, 2010).

De acordo com Callegari-Jacques (2009) muitas relações de causa e efeito são resumidas apenas por uma linha reta. A análise de regressão linear visa exatamente

fornecer por meio de linhas retas (por isso o termo linear) explicações entre as relações com apenas uma variável preditiva (por isso o termo simples). Contudo, existem outros tipos de regressão, como regressão linear múltipla e regressão linear logística.

A regressão linear simples é dada por:

$$\hat{y} = a + bx \tag{2.9}$$

em que, a e b são os coeficientes linear e angular, respectivamente e $\hat{y}$ é uma estimativa para y. O coeficiente angular representa a inclinação da reta, tal que, para cada valor acrescido em x representa um acréscimo ou decréscimo em $\hat{y}$. Estes parâmetros do modelo podem ser calculados através da equação:

$$b = \frac{n\sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{2.10}$$

e

$$a = \bar{y} - b\bar{x} \tag{2.11}$$

assim, para cada valor em $x_i (i = 1, 2, \cdots, m)$ a estimativa pode ser dada por:

$$\hat{y}_i = a + bx_i \tag{2.12}$$

o resíduo entre o valor observado e predito é obtido pela equação:

$$e = y_i - \hat{y}_i \tag{2.13}$$

de modo geral, ao plotar o gráfico de dispersão, geralmente a reta obtida pela equação não passa exatamente sobre todos os pontos observados sempre. Isso ocorre por causa da existência de fatores não controláveis no processo. Nessas situações, $\sum_{i=1}^{n} e_i \neq 0$.

2.4.1 Transformações

Conforme Barbetta, Reis e Bornia (2010) apresentam em sua obra, é comum diversas variedades de estudos envolvendo dados com distribuição assimétrica. Nesses casos, valores grandes em X irão influenciar muito na inclinação da reta. Assim, nessas situações a dispersão dos dados sugere uma relação não linear o que dificulta a aplicação da análise regressão linear. Para a solução deste problema, recomenda-se a utilização de transformações logarítmicas que podem ocorrer tanto em valores de X como em valores de Y:

$$\log(Y) = a + b\log(x_i) + \varepsilon_i, \tag{2.14}$$

onde, ε_i é um erro aleatório no instante i. Após a estimativa dos parâmetros a e b, nesta ordem, pode-se aplicar a transformação inversa:

$$\hat{y} = e^{a+b\times\log(x)}. \tag{2.15}$$

Exitem situações típicas em que podem ser recomendada a transformação logarítmica apenas na variável X, assim, essa modelagem é descrita por:

$$\hat{y}_i = a + b\log(x_i) + \varepsilon_i, \tag{2.16}$$

Barbetta, Reis e Bornia (2010) ainda sugerem nos casos onde os dados apresentam uma relação não linear com um aumento da variância à medida que X aumenta, a aplicação da transformação logarítmica nos valores da variável Y, ajustando ao seguinte modelo:

$$\log(y_i) = a + bx_i + \varepsilon_i \tag{2.17}$$

A modelagem apresentada nesta seção é capaz de estimar valores futuros para uma determinada variável de interesse. Neste trabalho esse tipo de modelagem tem um papel importante para prever o erro que será cometido pelo método proposto. A Seção 4.4 traz mais detalhes de como essa modelagem será abordada neste trabalho.

2.5 Combinando Previsões de Séries Temporais

2.5.1 Média Simples

Entre as alternativas de combinação de previsão de séries temporais, as mais adotadas geralmente na literatura são caracterizadas como funções de combinação linear (LCFs). Na literatura de LCFs, o método de média simples (S*imple Average* - SA) dos modelos individuais têm sido enfatizado por Menezes, Bunn e Taylor (2000), Zou e Yang (2004), Firmino, Neto e Ferreira (2014). A abordagem do SA assume que os pesos dos modelos individuais são constantes e iguais, isto é, os pesos de cada modelo são dados por $\omega_j = \frac{1}{k}$. De modo geral, o método SA é apresentado por:

$$\mathrm{SA}_t = \frac{1}{k}\sum_{j=1}^{k} x_{t,j} \tag{2.18}$$

onde k é o número de modelos individuais e $x_{t,j}$ é a previsão do j-ésimo modelo individual no instante de tempo t.

2.5.2 Média Ponderada

Enquanto o método de média simples assume que todos os modelos individuais contribuem igualmente para o processo de combinação, o método de média ponderada atribui para cada modelo um peso em função da sua eficiência estatística. Assim, os métodos de média

ponderada são essencialmente mais sofisticados que o SA (KRISHNAMURTI et al., 2000). Especificamente, o método de média ponderada pode ser genericamente apresentado como:

$$\mathrm{WA}_t = \sum_{j=1}^{k} \omega_j \times x_{t,j} \tag{2.19}$$

onde, k é conhecido como o número de modelos individuais, ω_j é o peso atribuído ao j-ésimo modelo individual $x_{t,j}$ no instante t. Especificamente, os pesos tem usualmente assumido valores no intervalo 0 a 1, tal que $0 \leq \omega_j \leq 1$ e $\sum_{j=1}^{k} \omega_j = 1$. Kuncheva (2014) e Filho (2015) apresentam em suas obras uma forma de como calcular os pesos, sendo:

$$\omega_i = \frac{\dfrac{1}{\sigma_i}}{\sum_{j=1}^{k} \dfrac{1}{\sigma_j}}, i = 1, \cdots, k, \tag{2.20}$$

onde σ_i e σ_j são respectivamente as variâncias dos erros para o i-ésimo e j-ésimo modelos de previsão. Em Araripe (2008) é apresentada uma outra alternativa de *ensemble* baseado na média ponderada para combinar de dois em dois modelos, isto é $k = 2$. Na obra, o modelo é dado por:

$$\mathrm{WA}_t = \omega \times x_1 + (1 - \omega) \times x_2, \tag{2.21}$$

onde, ω corresponde ao fator que minimiza a variância do erro da previsão combinada, que pode ser escrita da seguinte forma:

$$\omega = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \tag{2.22}$$

assim, ρ é o coeficiente de correlação linear entre os erros de previsão. O σ_1 e σ_2 corresponde ao desvio padrão dos erros de previsão dos modelos individuais x_1 e x_2.

Existem também abordagens através da mediana e moda apresentadas por Kourentzes, Petropoulos e Trapero (2014) que mostram em seu trabalho que a combinação de modelos individuais com diferentes frequências pode melhorar as previsões. Além disso, o autor encontrou pequenas diferenças em utilizar a média ou mediana. Muitos pesquisadores Hillebrand e Medeiros (2010), Chen e Ren (2009), Kourentzes, Petropoulos e Trapero (2014) também têm estudado o método de *Bagging* (*Bootstrap aggregation and combination*) para prover melhores previsões combinadas. A ideia básica do *Bagging* é treinar vários modelos individuais e então combinar as previsões obtidas pelos modelos, usualmente através de uma função de combinação linear.

Desta forma, existem muitos trabalhos que vêm utilizando como critério de combinação a média, moda ou mediana. Contudo, como a média, moda e mediana são basicamente medidas de tendência central, onde a qualidade da estimativa final pode ser muito prejudicada quando os modelos individuais envolvem uma variância alta. As combinações via média também são sensíveis aos *outliers* e não é capaz de corrigir o erro quando os modelos individuais subestimam ou superestimam a série temporal (OLIVEIRA, 2014).

2.5.3 Mediana

A combinação dos modelos individuais de previsão por meio do método da mediana trata-se basicamente de uma função $f(x_t)$ onde $x_t = x_{t,1}, x_{t,2}, \cdots, x_{t,i}, \cdots, x_{t,k}$ que calcula a mediana das previsões dos modelos individuais. Neste sentido, primeiramente a função ordena todos os valores em ordem crescente ou decrescente e posteriormente seleciona o elemento que estiver posicionado no índice do meio do conjunto de previsões. Em outras palavras, por exemplo, assumindo $x_t = 15, 15, \mathbf{27}, 32, 44$ a previsão agregada via mediana seria $f(x_t) = 27$.

As combinações via mediana para conjunto de dados na qual a quantidade de previsões (observações) for par, o cálculo da mediana será dado pela média dos dois elementos que estiverem no índice do meio do conjunto de previsões, logo podemos exemplificar assumindo que $x_t = 12, 15, \mathbf{15}, \mathbf{27}, 32, 44$ a previsão agregada pela mediana seria $f(x_t) = (\mathbf{15} + \mathbf{27})/2 = 21$ (veja mais detalhes em Vieira (2015)).

2.5.4 Moda

As agregações de previsões via moda são obtidas por meio do cálculo da observação que mais aparece no conjunto de dados. Assim, a função $f(x_t)$ para $x_t = x_{t,1}, x_{t,2}, \cdots, x_{t,i}, \cdots, x_{t,k}$ pode ser exemplificada assumindo que $x_t = \mathbf{15}, \mathbf{15}, 27, 32, 44$, assim, $f(x_t) = 15$, porque a observação 15 foi a que mais se repetia no conjunto de dados.

Contudo, quando as previsões são compostas por números reais com diversas casas decimais, pode ocorrer de não ser possível definir uma moda para o conjunto de previsões, assim por exemplo, assumindo que $x_t = 15.957, 15.969, 27.104, 32.988, 44.276$ a previsão agregada para este conjunto não existiria. Além disso, se o conjunto de previsões não tiver valores repetidos, a agregação via moda não será possível, uma vez que há a restrição da necessidade de existir pelo menos um dado repetido. Neste sentido, uma forma para corrigir este problema é recorrer a modelos de moda diferentes do tradicional, assim uma alternativa é a moda de Czuber (FÁVERO; BELFIORE, 2017), que consiste em agrupar o conjunto de previsões em classes, que a partir da classe com maior aparição (frequência) de previsões, denominada classe modal, a moda de Czuber será calculada. A quantidade de classes pode ser definida por:

$$NC = \sqrt{n} \tag{2.23}$$

ou

$$NC = 1 + 3.3 \times \log(n), \tag{2.24}$$

tal que n é o número de previsões (observações). As classes são divididas em amplitudes (intervalos de valores) iguais e podem ser calculadas da seguinte forma:

$$A_{MO} = \frac{\max(x) - \min(x)}{NC}. \tag{2.25}$$

Desta forma, a moda de Czuber pode ser calculada como segue (FÁVERO; BELFIORE, 2017):

$$\text{MO} = LI_{MO} + \frac{F_{MO} - F_{MO-1}}{2 \times F_{MO} - (F_{MO-1} - F_{MO+1})} \times A_{MO} \tag{2.26}$$

em que:
LI_{MO} é o limite inferior da classe modal;
F_{MO} é a frequência absoluta da classe modal;
F_{MO-1} é a frequência absoluta da classe anterior à classe modal;
F_{MO+1} é a frequência absoluta da classe posterior à classe modal;
A_{MO} amplitude da classe modal.

2.6 Máquinas de Aprendizado Extremo

As redes neurais artificiais têm sido largamente utilizadas em vários ramos da ciência para solucionar problemas complexos (HAYKIN, 1999). Isso ocorre porque as redes neurais têm a propriedade de atuar como uma função de aproximação universal, em outras palavras, capaz de mapear um conjunto de entradas em uma respectiva saída desejada com uma taxa de sucesso associada. Desta forma, as redes neurais se tornaram uma ferramenta poderosa, principalmente no ramo da computação, onde é capaz de tratar diversos tipos de problemas como de classificação, previsão, reconhecimento, entre vários outros. Tradicionalmente, os algoritmos de aprendizagem presentes na literatura, como o *backpropagation* (veja mais detalhes em Haykin (1998), Haykin (2001), Braga, Carvalho e Ludermir (2000)), por exemplo, realizam uma abordagem de retropropagação do erro de treinamento através das camadas da rede, e desta forma ajustando os pesos sinápticos. Contudo, estas metodologias tradicionais de treinamento requerem um processo iterativo que pode custar vários minutos, horas ou inúmeros dias para treinar uma determinada rede neural. Este processo pode ser tão custoso porque todos os parâmetros da rede precisam ser ajustados, isto é, peso sináptico e *bias*, criando uma dependência entre as diferentes camadas (HUANG; ZHU; SIEW, 2006).

Huang e Babri (1998) propuseram uma rede neural com uma única camada escondida (denominada *Single-hidden layer feedforward neural network* - SLFN), utilizando praticamente qualquer função de ativação não-linear é capaz de aprender. Posteriormente, Huang (2003) mostra que as SLFNs com vários neurônios na camada escondida podem ter seus pesos de entrada e o *bias* ajustados aleatoriamente. Esse novo método mostra-se interessante em termos computacionais, pois a rede neural continua com propriedades de generalização utilizando observações distintas. Neste sentido, diferentemente das redes neurais tradicionais, em que o processo de aprendizagem ajusta todos os parâmetros da rede neural através de inúmeras iterações, este método visa ajustar apenas os pesos da camada de saída. Huang, Zhu e Siew (2003) realizaram simulações envolvendo aplicações

reais e artificiais que mostram a eficiência do método proposto, bem como destaca o desempenho em termos do tempo computacional, mostrando que é necessário menos tempo em relação ao *backpropagation* para rede produzir generalizações acuradas.

As redes neurais tipicamente com uma única camada escondida, continuaram sendo alvo de estudos. Huang, Zhu e Siew (2004) propõem um novo método chamado por *Extreme Learning Machine* (ELM), em português Máquina de Aprendizado Extremo, que consiste em atribuir os pesos de entrada e o *bias* aleatoriamente se as funções de ativação na camada oculta são infinitamente diferenciáveis. Enquanto que os pesos da camada de saída podem ser considerados basicamente como um sistema linear, o qual pode ser resolvido através da pseudo-inversa generalizada de Moore-Penrose, descrita em Penrose (1955), Albert (1972), Ben-Israel (2001). Huang, Zhu e Siew (2006) mencionam que sua principal contribuição foi propor um algoritmo de treinamento para rede neural com uma camada escondida, cuja velocidade de treinamento pode ser milhares de vezes mais rápida comparada com os métodos tradicionais de treinamento de redes neurais como o algoritmo *backpropagation*.

2.6.1 Conceitos Básicos

Inicialmente, os dados de entrada da rede neural devem ser arranjados em uma matriz X com n colunas e p linhas, conforme descrito abaixo:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \dots & x_{p,n} \end{bmatrix} \tag{2.27}$$

onde, $x_{p,n} \in \mathbb{R}$, p trata-se da quantidade de características da entrada e n é o número de padrões de entrada.

Seja T a matriz de saída, que corresponde às observações que a rede neural terá acesso para realizar o treinamento e consequentemente extrair aprendizado. A matriz de saída é composta por n colunas e L linhas, tal que, as colunas representam a quantidade de padrões de entrada, enquanto as linhas representam as respectivas saídas da rede, conforme é apresentado abaixo:

$$T = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,n} \\ t_{2,1} & t_{2,2} & \dots & t_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{L,1} & t_{L,2} & \dots & t_{L,n} \end{bmatrix} \tag{2.28}$$

onde, $t_{L,n} \in \Re$, L trata-se da quantidade de saídas e n é o número de padrões de entrada.

Supõem-se que existe uma função matemática capaz de mapear uma matriz de entrada qualquer X para uma matriz de saída $\hat{T}$. A Equação 2.29 representa esse mapeamento de forma matemática:

$$\hat{T} = F(X) \tag{2.29}$$

onde, a função *F(X)* é desconhecida a *priori*. Logo, não existem precedentes da formulação matemática capazes de associar a matriz de entrada X com sua respectiva matriz de saída $\hat{T}$. Desta forma, $F(\cdot)$ tem a função de realizar o mapeamento da matriz de entrada para matriz de saída.

Assim, diferentemente das redes neurais convencionais a *Extreme Learning Machine* (ELM) é um algoritmo representado basicamente por matrizes. Dessa forma, as operações realizadas pelos neurônios artificias são sob matrizes.

2.6.2 Aprendizagem para Problemas de Previsão

O treinamento consiste em computar todos os pesos (parâmetros) da rede neural artificial. Basicamente, é preciso calcular a matriz de pesos W, a matriz de *bias* B e a matriz de pesos da camada de saída β. O Algoritmo 1 exemplifica o treinamento de uma rede neural do tipo SLFN utilizando o método *Extreme Learning Machine* para ajustar os pesos da camada de saída.

O treinamento da rede ELM aplicada ao problema de séries temporais pode ser dado em quatro etapas. A primeira etapa consiste na normalização dos dados, que em muitos casos é necessária para contribuir na qualidade das generalizações da rede. Assim, esta etapa é tida principalmente como um pré-processamento dos dados que serão utilizados no processo de ajuste da ELM. A normalização pode ser expressa como segue:

$$novo = \frac{x - min(x)}{max(x) - min(x)}. \tag{2.30}$$

Posteriormente, os dados são arranjados dentro da matriz de entrada X. Neste sentido, pode-se exemplificar uma matriz de entrada arranjada com dados de uma série temporal $\mathbf{u} = (1, 2, 3, 4, 5, 6, 7)$ com *lag* igual a 4. Assim, a matriz de saída T corresponde as previsões da série $\mathbf{u}$, sendo respectivamente $T = (5, 6, 7)$. Portanto, para a entrada $\mathbf{x}_1 = (1, 2, 3, 4)$ a saída será $T_{1,1} = (5)$, para a entrada $\mathbf{x}_2 = (2, 3, 4, 5)$ a saída corresponderá a $T_{1,2} = (6)$, e assim segue o mesmo raciocínio para as demais entradas, conforme

especificado a seguir:

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix}; \qquad T = \begin{bmatrix} 5 & 6 & 7 \end{bmatrix}. \tag{2.31}$$

Na segunda etapa, os pesos da camada escondida são inicializados com valores escolhidos aleatoriamente no intervalo entre 0 e 1. Os pesos são formalmente representados pela matriz W_{ij}, $i = 1, \cdots, m$ e $j = 1, \cdots, p$, tal que, m é a quantidade de neurônios na camada escondida e p trata-se da quantidade de atributos (características) do padrão de entrada. Em seguida, inicializa-se o *bias* da rede neural, ou seja, o vetor *bias* assume valores podendo ser -1 ou 1. Através do vetor *bias* é possível replicar os valores para a respectiva matriz B_{ij}, $i = 1, \cdots, m$ e $j = 1, \cdots, n$ de *bias*, tal que, n é o número de padrões de entrada da rede neural. Logo, assumindo que $bias = (1, 1, -1, 1, -1)$ e $n = 3$ a matriz B pode ser expressada da seguinte forma:

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \end{bmatrix}. \tag{2.32}$$

A terceira etapa consiste em calcular a saída da camada oculta. Esse fluxo corresponde em encontrar a matriz **H**. Esta saída em uma rede neural tradicional pode ser calculada através de alguma função de ativação. Contudo, esta equação para ELM pode ser reescrita da seguinte forma:

$$\mathbf{H} = \varphi(B + W \times X) \tag{2.33}$$

onde, **H** é a matriz resultante da saída da função de ativação $\varphi(\cdot)$ aplicada a cada um dos elementos da matriz obtida da operação $B + W \times X$. Assim, H_{ij} possui dimensão $i = 1, \cdots, m$ e $j = 1, \cdots, n$, é a saída da camada oculta da rede neural.

Por fim, na última etapa, a ideia é calcular os pesos da camada de saída da rede neural. Assim, buscando a melhor representação para a Equação 2.29, pode-se dizer que a função $F(X)$ tem como objetivo encontrar os pesos da camada de saída, que seja possível mapear a entrada para saída desejada, uma vez que os pesos anteriores da rede já foram obtidos aleatoriamente.

De forma geral, uma alternativa para representar a função $F(X)$ é através da resolução de um sistema de equações lineares, implicando em encontrar os pesos da matriz

β que sejam capazes de realizar o mapeamento da entrada para a saída. Contudo, em tal mapeamento os pesos da camada de saída são desconhecidos ou podem não existir. Neste sentido, uma solução é utilizar o método de mínimos quadrados para o sistema linear e encontrar a matriz β:

$$\beta = \mathbf{I}^{\dagger} \times T^{\dagger} \tag{2.34}$$

onde, $\mathbf{I}$ é a matriz inversa generalizada de Moore-Penrose da matriz $\mathbf{H}$. O $\dagger$ significa transposto.

O Algoritmo 1 ilustra o treinamento da rede neural via *Extreme Learning Machine*, ou seja, um pseudo-código do conteúdo apresentado na Seção 2.6. O algoritmo basicamente recebe os padrões de entrada X e saída T, em seguida inicializa os pesos da camada escondida W e *bias* B, e calcular a matriz $\mathbf{H}$ conforme Equação 2.33. Posteriormente, calcula-se os pesos da camada de saída da rede neural através do produto da matriz inversa com a matriz de saída, conforme a Equação 2.34. Cabe ressaltar que a base de dados com a série temporal *db* pode ser um vetor, matriz ou alguma outra estrutura de dados. Conforme mencionado anteriormente a normalização apresentada no algoritmo é importante para melhor ajustar os parâmetros da rede e as funções para arranjar os dados servem para organizar os dados como mostra a Equação 2.31.

2.6.3 Generalização

Na fase de generalização da Máquina de Aprendizado Extremo, o principal objetivo consiste em inferir sobre um determinado problema, sem o prévio conhecimento da rede neural artificial sobre um conjunto de entrada. Neste sentido, para a rede inferir sobre um conjunto de entradas desconhecidas, é necessário que anteriormente já tenha sido calculado os *bias*, e os pesos da camada escondida e de saída.

Primeiramente, os padrões de entrada são normalizados e arranjados na matriz de entrada X, em seguida o vetor *bias* é replicado para formar a matriz $\mathbf{B}$. Posteriormente, a matriz $\mathbf{H}$ é calculada conforme:

$$H = \varphi(B + W \times X) = \varphi(tempH) \tag{2.35}$$

onde, $tempH = B + W \times X$.

Por fim, a rede neural via ELM é capaz de inferir sob um conjunto de entradas desconhecidas através da seguinte equação:

$$\hat{u} \leftarrow H^{\dagger} \times \beta \tag{2.36}$$

Entrada : $db \equiv$ base de dados com as séries temporais;
$dbTarget \equiv$ são as saídas desejadas;
$n \equiv$ número de padrões de entrada da rede;
$m \equiv$ quantidade de neurônios da camada escondida;
$p \equiv$ quantidade de atributos (características) do padrão de entrada.
Saída : $\beta \equiv$ matriz obtida via Equação 2.34;
$W \equiv$ matriz de pesos da camada escondida;
$bias \equiv$ vetor de *bias*.
Início
$db \leftarrow$ `NormalizarVetorDados` (db) ;
$X \leftarrow$ `ArranjarMatrizEntrada` (db) ;
$T \leftarrow$ `ArranjarMatrizSaida` $(dbTarget)$;
$W \leftarrow$ `InicializarMatrizPesos` (p,m) ;
$bias \leftarrow$ `InicializarVetorBias` (m) ;
$B \leftarrow$ `Replicar` $(bias, n)$;
$tempH \leftarrow W \times X$;
$tempH \leftarrow tempH + B$;
$H \leftarrow \varphi(tempH)$;
$I \leftarrow$ `PseudoInversa` (H) ;
$\beta \leftarrow I^{\dagger} \times T^{\dagger}$;
`retorna`$(\beta, W, bias)$;
Fim

Algoritmo 1: Algoritmo de treinamento para uma rede neural artificial do tipo SLFN através do método *Extreme Learning Machine*.

onde, β já foi previamente ajustado na fase de treinamento da rede e $\hat{u}$ é a matriz com os valores inferidos sobre a matriz de entrada X.

O Algoritmo 2 apresenta a metodologia de generalização de novos padrões a partir de um conjunto de entrada desconhecidas pela rede neural, conforme mostrada nesta seção. Assim, o algoritmo basicamente recebe como entrada os novos padrões, desconhecidos a *priori* pela rede, bem como os pesos da camada escondida (W) e de saída (β). A seguir, o algoritmo normaliza os padrões de entrada e os arranja na matriz X, enquanto o *bias* é replicado para matriz **B**. Posteriormente, calcula-se a matriz **H** tal como apresenta a Equação 2.35, em seguida é possível inferir sobre um novo padrão de entrada através do produto da transposta da matriz **H** com a matriz β, conforme Equação 2.36.

2.7 Cópulas

Historicamente, Sklar (1959) *apud* Nelsen (2006) propôs uma distribuição acumulada conjunta multivariada composta por k-distribuições que chamou de Cópula. Assim, cópulas são funções de distribuição acumulada multivariada em que seus argumentos são distribuições acumuladas marginais univariadas de k variáveis aleatórias.

Entrada : $db \equiv$ base de dados com as séries temporais;
$n \equiv$ número de padrões de entrada da rede;
$\beta \equiv$ matriz obtida via Equação 2.34;
$W \equiv$ matriz de pesos da camada escondida;
$bias \equiv$ vetor de *bias*.
Saída : $\hat{u} \equiv$ matriz com as previsões estimadas para db.
Início
$db \leftarrow$ `NormalizarVetorDados` (db) ;
$X \leftarrow$ `ArranjarMatrizEntrada` (db) ;
$B \leftarrow$ `Replicar` $(bias, n)$;
$tempH \leftarrow W \times X$;
$tempH \leftarrow tempH + B$;
$H \leftarrow \varphi(tempH)$;
$\hat{u} \leftarrow H^{\dagger} \times \beta$;
retorna $\hat{u}$;
Fim

Algoritmo 2: Algoritmo de teste para uma rede neural artificial do tipo SLFN através do método *Extreme Learning Machine*.

Desta forma, uma variável aleatória pode ser considerada como um valor único ao espaço amostral que é determinado aleatoriamente por uma função que associa cada resultado do experimento realizado. De maneira geral, o valor de uma variável aleatória só é conhecido após a realização de um experimento, como por exemplo, o número de caras obtido após o lançamento de duas moedas (BARBETTA; REIS; BORNIA, 2010).

Quando se trata de variáveis quantitativas aleatórias, entende-se que as variáveis quantitativas podem assumir valores discretos (números inteiros) ou contínuos (números reais). Neste sentido, quando os valores da variável aleatória são discretos é dito que a distribuição de probabilidade de uma variável aleatória é a probabilidade de cada valor de um evento qualquer ocorrer. Por outro lado, quando a variável é contínua, a distribuição de probabilidade é calculada a partir de uma função de densidade de probabilidade (BARBETTA; REIS; BORNIA, 2010). Então, sejam Y e X duas variáveis aleatórias contínuas e sejam $F(y) = P(Y \leq y)$ e $G(x) = P(X \leq x)$ as funções de distribuição acumulada marginal univariada de Y e X, respectivamente. Ressaltando que os valores das distribuições marginais estão no intervalo [0,1].

2.7.1 Conceitos

Este método de combinação visa agregar as distribuições marginais, de tal maneira que a distribuição acumulada multivariada tenha todas as informações contidas nas distribuições marginais, isto é, com o mínimo de perda de informação.

Assim, a proposta de Sklar (1959) *apud* Nelsen (2006) visa apresentar uma função $\mathcal{C}(\cdot)$ que através de um conjunto de tamanho qualquer de distribuições acumuladas

marginais univariadas (como $F(y)$ e $G(x)$, por exemplo) possam ser combinadas em uma distribuição acumulada multivariada (YAGER, 2016):

$$H(y, x) = \mathcal{C}(F(y), G(x)) \tag{2.37}$$

para os casos que não sejam bivariados pode ser representada a Equação 2.37 como um conjunto de k variáveis aleatórias $\mathbf{Y}$ escrito da seguinte forma:

$$F_Y(y_1, y_2, \cdots, y_k) = C(F_{Y_1}(y_1), \cdots, F_{Y_i}(y_i), \cdots, F_{Y_k}(y_k)) \tag{2.38}$$

onde, $F_Y(\cdot)$ é a distribuição acumulada multivariada e $C(\cdot)$ é a função de cópulas. As Equações 2.37 e 2.38 são fundamentadas através do teorema de Sklar (1959) que afirma:

Teorema 1. *Seja* $\mathbf{Y} = (Y_1, \cdots, Y_i, \cdots, Y_k)$ *um vetor de variáveis aleatórias com função de distribuição acumulada multivariada* $F_{\mathbf{Y}}$ *e função de distribuição acumulada univariada (ou também conhecida como distribuição de probabilidade marginal univariada)* $F_{Y_i}(y_i)$*, sendo* $i = 1, 2, \cdots, k$*. Portanto, para todo* $\boldsymbol{y} = (y_1, \cdots, y_k) \in \mathbb{R}^k$*. Então, a cópula* C *é uma função de distribuição acumulada* k*-dimensional, tal que esta é escrita da seguinte forma:*

$$\begin{aligned} F_Y(y_1, y_2, \cdots, y_k) &= C(F_{Y_1}(y_1), \cdots, F_{Y_i}(y_i), \cdots, F_{Y_k}(y_k)) \\ &= C(v_1, \cdots, v_i, \cdots, v_k) \\ &= P(Y_1 \leq y_1, \cdots, Y_i \leq y_i, \cdots, Y_k \leq y_k) \end{aligned} \tag{2.39}$$

onde, $v_1, \cdots, v_i, \cdots, v_k$ *são instâncias de* k *distribuições de probabilidade marginal univariada e* $v_i \in [0, 1]$*. Assim, as variáveis aleatórias* $Y_1, Y2, \cdots, Y_k$ *são mapeadas para* $v_1, v_2, \cdots, v_k$ *através da seguinte equação:*

$$v_i = F_{Y_i}(y_i). \tag{2.40}$$

■

As distribuições marginais univariadas de entrada da cópula, conforme mencionado anteriormente, estão no intervalo [0,1], por isso $v_i \in [0, 1]$ tal que $(i, \cdots, k)$. Portanto, v_i representa o valor da distribuição de probabilidade marginal univariada dada pela função $F_{Y_i}(\cdot)$ de uma instância qualquer y_i que será o argumento de entrada da função de cópulas. Desta maneira, existe um mapeamento entre os valores das variáveis aleatórias $(y_1, \cdots, y_i, \cdots, y_k)$ para os valores das distribuições marginais de $(v_1, \cdots, v_i, \cdots, v_k)$ através das Equação 2.40.

A função de densidade de probabilidade multivariada das variáveis aleatórias em $\mathbf{Y}$ pode ser descrita da seguinte forma:

$$\begin{aligned} p_{\mathbf{Y}}(y_1, \cdots, y_i, \cdots, y_k) = \frac{\partial^k F_{\mathbf{Y}}(y_1, \cdots, y_k)}{\partial y_1 \cdots \partial y_k} = \\ \frac{\partial^k \mathcal{C}(F_{Y_1}(y_1), \cdots, F_{Y_k}(y_k))}{\partial y_1 \cdots \partial y_k} = \frac{\partial^k \mathcal{C}(v_1, \cdots, v_k)}{\partial v_1 \cdots \partial v_k} \times \prod_{i=1}^{k} \frac{\partial v_i}{\partial y_i} = \\ c(F_{Y_1}(y_1), \cdots, F_{Y_k}(y_k)) \prod_{i=1}^{k} p_{Y_i}(y_i) \end{aligned} \tag{2.41}$$

tal que,

$$p_{Y_i}(y_i) = \frac{\partial F_{Y_i}(y_i)}{\partial y_i} \tag{2.42}$$

e

$$c(v_1, \cdots, v_k) = \frac{\partial^k \mathcal{C}(v_1, \cdots, v_k)}{\partial v_1 \cdots \partial v_k} \tag{2.43}$$

onde, $p_{\mathbf{Y}}(y_1, \cdots, y_i, \cdots, y_k)$ é a função de densidade de probabilidade multivariada a partir de $\mathbf{Y}$, enquanto $c(\cdot)$ é a função de densidade da cópula e $p_{Y_i}(y_i)$ é a função de densidade de probabilidade marginal univariada de Y_i. Nas próximas seções serão apresentadas algumas das diferentes cópulas existentes na atualidade, como Gumbel-Hougaard, Clayton, Normal e Cacoullos, bem como suas respectivas funções de densidades de probabilidade que poderão ser utilizadas na Equação 2.43.

Atualmente, existem diversas famílias e tipos diferentes de cópulas presentes na literatura (ASSIS, 2016; OLIVEIRA, 2014; LEAL, 2010). A diversidade de cópulas se dá pelo fato de cada tipo apresentar comportamentos diferentes, ou seja, existem cópulas com maior capacidade de captar a dependência em um dos lados da distribuição, isto é, quando tal dependência é positiva ou negativa na cauda da distribuição. Enquanto que existem outros tipos de cópulas que são capazes de capturar a dependência tanto negativa quanto positiva. Existem ainda as cópulas paramétricas e não paramétricas. As cópulas paramétricas possuem um ou vários parâmetros conhecidos como parâmetro de dependência da cópula que é responsável por medir o grau de associação entre as marginais, o valor do parâmetro é estabelecido por uma restrição matemática previamente estabelecida pela cópula (ASSIS, 2016; OLIVEIRA, 2014).

Uma das famílias de cópulas mais conhecidas são as arquimedianas que abrangem uma grande variedade de estruturas de dependência como linear, exponencial, parabólica, entre outras. De modo geral, as cópulas arquimedianas são muito conhecidas pela capacidade de captar dependência caudal assimétrica, sendo uma das propriedades a favor quando aplicada em modelagens de dados com estrutura de dependência assimétrica. Dentre as diversas cópulas existentes da família arquimediana, destacam-se as cópulas de Gumbel-Hougaard, Clayton e Frank. Estas possuem por definição, apenas um

parâmetro de dependência, o que torna essas cópulas razoavelmente simples de serem aplicadas (OLIVEIRA, 2014; ASSIS, 2016).

2.7.2 Cópula Normal

A cópula Normal (ou Gaussiana) é membro da família das cópulas Elípticas. Esta é simétrica e pode capturar não apenas a dependência positiva mas também a negativa nas caudas da distribuição, ao contrário do que ocorre com as cópulas de Gumbel-Hougaard e Clayton. Neste sentido, esta cópula é capaz de modelar estruturas de dependência em ambas as extremidade da distribuição, ou seja, na cauda positiva e negativa. A cópula Normal pode ser escrita da seguinte forma:

$$\mathcal{C}(v_1, \cdots, v_k) = \Phi(\varphi^{-1}(v_1), \cdots, \varphi^{-1}(v_k)|\rho), \tag{2.44}$$

onde $\varphi^{-1}(\cdot)$ é a função inversa da distribuição acumulada unitária da distribuição Normal e $\Phi(\cdot)$ é a função de distribuição acumulada de uma distribuição Normal multivariada com vetor de média zero e matriz de covariância igual para ρ.

A função de densidade de probabilidade da cópula Normal é:

$$c(v_1, \cdots, v_k) = c(\varphi^{-1}(v_1), \cdots, \varphi^{-1}(v_k)|\rho) = \frac{1}{\sqrt{|\rho|}} \times exp\left(-\frac{1}{2}(\varphi^{-1}(v_1), \cdots, \varphi^{-1}(v_k))(\rho^{-1} - \mathbb{I}) \begin{pmatrix} \varphi^{-1}(v_1) \\ \vdots \\ \varphi^{-1}(v_k) \end{pmatrix} \right) \tag{2.45}$$

onde $\mathbb{I}$ é a matriz identidade, $\varphi^{-1}(\cdot)$ é a inversa de v_j de acordo com uma distribuição Normal e ρ é a matriz de covariância. A matriz de covariância pode ser obtida através da fórmula de correlação de Person.

Neste caso, $\rho \in [-1, 1]$, onde $\rho_{i,j} = 0$ indicando independência entre as variáveis do índice i e j, $\rho_{i,j} = -1$ implicando na perfeita dependência negativa, e $\rho_{i,j} = 1$ reflete na perfeita dependência positiva (RENARD; LANG, 2007; AAS, 2004).

2.7.3 Cópula de Cacoullos

A cópula de Cacoullos proposta por Oliveira et al. (2018) é uma abordagem semi-paramétrica para $c(\cdot)$. A função de densidade de probabilidade de Cacoullos faz uso do kernel multivariado proposto por Cacoullos (1964). Este autor mostrou como estender os resultados de Parzen (1962) para estimar funções de densidade de probabilidade

multivariada. Parzen trabalhou em casos univariados envolvendo funções de kernel enquanto Cacoullos introduziu a modelagem multivariada pelo produto das funções de kernel univariados.

Posteriormente, Specht (1990) propõe um novo modelo de Rede Neural, intitulada como Rede Neural Probabilística que é baseada na função multivariada de Cacoullos. Neste sentido, a função de ativação da Rede Neural Probabilística trata-se da equação de kernel Gaussiano multivariado. Em particular, para os casos onde o kernel é Gaussiano o estimador multivariado pode ser expresso por (SPECHT, 1990):

$$f(\mathbf{T}) = \frac{1}{(2\pi)^{k_f/2}\lambda^{k_f}} \frac{1}{n_f} \sum_{i=1}^{n_f} \exp\left[-\frac{(\mathbf{T}-\mathbf{T_i})^{\dagger}(\mathbf{T}-\mathbf{T_i})}{2\lambda^2}\right] \tag{2.46}$$

onde,
$i \equiv$ índice do padrão;
$n_f \equiv$ número total de padrões de treinamento;
$T_i \equiv i$-ésimo padrão de treinamento em T;
$\lambda \equiv$ parâmetro de suavização;
$k_f \equiv$ dimensionalidade do espaço de medição.

Desta forma, a função de densidade de probabilidade da cópula de Cacoullos é uma adaptação da função de ativação apresentada por Specht (1990):

$$c(v_1, \cdots, v_k) = \frac{1}{2\pi^{k/2}\lambda^k} \frac{1}{n} \sum_{t=1}^{n} \exp\left[-\sum_{i=1}^{k} \frac{(v_i - F_{Y_i}(y_{ti}|\hat{\alpha}_i))^2}{2\lambda^2}\right] \tag{2.47}$$

onde,
$k \equiv$ é o número de variáveis aleatórias;
$n \equiv$ é o tamanho do conjunto de treinamento;
$\lambda \equiv$ parâmetro da cópula;
$y_{ti} \equiv t$-ésima observação da variável Y_i do conjunto de treinamento;
$F_{Y_i}(\cdot) \equiv$ é a função de distribuição acumulada univariada da variável Y_i obtida do conjunto de treinamento;
$\hat{\alpha}_i \equiv$ vetor de parâmetros de $F_{Y_i}(\cdot)$, estimado a partir do conjunto de treinamento;
$v_i \equiv$ é uma instância da distribuição acumulada univariada do conjunto de teste.

