UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENERGIA NUCLEAR
PÓS-GRADUAÇÃO EM TECNOLOGIAS ENERGÉTICAS E NUCLEARES

GABRIEL DANTAS DE OLIVEIRA ROLIM

DOWNSCALING ESTATÍSTICO DO VENTO LOCAL BASEADO EM DEFINIÇÃO OBJETIVA DO CONJUNTO DE VARIÁVEIS REGRESSORAS

Recife
2020

GABRIEL DANTAS DE OLIVEIRA ROLIM

DOWNSCALING ESTATÍSTICO DO VENTO LOCAL BASEADO EM DEFINIÇÃO OBJETIVA DO CONJUNTO DE VARIÁVEIS REGRESSORAS

Dissertação submetida ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco para obtenção do título de Mestre em Ciências.

Área de Concentração: Fontes renováveis de energia.

Orientadora: Profa. Dra. Olga de Castro Vilela

Coorientador: Prof. Dr. Alexandre Carlos Araújo da Costa

Recife

2020

Catalogação na fonte
Bibliotecário Gabriel Luz, CRB-4 / 2222

R748d Rolim, Gabriel Dantas de Oliveira

Downscaling estatístico do vento local baseado em definição objetiva do conjunto de variáveis regressoras / Gabriel Dantas de Oliveira Rolim – Recife, 2020.

93 f.: figs., tabs.

Orientadora: Profa. Dra. Olga de Castro Vilela.

Coorientador: Prof. Dr. Alexandre Carlos Araújo da Costa.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, 2020.

Inclui referências e apêndices.

1. Tecnologias Energéticas e Nucleares. 2. Energia eólica. 3. Downscaling estatístico. 4. Seleção de variáveis. 5. Screening regression. 6. Região Nordeste do Brasil. I. Vilela, Olga de Castro (Orientadora). II. Costa, Alexandre Carlos Araújo da (Coorientador). III. Título.

UFPE

621.042 CDD (22. ed.) BCTG / 2021 - 124

GABRIEL DANTAS DE OLIVEIRA ROLIM

DOWNSCALING ESTATÍSTICO DO VENTO LOCAL BASEADO EM DEFINIÇÃO OBJETIVA DO CONJUNTO DE VARIÁVEIS REGRESSORAS

Dissertação submetida ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco para obtenção do título de Mestre em Ciências.

Aprovada em: 02 / 10 / 2020

BANCA EXAMINADORA

_____

Profa. Dra. Dóris Regina Aires Veleda
Universidade Federal de Pernambuco

_____

Prof. Dr. Tsang Ing Ren
Universidade Federal de Pernambuco

_____

Prof. Dr. Gilney Figueira Zebende
Universidade Estadual de Feira de Santana

AGRADECIMENTOS

Agradeço:

À Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE), projeto HPC4E (no âmbito do Programa H2020 da União Europeia a 3ª chamada de TIC da RNP/MCTI) (https://hpc4e.eu/the-project/work-plan/wp4) e Projeto IBITU.INTELIPREV (no âmbito do Programa de P&D ANEEL) pelo apoio financeiro para a execução do trabalho aqui apresentado.

Ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da UFPE (PROTEN-UFPE) pelo interesse e atenção dispensados durante o mestrado.

Ao Pierre Pinson e à Danmarks Tekniske Universitet (Universidade Técnica da Dinamarca - DTU) pela orientação e suporte durante minha estadia na Dinamarca.

À minha família pela inspiração, motivação e suporte incondicional em todos os momentos de minha vida, a vocês eu devo tudo e espero, algum dia, conseguir retribuir.

À Renata Boschi pelo companheirismo, inspiração e apoio (sem esquecer do hygge, claro!). À Sylvia e Alfredo por todo o apoio.

À Nina, Thiago Tavares, Felipe, Marília, Gabriel Muniz, Luciano, Runá, Nataly, Daniel Marquim, e demais companheiros de longa data pelas lições e por transformar qualquer que seja a jornada em uma boa lembrança.

À Alexandre e Olga pela inspiração, oportunidade e por revelar que é possível acordar todos os dias para fazer aquilo que se ama. Espero poder, algum dia, ser capaz de inspirar e conceder a outros estudantes a oportunidade que vocês me concederam.

Obviamente não se faz ciência (ou qualquer outra coisa) sozinho. Por isso, devo agradecer aos meus companheiros do CER pelo imenso companheirismo e ideias geniais que me fazem perder sono e acordar todos os dias com uma imensa vontade de viver. Obviamente tenho que citar alguns nomes fundamentais em minha jornada: Valentin Perruci (meu parceiro de longa data e interminável fonte de ideias brilhantes), Leonardo Aquino (que permanece genial, mesmo estando no lado negro da força), Janis Joplin (detentora de uma inteligência imensa e de uma gentileza ainda maior), Evelyn (catalisadora da felicidade existente no CER), Charles (genial no que quer que faça), Leonardo Petribú (uma mente brilhante com uma didática incrível), Emily (uma pessoa excepcional no que quer que faça), Renan (detentor da mente mais ágil e barista do café mais cheiroso da cidade), Thiago (o mago da atmosfera, capaz de ocupar Terabytes em poucos minutos), Marcos e Henrique (companheiros atenciosos).

RESUMO

O rápido crescimento da energia eólica no mundo e, particularmente, no Brasil, vem demandando uma descrição cada vez mais acurada do comportamento do vento nos possíveis locais de centrais eólicas. Com relação a esse tema, técnicas objetivas de *downscaling* estatístico, principalmente aquelas baseadas em modelos regressivos, desempenham um papel fundamental. Tais técnicas baseiam-se na minimização de uma função de custo, que permite ao modelo extrair informações oriundas de previsão numérica do tempo (e.g., *General Circulation Models* outputs) para descrever o comportamento do vento na microescala. Contudo, até o momento, apenas poucos estudos têm concentrado seus esforços em desenvolver e/ou aplicar técnicas objetivas para selecionar quais variáveis, quando utilizadas como entrada para os modelos regressivos, promovem a maior acurácia do *downscaling* estatístico. Nesse contexto, este trabalho propõe uma nova metodologia para a seleção do conjunto de variáveis regressoras para o *downscaling* da magnitude da velocidade do vento horizontal, que demanda um baixo esforço computacional, sendo bastante flexível a ponto de ser aplicada a qualquer região de interesse, em conjunto com qualquer modelo regressivo e empregando dados de reanálise (visando, por exemplo, a avaliação do recurso eólico) ou dados de previsão (visando, por exemplo, o despacho da potência de centrais eólicas junto ao operador do sistema elétrico). Particularmente, o modelo regressivo utilizado neste trabalho foi a regressão linear múltipla devido à sua simplicidade. Os resultados deste trabalho constatam a grande viabilidade de utilização da metodologia desenvolvida tanto em relação a modelos de seleção específicos da regressão linear múltipla (i.e., *Stepwise Regression* e Regressão Lasso), quanto a um modelo geral de seleção de variáveis regressoras que demanda um grande esforço computacional (i.e., *Forward Selection*).

Palavras-chave: Energia eólica. Downscaling estatístico. Seleção de variáveis. Screening regression. Região Nordeste do Brasil.

ABSTRACT

The fast growth of wind energy in the world and, mainly, in Brazil, has demanded an increasingly accurate description of the wind's behavior in wind farms' possible locations. Regard to such themes, statistical downscaling objective techniques, especially those based on regressive models, perform a fundamental role. Such techniques are based on the minimization of a cost function, which allows the model to extract information derived from numerical weather prediction models (e.g., General Circulation Models outputs) to describe the wind's behavior at the microscale. However, so far, only a few studies have concentrated their efforts on developing or applying objective techniques to select which variables, when used as input for regressive models, promote better statistical downscaling accuracy. In this context, this work proposes a new methodology for selecting the set of regression variables for the downscaling of the horizontal wind speed magnitude. This methodology requires a low computational effort, being very flexible to the point of being applied to any region of interest, in conjunction with any regressive model and employing reanalysis data (aiming, for example, the wind resource assessment) or forecasting data (aiming, for example, the power dispatched by wind farms to the transmission system operator). The regressive model used in this work was the multiple linear regression due to its simplicity. The results of this work confirm the great feasibility of using the developed methodology both concerning models that only can be applied in conjunction with multiple linear regression (i.e., Stepwise Regression and Lasso Regression), as well as a general model of regressor variables selection that demands a higher computational effort (i.e., Forward Selection).

Keywords: Wind energy. Statistical downscaling. Variable selection. Screening regression. Northeastern Brazil.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A crescente demanda mundial por energia, com vistas a suprir a demanda social, aliada à redução da dependência de combustíveis fósseis vem promovendo um rápido crescimento na implantação de fontes renováveis de energia (REN21, 2020).

No Brasil, a necessidade de diversificação da matriz elétrica nacional, através da inserção de fontes limpas e renováveis de energia, vem dando origem a diversas políticas de incentivo (e.g. Resoluções Normativas ANEEL n° 482/2012 e 687/2015). Por esse motivo, o setor eólico vem vivenciando uma rápida expansão. O Brasil já conta com uma potência eólica instalada de 15,37 GW, contribuindo com 8,6% da matriz elétrica nacional (EPE, 2020). Nesse sentido, a Região Nordeste do Brasil ocupa um local de destaque, pois detém mais de 80% da capacidade eólica instalada (ABEEólica, 2020).

Esse rápido crescimento experienciado pela energia eólica vem demandando uma descrição cada vez mais precisa do comportamento do vento nos possíveis locais de instalação de centrais, seja com o intuito de caracterizar climatologicamente a variável no local de interesse, ou com o intuito de prever o recurso disponível, auxiliando na tomada de decisão dos operadores de usinas e do sistema elétrico.

Diversas aplicações que demandam a descrição do comportamento de variáveis atmosféricas fazem uso dos resultados concedidos por Modelos de Circulação Geral Atmosférica (GCM). Esses modelos simulam – através da solução numérica das equações de conservação de massa, energia e quantidade de movimento – o comportamento de diversas grandezas atmosféricas com ampla cobertura espacial (ao longo de todo o globo, em diversos níveis verticais) e temporal (KALNAY *et al.*, 1996).

No entanto, as simulações realizadas por esses modelos demandam um alto esforço computacional, principalmente tendo em vista a ampla cobertura espacial e temporal. Portanto, torna-se imprescindível o emprego de baixa resolução espacial (da ordem de centenas de quilômetros) e temporal (da ordem de horas), para que a simulação do modelo seja realizada em tempo hábil.

Em virtude da baixa resolução espacial, os resultados oriundos dos Modelos de Circulação Geral Atmosférica são insuficientes para a maioria das aplicações em energia eólica, tendo em vista que as centrais eólicas possuem dimensão típica de 10 km e os rotores aerodinâmicos dos aerogeradores possuem diâmetro típico de 100m.

O aumento da resolução das saídas dos GCMs é realizado através do uso de técnicas de *downscaling*. Essas técnicas se alimentam das saídas do GCM com vistas a descrever o

comportamento da variável atmosférica de interesse com uma resolução espacial mais alta (PANOFSKY e DUTTON, 1987; ORLANSKY, 1975).

Nesse sentido, as técnicas estatísticas de *downscaling*, principalmente aquelas que fazem uso de modelos regressivos, desempenham um papel fundamental, pois elas se baseiam na minimização de uma função de custo (WILKS, 2019), possibilitando que o modelo extraia informações provenientes de uma base de dados com baixa resolução para descrever o comportamento da variável de interesse com uma resolução mais alta.

Em geral, as variáveis empregadas como entrada para os modelos regressivos de *downscaling* (i.e., variáveis regressoras) são definidas subjetivamente, optando-se preferencialmente por aquelas mais próximas ao local de interesse (HEWITSON e CRANE, 1996; CURRY *et al.*, 2012; DUPRÉ *et al.*, 2020). No entanto, alguns estudos revelam que variáveis localizadas em pontos da malha do GCM distantes do local de interesse podem contribuir para uma maior acurácia do *downscaling* estatístico (HOFER *et al.*, 2010; SAUTER e VENEMA, 2011). Isso porque, esses pontos podem possuir informações complementares àquelas contidas nos pontos de malha mais próximos ao local de interesse, ou podem apresentar uma melhor descrição do comportamento da série observada (FOWLER *et al.*, 2007; HOFER *et al.*, 2010)**.**

Poucos estudos se dedicaram a desenvolver ou aplicar métodos objetivos[1] (i.e., automáticos) para a seleção do conjunto de variáveis regressoras, de tal forma a maximizar o desempenho do *downscaling* estatístico (RADANOVICS *et al.*, 2013). Além disso, grande parte desses estudos necessita da definição subjetiva de parâmetros (e.g., HOFER *et al.*, 2010; SAUTER e VENEMA, 2011; GUO *et al.*, 2012), ou fazem uso de métodos que apenas podem ser aplicados em conjunto com um determinado modelo regressivo (e.g., HAMMAMI *et al.*, 2012; GAO *et al.*, 2014; YANG *et al.*, 2017; TEEGAVARAPU e GOLY, 2018). Já os métodos de seleção que podem ser aplicados em conjunto com qualquer modelo regressivo, em geral demandam um alto esforço computacional, inviabilizando sua utilização em diversas aplicações que exigem uma resposta rápida por parte do modelo (e.g., SAUTER e VENEMA, 2011).

Assim, o objetivo deste trabalho é o desenvolvimento de uma metodologia que selecione quais variáveis, ao serem utilizadas como variáveis regressoras, promovem o melhor

[1] Diversos estudos empregam o termo "subjetivo" fazendo referência a métodos que carecem, para sua execução, de conhecimento subjetivo (e.g., GILCHRIST e CRESSMAN, 1954; GUSTAFSSON, 1981; EISCHEID et al., 1995). Nesse sentido, desde o ponto de vista da execução do método, o termo "objetivo" pode ser compreendido como "automático". Vale destacar que, mesmo sendo o método objetivo (automático) em sua execução, a inferência com respeito a um ou mais parâmetros (entre aqueles que compõem o algoritmo) pode se dar de forma subjetiva

desempenho do *downscaling* estatístico, fazendo uso de um baixo esforço computacional e que possa ser empregada em conjunto com qualquer modelo regressivo. Além do mais, é necessário que essa metodologia seja objetiva, ou seja, não dependa da definição subjetiva de parâmetros. Dessa forma, essa metodologia poderá ser empregada em uma vasta gama de aplicações.

Nesse contexto, a metodologia proposta neste trabalho tem por objetivo ordenar o conjunto de possíveis variáveis regressoras de acordo com sua relevância, que será dada pelas funções de ordenamento e, em seguida, indicar quais dessas variáveis, quando utilizadas como variáveis regressoras, produzem o melhor desempenho do *downscaling* estatístico.

As inovações deste trabalho se verificam através de três pontos principais. O primeiro deles é a criação de métricas que definam a relevância de cada uma das variáveis do domínio de busca (i.e., funções de ordenamento). Essas métricas devem definir a relevância de tal forma que a utilização das variáveis mais relevantes do domínio de busca como entrada para o método regressivo promova a melhor acurácia do *downscaling* estatístico.

A segunda é comparação do desempenho obtido pela metodologia desenvolvida neste trabalho com aquele obtido por metodologias consagradas na literatura.

A terceira é o estudo da melhor forma de dividir a base de dados com vistas a uma melhor tomada de decisão por parte da metodologia desenvolvida neste trabalho. Além disso, foi analisada a robustez da metodologia com respeito a diferentes formas de divisão da base de dados.

A variável atmosférica a ser estudada ao longo do trabalho será a magnitude da velocidade do vento horizontal com vistas a aplicações relacionadas à área de energia eólica. Como principais exemplos dessas aplicações, temos: a reanálise com vistas à tomada de decisão sobre implantação de centrais eólicas e a previsão com vistas ao despacho da produção de centrais eólicas. Contudo, é importante frisar que essa metodologia pode ser aplicada a qualquer variável atmosférica, abrangendo um número ainda maior de aplicações.

Este trabalho está dividido em seis capítulos principais: 1) Introdução; 2) Conceitos Preliminares, onde são introduzidos conceitos básicos necessários para a compreensão do trabalho; 3) Revisão Bibliográfica, na qual são examinados os estudos mais relevantes e atuais relacionados ao tema proposto nesta dissertação; 4) Metodologia e Modelos; 5) Resultados e Discussão, na qual são apresentados e discutidos os principais resultados obtidos ao longo deste trabalho; 6) Conclusões e Perspectivas. Ao final do trabalho, situa-se a lista de referências bibliográficas, seguida pelos Apêndices A e B. O Apêndice A apresenta figuras relacionadas com a discussão apresentada no capítulo de resultados. Já o Apêndice B, apresenta resultados

relacionados à utilização de componentes principais com o intuito de diminuir a colinearidade das variáveis a serem selecionadas.

2 CONCEITOS PRELIMINARES

Este capítulo apresentará conceitos preliminares de fundamental importância para a compreensão dos temas abordados ao longo deste trabalho.

2.1 Modelos de Circulação Geral Atmosférica

Os Modelos de Circulação Geral Atmosférica (*General Circulation Models – GCM*) são modelos que simulam - através da solução numérica das equações de conservação de massa, energia e momento linear - o comportamento de diversas grandezas atmosféricas com ampla cobertura espacial (ao longo de todo o globo em diversos níveis verticais – ver Figura **1**) e temporal (suficientemente longos para caracterizar o comportamento climático das variáveis atmosféricas). Esses modelos são desenvolvidos e disponibilizados por grandes centros de pesquisa, tendo como principais representantes o European Centre for Medium-Range Weather Forecasts (ECMWF) e o National Centers for Environmental Prediction (NCEP).

Figura 1 – Exemplo de malha de pontos de um GCM

Fonte: https://str.llnl.gov/december-2017/bader.

Dentre os diversos produtos disponibilizados pelos Modelos de Circulação Geral Atmosférica, destacam-se: análise, reanálise e previsão. A análise é fruto de modelos operacionais (i.e., opera em tempo real) e utiliza uma base dados observacionais disposta de forma irregular sobre o terreno (que pode ser alterada ao longo do tempo de simulação) com

vistas a descrever o comportamento atmosférico em uma malha regular. Dessa forma, os dados oriundos da análise constituem uma base heterogênea, tanto devido a mudanças na base de dados assimilada pelo modelo quanto devido às mudanças na parametrização do modelo ao longo das simulações.

Já nos modelos de reanálise, as simulações são realizadas de forma não operacional e a base dados assimilada, bem como as parametrizações empregadas, não são modificadas ao longo do período de simulação, constituindo, dessa forma, uma base de dados homogênea (i.e. as bases de dados de reanálise não são afetadas por mudanças no método ao longo do período de simulação) (DEE, UPPALA, *et al.*, 2011). A previsão, por outro lado, emprega modelos operacionais com vistas a descrever o estado da atmosfera em instantes futuros. Assim como a base de dados de análise, a base de dados de previsão também é heterogênea, pois as parametrizações de modelo empregadas, bem como os dados assimilados podem ser alterados ao longo da simulação.

Todo o processo envolvido na simulação dos estados atmosféricos realizada por um GCM requer um alto esforço computacional, principalmente tendo em vista a ampla cobertura espacial e temporal. Devido a isso, torna-se imprescindível o emprego de baixa resolução espacial (da ordem de centenas de quilômetros) e temporal (da ordem de horas), para que a simulação do modelo seja realizada em tempo hábil.

Mesmo com baixa resolução espacial e temporal, as saídas dos GCMs são amplamente empregadas como descrição preliminar dos estados atmosféricos em diversas aplicações. Contudo, a sua resolução espacial é insuficiente para a maioria das aplicações em energia eólica, tendo em vista que as centrais eólicas possuem dimensão típica de 10 km e os rotores aerodinâmicos dos aerogeradores possuem diâmetro típico de 100m.

2.2 Aumento da Resolução (*Downscaling*)

O termo *downscaling* refere-se ao conjunto de técnicas utilizadas para descrever o comportamento de uma variável atmosférica em uma escala menor (microescala) a partir de informações oriundas de escalas maiores (PANOFSKY e DUTTON, 1987) (ORLANSKI, 1975). Nos casos de estudo apresentados neste trabalho isso se faz necessário, pois a resolução das saídas de simulações atmosféricas macroescalares (como os GCMs) não permitem a reprodução de importantes fenômenos atmosféricos nos pontos de interesse em terra, ou seja, na microescala (WILBY e WIGLEY, 1997). Além disso, Wilby & Wigley (1997) indicam que por mais que a tecnologia e a resolução espacial e temporal dos GCMs aumentem, as técnicas

de *downscaling* sempre serão necessárias, porque, ainda assim, os GCMs não serão capazes de enxergar diversos fenômenos microescalares.

As técnicas de *downscaling* podem ser classificadas em duas principais abordagens: técnicas dinâmicas e técnicas estatísticas.

2.2.1 Abordagem dinâmica

As técnicas dinâmicas de *downscaling* são aquelas que fazem uso de variáveis macroescalares como condições iniciais e de contorno e resolvem numericamente as equações físicas que governam o comportamento da atmosfera (COSTA, 2005). Podemos dividir a abordagem dinâmica de *downscaling* em dois subgrupos: a modelagem de camada limite planetária (PBL, *Planet Boundary Layer*) (LANDBERG e WATSON, 1994; AQUINO, 2017) que realiza o acoplamento entre a macroescala (GCM) e a microescala (escala com resolução espacial da ordem de centenas de metros); e os Modelos de Área Limitada (*Limited Area Models*, *LAM;* ou *Nested Models*) (WILBY e WIGLEY, 1997) que fazem uso das saídas dos GCMs como condições de contorno para caracterizar o comportamento do vento na mesoescala (escala cuja resolução espacial é da ordem de dezenas de quilômetros).

Os Modelos de Camada Limite Planetária se baseiam na simplificação da Lei da Conservação do Momento Linear com vistas a caracterizar o comportamento do vento e de outras variáveis atmosféricas no interior da PBL. Essa simplificação se dá através de análise dimensional, desprezando-se os termos de menor importância (PEDLOSKY, 1984), e isso faz com que os Modelos de Camada Limite Planetária demandem um baixo esforço computacional na realização das simulações. Os Modelos de PBL podem se apresentar de duas formas: a primeira delas e mais simples, denominada Regime Neutro, parte do pressuposto que o efeito da troca de calor entre a atmosfera e a superfície terrestre é desprezível (LANDBERG e WATSON, 1994); A segunda, denominada Regime Não Neutro, leva em consideração a troca de calor entre a atmosfera e a superfície terrestre (LANDBERG *et al.*, 2001), sendo de fundamental importância na descrição do comportamento atmosférico em locais com alta instabilidade atmosférica (i.e. locais em que há um fluxo ascendente de calor, partindo do solo em direção à atmosfera), como, por exemplo, a Região Nordeste do Brasil.

Os modelos de Área Limitada, também denominados Modelos de Circulação Regional (*Regional Circulation Models*, *RCM*) (GRELL *et al.*, 1995), são modelos que se assemelham aos GCMs e utilizam as saídas do GCM como condições de contorno na simulação do comportamento atmosférico em uma malha com resolução maior que a do GCM. Esse aumento

de resolução faz com que o modelo incorpore informações regionais com maior grau de detalhamento (e.g., orografia). Os Modelos de Circulação Regional, assim com os GCMs, também podem realizar assimilação de dados observacionais. Contudo, devido à grande complexidade envolvida na solução numérica das equações de conservação de massa, energia e momento linear, suas simulações demandam um grande esforço computacional, sendo essa uma de suas principais desvantagens.

2.2.2 Abordagem estatística

As técnicas estatísticas de *downscaling* são aquelas que se utilizam de relações empíricas entre os dados observacionais e a circulação na escala sinóptica (saída do GCM) com o intuito de descrever o comportamento microescalar da variável de interesse (WILBY, CHARLES, *et al.*, 2004). Elas podem ser classificadas em três grandes grupos, são eles: Modelos de Classificação; Geradores Estocásticos no Tempo e Modelos Regressivos. É válido salientar que é bastante comum a combinação de técnicas distintas de *downscaling* estatístico com vistas a realizar a descrição de uma variável atmosférica na microescala.

Os modelos de classificação dividem os instantes de tempo em um número finito de estados atmosféricos, seja aplicando análise de agrupamento (*cluster analysis*) às variáveis macroescalares ou utilizando classificações subjetivas dos estados atmosféricos (BÁRDOSSY e CASPARY, 1990). Dentre as principais técnicas empregadas no intuito de realizar classificação dos estados sinópticos, podemos citar: K-Médias (*k-means*) (KANNAN e GHOSH, 2011); Análise de Componentes Principais (*Principal Component Analysis*, *PCA*) (HUTH, 1996; HUTH, 2000); Análogos (VAN DEN DOOL, 1989; VON STORCH e ZWIERS, 2003; TIMBAL e MCAVANEY, 2001); Modelo *Hiden* Markov (BELLONE, HUGHES e GUTTORP, 2000; ROBERTSON, KIRSHNER e SMYTH, 2004) e Mapas Auto-Organizados (HEWITSON e CRANE, 2002; SAUTER e VENEMA, 2011).

Os Geradores Estocásticos no Tempo são modelos que tem como objetivo replicar características estatísticas da variável de interesse na microescala (e.g. média e desvio padrão). Contudo, são incapazes de reproduzir a sequência de eventos observados (i.e. séries temporais observacionais da variável de interesse) (WILKS, 2019).

Os modelos regressivos relacionam as variáveis macroescalares (saídas do GCM) e o comportamento da variável de interesse na microescala (observações) através de funções de transferência lineares ou não lineares. Dentre os principais métodos regressivos adotados, temos a Regressão Linear Múltipla (*Multiple Linear Regression*, *MLR*) (VAN DER KAMP,

CURRY e MONAHAN, 2012) e as Redes Neurais Artificiais (*Artificial Neural Networks, ANN*) (SAILOR, HU, *et al.*, 2000).

Neste trabalho iremos nos concentrar na utilização de Métodos Regressivos de *downscaling* com o intuito de realizar o downsclaing da magnitude da velocidade do vento horizontal.

2.3 Regressão Linear Múltipla

A Regressão Linear Múltipla é a forma mais simples e mais geral de relacionar variáveis através de um método regressivo linear (WILKS, 2019). Assim como em uma regressão linear simples, existe apenas um preditando (e.g., dados observacionais na microescala), contudo, a regressão linear múltipla relaciona o preditando a mais de uma variável regressora (e.g., variáveis de saída dos Modelos de Circulação Geral Atmosférica).

A Regressão Linear Múltipla pode ser descrita través da Equação 1:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (1)$$

Onde y_i são os dados que compõem a série temporal do preditando; x_{ij} são os dados que compõem a série temporal da variável regressora *j*; β_j é o parâmetro da regressão associado à variável x_j e ε_i são os erros associados ao modelo.

A forma mais usual de estimar os parâmetros da regressão linear - e que também foi adotada neste trabalho - é utilizando o método dos mínimos quadrados, que tem por objetivo a estimativa dos parâmetros com vistas à minimização do erro médio quadrático (*mean squared error, MSE*), descrito na Equação 2:

$$MSE = \sum_{i=1}^{T} \varepsilon_i^2 \quad (2)$$

Onde ε são os erros entre a série estimada pelo modelo de regressão linear múltipla e o preditando.

3 REVISÃO DE LITERATURA

O rápido crescimento das energias renováveis no mundo (REN21, 2020), com ênfase nas fontes solares e eólicas, vem demandando uma descrição cada vez mais precisa do comportamento do vento e da radiação solar nos possíveis locais de instalação de centrais, seja com o intuito de caracterizar climatologicamente a variável no local de interesse, ou com o intuito de prever o recurso disponível, auxiliando na tomada de decisão dos operadores de usinas e do sistema elétrico.

Com relação a essa discussão, as técnicas objetivas (i.e. automáticas) de *downscaling* estatístico desempenham um papel fundamental. Como visto no capítulo de conceitos preliminares, essas técnicas baseiam-se na minimização de uma função de custo (WILKS, 2019), possibilitando que o modelo extraia informações da macroescala (e.g., saídas do GCM) para descrever o comportamento da variável de interesse na microescala (WILBY e WIGLEY, 1997; ORLANSKY, 1975). Nesse sentido, muitos estudos que utilizam métodos regressivos de *downscaling* adotam subjetivamente como variáveis regressoras as variáveis macroescalares localizadas nos pontos de malha do GCM mais próximos ao local de interesse (e.g., HEWITSON e CRANE, 1996; CURRY *et al.*, 2012; DUPRÉ *et al*., 2020).

