Artificial Intelligence and the Multivariate Approach in Predictive Analysis of the Small Cap Index of the Brazilian Stock Exchange

Bianca Kaczorowski, Mariana Kleina, Marcos A. M. Mendes and Wiliam de A. Silva

***Abstract*—Looking for greater returns, the Brazilian investors have increasingly resorted to the stock exchange. In this way, companies classified as Small Cap appear as a great opportunity for significative profitability given the higher volatility. In this scenario, studies aiming to explain the behavior of the stock market are becoming even more important. This article aims to develop prediction models for the Small Cap Index in order to assess whether the series under study can be explained by economic and financial variables. The monthly data used were collected from October 2005 to December 2019. The methods applied for the development of predictive models were multiple linear regression and the multilayer perceptron artificial neural network with backpropagation supervised learning algorithm. For the first forecasting method, the model presents, based on 10 predicting variables, approximately 91% of the explanatory capacity of the Small Cap Index variability, against approximately 98% of the second method, which is based on 16 variables for the forecast. Although the artificial neural network presents better prediction results, both models are satisfactory to explain the behavior of the index under study.**

***Index Terms*—Small Cap Index; Multiple linear regression; Artificial Neural Network; Prediction.**

I. Introdução

O número de investidores ativos na bolsa de valores brasileira chegou, em dezembro de 2019, a 1,690 milhão [1], um crescimento de 102,4% frente ao mesmo período do ano de 2018. Esses dados são reflexo de inúmeros fatores, dentre os principais pode-se citar a manutenção da taxa básica de juros (Selic), a disseminação da educação financeira pelo Brasil juntamente com o desenvolvimento de novas plataformas por parte das corretoras e assessorias de investimentos e as aprovações de reformas no Congresso Nacional do Brasil, sobretudo da previdência.

A Taxa Selic é um instrumento de política monetária, isto é, sofre alterações estratégicas para que as atividades econômicas sejam estimuladas ou para que a inflação seja esma contida. Em agosto de 2020, o Comitê de Política Monetária (Copom) reduziu a taxa de juros brasileira para 2% ao ano, atingindo seu menor patamar histórico [2].

B. Kaczorowski, Universidade Federal do Paraná, Curitiba, PR, Brasil (biakac@hotmail.com).

M. Kleina, Universidade Federal do Paraná, Curitiba, PR, Brasil (marianakleina@ufpr.br).

M. A. M. Marques, Universidade Federal do Paraná, Curitiba, PR, Brasil (marquesammarcos@gmail.com)

W. A. Silva, Universidade Federal do Paraná, Curitiba, PR, Brasil (wiliamdeassis@gmail.com)

Corresponding author: Bianca Kaczorowski.

Em outubro de 2016, a mesma taxa era de 14,25% e desde então cortes vêm sendo realizados. Essa contínua queda reflete a situação econômica que o país vem apresentando de recuperação da crise enfrentada no ano de 2015.

Ciclos de juros menores estimulam o consumo e reduzem as taxas de empréstimos, em contrapartida também reduzem a rentabilidade de aplicações financeiras indexadas à taxa básica de juros, como cadernetas de poupança e títulos públicos [3]. Em busca de maiores retornos, as pessoas físicas migram de investimentos mais conservadores para investimentos com maior volatilidade e risco, como o mercado acionário.

Nesse contexto, as *small caps*, empresas com menor capitalização listadas na B3 (Brasil, Bolsa, Balcão – a bolsa de valores oficial brasileira), surgem como uma grande oportunidade de investimento dado que as empresas classificadas como tal possuem maior potencial de crescimento e valorização na bolsa, ao mesmo passo que possuem uma menor liquidez, o que acarreta uma maior volatilidade e consequente elevação do risco.

A literatura brasileira acerca do tema é escassa e concentrada em estudos sobre otimização de carteiras por meio da avaliação do desempenho de empresas que fazem parte do Índice *Small Cap*. Entretanto, há diversos trabalhos, nacionais e internacionais, voltados a séries temporais financeiras em geral que afirmam que a sua formação segue o chamado passeio aleatório (*random walk*), ou seja, que não podem ser previstos a partir de seu comportamento passado, elas possuem ruídos e são deterministicamente caóticas e não-estacionárias [4][5].

Desta forma, o presente artigo visa estimar o comportamento da série de dados do Índice *Small Cap* por meio do estudo comparativo da acurácia dos resultados de predição do método de regressão linear multivariada e de um método de inteligência artificial. A escolha do comparativo entre o método estatístico e o método computacional se deu pelo fato da regressão linear ser muito difundida por apresentar intervalos de confiança e o impacto e correlação de cada variável na predição da variável resposta e, ao mesmo tempo, pelo crescimento das aplicações de redes neurais artificiais nos mais diversos meios. O propósito é analisar se o comportamento do mercado acionário no país pode ser explicado pelas diversas variáveis econômico-financeiras e verificar aplicabilidade dos métodos em estudo no caso de séries financeiras com maior volatilidade.

II. Referencial Teórico

A corrente seção visa elucidar conceitualmente temas que

serão aplicados no estudo e está seccionada em 3 partes. São elas: (A) Índice *Small Cap*; (B) Regressão linear múltipla e (C) Redes neurais artificiais.

A. Índice Small Cap

O Índice *Small Cap*, criado em 2005, é um dos diversos índices de retorno total da B3. Classificado como índice de segmento e setoriais, tem por objetivo medir o desempenho médio das cotações dos ativos da carteira composta por empresas de menor capitalização, ou seja, com baixo valor de mercado listadas na bolsa de valores brasileira [6].

A carteira teórica de ativos é elaborada com base na liquidez das empresas e são ponderadas pelo valor de mercado, isto é, a multiplicação entre o número de ações emitidas e a cotação do ativo. A pontuação do índice então, é dada pela soma do valor de mercado dessas empresas [7].

Entre as características comuns a essas empresas têm-se: O menor valor de mercado – variando entre 300 e 2 bilhões de dólares; a menor liquidez – o volume de papéis negociados ao longo do pregão é menor se comparado a outras empresas; a menor cobertura pelo mercado – são menos acompanhadas pelo mercado; possuem maior volatilidade na cotação – a cotação é afetada com maior intensidade pelo movimento de compra ou venda das ações; possuem um maior risco intrínseco – são empresas que possuem mais dívidas pelo fato de financiarem suas respectivas expansões e crescimento, expondo-se às mudanças na economia, a título de exemplo tem-se o aumento da taxa de juros [8].