O parâmetro da cópula de Cacoullos é $\lambda > 0$. Specht (1966) *apud* Specht (1990) discute como estimar o parâmetro envolvendo problemas de classificação. Os autores notam que na prática não é difícil encontrar um bom valor para o parâmetro. O trabalho de (OLIVEIRA et al., 2018) apresenta uma abordagem para estimar o melhor valor para λ através do método de mínimos quadrados. Este método visa encontrar o melhor ajuste

para um conjunto de dados pela minimização da soma dos quadrados da diferença entre o valor estimado e o respectivo valor observado.

2.7.4 Método de Estimação do Parâmetro de Cópulas

Uma das principais alternativas de estimação dos parâmetros de cópula é o método conhecido como inferência de marginais ou do inglês *Inference Function for Margins* (IFM). Este método sugere que os parâmetros sejam estimados em duas etapas, isto é, os parâmetros da distribuição de probabilidade acumulada são estimados separadamente para cada modelo. Posteriormente, o IFM sugere que sejam obtidos os parâmetros de dependência da cópula, isto é, θ representando o parâmetro para as cópulas Gumbel-Hougaard, Frank e Clayton, enquanto ρ é dito o parâmetro de dependência da cópula Normal. Maiores detalhes sobre o método IFM podem ser encontrados em (JOE; XU, 1996).

Outra técnica de estimação que vem sendo discutida na literatura (OLIVEIRA, 2014) é o método de máxima verossimilhança (MMV) tipicamente utilizado para estimar tanto os parâmetros das distribuições de probabilidade acumulada quanto o parâmetro de dependência da cópula. Diferentemente do método IFM, o MMV estima simultaneamente os parâmetros das distribuições marginais e da cópula. Neste sentido, o MMV visa maximizar o resultado da função de densidade de probabilidade multivariada dada pela Equação 2.41. Assim, o método MMV pode ser descrito da seguinte forma:

$$\begin{aligned}\text{MMV} &= \sum_{t=1}^{n} ln\left(p_{Y_1,\cdots,Y_k}(y_{t,1},\cdots,y_{t,i},\cdots,y_{t,k}|\alpha_1,\cdots,\alpha_i,\cdots,\alpha_k,\gamma)\right) \\ &= \sum_{t=1}^{n} ln\left(c(F_{Y_1}(y_{t,1}|\alpha_1),\cdots,F_{Y_k}(y_{t,k}|\alpha_k)|\gamma)\right) + \sum_{t=1}^{n}\sum_{i=1}^{k} ln\left(p_{Y_i}(y_{t,i}|\alpha_i)\right) \qquad (2.48)\end{aligned}$$

onde α_i é o vetor que armazena o conjunto de parâmetros da distribuição de probabilidade acumulada $F_{Y_i}(y_i)$ e γ representa o vetor com os parâmetros de dependência da cópula. Neste sentido, θ representa o parâmetro de dependência das cópulas Gumbel-Hougaard, Clayton e Frank, enquanto que ρ indica o parâmetro da cópula Normal, por exemplo. Já o vetor $\mathbf{y}_t = (y_{t,1}, ..., y_{t,i}, ..., y_{t,k})$ representa a t-ésima observação do vetor de variáveis aleatórias $\mathbf{Y}$. Assim, o método MMV, estima simultaneamente os parâmetros γ e α_i, de modo a maximizar a função MMV da Equação 2.48. A estimação simultânea torna a computação para obter os parâmetros corretos uma tarefa bastante complexa (LEAL, 2010).

2.8 Resumo do Capítulo

Este capítulo mostrou os conceitos básicos de séries temporais, bem como as principais métricas para avaliar o desempenho dos modelos individuais. Também, foram abordadas as técnicas para medir o coeficiente de correlação entre duas variáveis para analisar os experimentos realizados neste trabalho com interesse de investigar o grau de associação entre o erro cometido pelo método que será proposto no próximo capítulo e a quantidade de modelos individuais.

Foram abordadas algumas das principais técnicas de combinação de previsões de séries temporais encontradas na literatura. As redes neurais artificiais estão entre as metodologias mais conhecidas de previsão de séries temporais, e estas foram adotadas juntamente com o algoritmo de aprendizado ELM, por sua capacidade de generalização e custo computacional. As RNAs são responsáveis por produzir as previsões que serão utilizadas na combinação. Finalmente, a metodologia de agregação via cópulas é descrita para diferentes tipos de cópulas.

3 MÉTODO PROPOSTO

Neste capítulo é apresentado um *ensemble* para combinar vários modelos individuais de previsão de séries temporais por meio de cópulas. O método proposto é introduzido inicialmente expondo sua arquitetura, em seguida, são definidos os conjuntos de treinamento e teste. Posteriormente, são apresentadas as propriedades das redes neurais artificiais utilizadas neste trabalho, bem como o procedimento proposto para calcular os erros de previsão e ajuste da cópula. E por fim, é formulado o processo utilizado para combinar as previsões obtidas pelas redes neurais através de cópulas.

3.1 Contextualização

A literatura mostra que é possível combinar metodologias pré-estabelecidas, isto é, combinar modelos individuais como redes neurais artificiais, ou ainda, os autorregressivos integrados de médias móveis (*Autoregressive Integrated Moving Average* - ARIMA) (BOX; JENKINS; REINSEL, 1994), por meio de alguma metodologia de combinação, como média simples, média ponderada e várias outras, já discutidas. Neste sentido, o método proposto, visa utilizar redes neurais artificiais como estratégia para prever determinada série temporal de interesse. E posteriormente, as previsões realizadas pelos preditores serão combinadas via cópulas. Para o método proposto, especificamente, deseja-se investigar como se comporta quando recebe várias distribuições de probabilidade marginal unitária, em outras palavras, inúmeros modelos individuais e verificar o que ocorre com o erro de previsão na medida em que estes modelos são agregados no processo combinatório.

A arquitetura do método proposto se divide em duas partes (treinamento e combinação). Sendo a primeira parte destinada ao treinamento dos modelos individuais, assim esta etapa visa construir os modelos individuais e ajustar a cópula. Enquanto que a segunda parte tem a finalidade de combinar as previsões.

3.2 Arquitetura do Método Proposto

A Figura 3 ilustra a parte de treinamento do método proposto. Primeiramente, a abordagem proposta cria um conjunto finito de preditores. Os preditores criados são baseados nas Redes Neurais Artificiais (RNA) e utiliza o algoritmo de treinamento Máquinas de Aprendizagem Extrema (*Extreme Learning Machine* - ELM). Cada RNA criada pelo método proposto é ajustada com parâmetros diferentes (pesos e número de neurônios na camada escondida). De modo, a buscar a diversidade dos preditores (maiores detalhes sobre a diversidade das RNAs via ELM pode ser encontrada em Huang, Zhu e Siew (2006)).

A figura apresenta as etapas para obtenção dos preditores. Desta forma, o método extrai da série temporal de interesse um conjunto de observações, ou seja, é selecionado um intervalo de observações da série temporal que serão utilizadas para treinar e testar as RNAs. Após a etapa de treinamento das RNAs, estas são denominadas como preditores ou modelos individuais. Assim, sejam $\mathrm{RNA}_1, \cdots, \mathrm{RNA}_i, \cdots, \mathrm{RNA}_k$ as redes neurais treinadas e $x_{t,1}, \cdots, x_{t,i}, \cdots, x_{t,k}$ as previsões das RNAs para o conjunto de treinamento da série temporal observada $(u_t, t = 1, \cdots, n)$. Cabe enfatizar que os k preditores fazem a previsão da mesma série temporal.

As previsões das RNAs são dadas por $x_{t,i}$ no instante t, $t = 1, \cdots, n$ para a i-ésima RNA (onde, $i = 1, \cdots, k$) o método proposto calcula o erro de previsão para cada RNA, denominado por $e_{t,1}, \cdots, e_{t,i}, \cdots, e_{t,k}$. Os erros são utilizados no processo de ajuste da cópula, visto que por meio do erro pode-se obter diversas informações valiosas como a variância, média, correlação dos erros, entre outras informações que contribuem para a combinação das RNAs.

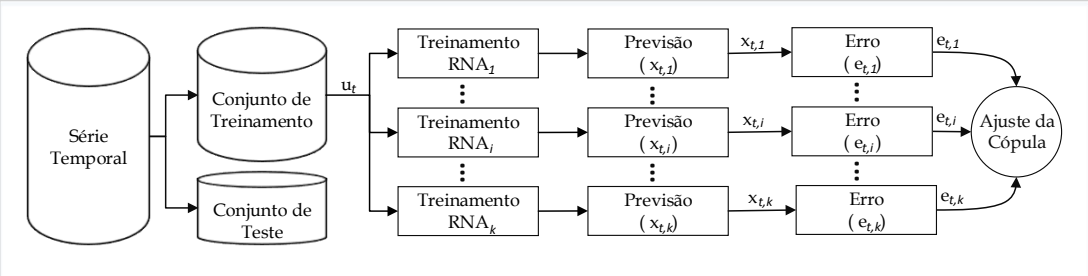


Figura 3 – Treinamento do método proposto. Inicialmente a série é dividida em dois conjuntos (treinamento e teste), as observações exclusivamente do conjunto de treinamento são utilizadas para treinar as RNAs e prever a série. Posteriormente, os erros de previsão são calculados e usados para ajustar a cópula.

A Figura 4 apresenta a parte de combinação da arquitetura do método proposto, aonde as RNAs já estão previamente treinadas e prontas para preverem a série temporal a ser estudada. Neste sentido, observe que os preditores (isto é, as RNAs) recebem como entrada as observações do conjunto de teste da série temporal. Esse conjunto foi previamente dividido na primeira parte da arquitetura. Em seguida, os k preditores realizam as previsões e passam para a etapa de calcular os erros dos preditores. Neste sentido, o erro é a diferença entre o valor predito (conjunto de teste) e um valor esperado da série temporal (os detalhes de como obter este valor são descritos na Seção 3.7.1). Posteriormente, os erros são modelados em uma distribuição de probabilidade marginal acumulada, ou seja, os erros são mapeados para a distribuição acumulada marginal via Função de Distribuição Acumulada (FDA). Finalmente, as distribuições acumuladas dos erros são combinadas através da função de densidade de probabilidade conjunta de cópulas $c(v_1, \cdot, v_i, \cdots, v_k)$, esta ajustada anteriormente na primeira parte da arquitetura com os

dados do conjunto de treinamento. Por fim, $\hat{u}_t$ é estimado por meio, por exemplo, do método de máxima verossimilhança ou mínima variância. Todos os passos apresentados nesta arquitetura são descritos em detalhes nas seções a seguir.

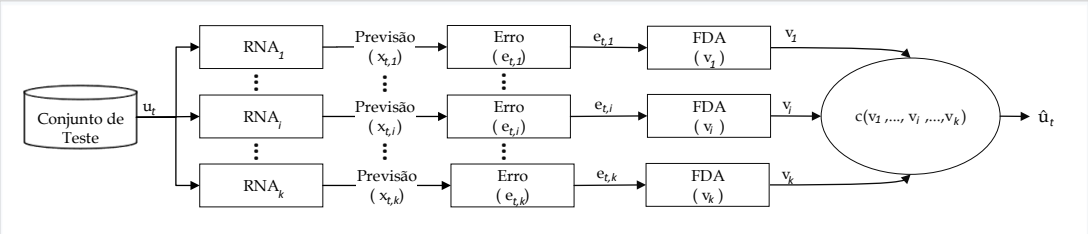


Figura 4 – Combinação via método proposto. As observações u_t do conjunto de teste são a entrada para as RNAs preverem a série e posteriormente os erros são calculados e modelados via FDA. Finalmente, $v_1, \cdots, v_k$ são agregados por meio de cópula para obter $\hat{u}_t$.

3.3 Divisão das Observações da Série Temporal

Em problemas de previsão de séries temporais são muitas as metodologias tradicionais como Oliveira et al. (2018), Firmino, Neto e Ferreira (2014), Oliveira et al. (2017), Ferreira, Vasconcelos e Adeodato (2008) que dividem os dados em apenas dois conjuntos (treinamento e teste), cuja finalidade dos dados de treinamento é ajustar o modelo de predição, ou seja, esses dados são utilizados para encontrar os parâmetros do modelo que o torne eficiente e acurado. Por outro lado, o conjunto de teste é aplicado para avaliar a qualidade do modelo previamente treinado. Contudo, em aprendizagem de máquina alguns autores tem buscado utilizar uma terceira parte dos dados, conhecido por conjunto de validação. Essa técnica visa ajustar os parâmetros para produzir medidas acuradas pelo modelo. Basicamente, essa técnica divide o conjunto de dados em três partes: treinamento, validação e teste. Assim, o conjunto de validação é utilizado para avaliar se o modelo não está super ajustado, em outras palavras, se o modelo apresenta generalizações acuradas para novos padrões de entrada.

O conjunto de validação levanta preocupações teóricas, pois os dados possuem dependência e podem ocorrer efeitos que evoluem ao longo do tempo. Neste sentido, ao remover parte da série temporal para utiliza-lá no processo de validação do modelo, teoricamente causa um efeito de deterioração da dependência da série por interromper a evolução dos valores observados ao longo do tempo. Bergmeir e Benítez (2012) trata deste problema teórico, apresentando um estudo com séries temporais envolvendo duas técnicas de validação cruzada, entre elas o *n-folds cross-validation* aplicada em diversos trabalhos de aprendizagem de máquina. O autor relata que não encontrou nenhum problema prático com a validação cruzada, bem como os resultados não apontaram nenhum efeito da dependência

em relação a evolução do tempo. Porém, cabe ressaltar que o estudo apresentado pelo autor se limitou apenas as séries temporais estacionárias.

Considerando o problema teórico apresentado e a larga utilização tradicional da divisão da série temporal em apenas dois conjuntos (treinamento e teste), e ainda, que Huang, Zhu e Siew (2006) ao proporem o algoritmo de aprendizagem ELM, não preveem a necessidade de utilização da técnica de validação para melhorar a acurácia dos parâmetros da rede. E este trabalho não se limita a estudar a acurácia dos modelos individuais de previsão e considerando ainda, que a técnica de validação não se aplica ao processo de combinação via cópula, uma vez que os parâmetros (média, desvio padrão e matriz de covariância) em termos estatísticos para serem estimados com alta veracidade requerem o maior número disponível de observações para estimar os parâmetros. Assim, esse trabalho adotou apenas a divisão tradicional da série em duas partes.

Logo, as séries temporais de entrada do método proposto são divididas em dois conjuntos: treinamento e teste. A parte utilizada para ajustar os pesos (treinamento) é formada por n pontos da série. Enquanto que a parte utilizada para testar os modelos individuais é composta por m observações. Assim, respectivamente, podem ser propostos n pontos da série para treinamento dos modelos individuais e m pontos para teste. Cabe ressaltar que o tamanho total da série é representado por $n + m$.

3.4 Redes Neurais Artificias (Preditores Individuais)

Sejam u_t o valor de uma dada série temporal no instante t e $\mathbf{X}_t = (X_{t,1}, \cdots, X_{t,i}, \cdots, X_{t,k})$ o vetor com k preditores individuais (RNAs) para u_t. Assim, seja $\mathbf{x}_t = (x_{t,1}, \cdots, x_{t,i}, \cdots, x_{t,k})$ o vetor com as respectivas previsões dos preditores individuais. Neste trabalho, o conjunto de treinamento utilizado para ajustar o *ensemble* baseado em cópulas é $\{u_t, \mathbf{x}_t\}_{t=1}^{n}$ e para o conjunto de teste é $\{u_t, \mathbf{x}_t\}_{t=n+1}^{n+m}$, tal que n é o número de observações do conjunto de treinamento e m a quantidade do conjunto de teste. Desta forma, $\{u_t, \mathbf{x}_t\}_{t=n+1}^{n+m}$ poderia ser reescrito da seguinte forma $\{u_t, x_{t,1}, \cdots, x_{t,i}, \cdots, x_{t,k}\}_{t=n+1}^{n+m}$.

3.4.1 Camada Escondida

Os modelos individuais de previsão de séries temporais foram construídos utilizando redes neurais artificiais com algoritmo de treinamento ELM que por padrão sua arquitetura possui uma única camada escondida. As RNAs adotadas utilizam a função de ativação sigmóide por ser uma função amplamente utilizada no ramo da aprendizagem de máquina, além de possuir características que a tornam suave e continuamente diferenciável e principalmente, pela função ser não linear variando sua forma em formato de "S". O que é mais sensato uma vez que os neurônios biológicos atuam de forma binária (ativado ou não ativado) a sigmóide se mostra como uma boa opção pelo fato de sua saída apresentar valores entre 0

e 1, representando a ativação ou não ativação do neurônio. Por outro lado, também seria possível utilizar outras funções de ativação como a tangente hiperbólica, por exemplo, que é similar a sigmóide possuindo forma de "S", mas apresenta saída com valores entre -1 e 1 ao invés de 0 e 1 o que pode ser mais vantajoso em certas situações.

Neste método, podem ser gerados k modelos individuais, isto é, k redes neurais. A generalização dos modelos individuais foi garantida através da geração aleatória dos pesos da camada de entrada, bem como as arquiteturas das redes são selecionadas aleatoriamente. Assim, as redes podem assumir uma arquitetura entre 1 e L neurônios na camada escondida. A diversidade dos modelos individuais também foi buscada por meio do número de pontos das séries temporais estudadas e o tamanho da janela de treinamento (lag). A janela (lag) de entrada das RNAs foi previamente selecionada para cada série temporal de interesse por meio da função de autocorrelação. Cabe ressaltar, que não foi proposto um lag variável para criar os modelos individuais, visto que o foco do trabalho é combinar os modelos independente de sua qualidade.

3.5 Calculando os Erros dos Preditores

Conforme ilustrado na Figura 4, a etapa de calcular os erros dos k preditores, consiste em calcular o vetor de erros $\mathbf{E}_t = (E_{t,1}, \cdots, E_{t,i}, \cdots, E_{t,k})$ tal que t é o índice da observação da série temporal e k é a quantidade de modelos individuais. Desta forma, $\mathbf{E}_t$ é o vetor de erros dos modelos individuais $\mathbf{X}_t$ que pode ser dado pelo erro aditivo ou multiplicativo, por exemplo. Neste sentido, o erro aditivo é dado pela equação:

$$\mathbf{E}_t = \mathbf{X}_t - u_t \tag{3.1}$$

ou

$$E_{t,i} = X_{t,i} - u_t. \tag{3.2}$$

De maneira geral, após a construção dos modelos individuais de previsão $\mathbf{X}_t$, pode ser reescrito o cálculo dos erros de previsão utilizando as instâncias de $\{x_t\}_{t=1}^{n}$ e $\{u_t\}_{t=1}^{n}$ através da seguinte equação:

$$e_{t,i} = x_{t,i} - u_t \tag{3.3}$$

assim, $\mathbf{e}_t = (e_{t,1}, \cdots, e_{t,i}, \cdots, e_{t,k})$ é o vetor de erros de $\mathbf{x}_t$.

3.6 Ajuste da Cópula

Para a cópula preceder com a combinação é necessário que ela seja ajustada com alguns parâmetros. Assim, o primeiro passo é identificar qual a distribuição de probabilidade para cada $\mathbf{E}_i = (E_{1,i}, \cdots, E_{t,i}, \cdots, E_{n,i})$. Essa identificação pode ser realizada no conjunto de treinamento por meio de teste estatístico conhecido como teste de aderência (CHWIF;

MEDINA, 2014). Este teste é capaz de verificar se a distribuição é normal, por exemplo, é estatisticamente adequada para representar os erros coletados. Assim, para cada $\mathbf{E}_i$ pode ser verificada a distribuição de probabilidade que melhor se ajusta ao erro.

Identificada a distribuição de probabilidade do erro (normal, exponencial, logarítmica, entre outras), se faz necessário estimar para cada distribuição seus parâmetros. Logo, seja $\hat{\alpha}_i$ o vetor de parâmetros devidamente estimado para a distribuição de probabilidade do $\mathbf{E}_i$.

Portanto, $e_{t=8,i=3}$ é o erro da oitava observação da série temporal cometido pelo terceiro modelo individual. Após a abordagem proposta computar os erros dos modelos individuais, é estimada a distribuição de probabilidade acumulada marginal dos erros através da função:

$$v_i = F_{E_i}(e_{t,i}|\alpha_i), \tag{3.4}$$

logo, $v_i = (v_1.\cdots, v_i, \cdots, v_k)$ é o vetor que contém a distribuição acumulada marginal dos erros do i-ésimo $\mathbf{E}_i$ que por sua vez as instâncias de $\mathbf{E}_i$ são dadas por $\mathbf{e}_i = (e_{1,i}, \cdots, e_{2,i}, \cdots, e_{n,i})$, α_i é o vetor de parâmetros da distribuição. Neste contexto, para uma distribuição normal de probabilidade, por exemplo, o vetor α_i possuiria duas posições, visto que a distribuição normal é composta por dois parâmetros, sendo a média e o desvio padrão. Em termos estatísticos, é necessário estimar os parâmetros das distribuições. Especificamente, para distribuição normal os parâmetros $\hat{\alpha}_i$ podem ser estimados da seguinte forma:

$$\mu_{E_i} = \frac{1}{n}\sum_{t=1}^{n} e_{t,i}, \tag{3.5}$$

onde, μ_{E_i} é a média dos erros $\mathbf{E}_i$. Enquanto que:

$$\sigma_{E_i} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(e_{t,i} - \mu_{e_{t,i}})^2}, \tag{3.6}$$

isto é, σ_{E_i} representa a estimativa do desvio padrão do erro. Assim, $\hat{\alpha}_i$ para uma distribuição normal é dado por:

$$\hat{\alpha}_i = \{\mu_{E_i}, \sigma_{E_i}\}. \tag{3.7}$$

Por fim, conforme visto anteriormente o parâmetro da cópula pode ser estimado através do método IFM ou MMV. Assim, o método MMV ilustrado na equação 2.48 pode ser reescrita em termos de v_i da seguinte forma:

$$\sum_{t=1}^{n} \ln p_{\mathbf{E}_t}(v_1, \cdots, v_i, \cdots, v_k|\alpha_1, \cdots, \alpha_i, \cdots, \alpha_k, \gamma) \prod_{i=1}^{k} p_{E_{t,i}}(e_{t,i}), \tag{3.8}$$

onde, γ trata-se do parâmetro da cópula. Assim, para a cópula de Gumbel-Hourgaard o $\gamma = \theta$.

3.7 Combinação

Na etapa de combinação ilustrada na Figura 4 por $c(\cdot)$ a proposta é combinar vários modelos individuais através dos erros de previsão cometidos por cada modelo. Assim, neste trabalho, o formalismo de cópula é proposto para combinar tais modelos, onde os argumentos de entrada são as distribuições de probabilidade marginal dos erros obtidos pela Equação 3.4 oriundos das RNAs vistas na Figura 3. Contudo, o método proposto não se limita a combinar apenas RNAs, isto é, pode ser adotada outra abordagem como ARIMA (BOX; JENKINS; REINSEL, 1994), por exemplo.

A combinação via cópulas ocorre através da Função de Densidade de Probabilidade Conjunta (FDP) dos erros $\mathbf{E}_t$ dado por:

$$p_{\mathbf{E}_t}\left(e_{t,1}, \cdots, e_{t,k}\right) = c_{\mathbf{E}_t}\left(v_1, \cdots, v_k\right) \prod_{i=1}^{k} p_{E_{t,i}}(e_{t,i}) \tag{3.9}$$

onde, $p_{\mathbf{E}_t}(\cdot)$ é a FDP marginal de $E_{t,i}$ e $c_{\mathbf{E}_t}(\cdot)$ representa a FDP conjunta da cópula. Cabe ressaltar, que nesta equação são propostos k modelos individuais. O horizonte de previsão para o conjunto de teste é dado por $t = (n+1, n+2, \cdots, n+m)$.

Do lado direito da Equação 3.9 podem ser propostos diferentes tipos de cópulas, onde estas podem variar quando se trata da quantidade de parâmetros, ou seja, as cópulas Gumbel-Hougaard, Frank e Clayton fazem uso de um único parâmetro de dependência, por exemplo. Enquanto a cópula Normal pode fazer uso de apenas um ou vários parâmetros. Desta forma, têm sido estimados os parâmetros das distribuições de probabilidade acumulada $(\hat{\alpha}_1, \cdots, \hat{\alpha}_k)$ e posteriormente os parâmetros da cópula são intitulados por $\hat{\gamma}$. A Equação 3.9 da FDP conjunta pode ser reescrita da seguinte forma:

$$p_{\mathbf{E}_t}\left(\mathbf{e}_t|\hat{\alpha}_1, \cdots, \hat{\alpha}_k, \hat{\gamma}\right) = c_{\mathbf{E}}\left(F_{E_1}(e_{t,1}|\hat{\alpha}_1), \cdots, F_{E_k}(e_{t,k}|\hat{\alpha}_k)|\hat{\gamma}\right) \prod_{i=1}^{k} p_{E_i}(e_{t,i}|\hat{\alpha}_i). \tag{3.10}$$

3.7.1 Combinação pelo Método de Máxima Verossimilhança

O *copula-based ensemble* proposto (CB) faz o uso de cópulas para combinar as distribuições de probabilidade acumulada dos modelos individuais e utiliza o método de máxima verossimilhança para estimar o valor de CB que maximiza a Equação 3.10. Assim, o valor estimado é a previsão combinada pelo *ensemble* proposto, onde o conjunto de parâmetros $(\hat{\alpha}_1, \cdots, \hat{\alpha}_k, \hat{\gamma})$ são dados para os k modelos individuais de previsão, e especificamente, para este trabalho é assumido a normalidade entre as distribuições de probabilidade marginal dos modelos individuais e adotada a estrutura de erro aditivo. Logo, a Equação 3.10 pode

ser reescrita da seguinte forma:

$$\mathrm{CB}_t = \arg\max_u f(u) = p_{\mathbf{E}_t}\left(\mathbf{x}_t - \mathbf{u}|\hat{\alpha}_1, \cdots, \hat{\alpha}_k, \hat{\gamma}\right) = \\ c_{\mathbf{E}}\left(F_{E_1}(x_{t,1} - u|\hat{\alpha}_1), \cdots, F_{E_k}(x_{t,k} - u|\hat{\alpha}_k)|\hat{\gamma}\right) \prod_{j=1}^{k} p_{E_j}(x_{t,j} - u|\hat{\alpha}_j). \quad (3.11)$$

Desta forma, uma fórmula genérica para o *ensemble* baseado em cópulas via o método de máxima verossimilhança com outras estruturas de erro pode ser escrita como segue:

$$\mathrm{CB}_t = \arg\max_u f(u) = p_{\mathbf{E}_t}\left(\mathbf{x}_t \oplus \mathbf{u}|\hat{\alpha}_1, \cdots, \hat{\alpha}_k, \hat{\gamma}\right) = \\ c_{\mathbf{E}}\left(F_{E_1}(x_{t,1} \oplus_1 u|\hat{\alpha}_1), \cdots, F_{E_k}(x_{t,k} \oplus_k u|\hat{\alpha}_k)|\hat{\gamma}\right) \prod_{i=1}^{k} p_{E_i}(x_{t,i} \oplus_i u|\hat{\alpha}_i), \quad (3.12)$$

onde, $\oplus_i$ assume o operador aritmético '-' se a estrutura de erro for aditiva ou '/' caso seja multiplicativa, sempre sendo levado em consideração $E_i (i = 1, 2, \cdots, k)$.

A partir da Equação 3.12, pode ser visto que dada as previsões dos modelos individuais $\mathbf{x}_t = (x_{t,1}, \cdots, x_{t,i}, \cdots, x_{t,k})$ e a FDP conjunta com parâmetros ajustados ($\hat{\alpha}_1$, $\cdots, \hat{\alpha}_k, \hat{\gamma}$), o único elemento da equação desconhecido é o u. Vale ressaltar que a eventual presença de viés nos modelos individuais é levada em conta, uma vez que CB se baseia nas discrepâncias $x_{t,i} - u$ (ou $x_{t,i}/u$). Desta forma, podemos notar que de fato a Equação 3.12 pode conduzir a um intrínseco problema de otimização computacional, onde algoritmos de aproximação podem ser requeridos.

O *ensemble* proposto estima o valor de $\hat{u}_t$ através da Equação 3.12 utilizando para o valor de u apresentado na equação, um conjunto de possíveis valores que a série real pode assumir, tal que, estes valores são dados pelo valor mínimo e máximo da base de dados de treinamento $\{u_t, x_{t,1}, \cdots, x_{t,k}\}_{t=1}^{n}$. Assim, o conjunto de valores de u é dado por um vetor com mil valores igualmente espaçados entre o valor mínimo e máximo da base de dados apresentada anteriormente. Neste sentido, se por exemplo, os valores mínimo e máximo fossem 10 e 30 respectivamente, o vetor u seria preenchido com os seguintes valores $(10, 10.02, 10.04, 10.06, \cdots, 29.98, 30)$. Cabe ressaltar, que o conjunto com mil valores para u poderia ter tamanho maior ou menor. Todavia, um conjunto maior exige mais tempo computacional e não representa melhor precisão, e por outro lado, tamanho menor iria diminuir a precisão da previsão combinada. O tamanho proposto foi o que melhor uniu precisão e tempo computacional (este mesmo também foi proposto por Oliveira et al. (2017), Oliveira et al. (2016)). Os Algoritmos 3 e 4 ilustram o processo de combinação

através do método de máxima verossimilhança por meio de cópula.

Entrada: $x_{t,1}, \cdots, x_{t,k}, u_t, n, m$.
Resultado: Este algoritmo resulta em um vetor $\hat{u}_t$ processado com as combinações de k previsões.
Início
  Para $i \leftarrow 1 \in k$ **faça**
    Para $t \leftarrow 1 \in n$ **faça**
      $e_{t,i} \leftarrow x_{t,i} - u_t$;
    fim para
    $\hat{\alpha}_i \leftarrow$ é o vetor de parâmetros estimado via IFM;
    $F_{e_{j,i}} \leftarrow$ `ConstróiDistribuiçãoMarginal` $(e_{j,i}, \hat{\alpha}_i)$;
  fim para
  $\hat{\gamma} \leftarrow$ Ajusta o parâmetro da cópula utilizando $F_{e_{j,1}}, \cdots, F_{e_{j,k}}$ via IFM ;
  $c \leftarrow$ `AjustaCópula` $(\hat{\alpha}_1, \cdots, \hat{\alpha}_k, \hat{\gamma})$;
  Para $t \leftarrow \mathrm{n}+1 \in (\mathrm{n}+\mathrm{m})$ **faça**
    $\hat{u}_t \leftarrow$ `Combinar`$(x_{t,1}, \cdots, x_{t,k}, F_{e_{j,1}}, \cdots, F_{e_{j,k}}, \hat{\alpha}_1, \cdots, \hat{\alpha}_k, c)$;
  fim para
  retorno $\hat{u}_t$;
fim

Algoritmo 3: Algoritmo de combinação pelo método de máxima verossimilhança via cópulas (OLIVEIRA, 2014).

Função `Combinar(...)`
  Entrada: $x_1, \cdots, x_k, F_{e_{j,1}}, \cdots, F_{e_{j,k}}, \hat{\alpha}_1, \cdots, \hat{\alpha}_k, c$.
  Resultado: $\hat{u}_t$
  Início
    $n_l \leftarrow 1000$;
    $l_t \leftarrow$ é um vetor de tamanho qualquer n_l com valores igualmente espaçados entre os valores mínimo e máximo de $\{x_{t,1}, \cdots, x_{t,k}, u_t\}_{t=1}^{n}$ (conjunto de treinamento);
    Para $t \leftarrow 1 \in n_l$ **faça**
      Para $i \leftarrow 1 \in k$ **faça**
        $e_i \leftarrow x_i - l_t$;
      fim para
      fl $\leftarrow$ c.densidade$(e_1, \cdots, e_k, F_{e_{j,1}}, \cdots, F_{e_{j,k}}, \hat{\alpha}_1, \cdots, \hat{\alpha}_k)$;
    fim para
    $s \leftarrow$ é o índice do maior elemento contido no vetor **fl**;
    retorno l_s;
  fim
fim

Algoritmo 4: Algoritmo ilustra a função que realiza a combinação por meio do método de máxima verossimilhança (OLIVEIRA, 2014).

3.7.2 Combinação pelo Método de Mínima Variância

O método de Mínima Variância (MV) é um caso especial da cópula normal, onde o método supõe que a distribuição de probabilidade marginal dos erros é normal para qualquer modelo individual, bem como a dependência é modelada pela cópula normal. Assim, o MV utiliza uma combinação linear dada pela média ponderada:

$$\mathrm{CB}_t = \sum_{i=1}^{k} \omega_i x_{t,i} \tag{3.13}$$

onde k é o número de modelos individuais e ω_i é o peso atribuído para o i-ésimo modelo individual no instante t, representado por $x_{t,i}$ e a soma de todos os pesos deve ser um, isto é, $\sum_{i=1}^{k} \omega_i = 1$ (OLIVEIRA et al., 2017). Assim, cada peso é atribuído com base no desempenho de cada modelo individual $\mathbf{x}_i$. Neste sentido, ω_i é uma função entre a eficiência e correlação entre os modelos individuais, dada por:

$$\omega_i = \frac{\sum_{l=1}^{k} a_{li}}{\sum_{l=1}^{k}\sum_{i=1}^{k} a_{li}}$$

onde a_{li} é o i-ésimo elemento da l-ésima linha da matriz inversa de covariância M^{-1} em relação aos modelos individuais. Assim, M^{-1} foi obtida por meio da matriz inversa generalizada de Moore-Penrose (PENROSE, 1955).

Para um caso especifico, por exemplo, com dois modelos, isto é, $k = 2$, a matriz inversa de covariância pode ser dada por:

$$M^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}$$

onde $\rho = \sigma_{12}/\sigma_1\sigma_2$ é a correlação entre as previsões $X_{t,1}$ e $X_{t,2}$.

3.8 Resumo do Capítulo

Neste capítulo foram abordados os conceitos e definições do método proposto. Inicialmente, uma contextualização e a arquitetura do método foram apresentadas. Posteriormente, as propriedades do conjunto de treinamento e teste, bem como a formalização das RNAs foram expostas. Finalmente, o capítulo apresentou o processo para calcular os erros de previsão, ajuste da cópula e por fim, descreveu duas metodologias que utilizam cópulas para combinar as previsões das RNAs, podendo ser combinados via método de máxima verossimilhança e método de mínima variância. Ressaltando que ambos os métodos produzem teoricamente os mesmos resultados.

4 METODOLOGIA DOS EXPERIMENTOS

Neste capítulo é descrito em detalhes como será ajustado e realizado os experimentos para avaliar o desempenho do método proposto. Neste sentido, as séries temporais utilizadas, bem como o conjunto de treinamento e testes são expostos. Assim, os parâmetros utilizados para executar o experimento também são informados, tal como: a quantidade de modelos individuais adotados, número de neurônios das redes neurais, medidas de desempenho e teste de hipótese. Além disso, o modelo de regressão linear utilizado para explicar a causalidade entre o erro e a quantidade de modelos é formalizado.

4.1 Descrição Geral dos Experimentos

Com a finalidade de avaliar o desempenho do *ensemble* proposto neste trabalho, serão apresentados alguns experimentos envolvendo séries temporais de diversas naturezas, como financeira e demográfica, por exemplo. A série temporal u_t foi dividida em duas partes neste trabalho, sendo a primeira parte destinada a fase de treinamento dos modelos individuais com 75% da série, que representa n pontos (assim $t = 1, 2, \cdots, n$). A segunda parte utiliza os últimos 25% da série, representada por m pontos para realizar as previsões, em outras palavras, m pontos são utilizados para testar os modelos individuais (assim $t = n+1, n+2, \cdots, n+m$). Logo, a série temporal completa é dada por $u_t, t = 1, 2, \cdots, n+m$.

Os experimentos realizados neste trabalho envolveram 1000 modelos individuais, ou seja, $X_{t,1}, X_{t,2}, \cdots, X_{t,1000}$, deste modo $k = 1000$. As combinações foram realizadas a cada 10 modelos. Desta maneira, inicialmente foram combinados 10 modelos, depois 20 modelos, e assim por diante, sendo $k = 10, 20, \cdots, 1000$. A constante máxima $k = 1000$ foi adotada por ser uma quantidade geralmente maior que as adotadas na literatura, como pode ser visto nas obras de Kourentzes, Barrow e Crone (2014) que conduziram um estudo envolvendo 10 e 100 modelos individuais e Kuncheva (2014) que apresenta um ensaio com 100 preditores, por exemplo. Desta forma, o objetivo é estudar detalhadamente o que acontece com valores consideravelmente grandes para k, embora o número de modelos não se limita apenas ao valor exposto neste trabalho.

Para todos os experimentos apresentados, as arquiteturas das redes neurais artificiais puderam assumir L neurônios na camada escondida, onde o valor de L foi selecionado aleatoriamente no intervalo entre 1 e 1000. A quantidade de neurônios na camada escondida é levada em consideração para ser possível obter uma variedade de modelos diferentes, não apenas na sua arquitetura, mas também no desempenho dos modelos individuais. Assim, foi possível observar que $L \in [1, 1000]$ possibilita a generalização de modelos pouco e muito acurados. Contudo, este intervalo não se limita apenas ao valor adotado neste trabalho.

Os modelos individuais referidos como $X_{t,1}, X_{t,2}, \cdots, X_{t,k}$ são respectivamente redes neurais artificiais treinadas por meio do algoritmo *Extreme Learning Machine*. A justificativa do *ensemble* proposto fazer uso desses modelos individuais, se dá pela capacidade de generalização das RNAs treinadas via ELM, bem como sua eficiência e acurácia em termos estatísticos e também pelo curto tempo de treinamento. Porém, o *ensemble* proposto não se limita a agregar exclusivamente RNAs, é possível combinar qualquer tipo de modelo individual, no entanto como a maioria dos preditores apresentados na literatura são substancialmente mais lentos quando comparados com as RNAs com ELM, que foram escolhidas com base nesse critério.