Em muitos casos, no entanto, variáveis macroescalares que se encontram distantes do local de interesse apresentam grande capacidade de representar o comportamento do vento nesse local (e.g., HOFER *et al*., 2010; SAUTER e VENEMA, 2011; HORTON *et al.*, 2012; GUO *et al.,* 2012; CURRY *et al.*, 2012). Nesse contexto, se faz necessária a utilização de um conjunto mais amplo de variáveis regressoras, de modo que as variáveis mais distantes, geralmente desprezadas, possam ser utilizadas, conseguindo, assim, uma melhor descrição do comportamento da variável de interesse na microescala.

Por outro lado, quando apenas aumentamos a área de variáveis macroescalares a serem utilizadas, não levando em consideração a contribuição individual de cada uma delas, provavelmente estamos inserindo no conjunto de variáveis regressoras elementos com baixa capacidade de representar o comportamento da variável de interesse no local de estudo. Portanto, torna-se necessário selecionar as variáveis do domínio que serão utilizadas, uma vez que grande parte dos métodos regressivos comumente empregados são incapazes de selecionar as variáveis regressoras que promovem a maior acurácia do *downscaling*.

A seleção de variáveis regressoras pode se dar de duas formas distintas: subjetiva e objetiva. A seleção subjetiva é aquela em que as variáveis regressoras são selecionadas subjetivamente a partir do conhecimento empírico acerca do local de interesse, e por isso seu

uso se torna bastante restrito. Já a seleção objetiva de variáveis regressoras (i.e. automática) não demanda conhecimento empírico sobre o local de estudo, e, por isso, pode ser largamente empregada, pois a aplicação da metodologia já é autossuficiente, não necessitando de intervenções.

Os métodos de seleção objetiva de variáveis regressoras, por sua vez, podem ser divididos em duas abordagens principais: classificação/compressão e seleção direta. Com o intuito de facilitar a compreensão das classificações de métodos que será realizada a seguir, indicamos a visualização da Figura **2**.

A abordagem de classificação/compressão é aquela em que o conjunto original de variáveis regressoras sofre uma compressão para reduzir o número de elementos, com vistas a preservar alguma característica desejada. A respeito desse subconjunto de técnicas, podemos citar a Análise de Componentes Principais (JOLLIFFE, 1986), que leva em consideração apenas a variância de todo o domínio de busca[2], a fim de condensar a maior parte dessa variância em um menor número de variáveis (e.g., WETTERHALL *et al.*, 2004; VAN DER KAMP *et al.*, 2012); e a Análise de Correlação Canônica, que seleciona a informação relevante do domínio de busca levando em consideração apenas a correlação (e.g. BARNETT e PREISENDORFER, 1987; VON STORCH *et al.*, 1993; LANDMAN *et al.*, 2001; HUTH, 2004).

A abordagem de seleção direta contempla técnicas em que as variáveis são selecionadas a partir do domínio de busca. No entanto, essas variáveis selecionadas não passam por nenhum tipo de transformação – são mantidas originais. Ao contrário das técnicas presentes na abordagem de classificação/compressão, que são bem difundidas e amplamente utilizadas, as técnicas presentes na abordagem de seleção direta contam com um menor número de publicações.

É possível, ainda, subdividir o conjunto de técnicas de seleção direta em dois subconjuntos principais. São eles: seleção direta regional e seleção direta individual.

As técnicas de seleção direta regional são aquelas em que grupos do domínio de busca são avaliados como possíveis variáveis regressoras, levando-se em consideração características de todo o grupo em detrimento das características pontuais de cada indivíduo. Possivelmente, a forma mais simples de se realizar esse tipo de seleção de variáveis regressoras é testando a acurácia do modelo regressivo para diferentes áreas do domínio, seja escolhendo essas áreas de forma subjetiva (e.g., TIMBAL e MCAVANEY, 2001; D'ONOFRIO *et al.*, 2010), ou

[2] O domínio de busca é o conjunto de variáveis utilizadas como entrada para o método de seleção de variáveis regressoras.

escolhendo-as de forma objetiva (e.g., RADANOVICS *et al.*, 2013). Outras maneiras viáveis de realizar essa filtragem são: selecionando como variáveis regressoras aquelas pertencentes a grupos que possuem a média da correlação entre as variáveis do domínio de busca e a observação acima de um determinado limite estabelecido subjetivamente (e.g., HOFER *et al.*, 2010); ou selecionando como variáveis regressoras as variáveis do domínio de busca dentro de um perímetro ao redor do local estudado, sendo esse perímetro formado por variáveis do domínio que possuem correlação com os dados observacionais acima de um limite estabelecido subjetivamente (GUO *et al.*, 2012).

Em geral, as técnicas de seleção regional estão associadas à definição de limites subjetivos. Além disso, o fato de selecionar áreas do domínio de busca pode ocasionar a escolha por áreas muito restritas ao redor do local de interesse, desprezando variáveis distantes que poderiam contribuir para uma melhora na acurácia do *downscaling*. Por esses motivos, a seleção de variáveis regressoras através de técnicas de seleção direta regional contempla um conjunto restrito de aplicações.

As técnicas de seleção direta individual são aquelas em que as características das variáveis do domínio de busca são analisadas individualmente. Em muitos casos, a seleção de variáveis regressoras fazendo uso desse conjunto de técnicas é encontrado na literatura como *Screening Regression* (WILKS, 2019).

A maior parte dos artigos que fazem uso de técnicas de seleção direta individual (*Screening Regression*) concentram sua atenção na escolha da natureza da variável do GCM a ser utilizada (e.g., escolha entre temperatura, velocidade do vento e pressão atmosférica). Entretanto, não realizam seleção dos pontos da malha do GCM a serem utilizados como variáveis regressoras (e.g., ENKE *et al.*, 2005; RANABOLDO *et al.*, 2013; BAGHANAM *et al.*, 2019), desprezando a potencial melhora na acurácia do *downscaling* devido à seleção do melhor conjunto de pontos de malha do GCM.

A procura individual por variáveis que maximizem a acurácia do *downscaling* estatístico pode se dar de duas formas: através de métodos *embedded* e métodos *wrappers* (GUYON e ELISSEEFF, 2003).

Os métodos *embedded* realizam a seleção de variáveis regressoras ao longo do processo de treinamento (calibração) dos parâmetros do modelo regressivo. Por isso, cada método *embedded* diz respeito a um modelo de regressão, não podendo ser aplicado a outros modelos (GARCÍA-HINDE *et al.*, 2016). Podemos citar como métodos *embedded* mais utilizados o *Stepwise Regression* (SWR) (e.g., HESSAMI *et al.*, 2008; YANG *et al.*, 2017) e a Regressão

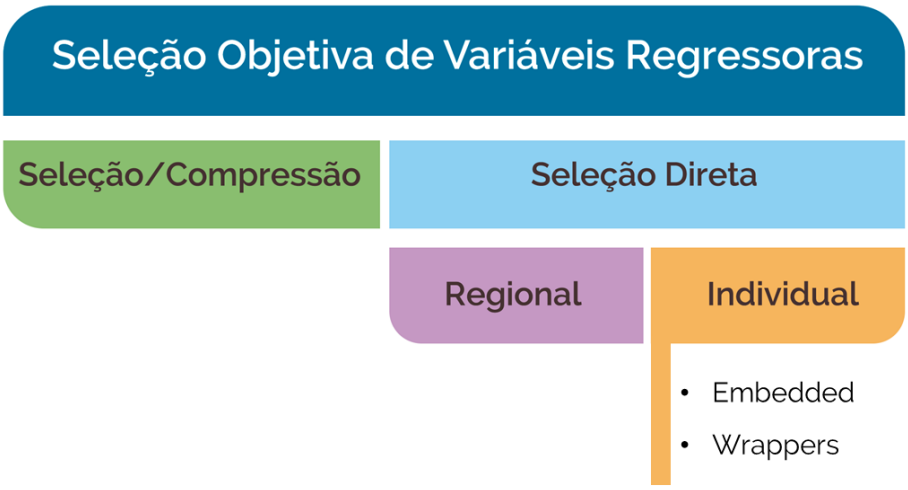
Lasso (TIBSHIRANI, 1996) (e.g., HAMMAMI *et al.*, 2012; GAO *et al.*, 2014), ambos referentes à Regressão Linear Múltipla.

Os métodos *wrappers*, por outro lado, enxergam a tarefa de selecionar o melhor conjunto de variáveis regressoras como um problema np-completo (KOHAVI e JOHN, 1997; GUYON e ELISSEEFF, 2003) e utilizam o modelo regressivo como uma caixa preta com o intuito de ordenar subconjuntos de variáveis de acordo com a acurácia por eles obtido. Por esse motivo, os modelos *wrappers* são mais gerais e podem ser aplicados em associação com qualquer modelo regressivo. É importante ressaltar que apenas um pequeno número de artigos publicados faz uso de métodos *wrappers* para a seleção dos melhores pontos de malha do GCM. Como principal representante temos Sauter e Venema (2011) que fazem uso do método de Recozimento Simulado (KIRKPATRICK *et al.*, 1983).

As técnicas de seleção individual utilizando métodos *wrappers*, como Recozimento Simulado ou *Forward Selection* (RANABOLDO *et al.*, 2013) demandam um grande esforço computacional, o que torna inviável sua aplicação para grandes domínios de busca. Além disso, a utilização de Recozimento Simulado necessita da definição de alguns parâmetros que podem variar de acordo com o local estudado e a variável atmosférica de interesse, reduzindo sua gama de aplicações.

Um breve resumo da classificação das metodologias de seleção objetiva de variáveis regressoras realizada neste trabalho é mostrado na Figura **2.**

Figura 2 – Classificação dos tipos de seleção objetiva de variáveis regressoras



Fonte: O autor (2020).

Na Tabela 1 encontra-se um resumo das principais referências bibliográficas relacionadas com o tema abordado neste trabalho.

Tabela 1 – Resumo das principais referências bibliográficas

Artigo	Variável estudada	Modelo regressivo	Tipo de seleção	Objetividade
Timbal e McAvaney, 2001	Temperatura	Análogos	Regional	Subjetivo
Hessami et al., 2008	Temperatura e precipitação	MLR	Individual (*embedded*)	Objetivo
Hofer et al., 2010	Temperatura e Umidade	PCA + MLR	Regional	Semi objetivo
D'Onofrio et al., 2010	Precipitação	Seleção aleatória de elementos do agrupamento	Regional	Subjetivo
Sauter e Venema, 2011	Precipitação	Redes Neurais Artificiais	Individual (*wrapper*)	Semi objetivo
Horton et al., 2012	Precipitação	Análogos	Regional	
Guo et al., 2012	Precipitação	Regressão escalonada baseada na correlação	Regional	Semi objetivo
Hammami et al., 2012	Temperatura	MLR	Individual (*embedded*)	Objetivo
Radanovics et al., 2013	Precipitação	Análogos	Regional	Objetivo
Gao et al., 2014	Precipitação	MLR	Individual (*embedded*)	Objetivo
Yang et al., 2017	Temperatura e precipitação	MLR	Individual (*embedded*)	Objetivo
Teegavarapu e Goly, 2018	Precipitação	MLR	Individual (*embedded*)	Objetivo
Dantas et al., 2020	**Velocidade do vento**	**MLR**	**Individual (*wrapper*)**	**Objetivo**

Fonte: O autor (2020).

Considerando tudo o que foi exposto anteriormente, é fundamental a utilização de um método que selecione de forma direta e individual quais variáveis do domínio de busca, ao serem utilizadas como variáveis regressoras, otimizam o resultado do *downscaling* estatístico. Isso porque os métodos regionais podem não ser efetivos na seleção de variáveis do domínio de busca distantes do local de interesse e geralmente necessitam da definição subjetiva de limites **(**TIMBAL e MCAVANEY, 2001; D'ONOFRIO *et al.*, 2010; HOFER *et al.*, 2010; GUO *et al.*, 2012; RADANOVICS *et al.*, 2013).

Outra caraterística que deve ser levada em consideração é a possibilidade de utilizar o método de seleção de variáveis regressoras desenvolvido neste trabalho em associação com qualquer tipo de modelo regressivo. Dessa forma, o modelo pode ser empregado em um maior número de aplicações. No entanto, os métodos de seleção que atendem a esse requisito (i.e.,

métodos *wrappers*), em geral, demandam um alto esforço computacional (SAUTER e VENEMA, 2011). Por esse motivo, o método proposto deve demandar um baixo esforço computacional, viabilizando sua utilização em aplicações operacionais (em tempo real).

Nesse contexto, a metodologia proposta neste trabalho, denominada Seleção Baseada no Ordenamento (SOB), se enquadra nos métodos *wrappers* e tem como objetivo ordenar as variáveis do domínio com respeito à sua relevância, que será dada pelas funções de ordenamento e, em seguida, indicar quais dessas variáveis, quando utilizadas como variáveis regressoras, produzem o melhor resultado.

A variável atmosférica a ser estudada ao longo do trabalho será a magnitude da velocidade do vento horizontal com vistas a aplicações relacionadas à área de energia eólica. Como principais exemplos dessas aplicações, temos: a reanálise com vistas à tomada de decisão sobre implantação de centrais eólicas, e a previsão com vistas ao despacho da produção de centrais eólicas.

Contudo, é importante frisar que essa metodologia pode ser aplicada a qualquer variável atmosférica, abrangendo um número ainda maior de aplicações.

3.1 Objetivo do trabalho

Propor uma nova metodologia de seleção de variáveis regressoras para o *downscaling* de variáveis atmosféricas (particularmente, neste trabalho, aplicada à magnitude do vento horizontal), aqui chamada Seleção Baseada no Ordenamento (SBO), que apresente as seguintes características:

- Possibilita uma seleção objetiva, direta e individual no domínio de busca;
- Possibilita a utilização de qualquer tipo de modelo regressivo;
  Apresenta baixo esforço computacional.

Além disso, este trabalho se propõe a:

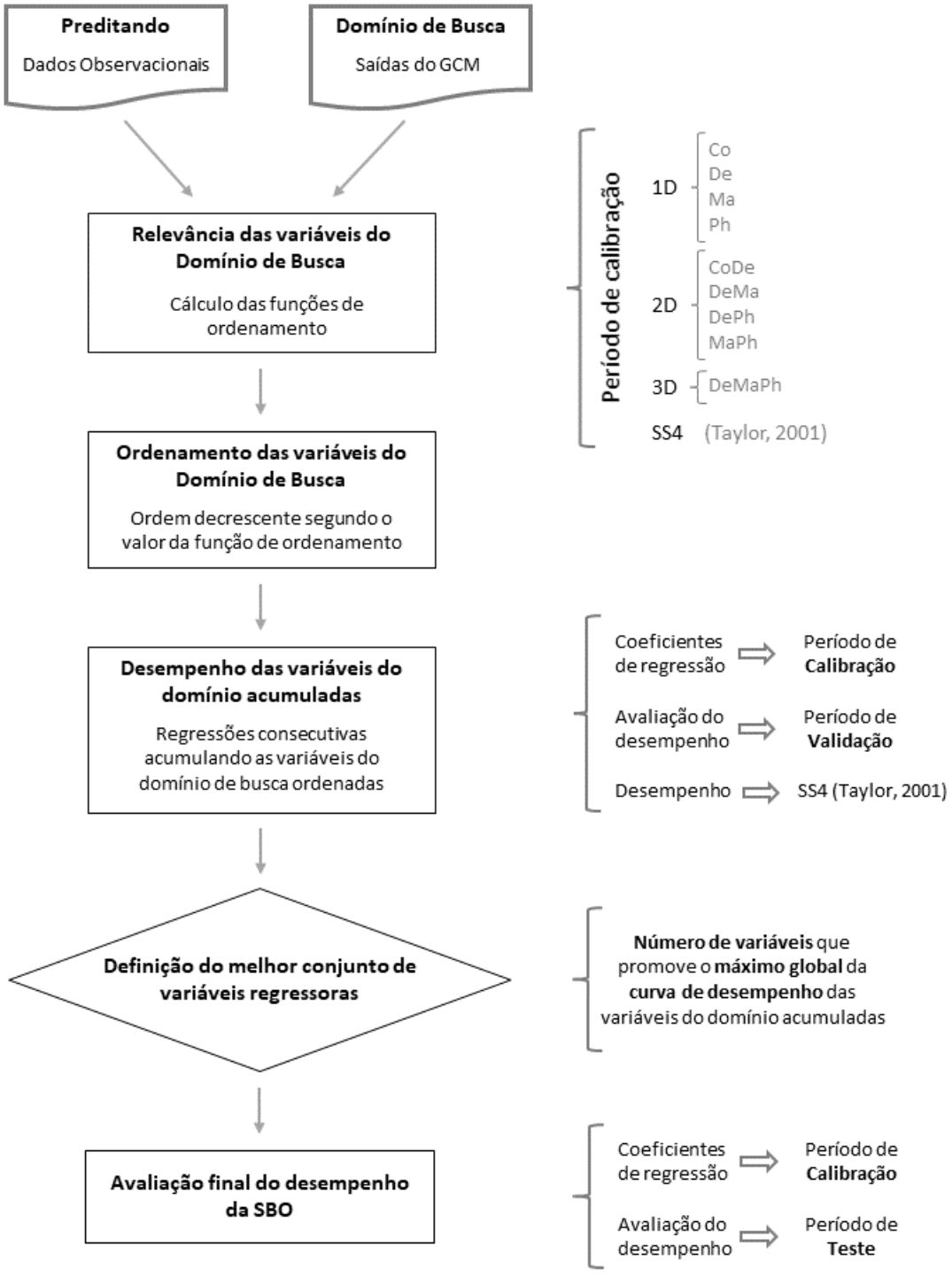
- Desenvolver novas métricas para avaliação da relevância das variáveis do domínio de busca;
- Realizar uma comparação entre a metodologia proposta (SBO) e as metodologias consagradas na literatura;
- Identificar a melhor maneira de dividir o conjunto de dados com vistas à realização da validação da metodologia.

4 MATERIAL E MÉTODOS

Com o objetivo de proporcionar uma melhor compreensão, a metodologia proposta neste trabalho será dividida em quatro seções. Na primeira, será abordado o cálculo das funções de ordenamento, que relacionam cada variável do domínio de busca com o preditor (dados observacionais) e descrevem o grau de relevância de cada uma dessas variáveis. Na segunda, será discutido como determinar quais variáveis do domínio de busca, ao serem utilizadas como variáveis regressoras, produzem o melhor resultado. Na terceira, será abordado o procedimento de validação cruzada utilizado. Na quarta, será abordado quais modelos serão utilizados como modelos de referência com os quais serão comparados os resultados obtidos pela metodologia desenvolvida neste trabalho. A metodologia deste trabalho está mostrada no diagrama da Figura **3**.

Vale destacar que, adicionalmente aos temas citados acima, avaliou-se a utilização da análise de componentes principais com vistas a reduzir o grau de colinearidade entre as variáveis que compõem o domínio de busca. O teste realizado, apresentado no Apêndice B, não apresentou resultados relevantes, conduzindo à seguinte conclusão: O uso de componentes principais é bastante sensível a variações desiguais ao longo das variáveis do domínio de busca, por esse motivo, a utilização de componentes principais como entrada no método de seleção de variáveis regressoras proposta neste trabalho é bastante limitada.

Figura 3 – Fluxograma resumindo a metodologia proposta neste trabalho



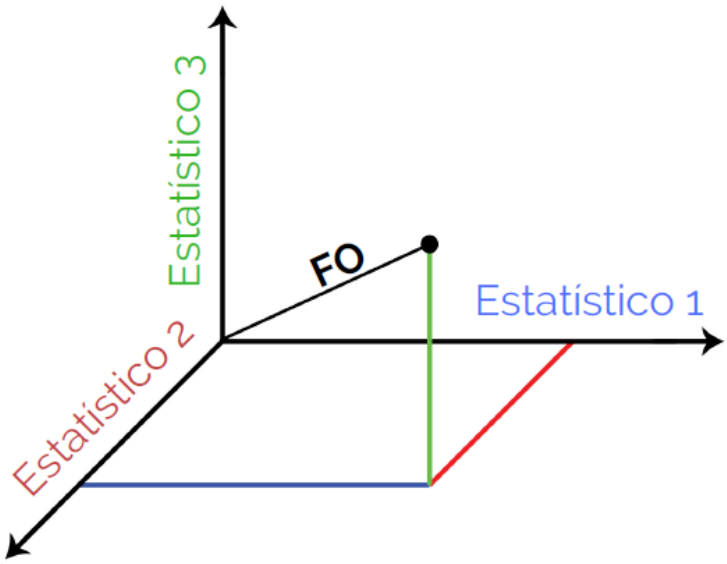
Fonte: O autor (2020).

4.1 Cálculo das funções de ordenamento

Neste trabalho, propõem-se dez diferentes funções de ordenamento, definindo-se a função de ordenamento por meio da norma de um vetor pertencente a um espaço de n dimensões em que a projeção sobre cada eixo coordenado é dada por uma estatística que relaciona a variável do domínio de busca e o preditando (dados observacionais) (ver Figura 4). As estatísticas que compõem os eixos coordenados são dimensionadas de tal forma a possuírem valor mínimo de zero e aumentam à medida em que a variável do domínio de busca melhor representa o preditando, atingindo valor máximo de um. Isso faz com que todas as estatísticas que constituem os eixos coordenados possuam a mesma relevância na norma do vetor que define a função de ordenamento.

É importante salientar que, com vistas a simplificar a análise a ser realizada, uma vez que se trata de uma primeira abordagem ao tema, os estatísticos que compõem as funções de ordenamento não levam em consideração características importantes dos sinais, como, por exemplo, a não estacionariedade e a similaridade através de relações não lineares.

Figura 4 - Demonstração de um vetor que define uma função de ordenamento



Demonstração de um vetor que define uma função de ordenamento composta pelos *Estatístico 1*, *Estatístico 2* e *Estatístico 3*.

Fonte: O autor (2020).

Neste trabalho, foram propostas dez funções de ordenamento diferentes: quatro unidimensionais, quatro bidimensionais, uma tridimensional e o Skill Score definido em (TAYLOR, 2001). Essas funções serão explicadas nos itens abaixo.

4.1.1 Funções de ordenamento unidimensionais

As funções de ordenamento propostas neste trabalho têm como objetivo determinar a relevância das variáveis do domínio de busca. Nesse sentido, foram desenvolvidas funções de ordenamento que identificam características desejáveis do preditando nas variáveis do domínio de busca.

A primeira característica a ser identificada é a similaridade entre os sinais no domínio da frequência. Por isso, foi utilizado o valor absoluto do coeficiente de correlação (WILKS, 2019) como função de ordenamento (Co) (ver Equação 3).

No entanto, a caracterização de uma série temporal no domínio da frequência se dá através da magnitude e da fase da transformada da série. Por isso, foram criadas duas funções de ordenamento distintas para a magnitude e fase. A função de ordenamento que diz respeito à magnitude é baseada na raiz do erro médio quadrático entre as magnitudes da transformada de Fourier dos sinais do preditando e da variável do domínio de busca (Ma) (ver Equação 7). A função de ordenamento que diz respeito à fase é baseada na raiz do erro médio quadrático entre as fases da transformada de Fourier dos sinais do preditando e da variável do domínio de busca (Ph) (ver Equação 8).

A última função de ordenamento unidimensional tem o intuito de identificar a capacidade de representar a variabilidade do preditando. Por isso, foi desenvolvida uma função de ordenamento baseada na razão do desvio padrão (De) (ver Equação 4).

A modificação aplicada aos estatísticos para a criação do De, Ma e Ph é basicamente a normalização do estatístico pelo seu valor máximo no domínio de busca e a adequação para que a função de ordenamento obedeça a uma ordem decrescente com valor máximo de um e valor mínimo de zero.

$$Co = \frac{cov(d,p)}{\sigma_d \sigma_p} \tag{3}$$

$$De = 1 - \left[\frac{\left| 1 - \left(\frac{\sigma_d}{\sigma_p} \right) \right|}{\max \left| 1 - \left(\frac{\sigma_d}{\sigma_p} \right) \right|_{Domínio}} \right] \tag{4}$$

$$RMSE_{Magnitude} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(|D_t| - |P_t|)^2} \quad (5)$$

$$RMSE_{Fase} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\sin(\theta_{D_t}) - \sin(\theta_{P_t})\right)^2} + \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\cos(\theta_{D_t}) - \cos(\theta_{P_t})\right)^2} \quad (6)$$

$$Ma = 1 - \frac{RMSE_{Magnitude}}{\max(RMSE_{Magnitude})_{Domínio}} \quad (7)$$

$$Ph = 1 - \frac{RMSE_{Fase}}{\max(RMSE_{Fase})_{Domínio}} \quad (8)$$

Onde: d significa variável do domínio de busca; p significa preditando; σ_d significa o desvio padrão de d; σ_p significa o desvio padrão de p; D é a transformada de Fourier do sinal de d; P é a transformada de Fourier do sinal de p; θ_D é o ângulo de fase de D e θ_P é o ângulo de fase de P.

4.1.2 Funções de ordenamento bidimensionais

As quatro funções de ordenamento bidimensionais utilizadas neste estudo (ver Equações 9 a 12) são todas constituídos pelas funções de ordenamento unidimensionais mostradas anteriormente. Contudo, é importante notar que, com a intenção de evitar redundâncias nas funções de ordenamento, uma vez que a correlação está intrinsecamente relacionada à estrutura de fase e frequência dos sinais, nenhuma função de ordenamento foi composta simultaneamente por Co e Ma ou Co e Ph.

Para facilitar a análise e compreensão dos resultados, as expressões das funções de ordenamento bidimensionais e tridimensionais foram multiplicadas por constantes, de modo que o valor máximo das funções de ordenamento seja igual a um.

$$CoDe = \left(\frac{1}{\sqrt{2}}\right)\sqrt{Co^2 + De^2} \quad (9)$$

$$DeMa = \left(\frac{1}{\sqrt{2}}\right)\sqrt{De^2 + Ma^2} \quad (10)$$

$$DePh = \left(\frac{1}{\sqrt{2}}\right)\sqrt{De^2 + Ph^2} \quad (11)$$

$$MaPh = \left(\frac{1}{\sqrt{2}}\right)\sqrt{Ma^2 + Ph^2} \quad (12)$$

4.1.3 Funções de ordenamento tridimensionais

A função de ordenamento tridimensional utilizada neste estudo é constituída por De, Ma e Ph (ver Equação 13).

$$DeMaPh = \left(\frac{1}{\sqrt{3}}\right)\sqrt{De^2 + Ma^2 + Ph^2} \quad (13)$$

4.1.4 Função de ordenamento extra

Para este estudo, também foi utilizado o *skill-score* descrito em Taylor (2001) (SS4) como uma função de ordenamento (ver Equação 14). Assim como as funções de ordenamento mostradas anteriormente, o *SS4* varia de zero a um, ou seja, da situação de menor representatividade até a situação de maior representatividade, e possui os seguintes atributos: "para qualquer variância modelada dada, o escore aumenta monotonicamente com o aumento da correlação (entre a saída do modelo e a observação) e, para qualquer correlação dada, o escore aumenta à medida em que a variância modelada se aproxima da variância observada" (TAYLOR, 2001). O valor do SS4 é definido por:

$$SS4 = \frac{\left(1 + \frac{cov(d,p)}{\sigma_d \sigma_p}\right)^4}{4\left(\left(\frac{\sigma_d}{\sigma_p}\right) + \left(\frac{\sigma_p}{\sigma_d}\right)\right)^2} \quad (14)$$

É importante salientar que estatísticos como RMSE e BIAS não foram utilizados diretamente nos sinais originais (sem qualquer tipo de transformação) com o intuito de compor as funções de ordenamento. Isso porque, no presente estudo, estamos mais interessados em

caracterizar a estrutura de fase e frequência, bem como o desvio padrão do preditando, uma vez que a correção dessas características em processos posteriores pode se dar de maneira muito mais difícil que a correção do BIAS.