B. Regressão Linear Múltipla

Devido à complexidade dos fenômenos, o estudo de inúmeras variáveis simultaneamente tornou-se relevante para a tomada de decisões. Dentre os objetivos da aplicação de técnicas multivariadas, os principais são: Redução de dados; classificação e discriminação das variáveis; investigação da relação entre as variáveis; predição; construção de testes de hipóteses [10].

Os resultados derivados da aplicação de técnicas multivariadas são combinações lineares das variáveis aos quais são atribuídos pesos determinados empiricamente. As variáveis devem ser aleatórias e interdependentes, de forma que seus efeitos não possam ser substancialmente interpretados separadamente [9].

Dentre as técnicas multivariadas de análise, a mais difundida é a regressão linear múltipla [9]. O objetivo do método é a elaboração de um modelo matemático que estime a variável resposta a partir de duas ou mais variáveis explicativas [11].

O grau de ajuste do modelo aos dados é determinado pelo coeficiente de determinação ajustado – $\bar{R}^2$, que pertence ao intervalo $[0;1]$, sendo mais próximo de 1 o coeficiente do modelo que melhor se ajusta [12].

1) Redução do número de variáveis

Frente ao grande número de variáveis utilizadas nos estudos em geral, torna-se inviável a manipulação e interpretação do conjunto de dados sem o auxílio de uma ferramenta que ajude a identificar as inter-relações presentes de forma a reduzir a dimensão dos dados sem grande perda de informação.

Portanto, o objetivo principal da análise de componentes principais (ACP) é redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados. As primeiras componentes principais, combinação linear de todas as variáveis originais, respondem pela maior parte da variação total do conjunto de dados original de tal forma que a dimensionalidade dos dados pode ser reduzida [13].

Na determinação do número de componentes principais, aplica-se o método B4 de Jolliffe onde excluem-se variáveis a partir da seleção das componentes principais com autovalores maiores que 0,7, dado que valores abaixo desse limite não geram impacto significativo na variabilidade dos dados em análise. A partir da matriz de correlação entre as componentes principais selecionadas e as variáveis do conjunto de dados original, as variáveis que possuem a maior correlação, em módulo, com cada componente principal devem ser mantidas no estudo. Caso uma variável apresente maior correlação em mais de uma componente principal, a próxima variável com maior correlação deve ser selecionada [14].

2) Seleção das variáveis

Na construção do modelo matemático, quanto maior a quantidade de variáveis, mais dependente dos dados observados o mesmo pode ficar. Nesse contexto, a seleção de variáveis é fundamental para a acurácia do estudo, sendo o método *stepwise* (passo a passo) o mais difundido dentre as abordagens aplicáveis à regressão linear múltipla [15].

O método é operacionalizado computacionalmente e consiste na inclusão e remoção de variáveis baseado na medida de significância estatística do coeficiente associado à variável, selecionando assim somente as variáveis que melhor se adaptam à modelagem [16].

3) Análise de adequação do modelo

De forma a estudar a qualidade e a confiabilidade do ajuste proposto, análises de adequação são propostas. São elas: Análise dos resíduos e multicolinearidade.

Os resíduos representam a diferença entre o valor de Y_j observado e o valor $\hat{Y}_j$ previsto, sendo a principal mensuração do erro do modelo matemático criado. A análise dos resíduos evidencia a conformidade da modelagem às suposições intrínsecas de linearidade, normalidade, homocedasticidade e independência [17].

O pressuposto de linearidade estipula que a relação entre as variáveis deve ser linear, isto é, os modelos de regressão linear preveem valores que se ajustam a uma linha reta [18].

Com o propósito de verificar a existência de alguma relação linear entre a variável resposta e as variáveis preditoras, ou seja, a validade do modelo – medida a partir da significância global da regressão linear múltipla, aplica-se a análise de variância – ANOVA. Para que o critério seja satisfeito, o valor parâmetro (p-valor) deve ser inferior a 0,05 para um nível de confiança de 95% [19].

A suposição de normalidade dos resíduos é fundamental para a assegurar a confiabilidade das predições propostas pelo modelo ajustado. Com a elaboração de um histograma é

possível visualizar se os dados se ajustam a uma curva de distribuição normal de Gauss, entretanto há métodos mais exatos para aferir a normalidade do conjunto de dados [9].

Dentre os principais testes estatísticos utilizados estão Anderson-Darling, Kolmogorov-Smirnov, Ryan-Joiner e Shapiro-Wilk, sendo esse último o teste mais difundido para avaliação da normalidade [20].

Para que o critério da normalidade seja satisfeito, o valor parâmetro (p-valor) deve ser superior a 0,05 para um nível de confiança de 95% [17].

Quando a variância do erro se apresenta constante para observações distintas tem-se homocedasticidade dos resíduos. Isso significa que a variável resposta tem o mesmo grau de variância ao longo do domínio das variáveis preditoras. Na ausência de homocedasticidade, a análise de regressão pode ser invalidada [18].

O gráfico dos resíduos *versus* valores ajustados é uma das ferramentas utilizadas para a verificação da homocedasticidade. Todavia, testes estatísticos também podem ser usados com tal finalidade, dentre eles o teste de Breush-Pagan baseado no teste multiplicador de Lagrange. Caso o p-valor seja superior a 0,05, para um nível de confiança de 95%, tem-se o critério da homocedasticidade satisfeito [17].

Para que o modelo seja validado estatisticamente, é necessário que os resíduos sejam independentes. Tal suposição pode ser determinada por meio do gráfico resíduos *versus* ordem de coleta. Se houver uma tendência no comportamento dos pontos conclui-se que há dependência dos resíduos [9].

Estatisticamente, é frequente o uso do teste Durbin-Watson para determinar a existência de independência dos resíduos. O p-valor deve ser maior que 0,05 para um nível de significância de 95%, para que o critério seja satisfeito [17].