O *ensemble* proposto neste trabalho utiliza a cópula normal para realizar as combinações apresentadas nos experimentos. Deve-se enfatizar que a escolha por esta cópula justifica-se devido a capacidade da mesma, em captar dependências positivas e negativas, além de sua eficiência e acurácia estatística já ter sido previamente ilustrada em trabalhos como Oliveira et al. (2017) e Oliveira et al. (2016). Cabe ressaltar também, que o *ensemble* proposto realiza as combinações através da MV apresentado anteriormente pelo fato de consumir menos tempo computacional. O desempenho em termos do tempo computacional é destacado por Oliveira (2014) no qual menciona que algumas cópulas consumem menor tempo computacional que outras, desta forma, o autor aponta que a cópula Cacoullos apesar de sua qualidade e ser não-paramétrica, tem um custo computacional superior a cópula normal. Cabe ressaltar, que este trabalho visa treinar, prever e combinar inúmeros modelos individuais, e portanto, foi focado nos métodos mais eficazes em termos do tempo computacional para essa finalidade. Neste sentido, foi assumido que as distribuições de probabilidade marginal dos modelos individuais seguem uma distribuição normal. Essa condição se faz necessária uma vez que os dados podem não seguir tal distribuição, e nesse sentido caberia estudar a distribuição de cada modelo individual, por meio dessa condição, pode-se assumir a normalidade entre os modelos individuais, e reduzir o tempo computacional, uma vez que essa ação dispensa a necessidade de computar a distribuição de cada modelo. Ressaltando que a condição adotada já é comumente praticada na literatura por meio do método de mínima variância, que tem apresentado bons resultados conforme é discutido por Firmino, Neto e Ferreira (2014).

O método IFM foi aplicado neste estudo para ajustar os parâmetros (μ e σ) das distribuições marginais dos erros, além de ajustar o parâmetro de dependência da cópula normal. Desta forma, os experimentos que serão apresentados, foram desenvolvidos por meio da linguagem de programação R (VENABLES; SMITH; TEAM, 2019) com o auxilio da biblioteca "elmNN" (GOSSO, 2013). Cabe destacar que a biblioteca foi utilizada basicamente na parte I do *ensemble* proposto para calcular a matriz inversa generalizada de Moore-Penrose através da função $ginv(\cdot)$ e implementadas as demais funcionalidades da rede neural via algoritmo ELM. O *software* também foi aplicado na implementação da parte II voltada para a combinação das redes neurais via cópula.

Neste trabalho, são levados em consideração quatro tipos distintos de séries temporais, sendo: financeira, demográfica, hidrológica e meteorológica (veja a Tabela 1). Na série financeira S&P f500 (SP), as observações ocorrem entre 3 de Janeiro de 1950 até 3 de Fevereiro de 2017. A série temporal Nasdaq Index (ND), consiste em outro fenômeno financeiro adotado neste estudo, tal que suas observações são levadas em consideração entre 8 de Novembro de 1996 a 24 de Fevereiro de 2017. A Dow Jones Industrial Average (DJ) é uma série do mercado financeiro que foi observada no período de 1 de Janeiro de 1998 até 26 de Agosto de 2003. A série Quebec (QB) trata-se de um fenômeno de demografia que remete ao número diário de nascimentos ocorridos na cidade de Quebec no Canadá a partir de 1 de Janeiro de 1977 até 31 de Dezembro de 1990. Rio Saugeen (RS) trata-se de um fenômeno da hidrologia referente às medições diárias da média do fluxo do rio Saugeen em Walkerton e Ontario no Canadá, os pontos desta série temporal foram coletados entre 1 de Janeiro de 1915 até 16 de Janeiro de 1980. A série de Precipitação em Melbourne (PM) corresponde ao fenômeno de precipitação de chuva na cidade de Melbourne na Austrália, as medições ocorreram entre 1 de Janeiro de 1981 até 1 de Janeiro de 1991 os dados foram obtidos da agência de meteorologia da Austrália. A série Rio Oldman (RO) trata-se de um evento hidrológico, que mede o fluxo diário médio na bacia do rio Oldman entre 1 de Janeiro de 1988 a 31 de Dezembro de 1991. A Temperatura de Oldman (TO) trata-se de um evento de meteorologia na bacia do rio Oldman entre 1 de Janeiro de 1988 a 31 de Dezembro de 1991. Rio Fisher (RF) trata-se do fenômeno hidrológico que mede o fluxo diário médio do rio Fisher próximo de Dallas nos Estados Unidos da Àmerica, o evento foi observado entre 1 de Janeiro de 1988 a 31 de Dezembro de 1991. Finalmente, a série Precipitação em Oldman (PO), considerada neste trabalho, trata-se da medição total diária de precipitação na bacia do rio Oldman entre 1 de Janeiro de 1988 a 31 de Dezembro de 1991.

As séries temporais SP, ND e DJ podem ser facilmente encontradas em sites que apresentam as informações do mercado financeiro [1]. As séries QB, RS, RO, TO, RF e PO foram obtidas por meio do repositório de séries temporais DataMarket [2]. Contudo, este grupo de séries temporais foram retiradas da obra de Hipel e McLeod (1994). Por fim, a série PM foi coletada pela agência de meteorologia Australiana, em que os dados podem ser obtidos por meio do repositório DataMarket [3].

A Tabela 1 apresenta resumidamente as propriedades de cada série temporal estudada, com sua sigla, descrição, tipo, unidade de tempo, data de início da coleta da série, a data de início considerada para realizar os testes dos experimentos realizados

[1] Um site que apresenta este tipo de informação é o Yahoo! Finanças, por exemplo. O site está disponível no endereço: https://br.financas.yahoo.com/

[2] As séries apresentadas podem ser encontradas no endereço: https://datamarket.com/data/list/?q=provider:tsdl

[3] Os dados desta série temporal estão disponíveis para consulta em: https://datamarket.com/data/set/2328/daily-rainfall-in-melbourne-australia-1981-1990#!ds=2328&display=line

neste trabalho. Mostra também a data de término, ou seja, data da última coleta da série temporal. E finalmente, é apresentado o tamanho do conjunto de treinamento (n), teste (m) e o valor do Lag adotado.

Tabela 1 – Propriedades das séries temporais estudadas.

Sigla	Descrição	Tipo	Unidade	Início	Início (teste)	Término	n	m	Lag
SP	S&P 500	Financeira	diária	3 de Jan, 50	17 de Out, 66	3 de Fev, 17	12661	4220	40
ND	Nasdaq	Financeira	diária	8 de Nov, 96	5 de Dez, 01	24 de Fev, 17	3831	1276	40
DJ	Dow Jones	Financeira	diária	1 de Jan, 98	1 de Abr, 02	26 de Ago, 03	1065	355	40
QB	Quebec	Demografia	diária	1 de Jan, 77	2 de Jul, 87	31 de Dez, 90	3835	1278	25
RS	Rio Saugeen	Hidrologia	diária	1 de Jan, 15	13 de Out, 63	16 de Jan, 80	17806	5935	40
PM	Melbourne	Meteorologia	diária	1 de Jan, 81	2 de Jul, 88	1 de Jan, 91	2740	913	40
RO	Rio Oldman	Hidrologia	diária	1 de Jan, 88	10 de Abr, 91	31 de Dez,91	1096	365	40
TO	Rio Oldman	Meteorologia	diária	1 de Jan, 88	10 de Abr, 91	31 de Dez,91	1096	365	40
RF	Rio Fisher	Hidrologia	diária	1 de Jan, 88	10 de Abr, 91	31 de Dez, 91	1096	365	40
PO	Rio Oldman	Meteorologia	diária	1 de Jan, 88	10 de Abr, 91	31 de Dez, 91	1096	365	40

A escolha do Lag (janela) utilizado neste trabalho foi obtido por meio da análise de autocorrelação. Assim, para cada série temporal apresentada anteriormente, foi realizada a análise de autocorrelação, para entender, a aleatoriedade das séries e encontrar um lag significante para prever a série temporal. A Tabela 1 mostra que na grande maioria dos casos o lag significante utilizado para a previsão das séries temporais foi $Lag = 40$, com exceção da série QB que utilizou $Lag = 25$. Cabe ressaltar, que não foi adotado um lag variável para criar os modelos individuais, visto que o foco do trabalho é combinar os modelos independente de sua eficiência e acurácia estatística.

4.2 Medidas de Desempenho

As métricas que serão adotadas para avaliar o desempenho do CB serão obtidas por meio da média aritmética de cada uma das medidas de desempenho MSE, MAE, THEILU e RMSE descritas matematicamente pelas Equações 2.1, 2.2, 2.3 e 2.4. Essas métricas foram escolhidas por serem utilizadas largamente na literatura para mensurar a qualidade das estimativas obtidas pelos modelos, o MSE trata-se de uma medida capaz de unir a acurácia e eficiência das estimativas em um só medida, o MAE funciona de maneira semelhante ao MSE, porém dispensa elevar o erro ao quadrado para força que os erros não tenham sinal, esta métrica faz isso por meio do erro absoluto, por outro lado a métrica THEILU já proporciona um coeficiente capaz de indicar a qualidade das estimativas através dos valores 0 e 1, aonde 0 indica uma estimativa perfeita e 1 representa estimativas com valores distorcidos da realidade. Finalmente, RMSE é a raiz do MSE e possui uma capacidade medir erros geralmente quando a variância for considerada pequena, por exemplo, menor um. Neste sentido, os experimentos que serão apresentados foram combinados N vezes com a mesma quantidade k de modelos individuais, porém, alternando a ordem de combinação dos modelos individuais para evitar que o processo de medição seja prejudicado por algum grupo de modelos que esteja nas extremidades ou no centro do vetor de modelos individuais,

em outras palavras, as combinações são realizadas aleatoriamente alternando a ordem das combinações dos modelos individuais, por exemplo, para $k = 10$ combina primeiramente nesta ordem $X_{t,1}, X_{t,2}, X_{t,3}, X_{t,4}, X_{t,5}, X_{t,6}, X_{t,7}, X_{t,8}, X_{t,9}, X_{t,10}$, posteriormente $X_{t,101}, X_{t,15}, X_{t,95}, X_{t,71}, X_{t,4}, X_{t,7}, X_{t,156}, X_{t,14}, X_{t,881}, X_{t,714}$, depois seguindo a mesma linha de raciocínio obtemos $X_{t,501}, X_{t,22}, X_{t,856}, X_{t,1}, X_{t,5}, X_{t,711}, X_{t,921}, X_{t,408}, X_{t,50}, X_{t,31}$ e assim por diante até completar 50 combinações distintas para calcular a média das métricas utilizadas. Consequentemente, será obtida uma amostra com N métricas distintas para cada valor de k (sendo, $k = 10, 20, 30, \cdots, 1000$). O valor adotado para o tamanho da amostra de medidas foi $N = 50$. Desta forma, a média do MSE para uma dada quantidade k de modelos individuais pode ser descrita por:

$$\overline{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\text{MSE}_i, \tag{4.1}$$

para MAE, RMSE e THEILU pode ser escrito como segue:

$$\overline{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}\text{MAE}_i, \tag{4.2}$$

$$\overline{\text{RMSE}} = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}_i, \tag{4.3}$$

$$\overline{\text{THEILU}} = \frac{1}{N}\sum_{i=1}^{N}\text{THEILU}_i. \tag{4.4}$$

Assim, o desempenho alcançado pelo *ensemble* é medido nesse trabalho por meio do $\overline{\text{MSE}}, \overline{\text{MAE}}, \overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$.

4.3 Teste de Hipótese da Correlação

Depois de calcular o desempenho do *ensemble* para cada série temporal por meio das métricas anteriormente apresentadas, pode ser investigada a força de associação entre as variáveis k e cada uma das métricas ($\overline{\text{MSE}}, \overline{\text{MAE}}, \overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$). Essa associação foi medida neste trabalho por meio do coeficiente de correlação de Spearman. O coeficiente de Spearman foi adotado por ser não-paramétrico e dispensar a exigência de normalidade entre as variáveis.

Contudo, para garantir que o coeficiente de correlação encontrado não seja apenas casual, em virtude de algum erro de desvio da amostragem, será aplicado o teste de hipótese para o coeficiente de correlação de Spearman. Este teste parte do pré-suposto que a hipótese H_0 é verdadeira, indicando que a correlação entre as variáveis é nula, ou seja, $\rho = 0$ indicando que não existe correlação entre k e $\overline{\text{MSE}}$.

Por outro lado, a hipótese H_1 caracteriza $\rho \neq 0$ que indica a existência de correlação entre as variáveis. Desta forma, quando $\rho > 0$ a correlação é positiva, representando que quando k cresce a métrica tende a crescer, em outras palavras, k e $\overline{\text{MSE}}$ crescem no mesmo sentido. Para $\rho < 0$ carateriza uma correlação estatisticamente negativa que implica em valores contrários para as variáveis, isto é, quando k cresce o $\overline{\text{MSE}}$ diminui. Neste sentido, cabe salientar que este trabalho almeja alcançar resultados que satisfaçam a segunda condição, ou seja, uma correlação negativa ($\rho < 0$) entre o número de modelos individuais e o erro cometido pelos modelos.

4.4 Regressão Linear Simples

A regressão linear será adotada neste trabalho para explicar a relação de causa e efeito das variáveis: número de modelos individuais (k) e a métrica que mensura o erro de previsão ($\overline{\text{MSE}}$, por exemplo). Cabe ressaltar, que a análise de regressão parte da suposição de que as variáveis são correlacionadas, por isso, anteriormente, um teste de correlação foi aplicado entre as variáveis, para garantir que os resultados obtidos via regressão linear não sejam meramente casuais.

Assim, sejam as observações pareadas (k, $\overline{\text{MSE}}$) para o MSE, tal que, a regressão linear simples para este conjunto de dados pode ser escrita como:

$$\widehat{\overline{\text{MSE}}} = a + b \times k, \tag{4.5}$$

onde, a e b são os parâmetros da regressão, k é a variável preditiva que se trata da quantidade de modelos individuais, e finalmente, $\widehat{\overline{\text{MSE}}}$ é o valor estimado para a média do erro quadrático médio da combinação com k modelos.

Quando a relação entre k e a métrica não for linear, será necessário aplicar alguma transformação para aproximar ao máximo a linearidade da relação. Esse processo se faz necessário uma vez que a análise de regressão linear é aplicada para dados com tal tipo de relação. Dessa forma, uma alternativa, para os casos em que a relação entre k e a métrica não for linear, é aplicar uma modelagem exponencial, no intuito de transformar a relação entre as variáveis em aproximadamente log-linear. Assim, o modelo exponencial pode ser definido matematicamente como:

$$\widehat{\overline{\text{MSE}}} = e^{a+b\times k}. \tag{4.6}$$

As Equações 4.5 e 4.6 apresentam a definição matemática para aplicar a regressão linear para estimar o $\widehat{\overline{\text{MSE}}}$, cabendo ressaltar que para as demais métricas $\widehat{\overline{\text{MAE}}}$, $\widehat{\overline{\text{RMSE}}}$ e $\widehat{\overline{\text{THEILU}}}$ seguem a mesma ideia. A regressão para estimar a quantidade de modelos individuais é basicamente a mesma sendo em termos do $\overline{\text{MSE}}$:

$$\hat{k}^{(\overline{\text{MSE}})} = a + b \times \overline{\text{MSE}} \tag{4.7}$$

onde $\hat{k}^{(\overline{\text{MSE}})}$ é a estimativa do número de modelos individuais necessário para obter o $\overline{\text{MSE}}$, por outro lado, quando os dados não seguirem a linearidade uma opção é adotar a seguinte definição:

$$\hat{k}^{(\overline{\text{MSE}})} = a + b \times \log(\overline{\text{MSE}}). \tag{4.8}$$

na qual $\log(\overline{\text{MSE}})$ foi utilizado no processo de estimação dos parâmetros a e b.

De modo geral, os experimentos adotados neste trabalho, visam explicar por meio da análise de regressão, a relação de causa e efeito entre a quantidade de modelos e o erro de previsão. Neste sentido, a análise de regressão apresenta papel fundamental para explicar a causalidade entre o erro e k.

4.5 Resumo do Capítulo

Neste capítulo foi descrita a metodologia dos experimentos que foram conduzidos no trabalho. Particularmente, o capítulo apresenta as definições adotadas para realização dos experimentos, bem como demonstra como será medido o desempenho do método proposto. Um teste estatístico e suas características para analisar o grau de associação entre o erro comedido pelo método proposto e a quantidade de modelos também é exibido. Além, de ilustrar uma metodologia via modelos de regressão linear simples para encontrar a quantidade de modelos individuais a partir da métrica, bem como encontrar a métrica por meio do número de modelos.

5 ANÁLISE E RESULTADOS

Neste capítulo são apresentados os resultados experimentais para o método proposto. As análises de regressão linear simples, para estimar o erro do método proposto têm sido expostas neste capítulo. Os resultados e análises da comparação entre CB e metodologias previamente estabelecidas na literatura são apresentadas.

A análise de regressão linear é uma das técnicas mais populares para estudar o comportamento de duas variáveis no ramo da estatística, sendo um ponto de partida a ser considerado para um entendimento mais detalhado e correto sobre o comportamento de causalidade entre o erro do método proposto e a quantidade de modelos individuais. Nesse sentido, a análise de regressão aparece como uma técnica para estimar o erro do método a partir da quantidade de modelos individuais. O coeficiente de correlação é uma técnica para compreender o comportamento entre o erro do método proposto e a quantidade de modelos individuais e vice-versa. Desta forma, através do teste de hipótese do coeficiente de correlação é possível saber o tipo (positiva ou negativa) e grau de correlação entre o erro e o número de modelos.

5.1 Resultados do Método Proposto

Especificamente neste trabalho, foi observado ao longo dos estudos que quando a série temporal é pequena, ocasiona um super ajustamento (*overfitting*) nas redes neurais que possibilita a geração de modelos super especializados, ou seja, preditores que não apresentam viés estatístico, em outras palavras, modelos que preveem corretamente todas as observações do conjunto de treinamento. Obtendo, desta forma, a média e a variância do erro, próxima de zero ou propriamente zero. Estatisticamente é inviável combinar este tipo de modelo por meio de cópulas uma vez que matematicamente é necessário que haja variabilidade entre os erros dos modelos individuais para que seja possível combinar. Neste sentido, séries pequenas foram evitadas neste trabalho para evitar a geração de modelos sem viés estatístico.

Uma particularidade observada no processo de geração dos modelos individuais foi que para janelas de previsão pequenas (por exemplo, lag=4), as redes neurais artificiais não eram capazes de criar preditores individuais diversificados. Em experimentos realizados ao longo deste trabalho, foi possível observar que na maior parte das vezes, quando os modelos individuais eram parecidos, à medida em que estes estavam sendo combinados os resultados pioravam. Certamente, porque a cada combinação a contribuição do novo modelo era menor que o erro associado ao mesmo. Logo, $lags$ pequenos foram evitados

nos experimentos apresentados, de modo, que seja melhorada a diversidade dos modelos individuais.

Os resultados que serão apresentados são mensurados através da média das seguintes métricas apresentadas anteriormente: $\overline{\text{MSE}}$, $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$ envolvendo apenas o conjunto de teste. Os resultados exibidos incluem os experimentos com CB e outras metodologias da literatura, que preveem, respectivamente, as séries temporais SP, ND, DJ, QB, RS, PM, RO, TO, RF e PO. De modo, a demonstrar a eficiência e acurácia estatística dos *ensembles*, as medidas de desempenho apresentadas correspondem a média de 50 combinações para cada valor de k (ver Seção 4.2).

Inicialmente será avaliado o grau de associação entre as métricas mencionadas anteriormente e o número de modelos combinados. Este procedimento se faz necessário para observar se existem indícios de causa e efeito entre as variáveis: erro (isto é, a métrica) e número de modelos combinados. Desta forma, foi aplicado o teste de hipótese para os coeficientes de correlação de Spearman (ρ) e Kendall (τ), sendo ambos testes não paramétricos.

As Tabelas 2 e 3 apresentam os resultados do *Ensemble* proposto - CB obtido por meio do teste de hipótese de Spearman e Kendall que mede o nível de associação entre as métricas ($\overline{\text{MSE}}$, $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$) e k (número de modelos individuais combinados). Assim, o teste indica a grandeza de associação entre as métricas e k.

Os resultados para ρ e τ dos respectivos testes rejeitaram a hipótese H_0 indicando que existe correlação entre as variáveis métrica e k. Assim, os ensaios ilustrados nas Tabelas 2 e 3 mostram que existe evidência de correlação negativa entre as variáveis com nível de significância de 5% para todos os casos. Conforme pode ser notado, para a métrica $\overline{\text{MSE}}$ e série SP o valor de ρ= -0.9648845 e $p\text{-}value < 2.2e-16$, ou seja, indicando que a correlação entre métrica e k é negativa a um nível de significância ainda menor que 5%.

De fato, os resultados da Tabela 2 sugerem que o CB possui uma correlação negativa, a um nível de significância de 5% para as séries SP, ND, QB, RS, PM, RO, TO, RF e PO, ou seja, para todas as séries temporais em estudo. Na Tabela 3 pode ser notado que CB também possui uma associação negativa para todas as séries temporais com nível de significância de 5%.

De modo geral, os resultados mostram que existe correlação negativa entre o erro e a quantidade de modelos individuais inseridos na combinação. O que suporta a premissa da quantidade de modelos individuais impactar na qualidade de CB.

Os resultados que serão apresentados nas figuras a seguir foram alcançados pelo CB em termos do $\overline{\text{MSE}}$, $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$ para prever as séries temporais apresentadas anteriormente. Observe que o eixo "x" representa k (a quantidade de modelos individuais que estão sendo incluídos na combinação) e o eixo "y" corresponde ao valor médio da

Tabela 2 – Teste de hipótese para o coeficiente de correlação de Spearman (ρ) entre as métricas obtidas via CB e k (conjunto de teste).

Séries	Métricas							
	$\overline{\text{MSE}}$		$\overline{\text{MAE}}$		$\overline{\text{RMSE}}$		$\overline{\text{THEILU}}$	
	ρ	*p-value*	ρ	*p-value*	ρ	*p-value*	ρ	*p-value*
SP	-0.96488	<2.2e-16	-0.91381	<2.2e-16	-0.96488	<2.2e-16	-0.96394	<2.2e-16
ND	-0.64846	<2.2e-16	-0.42594	1.213e-05	-0.64846	<2.2e-16	-0.65602	1.279e-13
DJ	-0.61836	<2.2e-16	-0.70799	<2.2e-16	-0.60824	<2.2e-16	-0.60093	<2.2e-16
QB	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16
RS	-1	<2.2e-16	-0.91702	<2.2e-16	-1	<2.2e-16	-0.99995	<2.2e-16
PM	-0.99995	<2.2e-16	-0.99927	<2.2e-16	-0.99995	<2.2e-16	-0.99998	<2.2e-16
RO	-0.99985	<2.2e-16	-0.99947	<2.2e-16	-0.99984	<2.2e-16	-0.99984	<2.2e-16
TO	-0.71015	<2.2e-16	-0.71842	<2.2e-16	-0.75595	<2.2e-16	-0.74905	<2.2e-16
RF	-0.92998	<2.2e-16	-0.61946	<2.2e-16	-0.90701	<2.2e-16	-0.91089	<2.2e-16
PO	-0.60156	<2.2e-16	-0.78787	<2.2e-16	-0.60963	<2.2e-16	-0.60733	<2.2e-16

Tabela 3 – Teste de hipótese para o coeficiente de correlação de Kendall (τ) entre as métricas obtidas via CB e k (conjunto de teste).

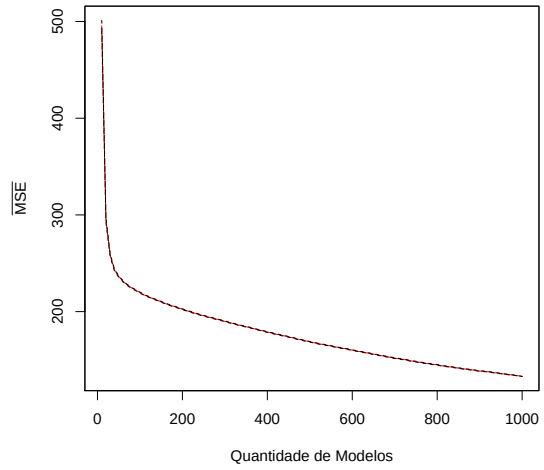
Séries	Métricas							
	$\overline{\text{MSE}}$		$\overline{\text{MAE}}$		$\overline{\text{RMSE}}$		$\overline{\text{THEILU}}$	
	τ	*p-value*	τ	*p-value*	τ	*p-value*	τ	*p-value*
SP	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16
ND	-0.99838	<2.2e-16	-0.99636	<2.2e-16	-0.99838	<2.2e-16	-0.99838	<2.2e-16
DJ	-0.43434	1.524e-10	-0.51757	2.349e-14	-0.41979	6.073e-10	-0.40484	2.4e-09
QB	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16	-1	<2.2e-16
RS	-1	<2.2e-16	-0.81414	<2.2e-16	-1	<2.2e-16	-0.99878	<2.2e-16
PM	-0.99878	<2.2e-16	-0.98949	<2.2e-16	-0.99878	<2.2e-16	-0.99959	<2.2e-16
RO	-0.99595	<2.2e-16	-0.98989	<2.2e-16	-0.99555	<2.2e-16	-0.99555	<2.2e-16
TO	-0.59515	<2.2e-16	-0.57818	<2.2e-16	-0.61979	<2.2e-16	-0.61292	<2.2e-16
RF	-0.81696	<2.2e-16	-0.47353	2.937e-12	-0.78545	<2.2e-16	-0.79030	<2.2e-16
PO	-0.42868	2.623e-10	-0.65050	<2.2e-16	-0.43272	1.781e-10	-0.42828	2.726e-10

métrica. Nas Figuras entre 5 e 14 serão realizadas 100 combinações envolvendo entre 10 e 1000 modelos individuais, ou seja, a primeira combinação é realizada com 10 modelos, posteriormente, com 20, em seguida com 30 até chegar a última combinação com 1000 modelos. A linha contínua representa o valor da métrica, enquanto que as linhas tracejadas indicam os intervalos superior e inferior de confiança.

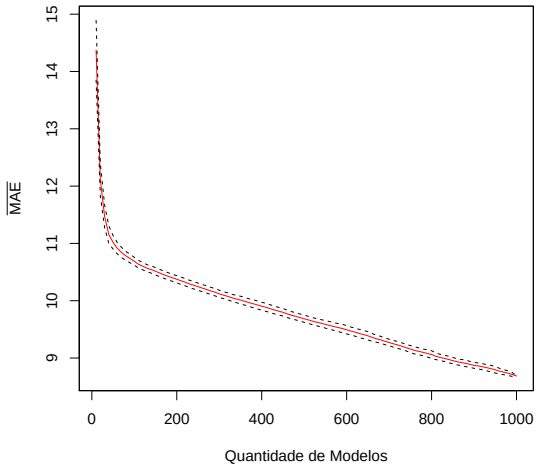
A Figura 5 ilustra o comportamento de CB para prever a série temporal SP (utilizando o conjunto de teste). Como pode ser observado, o comportamento de CB para todas as métricas é decrescente, ilustrando uma associação negativa entre as métricas e k, conforme suportado anteriormente pelo teste de hipótese. Neste sentido, a medida que as métricas decrescem o número de modelos individuais cresce. O intervalo de confiança para a série observada apresentou um intervalo relativamente pequeno, bem como os intervalos inferiores e superiores tiveram poucas variações ao longo da agregação de mais modelos individuais. Com exceção do $\overline{\text{THEILU}}$ que é possível notar uma diminuição do intervalo

de confiança. Ao observar o $\overline{\text{MSE}}$, podemos notar uma melhoria significativa reduzindo o erro de aproximadamente 500 (para $k = 10$) para menos de 200 (para $k = 1000$).

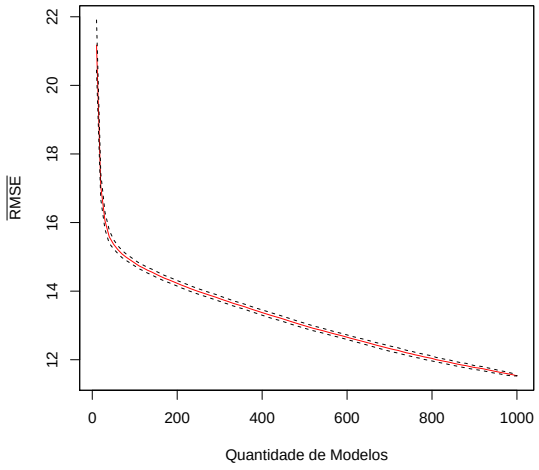
Através dos ensaios ilustrados nas Figuras 5, 6, 7, 8, 10, 9, 11, 13, 12, 14 não podem ser afirmadas algumas suposições de interesse deste trabalho, como por exemplo, se o decaimento do erro tende a zero com k tendendo ao infinito, ou ainda, se existe um número ótimo para k. As análises para tais suposições serão discutidas a seguir. Contudo, os resultados que estão sendo apresentados, demostram o comportamento do erro e intervalo de confiança a medida que k cresce.



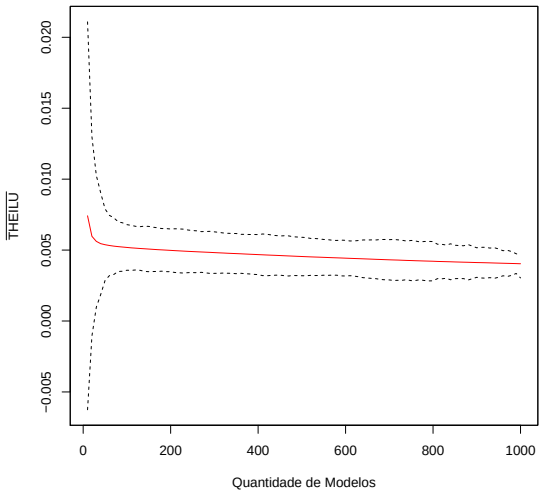
(a) $\overline{\text{MSE}}$



(b) $\overline{\text{MAE}}$



(c) $\overline{\text{RMSE}}$



(d) $\overline{\text{THEILU}}$

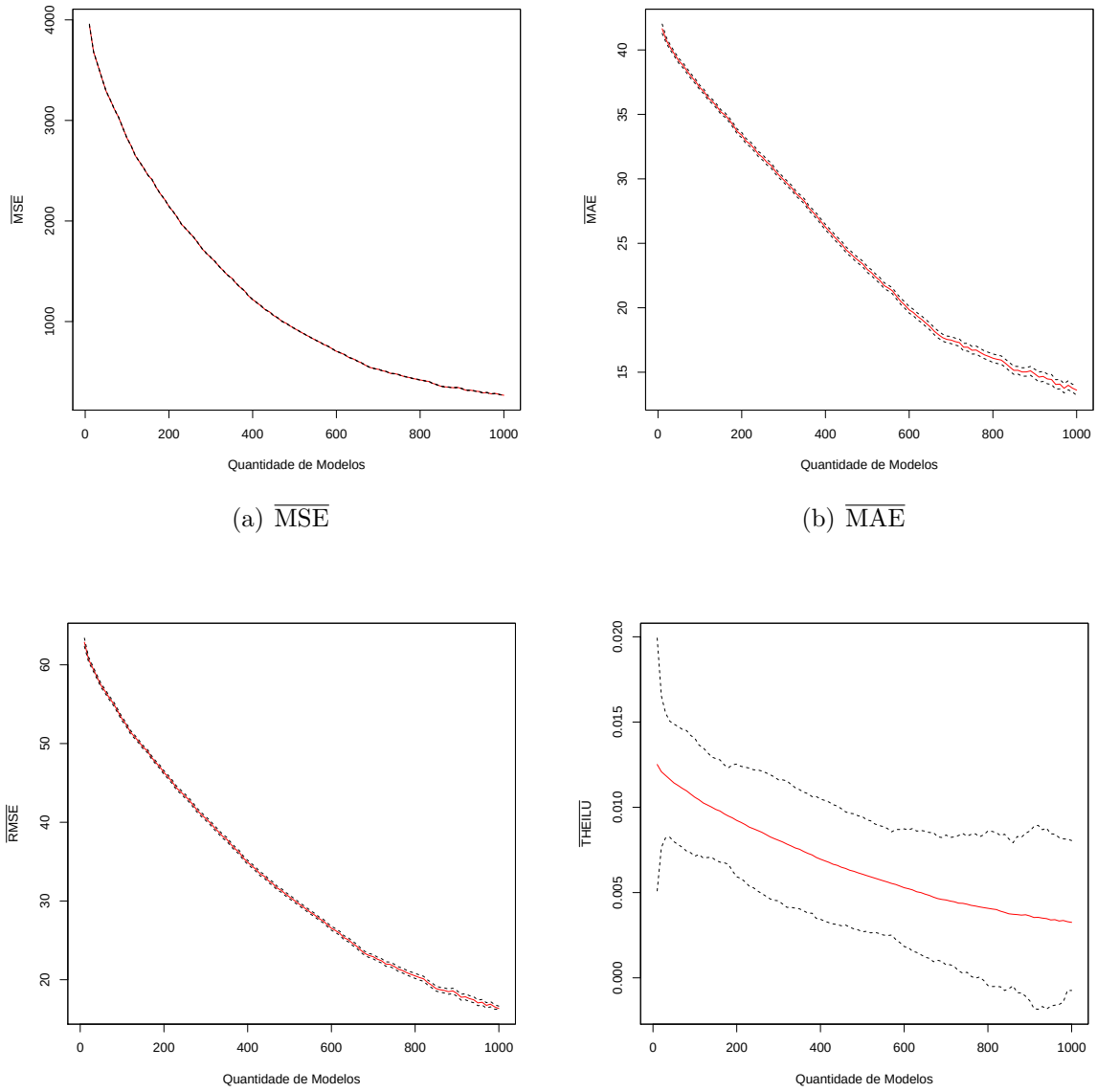
Figura 5 – CB para prever a série temporal SP (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

A Figura 6 ilustra o desempenho de CB para prever a série temporal ND. O comportamento para esta série corresponde ao mesmo mostrado para SP, isto é, uma correlação negativa entre a métrica e o número de modelos individuais, o que resulta em uma curva de decaimento. A curva é aproximadamente exponencial. Para o $\overline{\text{MSE}}$ a variação refletida no intervalo de confiança é pouco notável. Enquanto que para $\overline{\text{MAE}}$ e $\overline{\text{RMSE}}$ a variação é relativamente maior quando $k > 700$. Por outro lado, para $\overline{\text{THEILU}}$ a variação do intervalo de confiança fica mais evidente. A melhoria de CB fica evidenciada ao observar a convergência de aproximadamente $\overline{\text{MSE}} = 4000$ (quando $k = 10$) para $\overline{\text{MSE}} < 1000$ (quando $k = 1000$).

A Figura 7 apresenta o desempenho de CB aplicado à combinação das previsões para a série temporal DJ. O comportamento para esta série é diferente das séries SP e ND, ou seja, para $k > 400$ é razoável dizer que o erro é constante. Este comportamento ocorre para todas as métricas utilizadas. Assim, dado que a correlação apresentada anteriormente (veja Tabela 2) entre as métricas e k é negativa com nível de 5% de significância, pode-se dizer que estes resultados podem conduzir a futuras conclusões de um número específico de modelos individuais para esta série temporal. Neste sentido, para tais conclusões, se faz necessário estudos sobre regressões lineares e testes estatísticos que serão apresentados na Seção 5.2. O resultado revela uma redução da métrica ao longo das combinações de $\overline{\text{MSE}} > 2.5e + 19$ (quando $k = 10$) para $\overline{\text{MSE}} < 5.0e + 18$ (quando $k \geq 400$). O intervalo de confiança dá uma boa ideia quanto ao comportamento do erro em relação ao valor de k, demostrando algumas diferenças entre o limite superior e inferior, o que indica estatisticamente, a possibilidade da métrica se comportar de maneira parecida dos resultados apresentados.

As Figuras 8 e 9 mostram o desempenho de CB para as séries temporais QB e PM. Analisando os resultados pode-se destacar que o erro é substancialmente reduzido ao longo das combinações. Em alguns casos, quase formando um comportamento linear. Logo, CB apresenta menor erro para $k = 1000$ ao invés de $k < 1000$, ou seja, existe um ganho razoavelmente significativo sendo, por exemplo, para QB o $\overline{\text{MSE}} > 400$ (quando $k = 10$) e $\overline{\text{MSE}} < 100$ (quando $k = 1000$). Todavia, o intervalo de confiança é razoavelmente curto para todas as métricas. Mesmo quando trata-se da métrica $\overline{\text{THEILU}}$ a diferença entre o limite superior e inferior fica em 0.008 (limite superior - limite inferior, quando $k = 10$) para série QB.