4.2 Encontrando o melhor conjunto de variáveis regressoras

Concluída a fase anterior, as variáveis do domínio de busca são dispostas de acordo com o valor de cada uma das as funções de ordenamento (definidos na fase anterior) de forma decrescente, de modo que a primeira variável será aquela que possui o mais alto valor da função de ordenamento e a última será aquela que possui o mais baixo valor da função de ordenamento. Posteriormente, são realizadas regressões consecutivas, começando exclusivamente com a primeira variável do domínio de busca – a que possui a função de ordenamento mais alta – e, em cada passo, a próxima variável do domínio é adicionada (acumulada), seguindo a ordem acima mencionada, dando origem ao que se chama aqui de conjunto de variáveis regressoras acumuladas. De posse do resultado do desempenho de tal conjunto, torna-se possível decidir qual é o melhor número de variáveis do domínio de busca ordenadas segundo uma determinada função de ordenamento, caracterizando o melhor conjunto de variáveis regressoras. Neste trabalho, o resultado mais satisfatório é aquele que apresenta o mais alto valor do *skill score* definido por Taylor (2001), i.e., o SS4 (Equação 14) é aqui empregado como função de ordenamento bem como métrica para avaliação e intercomparação do desempenho dos modelos.

Como as variáveis de saída do GCM utilizadas neste trabalho estão dispostas em dez níveis verticais, foram realizadas seleções de variáveis regressoras distintas para cada um dos níveis verticais de modelo. Dessa forma, o melhor resultado de um local é dado pelo melhor conjunto de variáveis regressoras do nível que apresentou melhor performance.

Neste estudo, foi utilizada a regressão linear múltipla (MLR) como modelo regressivo devido à sua simplicidade e ao baixo esforço computacional requerido (VON STORCH *et al.*, 2003; WILKS, 2019) o que facilita as análises que serão efetuadas posteriormente. Além disso, esse modelo é largamente aplicado em *downscaling* estatístico de variáveis atmosféricas, obtendo resultados satisfatórios em diversos estudos (MURPHY, 1999; CURRY *et al.*, 2012).

No entanto, é importante enfatizar que qualquer outro modelo regressivo poderia ser empregado, incluindo o uso de diferentes modelos, nas etapas de encontrar o melhor conjunto de variáveis regressoras e realizar o *downscaling* com as variáveis já selecionadas.

4.3 Validação cruzada

A validação cruzada é um procedimento que objetiva a divisão das séries temporais utilizadas de tal forma que o período empregado para avaliar a performance do modelo (período de teste) seja diferente daqueles utilizados para estimar seus parâmetros (períodos de calibração e validação) (MICHAELSEN, 1987).

No presente trabalho, as séries temporais foram divididas em três períodos distintos: calibração, validação e teste. Os dados contidos no período de calibração são utilizados no cálculo das funções de ordenamento e no cálculo dos coeficientes da regressão sobre os conjuntos de variáveis regressoras acumuladas. Os coeficientes da regressão estimados utilizando o conjunto de calibração foram aplicados ao conjunto de validação de tal forma a intercomparar a performance das variáveis regressoras acumuladas e, com isso, selecionar o melhor conjunto de variáveis regressoras. Os dados contidos no período de teste foram utilizados apenas para avaliar o desempenho final da metodologia de seleção de variáveis regressoras, sem que nenhum parâmetro fosse estimado ao longo deste período.

Um resumo dos procedimentos executados em cada um dos períodos é mostrado na Tabela 2.

Tabela 2 – Procedimentos realizados em cada um dos períodos das séries temporais

Período	Procedimentos Realizados
Calibração	• Cálculo das funções de ordenamento • Cálculo dos parâmetros das regressões das variáveis regressoras acumuladas
Validação	• Avaliação da performance das regressões das variáveis regressoras acumuladas (escolha do melhor conjunto baseada no SS4)
Teste	• Avaliação da performance do método regressivo utilizando as variáveis regressoras selecionadas (melhor conjunto de variáveis regressoras)

Fonte: O autor (2020)

Devido à escassez de publicações relacionadas à comparação entre diferentes formas de dividir os períodos utilizados na validação cruzada para o *downscaling* estatístico, foi necessário avaliar diferentes abordagens para a divisão dos dados.

Esse teste não visa apenas avaliar a acurácia, mas também a robustez do modelo desenvolvido neste trabalho, isso é, a independência da acurácia do modelo com respeito ao período dos dados utilizados na calibração dos parâmetros.

Foram avaliadas sete maneiras distintas de realizar a validação cruzada. Em todas elas, o período de teste compreende o último terço dos dados da série temporal, pois dessa forma é possível ter um período comum de comparação entre todos os tipos de validação cruzada. A distribuição dos períodos de calibração e validação são mostrados na e esquemas exemplificando os tipos de validação cruzada são mostrados nas Figuras 5 a 11. Contudo, em alguns tipos de validação cruzada, existem intervalos concomitantes entre os períodos de calibração e validação (ver Figuras 6, 9, 10 e 11). Essa escolha de períodos foi intencional para que fosse possível compreender se a utilização de períodos mais extensos acarretaria numa melhora substancial da performance do *downscaling*.

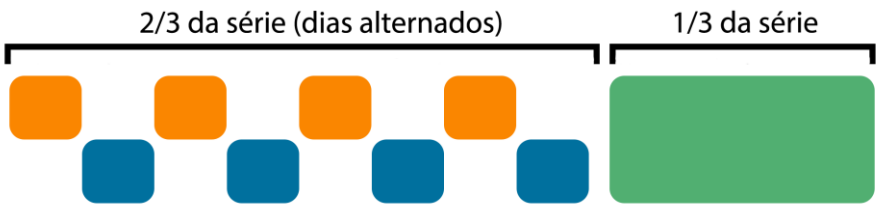
Tabela 3 – Períodos das séries temporais empregados na calibração e validação

Esquema de validação cruzada	**Calibração**	**Validação**
Tipo 1	Dados contidos nos dias ímpares (dos primeiros 2/3 da série)	Dados contidos nos dias pares (dos primeiros 2/3 da série)
Tipo 2	Dados contidos nos dias ímpares (dos primeiros 2/3 da série)	Primeiros 2/3 da série
Tipo 3	Primeiro 1/3 da série	Segundo 1/3 da série
Tipo 4	Segundo 1/3 da série	Primeiro 1/3 da série
Tipo 5	Primeiro 1/3 da série	Primeiros 2/3 da série
Tipo 6	Segundo 1/3 da série	Primeiros 2/3 da série
Tipo 7	Primeiros 2/3 da série	Segundo 1/3 da série

Fonte: O autor (2020)

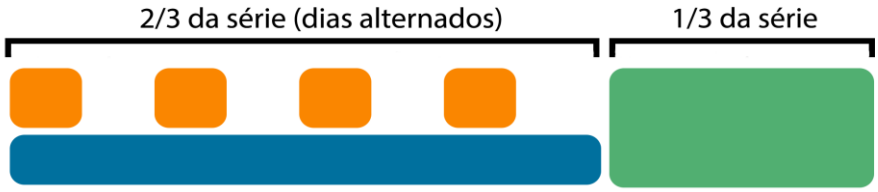
As Figuras 5 a 11 foram criadas com o intuito de tornar mais fácil a visualização das parcelas das séries temporais utilizadas nos conjuntos de calibração (em cor laranja), validação (em cor azul) e teste (em cor verde).

Figura 5 – Divisão dos períodos de calibração, validação e teste (Esquema 1)



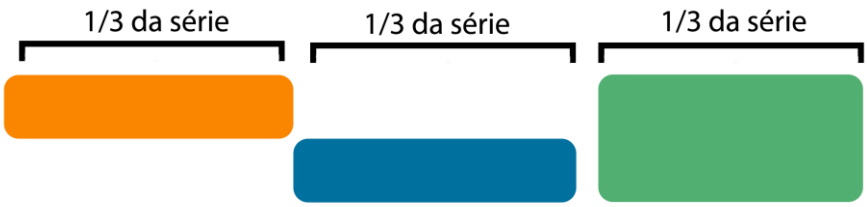
Fonte: O autor (2020).

Figura 6 - Divisão dos períodos de calibração, validação e teste (Esquema 2)



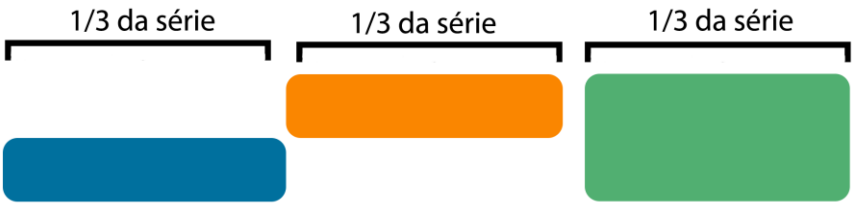
Fonte: O autor (2020).

Figura 7 - Divisão dos períodos de calibração, validação e teste (Esquema 3)



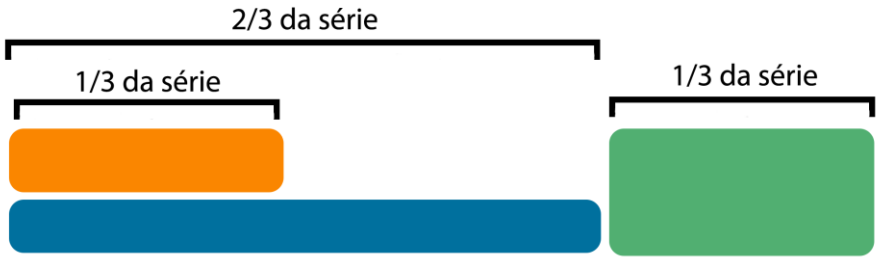
Fonte: O autor (2020).

Figura 8 - Divisão dos períodos de calibração, validação e teste (Esquema 4)



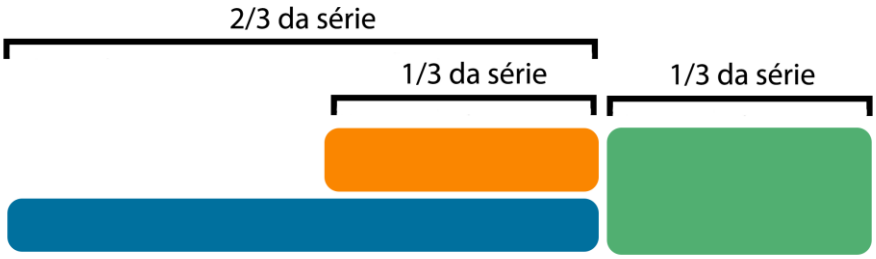
Fonte: O autor (2020).

Figura 9 - Divisão dos períodos de calibração, validação e teste (Esquema 5)



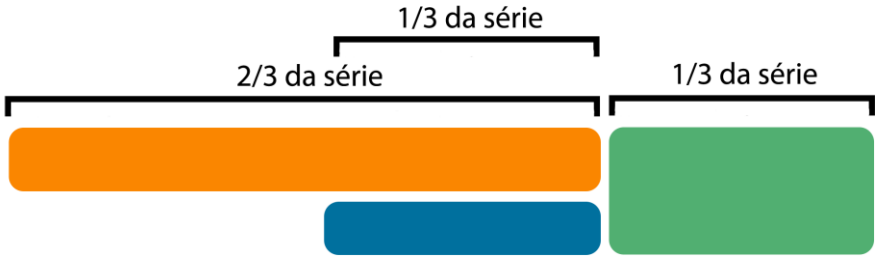
Fonte: O autor (2020).

Figura 10 - Divisão dos períodos de calibração, validação e teste (Esquema 6)



Fonte: O autor (2020).

Figura 11 - Divisão dos períodos de calibração, validação e teste (Esquema 7)



Fonte: O autor (2020).

4.4 Modelos de referência

Com vistas a compreender a melhora na acurácia do *downscaling* e a viabilidade de se utilizar o modelo de seleção objetiva de variáveis regressoras desenvolvido neste trabalho, se faz necessária a comparação dele com modelos de referência (modelos mais simples e/ou largamente utilizados).

O primeiro modelo de referência é a Interpolação Bilinear dos quatro pontos da malha horizontal do GCM mais próximos ao local de estudo (IBL4) (ACCADIA *et al.*, 2003). Esse modelo representa a estimativa do GCM sem que qualquer coeficiente seja estimado estatisticamente, pois os coeficientes da interpolação dependem exclusivamente da distância entre os pontos do GCM e a posição do local de estudo.

O segundo modelo de referência é a regressão linear múltipla dos quatro pontos da malha horizontal do GCM (WILKS, 2019). Esse modelo é largamente empregado e, por não realizar qualquer tipo de seleção de variáveis regressoras, ele representa o resultado mais simples que pode ser obtido pelo modelo regressivo.

Os próximos três modelos de referência são modelos estatísticos que fazem uso da regressão linear múltipla e realizam seleção do conjunto de variáveis regressoras. Esses modelos irão ter como entrada o mesmo domínio de busca aplicado à metodologia proposta neste trabalho. Dois desses modelos são *embedded*, isso é, apenas podem ser aplicados em conjunto com a regressão linear múltipla. São eles a *stepwise regression* (SWR) e a Regressão Lasso. Eles serão utilizados para verificar se a acurácia da metodologia proposta neste trabalho (que pode ser aplicada em conjunto com qualquer método regressivo) é equivalente ou superior à acurácia de métodos elaborados exclusivamente para a regressão linear múltipla.

O último modelo de referência, o *Forward Selection*, é um modelo *wrapper* objetivo. Esse modelo pode ser empregado em conjunto com qualquer modelo regressivo, mas faz uso de um grande esforço computacional. Ele será utilizado para compreender se a metodologia desenvolvida neste trabalho consegue obter desempenho equivalente ao obtido por métodos que demandam um grande esforço computacional.

Como forma de comparar diretamente a acurácia apresentada pelo modelo de seleção de variáveis regressoras com a acurácia apresentada pelos modelos de referência, foram utilizados como estatísticos a melhora sobre o IBL4 (ver Equação 15) e a melhora sobre o MLR4 (ver Equação 16).

$$Melhora\ sobre\ IBL4 = \frac{SS4_{Modelo} - SS4_{IBL4}}{SS4_{IBL4}} \tag{15}$$

$$Melhora\ sobre\ MLR4 = \frac{SS4_{Modelo} - SS4_{MLR4}}{SS4_{MLR4}} \tag{16}$$

Onde $SS4_{Modelo}$ é o $SS4$ entre a saída do modelo estudado (seja o modelo proposto neste trabalho ou algum dos modelos de referência) e a observação, $SS4_{IBL4}$ é o $SS4$ entre a saída do IBL4 e a observação e $SS4_{MLR4}$ é o $SS4$ entre a saída do MLR4 e a observação.

Outro estatístico de extrema importância na comparação dos resultados foi a razão entre os desvios padrão (STD_{ratio}) entre um modelo e a observação (ver Equação 17)

$$STD_{ratio} = \frac{\sigma_{Modelo}}{\sigma_{Oberservação}} \tag{17}$$

Onde σ_{Modelo} é o desvio padrão do modelo a ser analisado e $\sigma_{Oberservação}$ é o desvio padrão da série observacional.

5 RESULTADOS E DISCUSSÃO

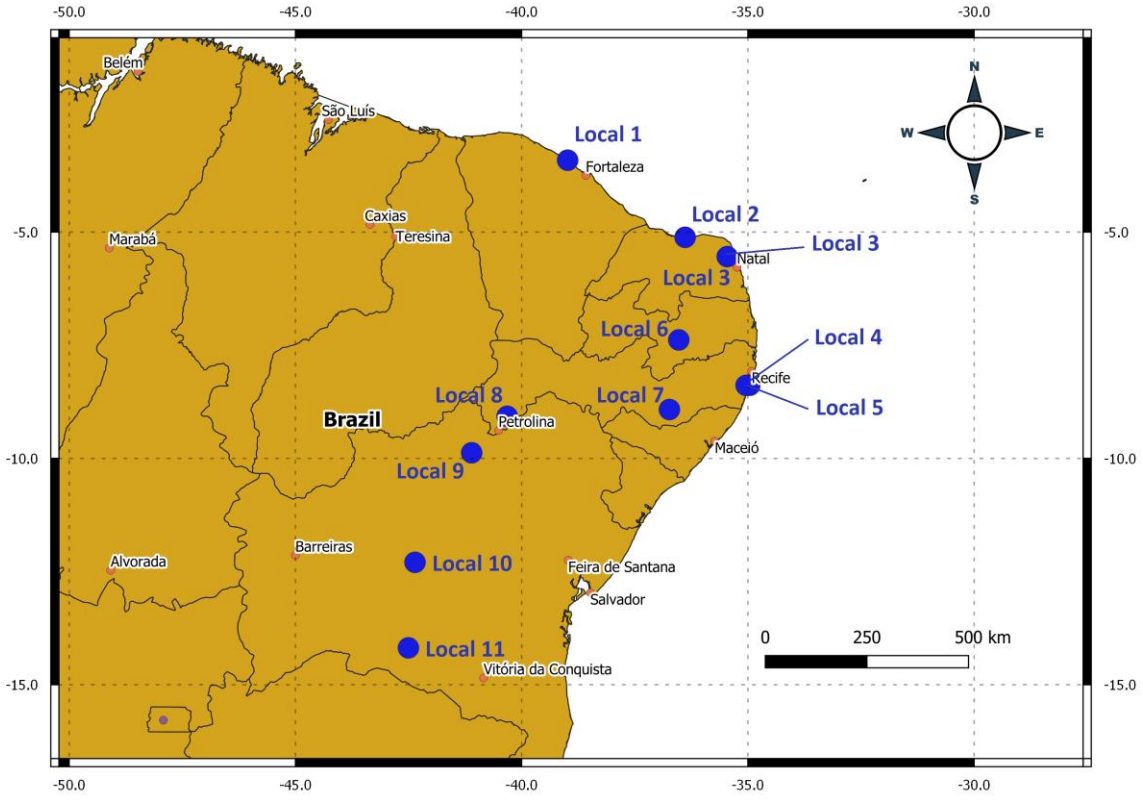
Com o intuito de avaliar a metodologia aqui proposta para a seleção de variáveis regressoras aplicada ao *downscaling* da magnitude da velocidade do vento local horizontal em superfície, foram realizados 11 estudos de caso, baseados em dados observacionais de 11 torres anemométricas localizadas na Região Nordeste do Brasil, região que conta, atualmente, com aproximadamente 85% da potência eólica instalada no país (EPE, 2020).

Com vistas a uma melhor compreensão, os estudos de caso são aqui apresentados em sete seções. Na primeira seção, são descritos os dados observacionais utilizados (i.e., preditando). Na segunda, são descritos os dados macroescalares utilizados (i.e., os dados numéricos do GCM – as variáveis do domínio de busca). Na terceira, são avaliados os diferentes esquemas de validação cruzada levando-se em consideração a acurácia dos modelos MLR4 e da metodologia proposta neste trabalho (cabe ressaltar que, ao final da terceira seção, decide-se sobre o melhor esquema de validação cruzada. A partir daí, as seções seguintes apenas apresentam resultados referentes ao melhor esquema de validação cruzada). Na quarta, discute-se sobre a distribuição espacial das funções de ordenamento. Na quinta, discute-se sobre o desempenho das variáveis regressoras acumuladas. Na sexta, são comparados os resultados obtidos através do uso das diferentes funções de ordenamento. Na sétima e última das seções do capítulo 5, são comparados os resultados obtidos pela metodologia proposta neste trabalho com aqueles obtidos pelos modelos de referência.

5.1 Dados observacionais

Os dados observacionais utilizados nos estudos de caso são provenientes de 11 torres anemométricas situadas na região Nordeste do Brasil, como pode ser visto no mapa da Figura **12**. Todas as torres fornecem dados de direção e magnitude do vento horizontal – obtidos a partir de birutas e anemômetros situados a uma altura da ordem da altura típica do cubo de um aerogerador de grande porte (da ordem de 100 metros sobre o nível do solo) – integrados em intervalos de 10 minutos. As fontes dos dados observacionais estão listadas na Tabela **4** e o intervalo de tempo no qual foram realizadas as campanhas de medição estão expostos na Figura **13**.

Figura 12 – Disposição geográfica dos locais estudados



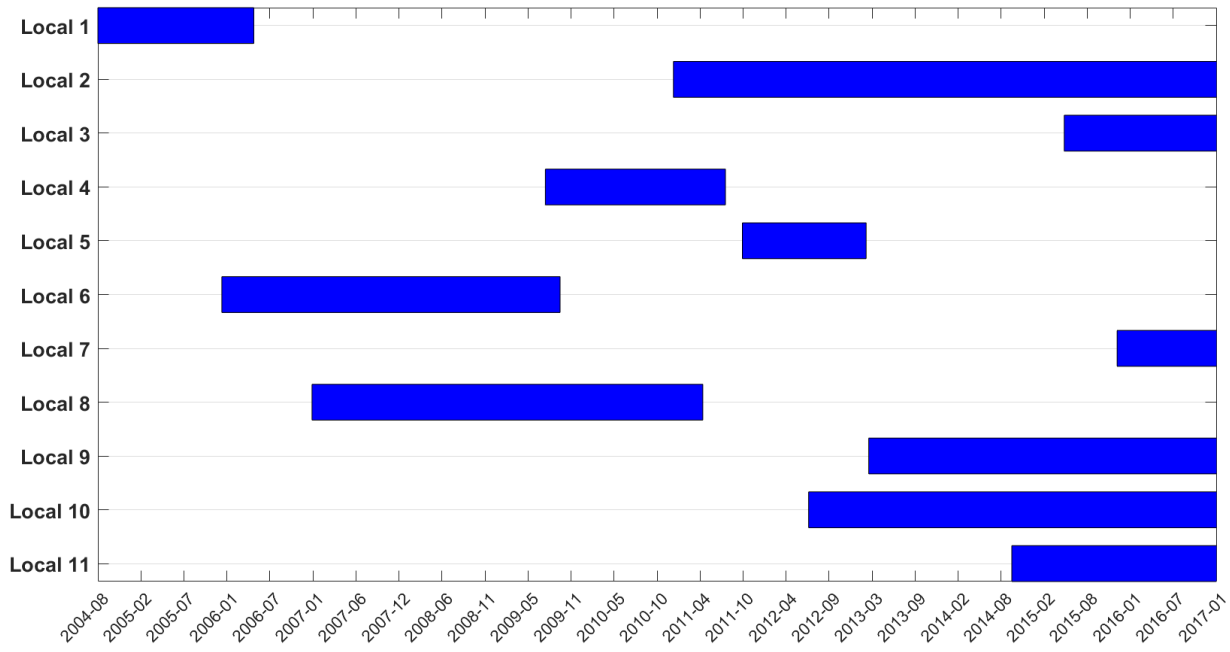
Fonte: O autor (2020).

Tabela 4 – Fontes dos dados observacionais (eólica)

Locais	Fonte dos Dados
1	SEINFRA-CE (https://www.seinfra.ce.gov.br/)
6 e 8	Projeto SONDA (http://sonda.ccst.inpe.br/)
4 e 5	Grupo de Mecânica dos Fluidos Ambiental no âmbito do Projeto Pilacas (https://www.ufpe.br/mecfluamb)
2, 3, 7, 9, 10 e 11	Operador Nacional do Sistema Elétrico (ONS) no âmbito do projeto HPC4E (https://hpc4e.eu/the-project/work-plan/wp4)

Fonte: O autor (2020).

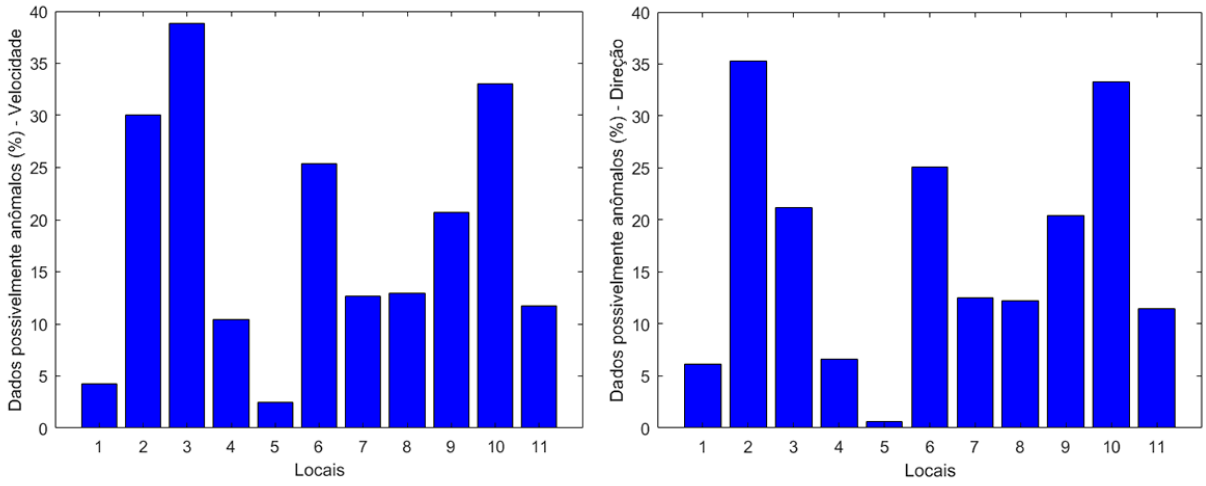
Figura 13 - Período da campanha de medição dos dados observacionais



Fonte: O autor (2020).

O procedimento de garantia de qualidade desenvolvido por Moraes (2015) foi empregado com o objetivo de identificar comportamentos suspeitos e anômalos ao longo das séries temporais de magnitude e direção do vento horizontal. Tal procedimento está baseado em uma sequência de testes globais (que avaliam a qualidade de uma série temporal como um todo) e testes locais (que avaliam a qualidade dos dados observacionais de forma individual ou em pequenos agrupamentos). O resultado de tal procedimento aplicado aos dados observacionais utilizados neste trabalho é mostrado no gráfico da Figura **14**. É importante salientar que a maior parte dos instantes em que os dados são classificados como anômalos diz respeito à ausência de dados observacionais nas séries originais.

Figura 14 – Quantidade de dados observacionais considerados possivelmente anômalos



Fonte: O autor (2020).

5.2 Dados macroescalares (GCM)

Como variáveis regressoras, foram utilizadas as magnitudes da velocidade do vento horizontal, compostas a partir das componentes zonal e meridional da velocidade do vento (ver Equação 17), oriundas do programa de reanálise ERA-Interim, disponibilizadas pelo European Centre for Medium-Range Weather Forecasts (ECMWF) (DEE *et al*, 2011; BERRISFORD *et al.*, 2011), com resolução espacial de 0,75° (da ordem de 80 km) e resolução temporal de 6 horas:

$$M = \sqrt{U^2 + V^2} \tag{17}$$

Onde M é a magnitude do vento horizontal, U é a componente zonal da velocidade do vento horizontal e V é a componente meridional. As componentes zonal e meridional do vento são uma representação da velocidade do vento horizontal em um plano cartesiano. Onde a componente zonal apresenta valores positivos no sentido oeste-leste e a componente meridional apresenta valores positivos no sentido sul-norte.

Os estudos e aplicações relacionados à área de eólica usualmente descrevem a velocidade do vento horizontal em um sistema de coordenadas polares, em que a direção zero graus descreve o vento proveniente da direção norte. E o ângulo que descreve a direção cresce no sentido anti-horário. Assim, este trabalho está dedicado à descrição da magnitude da velocidade do vento horizontal devido a sua importância para aplicações na área de eólica.

Para cada um dos locais de interesse, foi utilizado um domínio que compreende os dez níveis verticais de modelo mais próximos ao solo (níveis de modelo de 1012,05 hPa até 908,65

hPa, correspondendo a altura geopotencial de 10 metros até 760 metros acima do nível do solo, respectivamente) e 30x30 pontos da malha horizontal do GCM centrados na localização da torre anemométrica.

5.3 Avaliação dos diferentes esquemas de validação cruzada

Nesta seção, são comparadas as diferentes formas de dividir os dados com o intuito de realizar a validação cruzada. É importante salientar que, no caso do modelo MLR4 os esquemas de validação cruzada 1 e 2, 3 e 5, e 4 e 6 possuem, respectivamente, o mesmo resultado, pois o período de calibração foi o mesmo (i.e., o período de calibração empregado nos esquemas 1 e 2 é o mesmo. Isso vale para os esquemas 3 e 5, bem como para os esquemas 4 e 6) e nenhuma decisão foi tomada no período de validação, pois o modelo MLR4 não realiza seleção de variáveis regressoras.