A multicolinearidade, correlação entre as próprias variáveis preditoras, é um problema frequente nos estudos de regressão linear múltipla, pois prejudica a estimação final. Nesse contexto, a variável resposta deve ter uma forte correlação com as variáveis preditoras ao mesmo passo que as variáveis preditoras, entre si, devem ter a menor multicolinearidade possível [21].

Pode-se determinar a multicolinearidade por meio do fator de inflação de variâncias – VIF, que representa o incremento da variância devido à presença da multicolinearidade. Um VIF maior que 10 demonstra indícios de multicolinearidade [21].

C. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são sistemas matemáticos não lineares que reproduzem neurônios unidos por meio de conexões associadas a pesos. As RNAs visam simular o funcionamento do cérebro humano, sendo as mesmas capazes de reconhecer padrões, detectar relações em um conjunto de dados, operar com dados imprecisos, prever séries temporais com elevado grau de precisão e ainda possuem a capacidade de aprendizagem em virtude dos processos iterativos. Nesse contexto, sua aplicação no setor financeiro torna-se extremamente relevante [22].

Os parâmetros que determinam a arquitetura das RNAs são: Número de camadas; número de neurônios existentes em cada camada; tipo de conexão entre os neurônios e topologia da rede [23].

Os neurônios, por sua vez, são compostos por três elementos: Pesos sinápticos; função soma e função de ativação ou transferência, conforme apresentados na Fig. 1 [23].

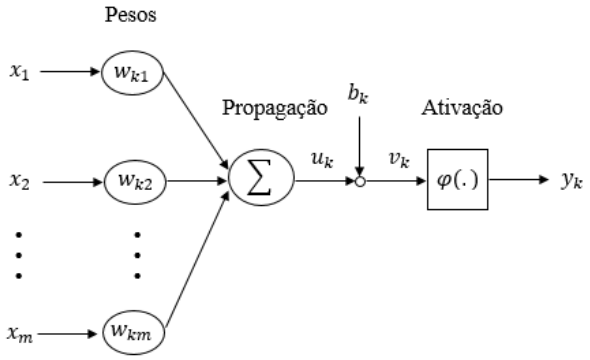

Fig. 1. Funcionamento da rede neural artificial [24].

O processamento das informações ocorre nos neurônios. Sinais x_m são transmitidos entre os k neurônios por meio de sinapses, que possuem um peso w_{km} associado, usado como multiplicador do sinal transmitido. Então é "aplicada uma função de ativação $\varphi(.)$ sobre a soma ponderada u_k dos pesos e a partir do resultado é determinado se o neurônio será ativado tendo um comportamento excitatório ou se terá um comportamento inibitório permanecendo inativo. Este sinal de ativação y_k será propagado para os demais neurônios componentes da RNA de acordo com a topologia e as interconexões da rede" [25]. A variável b_k é o *bias*, uma entrada externa adicionada à função de ativação que tem por objetivo aumentar ou diminuir a entrada líquida da função sendo uma transformação afim v_k.

O aprendizado da RNA, processo pelo qual os pesos são alterados gradativamente, é fundamental para a acurácia das saídas e o método escolhido define a maneira como eles serão modificados. As três principais formas são: Aprendizado supervisionado; aprendizado não-supervisionado; aprendizado por reforço [23].

1) Multilayer perceptron

Existem inúmeros modelos de redes neurais artificias, dentre eles o *perceptron* de múltiplas camadas ou seu termo em inglês *multilayer perceptron* (MLP). Esse modelo é caracterizado pela sua composição, dado que possui, além da camada de entrada e de saída, uma ou mais camadas ocultas. Em comparação aos modelos sem camadas intermediárias, o MLP apresenta melhores resultados.

É comum utilizar para esse modelo o método *feedforward* para a alimentação, ou seja, alimentação para frente dos padrões de entrada e o método de *backpropagation* como forma de treinamento, que é baseado na aprendizagem por correção do erro, ou seja, consiste no cálculo do gradiente do erro para cada vetor de entrada. É um dos principais algoritmos de aprendizagem supervisionada e que, por sua vez, é composto por duas etapas, são elas: Propagação *forward* e propagação *backward*. Na propagação *forward* os pesos sinápticos são fixos e os sinais se propagam pela rede até a saída, ao passo que na propagação *backward* é feita a comparação entre a saída obtida e a desejada – é o sinal de erro, que é então enviado para trás, no sentido contrário das conexões sinápticas, ajustando assim os

pesos de forma a minimizar o erro à medida que a rede aprende [23].

III. METODOLOGIA

A presente seção tem por objetivo descrever a metodologia aplicada na construção dos modelos de previsão do Índice *Small Cap* englobando a seleção do conjunto de dados, redução e seleção das variáveis, otimização dos modelos e avaliação dos resultados, dentre eles os métodos estatísticos aplicados na análise de adequação do modelo de regressão.

A. Seleção das Variáveis

Uma matriz A foi criada a partir da consolidação dos dados coletados mensalmente desde a criação do Índice *Small Cap*, em outubro de 2005, até o último mês do ano de 2019, totalizando 171 observações. São dados referentes aos saldos no final do período ou dados do fechamento do último dia útil do mês, para os que possuem cotações diárias. Para o presente estudo foram selecionadas 90 variáveis dentre indicadores macroeconômicos e microeconômicos brasileiros, dados relacionados às *commodities* – dado que o Brasil é um grande produtor e exportador, índices de volatilidade e de rendimento do tesouro americano – por ser emitido pela maior economia do mundo esse tipo de investimento é considerado o mais seguro e serve como referência para todo o mercado financeiro. O conjunto de dados utilizado no trabalho pode ser encontrado em https://docs.ufpr.br/~marianakleina/SmallCap.html.

O conjunto de dados foi importado no *software* R para realização da análise de componentes principais a partir da matriz de correlação com os dados normalizados – $Z \sim N(0,1)$. Como resultado obtém-se os autovalores, isto é, a proporção de variância explicada e a matriz de autovetores. O número de componentes principais foi definido com base no valor do autovalor, ou seja, apenas componentes principais com autovalor maior que 0,7 foram selecionadas.