A Figura 10 ilustra o desempenho de CB para a série temporal RS. Para estas séries em particular, os resultados mostram um decaimento da métrica ao longo das combinações. O desempenho do CB pode ser notado através dos resultados obtidos ao longo das combinações, isto é, para $k = 10$ o $\overline{\text{MSE}} > 2500$, enquanto que para $k = 1000$ o $\overline{\text{MSE}} < 1500$, o que demostra a qualidade de CB na medida que mais modelos individuais são incluídos na agregação. Os resultados sugerem uma sutil diminuição do intervalo de



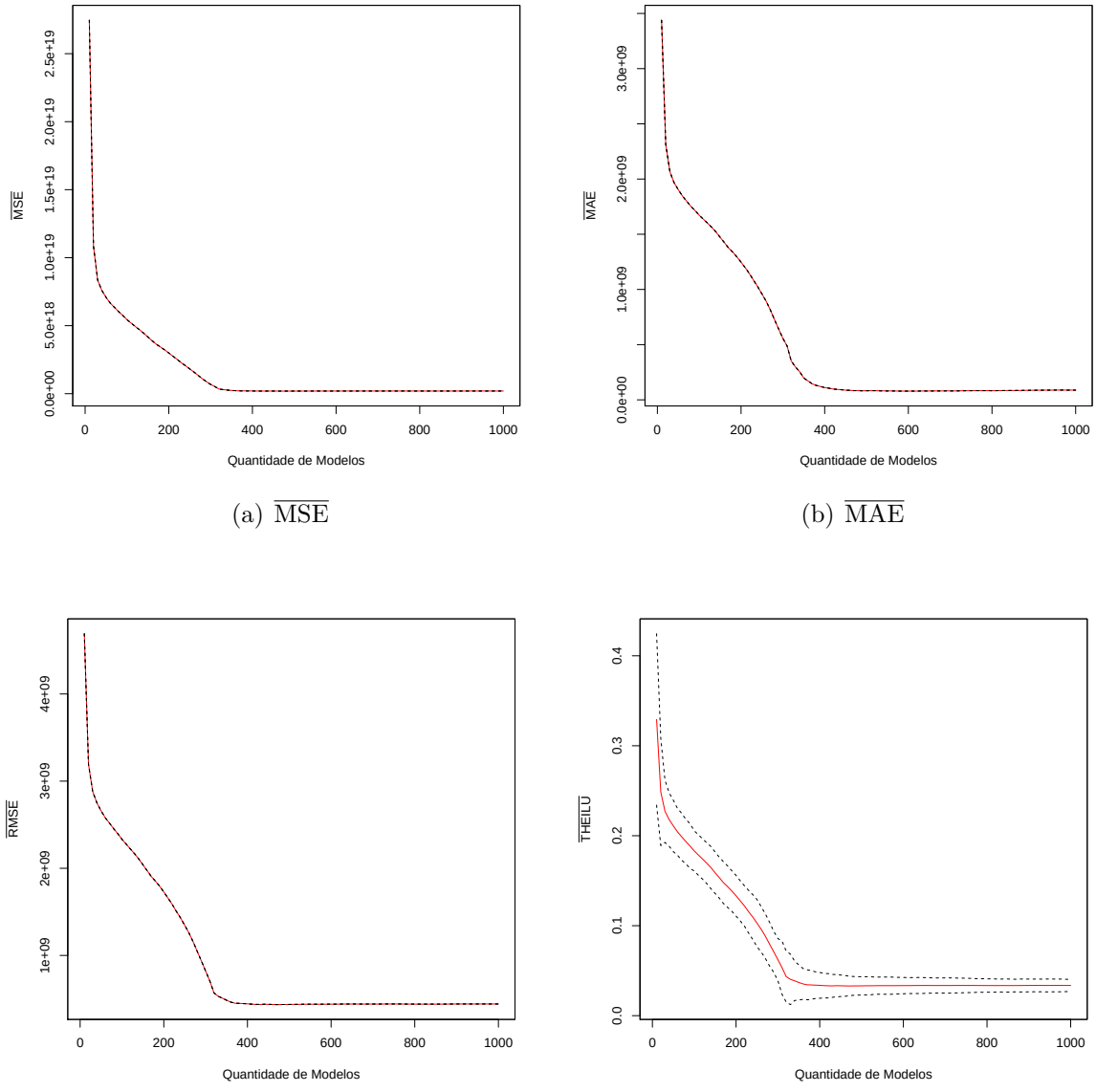
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 6 – CB para prever a série temporal ND (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

confiança na medida que o valor de k se aproxima de 1000.

As Figuras 11 e 12 apresentam os resultados obtidos para as seguintes séries temporais RO e RF. O decaimento do erro para essas séries ocorreu de maneira significativa, ainda com poucos modelos combinados, isto é, observando o $\overline{\text{MSE}}$ pode ser notado um decaimento com $k < 200$, o que pode sugerir uma quantidade específica para esta série, uma vez que $k > 200$ obtém erro aproximadamente constante para a série RF, enquanto que para RO os resultados sugerem um diminuição do erro assintotando o ponto com



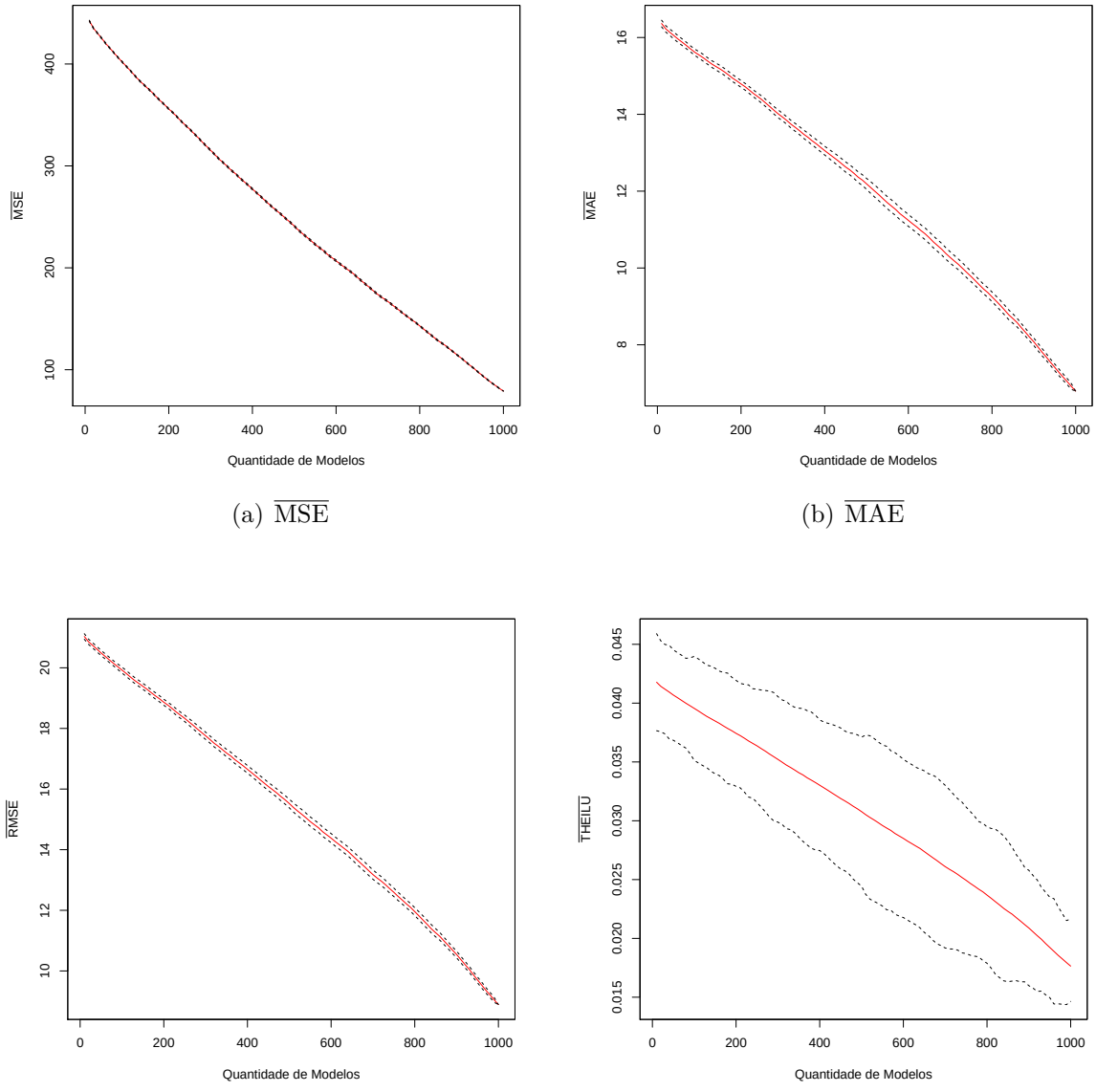
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 7 – CB para prever a série temporal DJ (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

métrica igual a zero. Contudo, uma análise estatística para investigar a quantidade ótima de k se faz necessária e será apresentada na Seção 5.2. O intervalo de confiança para ambas as séries apresenta uma curta distância entre os limites para todos os experimentos.

A Figura 13 para a série TO mostra uma variação com aproximadamente $k = 400$ para todas as métricas. Uma vez que os resultados demostraram correlação negativa, o eixo $\overline{\text{MSE}}$ apresenta valores relativamente pequenos (isto é, inferior a 1.5). É razoável assumir que estas variações ao longo das combinações não representa alterações significativas, dado



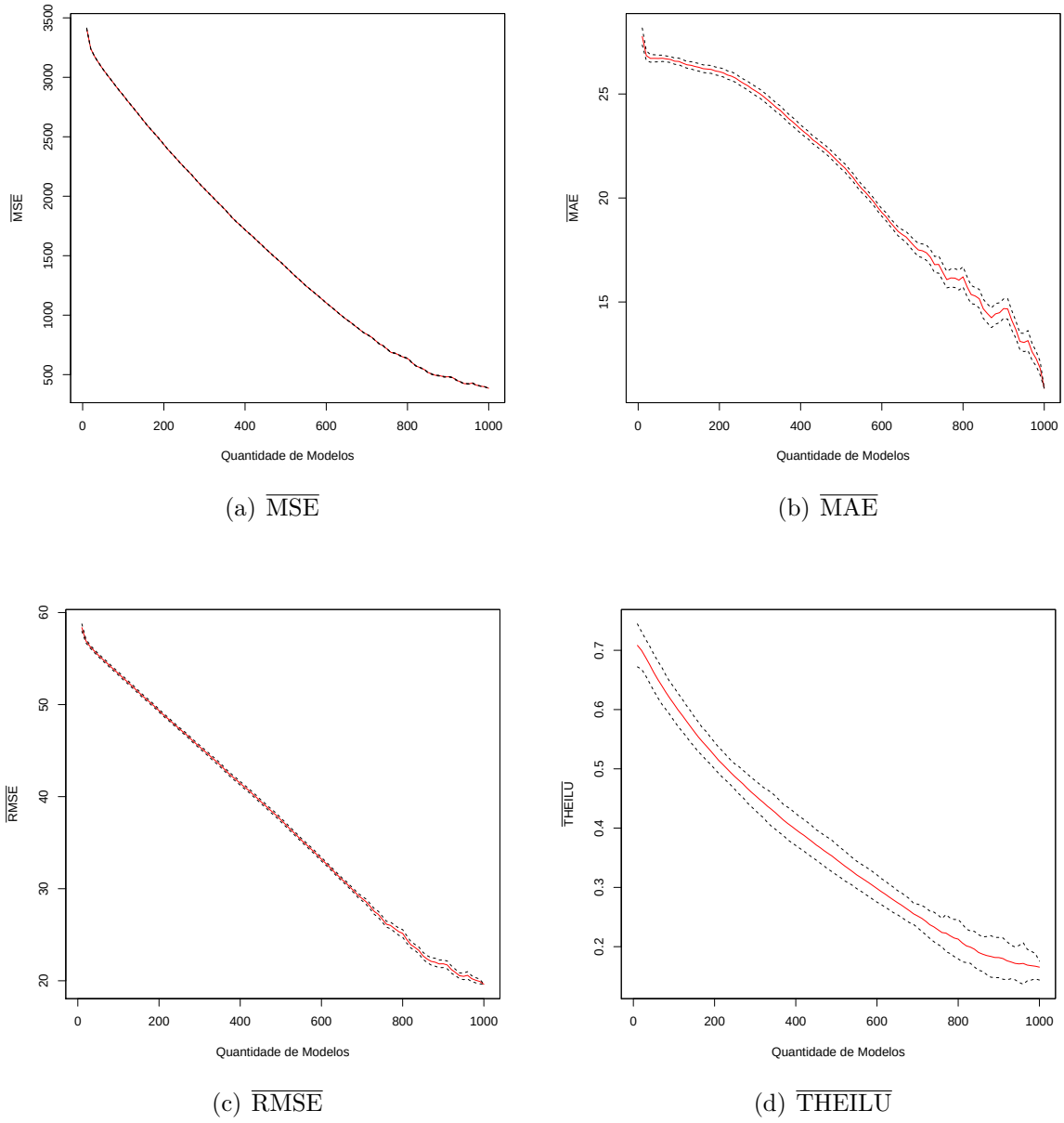
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 8 – CB para prever a série temporal QB (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

que o valor da métrica é relativamente pequeno.

Os resultados ilustrados na Figura 14 mostram a eficiência de CB para prever a série PO. Os resultados mostram que para esta série também foi obtido um comportamento assintótico, ou seja, quanto maior o valor de k mais próximo o desempenho de CB se aproxima do ponto com métrica zero. Contudo, os experimentos comprovam que para $k < 400$ já é possível obter uma métrica próxima de zero. Cabe destacar o intervalo de confiança que se mostra constante ao longo das combinações. Apesar disso, o intervalo é



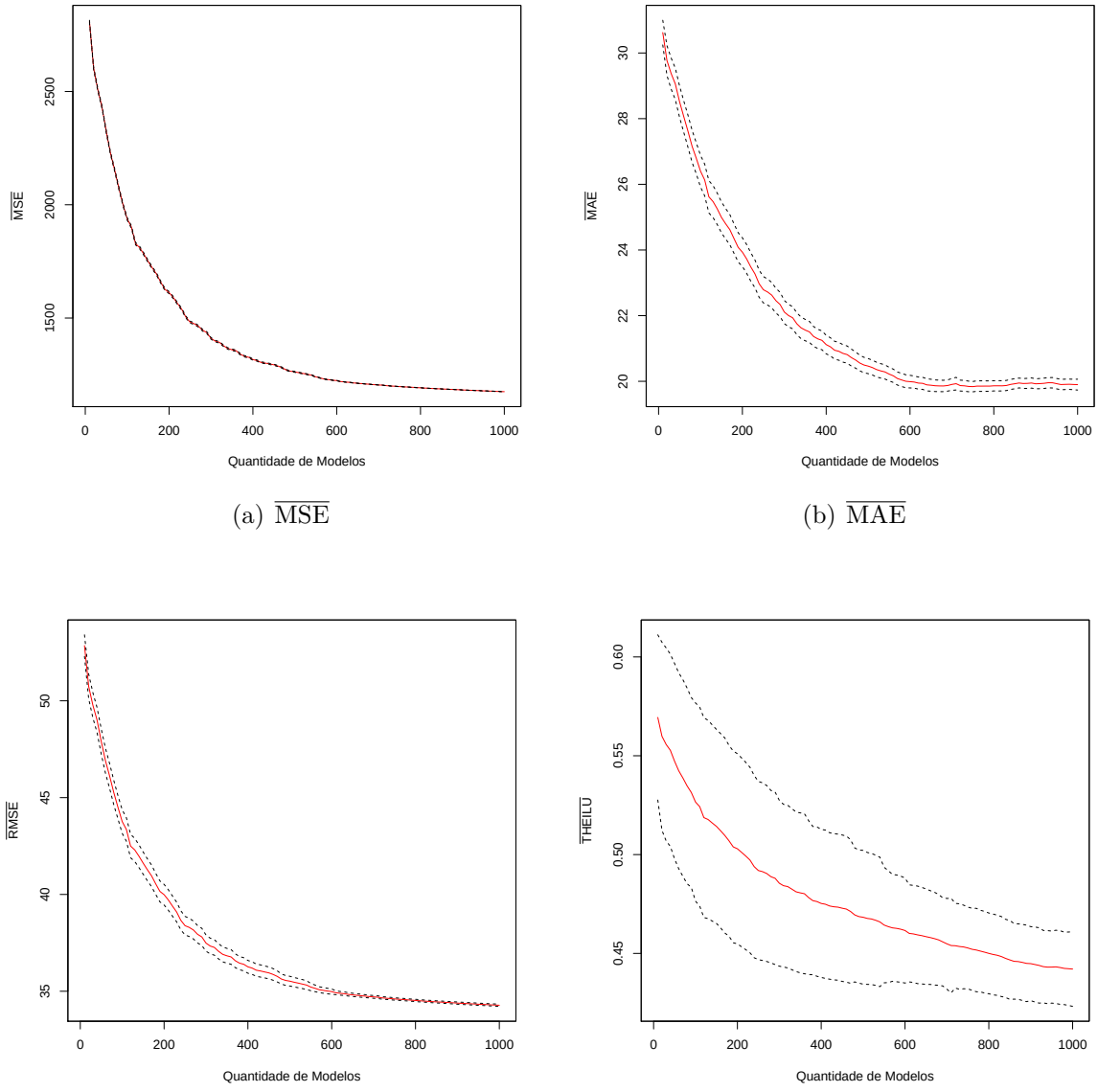
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 9 – CB para prever a série temporal PM (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

razoavelmente curto entre os limites.

De maneira geral, os resultados apresentados (veja as figuras citadas anteriores) sugerem que a quantidade de modelos individuais impactou na qualidade de CB. Em alguns casos a acurácia e eficiência de CB são substancialmente melhoradas com poucos modelos agregados, por outro lado, para outras séries o desempenho de CB é linearmente ou exponencialmente aprimorado.



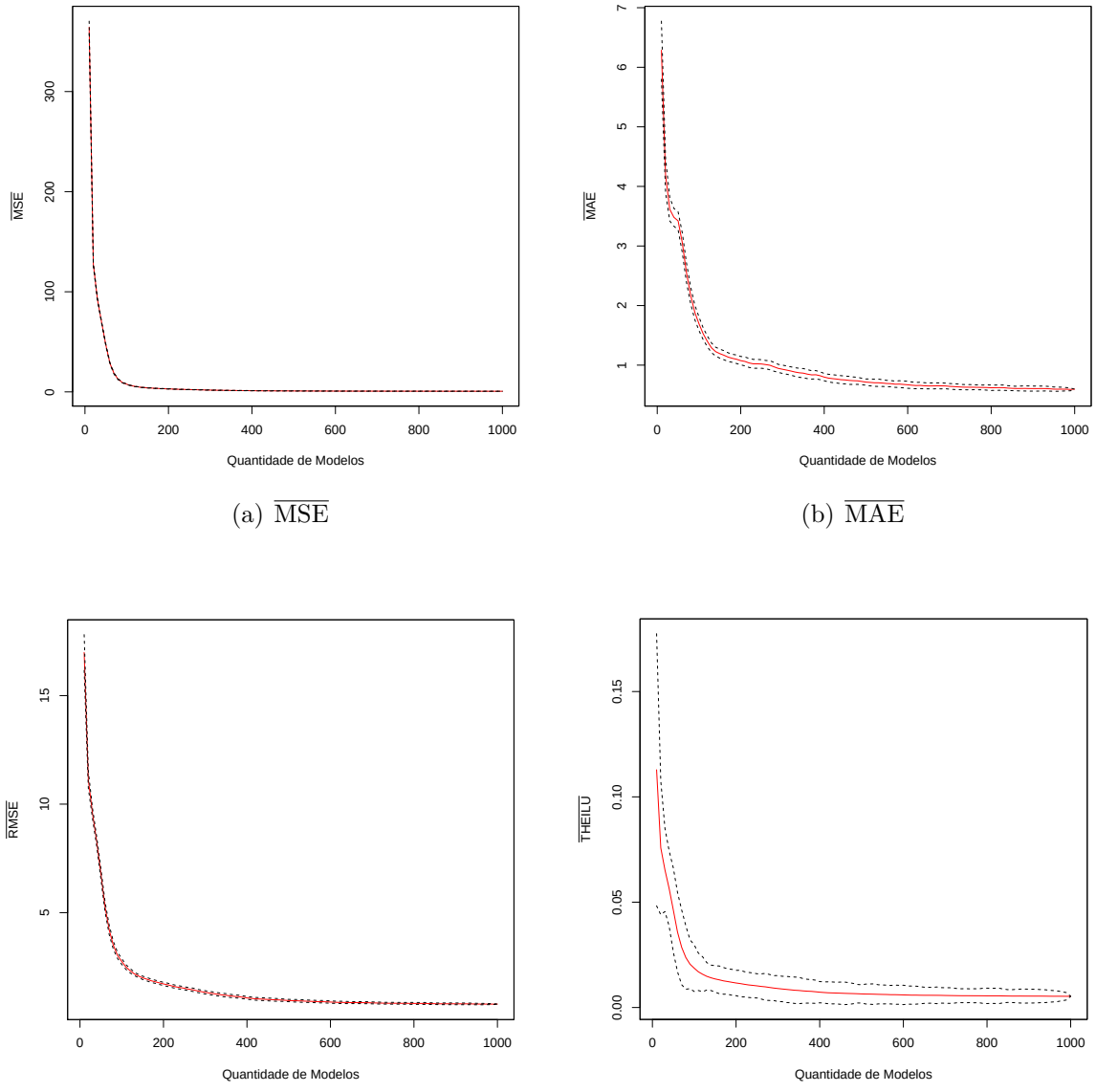
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 10 – CB para prever a série temporal RS (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

5.2 Regressão Linear para Estimar o Erro do Método Proposto

Os resultados apresentados anteriormente neste trabalho não são capazes de informar uma quantidade ótima de modelos individuais para ser adotado na agregação. Assim, também não foram mostradas evidências de que quantidades maiores para k representam um erro ainda menor. Logo, para sanar tais questões, requer a aplicação do modelo de regressão linear. Neste sentido, como anteriormente, foram mostrados os resultados do teste de hipótese para o coeficiente de correlação de Spearman e Kendall, que apontaram a existência



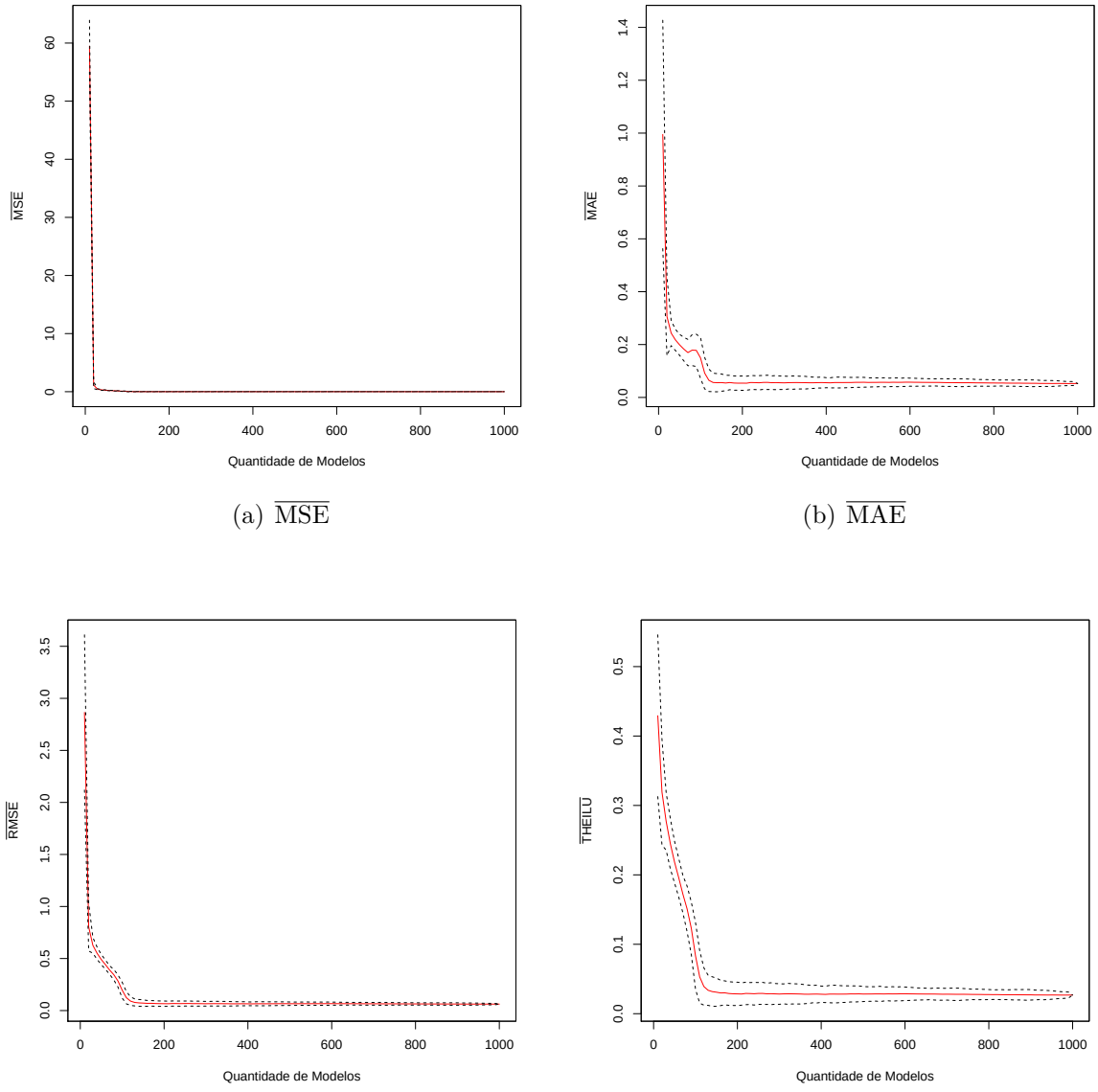
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 11 – CB para prever a série temporal RO (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

de correlação negativa entre k e o erro (métrica adotada). Desta forma, os resultados que serão aplicados para o modelo de regressão linear terão a finalidade de ilustrar o comportamento de CB para valores de k maiores que os apresentados anteriormente. Além disso, a análise de regressão proporciona a possibilidade de estimar o erro obtido por CB através da quantidade de modelos individuais que será considerado na combinação.

Nesse sentido, foram consideradas duas modelagens de regressão linear simples. A primeira intitulada como Linear (LI), ilustrada anteriormente na Equação 4.5, essa



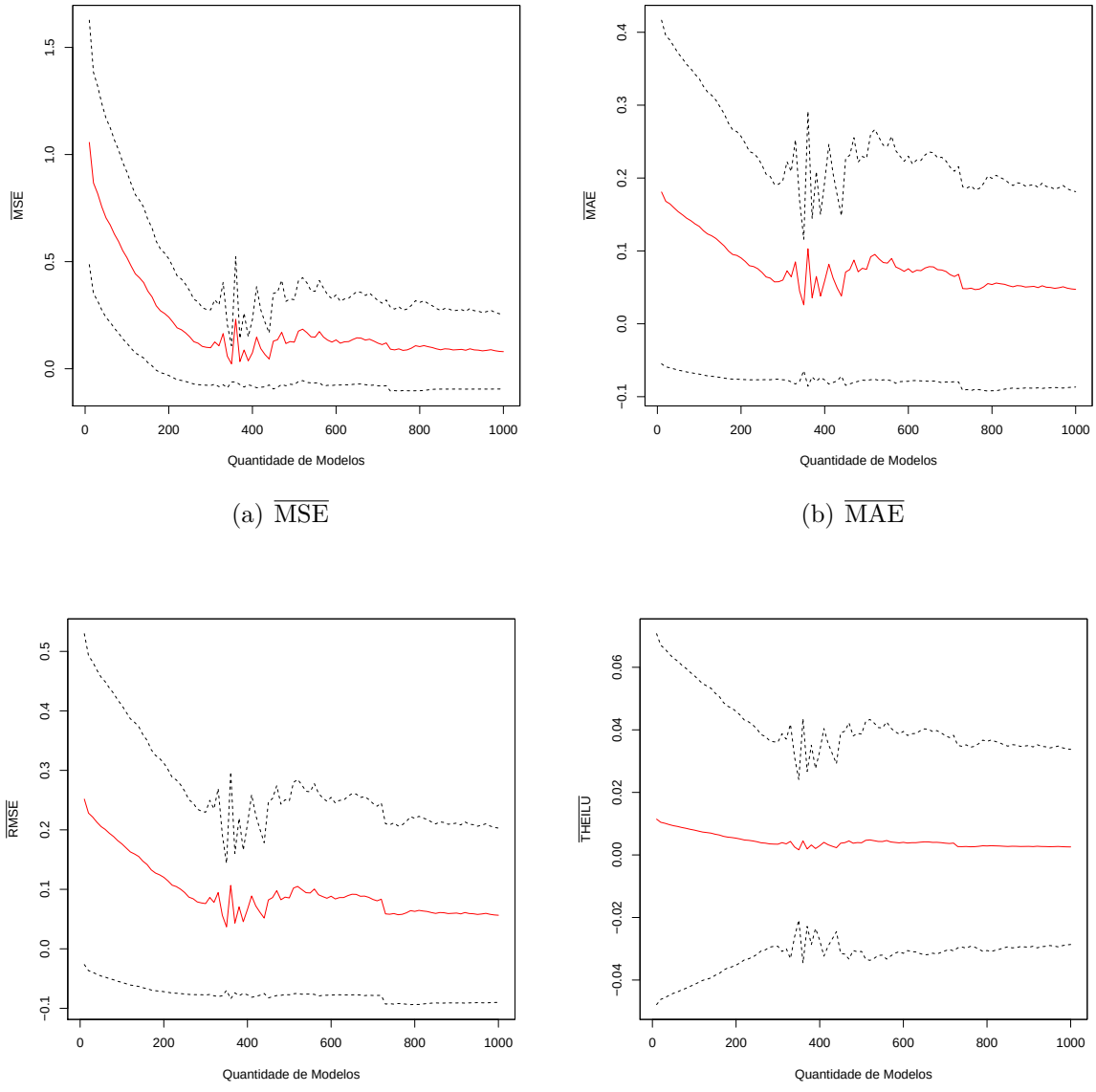
(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 12 – CB para prever a série temporal RF (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

é uma modelagem geralmente utilizada quando a relação é aproximadamente linear. A segunda foi intitulada como Log-Linear (LL), ilustrada via Equação 4.6, essa aplica o log na variável métrica ($\overline{\text{RMSE}}$, por exemplo) para calcular o parâmetro b, o que geralmente proporciona uma relação aproximadamente log-linear.

Portanto, foram aplicados dois modelos de regressão linear simples para estimar o erro cometido por CB quando for combinada uma determinada quantidade de modelos individuais. Este experimento envolveu apenas quatro séries temporais sendo: SP, ND,

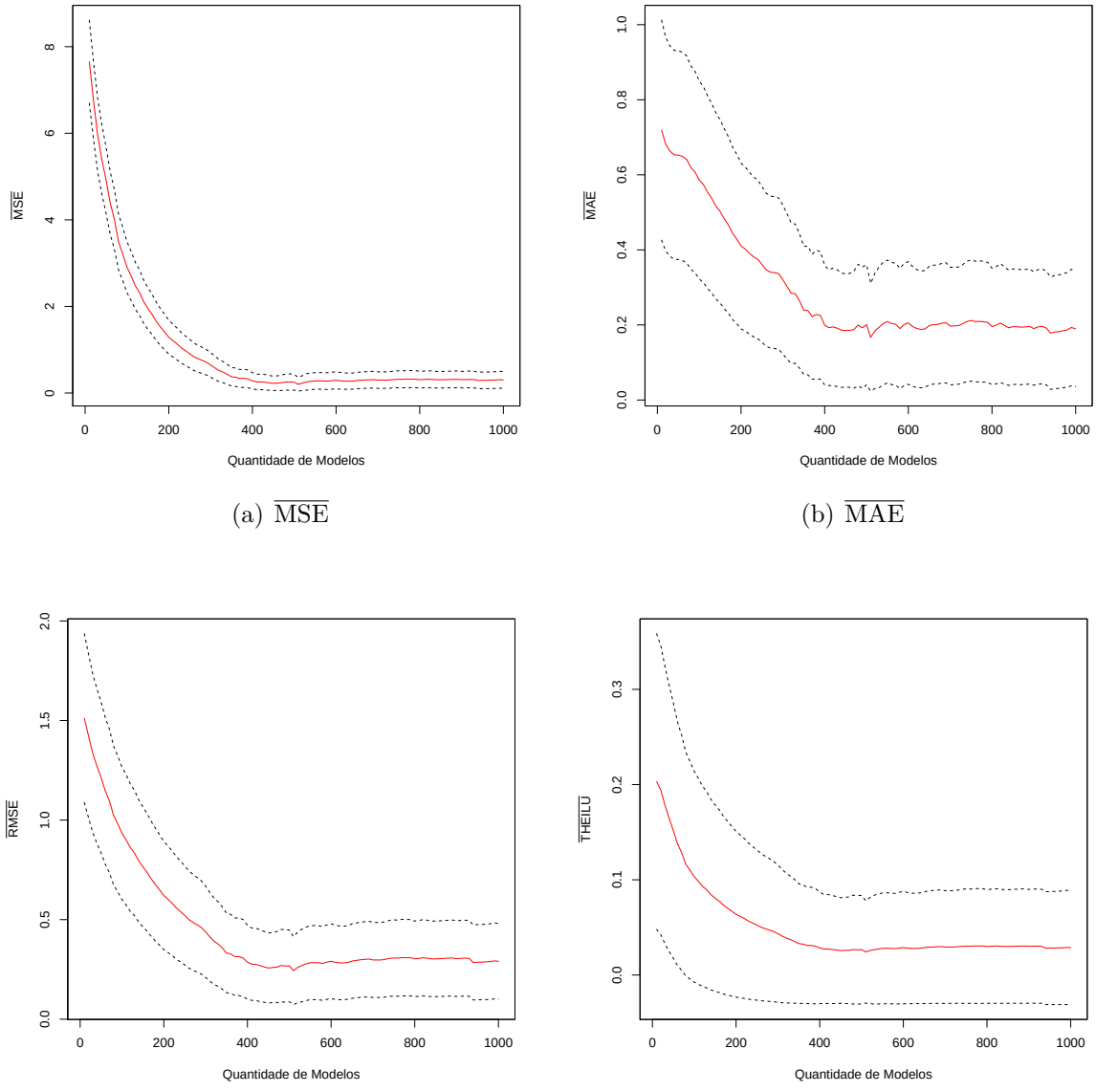


(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 13 – CB para prever a série temporal TO (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

QB e PM. Cabe ressaltar que pode ser necessário modelar diferentes tipos de modelos de regressão para ajustar a reta aos dados. No intuito de ilustrar um meio para possibilitar a estimação do erro obtido por CB para quantidades diversas de k, este trabalho apresenta dois modelos de regressão que melhor se ajustaram a estas quatro séries temporais. Contudo, a estimação do erro não se limita apenas as séries apresentadas, isto é, para DJ, RS, RO, TO, RF e PO pode-se estimar os erros também, mas para tanto se faz necessário a modelagem de outros modelos de regressão.



(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

Figura 14 – CB para prever a série temporal PO (conjunto de teste). As linhas tracejadas representam os intervalos de confiança (superior e inferior), enquanto que a linha vermelha (sólida) indica o erro de CB em cada uma das métricas em função de k.

A Tabela 4 apresenta o desempenho dos modelos de regressão Linear e Log-Linear para a série temporal SP. Conforme pode ser observado, os resultados sugerem que o modelo Log-Linear para todas as métricas é mais acurado em relação ao Linear. Desta forma, o modelo de regressão linear LL alcançou melhor desempenho em relação ao LI para a série SP. Logo, por meio do LL é possível obter o $\widehat{\overline{\text{MSE}}}$ informando como entrada da função de regressão o valor de k. Neste sentido, para obter o erro cometido por CB para uma determinada quantidade qualquer de modelos individuais, basta passar o valor de k

para a função retornar o valor estimado da métrica. Essa abordagem mostra-se interessante uma vez que é possível estimar o erro sem a necessidade de realizar as combinações para valores desconhecidos de k.

O desempenho dos modelos LI e LL para a série temporal ND também são apresentados na Tabela 4. Assim, como para a série SP o modelo LL mostrou melhor acurácia em todas as métricas. Neste sentido, os resultados indicam que o modelo LL será mais eficiente para estimar o erro cometido por CB para valores desconhecidos de k. Os modelos de regressão linear ajustados a série temporal ND mostram que o modelo LL é superior para todos os casos apresentados. Estes resultados indicam que o modelo LL é mais eficiente quando a inclinação da reta é similar a uma associação exponencial, favorecendo a modelagem desse modelo.

A Tabela 4 mostra ainda o desempenho dos modelos de regressão linear LI e LL ajustados a série temporal QB. Os resultados sugerem que o modelo LI para este ensaio apresenta um melhor ajuste em relação ao LL. Neste sentido, o erro quadrático médio para cada uma das métricas foi menor para LI, como pode ser notado pelo $\widehat{\overline{\text{MSE}}}$ para LL (322.047) e LI (37.453). Dado que os resultados alcançados por CB são lineares, isto é, o erro apresenta um comportamento bastante linear, o modelo LI que não possui qualquer ajuste para aproximação da linearidade dos dados, ao contrário de LL, que usa a função log para auxiliar na linearidade dos dados. Assim, possibilita ao modelo LI se ajustar melhor aos dados, sendo estatisticamente mais eficiente quando os dados são lineares.

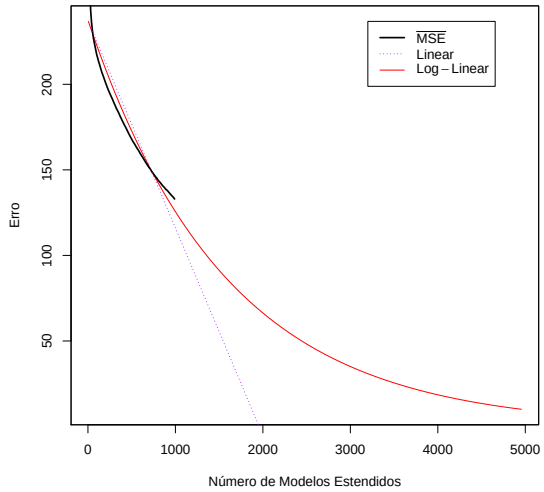
Por fim, a Tabela 4 mostra o desempenho dos modelos de regressão ajustado a série temporal PM. Neste ensaio em especial, obtivemos resultados que mostram situações em que o modelo LI é melhor que LL, além de uma situação oposta, com LL sendo melhor que LI. De maneira geral, LI mostra-se mais acurado através das métricas $\widehat{\overline{\text{MSE}}}$ com LI (18627.350) enquanto LL (22278.930), $\widehat{\overline{\text{MAE}}}$ com LI (0.50071) e LL (1.4394), e $\widehat{\overline{\text{RMSE}}}$ com 0.30299 e 2.46969 para LI e LL, respectivamente. Contudo, para a métrica $\widehat{\overline{\text{THEILU}}}$ o modelo LL foi melhor em relação ao LI.

Tabela 4 – Erro quadrático médio dos modelos LI e LL para estimar as métricas $\overline{\text{MSE}}$, $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$ (conjunto de teste).