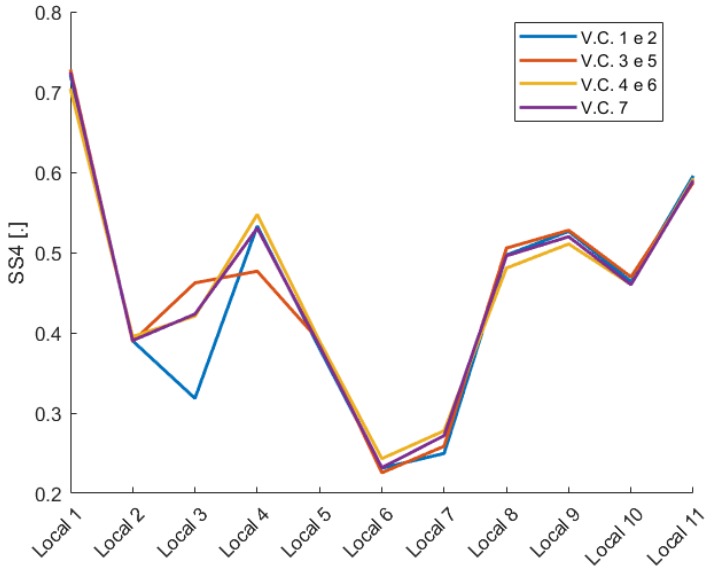
Foram utilizados apenas o MLR4 e a metodologia proposta neste trabalho para avaliar os diferentes esquemas de validação cruzada, isso porque, realizar a simulação fazendo uso de todos os modelos de referência para todos os esquemas de validação cruzada demandaria um alto esforço computacional. Além disso, o intuito dessa análise é comparar a acurácia da metodologia proposta neste trabalho nos diferentes esquemas de validação cruzada e comparar a sua robustez com aquela obtida por um modelo que não realiza tomada de decisão no período de validação (i.e., o modelo MLR4).

Ao final desta seção, decide-se sobre o melhor esquema de validação cruzada. Os resultados e análises das seções posteriores serão realizados apenas considerando esse esquema de validação.

O IBL4 não foi empregado nessa análise, pois ele não realiza qualquer tipo de calibração, logo, o resultado no período de teste é o mesmo independentemente do esquema de validação cruzada.

Na Figura **15** são mostrados os melhores resultados do MLR4 para todos os locais estudados, levando em consideração todos os esquemas de validação cruzada, i.e., os dados exibidos nesse gráfico dizem respeito à acurácia do MLR4 no nível de modelo (nível vertical do GCM) que proporcionou melhor acurácia para cada local e esquema de validação cruzada. É possível perceber que todos os esquemas de validação cruzada apresentam acurácias bastante similares, demonstrando a grande robustez do modelo MLR4. Contudo, destaca-se que os esquemas 4 e 6 por apresentarem acurácia ligeiramente superior que a dos demais, como pode ser visto na Figura **16**.

Figura 15 - Acurácia do MLR4 nos esquemas de validação cruzada (V.C.) estudados



Fonte: O autor (2020).

Figura 16 - Acurácia do MLR4 nos esquemas de validação cruzada (V.C.) (com ênfase sobre os esquemas 4 e 6)

Fonte: O autor (2020).

Quanto à metodologia de seleção de variáveis regressoras aqui proposta (SBO), as acurácias mais elevadas são mostradas na Figura 17. A metodologia proposta neste trabalho apresenta robustez menor que o MLR4, i.e., há uma maior flutuação dos resultados de SBO com respeito aos diferentes esquemas de validação cruzada. Isso ocorre devido à tomada de

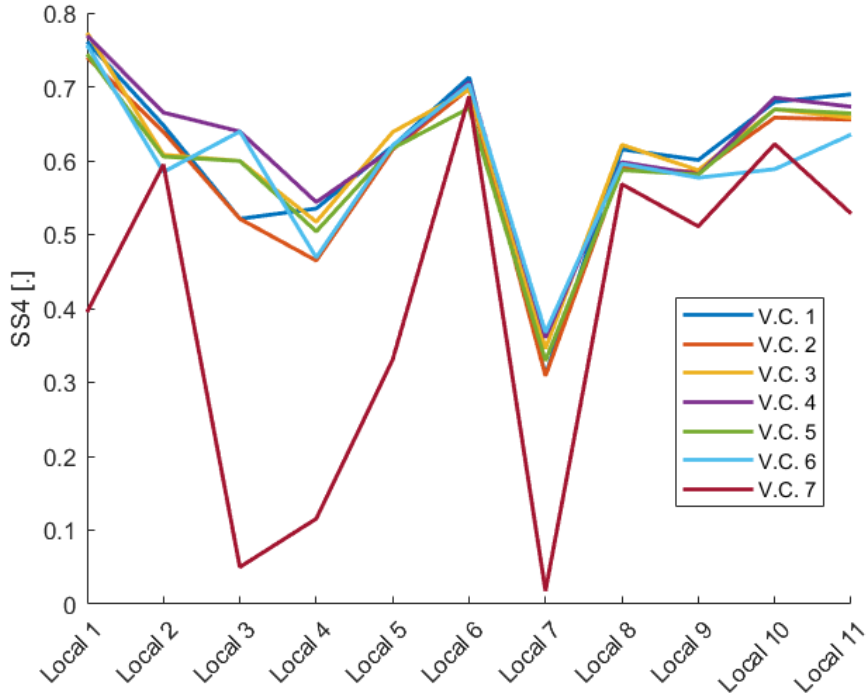
decisão realizada por SBO no período de validação, que aumenta a possibilidade de respostas do modelo[3].

Além disso, o esquema 7 de validação cruzada (V.C. 7) apresenta um resultado que difere bastante dos resultados correspondentes aos demais esquemas. Se o resultado de V.C. 7 é desprezado, a robustez do SBO se aproxima bastante daquela apresentada pelo MLR4. Cabe ressaltar que a baixa acurácia associada a V.C. 7 ocorre porque o período de validação está completamente contido no período de calibração. Por essa razão, os resultados obtidos fazendo-se uso dos dados no período de validação (i.e., desempenho do conjunto de variáveis regressoras acumuladas) não contribuem para uma boa tomada de decisão por parte do modelo, implicando em baixa acurácia.

De acordo com a metodologia SBO, a melhor forma de dividir os dados com vistas à validação cruzada se dá por meio de V.C. 4 (ver Figura **18**). Em segundo lugar, com desempenho muito próximo ao de V.C. 4, tem-se V.C. 1 (ver Figura **19**). Esses dois esquemas de validação cruzada foram mais acurados, porque seus períodos de calibração contemplam dados próximos ao período de teste, fazendo com que os coeficientes de regressão estimados sejam apropriados para descrever o comportamento da variável estudada no período de teste. Além disso, tais esquemas não utilizam dados concomitantes entre os períodos de calibração e validação, fazendo com que SBO tome melhores decisões. Como V.C. 4 apresenta a maior acurácia tanto para o MLR4 (modelo que não realiza tomada de decisão no período de validação) quanto para SBO, tal esquema foi escolhido como o para a validação cruzada. Portanto, as próximas seções de resultados e discussões analisam de forma profunda apenas os resultados obtidos fazendo uso desse esquema de validação cruzada.

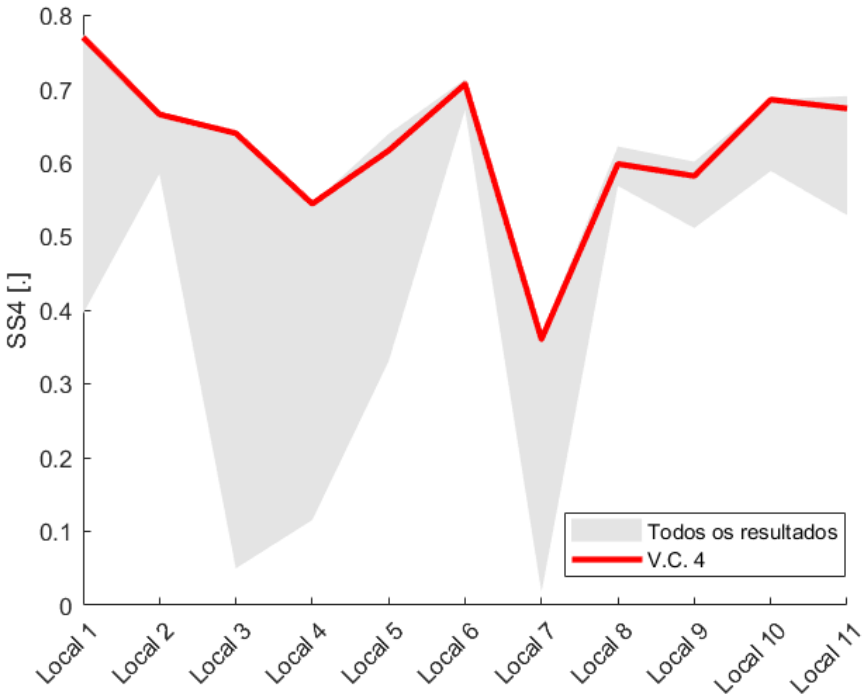
[3] O modelo MLR4, por não realizar tomada de decisão no período de validação, conta com quatro esquemas de validação cruzada distintos. No entanto, o modelo SBO, modelo que utiliza o período de validação na tomada de decisão sobre o melhor conjunto de variáveis regressoras, conta com sete esquemas distintos de validação cruzada. Esse fato implica numa maior possibilidade de respostas da SBO em relação ao MLR4.

Figura 17 - Acurácia da SBO nos esquemas de validação cruzada (V.C.) estudados



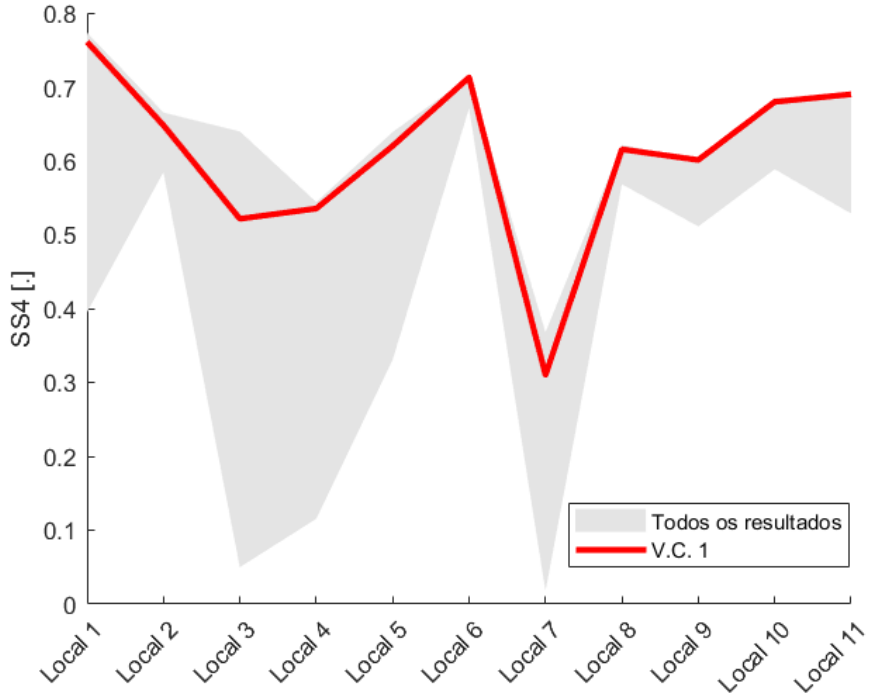
Fonte: O autor (2020).

Figura 18 - Acurácia da SBO nos esquemas de validação cruzada (V.C.) (com ênfase sobre o esquema 4)



Fonte: O autor (2020).

Figura 19 - Acurácia da SBO nos esquemas de validação cruzada (V.C.) (com ênfase sobre o esquema 1)



Fonte: O autor (2020).

5.4 Distribuição espacial dos valores das Funções de Ordenamento

Os valores das funções de ordenamento foram dispostos em mapas com o intuito de identificar quais aspectos físicos do domínio produzem efeito sobre eles.

É necessário salientar que a criação desses mapas não visa a intercomparação entre eles, mas sim a identificação de padrões independentemente dos valores obtidos, visto que o intuito das funções de ordenamento neste trabalho é única e exclusivamente ordenar as variáveis do domínio. É possível realizar várias análises com respeito à distribuição espacial desses valores, bem como sua variação ao longo do tempo. Essas análises, contudo, não serão abordadas neste trabalho.

Os principais aspectos que influenciam diretamente sobre a distribuição espacial do valor das funções de ordenamento são o tipo de cobertura (i.e., se o ponto do domínio se localiza sobre o continente ou sobre oceano) e a complexidade orográfica do local estudado.

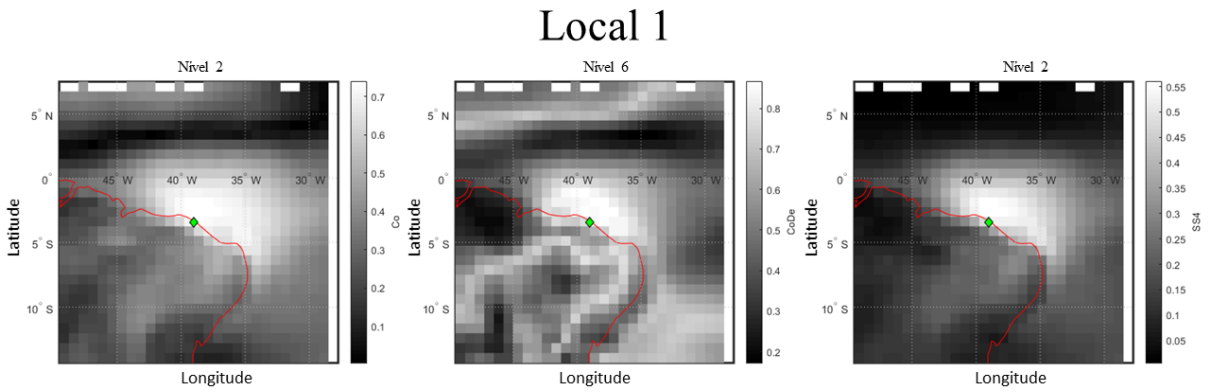
As funções de ordenamento mais influenciadas pelo tipo de cobertura são Co, CoDe e SS4 (ver seção 4.1), como pode ser visto nas Figuras 20 à 30[4]. Isso se deve principalmente à

[4] Os mapas apresentados nas Figuras 20 à 30 utilizam os níveis de modelo do GCM que obtiveram melhor desempenho para cada uma das funções de ordenamento.

diferença na inércia térmica entre o continente e o oceano, que atua diretamente na estrutura de fase e frequência dos sinais, sendo evidenciado através do coeficiente de correlação. Como essas duas funções de ordenamento são compostas diretamente pelo coeficiente de correlação, elas são as mais influenciadas. Esse efeito é perceptível mesmo em locais situados no interior do continente com orografia complexa, como os locais 10 e 11.

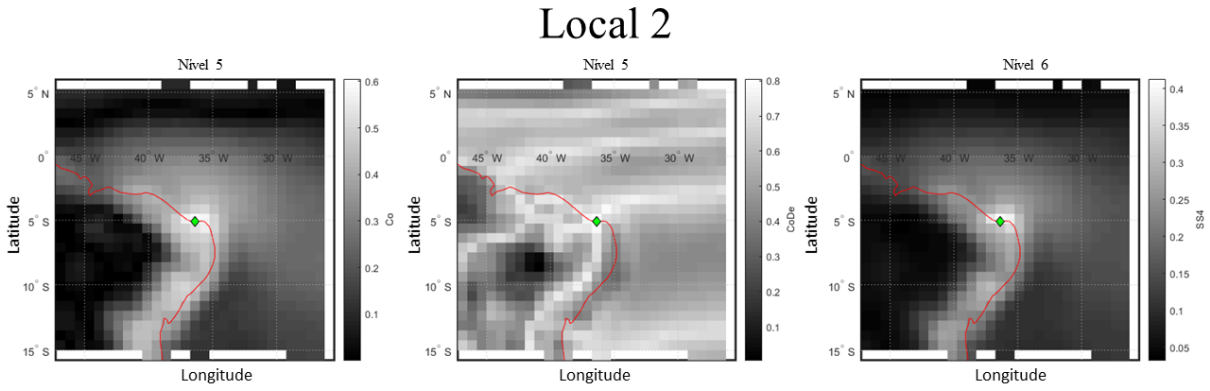
Nas Figuras 20 à 30, o mapa em tons de cinza descreve o valor da função de ordenamento para aquele ponto do GCM, o losango verde marca o local da torre anemométrica e a linha vermelha marca a interface oceano/continente.

Figura 20 – Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 1)



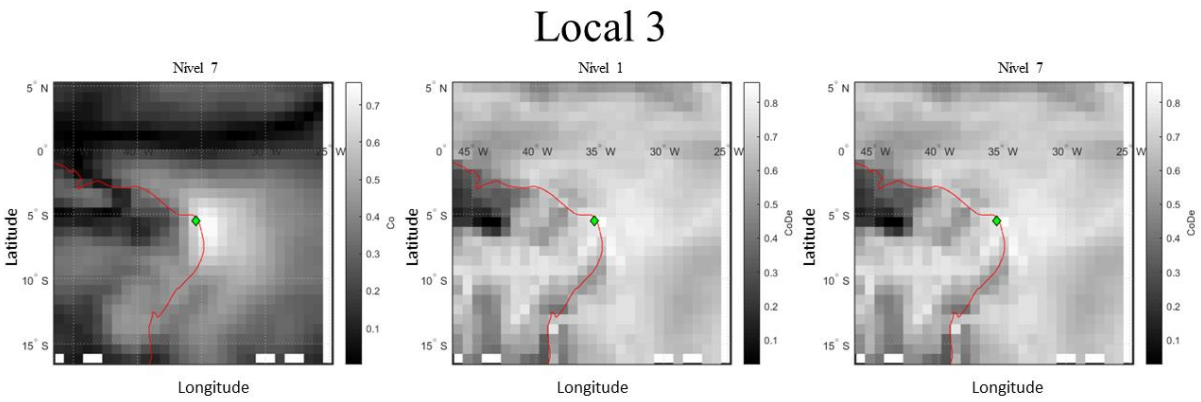
Fonte: O autor (2020).

Figura 21 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 2)



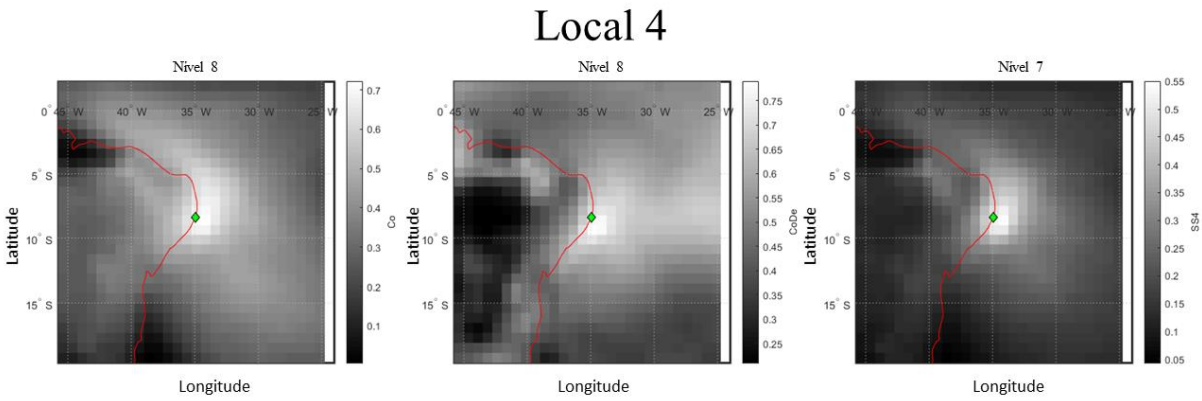
Fonte: O autor (2020).

Figura 22 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 3)



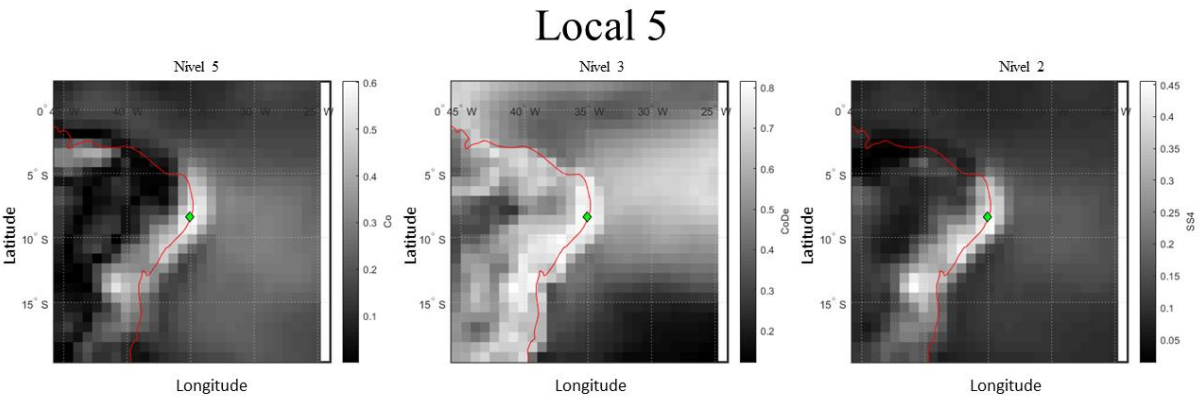
Fonte: O autor (2020).

Figura 23 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 4)



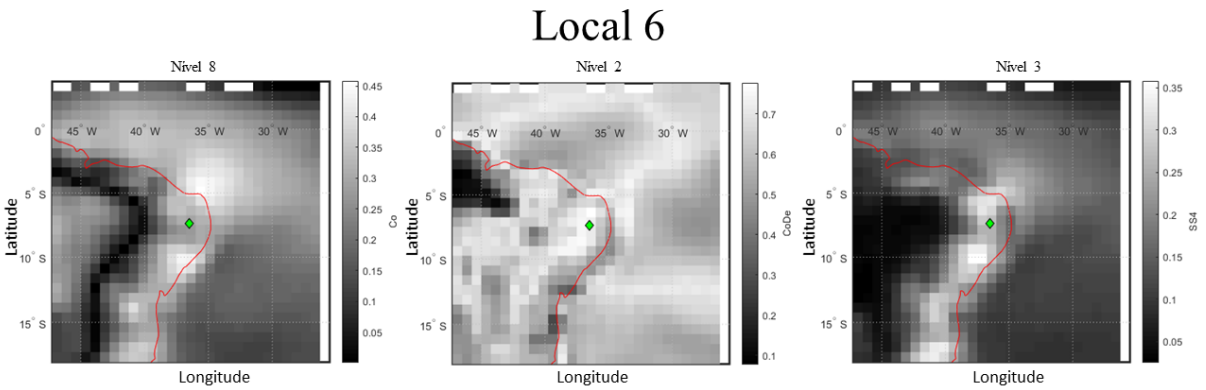
Fonte: O autor (2020).

Figura 24 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 5)



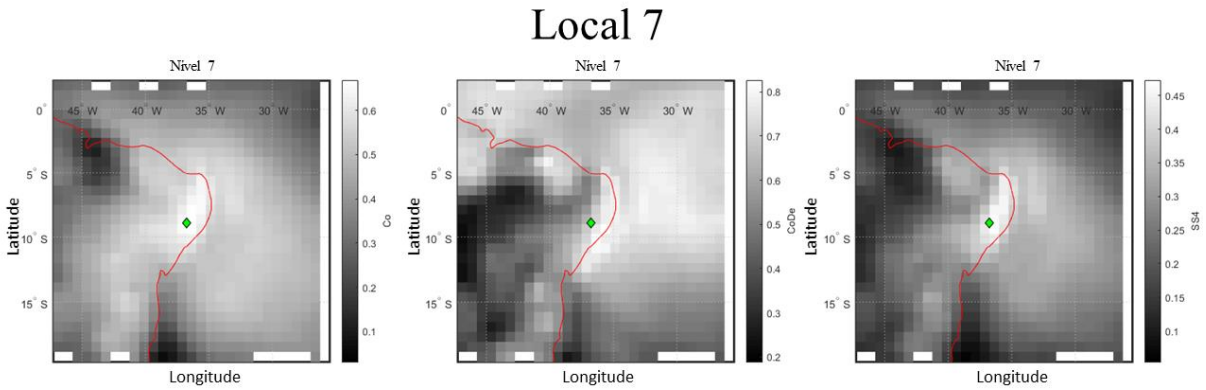
Fonte: O autor (2020).

Figura 25 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 6)



Fonte: O autor (2020).

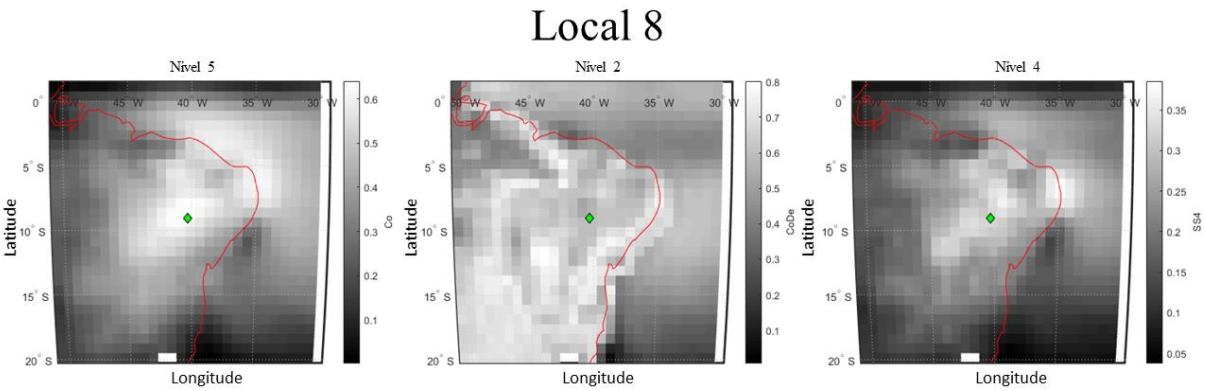
Figura 26 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 7)



Fonte: Própria

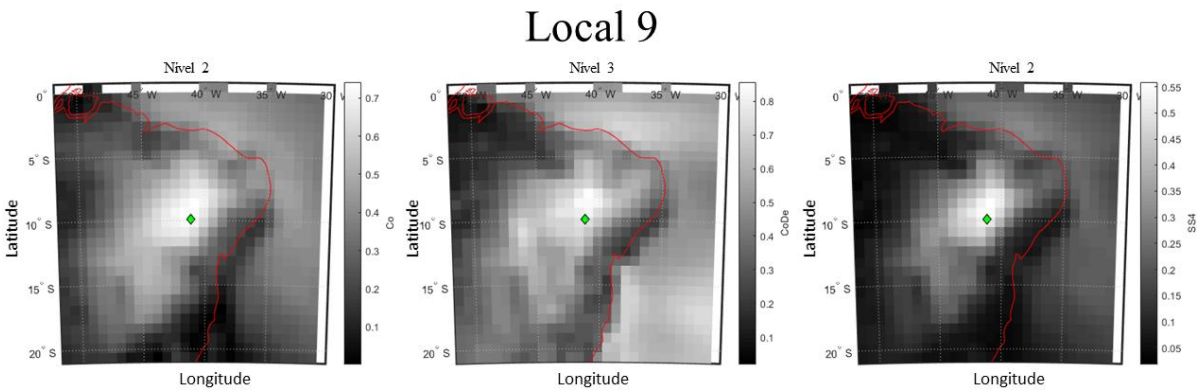
Fonte: O autor (2020).