Baseado nessas informações foi possível a elaboração da matriz de correlação entre as componentes principais selecionadas e as variáveis originais, fundamental para a aplicação do método B4 de redução de variáveis de Jolliffe, onde são selecionadas as variáveis originais com maior correlação absoluta com as componentes principais.

Uma matriz A' foi criada com os dados reduzidos, isto é, apenas com as 171 observações de cada uma das variáveis selecionadas a partir do método de Jolliffe, e foi importada no *software* R. A mesma matriz foi dividida aleatoriamente em dois conjuntos de dados, o conjunto de treino e o conjunto de teste. O primeiro, com 70% das observações de cada variável, foi utilizado para a modelagem, já o segundo, com os 30% das observações de cada variável restantes, foi utilizado para previsão da variável resposta.

B. Regressão Linear Múltipla

A seleção das variáveis foi realizada por meio do método *stepwise both*, uma combinação entre o método *backward* que inicia com um modelo geral, que inclui todas as variáveis, e depois vai eliminando as variáveis uma a uma de forma a encontrar o modelo ótimo e *forward* que inicia com o modelo mais simples, sem preditor, e vai adicionando as variáveis uma a uma até encontrar o modelo ótimo.

A partir do modelo ótimo gerado, avaliou-se os resultados por meio dos testes estatísticos, coeficiente de determinação ajustado, raiz do erro quadrático médio e multicolinearidade. Para um nível de significância de 95%, os seguintes testes foram realizados: t de *Student* e Teste F – ambos para analisar a significância estatística de cada variável, ANOVA para adequação do critério de linearidade, Anderson-Darling, Shapiro-Wilk e Kolmogorov-Smirnov para adequação do critério de normalidade, Breusch-Pagan para homocedasticidade e por fim Durbin-Watson para independência. O erro e o coeficiente de determinação não foram decisivos para a validação do modelo, mas sim para comparação entre o modelo de regressão e a rede neural artificial.

Em casos de não adequação do modelo aos critérios explicitados anteriormente, análises de valores atípicos, isto é, *outliers*, observações influentes e pontos de alavanca foram realizadas de forma a refinar o conjunto de dados de treino e então reaplicar o método *stepwise* e reavaliar os resultados até encontrar a adequação aos critérios. O método utilizado para identificação dos *outliers* foi os resíduos studentizados, em módulo, superiores a 2, conforme encontra-se comumente na literatura [26].

Com o modelo validado, o conjunto de dados de teste foi utilizado para a predição da variável resposta. Como método de avaliação de desempenho entre os resultados das previsões do modelo de regressão linear múltipla e da rede neural artificial, a raiz do erro quadrático médio entre o valor previsto e o valor original foi calculado.

C. Rede Neural Artificial

Os conjuntos de treino e de teste utilizados para a regressão linear múltipla também foram utilizados para a rede neural artificial. Porém, antes do treinamento da rede e da predição da variável resposta, os dados foram normalizados por meio da técnica min-máx de forma a se trabalhar com dados mais consistentes e homogêneos.

O número de neurônios na camada de entrada é igual ao número de variáveis presentes no estudo após a redução da dimensionalidade do conjunto de dados, ou seja, após a redução das 90 variáveis originais através da ACP e da aplicação do método B4 de Jollife. Já o número de neurônios na camada de saída é igual a um, dado que o objetivo da rede neural é a previsão de uma única variável.

O número de neurônios na camada oculta foi definido empiricamente. Variou-se o número de neurônios entre 1 e 200, selecionando o número de neurônios que gerou o erro quadrático médio mínimo.

A rede foi então treinada a partir da parametrização apresentada na Tabela I e a função utilizada para tal foi a *neuralnet*() do pacote *neuralnet*. O coeficiente de determinação ajustado e o erro quadrático médio foram calculados como forma de comparação com o modelo gerado pela regressão linear múltipla. Posteriormente a variável

resposta foi prevista com base nos valores do conjunto de dados de teste e a partir da configuração final da rede.

TABELA I
ARQUITETURA DA RNA

Tipo	Parâmetro adotado
Topologia da rede	*Multilayer Perceptron*
Algoritmo de alimentação	*Feedforward*
Algoritmo de treinamento	*Backpropagation*
Forma de aprendizado	Supervisionado
Função de ativação	Sigmóide logística
Função de saída	Sigmóide logística
Número de camadas ocultas	1
Threshold	0,01
Critério de parada	100.000

A avaliação dos resultados foi organizada a partir da análise da raiz do erro quadrático médio gerado pelo modelo na predição da variável resposta.

IV. RESULTADOS

A corrente seção tem como propósito apresentar os resultados obtidos a partir da aplicação da metodologia detalhada na seção III e está dividida em 3 partes. São elas: (A) Seleção das variáveis, (B) Regressão linear múltipla e (C) Rede neural artificial.

A. *Seleção das Variáveis*

Na análise de componentes principais, dentre as 90 componentes da matriz A, 16 apresentaram autovalor maior que 0,7. A variância acumulada do conjunto de dados, explicada por essas 16 componentes principais, é de 91,88%, indicando um bom grau de explicação. Isso significa que, as 90 variáveis iniciais, podem ser reduzidas a 16 variáveis com a menor perda de informações possível.

O método B4 de Jolliffe foi então aplicado na matriz de correlação entre as 16 componentes principais selecionadas e as 90 variáveis originais para identificar as 16 variáveis que seguiriam no estudo, são elas: 1) Macroeconômicas: Papel moeda emitido, relação câmbio/salário, Índice Nacional de Preços ao Consumidor – Amplo (IPCA), Índice Geral de Preços (IGP), indicador de movimento de cheques, Índice Nacional de Custos da Construção (INCC), Investimento Direto no Exterior (IDE), desembolsos do sistema BNDES – Setor Industrial, reservas bancárias, Índice de Preços por Atacado – Oferta Global (IPA-OG) – Produtos industriais e taxa de rolagem – Total. 2) Microeconômicas: Produção de aço bruto, valor de mão-de-obra e de materiais de construção (SINAPI) e vendas de veículos no mercado externo. 3) *Commodity*: Cotação de contratos futuros de algodão. 4) Volatilidade: Índice de Volatilidade (VIX).