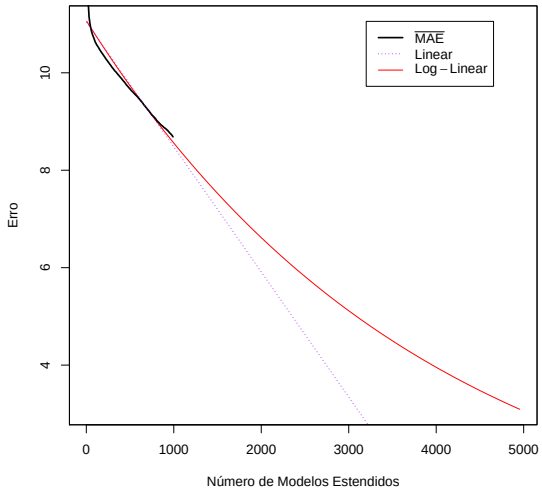
Séries	Métricas e Modelos de Regressão							
	$\widehat{\overline{\text{MSE}}}$		$\widehat{\overline{\text{MAE}}}$		$\widehat{\overline{\text{RMSE}}}$		$\widehat{\overline{\text{THEILU}}}$	
	LI	LL	LI	LL	LI	LL	LI	LL
SP	786.3889	**732.5744**	0.13974	**0.13254**	0.43435	**0.40492**	5.3448e-08	**4.9850e-08**
ND	112636.32	**2601.31**	2.10574	**0.22837**	6.73537	**0.26198**	2.6657e-07	**1.0001e-08**
QB	**37.453**	322.047	**0.03482**	0.24016	**0.03232**	0.28937	**1.179174e-07**	1.11955e-06
PM	**18627.350**	22278.930	**0.50071**	1.4394	**0.30299**	2.46969	0.00075	**4e-05**

A Figura 15 ilustra o desempenho dos modelos de regressão para a série SP. O valor de k adotado para esses ensaios foram 5 vezes maiores (isto é, $k <= 5000$) em relação

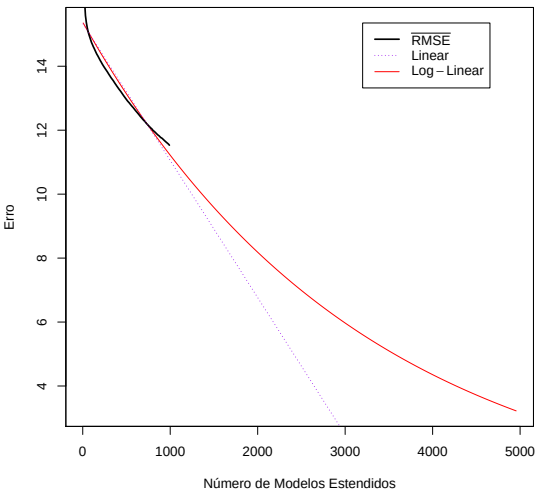
aos experimentos mostrados anteriormente. A linha preta na figura, representa a métrica, enquanto que a linha vermelha indica o modelo LL e a linha roxa tracejada indica o modelo LI. Os resultados sugerem que à medida que mais modelos individuais são agregados a combinação, menor será o erro cometido por CB. Assim, pode ser razoável dizer que a quantidade de modelos individuais influencia na qualidade da previsão combinada. Neste ensaio, para $k = 1000$ o $\widehat{\overline{\text{MSE}}} > 100$ enquanto que para $k > 3000$ o erro é substancialmente reduzido, chegando ao $\widehat{\overline{\text{MSE}}} < 50$.



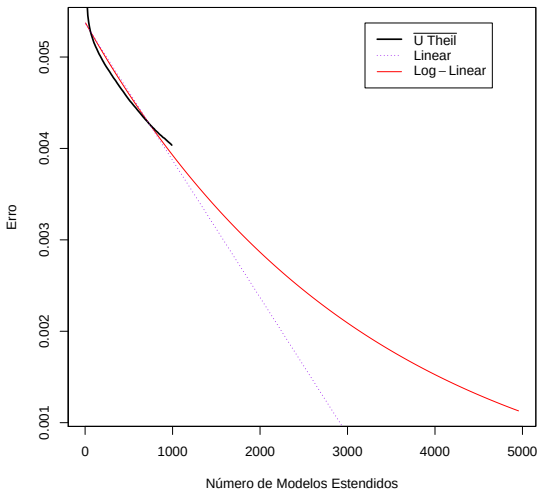
(a) $\widehat{\overline{\text{MSE}}}$



(b) $\widehat{\overline{\text{MAE}}}$



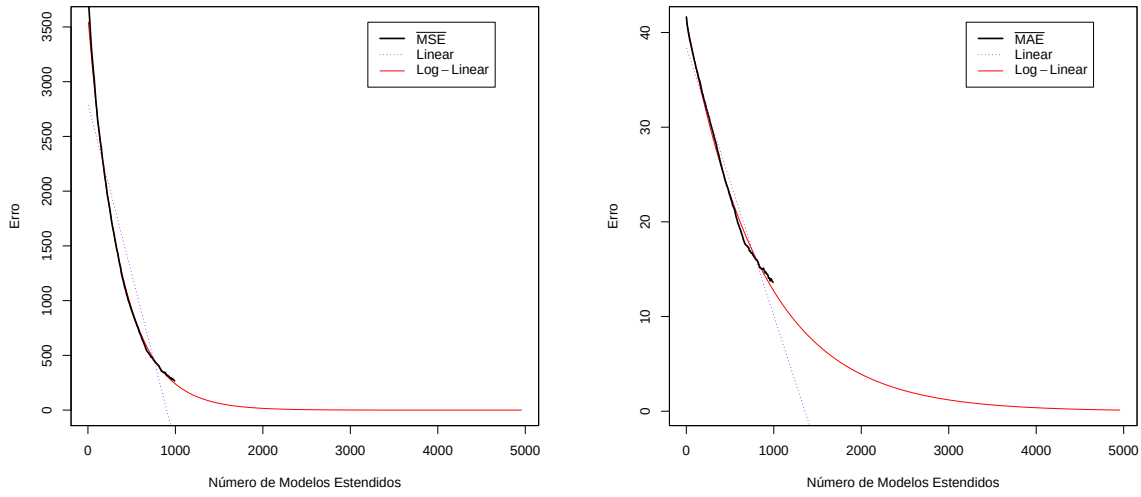
(c) $\widehat{\overline{\text{RMSE}}}$



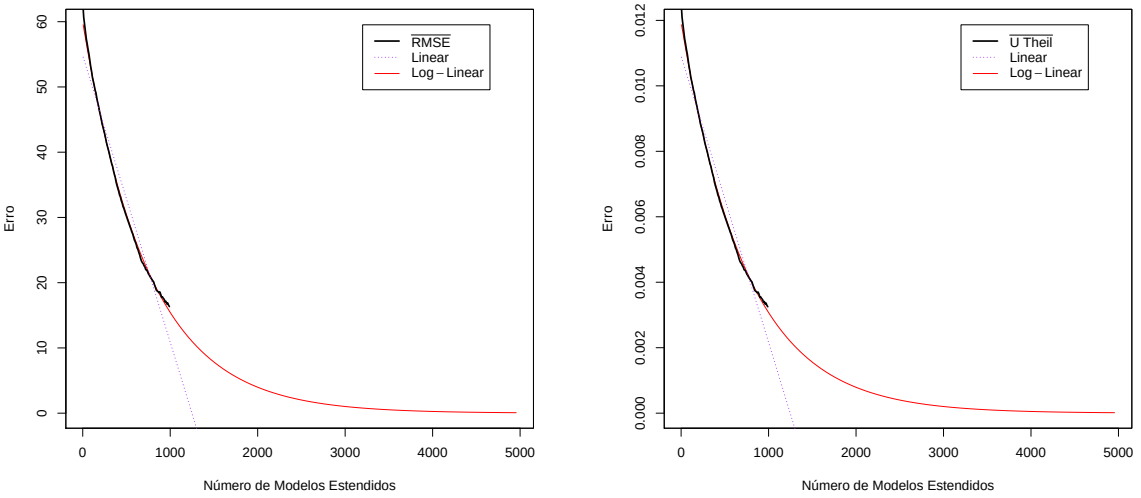
(d) $\widehat{\overline{\text{THEILU}}}$

Figura 15 – LI e LL para estimar o erro cometido por CB para a série temporal SP (conjunto de teste). A linha preta representa o valor da métrica (erro) em questão, a linha vermelha indica a estimativa de LL, enquanto que a linha roxa tracejada simboliza a estimativa de LI.

A Figura 16 ilustra os modelos LI e LL ajustados a série temporal ND. É visível que o modelo LL apresenta melhor ajuste a série ND para todas as métricas. De maneira geral, os resultados sugerem que a quantidade de modelos individuais influencia na eficiência de CB, uma vez que quanto maior a quantidade de modelos menor é o erro. A figura ilustra o comportamento do erro ao longo das agregações de mais modelos individuais e algo que pode ser observado é que o erro é assintótico ao eixo zero, ou seja, para $\widehat{\overline{\text{MAE}}}$, por exemplo, o erro ao longo das combinações se aproxima de zero tendendo ao infinito. Este mesmo comportamento aparece claramente para as demais métricas.



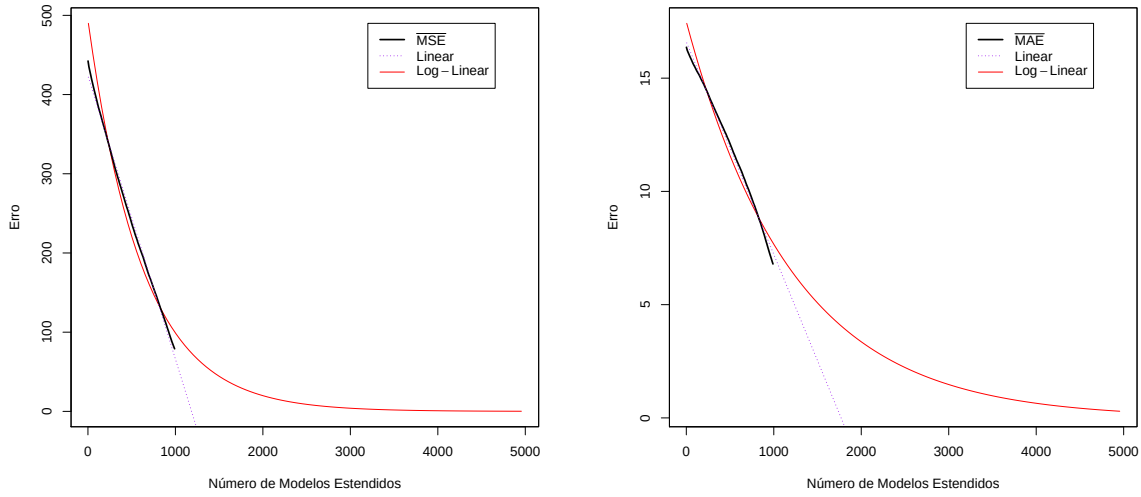
(a) $\widehat{\overline{\text{MSE}}}$ (b) $\widehat{\overline{\text{MAE}}}$



(c) $\widehat{\overline{\text{RMSE}}}$ (d) $\widehat{\overline{\text{THEILU}}}$

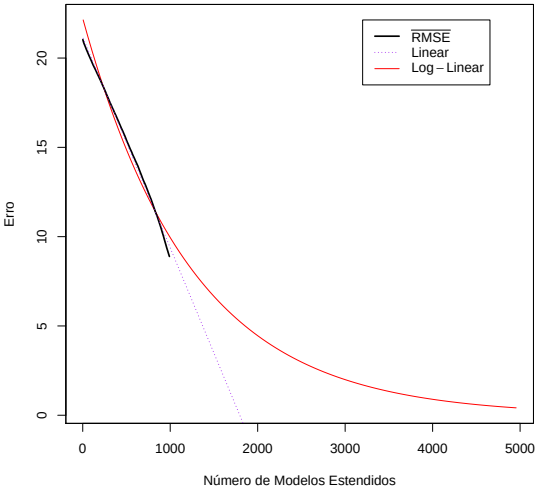
Figura 16 – LI e LL para estimar o erro cometido por CB para a série temporal ND (conjunto de teste). A linha preta representa o valor da métrica (erro) em questão, a linha vermelha indica a estimativa de LL, enquanto que a linha roxa tracejada simboliza a estimativa de LI.

A Figura 17 ilustra o desempenho dos modelos LI e LL ajustados a série QB. Conforme pode ser visto, o modelo LI mostra-se mais ajustado aos dados, como discutido anteriormente, isso ocorre porque a métrica possui um comportamento bastante linear. Apesar das métricas $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$ apresentarem uma suave curvatura, isso não impede que LI tenha um desempenho melhor que LL, e que dispense transformações de linearidade dos dados.

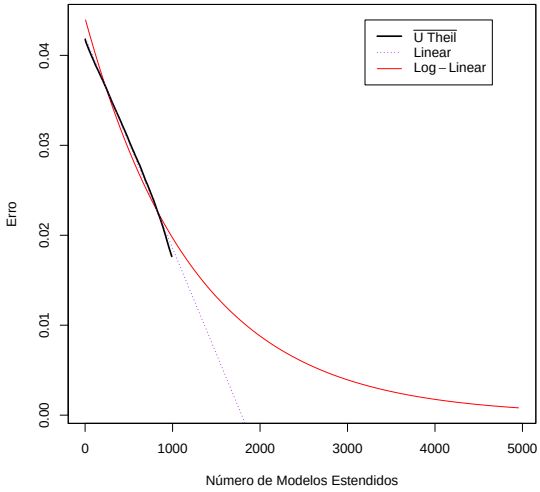


(a) $\widehat{\overline{\text{MSE}}}$

(b) $\widehat{\overline{\text{MAE}}}$



(c) $\widehat{\overline{\text{RMSE}}}$

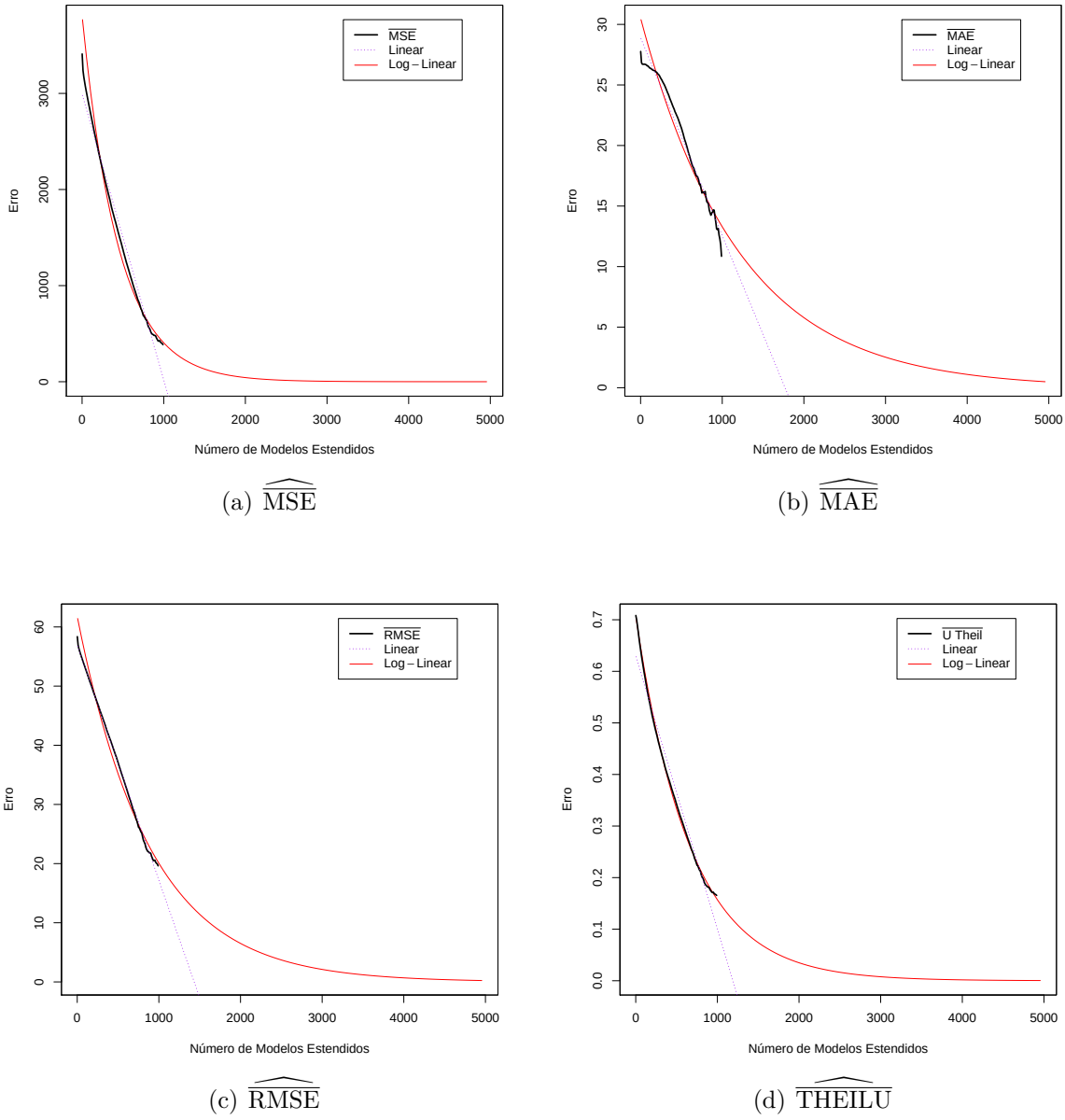


(d) $\widehat{\overline{\text{THEILU}}}$

Figura 17 – LI e LL para estimar o erro cometido por CB para a série temporal QB (conjunto de teste). A linha preta representa o valor da métrica (erro) em questão, a linha vermelha indica a estimativa de LL, enquanto que a linha roxa tracejada simboliza a estimativa de LI.

Os resultados para PM através das estimativas $\widehat{\overline{\text{MSE}}}$, $\widehat{\overline{\text{MAE}}}$ e $\widehat{\overline{\text{RMSE}}}$ apresenta

um comportamento semelhante ao obtido para estimar a série QB, em que os dados são aproximadamente lineares (veja Figura 18), o que possibilita que LI alcance um desempenho melhor em relação ao LL.



(a) $\widehat{\overline{\text{MSE}}}$ (b) $\widehat{\overline{\text{MAE}}}$

(c) $\widehat{\overline{\text{RMSE}}}$ (d) $\widehat{\overline{\text{THEILU}}}$

Figura 18 – LI e LL para estimar o erro cometido por CB para a série temporal PM (conjunto de teste). A linha preta representa o valor da métrica (erro) em questão, a linha vermelha indica a estimativa de LL, enquanto que a linha roxa tracejada simboliza a estimativa de LI.

Desta forma, o modelo LL mostra-se mais eficiente em relação a LI quando os dados apresentam curvaturas e podem ser transformados em aproximadamente lineares pelo processo de linearidade. Por outro lado, quando os dados são lineares o modelo LI apresenta melhor desempenho que LL.

Resumidamente, os resultados mostram a eficiência de dois modelos de regressão linear para estimar o erro cometido por CB para uma quantidade k de modelos individuais. Assim, por meio dos modelos LI ou LL é possível estimar o erro de CB assumindo um valor para k.

É razoável dizer que através destes resultados é factível responder a questão: É possível continuar melhorando a acurácia de CB aumentando o valor de k?. Assim, os resultados sugerem para as séries SP, ND, QB e PM que é possível obter uma acurácia ainda maior para CB na medida que k cresce. Uma vez que já foi comprovada a existência de correlação entre a métrica e k, também é possível por meio da regressão linear estimar o erro cometido por CB através de k.

A Tabela 5 ilustra os resultados para os modelos LI e LL para estimar $\hat{k}$ para as métricas abordadas neste trabalho. Nesses experimentos os modelo LI e LL foram implementados com algumas diferenças em relação ao que foi mostrado anteriormente, aonde LI é dado pela Equação 4.7, enquanto que LL é definido pela Equação 4.8. Os resultados mostram que LL é melhor ajustado e consequentemente superior para as séries SP e ND, por outro lado, LI mostra-se superior para série QB onde o erro tem um comportamento próximo da linearidade, o que não acontece para as outras séries. A série PM mostra que para as métricas $\overline{\text{MAE}}$ e $\overline{\text{RMSE}}$, LI foi melhor e para $\overline{\text{MSE}}$ e $\overline{\text{THEILU}}$ o modelo LL apresentou-se mais ajustado. Quando os dados não seguem um comportamento linear é natural que o modelo LL se ajuste melhor, como o que ocorreu neste experimento.

Tabela 5 – Erro quadrático médio dos modelos LI e LL para estimar k (conjunto de teste).

Séries	$\hat{k}$ e os Modelos de Regressão							
	$\hat{k}^{(\overline{\text{MSE}})}$		$\hat{k}^{(\overline{\text{MAE}})}$		$\hat{k}^{(\overline{\text{RMSE}})}$		$\hat{k}^{(\overline{\text{THEILU}})}$	
	LI	LL	LI	LL	LI	LL	LI	LL
SP	102375.7	**39663.7**	45638.3	**26684.4**	56307.7	**33170.6**	56523.1	**33288.0**
ND	53378.6	**1163.51**	19169.2	**3281.65**	20969.7	**1163.36**	20914.0	**1144.60**
QB	**1722.01**	27270.8	**4044.15**	33948.2	**2456.51**	27159.7	**2254.91**	26334.8
PM	14684.1	**7589.98**	**2610.18**	22050.4	**1224.46**	7894.82	15165.1	**728.45**

O Anexo A apresenta duas tabelas contendo os parâmetros utilizados pelos modelos LI e LL para realizar os experimentos ilustrado na Tabela 5. Trata-se dos parâmetros a e b, em que para estimar tais parâmetros foram obtidos do conjunto de treinamento de cada uma das séries temporais adotadas nesta seção.

5.3 Comparação entre SA, ME, WA, MO e CB

Os ensaios que serão apresentados a seguir, mostram uma comparação entre CB e os métodos Média Simple (SA), Mediana (ME), Média Ponderada (WA) e Moda (MO). Tais comparações são realizadas em termos das métricas previamente apresentadas, sendo $\overline{\text{MSE}}$,

$\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$. Seguindo a mesma metodologia dos experimentos anteriores, isto é, agregando de dez em dez modelos individuais até combinar mil modelos ($k = 10, 20, 30, \cdots, 1000$). Apesar deste trabalho não ter como objetivo comparar os resultados modelos combinados com os modelos individuais, um experimento envolvendo a comparação entre os modelos individuais (MI) e o melhor MI com os *ensembles* utilizados neste trabalho são apresentados no Anexo B.

O método SA consiste na soma das previsões de todos os preditores dividido pelo número de preditores, conforme foi descrito anteriormente na Equação 2.18. O método ME inicialmente ordena todas as previsões dos modelos individuais e escolhe a previsão localizada na posição do meio (veja Seção 2.5.3). O WA trata-se da média ponderada, isto é, a previsão de cada modelos individual será multiplicado por um peso, assim, cada modelo terá um peso atribuído a sua previsão com base na sua acurácia, as Equações 2.19 e 2.20 descrevem matematicamente este método. Por fim, o método MO calcula a moda das previsões dos modelos individuais, contudo, as previsões são números reais que muitas vezes para estes ensaios possuem seis casas decimais, o que na prática, não contém observações repetidas, desta forma, tornando o conjunto de k previsões sem moda. Neste sentido, para resolução deste problema, o método MO calcula a moda de Czuber que consiste em agrupar o conjunto de previsões em classes dividas em amplitudes iguais, assim, a moda será calculada com base na classe modal como apresentado na Seção 2.5.4 e descrita pela Equação 2.26.

A Tabela 6 apresenta os resultados do teste de hipótese para o coeficiente de correlação de Spearman para os métodos de Média Simples (SA), Mediana (ME), Média Ponderada (WA) e Moda (MO). Os valores indesejados (insatisfatórios) estão destacados em negrito. Assim, os resultados sugerem especificamente para métrica $\overline{\text{THEILU}}$ da série PM, que todos os métodos possuem correlação positiva, indicando que a medida que k cresce o erro aumenta proporcionalmente. Esse comportamento é indesejado, uma vez que, o objetivo é diminuir o erro a medida que k cresce, ou seja, uma correlação negativa entre as variáveis. O método ME não passou no teste de hipótese para o coeficiente de correlação de Spearman para a métrica $\overline{\text{MAE}}$ da série RS (com *p-value* igual a 0.09177), o que indica que mesmo o ρ sendo negativo não existe significância estatística para comprovar que esse resultado não seja uma mera causalidade. O método MO apresentou resultados indesejados para todas as métricas da série RS, isto é, ρ positivo, indicando uma correlação positiva entre o erro e o número de modelos individuais, tal resultado ainda aponta um nível de significância de 5% o que torna ainda mais insatisfatório o uso deste método para a série temporal em questão.

Para a série RF, novamente os resultados quase se repetem, isto é, a maioria dos métodos apresentaram uma correlação positiva para a métrica $\overline{\text{THEILU}}$. Contudo, WA não passou no teste de hipótese (com *p-value* de 0.9784), enquanto que SA também não

passou no teste para $\overline{\text{RMSE}}$ (com *p-value* de 0.3757). Os resultados também mostram que ME tem correlação positiva para $\overline{\text{RMSE}}$. Porém, de modo geral, o método MO apresentou correlação negativa para todas as métricas. Já o método ME apresentou correlação positiva para a série PO via $\overline{\text{THEILU}}$.

Tabela 6 – Teste de hipótese para o coeficiente de correlação de Spearman (ρ) entre as métricas obtidas via SA, ME, WA, MO e k. Os resultados indesejados (insatisfatórios) estão destacados em negrito (conjunto de teste).

Séries		Métricas							
		$\overline{\text{MSE}}$		$\overline{\text{MAE}}$		$\overline{\text{RMSE}}$		$\overline{\text{THEILU}}$	
		ρ	*p-value*	ρ	*p-value*	ρ	*p-value*	ρ	*p-value*
SP	SA	-0.93471	<2.2e-16	-0.27612	=0.00556	-0.52594	=3.1e-08	-0.53328	<1.879e-08
	ME	-0.97837	<2.2e-16	-0.92642	<2.2e-16	-0.94321	<2.2e-16	-0.94444	<2.2e-16
	WA	-0.99845	<2.2e-16	-0.99222	<2.2e-16	-0.99513	<2.2e-16	-0.99531	<2.2e-16
	MO	-0.99696	<2.2e-16	-0.99843	<2.2e-16	-0.99666	<2.2e-16	-0.99666	<2.2e-16
ND	SA	-0.75510	<2.2e-16	-0.73335	<2.2e-16	-0.73585	<2.2e-16	-0.73566	<2.2e-16
	ME	-0.96363	<2.2e-16	-0.96121	<2.2e-16	-0.95499	<2.2e-16	-0.95533	<2.2e-16
	WA	-0.81336	<2.2e-16	-0.71377	<2.2e-16	-0.74977	<2.2e-16	-0.75047	<2.2e-16
	MO	-0.82090	<2.2e-16	-0.88620	<2.2e-16	-0.81800	<2.2e-16	-0.81648	<2.2e-16
DJ	SA	-0.99471	<2.2e-16	-0.99678	<2.2e-16	-0.99446	<2.2e-16	-0.99599	<2.2e-16
	ME	-0.81348	<2.2e-16	-0.89646	<2.2e-16	-0.77066	<2.2e-16	-0.72882	<2.2e-16
	WA	-0.99973	<2.2e-16	-0.99937	<2.2e-16	-0.99971	<2.2e-16	-0.99914	<2.2e-16
	MO	-0.92558	<2.2e-16	-0.90738	<2.2e-16	-0.90754	<2.2e-16	-0.87207	<2.2e-16
QB	SA	-0.92049	<2.2e-16	-0.86678	<2.2e-16	-0.92034	<2.2e-16	-0.91621	<2.2e-16
	ME	-0.66078	<2.2e-16	-0.35287	0.000344	-0.65983	<2.2e-16	-0.64661	<2.2e-16
	WA	-0.75984	<2.2e-16	-0.61734	<2.2e-16	-0.75888	<2.2e-16	-0.76038	<2.2e-16
	MO	-0.62856	<2.2e-16	-0.68349	<2.2e-16	-0.62856	<2.2e-16	-0.65660	<2.2e-16
RS	SA	-0.25971	=0.00923	-0.26282	=0.00840	-0.25702	=0.01001	-0.27459	=0.005837
	ME	-0.50306	=1.467e-07	-0.16951	**=0.09177**	-0.503066	=1.467e-07	-0.75930	<2.2e-16
	WA	-0.89489	<2.2e-16	-0.76186	<2.2e-16	-0.89450	<2.2e-16	-0.88490	<2.2e-16
	MO	**0.6397**	<2.2e-16	**0.72039**	<2.2e-16	**0.64018**	<2.2e-16	**0.61678**	<2.2e-16
PM	SA	-0.97707	<2.2e-16	-0.96739	<2.2e-16	-0.97680	<2.2e-16	**0.95781**	<2.2e-16
	ME	-0.88902	<2.2e-16	-0.88068	<2.2e-16	-0.88702	<2.2e-16	**0.85480**	<2.2e-16
	WA	-0.86522	<2.2e-16	-0.82905	<2.2e-16	-0.86083	<2.2e-16	**0.86921**	<2.2e-16
	MO	-0.96442	<2.2e-16	-0.95964	<2.2e-16	-0.96442	<2.2e-16	**0.82581**	<2.2e-16
RO	SA	-0.95314	<2.2e-16	-0.95153	<2.2e-16	-0.91373	<2.2e-16	-0.91867	<2.2e-16
	ME	-0.99469	<2.2e-16	-0.99156	<2.2e-16	-0.99246	<2.2e-16	-0.99147	<2.2e-16
	WA	-0.99540	<2.2e-16	-0.99025	<2.2e-16	-0.99464	<2.2e-16	-0.99739	<2.2e-16
	MO	-0.98313	<2.2e-16	-0.98414	<2.2e-16	-0.98342	<2.2e-16	-0.98589	<2.2e-16
TO	SA	-0.99699	<2.2e-16	-0.98951	<2.2e-16	-0.99685	<2.2e-16	-0.99635	<2.2e-16
	ME	-0.97741	<2.2e-16	-0.98345	<2.2e-16	-0.97699	<2.2e-16	-0.97955	<2.2e-16
	WA	-0.99967	<2.2e-16	-0.99976	<2.2e-16	-0.99965	<2.2e-16	-0.99956	<2.2e-16
	MO	-0.93785	<2.2e-16	-0.91326	<2.2e-16	-0.93702	<2.2e-16	-0.93383	<2.2e-16
RF	SA	-0.31396	=0.00154	-0.74485	<2.2e-16	-0.08940	**=0.3757**	**0.69514**	<2.2e-16
	ME	-0.48322	=5.038e-07	-0.81663	<2.2e-16	**0.237455**	= 0.01756	**0.871599**	<2.2e-16
	WA	-0.72847	<2.2e-16	-0.76122	<2.2e-16	-0.712871	<2.2e-16	**0.002748**	**=0.9784**
	MO	-0.96518	<2.2e-16	-0.97082	<2.2e-16	-0.96423	<2.2e-16	-0.95552	<2.2e-16
PO	SA	-0.99605	<2.2e-16	-0.99918	<2.2e-16	-0.99885	<2.2e-16	-0.99885	<2.2e-16
	ME	-0.98333	<2.2e-16	-0.96786	<2.2e-16	-0.98316	<2.2e-16	**0.673327**	<2.2e-16
	WA	-0.99953	<2.2e-16	-0.99888	<2.2e-16	-0.99938	<2.2e-16	-0.99918	<2.2e-16
	MO	-0.97528	<2.2e-16	-0.97573	<2.2e-16	-0.97446	<2.2e-16	-0.92686	<2.2e-16

Assim, o teste de hipótese para o coeficiente de correlação de Spearman mostra que os erros obtidos via SA, ME, WA e MO não apresentam indícios de correlação negativa entre a métrica e a quantidade de modelos para todas as métricas aplicadas às séries temporais. Desta forma, os resultados sugerem, para os casos destacados, que os métodos SA, ME, WA e MO não são os melhores meios para combinar modelos individuais que permitam uma correlação negativa dos erros a medida que mais preditores são inseridos

na combinação. Desta forma, CB obteve um coeficiente de correlação negativo para todas as métricas e séries temporais em estudo, com um nível de significância de 5%, o que inicialmente comprova a superioridade de CB no aspecto da tendência de diminuição do erro ao longo do crescimento de k.

As Figuras 19, 20 21, 22, 23, 24, 25, 26, 27 e 28 ilustram as comparações entre os métodos SA, ME, WA, MO e CB para as séries temporais mencionadas na Tabela 6. Os ensaios adotaram as métricas $\overline{\text{MSE}}$, $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$ e $\overline{\text{THEILU}}$ para mensurar a acurácia dos métodos, tal que, o eixo "x" indica a quantidade de modelos individuais envolvidos na combinação, enquanto que o eixo "y" representa o valor da métrica.

A Figura 19 apresenta o desempenho do método SA, ME, WA e MO comparado com CB para a série temporal SP. Os resultados mostram a superioridade de CB em relação aos outros métodos para todas as métricas. Contudo, os resultados mostram que os métodos tem correlação negativa entre a métrica e k, o que sugere uma diminuição do erro a medida que mais preditores individuais são agregados. Os resultados sugerem que entre os métodos de combinação linear, o MO mostra-se com a alternativa mais acurada, em seguida o ME e posteriormente, WA e SA, apresentando ambos comportamentos bastante parecidos.

Os resultados alcançados por SA e WA são bastante parecidos. Cabe enfatizar que estes dois métodos se distinguem no processo adotado para obtenção dos pesos, isto é, apesar da equação para calcular o SA não apresentar pesos, estes existem implicitamente, ou seja, os pesos atribuídos são iguais para todos os modelos individuais (isto é, $w = 1/k$). Enquanto que para o método WA, os pesos são atribuídos através de equação matemática que visa atribuir os maiores pesos para os preditores que se mostrarem mais acurados. Os resultados apresentados neste trabalho, indicam para quase todas as séries temporais com exceção apenas de SP e PM que o método SA é mais acurado, ressaltando que nesta última série a superioridade foi constatada exclusivamente para a métrica $\overline{\text{THEILU}}$. Logo, o método WA na prática não conseguiu definir valores adequados para os pesos de modo que a acurácia fosse maximizada e consequentemente fosse melhor que atribuir pesos iguais para todos os preditores como ocorre com o método SA.

Os resultados apontados para a série temporal ND (veja Figura 20) mostram mais claramente a superioridade do método SA em relação ao WA. Apesar de diversos autores da literatura, relatarem que os métodos baseados na média ponderada serem, de modo geral, mais acurados em relação ao SA, os resultados deste trabalho mostram uma realidade diferente. Certamente, porque é necessário ter cuidado com a escolha do processo adotado para calcular os pesos, visto que este procedimento influencia nas previsões agregadas. Neste sentido, os resultados apontam que o procedimento adotado neste trabalho para calcular os pesos do WA não seria o ideal, visto que o procedimento não alcança uma qualidade superior, de maneira geral, para todos os experimentos em relação ao SA. Desta

forma, é razoável assumir que seria melhor adotar o processo de atribuir pesos iguais para todos os preditores, dado que o método SA é mais simples de implementar e exige um custo computacional menor. Por outro lado, vale destacar que o método MO, obteve o segundo melhor resultado e raramente esta metodologia é usada na prática.

Neste trabalho, o peso do método WA foi calculado conforme Equação 2.20 por uma questão apenas de simplicidade em termos de complexidade de implementação. Além disso, esse processo de calcular o peso foi citado em algumas obras da literatura. Contudo, existem diversas formas de calcular os pesos para o método WA, assim como são vários os autores que defendem a superioridade de WA em relação ao SA. Assim, fica clara a importância de escolher o procedimento adequado para calcular os pesos do método WA, pois caso este processo não seja adequado, as previsões agregadas podem ser comprometidas.

A Figura 20 ilustra o desempenho de SA, ME, WA, MO e CB para prever a série temporal ND. Conforme pode ser observado, CB apresenta um decaimento aproximadamente exponencial do erro, por outro lado, SA, ME, WA e MO mostram um erro aproximadamente constante. De modo geral, a melhoria de SA, ME, WA e MO é pouco notada, todavia o método SA para $k = 200$ o $\overline{\text{RMSE}} = 79.88610$ e $k = 1000$ o $\overline{\text{RMSE}} = 79.77672$, ou seja, uma diferença de 0.10938. Comparando com o CB a diferença para o mesmo intervalo de k foi de 29.97902, resultando em uma melhoria mais significativa. Entre os métodos clássicos de combinação linear o MO apresentou o melhor desempenho, mantendo-se com menor erro ao longo de quase todas as combinações. Enquanto que ME aparece em seguida e por fim, SA e WA obtiveram os piores resultados deste experimento.

O fato do ME apresentar melhores resultados em relação ao método SA basicamente comprova o que tem sido ilustrado em diversos trabalhos presentes na literatura de previsão de séries temporais. Uma vez que a média trata-se de um método sensível aos *outliers*. Neste sentido, os modelos individuais que apresentarem previsões distorcidas, com valores longes da realidade, poderão influenciar diretamente nas previsões agregadas do método SA, o tornando menos acurado. Por outro lado, o método ME após ordenar todos os modelos individuais pelos valores de suas respectivas previsões, considera apenas a previsão do modelo que estiver posicionado exatamente no centro, o que minimiza a possibilidade de ser selecionada uma previsão distorcida de algum modelo individual. Assim, o método ME é menos sensível aos *outliers* e consequentemente é capaz de minimizar os erros em relação ao SA quando o conjunto de dados apresenta estas distorções como nas séries SP, ND, DJ, QB, RO, TO, RF e PO, ilustradas nas Figuras 19, 20, 21, 22, 25, 26, 27, 28. Todavia, quando o conjunto de dados não apresenta distorções significativas o método SA pode apresentar superioridade sobre ME como ocorreu para a série RS (veja Figura 23) e PM, porém para esta última, a superioridade mostra-se apenas em termos da métrica $\overline{\text{THEILU}}$ (veja Figura 24(d)). Cabe ressaltar, que ME foi inferior também a WA para estes casos.