Figura 27 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 8)



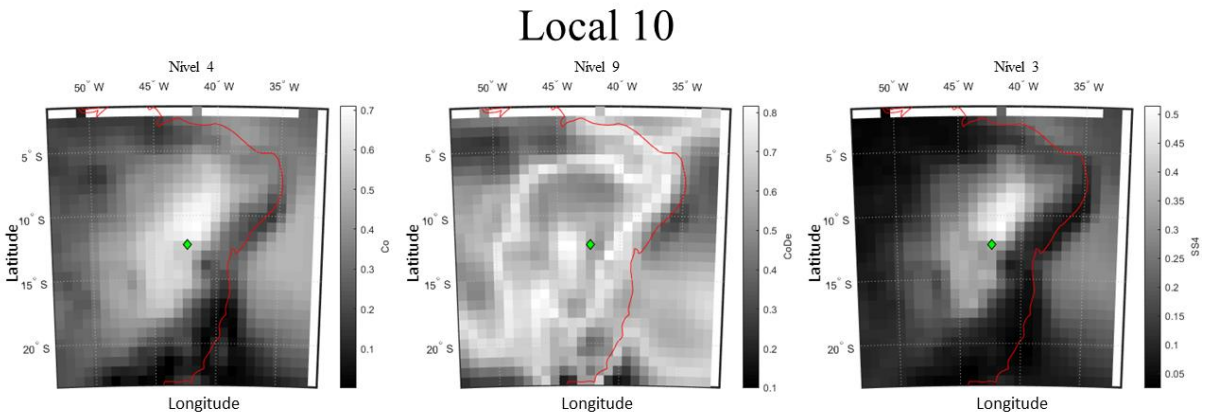
Fonte: O autor (2020).

Figura 28 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 9)



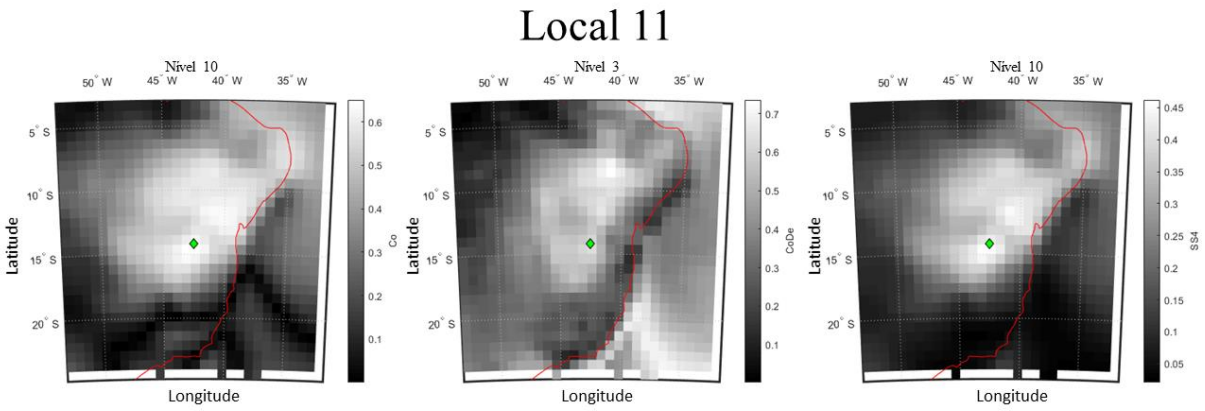
Fonte: O autor (2020).

Figura 29 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 10)



Fonte: O autor (2020).

Figura 30 - Mapas dos valores das funções de ordenamento Co, CoDe e SS4 (Local 11)



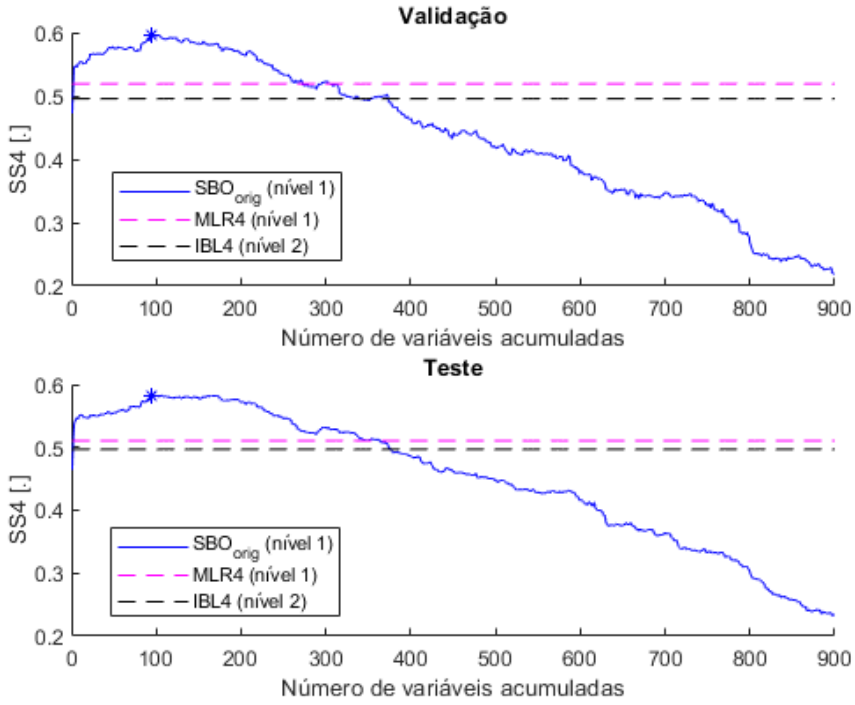
Fonte: O autor (2020).

5.5 Desempenho do conjunto de variáveis regressoras acumuladas

Nesta seção, serão discutidos os resultados do desempenho do conjunto de variáveis regressoras acumuladas (ver seção 4.2). Na Figura **31** é mostrada a performance do conjunto de variáveis regressoras acumuladas segundo a função de ordenamento Ph para o local 9. Esse local foi escolhido, pois permite fácil visualização do que será discutido ao longo desta seção. Além disso, a função de ordenamento Ph foi aquela que apresentou melhores resultados nesse local. No entanto, os aspectos apresentados através dessa figura podem ser visualizados em todos os demais locais (ver apêndice A).

No gráfico apresentado da Figura **31**, o eixo das abscissas informa quantas variáveis regressoras foram utilizadas no MLR, já o eixo das ordenadas mostra o desempenho do MLR. No gráfico, são mostradas as curvas para os períodos de validação e de teste; ambos são obtidos aplicando os coeficientes de regressão estimados no período de calibração sobre as variáveis regressoras do período em questão. É importante salientar que os resultados apresentados na Figura **31** e no apêndice A, dizem respeito ao nível de modelo do GCM que obteve o melhor desempenho para cada um dos modelos.

Figura 31 – Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 9)

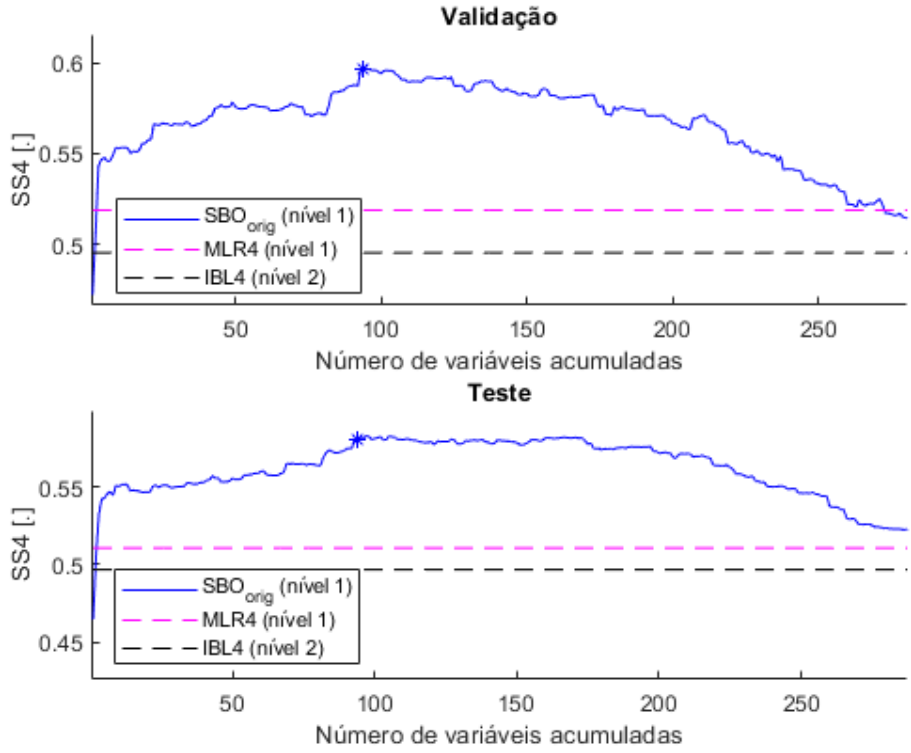


Fonte: O autor (2020).

A visualização da curva de desempenho do domínio acumulado sobre os períodos de validação e teste é de extrema importância, pois, através da comparação entre elas, é possível avaliar se as variáveis regressoras do período de validação se assemelham às do período de teste, ou seja, se a decisão acerca do melhor número de variáveis regressoras obtida durante o período de validação será válida para o período de teste. Esses fatos podem ser evidenciados através do formato da curva e da localização do ponto de melhor desempenho (máximo global da curva).

Outra característica importante a ser mencionada a respeito das curvas de performance das variáveis regressoras acumuladas é a existência de pequenas variações que ocorrem ao longo de todas as curvas (ver Figura **32**). Isso mostra que, para obtermos o melhor desempenho (máximo global da curva), será necessária a adição de variáveis que, imediatamente após serem adicionadas, diminuem o desempenho do conjunto de variáveis regressoras preexistente (ou seja, conjunto de variáveis regressoras acumuladas anterior ao da variável em questão), como pode ser visto na Figura **32**. Esse resultado está de acordo com os exemplos apresentados por Guyon e Elisseeff (2003), e, provavelmente, se deve à complementariedade de variáveis no domínio da frequência, isso é, o MLR é capaz de combinar variáveis regressoras que individualmente possuem baixa capacidade de descrever o sinal observado de forma a obter um sinal no domínio da frequência mais próximo daquele apresentado pelo preditando, aumentando, dessa forma, a acurácia do modelo.

Figura 32 – Curva de desempenho do conjunto de variáveis regressoras acumuladas com ampliação (Local 9)



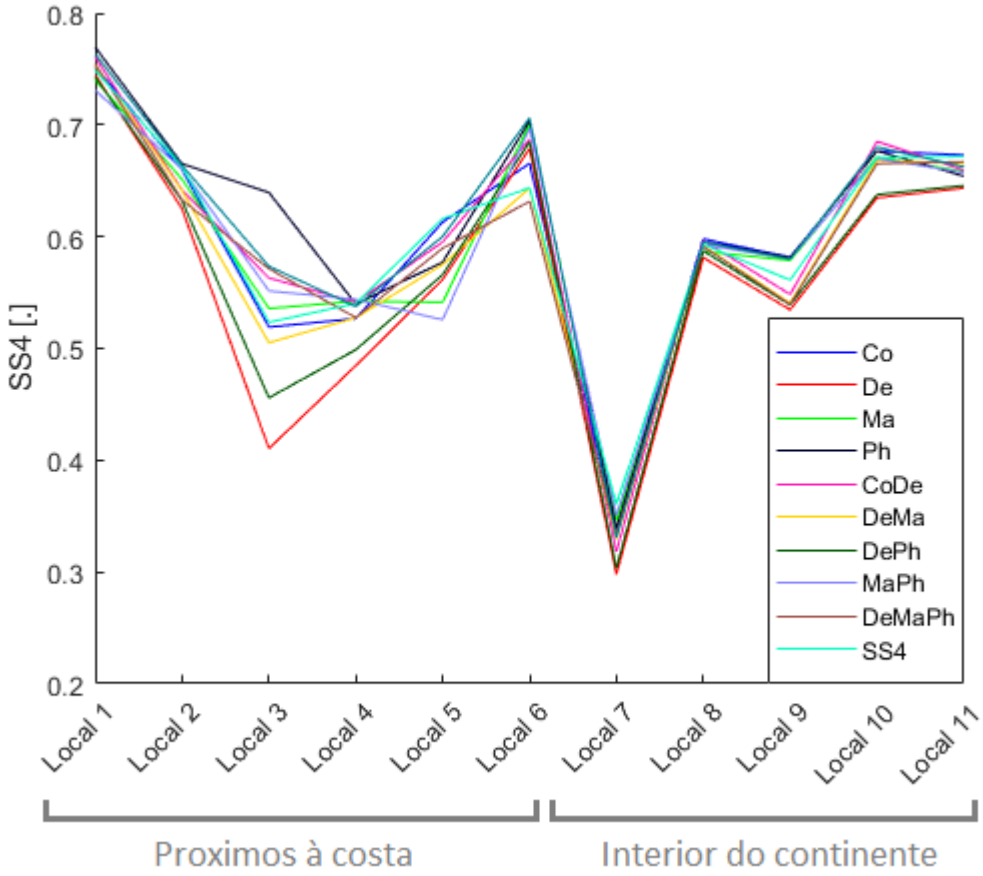
Fonte: O autor (2020).

Isso também mostra a necessidade de utilização de domínios com grandes áreas para realização da seleção das variáveis regressoras, pois, dessa forma, haverá um aumento da probabilidade de encontrar variáveis que, ao serem combinadas através de um método regressivo, permitem a filtragem de informações que contribuam para a melhora no desempenho do modelo.

5.6 Desempenho das funções de ordenamento

Nesta seção iremos analisar o desempenho da metodologia de seleção de variáveis regressoras desenvolvida neste trabalho segundo o tipo de função de ordenamento utilizada (ver seção 4.1). A Figura **33** mostra o desempenho da SBO nos locais estudados quando são utilizadas cada uma das funções de ordenamento. Os desempenhos apresentados no gráfico da Figura **33** dizem respeito ao nível de modelo do GCM que obteve a melhor performance para cada uma das funções de ordenamento.

Figura 33 – Melhores desempenhos de cada uma das funções de ordenamento

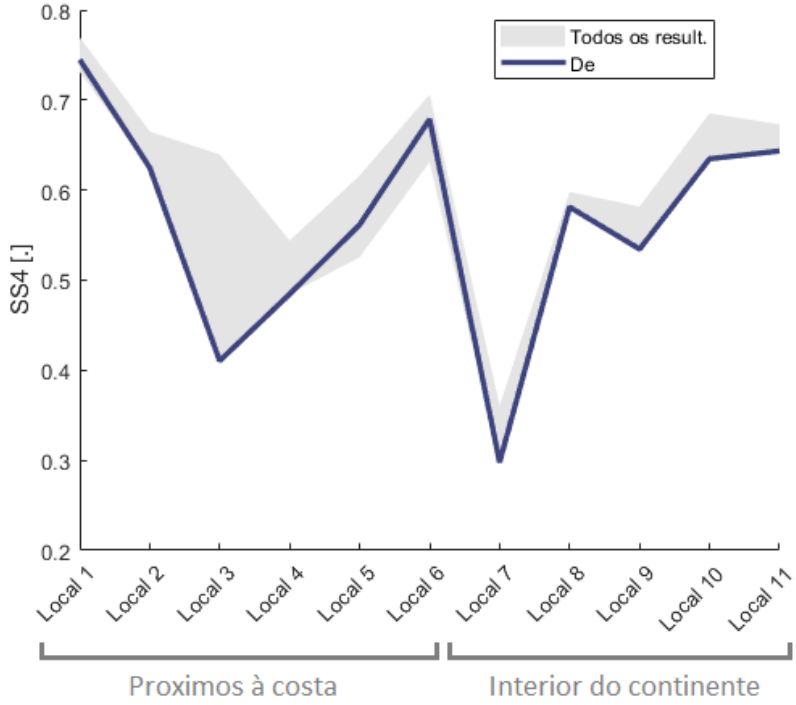


Fonte: O autor (2020).

É possível perceber que a acurácia do método pode variar sensivelmente, a depender da função de ordenamento utilizada. Isso demonstra a importância da função de ordenamento, pois sua escolha afetará drasticamente o resultado do *downscaling*.

A função de ordenamento que apresentou o pior acurácia foi a De, como pode ser visto na Figura **34.** Isso ocorre, pois essa função de ordenamento não leva em consideração a similaridade dos sinais da variável do domínio de busca e do preditando no domínio da frequência.

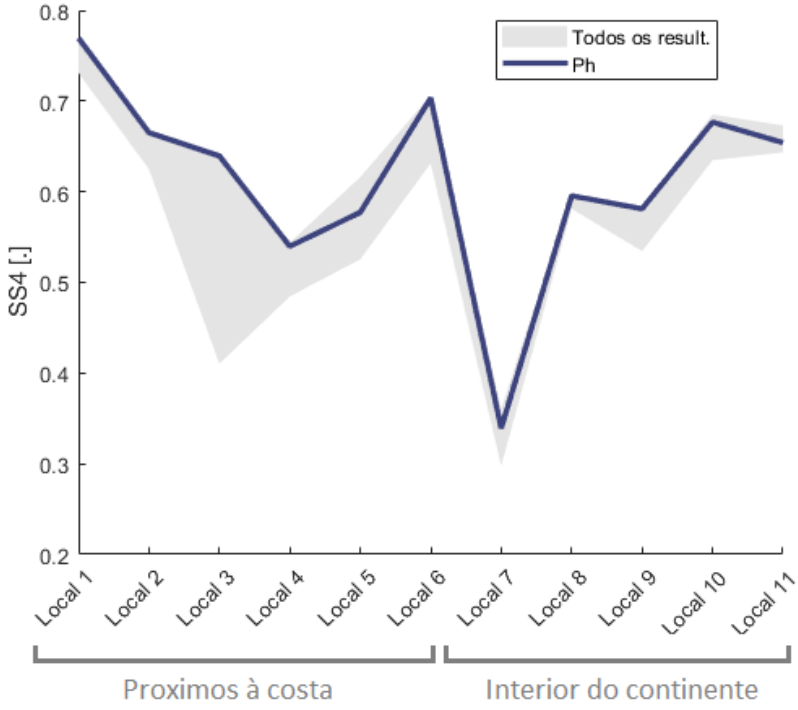
Figura 34 – Desempenho da função de ordenamento De



Fonte: O autor (2020).

Como melhores funções de ordenamento podemos citar Ph e Co, pois estas funções promoveram excelentes resultados. A função de ordenamento Ph foi responsável pelas melhores performances nos locais próximos a costa (i.e., locais 1 a 6), com exceção do local 5 (ver Figura **35**), pois apesar de estar situado em uma área com orografia simples, possui diversas construções em seu entorno (i.e., obstáculos). Além disso, essa função de ordenamento obteve resultados muito próximos aos melhores em locais de orografia simples situados no interior do continente (i.e., locais 8, 9 e 10).

Figura 35 - Desempenho da função de ordenamento Ph

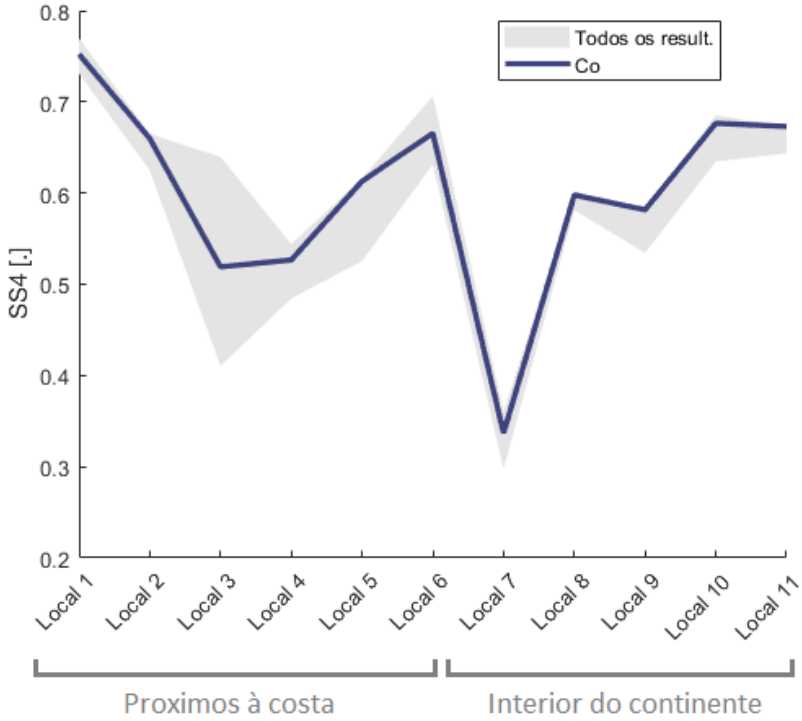


Fonte: O autor (2020).

A função de ordenamento Co foi responsável por obter os melhores resultados nos locais situados no interior do continente (i.e., locais 7 a 11) e no local 5, que fica situado na costa, porém possui diversos obstáculos em seu entorno (ver Figura **36**).

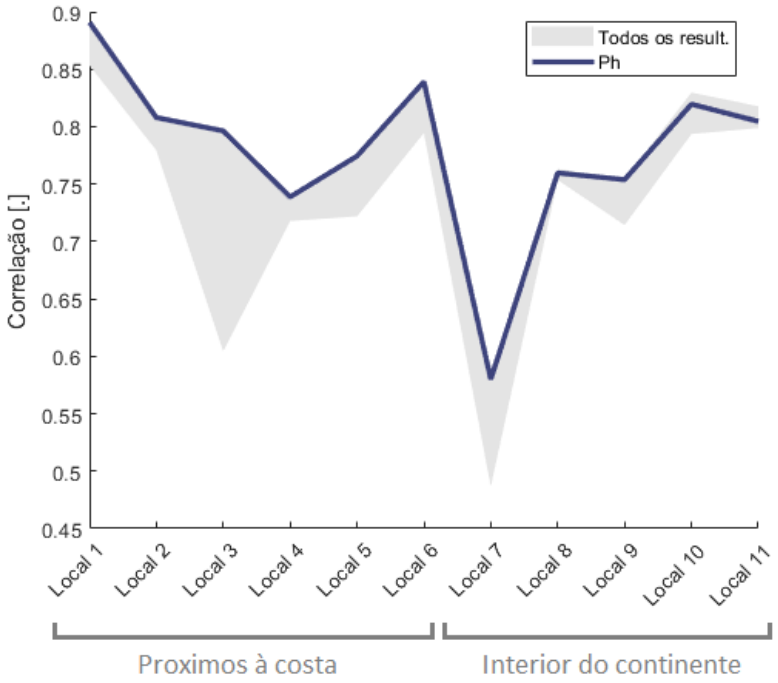
Um resultado bastante importante pode ser observado quando analisamos a correlação obtida pelo modelo SBO em função dos locais estudados. Isso porque a função de ordenamento Ph é responsável por obter os mais altos coeficientes de correlação na maioria dos locais estudados (ver Figura **37**). As únicas exceções são os locais 10 e 11, que são locais situados no interior do continente e que possuem orografia bastante complexa.

Figura 36 - Desempenho da função de ordenamento Co



Fonte: O autor (2020).

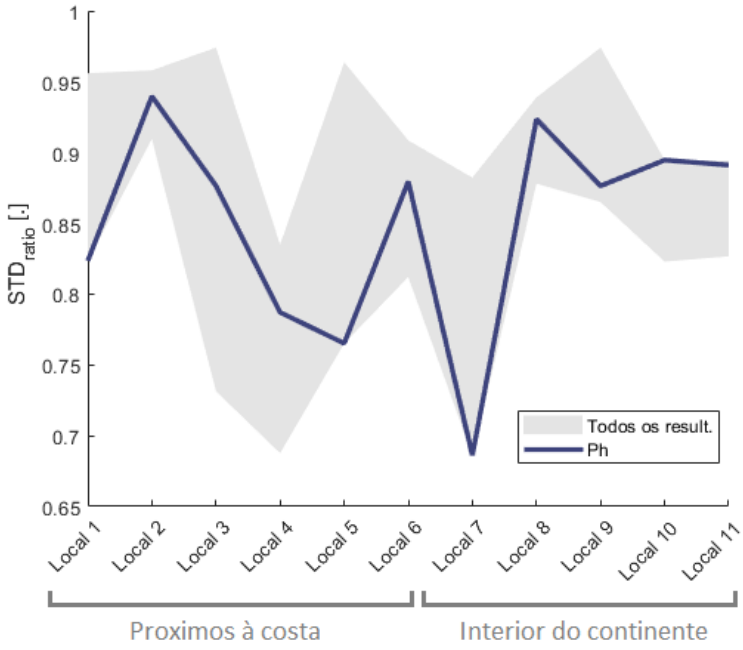
Figura 37 – Correlação obtida pela função de ordenamento Ph



Fonte: O autor (2020).

Tendo em vista que o SS4 é uma função do coeficiente de correlação e da razão dos desvios padrão entre o modelo e a observação, fica claro que a seleção executada fazendo uso da função de ordenamento Ph possui dificuldade em reproduzir o desvio padrão da observação, tendendo a subestimá-lo (ver Figura **38**).

Figura 38 – Razão entre os desvios padrão obtida pela função de ordenamento Ph



Fonte: O autor (2020).

No entanto, a correção do desvio padrão da série estimada no *downscaling* através de pós processamento estatístico se dá de forma muito mais simples que a correção do sinal no domínio da frequência. Logo, fica claro que a melhor função de ordenamento, ou seja, aquela que possui a maior capacidade de obter bons resultados, é a função Ph.

5.7 Comparação com modelos de referência

Nesta seção serão comparados os resultados obtidos pela metodologia proposta neste trabalho com os resultados obtidos pelos modelos de referência (ver seção 4.4). Primeiramente, os resultados da SBO serão comparados com aqueles obtidos pelos modelos que não realizam seleção de variáveis regressoras (i.e., IBL4 e MLR4). Em seguida, serão comparados os resultados da SBO com aqueles obtidos pelos modelos de referência que utilizam seleção

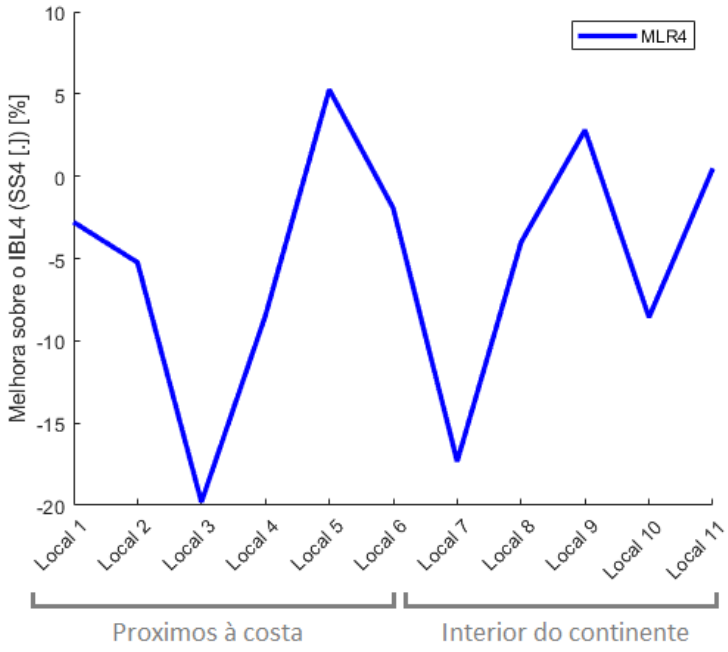
embedded (i.e., SWR e Lasso). Por fim, serão comparados os resultados da SBO com os aqueles obtidos pelos modelos de referência que utilizam seleção *wrapper* (i.e., *Forward Selection*).

É importante ressaltar que, assim como na seção anterior, os desempenhos apresentados nesta seção dizem respeito ao nível de modelo do GCM que obteve a melhor performance para cada um dos modelos.

5.7.1 Comparação com métodos que não realizam seleção de variáveis regressoras

Primeiramente iremos avaliar a acurácia do MLR4 frente ao IBL4, ou seja, iremos comparar a eficácia do método regressivo mais simples (i.e., o MLR4) com um modelo que reflete a saída do GCM sem que qualquer coeficiente estatístico seja estimado (i.e., IBL4). Na Figura **39** é mostrada a melhora do MLR4 sobre o IBL4 para todos os locais estudados.