De forma geral, todas as variáveis selecionadas tendem a influenciar o Índice *Small Cap* por estarem relacionadas diretamente com a economia e impactarem nos resultados financeiros das empresas que compõem o índice. Quando as expectativas do mercado com relação à economia do país são positivas, cresce a probabilidade do índice sofrer uma elevação.

A matriz A' foi então criada com as 171 observações das 16 variáveis selecionadas. Em sequência foi dividida aleatoriamente em dois conjuntos de dados, um de treino e um de teste, utilizados para ambos os modelos de predição.

B. *Regressão Linear Múltipla*

Com base nos dados de treino, o método *stepwise* foi aplicado e, dado que nem todos os critérios de adequação foram satisfeitos, os *outliers* foram retirados dos dados de treino sendo necessário reaplicar o método *stepwise*.

Esse processo foi realizado 11 vezes até que nenhum *outlier* fosse encontrado, porém realizando os testes estatísticos ainda restava um teste estatístico com valor não satisfatório. Como a retirada dos *outliers* pode prejudicar a capacidade de predição do modelo quando situações extremas ocorrem, optou-se por reiniciar as iterações e parar a retirada dos *outliers* quando todos os demais critérios fossem atendidos, com exceção do critério citado anteriormente que de nenhuma forma seria satisfeito, totalizando 6 iterações.

Os resultados dos testes de significância estatística t de *student* e Teste F indicam que, para um nível de significância estatística de 95%, cada variável presente no modelo matemático proposto explica, de maneira significativa, o comportamento do Índice *Small Cap*.

As correlações lineares entre as variáveis são baixas, com VIF bem inferiores a 10, dessa maneira pode-se concluir que nenhuma das variáveis preditoras pode ser explicada por outras variáveis presentes no modelo.

Analisando os resíduos, tem-se primeiramente o critério de linearidade. Segundo o resultado da ANOVA, p-valor menor que 2,2E-16, há uma relação linear entre a variável resposta e as demais variáveis preditoras. Assim sendo, o Índice *Small Cap* pode ser explicado por meio de uma equação linear.

A partir dos resultados do teste Durbin-Watson, p-valor igual a 2,40E-08, observa-se que o critério de independência não foi atendido, ou seja, os erros não possuem um comportamento totalmente independente. Na Fig. 2 é possível notar diversos pontos consecutivos acima e abaixo da linha central, indicando um comportamento sistêmico e comprovando a dependência dos resíduos.

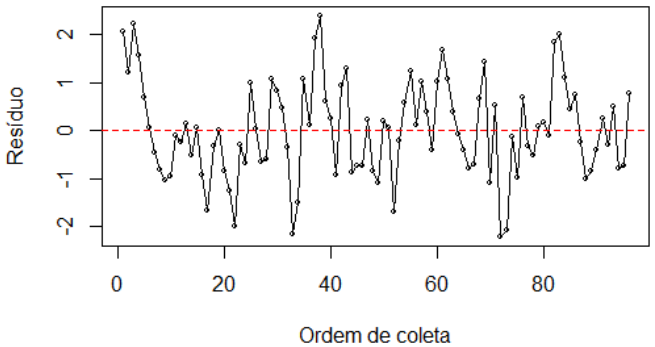

Fig. 2. Gráfico dos resíduos *versus* ordem de coleta – Independência.

De maneira geral, as violações desse critério são devido às influências de dependências temporais e espaciais [9]. Tal dependência pode estar associada à natureza dos dados utilizados – dados mensais de uma longa série temporal, e a determinação e análise dos intervalos de confiança das predições, que podem ser afetados pela independência dos

dados, não fazem parte do escopo do estudo.

Na Fig. 3 e Fig. 4 é possível visualizar que os resíduos se aproximam de uma distribuição normal, confirmando o atendimento do critério de normalidade (Anderson-Darling com p-valor igual a 0,5181; Shapiro-Wilk com p-valor igual a 0,4961; Kolmogorov-Smirnov com p-valor igual a 0,9897). Já na Fig. 5, verifica-se que os resíduos estão aleatoriamente distribuídos em torno da linha central, sem nenhum indicativo da presença de comportamentos ou tendências. Fato também comprovado pelo teste Breusch-Pagan com p-valor igual a 0,4411. Dessa forma conclui-se que a variância dos resíduos é homoscedástica e linear.

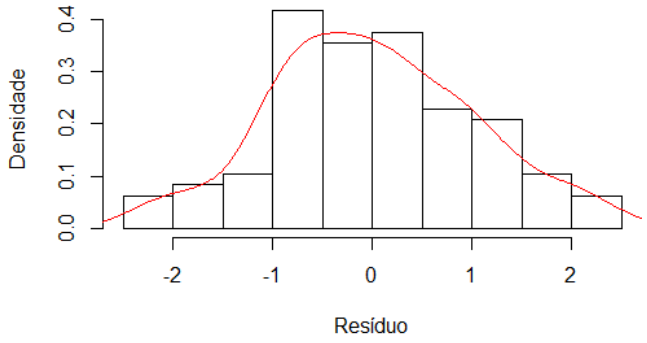

Fig. 3. Histograma dos resíduos – Normalidade.

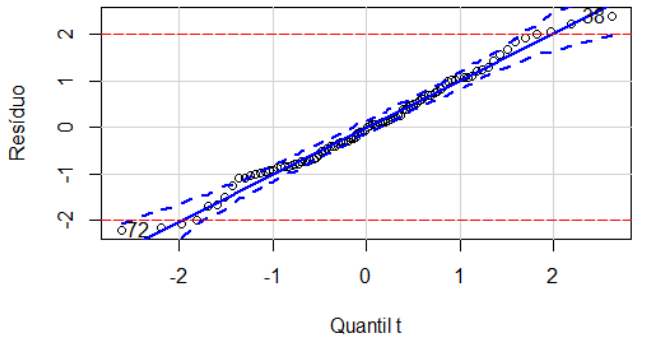

Fig. 4. Gráfico Quantil-Quantil – Normalidade.