O método MO assim como ME também é menos sensível aos *outliers* em comparação ao SA. Este método calcula a moda de Czuber, agrupando os dados em classes e a partir da classe com maior frequência de valores é calculada a previsão agregada. Neste caso, os valores distorcidos teoricamente são incluídos em classes nas extremidades e consequentemente com frequência menor, fazendo com que a moda não seja calculada a partir destas classes. Contudo, um ponto negativo deste método é o processo para calcular o número de classes e, consequentemente a amplitude da classe, tal processo pode resultar em previsões agregadas inferiores aos métodos SA e WA como ocorre para os ensaios com as séries temporais DJ, QB, RS e PM, para esta última especificamente, a inferioridade apenas é apresentada para a métrica $\overline{\text{THEILU}}$ (veja Figura 24(d)). Por outro lado, os demais ensaios mostram que o método MO foi superior, indicando uma possível situação aonde os dados possuem alguns *outliers* ou algum subconjunto de previsões individuais menos acuradas, resultando, consequentemente, em previsões agregadas via MO superiores em relação ao SA e WA. Desta forma, os métodos MO e ME tendem a ser opções mais robustas em termos estatísticos pelo fato de apresentarem maior acurácia para as séries temporais SP, ND, RO, TO, RF e PO como pode ser visto claramente nas Figuras 19, 20, 25, 26, 27 e 28.

Os resultados para as séries DJ, QB, RS e PM mostram que a melhoria dos métodos SA, ME, WA e MO ao longo das combinações é tão pequena que a métrica apresenta um comportamento aproximadamente contínuo. Assim, apesar destes métodos possuírem correlação negativa para a maioria das métricas utilizadas (conforme foi indicado pelo teste de hipótese de correlação de Spearman) estes métodos parecem não ser capazes de combinar diversos modelos individuais e resultar em previsões agregadas mais acuradas.

Os resultados de SA para as séries RO, TO, RF e PO apresentaram um melhor desempenho em relação as outras séries. Contudo, é razoável dizer que o desempenho continua sendo bastante inferior a CB, como pode ser observado na Figura 25 aonde SA possui valores $\overline{\text{MAE}} \cong 100$, enquanto que CB está fixado com $\overline{\text{MAE}} \cong 0$. Essa mesma discrepância de resultados aparece para as demais séries temporais.

De modo geral, os resultados obtidos pelo método SA são inferiores a CB em sua totalidade. A literatura de combinação de modelos individuais de previsão de séries temporais já tem apresentado por diversas vezes que o método SA trata-se de uma metodologia simples de combinação que não leva em consideração a eficiência e acurácia estatística de cada um dos modelos individuais. Neste sentido, todos os preditores são levados em consideração no processo de combinação de maneira igual, ou seja, possuem a mesma importância. Todavia, esse procedimento já tem mostrado deficiências em outros trabalhos.

Neste trabalho especificamente, são combinados mil modelos individuais de previsão de séries temporais diferentes. Como os preditores são gerados aleatoriamente, podem

ser criados modelos individuais com bom desempenho, bem como preditores pobres em acurácia. Na prática, os ensaios foram configurados para existir uma variedade grande de preditores considerados eficientes e ineficientes. Exatamente, com o intuito de observar o desempenho de SA, ME, WA, MO e CB para combinar tal variedade de modelos. Desta forma, conforme mencionado anteriormente, CB apresentou melhor desempenho para todas as séries temporais em relação aos demais métodos de combinação apresentados.

Apesar dos métodos SA, ME, WA e MO se mostrarem inferiores a CB, os resultados obtidos por meio deles para combinar as diversas séries temporais adotadas, mostram que os métodos são capazes para a maioria das séries de melhorar a acurácia das previsões agregadas a medida que mais modelos individuais são incluídos na combinação.

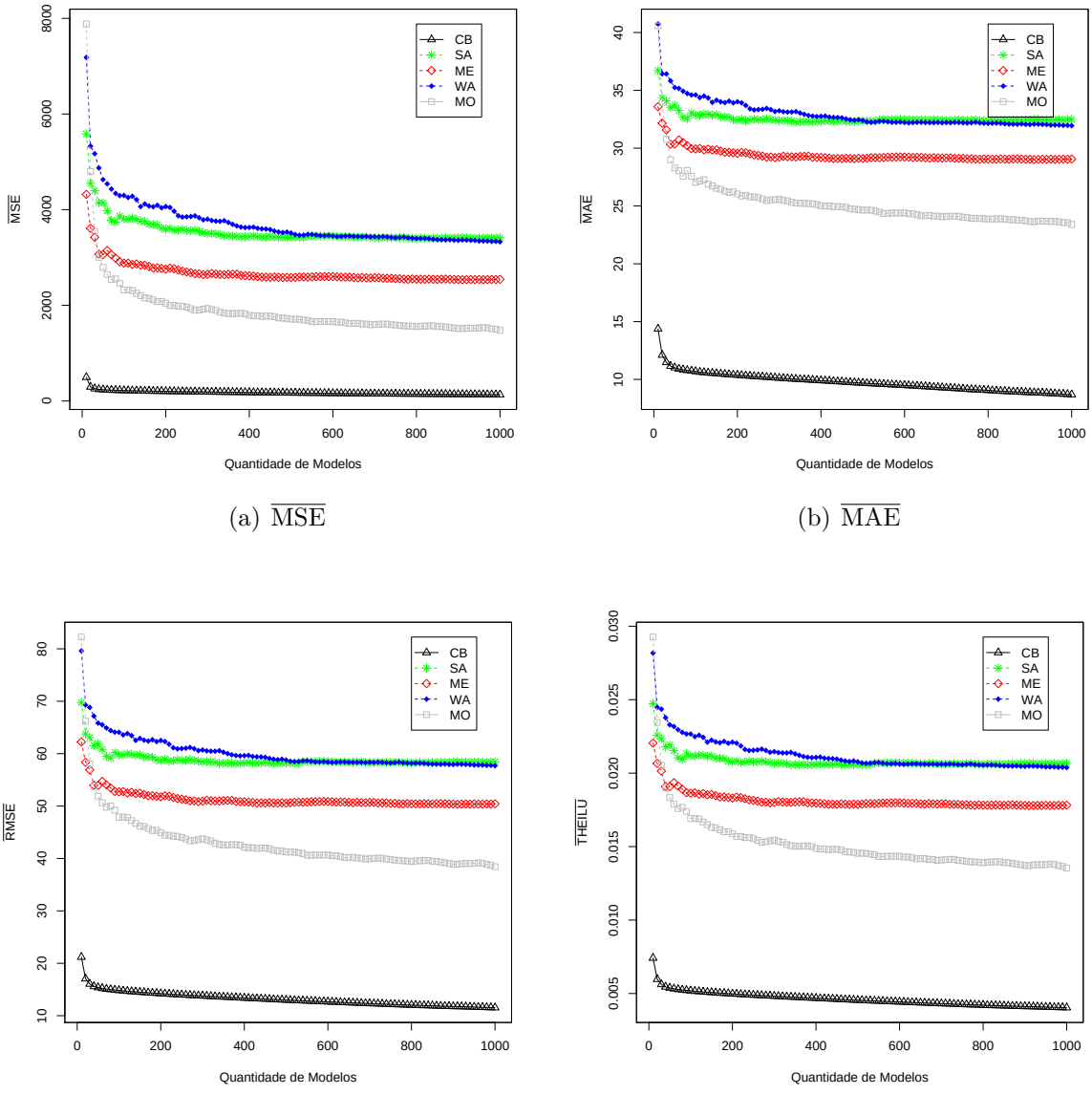
Dado que CB trata-se de uma metodologia de agregação de modelos individuais de previsão de séries temporais, considerando informações como a predição, desvio padrão, média, distribuição de probabilidade marginal e seus respectivos parâmetros, grau de associação entre os preditores. Logo, pode se observar que este processo de agregação é relativamente mais complexo que os adotados nos métodos de combinação linear difundidos na literatura de agregação de previsão de séries temporais. Neste sentido, presumir que tais informações adicionais estão contribuindo para causar tal superioridade do método proposto seja uma afirmação razoável.

5.4 Resumo do Capítulo

Neste capítulo foram discutidos os resultados obtidos pelo método proposto. Primeiramente, um teste de hipótese para o coeficiente de correlação de Spearman e Kendall foi realizado para analisar dois aspectos: (1) verificar se existe algum grau de associação entre o erro (métrica) e a quantidade de modelos individuais (k) e (2) investigar, caso exista algum grau de associação, se esta correlação é negativa, o que indicaria que a medida que a quantidade de modelos cresce menor será o erro. Logo, por meio do teste de hipótese para o coeficiente de correlação de Spearman e Kendall foi constatado a um nível de significância de 5% que o método proposto possui uma associação negativa entre o número de modelos individuais incluídos na combinação e o erro.

Suportado pelo teste de hipótese para o coeficiente de correlação de Spearman e Kendall, uma análise visual dos valores da métrica em relação a quantidade de modelos individuais mostra que o método proposto consegue reduzir o valor da métrica a medida que mais modelos são incluídos na combinação. Assim, as discussões deste capítulo descrevem o comportamento de cada um dos experimentos. De modo geral, os resultados mostram que o método proposto foi capaz de reduzir o valor da métrica para todas as séries temporais apresentadas.

A análise de regressão linear aplicada para estimar o erro do método proposto,

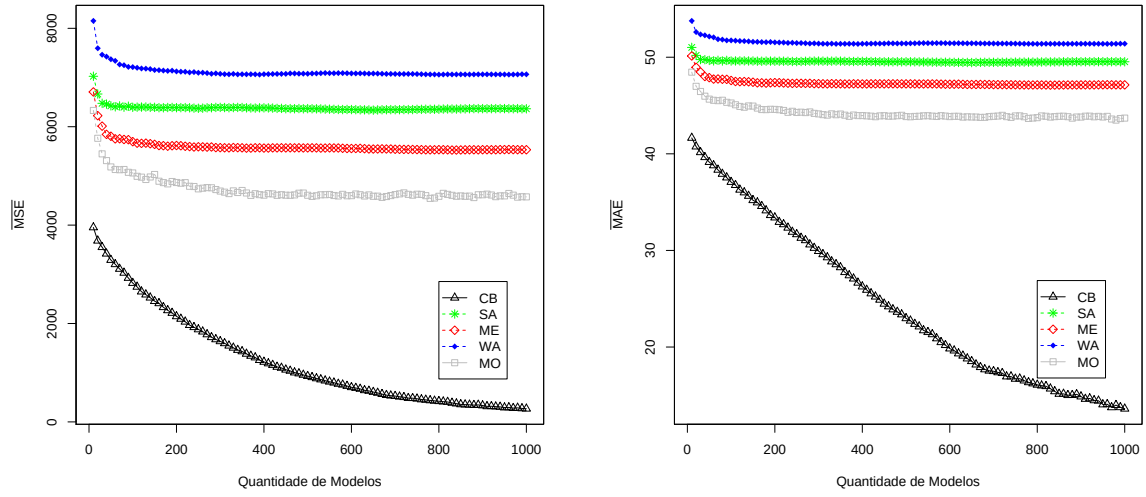


(a) $\overline{\text{MSE}}$ (b) $\overline{\text{MAE}}$

(c) $\overline{\text{RMSE}}$ (d) $\overline{\text{THEILU}}$

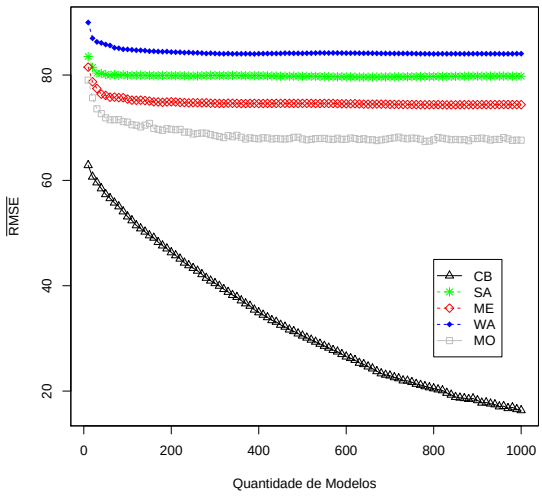
Figura 19 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal SP.

mostraram que os modelos de regressão ajustados as séries, tem a tendência de reduzir ainda mais o erro de previsão para quantidades superiores de preditores. Desta forma, por meio dos ensaios, foi possível discutir alguns pontos importantes neste capítulo, como: (1) se quantidades maiores de modelos individuais é capaz de melhorar as previsões combinadas, (2) é possível definir uma quantidade ótima de modelos individuais e (3) o impacto em termos de desempenho pode ser causado pela quantidade de modelos. Finalmente, são mostradas várias comparações entre SA, ME, WA, MO e CB. Os resultados mostraram a superioridade de CB para todas as séries temporais utilizadas.

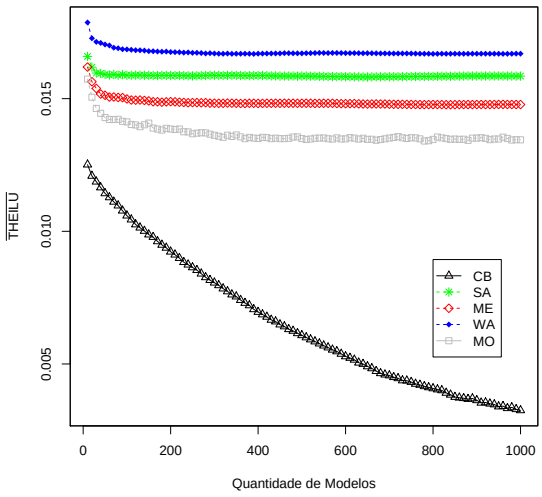


(a) $\overline{\text{MSE}}$

(b) $\overline{\text{MAE}}$

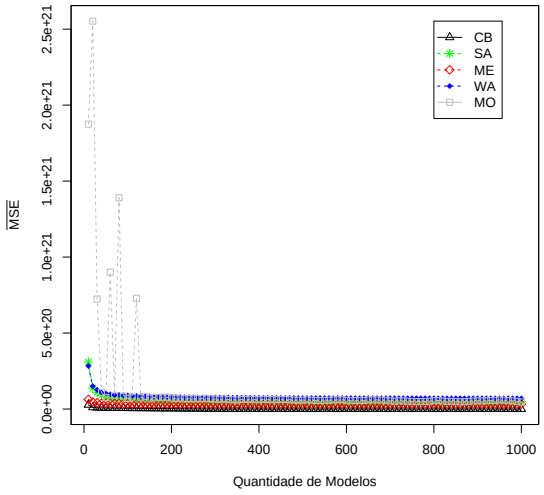


(c) $\overline{\text{RMSE}}$

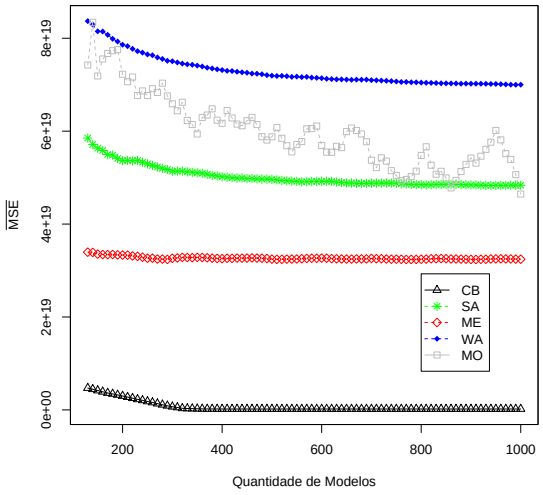


(d) $\overline{\text{THEILU}}$

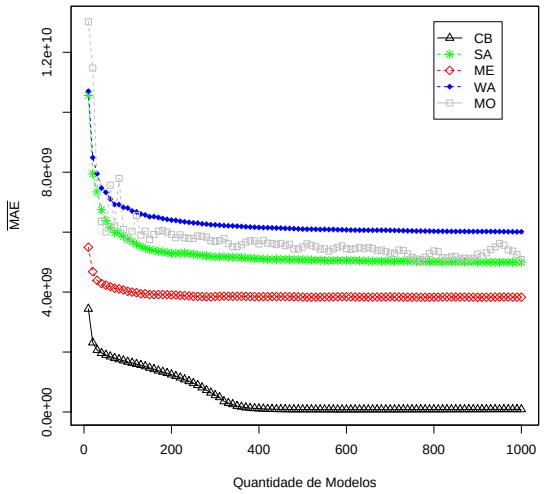
Figura 20 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal ND.



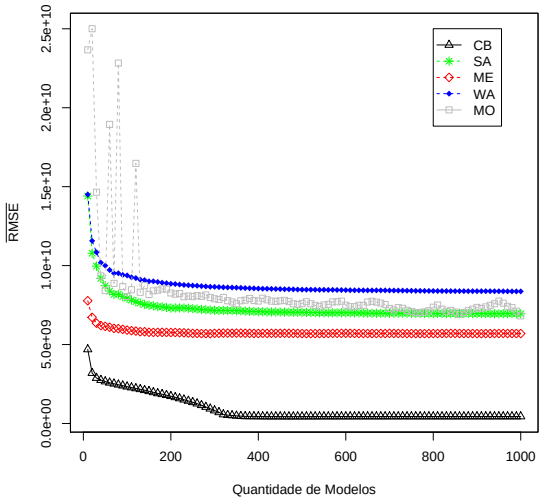
(a) $\overline{\text{MSE}}$



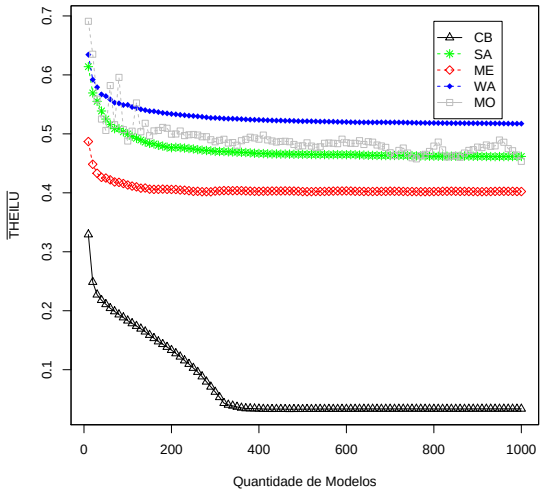
(b) $\overline{\text{MSE}}$ com $k = 130, \cdots, 1000$.



(c) $\overline{\text{MAE}}$

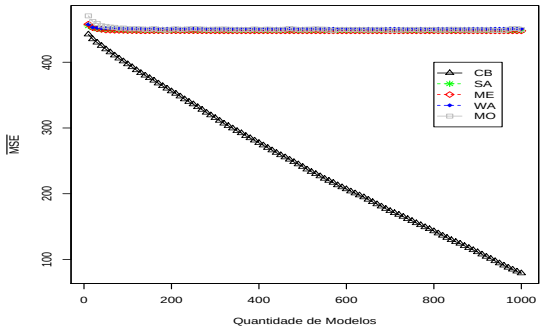


(d) $\overline{\text{RMSE}}$

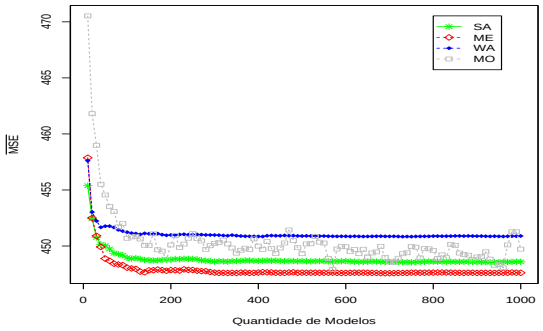


(e) $\overline{\text{THEILU}}$

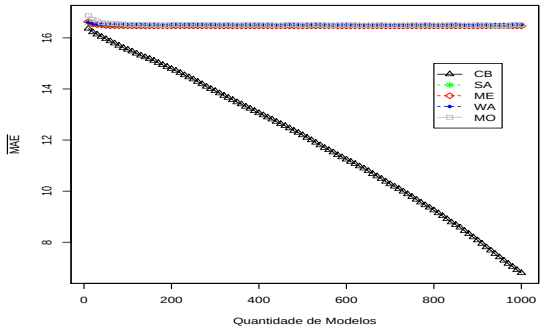
Figura 21 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal DJ. A Figura 21(b) destaca os resultados para $k > 130$.



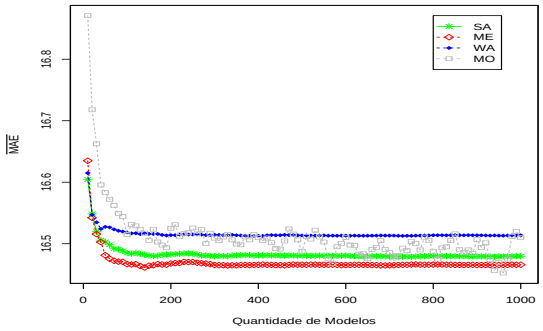
(a) $\overline{\text{MSE}}$



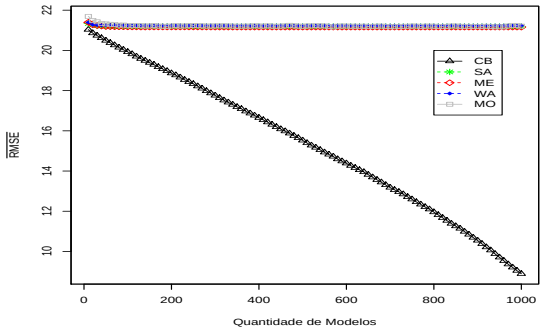
(b) $\overline{\text{MSE}}$



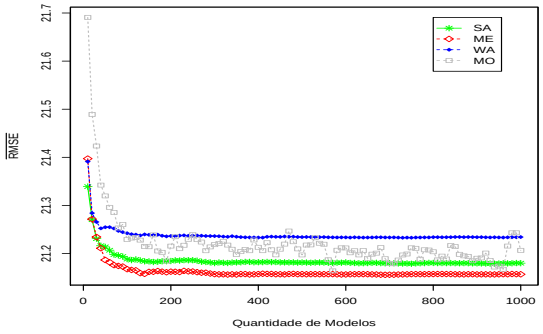
(c) $\overline{\text{MAE}}$



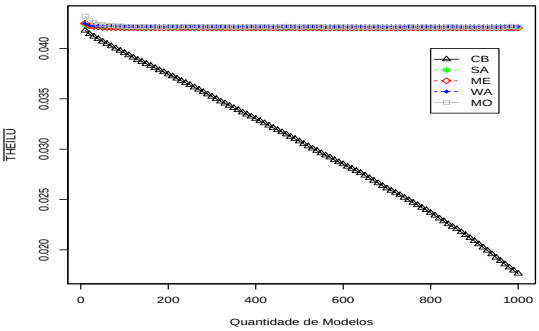
(d) $\overline{\text{MAE}}$



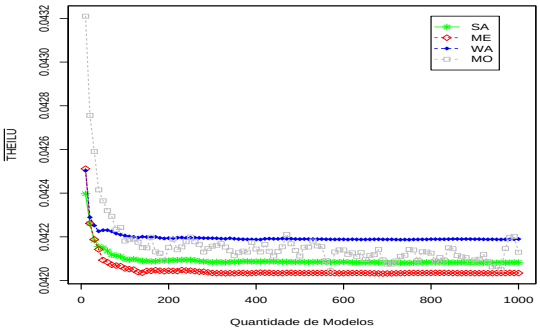
(e) $\overline{\text{RMSE}}$



(f) $\overline{\text{RMSE}}$

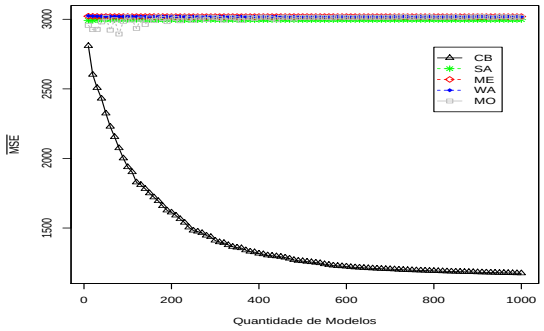


(g) $\overline{\text{THEILU}}$

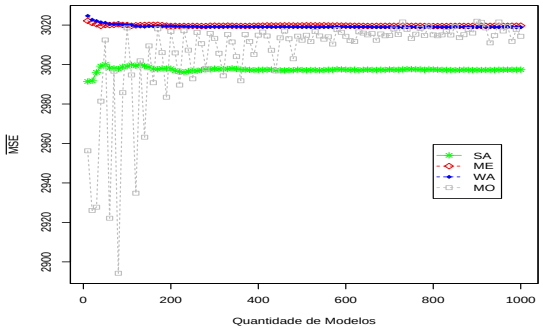


(h) $\overline{\text{THEILU}}$

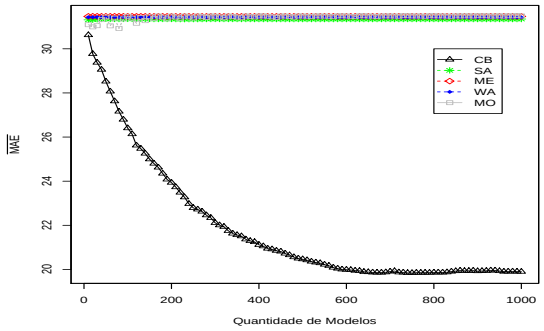
Figura 22 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal QB. As Figuras 22(b), 22(d), 22(f) e 22(h) apresentam os resultados exclusivamente para os métodos SA, ME, WA e MO.



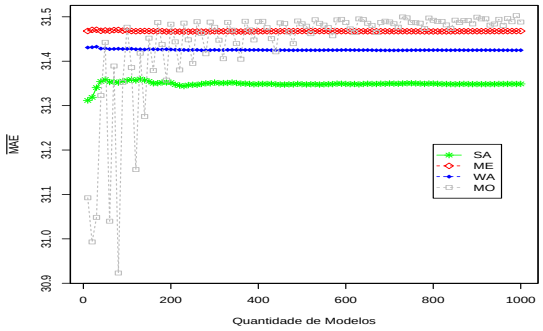
(a) $\overline{\text{MSE}}$



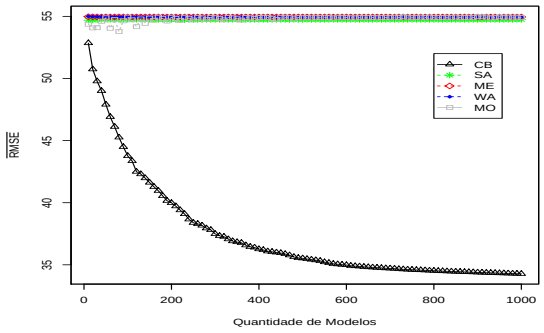
(b) $\overline{\text{MSE}}$



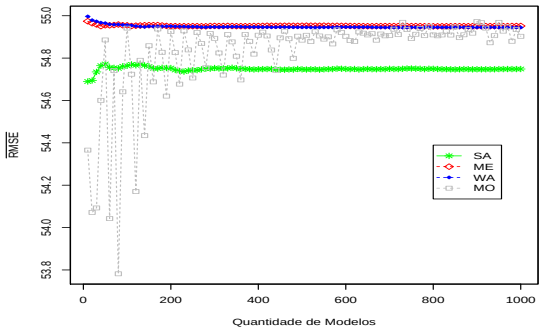
(c) $\overline{\text{MAE}}$



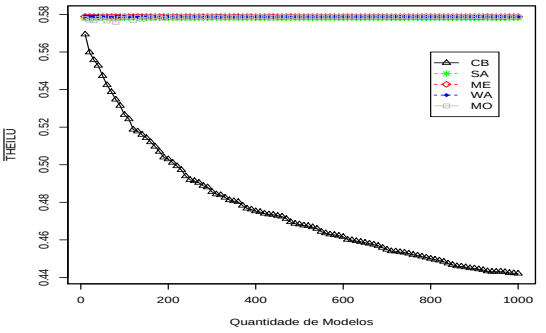
(d) $\overline{\text{MAE}}$



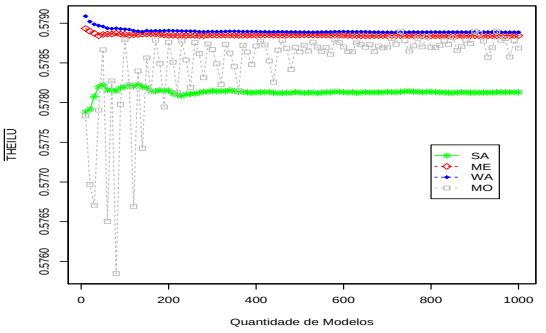
(e) $\overline{\text{RMSE}}$



(f) $\overline{\text{RMSE}}$

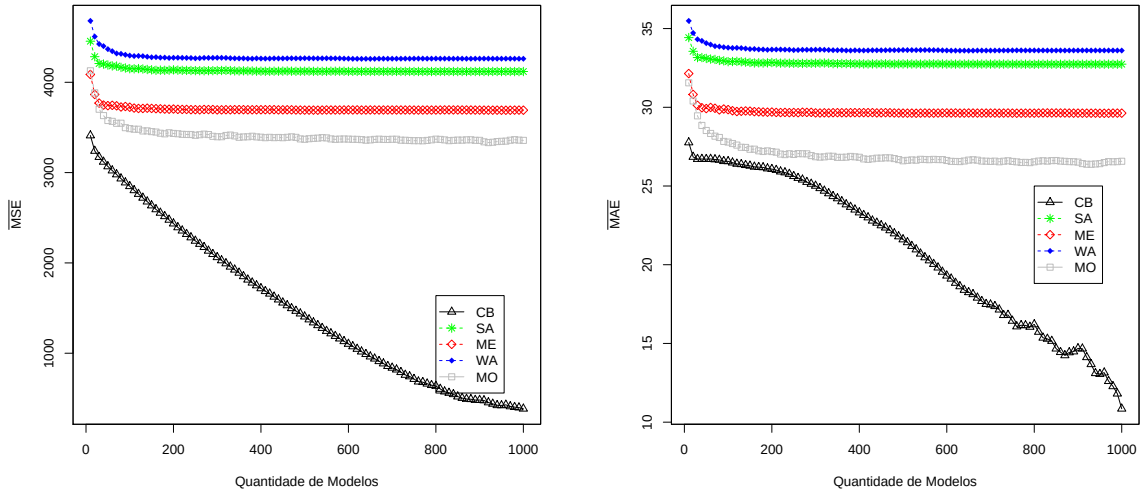


(g) $\overline{\text{THEILU}}$



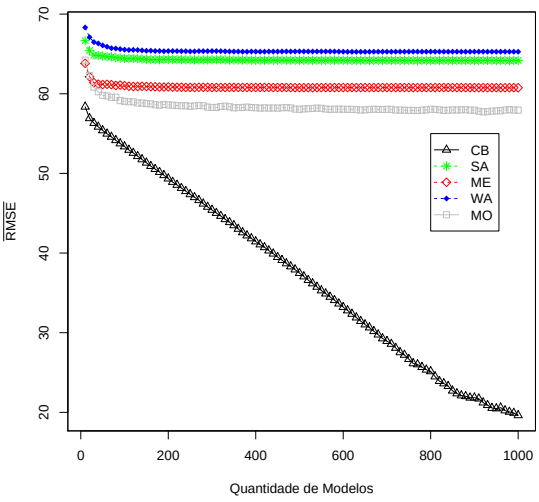
(h) $\overline{\text{THEILU}}$

Figura 23 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal RS.As Figuras 23(b), 23(d), 23(f) e 23(h) apresentam os resultados exclusivamente para os métodos SA, ME, WA e MO.

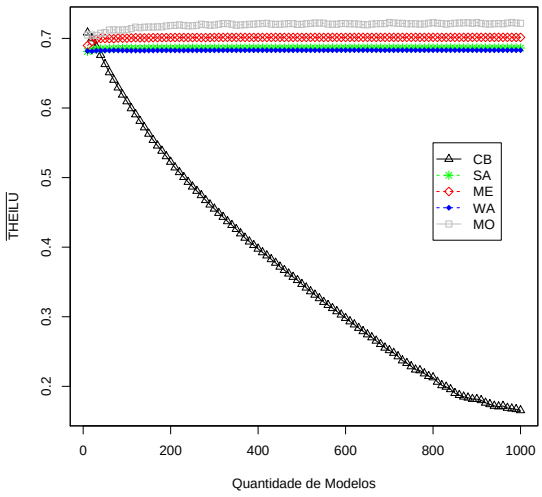


(a) $\overline{\text{MSE}}$

(b) $\overline{\text{MAE}}$

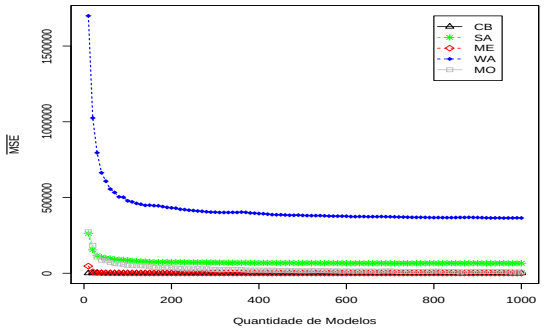


(c) $\overline{\text{RMSE}}$

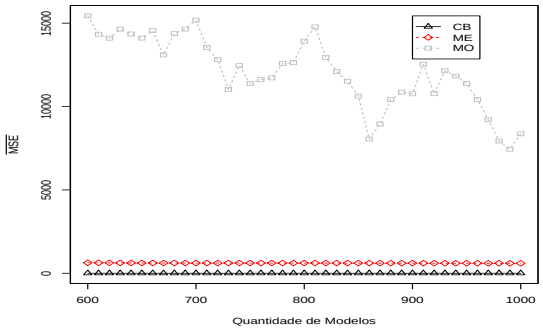


(d) $\overline{\text{THEILU}}$

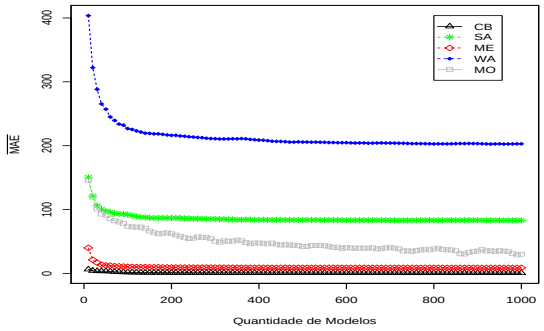
Figura 24 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal PM.



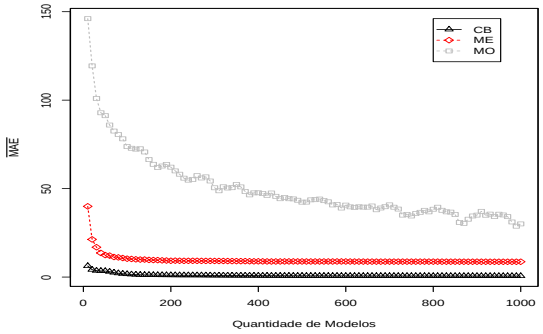
(a) $\overline{\text{MSE}}$



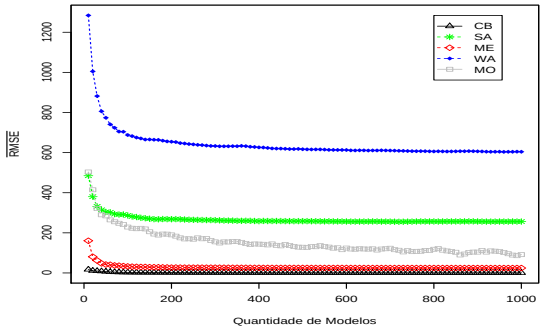
(b) $\overline{\text{MSE}}$ com $k = 600, \cdots, 1000$.