Figura 39 – Melhora sobre o IBL4 obtida pelo MLR4 (SS4)



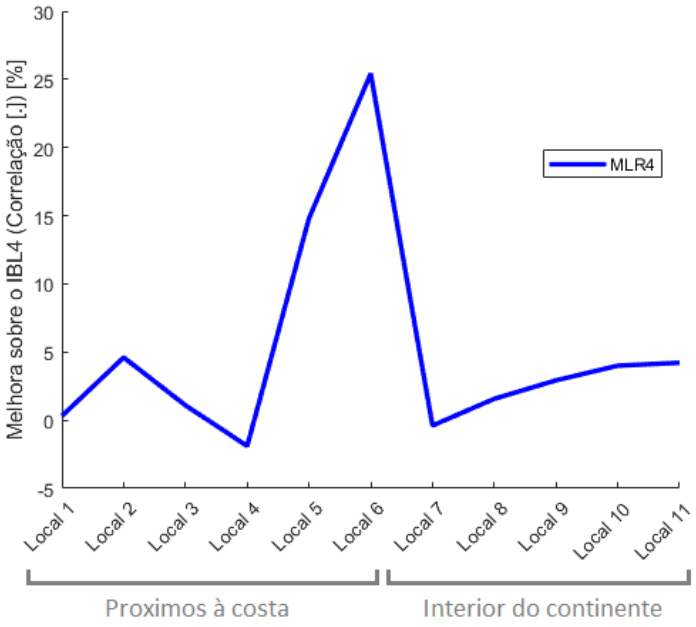
Fonte: O autor (2020).

O MLR4 supera o IBL4 em apenas dois dos onze locais estudados (i.e., locais 5 e 9) e apresenta acurácia próxima daquela obtida pelo IBL4 no local 11. No entanto, como mencionado na subseção 4.1.4, o SS4 é uma função do coeficiente de correlação e da razão entre os desvios padrão.

Dessa forma, analisando a melhora do MLR4 sobre o IBL4 de acordo com a correlação (ver Figura **40**), percebemos que o MLR4 supera o IBL4 em seis dos onze locais estudados (i.e., locais 2, 5, 6, 9, 10 e 11) e apresenta acurácia próxima daquela obtida pelo IBL4 em quatro locais (i.e., locais 1, 3, 7 e 8). Logo, é perceptível que a grande dificuldade do MLR4 em relação ao IBL4 se refere à descrição da variabilidade da série observada. Através da Figura **41**, percebemos que o MLR4 tende a subestimar a variabilidade da série estimada.

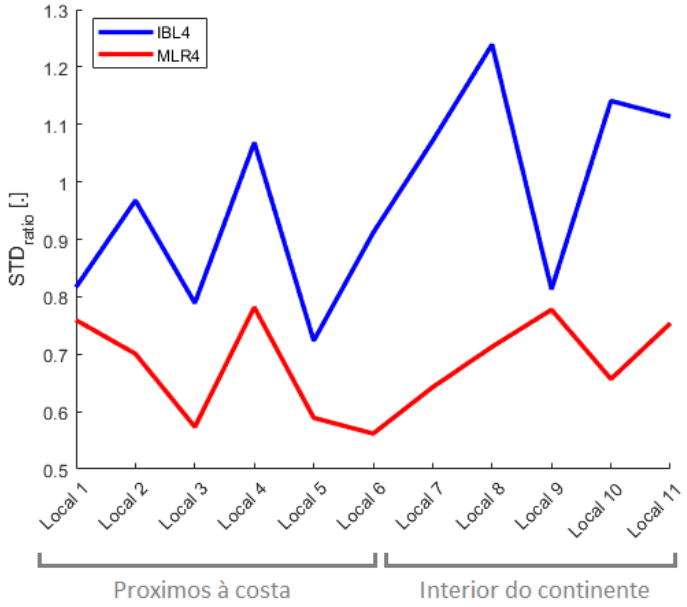
A Figura **42** apresenta a acurácia da metodologia proposta neste trabalho em relação ao IBL4. A SBO apresenta uma grande melhora em relação ao IBL4, pois supera a acurácia desse modelo em dez dos onze locais estudados (i.e., todos os locais exceto o local 4). Além do mais, obteve uma melhora mais expressiva do que aquela obtida pelo MLR4, com melhoras da ordem de vinte porcento.

Figura 40 - Melhora sobre o IBL4 obtida pelo MLR4 (correlação)



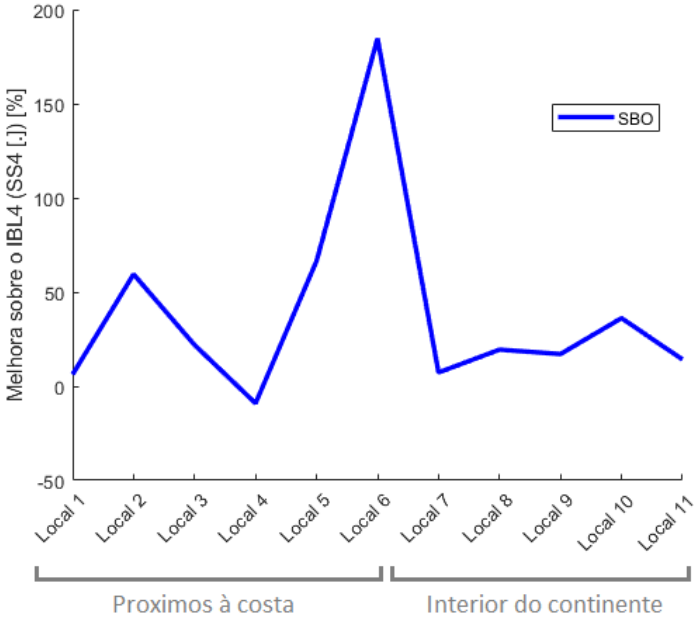
Fonte: O autor (2020).

Figura 41 – Comparação entre MLR4 e IBL4 (razão entre desvios padrão)



Fonte: O autor (2020).

Figura 42 - Melhora sobre o IBL4 obtida pela SBO (SS4)

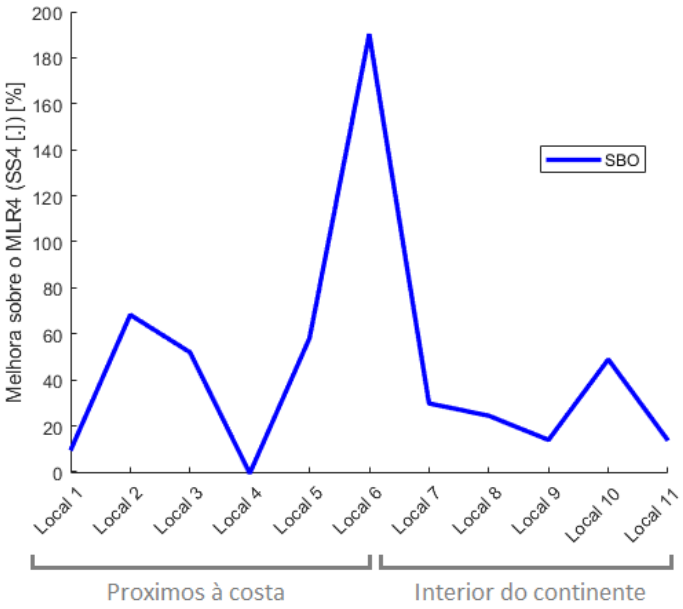


Fonte: O autor (2020).

A Figura **43** apresenta a melhora da SBO sobre o MLR4. A seleção de variáveis regressoras obteve excelente melhora sobre o MLR4, superando sua acurácia em dez dos onze locais estudados e obtendo acurácia similar em um local (i.e., local 4). Além disso, as melhoras

obtidas variam entre dez e setenta porcento, comprovando o grande benefício da utilização do método de seleção de variáveis regressoras proposto neste trabalho.

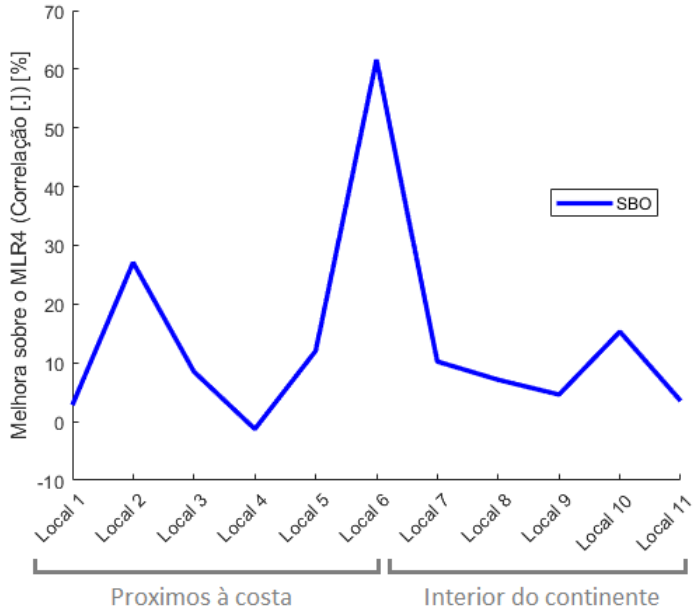
Figura 43 - Melhora sobre o MLR4 obtida pela SBO (SS4)



Fonte: O autor (2020).

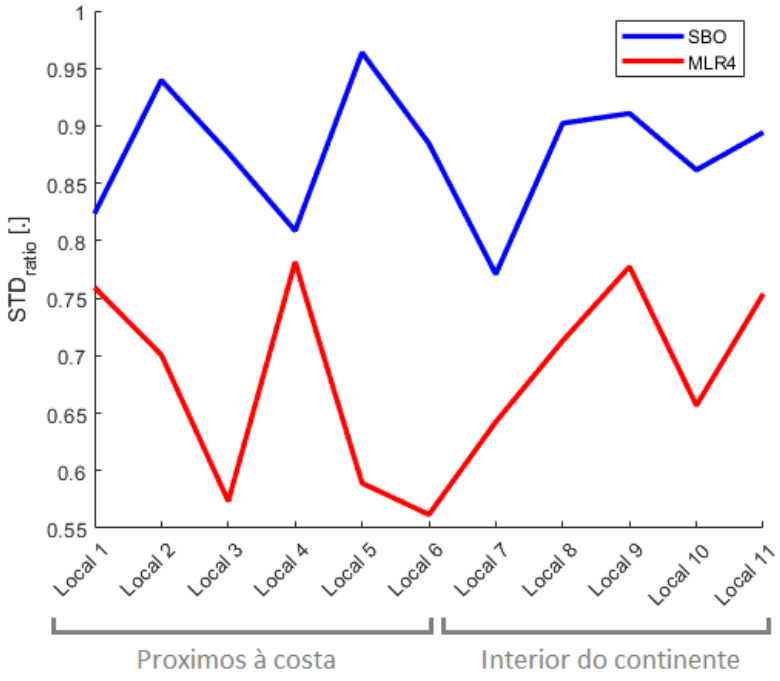
Essa melhora em relação ao MLR4 e ao IBL4 se deve à capacidade do SBO em melhorar a descrição do sinal observado tanto em relação ao domínio da frequência (visualizado através do coeficiente de correlação – ver Figura **44**) quanto em relação à variabilidade (visualizado através da razão entre os desvios padrão – ver Figura **45**).

Figura 44 - Melhora sobre o MLR4 obtida pela SBO (correlação)



Fonte: O autor (2020).

Figura 45 - Comparação entre MLR4 e SBO (razão entre desvios padrão)

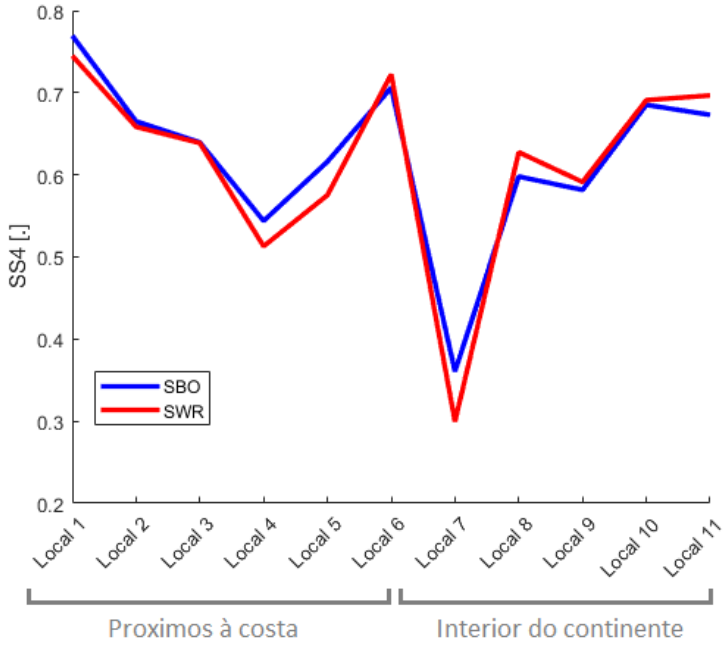


Fonte: O autor (2020).

5.7.2 Comparação com métodos de seleção de variáveis regressoras *embedded*

A Figura **46** apresenta as acurácias da SBO e do modelo SWR. É possível perceber que a SBO apresenta melhora em relação ao SWR nos locais situados na costa, no entanto, possui dificuldade em superar o SWR nos locais situados no interior, com exceção do local 7. No cômputo geral, a SBO superou o SWR em quatro dos onze locais estudados (i.e., locais 1, 4, 5 e 7), obteve resultados resultado semelhante em cinco locais (i.e., locais 2, 3, 6, 9 e 10) e foi superado pelo SWR em dois locais (i.e., locais 8 e 11).

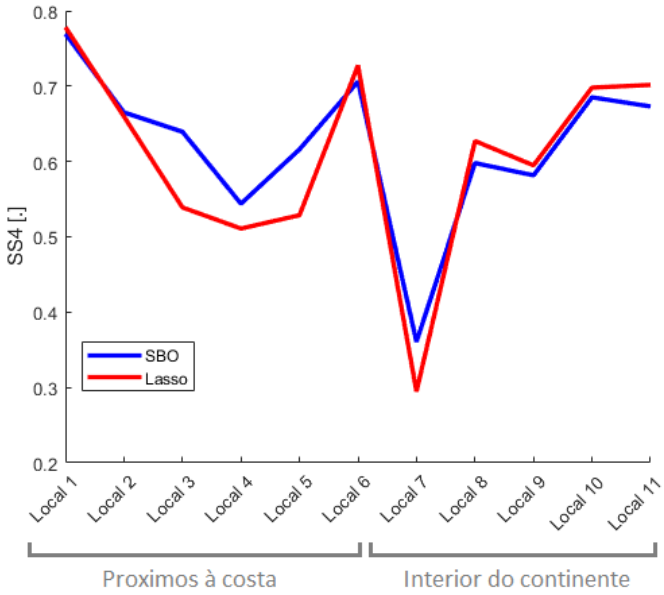
Figura 46 - Comparação entre SWR e SBO (SS4)



Fonte: O autor (2020).

Na Figura **47** são mostradas as acurácias da SBO e do modelo Lasso. Assim como no caso do SWR, a SOB apresenta melhores resultados nos locais próximos à costa, e possui dificuldade em superar o modelo Lasso nos locais situados no interior. No entanto, as melhoras da SBO sobre o modelo Lasso são mais expressivas que em relação ao SWR, exibindo melhoras de 0.1 no valor do SS4 para os locais 3 e 5. Além disso, a superação do Lasso e da SWR em relação a SBO não vai além de 0.03 no valor do SS4 (como pode ser visto nos locais 8 e 11, nas Figuras 49 e 50).

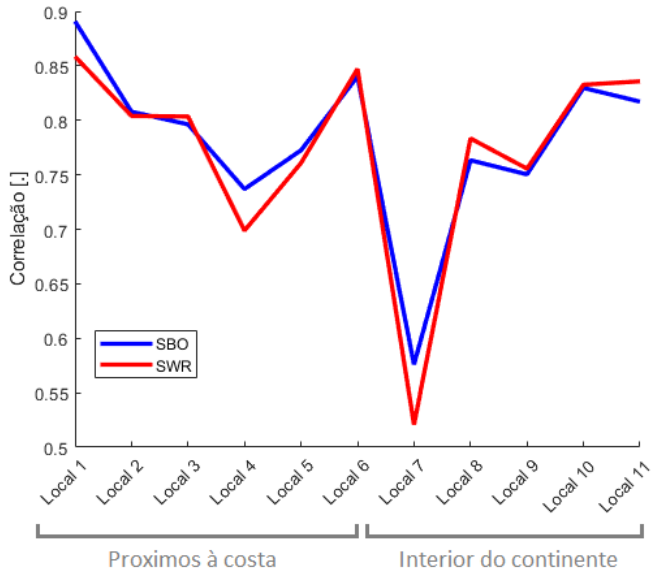
Figura 47 - Comparação entre Lasso e SBO (SS4)



Fonte: O autor (2020).

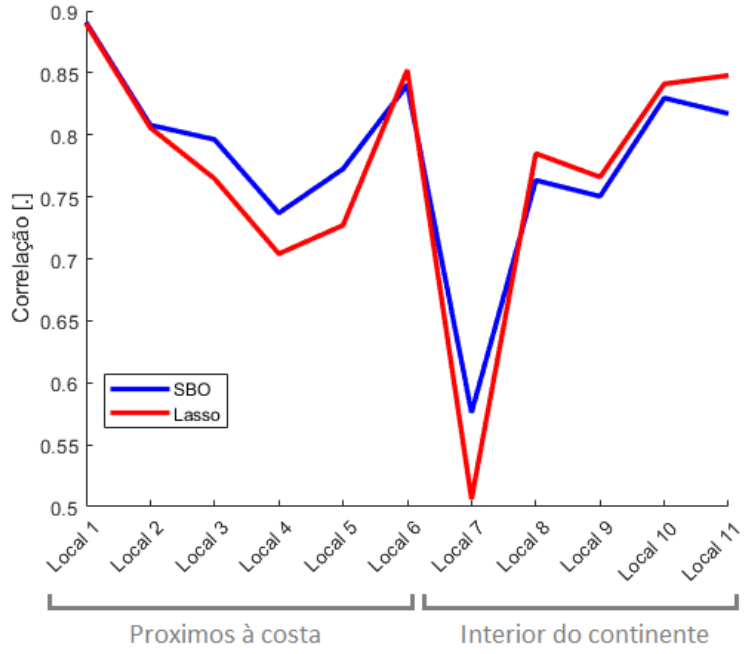
O motivo da melhora na acurácia em relação aos modelos de referência *embedded* se deve tanto à melhora da descrição das características da série observada no domínio da frequência (evidenciada através do coeficiente de correlação – ver Figuras 48 e 49), quanto à melhora na descrição da variabilidade do sinal observado (evidenciado através da razão entre os desvios padrão – ver Figuras 50 e 51).

Figura 48 - Comparação entre SWR e SBO (correlação)



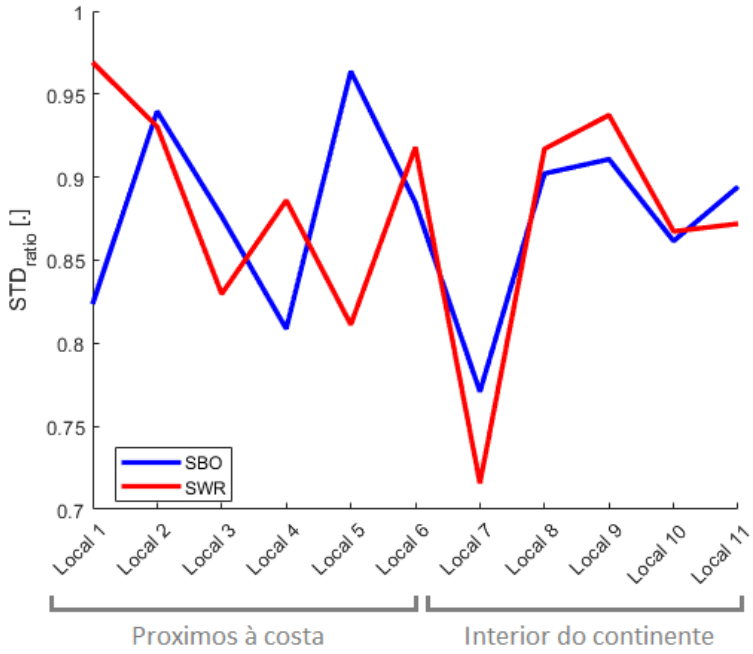
Fonte: O autor (2020).

Figura 49 - Comparação entre Lasso e SBO (correlação)



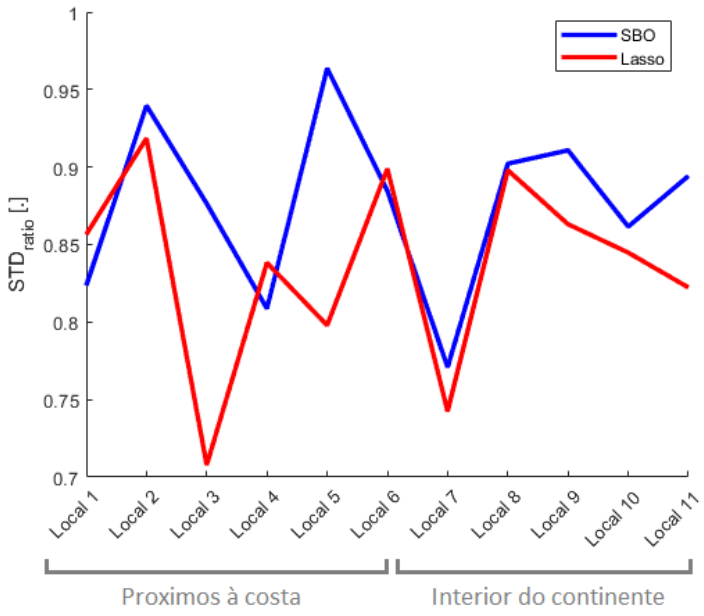
Fonte: O autor (2020).

Figura 50 - Comparação entre SWR e SBO (razão entre os desvios padrão)



Fonte: O autor (2020).

Figura 51 - Comparação entre Lasso e SBO (razão entre os desvios padrão)



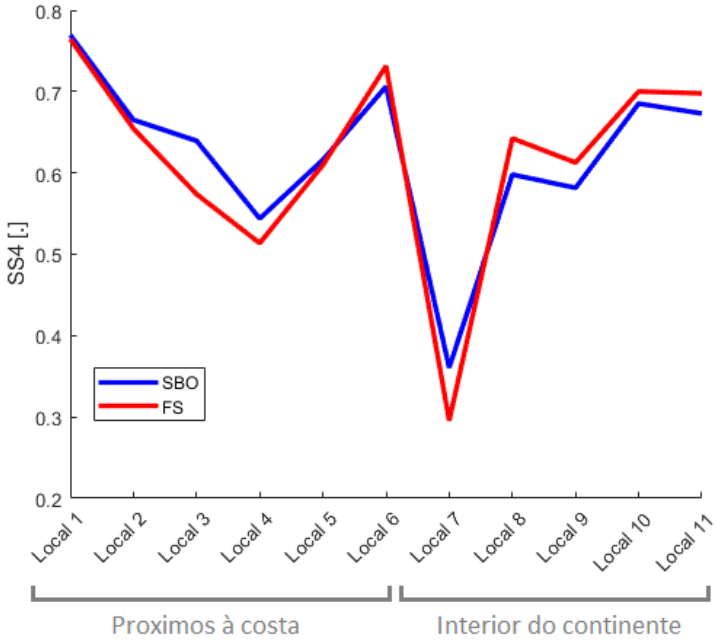
Fonte: O autor (2020).

A SBO obteve um bom resultado frente aos métodos *embedded*, pois esses métodos são extremamente específicos e foram desenvolvidos exclusivamente para maximizar a acurácia da regressão linear múltipla. Dessa forma, conseguimos mostrar que a SBO foi capaz de obter resultados similares ou melhores do que aqueles obtidos pelos modelos *embedded* em nove dos onze locais estudados, provando ser uma alternativa viável, mesmo em relação a modelos especializados para a regressão linear múltipla.

5.7.3 Comparação com método de seleção de variáveis regressoras *wrapper*

A Figura **52** apresenta a acurácia da SBO e do modelo Forward Selection. Assim como no caso dos modelos de referência *embedded*, a SBO tende a obter resultados melhores que o FS nos locais próximos a costa e resultados piores nos locais situados no interior. No entanto, a SBO supera o FS em apenas três dos onze locais estudados (i.e., locais 3, 4 e 7), apresenta acurácia similar em 3 locais (i.e., locais 1, 2 e 5) e é superado pela FS em cinco locais (i.e., locais 6, 8, 9, 10 e 11).

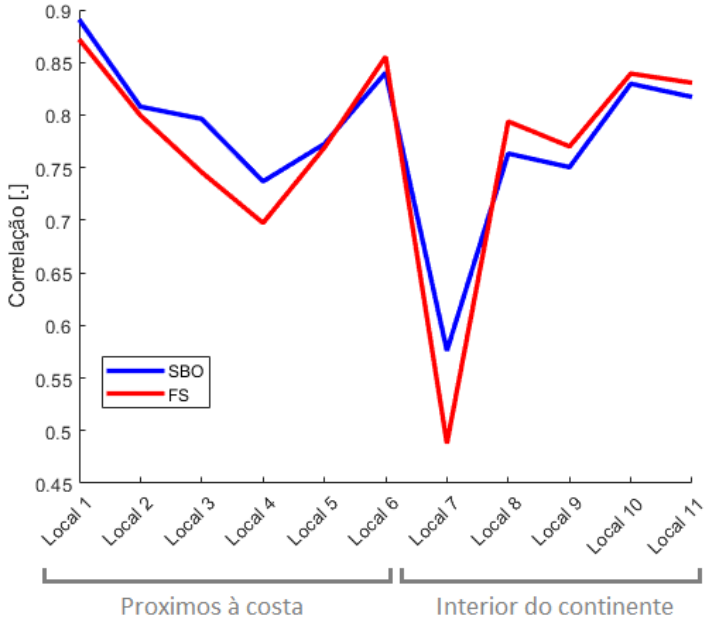
Figura 52 - Comparação entre FS e SBO (SS4)



Fonte: O autor (2020).

O modelo FS obtém melhores resultados do que a SBO, principalmente devido à melhor descrição do sinal observado no domínio da frequência, como pode ser visto através do coeficiente de correlação (ver Figura **53**).

Figura 53 - Comparação entre FS e SBO (correlação)



Fonte: O autor (2020).

Apesar do aparente baixo rendimento em relação ao FS (i.e., iguala ou supera o resultado do FS em seis dois onze locais estudados), a metodologia proposta neste trabalho possui uma grande vantagem quando levamos em consideração o esforço computacional necessário para a seleção do melhor conjunto de variáveis regressoras (Considerando apenas o nível de modelo que obteve o melhor resultado).

Na Tabela 5 é mostrado o número de regressões necessárias até o máximo global da curva de decisão sobre o melhor conjunto de variáveis regressoras. No entanto, são necessárias mais regressões para compreender que esse ponto é o ponto de máximo global da curva.