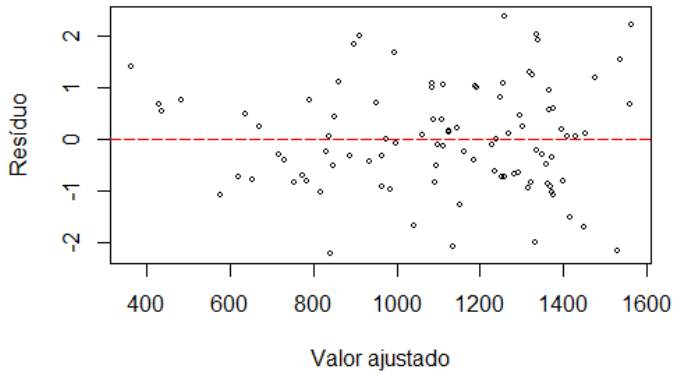

Fig. 5. Gráfico Resíduos *versus* valores ajustados – Homocedasticidade.

O modelo de regressão linear múltipla final gerado e validado, composto por 10 variáveis preditoras (resultante do método *stepwise*), apresenta um coeficiente de determinação ajustado de 0,9168 e raiz do erro quadrático médio de 79,33 (1). O ajuste do modelo para o conjunto de treino encontra-se na Fig. 6.

$$\hat{Y} = 1323{,}551253 + 0{,}00000235X_1 - 18{,}91241497X_2 - 247{,}4710496X_3 + 102{,}3952697X_4 + 1{,}262463925X_5 - 8{,}021711466X_6 - 25{,}74060595X_7 - 0{,}012377357X_8 + 0{,}00000485X_9 + 0{,}004474704X_{10} + \varepsilon \quad (1)$$

Onde:

$\hat{Y}$: Variável resposta prevista (Índice *Small Cap* – Pts);

X_1: Variável preditora (Papel moeda emitido – u.m.c., mil);

X_2: Variável preditora (Relação câmbio/salário – Índice, Junho/1994=100);

X_3: Variável preditora (IPCA – Var. %);

X_4: Variável preditora (IGP-DI – Var. %);

X_5: Variável preditora (Cotação de contratos futuros de algodão – US$/50.000 lb);

X_6: Variável preditora (VIX – US$);

X_7: Variável preditora (SINAPI – Var. %);

X_8: Variável preditora (Desembolsos do BNDES do Setor Industrial – R$ MM);

X_9: Variável preditora (Reservas bancárias – u.m.c., mil);

X_{10}: Variável preditora (Vendas de autoveículos no mercado externo – Unidades);

ε: Erro do modelo matemático.

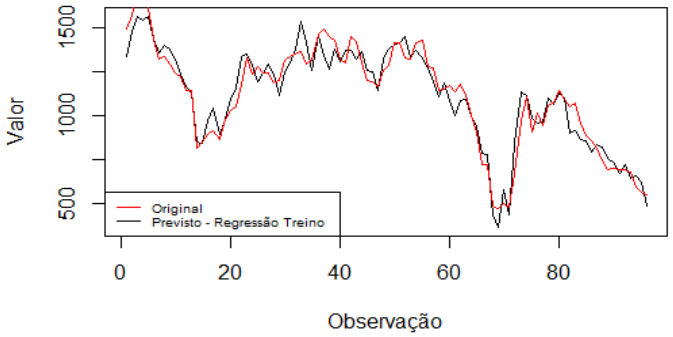

Fig. 6. Ajuste do modelo de regressão linear múltipla.

Com o modelo de regressão linear múltipla validado e com base nas observações das variáveis preditoras do conjunto de dados de teste, pode-se prever o Índice *Small Cap* obtendo a raiz do erro quadrático médio de 342,7254.

C. Rede Neural artificial

O conjunto de dados de treino normalizado foi utilizado, primeiramente, para a definição empírica do número de neurônios na camada oculta, quando se variou o mesmo entre 1 e 200. O erro quadrático médio foi mínimo com 50 neurônios, atingindo o valor de 2183,13.

Com todos os parâmetros definidos, a rede foi então treinada, ainda com os dados de treino normalizados. A rede gerada possui coeficiente de determinação ajustado de 0,9852 e raiz do erro quadrático médio de 46,72. A Fig. 7 ilustra o ajuste do conjunto de dados de treino do modelo criado.

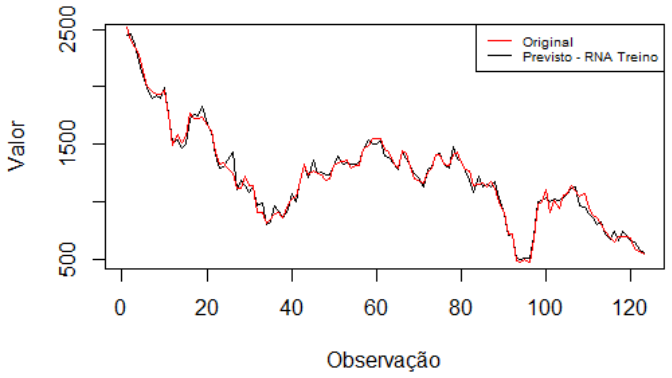

Fig. 7. Ajuste da rede neural artificial.

A partir do modelo treinado, usou-se o conjunto de dados de teste normalizado para a previsão do Índice *Small Cap*, obtendo uma raiz do erro quadrático médio de 223,99.

Na Fig. 8 é possível visualizar as previsões realizadas pelos dois modelos em comparação com os valores originais. Em ambos os casos o erro gerado na previsão com os dados de teste foi superior ao erro encontrado durante a modelagem, como era de se esperar tendo em vista que são dados nunca vistos pelos modelos.

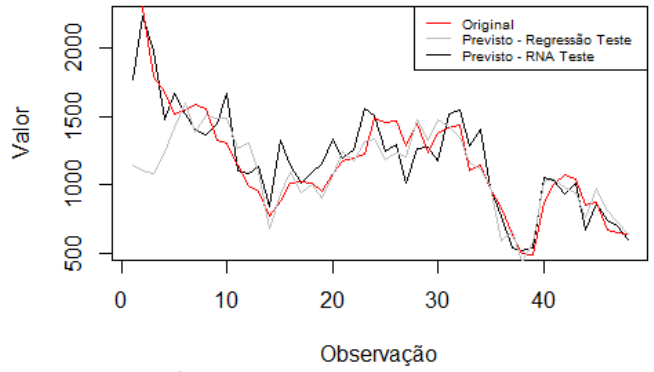

Fig. 8. Previsão do Índice *Small Cap*.