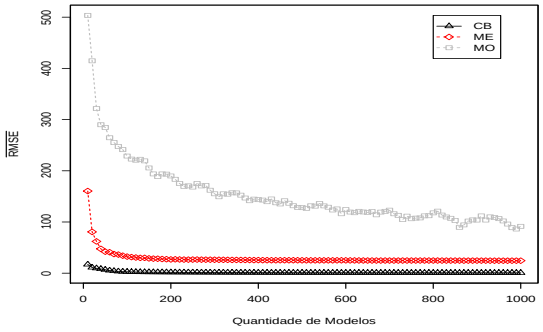
(c) $\overline{\text{MAE}}$



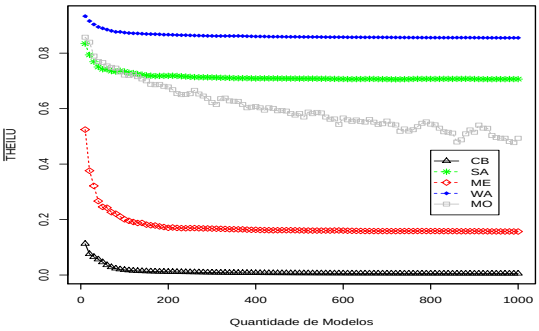
(d) $\overline{\text{MAE}}$



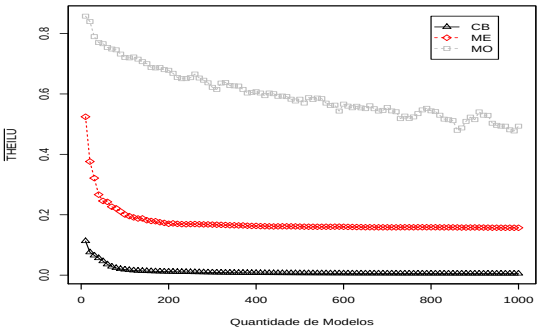
(e) $\overline{\text{RMSE}}$



(f) $\overline{\text{RMSE}}$

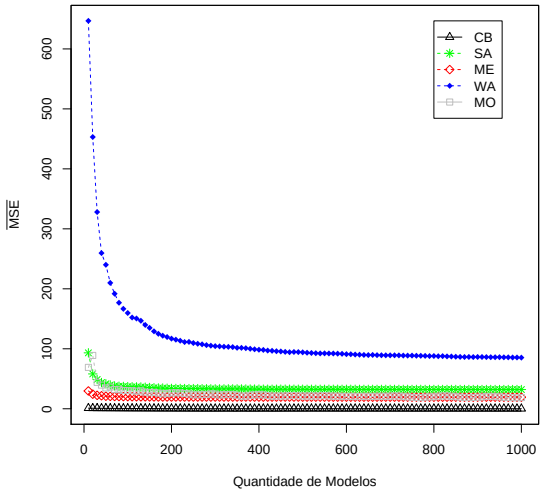


(g) $\overline{\text{THEILU}}$

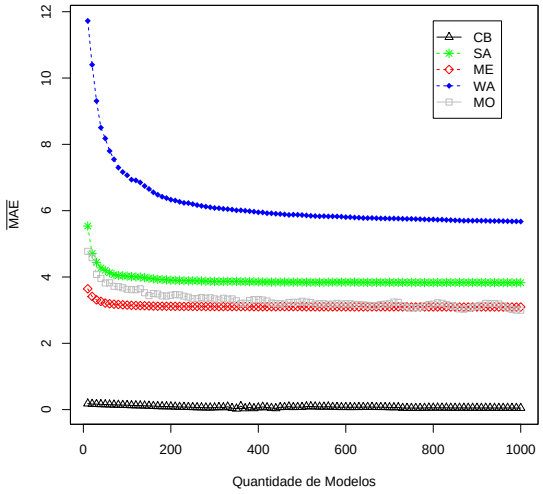


(h) $\overline{\text{THEILU}}$

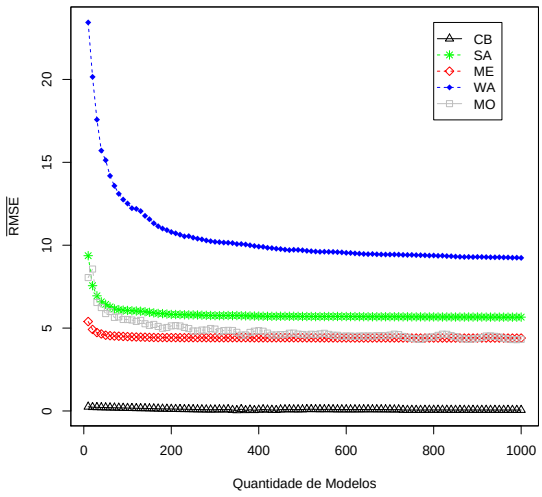
Figura 25 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal RO. As Figuras 25(b), 25(d), 25(f) e 25(h) apresentam os resultados exclusivamente para os métodos CB, ME e MO.



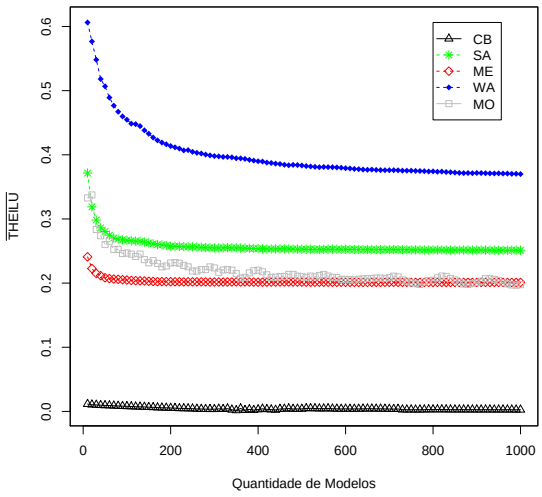
(a) $\overline{\text{MSE}}$



(b) $\overline{\text{MAE}}$

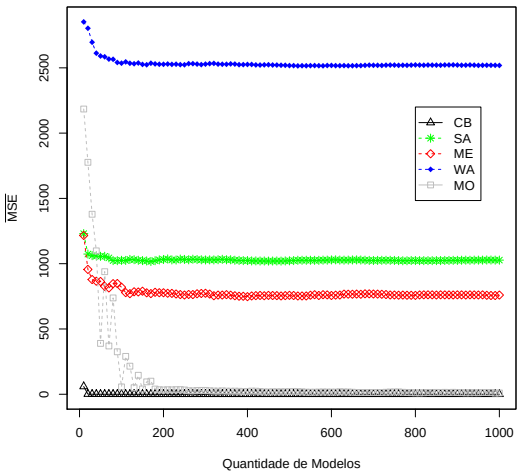


(c) $\overline{\text{RMSE}}$

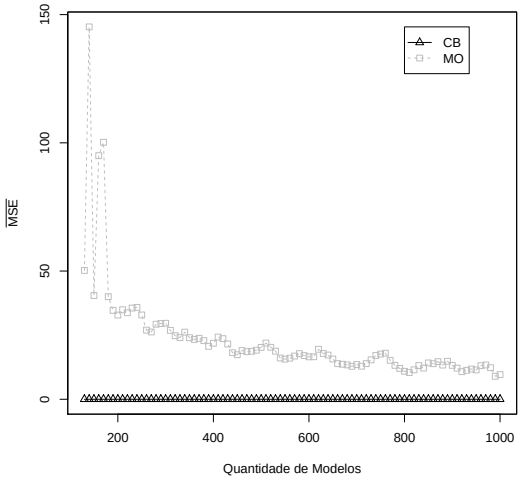


(d) $\overline{\text{THEILU}}$

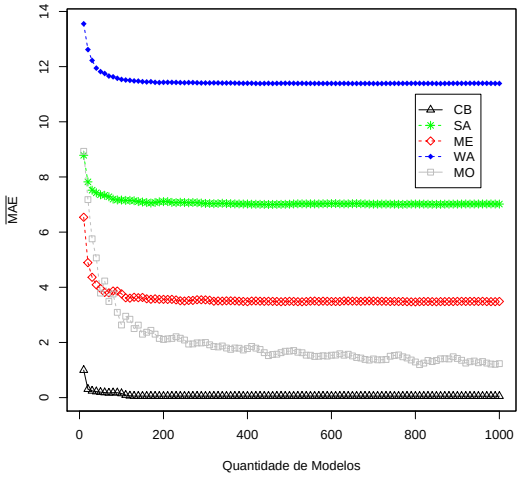
Figura 26 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal TO.



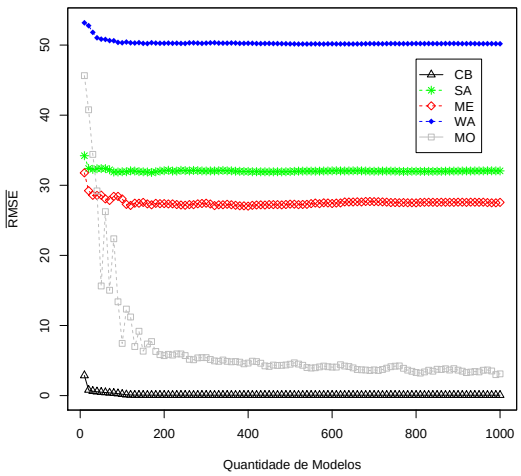
(a) $\overline{\text{MSE}}$



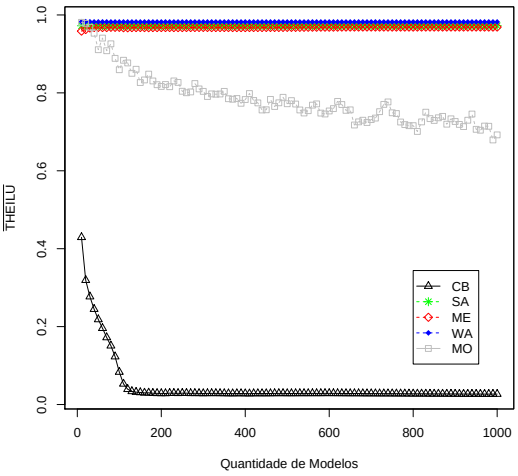
(b) $\overline{\text{MSE}}$ com $k = 130, \cdots, 1000$.



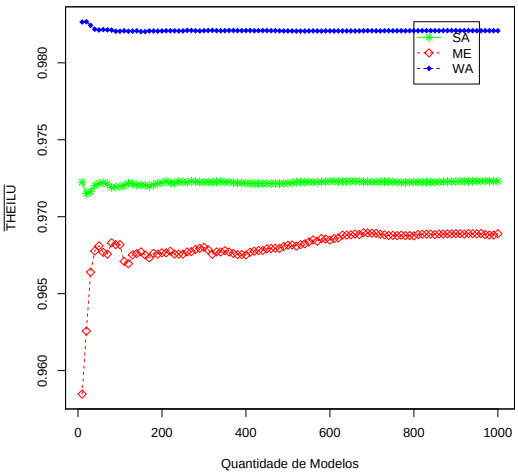
(c) $\overline{\text{MAE}}$



(d) $\overline{\text{RMSE}}$

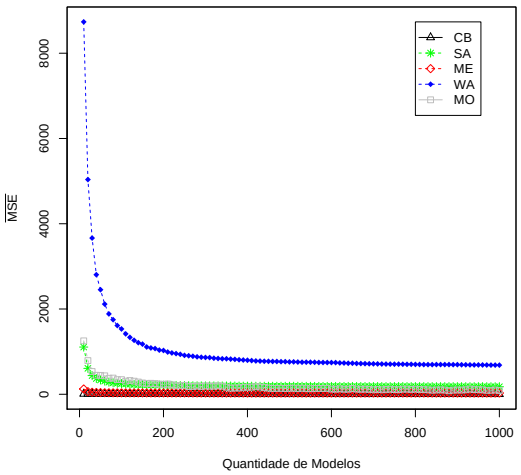


(e) $\overline{\text{THEILU}}$

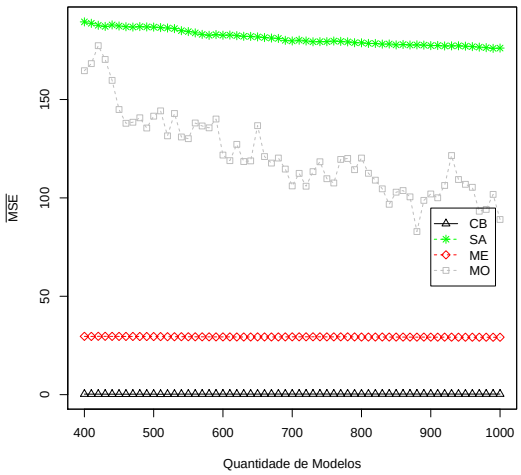


(f) $\overline{\text{THEILU}}$

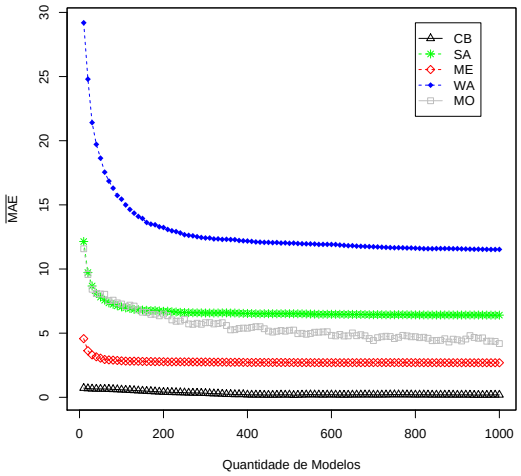
Figura 27 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal RF. A Figura 27(b) apresenta os resultados exclusivamente para os métodos CB e MO. Enquanto que a Figura 27(f) mostra para os métodos SA, ME e WA.



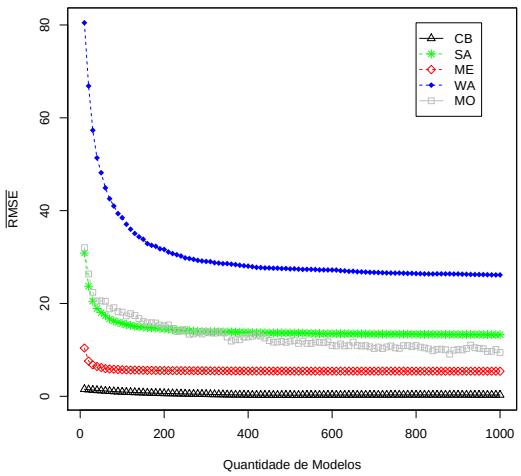
(a) $\overline{\text{MSE}}$



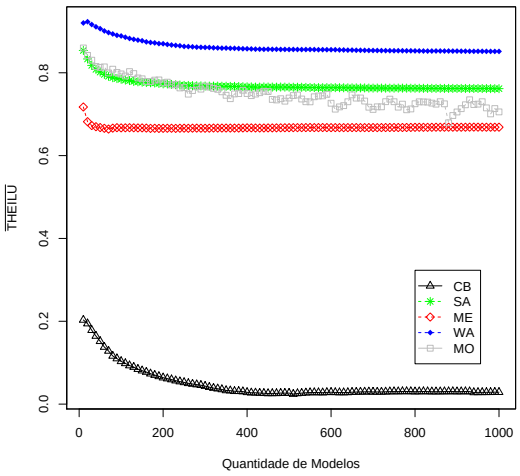
(b) $\overline{\text{MSE}}$ com $k = 400, \cdots, 1000$.



(c) $\overline{\text{MAE}}$



(d) $\overline{\text{RMSE}}$



(e) $\overline{\text{THEILU}}$

Figura 28 – Comparação entre os métodos SA, ME, WA, MO e CB para a série temporal PO. A Figura 28(b) apresenta os resultados exclusivamente para os métodos CB, SA, ME e MO.

6 CONCLUSÕES

Neste último capítulo são apresentadas as conclusões destes estudos e investigações experimentais sobre o impacto do número de modelos individuais sobre a qualidade das previsões combinadas. Desta forma, discussões sobre o método proposto são apresentadas, levando em conta a possibilidade de aplicar na prática o método para prever séries temporais do mundo real. As contribuições e limitações deste trabalho também são discutidas. Por fim, algumas sugestões de trabalhos futuros que podem estender este estudo são propostas.

6.1 Considerações Iniciais

Ensembles aplicados ao problema de previsão de séries temporais, estão entre as estratégias mais adotadas para produzir resultados mais eficientes em termos estatísticos (OLIVEIRA et al., 2017; OLIVEIRA et al., 2016; CLEMEN, 1989). Vários estudos têm mostrado que as previsões combinadas são estatisticamente superiores em relação as previsões obtidas por modelos individuais (AMENDOLA; STORTI, 2008). Diversas modelagens de *ensembles* propostas na literatura se diferenciam na metodologia de combinação, que tem a média simples e ponderada dentre todas as mais difundidas e estudadas (KOURENTZES; BARROW; CRONE, 2014).

Dentre os modelos individuais de previsão de séries temporais, as Redes Neurais Artificiais (RNAs) estão entre as mais encontradas nos estudos de previsão de séries temporais mencionadas na literatura (FERREIRA; VASCONCELOS; ADEODATO, 2008; FERREIRA, 2006). As RNAs normalmente aplicam o algoritmo de aprendizagem baseado no *backpropagation*. Contudo, um algoritmo alternativo foi proposto por Huang, Zhu e Siew (2004) que pode ser milhares de vezes mais rápido que a proposta tradicional, intitulado como máquina de aprendizagem extrema (ELM).

O formalismo matemático de cópulas aplicado ao problema de agregação de modelos individuais de previsão de séries temporais é novo na literatura. Apresentado pela primeira vez por Oliveira et al. (2013) tem se mostrado como ponto de partida para esta nova metodologia que estuda a combinação de diversos preditores.

Partindo da premissa que a quantidade de modelos individuais é capaz de influenciar na qualidade das previsões combinadas. Assumindo que a hipótese de que o *ensemble* torna-se cada vez mais acurado ao agregar mais modelos individuais. Esta hipótese sugere que o *ensemble* seja capaz de melhorar seu desempenho para números maiores de modelos, porque será capaz de obter mais informações sobre o fenômeno, convergindo para erros menores.

Deste modo, fazendo uma análise da literatura, aliado à hipótese levantada, este trabalho propôs um *ensemble* baseado em cópulas (CB) para combinar RNAs treinadas via ELM para prever séries temporais. O *ensemble* mais especificamente é baseado na cópula normal. Inspirado em buscar respostas para a hipótese exposta, um conjunto de séries temporais foi selecionado para análise de desempenho do *ensemble* proposto.

A abordagem proposta é avaliada através de um experimento que envolve um conjunto de séries temporais financeiras, de demografia, meteorologia e hidrologia. De modo a estabelecer uma referência do desempenho CB para o conjunto de séries temporais, foram realizados também experimentos com o método média simples (SA), média ponderada (WA), moda (MO) e mediana (ME).

Vários experimentos foram realizados com o *ensemble* baseado em cópulas. Os testes estatísticos do coeficiente de correlação mostraram para todas as séries temporais, que CB possui correlação negativa entre o erro e a quantidade de modelos individuais. Indicando basicamente um possível fenômeno de causa e efeito, em que quanto mais preditores são inseridos no processo de combinação menores serão os erros cometidos pelo *ensemble*. Supondo que esse fenômeno de causa e efeito seja verdade. Foram realizados experimentos utilizando modelos de regressão linear para estimar o comportamento CB com quantidades superiores de preditores. Os resultados suportam a suposição, tal que o erro cometido por CB foi substancialmente diminuído a medida que mais preditores eram combinados.

Os resultados obtidos pelo teste estatístico do coeficiente de correlação mostraram que os métodos de combinação via SA, WA, MO e ME pode ser acurado para alguns casos. Neste sentido, o erro obtido pelas previsões combinadas por meio destes métodos para as séries temporais são reduzidos ao longo das combinações, com exceção especificamente de apenas algumas métricas, que apresentam em particular resultados ineficientes, isto é, não mostram o fenômeno de causa e efeito em função da quantidade de preditores. Analisando os experimentos, embora que os métodos não tenham sido eficientes em sua totalidade, em alguns casos para métricas específicas o método MO e ME apresentam resultados próximos de CB como ocorre para série RO.

Analisando os experimentos, o método SA, WA, MO, ME e CB, de maneira geral, mostraram que o desempenho de CB é expressivamente superior, apesar dos métodos seguirem a mesma tendência decrescente apresentada por CB. Assim, os resultados dos métodos mostram que não foram capazes de capturar as informações mais valiosas de cada uma das previsões individuais para as séries temporais em sua totalidade. O que reflete em previsões combinadas com maior ruido, por possuir uma metodologia de agregação mais simples.

O método CB é uma metodologia que agrega os modelos de previsão de séries temporais levando em consideração informações como valor da previsão, média, desvio

padrão, distribuição de probabilidade marginal e seus respectivos parâmetros, grau de correlação entre os modelos individuais. Fazendo com que seu processo de agregação seja relativamente mais complexo em relação aos métodos de combinação linear previamente conhecidos na literatura. Assim, é razoável presumir que essas informações adicionais possam ser a razão que explique os resultados superiores do método proposto.

6.2 Contribuições da Tese

As contribuições alcançadas nesta Tese possibilitam responder às questões introduzidas no Capítulo de Introdução na Seção 1.2:

i **À medida que mais modelos individuais são incorporados à combinação, é possível produzir previsões mais acuradas?**

Após as análises, pode ser comprovado que dependendo do método utilizado, a quantidade de modelos individuais envolvidos na combinação é capaz de influenciar na qualidade das previsões agregadas para as séries temporais adotadas neste trabalho. A partir dos experimentos, o método baseado em cópulas mostrou-se acurado e eficiente para os diversos ensaios. Enfatizando, que para os ensaios com os métodos clássicos de combinação linear os resultados nem sempre mostraram evidências que existe influência na qualidade das previsões a partir da quantidade de modelos individuais agregados.

As obras apresentadas na literatura por Oliveira et al. (2018), Oliveira et al. (2017), Oliveira et al. (2016), Oliveira et al. (2013) mostram o formalismo de cópulas sendo aplicada, de modo geral, apenas para combinar dois modelos. Neste trabalho, CB apresentou a capacidade de agregar vários modelos individuais, e principalmente, a capacidade de extrair as vantagens oferecidas pelos preditores para as séries temporais adotadas. Desta forma, através dos experimentos realizados, o método proposto, mesmo sem utilizar nenhum processo de seleção de modelos, foi capaz de aperfeiçoar os resultados.

ii **Existe uma quantidade específica de modelos individuais que possa ser incorporada na combinação, de modo que, garanta previsões estatisticamente eficientes e acuradas?**

Por meio das análises estatísticas realizadas foi possível constatar que os métodos de regressão linear utilizados são uma alternativa para estimar a quantidade de modelos individuais que devem ser usados nos experimentos para alcançar a acurácia desejada. Assim, os métodos de regressão apresentados são capazes de estimar k a partir do erro que almeja-se obter. Além disso, o processo inverso também foi exposto, isto é, estimar o erro através da quantidade de modelos individuais utilizados na combinação.

Desta maneira, este trabalho traz uma alternativa para estimar a quantidade de modelos individuais necessária para garantir previsões estatisticamente acuradas.

iii **Combinar diversos modelos individuais de previsão de séries temporais, através de cópulas, conduz a melhores resultados comparados com outras abordagens presentes na literatura?**

As metodologias estudadas em questão, passaram pelos experimentos que foram conduzidos com previsões de séries temporais diversificadas envolvendo fenômenos da natureza, financeiros e demográfico. E uma das principais contribuições deste trabalho, trata-se da comparação apresentada com os métodos de combinação linear estabelecidos na literatura. Os resultados comprovaram a superioridade de CB sobre outros métodos de combinação linear. Os experimentos demonstraram para todas as métricas que CB é capaz de superar os métodos comparados. Estes resultados, sugerem que a combinação através dos métodos que não incorporam dependência entre os modelos não são capazes, em sua totalidade, de captar as principais vantagens de cada modelo individual, diferente do que ocorre com o *ensemble* baseado em cópula.

6.3 Limitações da Tese

A abordagem proposta apresenta uma limitação relacionada ao tamanho da série temporal. Durante a realização dos experimentos, foi observado que para séries pequenas ocorre um super ajustamento das RNAs, que ocasiona a geração de modelos extremamente especializados, ou seja, as redes neurais geradas não apresentavam viés estatístico. Assim, os modelos individuais eram capazes de prever todas as observações do conjunto de treinamento da série sem cometer nenhum erro. Desta forma, a média, bem como a variância dos erros das previsões foi aproximadamente zero ou propriamente zero.

A natureza do método CB requer que os modelos individuais sejam ruidosos. Neste sentido, estatisticamente é inviável combinar RNAs que não apresentem viés por meio de cópulas, dado que matematicamente o pré-requisito é que haja variabilidade entre os erros dos modelos individuais. Essa limitação faz com que o método CB seja evitado para aplicações em problemas com séries temporais pequenas.

Outra limitação observada é que utilizando janelas pequenas (por exemplo, $lag = 4$) no processo de treinamento das RNAs, o método CB não é capaz de construir preditores diversificados. Durante os experimentos realizados, foi possível verificar que na maior parte das vezes, quando os modelos individuais estavam sendo combinados, o desempenho de CB piorava ao invés de melhorar. Assim, as RNAs são aproximadamente iguais por causa da janela pequena, os erros cometidos pelos modelos individuais são aproximadamente os mesmos. Com baixa variação dos erros entre os modelos individuais, o procedimento

via cópula mostra-se incapaz de melhorá a cada nova RNAs incluída na combinação. O comportamento que ocorre para cada nova combinação, faz o erro de agregação aumentar e consequentemente o desempenho de CB piora. Logo, janelas pequenas foram evitadas.

6.4 Trabalhos Futuros

Embora o modelo CB tenha obtido resultados expressivos para obter previsões agregadas, inclusive nas comparações apresentadas com outras técnicas, ainda existem alguns pontos que podem ser estudados com maior cuidado.

Uma comparação entre CB e outras alternativas de combinação mais sofisticadas que os métodos clássicos se mostra importante para avaliar o desempenho do método proposto em relação as outras técnicas mais atuais da literatura de previsão de séries temporais.

Uma extensão deste trabalho, pode ser apontada pela necessidade de explorar o potencial de outros tipos de cópulas que não foram mencionadas neste trabalho. A cópula de Cacoullos é uma ótima alternativa por ser não paramétrica, além disso, esta mostrou-se superior a cópula normal em experimento realizado com um pequeno conjunto de dados (veja Oliveira et al. (2018)).

Vislumbra-se ainda a possibilidade de expandir estes estudos para o problema de classificação de padrões. A ideia seria combinar classificadores individuais, que também poderiam ser obtidos por meio de RNAs e treinadas com o auxílio do algoritmos de máquina de aprendizagem extrema.

Outro ponto importante, trata-se dos modelos individuais de previsão de séries temporais, que poderiam ser construídos levando em consideração diferentes técnicas. Assim, ao invés das RNAs comporem todo o conjunto de modelos individuais, este conjunto poderia ser preenchido por outros métodos como os modelos Box & Jenkins, metodologias híbridas, estatísticas, diferentes tipos de redes neurais, entre outros.

Finalmente, uma vez que é verdadeira a hipótese de que o *ensemble* torna-se cada vez mais acurado ao agregar mais modelos individuais e proporcionalmente inferior ao combinar menos preditores. Então um ponto interessante como trabalho futuro, seria investigar de forma comparativa se técnicas de seleção de modelos individuais poderiam produzir resultados mais eficientes. A ideia é analisar comparativamente o desempenho entre CB utilizando a técnica de seleção de preditores e CB combinando todos os modelos e avaliar o melhor método.

6.5 Produções Científicas

Ao longo desta pesquisa de doutorado foram produzidos os seguintes trabalhos:

1. *Aggregation of Time Series Forecasts via Cacoullos Copula*. Publicado no IEEE *International Joint Conference on Neural Networks* (IJCNN), 2018.

2. *Copulas-based time series combined forecasters*. Publicado na *Information Sciences*, 2017.

3. *Copulas-Based Ensemble of Artificial Neural Networks for Forecasting Real World Time Series*. Publicado no IEEE *International Joint Conference on Neural Networks* (IJCNN), 2016.

REFERÊNCIAS

AAS, K. *Modelling the dependence structure of financial assets: A survey of four copulas*. [S.l.], 2004. Citado na página 45.

ALBERT, A. *Regression and the Moore-Penrose pseudoinverse*. [S.l.]: Elsevier Science, 1972. v. 94. Citado 2 vezes nas páginas 22 e 37.

AMENDOLA, A.; STORTI, G. A gmm procedure for combining volatility forecasts. *Computational Statistics & Data Analysis*, v. 52, p. 3047 – 3060, 2008. Citado 3 vezes nas páginas 19, 23 e 102.

ARARIPE, A. A. d. *Prevendo inflação usando séries temporais e combinações de previsões*. Dissertação (Mestrado) — Escola de Pós-graduaão em Economia, 2008. Citado 2 vezes nas páginas 19 e 34.

ASSIS, T. F. O. de. *Cópula para combinação de modelos de séries temporais*. Tese (Doutorado) — Universidade Federal Rural de Pernambuco, 2016. Citado 4 vezes nas páginas 20, 21, 44 e 45.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística: para cursos de engenharia e informática*. [S.l.]: Atlas, 2010. Citado 5 vezes nas páginas 29, 31, 32, 33 e 42.

BARROW, D. K.; CRONE, S. F.; KOURENTZES, N. An evaluation of neural network ensembles and model selection fortime series prediction. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2010. Citado na página 21.

BARROW, D. K.; KOURENTZES, N. Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, v. 177, p. 24–33, 2016. Citado na página 23.

BEN-ISRAEL, A. The moore of the moore-penrose inverse. In: *International Linear Algebra Conference*. Haifa: [s.n.], 2001. Citado 2 vezes nas páginas 22 e 37.

BERGMEIR, C.; BENÍTEZ, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, v. 191, p. 192–213, 2012. Citado na página 51.

BLIEMEL, F. Theil's forecast accuracy coefficient: A clarification. *Journal of Marketing Research*, X, p. 444–446, 1973. Citado na página 28.

BONE, M. A. R.; CARDOT, H. A new boosting algorithm for improved time-series forecasting with recurrent neural networks. *Information Fusion*, v. 9, p. 41–55, 2008. Citado 2 vezes nas páginas 19 e 21.

BOX, G. E. P.; JENKINS, G. M. *Time series analysis: forecasting and control*. [S.l.]: Holden-Day, 1976. Citado 3 vezes nas páginas 8, 26 e 27.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time series analysis: forecasting and control*. 3rd. ed. [S.l.]: Prentice Hall, 1994. (Forecasting and Control Series). ISBN 9780130607744. Citado 3 vezes nas páginas 26, 49 e 55.

BRAGA, A. d. P.; CARVALHO, A.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Livros Técnicos e Científicos Editora S.A., 2000. Citado na página 36.

CABRERA, J. B. On the impact of fusion strategies on classification errors for largeensembles of classifiers. *Pattern Recognition*, v. 39, p. 1963–1978, 2006. Citado na página 23.

CACOULLOS, T. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, v. 18, p. 179–189, 1964. Citado na página 45.

CALLEGARI-JACQUES, S. *Bioestatística: Princípios e aplicações*. [S.l.]: Artmed Editora, 2009. Citado 3 vezes nas páginas 29, 30 e 31.

CHAPRA, S. C.; CANALE, R. P. *Métodos Numéricos para Engenharia*. [S.l.]: 7ªEdição, 2016. Citado na página 20.

CHEN, C.; DANTCHEVA, A.; ROSS, A. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, v. 32, Part B, p. 80 – 92, 2016. ISSN 1566-2535. Citado na página 19.

CHEN, T.; REN, J. Bagging for gaussian process regression. *Neurocomputing*, v. 72, p. 1605–1610, 2009. Citado na página 34.

CHWIF, L.; MEDINA, A. *Modelagem e simulação de eventos discretos: Teoria e aplicações*. [S.l.]: Elsevier Brasil, 2014. Citado na página 54.

CLEMEN, R. T. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, v. 5, p. 559–583, 1989. Citado 3 vezes nas páginas 19, 102 e 115.

DELL'AQUILA, R.; RONCHETTI, E. Stock and bond return predictability: the discrimination power of model selection criteria. *Computational Statistics & Data Analysis*, v. 50, p. 1478 – 1495, 2006. Citado na página 19.

ELIDAN, G. Copulas in machine learning. In: *CRM-CANSSI Workshop on New Horizons in Copula Modeling*. [S.l.: s.n.], 2014. Citado na página 20.

FÁVERO, L.; BELFIORE, P. *Manual de análise de dados: Estatística e modelagem multivariada com Excel, SPSS e Stata*. [S.l.]: Elsevier Editora Ltda, 2017. Citado 2 vezes nas páginas 35 e 36.

FERREIRA, T. A. E. *Uma Nova Metodologia Híbrida Inteligente para a Previsão de Séries Temporais*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2006. Citado na página 102.

FERREIRA, T. A. E.; VASCONCELOS, G. C.; ADEODATO, P. J. L. A new intelligent system methodology for time series forecasting with artificial neural networks. *Neural Processing Letters*, v. 28, p. 113–129, 2008. Citado 3 vezes nas páginas 23, 51 e 102.

FIELD, A. *Descobrindo a estatistica usando o SPSS*. [S.l.]: Bookman Editora, 2009. Citado na página 30.

FILHO, R. B. F. *Integração de modelos de previsão de demanda qualitativos e quantitativos e comparação com seus desempenhos individuais*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul Escola de Engenharia Programa de Pós-graduação em Engenharia de Produção, 2015. Citado na página 34.

FILHO, V. N.; PESSOA, S. R. *Previsão de séries temporais utilizando pools de preditores criados a partir do particionamento da série e da divisão da tarefa de previsão*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2015. Citado na página 27.

FIRMINO, P. R. A.; NETO, P. S. M.; FERREIRA, T. A. E. Correcting and combining time series forecasters. *Neural Networks*, v. 50, p. 1–11, 2014. Citado 5 vezes nas páginas 19, 21, 33, 51 e 60.

GOSSO, A. *Implementation of ELM (Extreme Learning Machine) algorithm for SLFN (Single Hidden Layer Feedforward Neural Networks)*. [S.l.], 2013. Citado na página 60.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132733501. Citado na página 36.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall, 1999. Citado na página 36.

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman, 2001. Citado na página 36.

HEESWIJK, M. et al. Adaptive ensemble models of extreme learning machines for time series prediction. In: *International Conference on Artificial Neural Networks - ICANN 2009*. [S.l.: s.n.], 2009. Citado na página 23.

HILLEBRAND, E.; MEDEIROS, M. C. The benefits of bagging for forecast models of realized volatility. *Econometric Reviews*, v. 29, p. 571–593, 2010. Citado na página 34.

HIPEL, K.; MCLEOD, A. *Time series modelling of water resources and environmental systems*. [S.l.]: Elsevier, 1994. Citado 3 vezes nas páginas 8, 27 e 61.

HOUGAARD, P. A class of multivanate failure time distributions. *Biometrika*, v. 73, p. 671 – 678, 1986. Citado na página 21.

HUANG, G.-B. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEE Transactions on Neural Networks*, v. 14, p. 274–281, 2003. Citado na página 36.

HUANG, G.-B.; BABRI, H. A. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEE Transactions on Neural Networks*, v. 9, p. 224–229, 1998. Citado na página 36.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. *Real-time learning capability of neural networks*. [S.l.], 2003. Citado na página 36.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. [S.l.], 2004. v. 2, p. 985–990. Citado 2 vezes nas páginas 37 e 102.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing*, v. 70, p. 489–501, 2006. Citado 5 vezes nas páginas 24, 36, 37, 49 e 52.

INOUE, A.; KILIAN, L. On the selection of forecasting models. *Journal of Econometrics*, v. 130, p. 273–306, 2006. Citado na página 23.

JEONG, D. I.; KIM, Y.-O. Combining single-value streamflow forecasts - a review and guidelines for selecting techniques. *Journal of Hydrology*, v. 377, p. 284–299, 2009. Citado na página 19.

JOE, H.; XU, J. J. *The estimation method of inference functions for margins for multivariate models*. [S.l.], 1996. Disponível em: <http://www.stat.ubc.ca/~harry/ifm.pdf>. Citado na página 47.

KIM, D.; KIM, C. Forecasting time series with genetic fuzzy predictor ensemble. *Fuzzy Systems, IEEE Transactions*, v. 5, p. 523–535, 1997. Citado na página 23.

KOURENTZES, N.; BARROW, D.; PETROPOULOS, F. Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, v. 209, p. 226–235, 2019. Citado 2 vezes nas páginas 19 e 23.

KOURENTZES, N.; BARROW, D. K.; CRONE, S. F. Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, v. 41, p. 4235–4244, 2014. Citado 6 vezes nas páginas 19, 20, 21, 23, 59 e 102.

KOURENTZES, N.; PETROPOULOS, F.; TRAPERO, J. R. Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, v. 30, p. 291–302, 2014. Citado na página 34.

KRISHNAMURTI, T. N. et al. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, v. 13, p. 4196–4216, 2000. Citado 2 vezes nas páginas 19 e 34.

KUNCHEVA, L. *Combining pattern classifiers: Methods and algorithms*. [S.l.]: Wiley, 2014. Citado 3 vezes nas páginas 23, 34 e 59.

KUTNER, M. H. et al. *Applied linear statistical models*. 5. ed. [S.l.]: McGraw-Hill Irwin Boston, 2005. v. 103. Citado na página 30.

LEAL, D. M. B. *Aplicação de cópulas ao ramo vida: Risco de resgate e risco de taxa de juro*. Dissertação (Mestrado) — Instituto Superior de Economia e Gestão, 2010. Disponível em: <http://pascal.iseg.utl.pt/~alfredo/ftp/papers/leal_msc.pdf>. Citado 2 vezes nas páginas 44 e 47.

LUX, T.; MORALES-ARIAS, L. Forecasting volatility under fractality, regime-switching, long memory and student-t innovations. *Computational Statistics and Data Analysis*, v. 54, p. 2676 – 2692, 2010. Citado 2 vezes nas páginas 19 e 23.

MENEZES, L. M. de; BUNN, D. W.; TAYLOR, J. W. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, v. 120, p. 190–204, 2000. Citado 2 vezes nas páginas 19 e 33.

NELSEN, R. B. *An introduction to copulas*. Second. Portland, USA: Springer, 2006. 272 p. (Springer Series in Statistics). Citado 4 vezes nas páginas 20, 21, 41 e 42.

OLIVEIRA, R. T. A. et al. Copulas-based ensemble of artificial neural networks for forecasting real world time series. In: *IEEE/INNS International Joint Conference on Neural Networks*. [S.l.: s.n.], 2016. Citado 6 vezes nas páginas 19, 20, 56, 60, 102 e 104.

OLIVEIRA, R. T. A. de. *Modelagem e simulação computacional da combinação de preditores de séries temporais por meio de cópulas*. Dissertação (Mestrado) — Universidade Federal Rural de Pernambuco, 2014. Citado 8 vezes nas páginas 19, 20, 34, 44, 45, 47, 57 e 60.

OLIVEIRA, R. T. A. de et al. Aggregation of time series forecasts via cacoullos copula. In: *International Joint Conference on Neural Networks*. [S.l.: s.n.], 2018. Citado 8 vezes nas páginas 19, 21, 23, 45, 46, 51, 104 e 106.

OLIVEIRA, R. T. A. de et al. Combining time series forecasting models via gumbel-hougaard copulas. In: *1st BRICS Countries Conference on Computational Intelligence*. [S.l.: s.n.], 2013. p. 1–6. Citado 3 vezes nas páginas 20, 102 e 104.

OLIVEIRA, R. T. de et al. Copulas-based time series combined forecasters. *Information Sciences*, v. 376, p. 110–124, 2017. Citado 9 vezes nas páginas 19, 20, 21, 51, 56, 58, 60, 102 e 104.

OLIVEIRA, T. F. et al. Combination of biased artificial neural network forecasters. In: *1st BRICS Countries Conference on Computational Intelligence*. [S.l.: s.n.], 2013. p. 1–6. Citado na página 21.

OMARI, A.; FIGUEIRAS-VIDAL, A. R. Post-aggregation of classifier ensembles. *Information Fusion*, v. 26, p. 96 – 102, 2015. ISSN 1566-2535. Citado na página 19.

PARZEN, E. On estimation of a probability density function and mode. *Annal of Mathematical Statistics*, v. 33, p. 1065–1076, 1962. Citado na página 45.

PENROSE, R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 51, p. 406–4013, 1955. Citado 3 vezes nas páginas 22, 37 e 58.