Tabela 5 - Número de regressões para encontrar o melhor conjunto de variáveis regressoras

	Número de regressões até o máximo global	
	Forward Selection	**Seleção Baseada no Ordenamento**
Local 1	95.931	8
Local 2	259.735	143
Local 3	55.456	58
Local 4	117.505	9
Local 5	100.620	104
Local 6	119.016	165
Local 7	115.990	45
Local 8	170.065	106
Local 9	182.871	118
Local 10	150.535	194
Local 11	209.275	159

Fonte: O autor (2020)

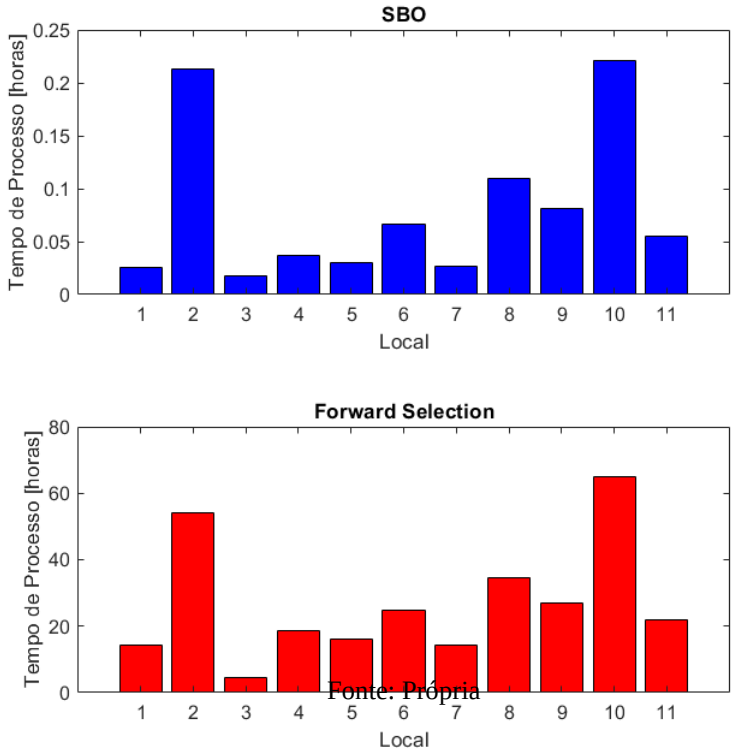
A metodologia de seleção de variáveis regressoras desenvolvida neste trabalho obteve uma diminuição drástica no esforço computacional necessário para a seleção do conjunto de variáveis regressoras, na ordem de mil vezes menor, chegando em alguns casos à ordem de dez mil vezes menor. Além disso, as iterações da SBO são completamente independentes, ou seja, a iteração i+1 não depende do resultado da iteração i, permitindo a paralelização das iterações, consequentemente, diminuindo o tempo de processamento. Já as iterações do FS não são

completamente independentes, exigindo um tempo de processamento radicalmente maior (ver Figura **54**).

Na Figura **54** é apresentado o tempo necessário de processamento para que a SBO e o *Forward Selection* realizem a seleção de variáveis regressoras. Esse tempo de processamento leva em consideração a seleção de variáveis regressoras para os dez níveis de modelo do GCM.

A máquina utilizada para o processamento possui processador AMD Opteron 6376 (2,3 GHz) com memória RAM de 32Gb.

Figura 54 - Tempo de processamento SBO e FS



Fonte: Própria

Fonte: O autor (2020).

O tempo de processamento varia sensivelmente a depender do local estudado, pois ele está intrinsecamente relacionado ao comprimento da série temporal estudada (ver Figura **13**). No entanto, a SBO reduziu o tempo de processamento em uma razão aproximada de quatrocentas vezes.

Tendo em vista a acurácia da SBO frente àquela obtida pelo FS e o esforço computacional demandado pelos dois métodos, fica clara a vantagem em se utilizar a metodologia de seleção de variáveis regressoras desenvolvida neste trabalho.

6 CONCLUSÕES E PERSPECTIVAS

Este trabalho apresenta uma metodologia de seleção de variáveis regressoras objetiva, direta e individual que pode ser aplicada em conjunto com qualquer tipo de modelo regressivo. Tal metodologia apresentou boa acurácia e um custo computacional bastante competitivo. Além disso, neste trabalho foram propostas métricas que definem a relevância das variáveis do domínio de busca (que permitiram resultados mais acurados que métricas já conhecidas) e foram avaliados diferentes esquemas para a divisão da base de dados com vistas à validação cruzada.

A metodologia de seleção de variáveis regressoras desenvolvida neste trabalho (SBO) apresentou bons resultados quando comparada a métodos criados especificamente para a regressão linear múltipla (métodos *embedded*), obtendo resultados iguais ou superiores a esses métodos em nove dos onze locais estudados (ver Tabela 6). Quando comparado com um método mais geral, que pode ser aplicado em conjunto com qualquer modelo regressivo (*Forward Selection*), a SBO apresentou boa acurácia, com uma diminuição drástica no esforço computacional requerido (ver Tabela 5), na ordem de mil vezes menor, chegando em alguns casos à ordem de dez mil vezes menor.

Tabela 6 – Comparação final da acurácia da SBO com a acurácia dos modelos de referência

	Número de locais em que a acurácia da SBO foi:		
	Inferior	**Similar**	**Superior**
IBL4	1	0	10
MLR4	0	1	10
SWR	2	5	4
Lasso	2	5	4
FS	5	3	3

Fonte: O autor (2020).

Também é importante salientar que a SBO obteve acurácia significativamente superior aos demais modelos de referência quando considerados locais próximos à costa, principalmente aqueles com orografia simples (ver Tabela 7). Esse resultado demonstra a grande aplicabilidade do método, visto que grande parte das centrais eólicas brasileiras estão localizadas próximas à costa. Além disso, a metodologia desenvolvida neste trabalho pode ser de grande aplicabilidade

na prospecção e operação dos futuros complexos de energia eólica *offshore*, tendo em vista a simplicidade orográfica e o posicionamento sobre o oceano.

Tabela 7 - Comparação final da acurácia da SBO com a acurácia dos modelos de referência para locais situados próximos à costa

	Número de locais próximos à costa em que a acurácia da SBO foi:		
	Inferior	**Similar**	**Superior**
IBL4	1	0	5
MLR4	0	1	5
SWR	0	3	3
Lasso	0	3	3
FS	1	3	2

Fonte: O autor (2020).

Dentre todas as métricas avaliadas com vistas à definição da relevância das variáveis do domínio de busca, aquela que apresentou melhores resultados foi a *Ph*, uma das métricas propostas neste trabalho, sendo esta métrica responsável pela melhor acurácia da SBO nos locais de orografia simples. Observou-se ainda que a métrica *Ph* promoveu a melhor descrição do sinal observado no domínio da frequência para todos os locais estudados, o que ficou demonstrado por meio dos correspondentes valores mais elevados do coeficiente de correlação.

Com respeito ao teste dos diferentes esquemas de validação cruzada, este trabalho demonstrou a necessidade de se utilizar o período de calibração (utilizado para estimar os coeficientes do modelo de regressão) próximo ao período de teste de forma a maximizar a acurácia. O trabalho também demonstrou a necessidade de se evitar dados concomitantes entre os períodos de calibração e validação. Tal concomitância pode vir a apresentar uma diminuição significativa na acurácia da metodologia. Por fim, vale ressaltar que os resultados desse experimento contribuem com o tema frente à escassez de publicações nessa área e podem ser empregados tanto com vistas à reanálise, quanto à previsão de variáveis atmosféricas.

Este trabalho apresenta uma primeira abordagem sobre a seleção objetiva dos pontos da malha do GCM baseada no ordenamento com vistas ao *downscaling* de variáveis atmosféricas. Embora os resultados sejam promissores, essa metodologia apresenta limitações. Dessa forma, sugerem-se futuras ações com vistas a mitigar tais limitações:

- Testar o desempenho da Seleção Baseada no Ordenamento em conjunto com métodos regressivos não lineares (e.g., Redes Neurais Artificiais) (HAYKIN, 2001). Além disso, testar a possibilidade de selecionar as variáveis regressoras fazendo uso de métodos regressivos lineares e utilizar essas variáveis como entrada para um método regressivo não linear, reduzindo o esforço computacional da seleção de variáveis regressoras;
- Testar novas funções de ordenamento que levem em consideração relações não lineares entre as variáveis do domínio de busca e o preditando, como por exemplo a Informação Mútua Normalizada (KRASKOV *et al.*, 2004);
- Testar funções de ordenamento que levem em consideração a não estacionariedade das séries do domínio de busca e do preditando (ZEBENDE, 2011; VASSOLER e ZEBENDE, 2012);
- Aprofundar o estudo com respeito às razões pelas quais a metodologia aqui proposta apresentou melhores resultados nos locais de orografia simples próximos à costa.

REFERÊNCIAS

ABEEÓLICA. **Boletim anual de geração eólica 2019**. Associação Brasileira de Energia Eólica. [S.l.], p. 15. 2020.

ACCADIA, C. et al. Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. **Weather and forecasting**, v. 18, n. 5, p. 918-932, 2003.

AQUINO, L. **Downscaling dinâmico do vento em superfície baseado em parametrizações da camada limite planetária no Nordeste brasileiro**. Universidade Federal de Pernambuco. Recife, p. 96. 2017.

BAGHANAM, A. H. et al. Conjunction of wavelet-entropy and SOM clustering for multi-GCM statistical downscaling. **Hydrology Research**, v. 50, n. 1, p. 1-23, 2019. ISSN 10.2166/nh.2018.169.

BÁRDOSSY, A.; CASPARY, H. J. Detection of climate change in Europe by analyzing European atmospheric circulation patterns from 1881 to 1989. **Theoretical and Applied Climatology**, v. 43, n. 3, p. 155-167, Stembro 1990.

BARNETT, T. P.; PREISENDORFER, R. Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis. **Monthly Weather Review**, v. 15, n. 9, p. 1825-1850, Setembro 1987.

BELLONE, E.; HUGHES, J. P.; GUTTORP, P. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. **Climate Research**, v. 15, n. 1-12, p. 1, Maio 2000.

BERRISFORD, P. et al. **ERA report series: The ERA-Interim archive**. European Centre for Medium Range Weather Forecasts. Reading, p. 27. 2011.

COSTA, A. **Mathematical/statistical and physical/meteorological models for short-term prediction of wind farms output**. Universidad Politécnica de Madrid. [S.l.]. 2005.

CURRY, C. L.; VAN DER KAMP, D.; MONAHAN, A. H. Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. I. Predicting wind speed. **Climate Dynamics**, v. 38, n. 7-8, p. 1281-1299, 2012.

D'ONOFRIO, A.; BOULANGER, J. -P.; SEGURA, E. C. CHAC: a weather pattern classification system for regional climate downscaling of daily precipitation. **Climatic Change**, v. 98, n. 3, p. 405-427, 2010.

DEE, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. **Quarterly Journal of the Royal Meteorological Society**, v. 137, n. 656, p. 553-597, 2011.

DUPRÉ, A. et al. Sub-hourly forecasting of wind speed and wind energy. **Renewable Energy**, v. 145, p. 2373-2379, 2020. ISSN 10.1016/j.renene.2019.07.161.

EISCHEID, J. K. et al. The Quality Control of Long-Term Climatological Data Using Objective Data Analysis. **Journal of Applied Metorology**, v. 34, n. 12, p. 2787–2795, 1995. ISSN 0894-8763.

ENKE, W.; SCHNEIDER, F.; DEUTSCHLÄNDER, T. A novel scheme to derive optimized circulation pattern classifications for downscaling and forecast purposes. **Theoretical and Applied Climatology**, v. I, n. 82, p. 51-63, 2005. ISSN 10.1007/s00704-004-0116-x.

EPE. **Balanço Energético Nacional**. Empresa de Pesquisa Energética. Rio de Janeiro, p. 264. 2020.

ESTEVEZ, P. A. et al. Normalized Mutual Information Feature Selection. **IEEE Transactions on Neural Networks**, v. 20, n. 2, p. 189 - 201, 2009. ISSN 10.1109/TNN.2008.2005601.

FOWLER, H. J.; BLENKINSOP, S.; TEBALDI, C. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. **International Journal of Climatology**, n. 27, p. 1547–1578, 2007.

GAO, L.; SCHULZ, K.; BERNHARDT, M. Statistical Downscaling of ERA-Interim Forecast Precipitation Data in Complex Terrain Using LASSO Algorithm. **Advances in Meteorology**, v. 2014, p. 1-16, 2014. ISSN 1687-9309.

GARCÍA-HINDE, O. et al. **Feature selection in Solar Radiation prediction using Bootstrapped SVRs**. 2016 IEEE Congress on Evolutionary Computation (CEC). [S.l.]: [s.n.]. 2016. p. 3638-3645.

GILCHRIST, B.; CRESSMAN, G. P. An Experiment in Objective Analysis. **Tellus**, v. 6, n. 4, p. 309-318, 1954. ISSN 0040-2826.

GRELL, G. A.; DUDHIA, J.; STAUFFER, D. R. **A description of the fifth-generation Penn State/NCAR mesoscale model (MM5)**. NCAR. 1995, p. 128. 1995.

GUO, Y.; LI, J.; LI, Y. A time-scale decomposition approach to statistically downscale summer rainfall over North China. **Journal of Climate**, v. 25, n. 2, p. 572-591, 2012.

GUSTAFSSON, N. A Review of Methods for Objective Analysis. In: BENGTSSON, L.; GHIL, M.; KÄLLÉN, E. **Dynamic Meteorology:** Data Assimilation Methods. New York: Springer, v. 36, 1981. p. 17-76.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research**, v. 3, p. 1157-1182, 2003.

HAMMAMI, D. et al. Predictor selection for downscaling GCM data with LASSO. **JOURNAL OF GEOPHYSICAL RESEARCH**, v. 117, n. D17, p. 1-11, 2012.

HAYKIN, S. **Redes neurais:** princípios e prática. 2nd. ed. [S.l.]: Bookman, 2001.

HESSAMI, M. et al. Automated regression-based statistical downscaling tool. **Environmental Modelling and Software**, v. 23, n. 6, p. 813-834, 2008. ISSN 13648152.

HEWITSON, B. C.; CRANE, R. G. Self-organizing maps: applications to synoptic climatology. **Climate Research**, v. 22, n. 1, p. 13-26, Agosto 2002. ISSN 1616-1572.

HEWITSON, B.; CRANE, R. Climate downscaling: techniques and application. **Climate Research**, v. 7, p. 85 - 95, Novembro 1996. ISSN 0936-577X.

HOFER, M. et al. Empirical-statistical downscaling of reanalysis data to high-resolution air temperature and specific humidity above a glacier surface (Cordillera Blanca, Peru). **Journal of Geophysical Research Atmospheres**, v. 115, n. 12, p. 1-15, Junho 2010.

HORTON, P. et al. Spatial relationship between the atmospheric circulation and the precipitation measured in the western Swiss Alps by means of the analogue method. **Natural Hazards and Earth System Science**, v. 12, n. 3, p. 777-784, 2012.

HUTH, R. Properties of the circulation classification scheme based on the rotated principal component analysis. **Meteorology and Atmospheric Physics**, v. 59, n. 3-4, p. 217-233, Setembro 1996.

HUTH, R. A circulation classification scheme applicable in GCM studies. **Theoretical and Applied Climatology**, v. 67, n. 1-2, p. 1-18, Outubro 2000.

HUTH, R. Sensitivity of Local Daily Temperature Change Estimates to the Selection of Downscaling Models and Predictors. **Journal of Climate**, v. 17, n. 3, p. 640-652, Fevereiro 2004.

JOLLIFFE, I. T. **Principal Component Analysis**. 1st. ed. New York: Springer Science+Business Media, 1986.

KALNAY, E. et al. The NCEP/NCAR 40-year reanalysis project. **Bulletin of the American Meteorological Society**, Washington, v. 77, n. 3, p. 437-471, Março 1996.

KANNAN, S.; GHOSH. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. **Stochastic Environmental Research and Risk Assessment**, v. 25, n. 4, p. 457-474, Maio 2011.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by Simulated Annealing. **Science**, v. 220, n. January 1983, p. 671-680, 1983. ISSN 9789812799371.

KOHAVI, R.; JOHN, G. H. Wrappers for Feature Subset Selection. **Artificial Intelligence**, v. 97, n. 1-2, p. 273-324, 1997. ISSN 00043702.

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. **Physical Review**, v. 69, n. 6, p. 1-16, 2004.

LABORATÓRIO de Mecância dos Fluidos - UFPE. Disponivel em: <https://www.ufpe.br/mecfluamb>. Acesso em: 7 dez. 2017.

LANDBERG, L.; WATSON, S. Short-term prediction of local wind conditions. **Boundary-Layer Meteorology**, v. 70, n. 1-2, p. 171-195, Julho 1994.

LANDMAN, W. A. et al. Statistical downscaling of GCM simulations to Streamflow. **Journal of Hydrology**, v. 252, n. 1-4, p. 221-236, Outubro 2001.

MICHAELSEN, J. Cross-Validation in Statistical Climate Forecast Models. **Journal of Climate and Applied Meteorology**, v. 26, n. 11, p. 1589-1600, 1987.

MORAES, C. F. W. D. C. B. **Procedimento objetivo para garantia de qualidade de dados observacionais de vento em superfície no litoral do Rio Grande do Norte**. Universidade Federal de Pernambuco. Recife. 2015.

MURPHY, J. An Evaluation of Statistical and Dynamical Techniques for Downscaling Local Climate. **Journal of Climate**, v. 12, n. 8, p. 2256-2284, 1999.

ONS. Operador Nacional do Sistema Elétrico. Disponivel em: <http://ons.org.br/>. Acesso em: 7 Dezembro 2017.

ORLANSKI, I. A Rotional Subdivision of Scales for Atmospheric Processes. **American Meteorological Society**, v. 56, p. 527 - 530, Maio 1975. ISSN 0003-0007.

PANOFSKY, H. A.; DUTTON, J. A. **ATMOSPHERIC TURBULENCE:** Models and Methods for Engineering Applications. New York: John Wiley and Sons, 1987.

PEDLOSKY, J. The Equations for Geostrophic Motion in the Ocean. **Journal of Physical Oceanography**, v. 14, p. 448-455, Fevereiro 1984.

PROJETO SONDA. SONDA, 2018. Disponivel em: <http://sonda.ccst.inpe.br/>. Acesso em: Janeiro 2018.

RADANOVICS, S. et al. Optimising predictor domains for spatially coherent precipitation downscaling. **Hydrology and Earth System Sciences**, v. 17, n. 10, p. 4189-4208, 2013.

RANABOLDO, M.; GIEBEL, G.; CODINA, B. Implementation of a Model Output Statistics based on meteorological variable screening for short-term wind power forecast. **Wind Energy**, v. 16, p. 811–826, 2013. ISSN 10.1002/we.1506.

REN21. **Renewables 2020: Global Status Report**. REN21. [S.l.], p. 367. 2020.

ROBERTSON, A. W.; KIRSHNER, S.; SMYTH, P. Downscaling of Daily Rainfall Occurrence over Northeast Brazil Using a Hidden Markov Model. **Journal of Climate**, v. 17, n. 22, p. 4407-4424, Novembro 2004.

SAILOR, D. J. et al. A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change. **Renewable Energy**, v. 19, n. 3, p. 960-1481, março 2000. ISSN 0960-1481.

SAUTER, T.; VENEMA, V. Natural Three-Dimensional Predictor Domains for Statistical Precipitation Downscaling. **Journal of Climate**, v. 24, p. 6132-6145, 2011.

SEINFRA-CE. **Secretaria de Infraestrutura do Estado do Ceará**. Disponivel em: <http://www.seinfra.ce.gov.br/index.php/downloads/category/6-energia>. Acesso em: 7 Dezembro 2017.

TAYLOR, K. E. Summarizing multiple aspects of model performance in a single diagram. **Journal of Geophysical Research**, v. 106, n. D7, p. 7183-7192, Abril 2001. ISSN 2156-2202.

TEEGAVARAPU, R. S. V.; GOLY, A. Optimal Selection of Predictor Variables in Statistical Downscaling Models of Precipitation. **Water Resources Management**, v. 32, n. 6, p. 1969-1992, 2018. ISSN 15731650.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267-288, 1996. ISSN 0035-9246.

TIMBAL, B.; MCAVANEY, B. J. An analogue-based method to downscale surface air temperature: application for Australia. **Climate Dynamics**, v. 17, n. 12, p. 947-963, Setembro 2001.

VAN DEN DOOL, H. M. A New Look at Weather Forecasting through Analogues. **Monthly Weather Review**, v. 117, n. 10, p. 2230-2247, Maio 1989.

VAN DER KAMP, D.; CURRY, C. L.; MONAHAN, A. H. Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. II. Predicting wind components. **Climate Dynamics**, v. 38, n. 7-8, p. 1301-1311, Abril 2012. ISSN https://doi.org/10.1007/s00382-011-1175-1.

VASSOLER, R. T.; ZEBENDE, G. F. DCCA cross-correlation coefficient apply in time series of air. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 7, p. 2438-2443, 2012.

VON STORCH, H.; ZORITA, E.; CUBASCH, U. Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. **Journal of Climate**, v. 6, n. 6, p. 1161-1171, Junho 1993.

VON STORCH, H.; ZWIERS, F. W. **Statistical Analysis in Climate Research**. 1ª. ed. [S.l.]: Cambridge University Press, 2003.

WETTERHALL, F.; HALLDIN, S.; XU, C.-Y. Statistical precipitation downscaling in central Sweden. **Journal of Hydrology**, v. 306, n. 1-4, p. 174-190, 2004.

WILBY, R. L. et al. **Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods**. Intergovernmental Panel on Climate Change. [S.l.], p. 27. 2004.

WILBY, R. L.; WIGLEY, T. M. L. Downscaling general circulation model output: a review of methods and limitations. **Progress in Physical Geography**, v. 21, p. 530 - 548, Dezembro 1997. ISSN 0309-1333.

WILKS, D. S. **Statistical Methods in the Atmospheric Sciences**. 4ª. ed. Cambridge: Elsevier, 2019.

YANG, C.; WANG, N.; WANG, S. A comparison of three predictor selection methods for statistical downscaling. **International Journal of Climatology**, v. 37, n. 3, p. 1238-1249, 2017. ISSN 10970088.

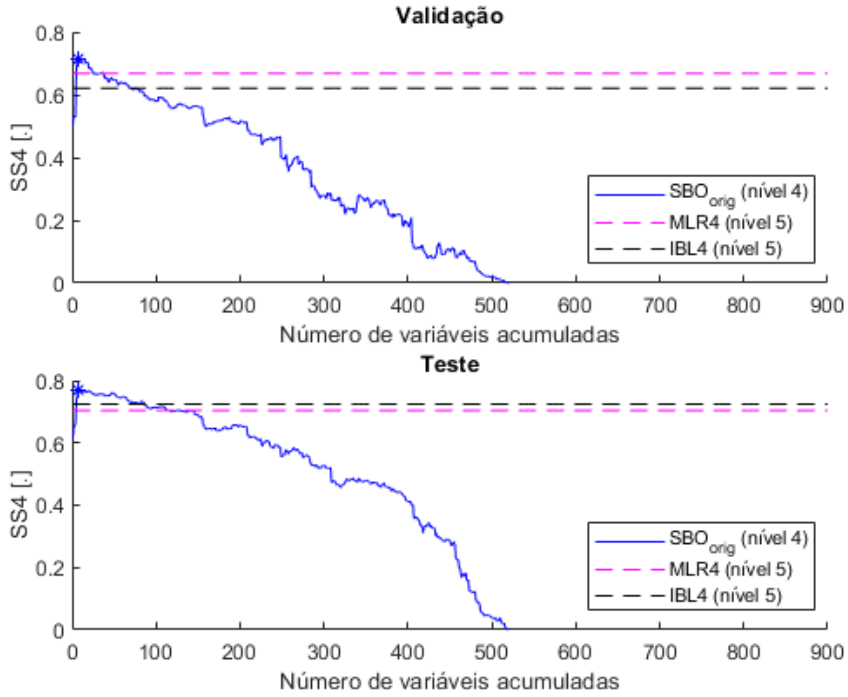
ZEBENDE, G. F. DCCA cross-correlation coefficient: quantifying level of cross-correlation.

Physica A: Statistical Mechanics and its Applications, v. 390, n. 4, p. 614-618, 2011.

APÊNDICE A - CURVAS DE DESEMPENHO DO CONJUNTO DE VARIÁVEIS REGRESSORAS ACUMULADAS

Neste apêndice são apresentados os gráficos contendo as curvas de desempenho das variáveis regressoras acumuladas. Para cada um dos locais estudados será mostrada a curva de desempenho de variáveis regressoras acumuladas obtida pela melhor função de ordenamento no nível de modelo do GCM que obteve a melhor performance. O melhor número de variáveis do domínio de busca acumuladas está marcado com um asterisco, no entanto, é importante ressaltar que a decisão sobre o melhor número de variáveis acumuladas foi tomada levando-se em consideração apenas os dados contidos no conjunto de calibração.

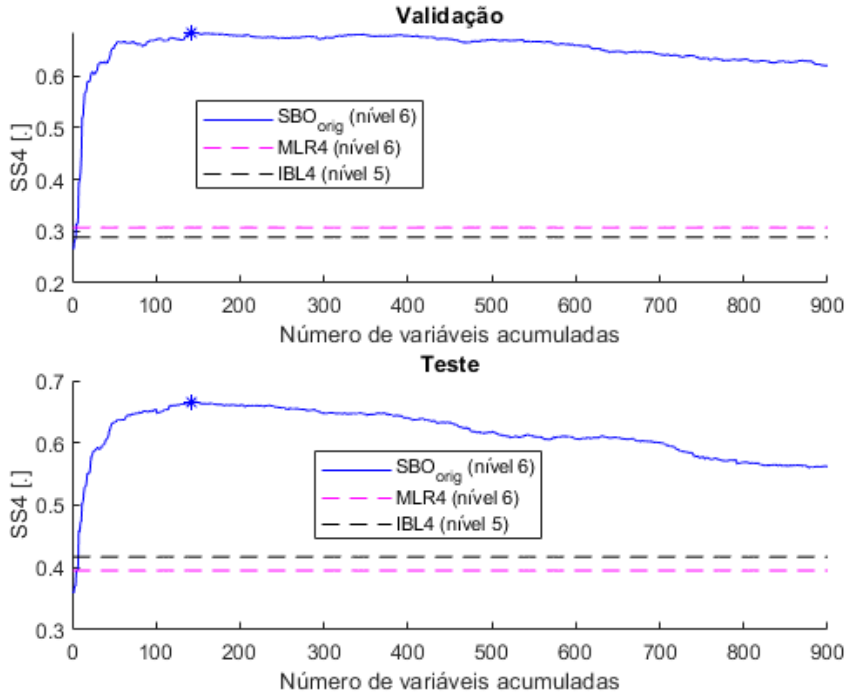
Figura A1 – Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 1)



Curva obtida fazendo uso da função de ordenamento Ph

Fonte: O autor (2020).

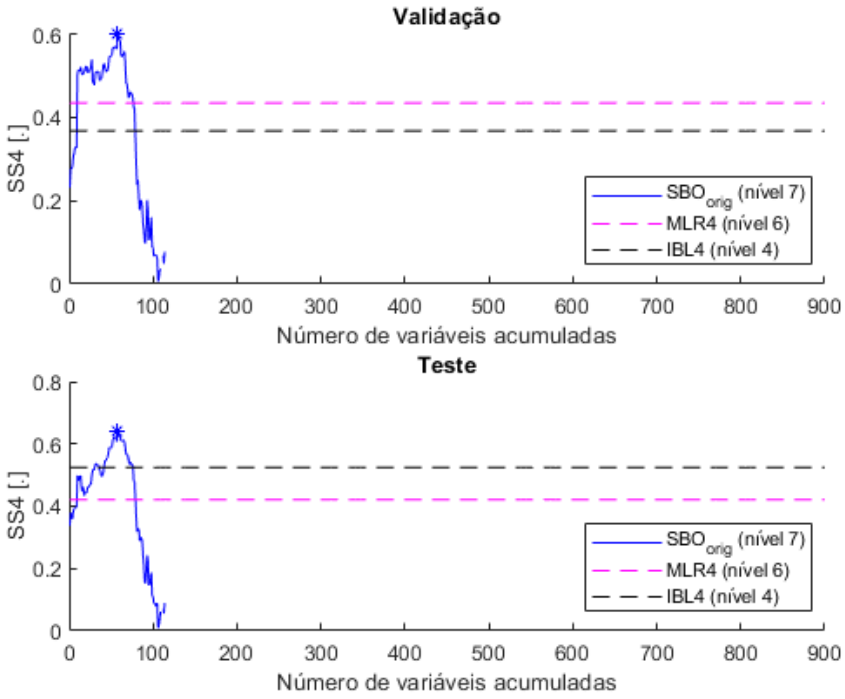
Figura A2 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 2)



Curva obtida fazendo uso da função de ordenamento Ph

Fonte: O autor (2020).