V. CONCLUSÃO

O presente estudo teve como objetivo a predição do Índice *Small Cap* da bolsa de valores brasileira, que possui como principal característica a elevada volatilidade, tendo em vista a menor negociação dos papéis e o risco intrínseco à maturidade das empresas que o compõe. Os modelos propostos levaram em consideração indicadores macro e microeconômicos brasileiros e dados relacionados à *commodity* e volatilidade.

Para tal foram utilizados métodos multivariados e de inteligência artificial, dentre eles, para a redução e seleção de variáveis, a análise de componentes principais, e para as predições a regressão linear múltipla e a rede neural artificial. Testes estatísticos foram igualmente utilizados para avaliação da adequação do modelo de regressão.

Como resultado, o modelo de regressão apresentou um coeficiente de determinação ajustado de 0,9168 e raiz do erro quadrático médio de 79,33, o que indica que as variáveis preditoras presentes na equação são capazes de explicar grande parte da variabilidade da variável resposta em estudo. Nesse modelo, o Índice *Small Cap* pode ser explicado por 10 indicadores, são eles: Papel moeda emitido, relação câmbio/salário, IPCA, IGP-DI, Cotação de contratos futuros de algodão, VIX, SINAPI, desembolsos do BNDES do setor industrial, reservas bancárias e vendas de autoveículos no mercado externo. É possível notar que a inflação, o dinheiro em circulação no mercado, as *commodities* e o risco percebido pelo mercado financeiro são fatores essenciais no comportamento da série. Na avaliação da capacidade de predição do modelo de regressão, a partir dos dados de teste, a equação apresentou a raiz do erro quadrático médio de 342,7254.

Para a rede neural, não são as variáveis presentes no modelo que indicam serem as mais significativas para explicar a variância da variável resposta, e sim os pesos sinápticos gerados durante o treino da rede. Assim sendo, o modelo contempla todas as variáveis selecionadas após a aplicação do método multivariado de redução dos dados. A rede neural artificial apresentou um coeficiente de determinação ajustado de 0,9852 e raiz do erro quadrático médio de 46,72. Entretanto, quando os dados de teste são utilizados para avaliar a capacidade predição do modelo a partir de dados nunca visto antes pelo mesmo, a raiz do erro quadrático médio atinge um valor de 223,99.

Desta forma, conclui-se que ambos os modelos foram satisfatórios para estimar o comportamento da série temporal em estudo. Embora apresente maior volatilidade, o índice pode ser explicado por indicadores econômico-financeiros, sendo a rede neural artificial a melhor opção para a predição apesar de não apresentar intervalos de confiança como uma regressão linear múltipla é capaz, assim como não possibilita identificar qual a variável que possui maior impacto, seja ele positivo ou negativo, sendo um dos grandes motivos pelos quais a regressão linear múltipla ainda é muito difundida.

O modelo de regressão apresentou uma raiz do erro quadrático médio maior, pois diversas observações atípicas foram retiradas dos dados para elaboração do modelo, consequentemente a capacidade preditiva do modelo na presença de dados atípicos é afetada. No caso da rede neural isso não acontece, o modelo consegue aprender com esses valores atípicos, apresentando melhores resultados posteriormente na predição.

Apesar dos resultados serem satisfatórios, ao longo do estudo constatou-se a dependência dos resíduos gerados pelo modelo de regressão linear múltipla, o que implica na fiabilidade dos intervalos de confiança para um determinado nível de significância. Nesse sentido, para trabalhos futuros, sugere-se a aplicação do Modelo Vetorial Auto Regressivo (VAR). Sua principal vantagem é a possibilidade de analisar o efeito da variação, ao longo do tempo, de determinada variável sobre as demais de modo a contornar a não independência dos resíduos. É comumente aplicado em estudos econométricos, onde busca-se entender a relação entre variáveis econômicas mediante um modelo matemático.

Por fim, vê-se ainda a oportunidade de aplicação de um método de validação cruzada para melhor avaliar a capacidade de generalização do modelo a partir de um conjunto de dados. Dado que a separação dos dados de treino e teste se deu de

maneira aleatória, a acurácia do modelo pode ser afetada, isto é, a assertividade do modelo poderia ser menor ou maior pelo fato de que nem sempre a separação dos dados é igual e eficaz a ponto de refletir um cenário mais provável. A validação cruzada se apresenta como uma ferramenta importante para avaliar o modelo em vários cenários de amostragem e então validar o mesmo de maneira mais efetiva.

Outras propostas para trabalhos futuros são a utilização de outras redes neurais artificiais, tais como a *Long Short-Term Memory* (LSTM) para previsão de séries temporais, e também outros algoritmos de aprendizagem do erro. Além disto, sugere-se a utilização de métodos de regressão não linear aplicados à metodologia proposta neste trabalho.

REFERÊNCIAS

[1] B3, "Destaques operacionais: Dezembro de 2019", 2019. [Online]. Available: https://ri.b3.com.br/ptb/4496/21610_729997.12.pdf..pdf.

[2] Bacen, "Ata da 232ª Reunião do Comitê de Política Monetária (Copom) do Banco Central do Brasil", 2020. [Online]. Available: https://www.bcb.gov.br/content/copom/atascopom/Copom232-not20200805232.pdf

[3] M. J. S. Oliveira, A. C. P. Assis, P. S. P. Júnior, S. W. Silva, and P. T Lemes, "Variação da Taxa Selic e a Rentabilidade de Fundos de Investimentos Referenciados: Uma Análise comparativa no período de 2013 a 2016", *Braz. J. Dev.,* Curitiba, vol. 4, no. 4, pp. 1449-1463, Julho 2018.

[4] E. Fama, "Capital markets: A review of theory and empirical work", *J. Finac*., vol. 25, no. 2, pp. 28–30, 1970.

[5] M. S. Fernandes, P. A. V. Hamberger, and A. C. M. Valle, "Análise técnica e eficiência dos mercados financeiros: Uma avaliação do poder de previsão dos padrões de candlestick", *Rev. Evid. Contáb. Financ.,* João Pessoa, vol. 3, no. 3, pp. 35-54, Novembro 2015.