RENARD, B.; LANG, M. Use of a gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources*, v. 30, p. 897–912, 2007. Citado na página 45.

SALINAS-GUTIÉRREZ, R. et al. Using gaussian copulas in supervised probabilistic classification. *Soft Computing for Intelligent Control and Mobile Robotics*, v. 318, p. 355–372, 2010. Citado na página 20.

SAMMUT, C.; WEBB, G. *Encyclopedia of machine learning*. [S.l.]: Springer US, 2011. (Encyclopedia of Machine Learning). ISBN 9780387307688. Citado na página 19.

SESMERO, M. P. et al. An ensemble approach of dual base learners for multi-class classification problems. *Information Fusion*, v. 24, p. 122 – 136, 2015. ISSN 1566-2535. Citado na página 19.

SINGH, V. P.; ZHANG, L. Idf curves using the frank archimedean copula. *Journal of Hydrologic Engineering*, v. 12, p. 651–662, 2007. Citado na página 21.

SKLAR, A. Fonctions de répartition à n dimensions et leurs marges. *l'Institut de Statistique de L'Université de Paris*, v. 8, p. 229–231, 1959. Citado 3 vezes nas páginas 41, 42 e 43.

SOBHANI, M. et al. Combining weather stations for electric load forecasting. *Energies*, v. 12, p. 11, 2019. Citado 2 vezes nas páginas 19 e 23.

SPECHT, D. F. *Generation of polynomial discriminant functions for pattern recognition.* Tese (Doutorado) — Stanford University, 1966. Citado na página 46.

SPECHT, D. F. Probabilistic neural networks. *Neural Networks*, v. 3, p. 109–118, 1990. Citado na página 46.

THEIL, H. *Economic forecasts and policy.* [S.l.]: North-Holland, 1965. Citado na página 28.

THEIL, H. *Applied economic forecasts.* [S.l.]: North-Holland, 1966. Citado na página 28.

VENABLES, W. N.; SMITH, D. M.; TEAM, R. C. *R: A language and environment for statistical computing.* [S.l.], 2019. Disponível em: <http://www.R-project.org>. Citado na página 60.

VIEIRA, S. *Introdução a bioestatistica.* [S.l.]: Elsevier, 2015. Citado 2 vezes nas páginas 29 e 35.

WALLIS, K. F. Combining forecasts: forty years later. *Applied Financial Economics*, v. 21, p. 33–41, 2011. Citado na página 19.

WANG, L. F. et al. An estimation of distribution algorithm based on clayton copula and empirical margins. In: *Life System Modeling and Intelligent Computing.* [S.l.: s.n.], 2010. Citado na página 21.

YAGER, R. R. Modeling multi-criteria objective functions using fuzzy measures. *Information Fusion*, v. 29, p. 105–111, 2016. Citado na página 43.

ZHANG, G. P. A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, v. 177, p. 5329–5346, 2007. Citado na página 21.

ZOU, H.; YANG, Y. Combining time series models for forecasting. *International Journal for Forecasting*, v. 20, p. 69–84, 2004. Citado na página 33.

ANEXO A – PARÂMETROS DOS MODELOS LINEAR E LOG-LINEAR

Tabela 7 – Parâmetros do modelo LI para estimar k (conjunto de treinamento).

Séries	$\hat{k}$, Modelos e os Parâmetros de Regressão							
	$\hat{k}^{(\overline{\text{MSE}})}$		$\hat{k}^{(\overline{\text{MAE}})}$		$\hat{k}^{(\overline{\text{RMSE}})}$		$\hat{k}^{(\overline{\text{THEILU}})}$	
	a	b	a	b	a	b	a	b
SP	1113.09	-3.789	2942.67	-253.67	2395.67	-146.5	2393.5	-418433.8
ND	756.8	-0.23626	1219.3	-30.37	1114.6	19.31	1114.2	-96909.3
QB	1156.5	-2.6782	1865.4	-112.8	1872.1	-88.668	1864.7	-44482.1
PM	931.05	-0.2957	1814.3	-62.57	1411.5	-24.48	1086.1	-1643.8

Tabela 8 – Parâmetros do modelo LL para estimar k (conjunto de treinamento).

Séries	$\hat{k}$, Modelos e os Parâmetros de Regressão							
	$\hat{k}^{(\overline{\text{MSE}})}$		$\hat{k}^{(\overline{\text{MAE}})}$		$\hat{k}^{(\overline{\text{RMSE}})}$		$\hat{k}^{(\overline{\text{THEILU}})}$	
	a	b	a	b	a	b	a	b
SP	6061.9	-1085.7	6995.8	-2865.2	6408.4	-2305.2	-11932.0	-2304.5
ND	2941.0	-356.99	3003.9	-800.20	2941.0	-713.9	-3140.1	-713.43
QB	4403.0	-715.58	4007.8	-1414.3	4400.9	-1430.4	-4476.3	-1423.2
PM	3945.0	-479.86	4514.8	-1325.4	3938.2	-957.9	-228.93	-675.63

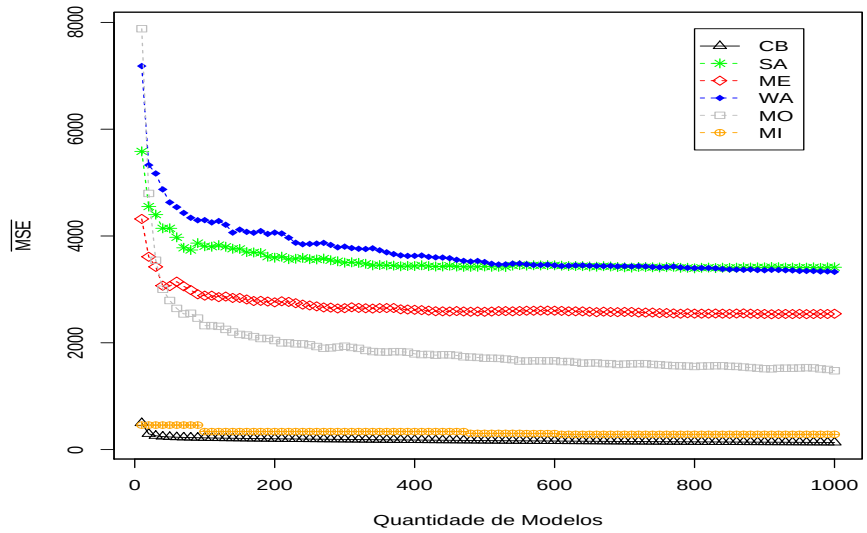
ANEXO B – COMPARAÇÃO ENTRE *ENSEMBLES* E MODELOS INDIVIDUAIS

As Figuras 29,30,31,32,33,34,35,36,37,38 apresentam os achados deste experimento. Os resultados obtidos através dos modelos individuais (MI) são comparados com os *ensembles*. As figuras (a) ilustram os resultados dos *ensembles* comparados com o melhor modelo individual encontrado entre os k modelos adotados no processo combinatório, assim quando $k = 10$, por exemplo, o experimento visa comparar o desempenho dos *ensembles* combinados com dez modelos individuais, com o desempenho do melhor MI representado pelo modelo individual com menor MSE (**min**(MI)). Por outro lado, as figuras (b) mostram as comparações entre os *ensembles* e o $\overline{\text{MSE}}$ de todos os modelos individuais.

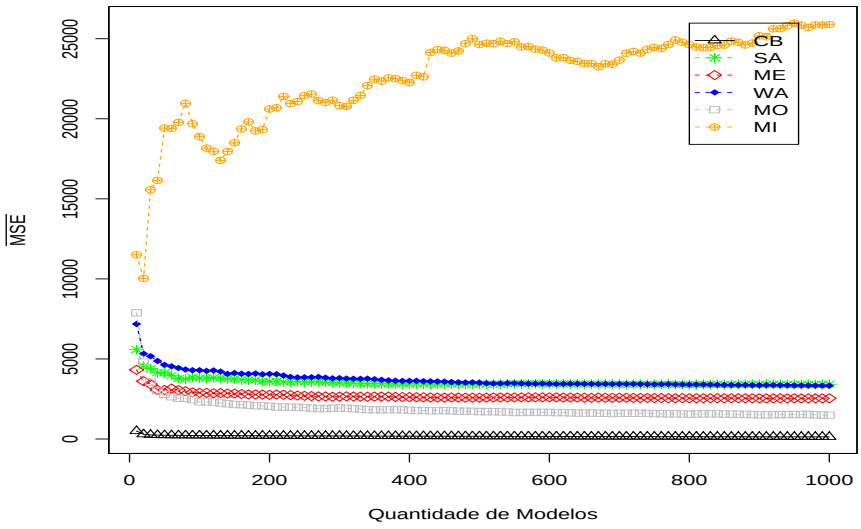
A Figura 29 (a), por exemplo, destaca a comparação entre CB, SA, ME, WA, MO e MI para a série SP, os melhores resultados para cada um dos modelos foram CB=132.9, SA=3414.8, WA=3328.2, ME=2543.5, MO=1476.1 e MI=284.5 para $k = 1000$, enfatizando que MI representa o MSE do melhor modelo individual dentre os k adotados na combinação dos *ensembles*. A figura (b) mostra o $\overline{\text{MSE}}$ dos *ensembles* e modelos individuais a medida que o valor de k cresce.

Os resultados de modo geral sugerem que os melhores modelos individuais são superiores aos *ensembles* baseados nas combinações lineares clássicas, o que não acontece com o CB que apresenta-se mais eficiente e acurado em relação ao melhor modelo individual. Contudo, quando o MI é comparado com as demais metodologias utilizando o $\overline{\text{MSE}}$ de todos os modelos individuais, ao invés de assumir apenas o melhor MI, os resultados são diferentes, indicando que MI é inferior a todos os *ensembles*, como mostra as figuras (b).

A literatura de combinação de modelos de previsão de séries temporais tem vários relatos quanto a superioridade dos *ensembles* em relação aos modelos individuais, contudo é normal que possa existir um modelo individual mais acurado e eficiente que os modelos combinados, porém a discussão que cerca este tema é o quão complicado possa ser encontrar tal modelo, ou ainda se o melhor modelos individual existe. Ao observar os resultados alcançados pela figura (b) fica claro que a maior parte dos modelos individuais são inferiores aos *ensembles* o que sugere que o processo de combinação é uma alternativa importante na busca de previsões mais acuradas e eficientes (maiores detalhes deste tema veja Clemen (1989)).

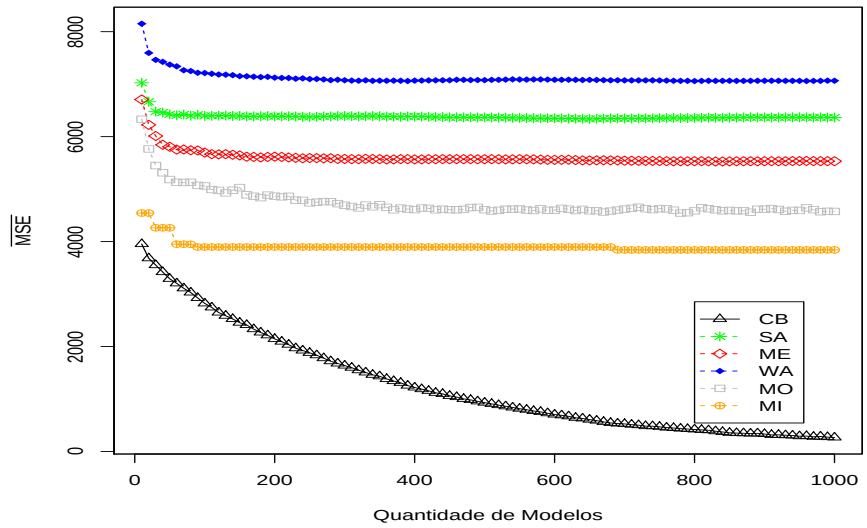


(a) Os melhores resultados são CB=132.9 e MI=284.5 (**min**(MI)).

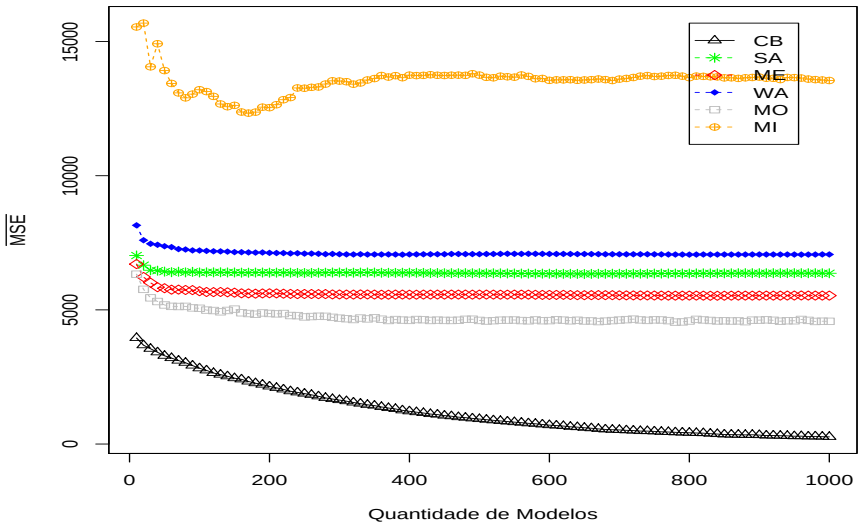


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 29 – Comparação entre os *Ensembles* e Modelos Individuais para a série SP.

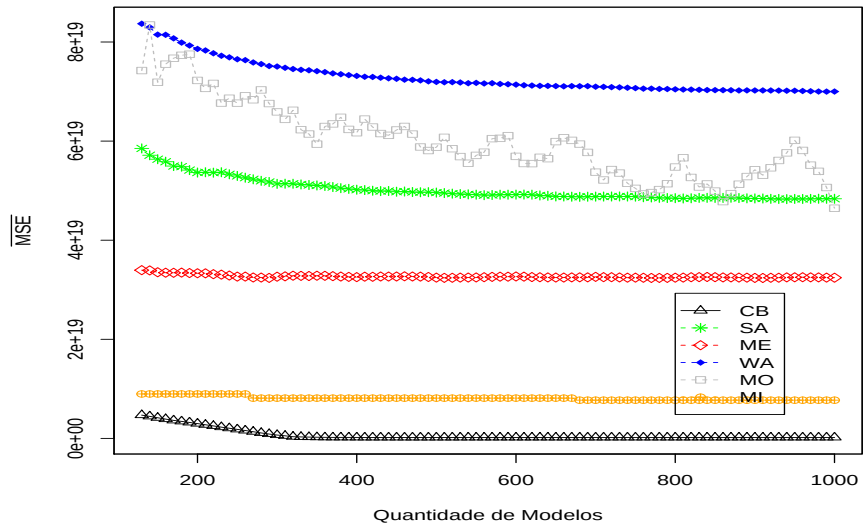


(a) *Ensembles* e o melhor Modelo Individual (**min**(MI)).

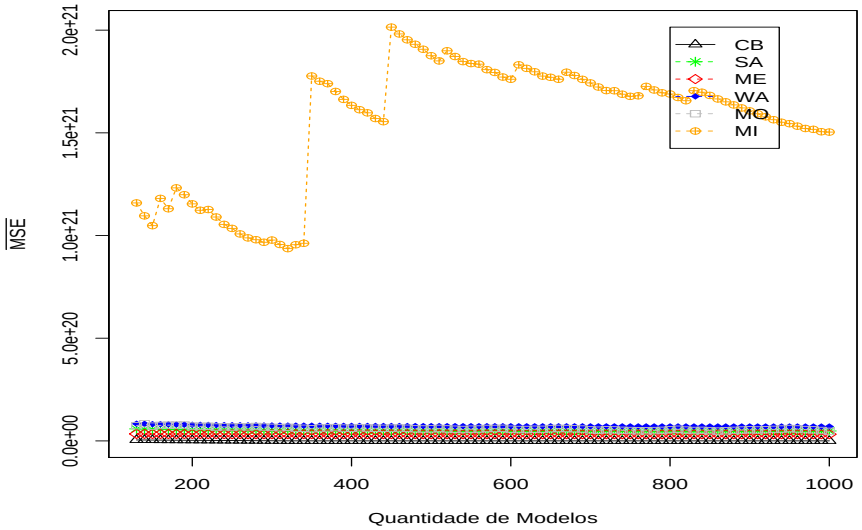


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 30 – Comparação entre os *Ensembles* e Modelos Individuais para a série ND.

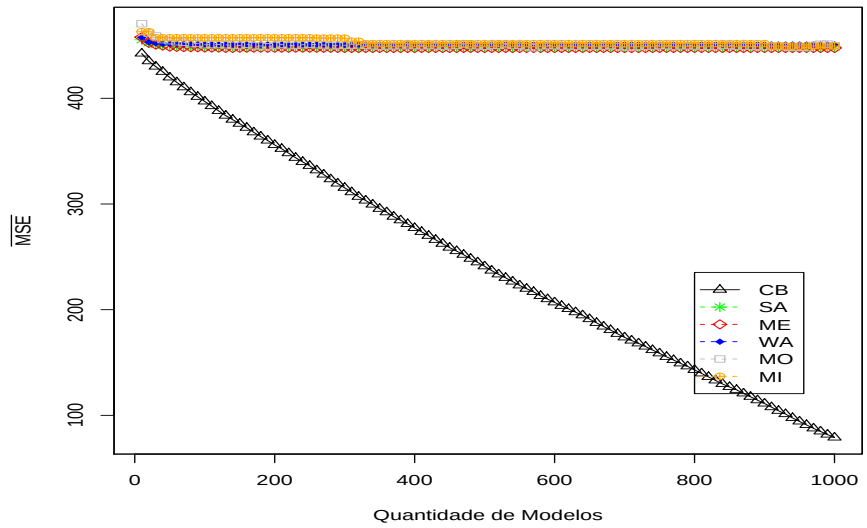


(a) *Ensembles* e o melhor Modelo Individual (**min**(MI)) para $k >= 130$.

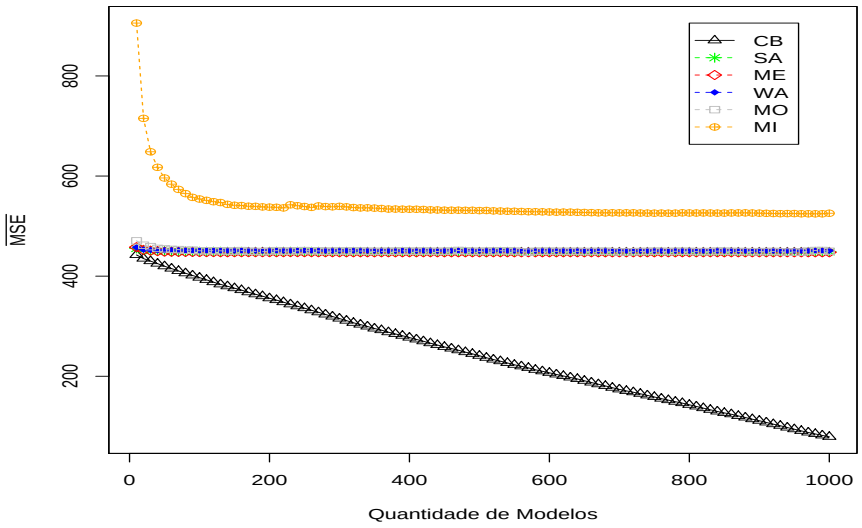


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 31 – Comparação entre os *Ensembles* e Modelos Individuais para a série DJ.

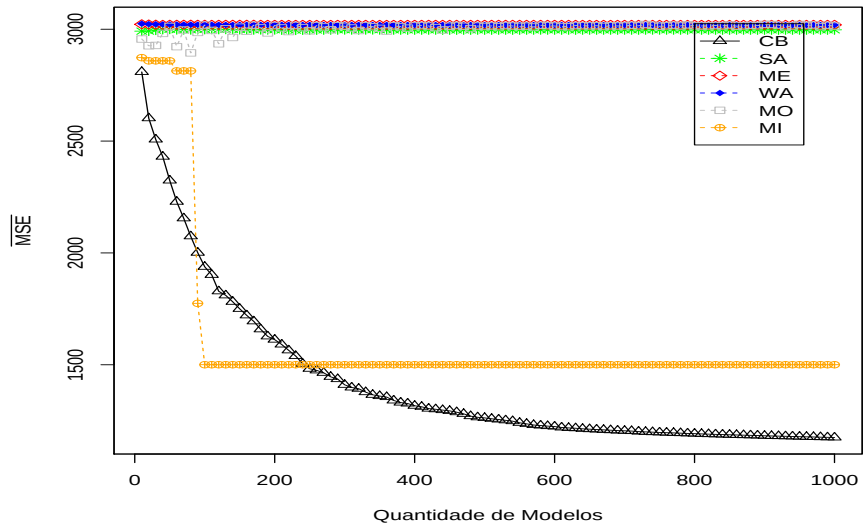


(a) Os melhores resultados são CB=79.0, SA=448.5, ME=447.6, WA=450.8, MO=449.7 e MI=449.5 (**min**(MI)).

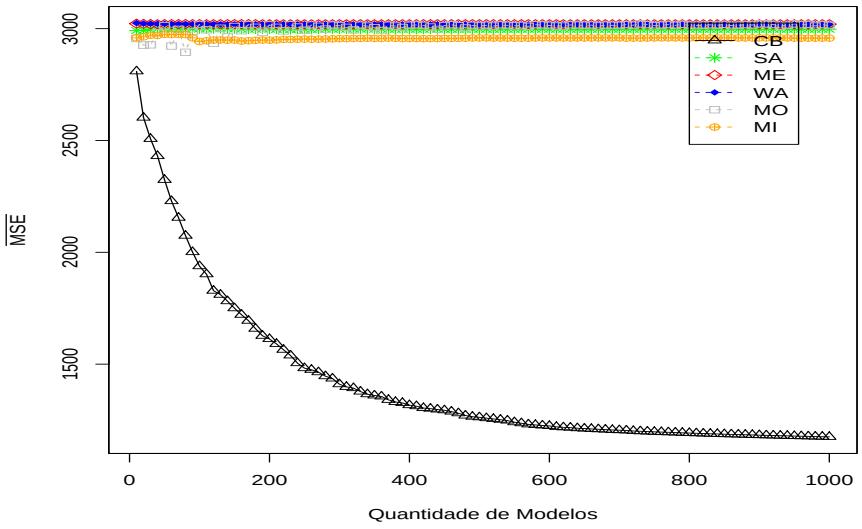


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 32 – Comparação entre os *Ensembles* e Modelos Individuais para a série QB.

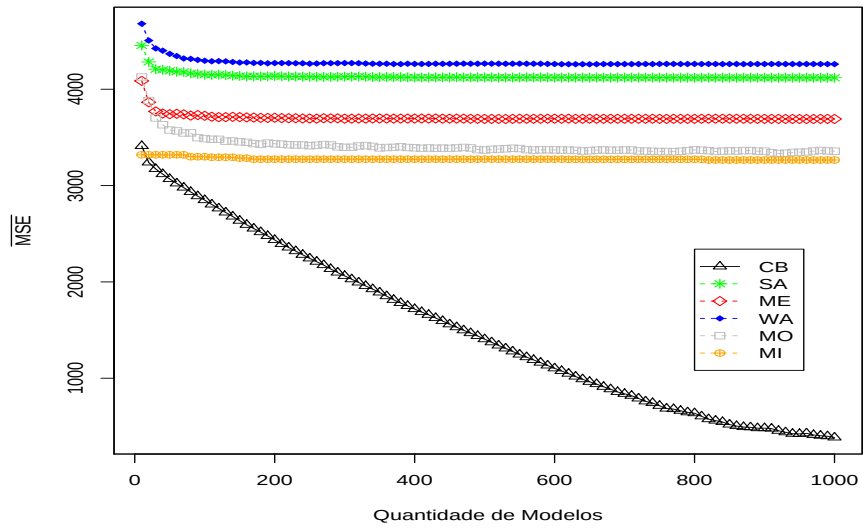


(a) *Ensembles* e o melhor Modelo Individual (**min**(MI)).

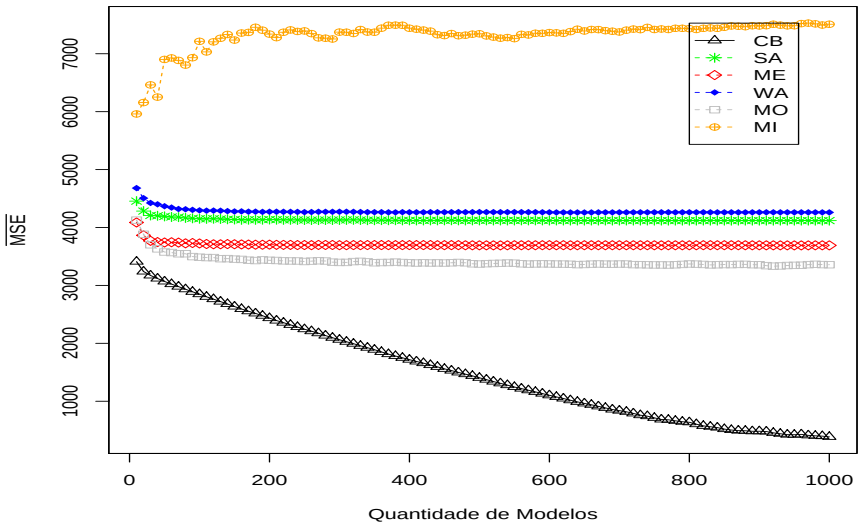


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 33 – Comparação entre os *Ensembles* e Modelos Individuais para a série RS.

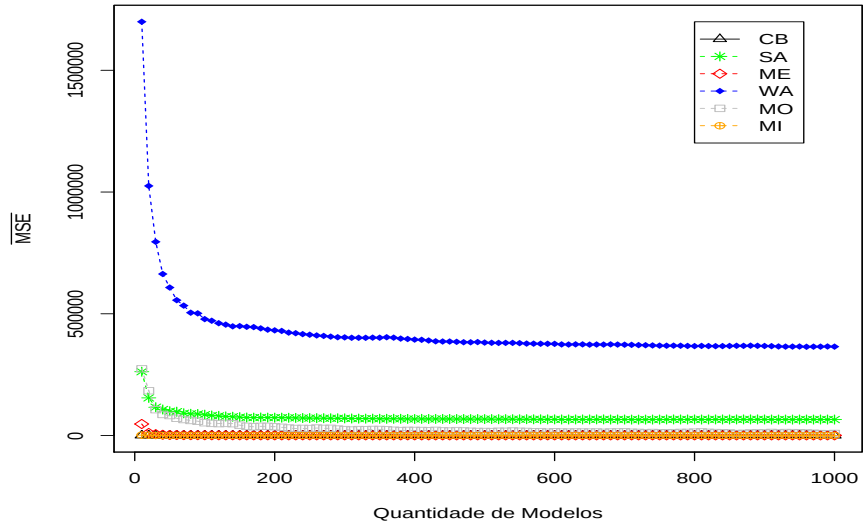


(a) *Ensembles* e o melhor Modelo Individual (**min**(MI)).

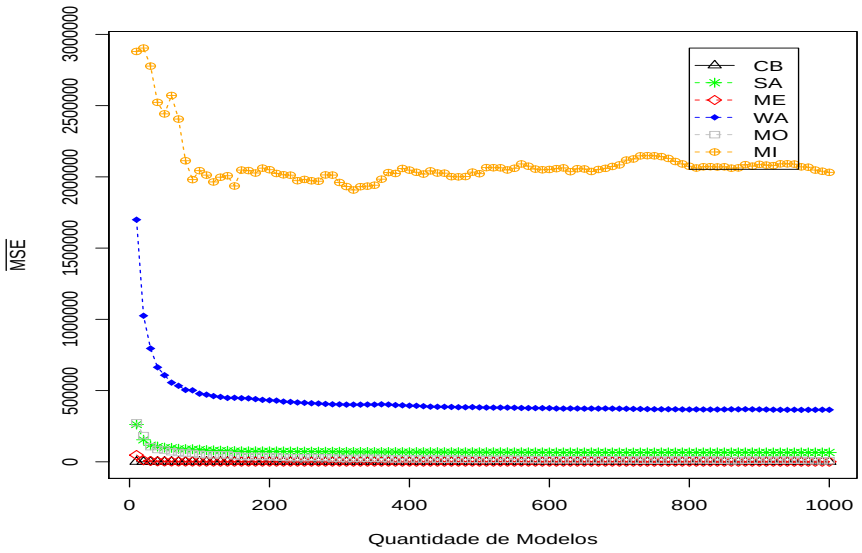


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 34 – Comparação entre os *Ensembles* e Modelos Individuais para a série PM.

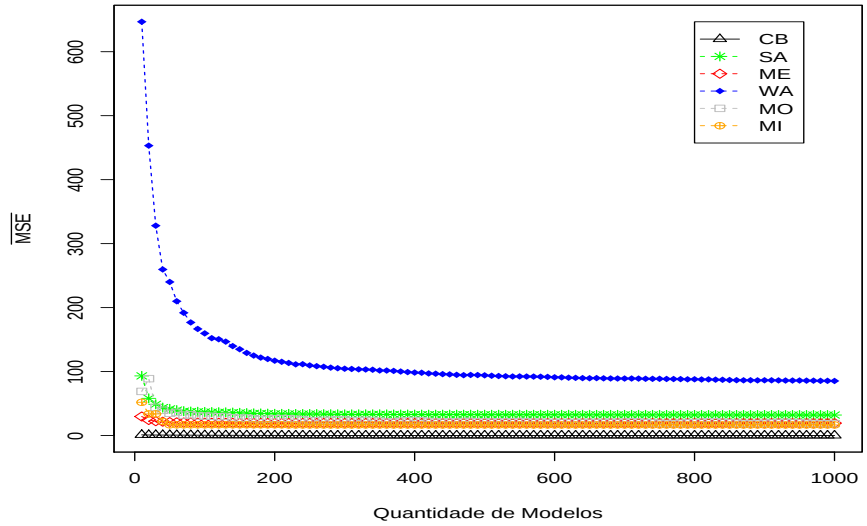


(a) Os melhores resultados são CB=0.6, ME=597.8, MO=8366.1 e MI=116.1 (**min**(MI)).

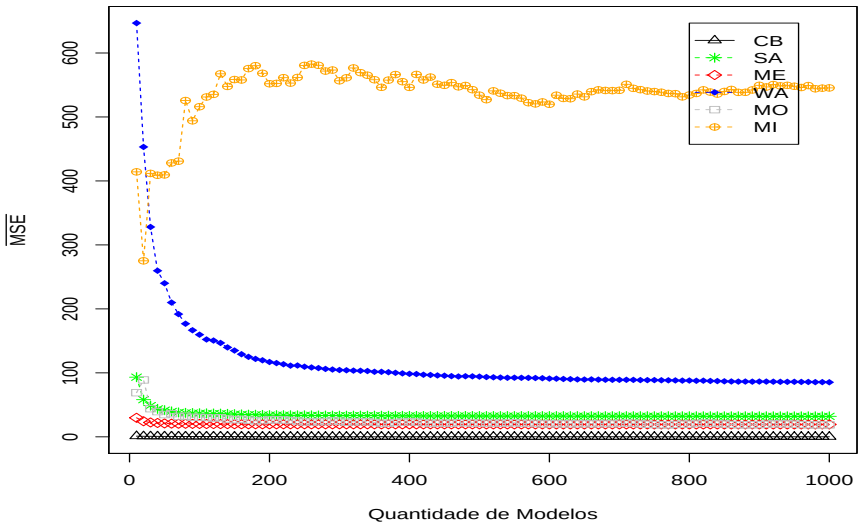


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 35 – Comparação entre os *Ensembles* e Modelos Individuais para a série RO.

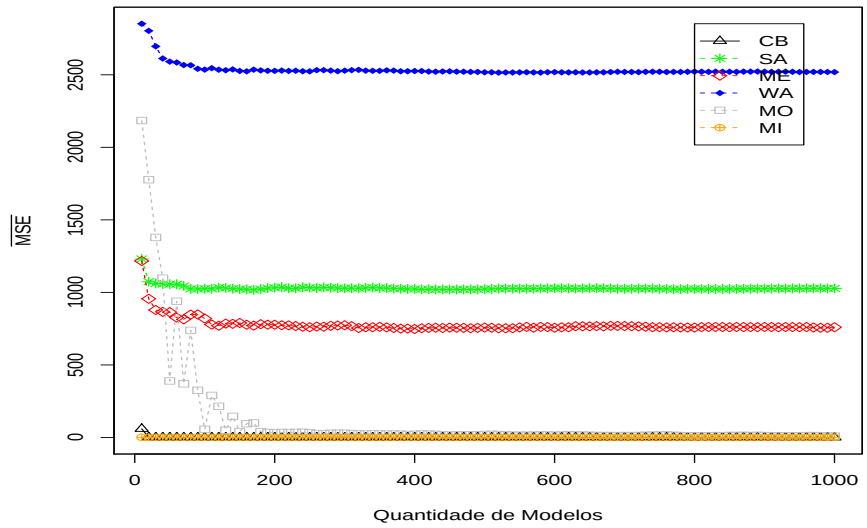


(a) Os melhores resultados são CB=0.07, ME=19.6, MO=18.7 e MI=16.5 (**min**(MI)).

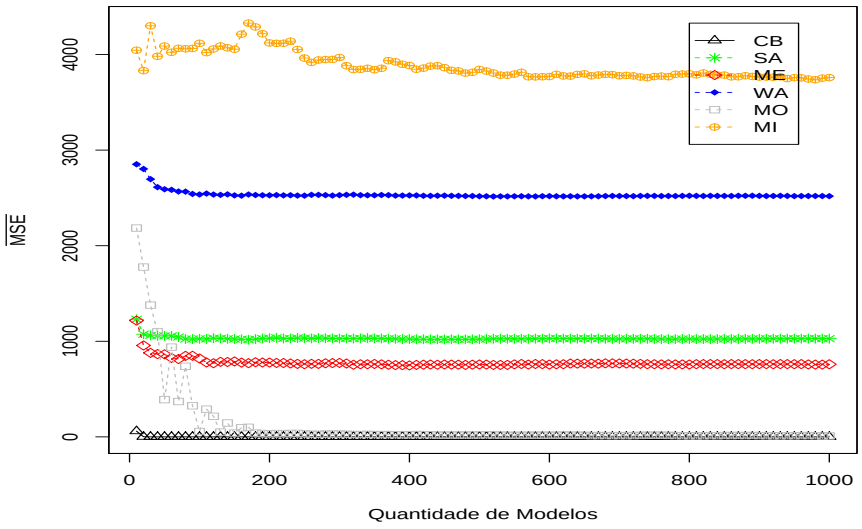


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 36 – Comparação entre os *Ensembles* e Modelos Individuais para a série TO.

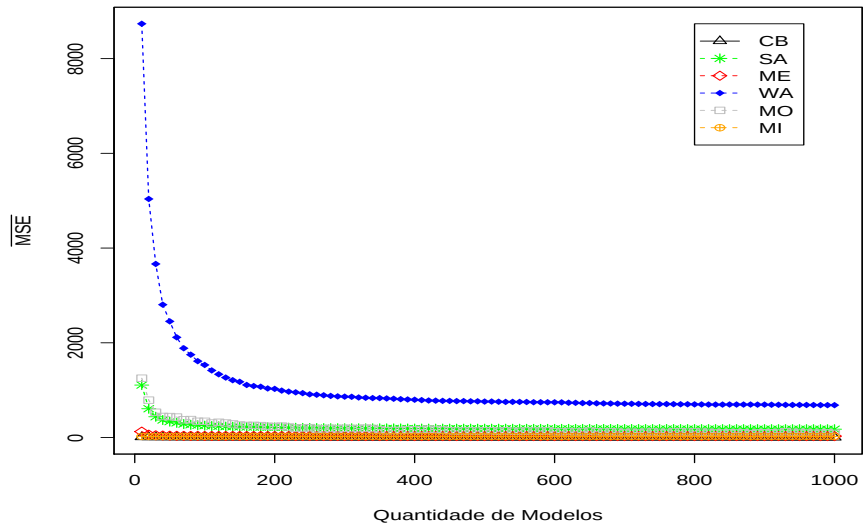


(a) Os melhores resultados são CB=0.003, MO=9.630 e MI=0.770 (**min**(MI)).

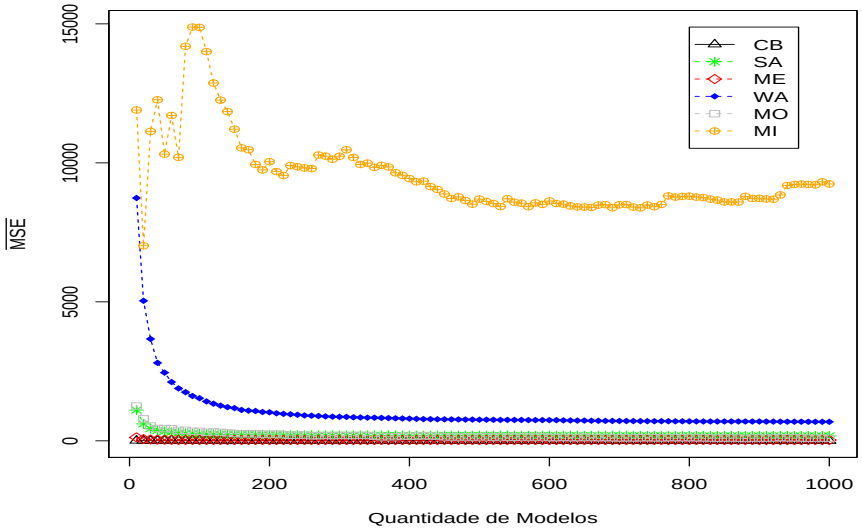


(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 37 – Comparação entre os *Ensembles* e Modelos Individuais para a série RF.



(a) Os melhores resultados são CB=0.3, SA=176.1, ME=29.1, MO=89.0 e MI=23.1 (**min**(MI)).



(b) $\overline{\text{MSE}}$ dos *Ensembles* e Modelos Individuais ($\frac{1}{n}\sum_{i=1}^{n}$ MI).

Figura 38 – Comparação entre os *Ensembles* e Modelos Individuais para a série PO.