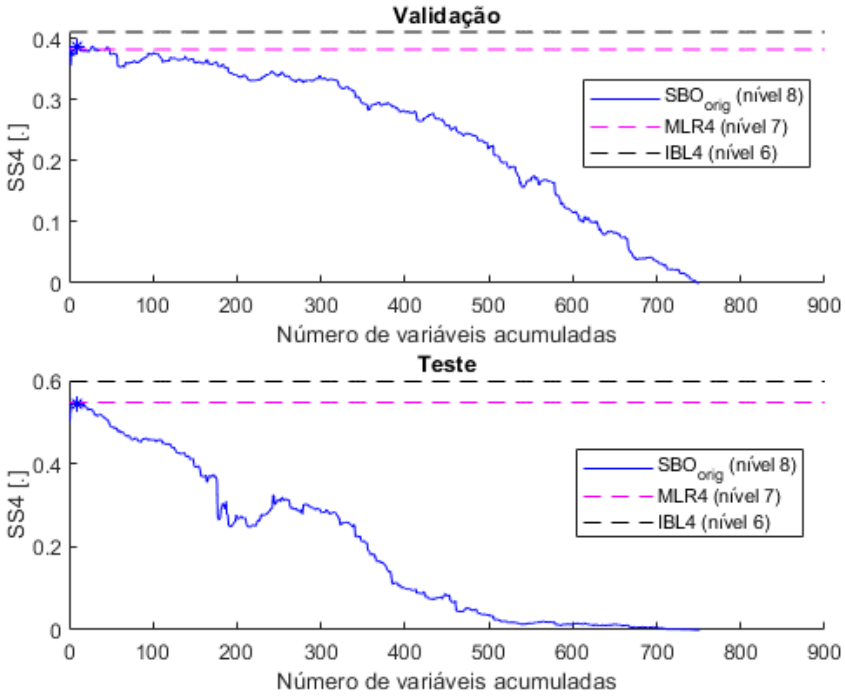
Figura A3 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 3)



Curva obtida fazendo uso da função de ordenamento Ph

Fonte: O autor (2020).

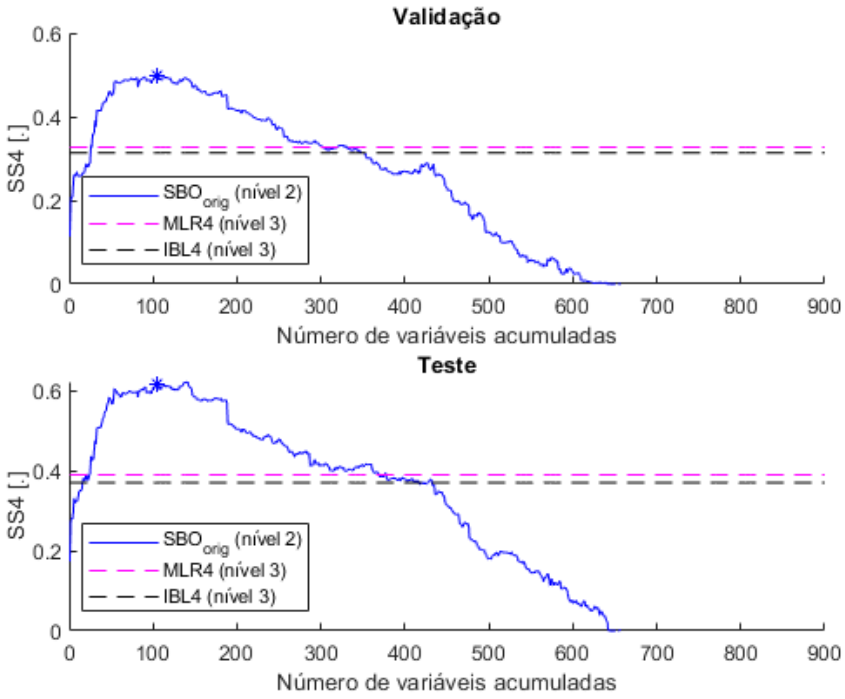
Figura A4 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 4)



Curva obtida fazendo uso da função de ordenamento Ph

Fonte: O autor (2020).

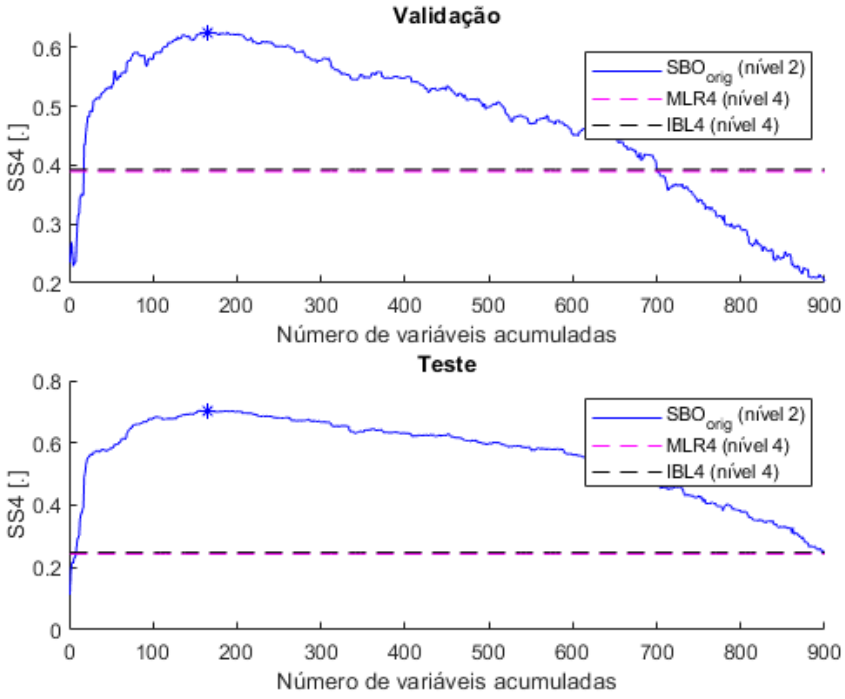
Figura A5 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 5)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020).

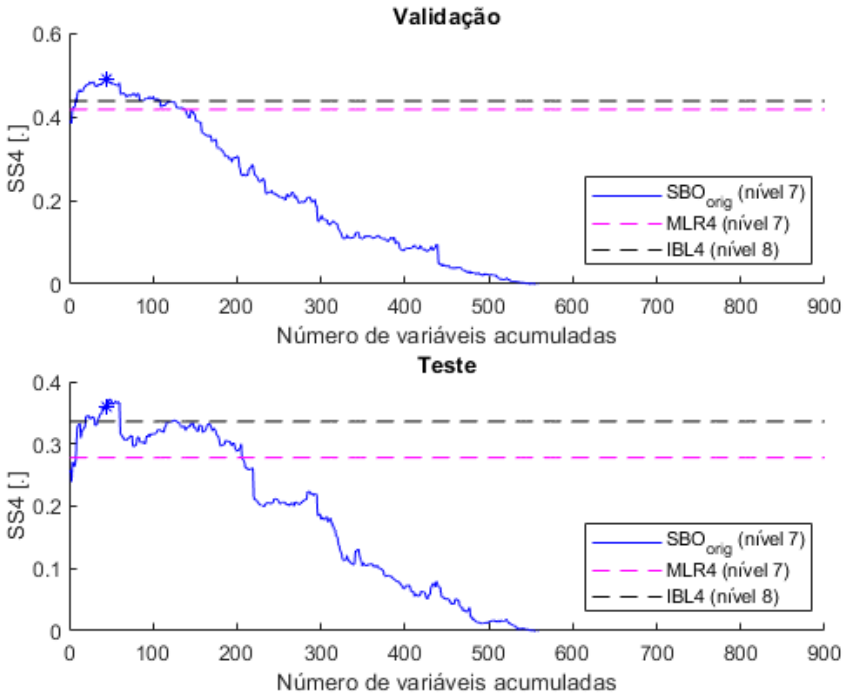
Figura A6 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 6)



Curva obtida fazendo uso da função de ordenamento Ph

Fonte: O autor (2020).

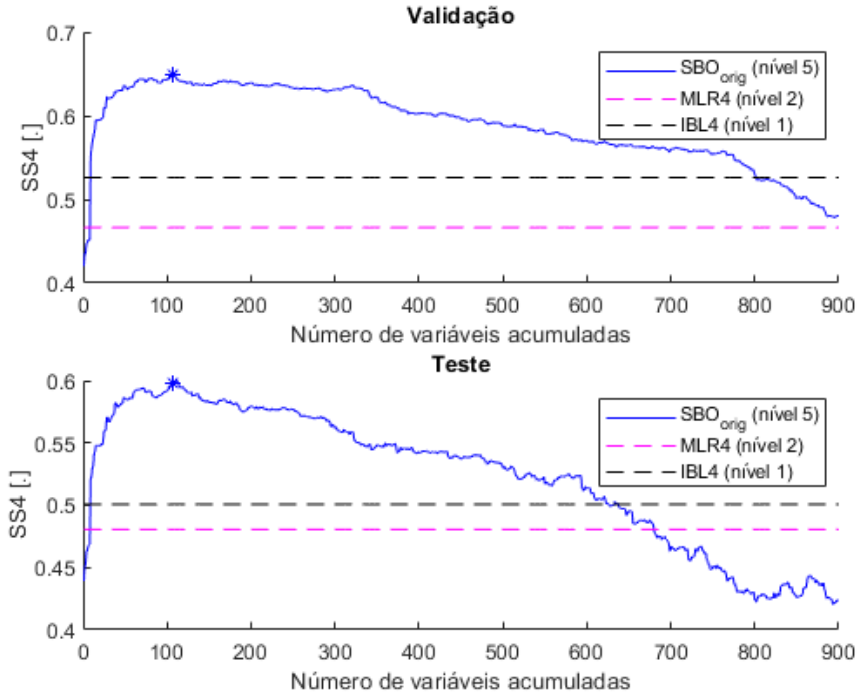
Figura A7 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 7)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020).

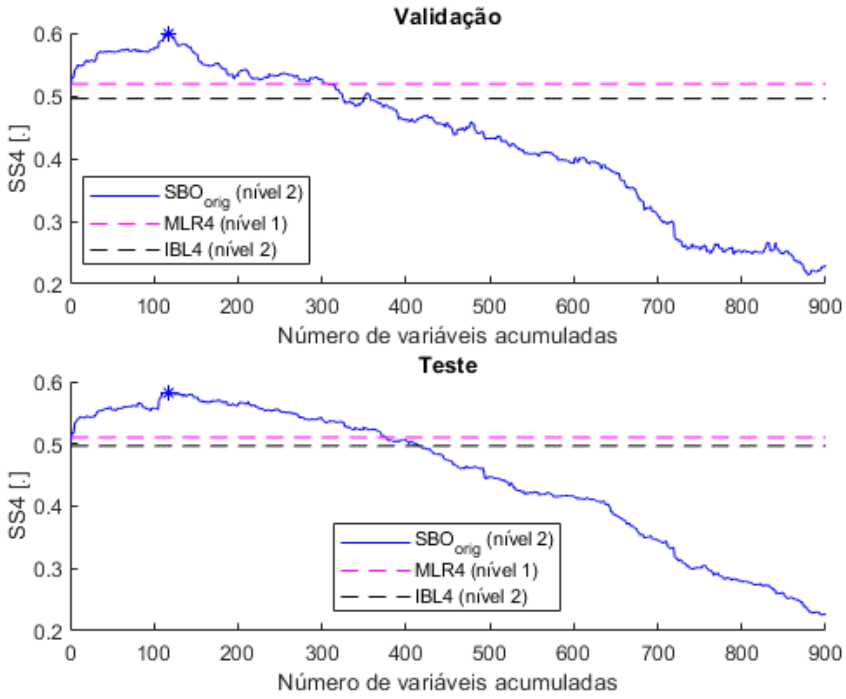
Figura A8 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 8)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020).

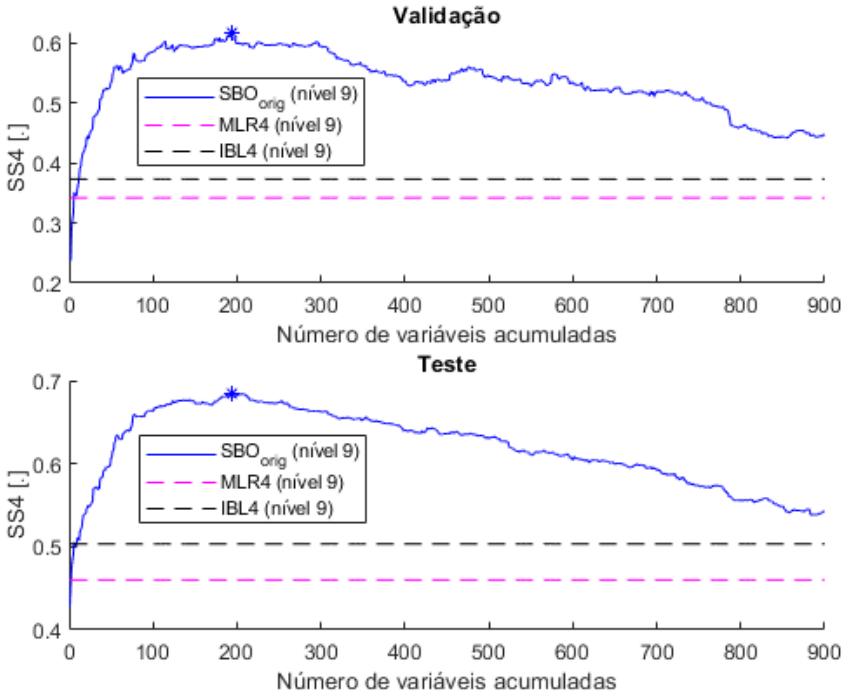
Figura A9 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 9)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020).

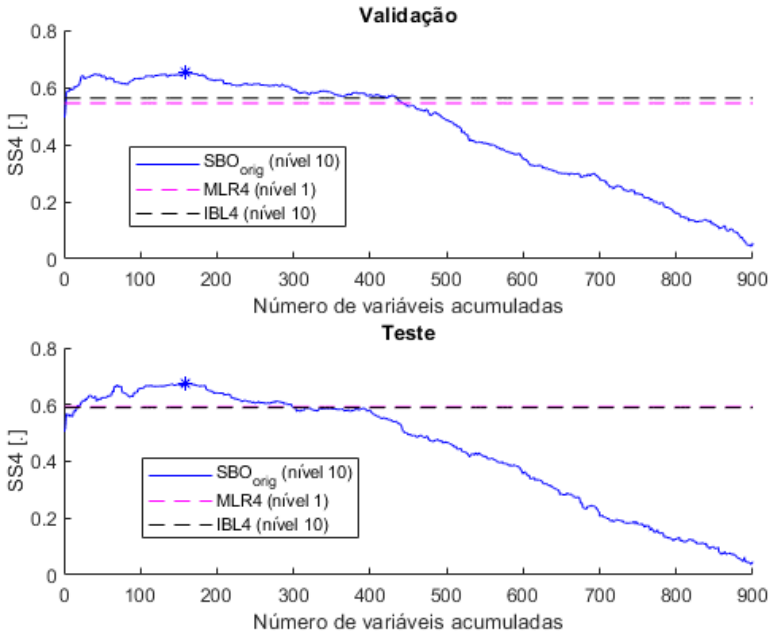
Figura A10 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 10)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020)

Figura A11 - Curva de desempenho do conjunto de variáveis regressoras acumuladas (Local 11)



Curva obtida fazendo uso da função de ordenamento Co

Fonte: O autor (2020).

APÊNDICE B – UTILIZAÇÃO DE COMPONENTES PRINCIPAIS COMO ENTRADA PARA O MÉTODO DE SELEÇÃO DE VARIÁVEIS REGRESSORAS

As variáveis macroescalares utilizadas no *downscaling* usualmente possuem alto índice de correlação entre si, e esse fato implica em diversos problemas no uso de métodos regressivos (VON STORCH e ZWIERS, 2003). Com vistas a reduzir o grau de colinearidade entre as variáveis que compõem o domínio de busca, isso é, tornar as variáveis ortogonais entre si, foram calculadas as componentes principais dessas variáveis (JOLLIFFE, 1986) e o método de seleção de variáveis regressoras foi aplicado utilizando como entrada essas componentes principais.

Inicialmente, utilizando os dados presentes no conjunto de calibração, foram calculados os autovetores da matriz de covariância das *m* variáveis que compõem o domínio de busca. Essa matriz de autovetores será a matriz de rotação do espaço com vistas a criar as componentes principais.

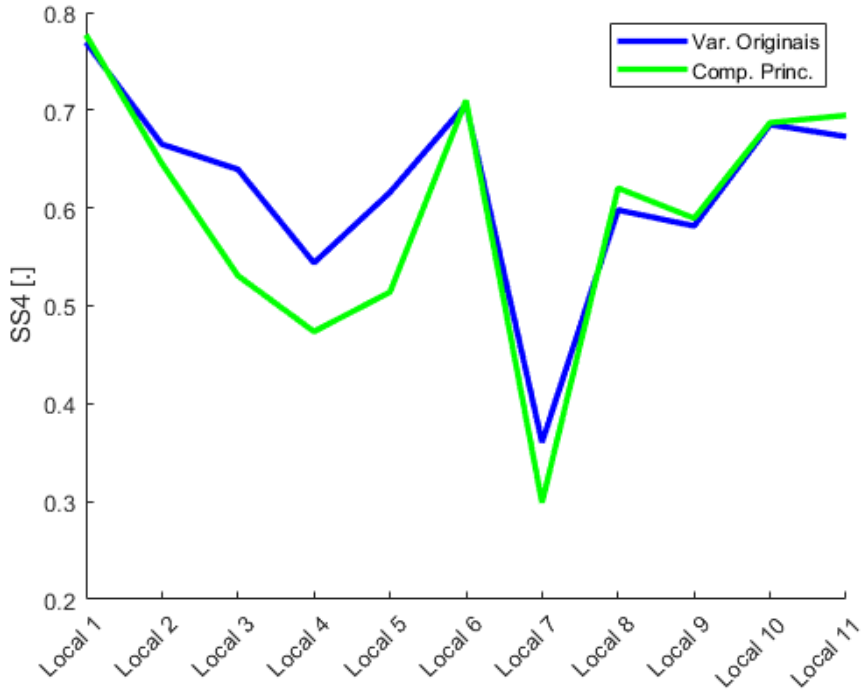
Neste estudo, o intuito da utilização das componentes principais não é a redução de dimensionalidade, mas sim a criação de variáveis ortogonais. Logo, nenhuma das componentes será eliminada, i.e., todas as componentes serão utilizadas como entrada para o método de seleção de variáveis regressoras.

Como, por princípio, não podemos levar em consideração os dados presentes no período de teste para inferir parâmetros do modelo, a matriz de rotação estimada utilizando os dados presentes no conjunto de calibração foi aplicada na rotação das variáveis do domínio de busca no período de teste.

A matriz de rotação estimada no período de calibração também foi aplicada ao período de validação, pois caso uma nova matriz viesse a ser utilizada no período de validação, as decisões tomadas no período de validação não seriam coerentes com o período de teste, diminuindo drasticamente a acurácia da seleção de variáveis regressoras.

Na Figura **55** são comparadas as melhores performances da SBO quando são utilizadas as variáveis originais do domínio de busca como entrada e quando são utilizadas as componentes principais do domínio de busca. Isso quer dizer que, para cada local foi adotada a acurácia obtida pela melhor função de ordenamento no melhor nível de modelo do GCM.

Figura 55 – Comparação entre a acurácia utilizando variáveis originais do domínio de busca e componentes principais do domínio de busca como entrada para a SBO



Fonte: O autor (2020)

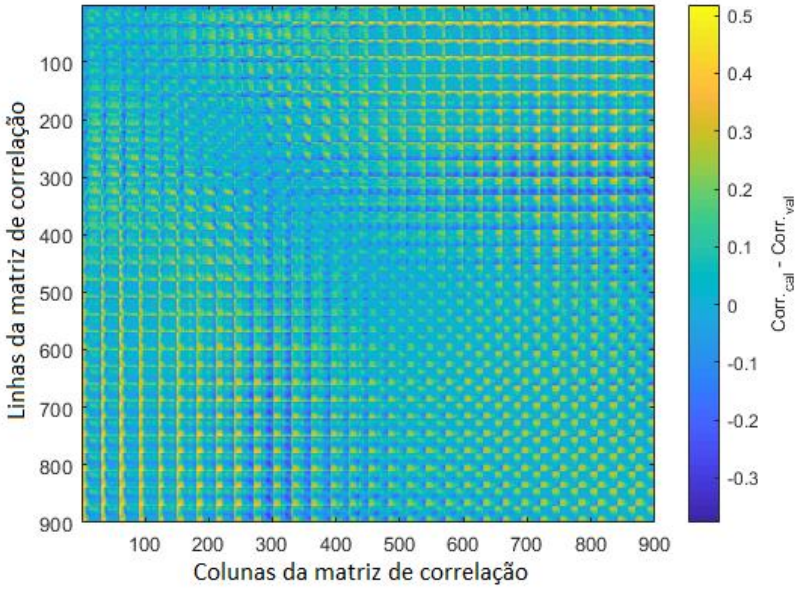
É possível perceber que ao utilizarmos componentes principais a acurácia da seleção baseada no ordenamento se torna inferior àquele apresentado quando utilizamos diretamente as variáveis do domínio de busca. Isso provavelmente se deve à não ortogonalidade das componentes principais nos períodos de validação e teste.

A não ortogonalidade é oriunda de diferenças na matriz de correlação (ou covariância) das variáveis do domínio, isso é, os autovetores da matriz de correlação (ou covariância) do período de calibração estão sendo empregados nos períodos de validação e teste com vistas a rotacionar todas as variáveis de forma a criar componentes principais. Contudo, quanto maior for a diferença entre as matrizes de correlação do período de calibração e dos períodos de validação e teste, maior será o erro associado a essa rotação, e por isso, mais distante da ortogonalidade essas novas variáveis estarão, ou seja, maior será a quantidade de informações redundantes, levando a um grande erro na combinação linear, pois os coeficientes da regressão linear foram estimados no período de calibração, onde todas as variáveis regressoras são ortogonais entre si.

As Figuras B2 e B3 apresentam a diferença na matriz de correlação entre os períodos de calibração e validação e entre os períodos de calibração e teste, respectivamente. Essa diferença ocasiona a falta de ortogonalidade ao se aplicar a matriz de rotação do período de

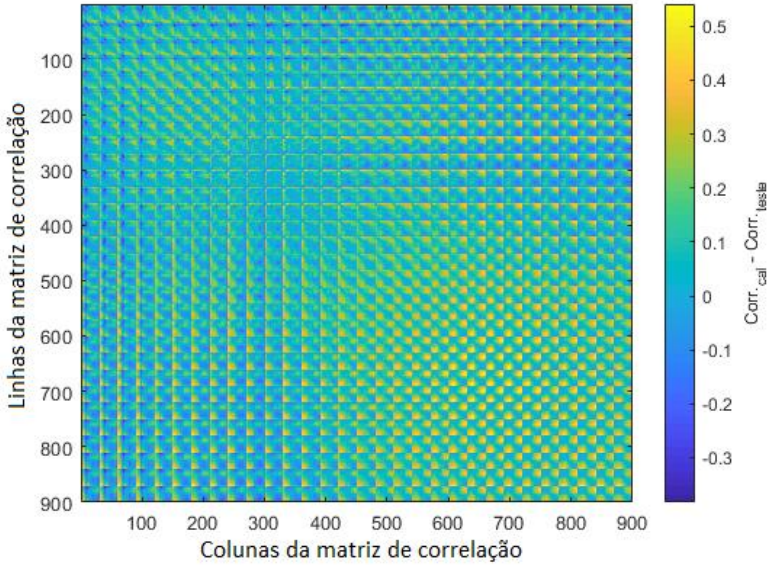
calibração nos demais períodos, como foi discutido no parágrafo anterior. No caso apresentado nas Figuras B2 e B3 a diferença entre as matrizes de correlação atinge valores altos, da ordem de 50% de correlação, provando a perda de ortogonalidade das componentes principais entre os períodos de calibração e validação e calibração e teste.

Figura B2 – Diferença entre as matrizes de correlação das variáveis do domínio de busca entre os períodos de calibração e validação (Local 4, nível 7)



Fonte: O autor (2020)

Figura B3 - Diferença entre as matrizes de correlação das variáveis do domínio de busca entre os períodos de calibração e teste (Local 4, nível 7)



Fonte: O autor (2020)

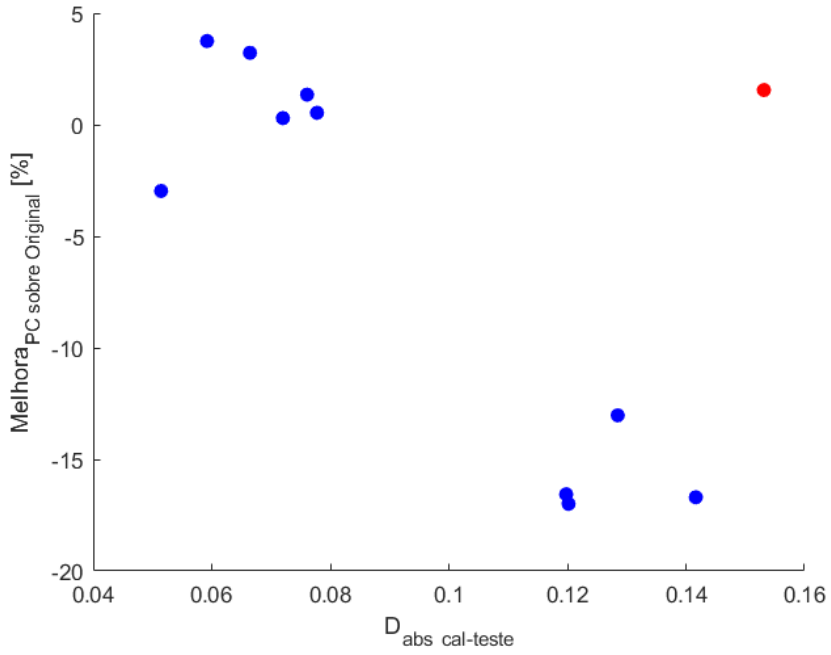
A Figura apresenta a dispersão da melhora na acurácia através do uso de componentes principais como entrada para a SOB (ver Equação 18) em função da média da diferença absoluta entre as matrizes de correlação das variáveis do domínio de busca nos períodos de calibração e teste (ver Equação 19).

$$Melhora_{PC\ sobre\ Original} = \frac{SS4_{PCA} - SS4_{Original}}{SS4_{Original}} \tag{18}$$

$$D_{abs\,cal-teste} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\rho_{cal_{ij}} - \rho_{teste_{ij}}\right| \tag{19}$$

Onde $SS4_{PCA}$ é o SS4 obtido pela SBO quando são utilizadas componentes principais do domínio de busca como entrada, $SS4_{Original}$ é o SS4 obtido pela SBO utilizando diretamente as variáveis do domínio de busca como entrada, $\rho_{cal_{ij}}$ e $\rho_{teste_{ij}}$ são os valores dos elementos na posição *ij* das matrizes de correlação entre as variáveis do domínio de busca no período de calibração e teste, respectivamente.

Figura B4 - Relação entre a acurácia da SBO e a diferença na matriz de correlação



Em azul são apresentados os resultados referentes aos locais 2 a 11, em vermelho é apresentado o resultado referente ao local 1.

Fonte: O autor (2020).

Através da Figura é perceptível uma relação entre a diferença na matriz de correlação e a acurácia obtida pela SBO quando utilizadas componentes principais do domínio como entrada. A única exceção é o Local 1 (ponto vermelho na Figura), que apresenta boa acurácia mesmo com uma grande diferença entre as matrizes de correlação.

O resultado observado no Local 1 demanda mais investigações. No entanto, considerando que a matriz de correlação (ou covariância) depende da magnitude e da fase dos sinais no domínio da frequência, ela será afetada sempre que algumas variáveis do domínio de busca apresentarem, entre dois períodos, comportamento diferente das demais. Sendo assim, a utilização de componentes principais como entrada no método de seleção de variáveis regressoras proposta neste trabalho é bastante limitada, visto que, idealmente, esse método deve utilizar grandes domínios de busca, o que por sua vez, acarretaria numa maior probabilidade de perda de ortogonalidade.