[6] B3, "Índice Small Cap", 2020. [Online]. Available: http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-de-segmentos-e-setoriais/indice-small-cap-smll.htm

[7] A. Assaf Neto, *Mercado financeiro*, 12th. ed. São Paulo: Atlas, 2014.

[8] G. C. Andrade and A. L. Sanches, "Otimização de Blue Chips com Small Caps na Formação de Carteiras Utilizando a Teoria de Markowitz e o Modelo Capm.", in *X SEGeT*., Brasília, pp. 1-16, Outubro 2013.

[9] J. F. Hair Jr., W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Análise Multivariada de Dados*, 6th ed. Porto Alegre: Bookman, 2009.

[10] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, 6th ed. Upper Saddle River: Pearson Prentice Hall, 2007.

[11] G. L. Rozza, R. G. da Silva, and S. I. M. G. Muller, "Estudo comparativo do uso de redes neurais artificiais e regressão linear múltipla para a previsão da concentração cáustica em uma etapa do processo de fabricação de alumina" *Rev. Prod. Online*, Florianópolis, vol. 15, no. 3, pp. 948-971, Setembro 2015.

[12] J. A. de Oliveira, B. Siqueira, and L. Brandão, "Comparação dos processos de micro rosqueamento por conformação e usinagem na liga de alumínio 7075-T651.", in *IX COBEF*, Joinville, Junho 2017.

[13] T. Moriggi, G. V. Loch, and M. A. M. Marques, "Product Performance: A Prediction Model for Compressive Strength of Composed Cements.", *IEEE Lat. Am. T*., vol. 18, no. 3, pp. 507-513, Março 2020.

[14] I. T. Jollife, "Discarding Variables in a Principal Component Analysis.", *J. Roy. Stat. Soc. C-App.,* vol. 21, no. 2, pp. 160-173, 1972.

[15] J. Zimmer and M. Anzanello, "Um novo método para seleção de variáveis preditivas com base em índices de importância", *Production*, vol. 24, no.1, pp. 84-93, Março 2014.

[16] M. F. Alves, A. D. P. Lotufo, and M. L. M. Lopes, "Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas.", *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics,* vol. 1 no. 1, 2013.

[17] H. N. Camelo, P. S. Lucio, O. M. Gomes, and J. B. V. Leal Jr., "Predição de velocidade do vento em regiões do nordeste brasileiro através de Regressão Linear e Não Linear para fins de geração eólica.", *Rev. Bras. Geogr. Fís.*, vol. 9, no. 3, pp. 927-939, Abril 2016.

[18] B. C. Santos, G. Andrade, and G. A. Del Conte, "A method for monitoring and diagnosing the circular trajectory error in micro milling.", *IEEE Lat. Am. T*., vol. 14, no. 12, pp. 4639-4645, Dezembro 2016.

[19] G. C. Leal, P. Mattiazzi, M. D. Hettwer and C. V. Silva, "Determinação de cádmio em suplementos alimentares por HR-CS SS GFAAS.", *Anais do X SIEPE da UNIPAMPA*, vol. 10, no. 2, Março 2020.

[20] E. O. Biu, M. T. Nwakuya, and N. Wonu, "Detection of Non-Normality in Data Sets and Comparison between Different Normality Tests.", *Asian J. Probab. Stat*., vol. 5, no. 4, pp. 1-20, Janeiro 2020.

[21] L. Oliveira, P. C. Tonin, and S. L. Vicenzi, "Comportamento dos custos totais de produção no segmento da avicultura de postura no estado do Paraná: Estudo baseado na análise de regressão linear múltipla.", *Rev. Prod. Online*, Florianópolis, vol. 20, no. 1, pp. 28-46, Março 2020.

[22] S. T. Zavadzki, M. Kleina, F. O. Drozda, and M. A. M. Marques, "Computational Intelligence Techniques Used for Stock Market Prediction: A Systematic Review.", *IEEE Lat. Am. T.*, vol. 18, no. 4, pp. 744-755, Abril 2020.

[23] L. Fleck, M. H. F. Tavares, E. Eyng, A. C. Helmann, and M. A. M. Andrade, "Redes neurais artificiais: Princípios básicos.", *Rev. Eletr. Científica Inovação e Tecnologia*, vol. 1, no. 13, pp. 47-57, 2016.

[24] S. Haykin, *Neural Networks and Learning Machines,* 3rd ed. USA: Prentice Hall, 2009.

[25] H. R. O. Rocha and M. A. S. Macedo, "Previsão do Preço de Ações Usando Redes Neurais." *Anais do VIII Congresso USP de Iniciação Científica em Contabilidade*, São Paulo, Julho 2011.

[26] C. D. Tunes, V. P. Gonçalves, D. B. Rodrigues, A. S. Almeida, J. B. Silva, and M. S. Franco, "Fosfito de potássio como indutor de resistência em mutantes de tomateiro contra Phytophthora infestans." *Rev. Verde de Agroecologia e Desenvolv. Sustent.,* vol. 14, no. 2, pp. 218 – 223, Junho 2019.

Bianca Kaczorowski é técnica em administração pelo Instituto Federal do Paraná (2014) e graduada em Engenharia de Produção pela UFPR (2020). Tem experiência nas áreas de projetos logísticos e de negócios digitais.

Mariana Kleina é professora do departamento de Engenharia de Produção da UFPR. É graduada em Matemática Industrial (2009) e doutora (2015) em Métodos Numéricos em Engenharia pela UFPR. Gosta das áreas de Pesquisa Operacional e Inteligência Artificial.

Marcos Augusto Mendes Marques é graduado em Engenharia Elétrica (2003) e doutor em Métodos Numéricos em Engenharia pela UFPR (2015). É professor do Departamento de Engenharia de Produção da UFPR. Tem experiência na área de Matemática e Estatística, com ênfase em Análise Numérica e Simulação.

Wiliam de Assis Silva é graduado em Engenharia de Produção Civil pela UTFPR (2010) e mestre em Engenharia de Produção pela UFPR (2018). Tem experiência na área de Gestão de Projetos e da Qualidade, Processos de Construção Civil e Sustentabilidade em Cadeias Logísticas.