BRUNO PÓVOA RODRIGUES

REAMOSTRAGEM EM REDES NEURAIS COM APLICAÇÃO A DADOS ESPACIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Agricultura e Informações Geoespaciais da Universidade Federal de Uberlândia, Campus Monte Carmelo, como parte das exigências para obtenção do título de "Mestre".

Orientador

Prof. Dr. Marcelo Tomio Matsuoka.

Coorientador

Prof. Dr. Vinicius Francisco Rofatto.

MONTE CARMELO
MINAS GERAIS - BRASIL
2021

BRUNO PÓVOA RODRIGUES

REAMOSTRAGEM EM REDES NEURAIS COM APLICAÇÃO A DADOS ESPACIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Agricultura e Informações Geoespaciais da Universidade Federal de Uberlândia, Campus Monte Carmelo, como parte das exigências para obtenção do título de "Mestre".

DEFESA em 30 de agosto de 2021.

Banca examinadora

Prof. Dr. Marcelo Tomio Matsuoka - Instituto de Geografia – (UFU)

Prof. Dr. Enio Tarso de Souza Costa - Instituto de Ciências Agrárias – (UFU)

Prof. Dr. Vinícius Amadeu Stuani Pereira - Curso de Agronomia – (UTFPR)

MARCELO TOMIO MATSUOKA
Instituto de Geografia – (UFU)
(Orientador)

MONTE CARMELO
MINAS GERAIS - BRASIL
2021

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

R696 2021

Rodrigues, Bruno Póvoa, 1990-
Reamostragem em redes neurais com aplicação a dados espaciais [recurso eletrônico] / Bruno Póvoa Rodrigues. - 2021.

Orientador: Marcelo Tomio Matsuoka.
Coorientador: Vinicius Francisco Rofatto.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Agricultura e Informações Geoespaciais.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2021.507
Inclui bibliografia.

1. Agronomia. I. Matsuoka, Marcelo Tomio,1978-, (Orient.). II. Rofatto, Vinicius Francisco,1986-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Agricultura e Informações Geoespaciais. IV. Título.

CDU: 631

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Agricultura e Informações Geoespaciais
Rodovia LMG 746, Km 01, s/nº, Bloco 1AMC, Sala 1A202 , Monte Carmelo-MG, CEP 38.500-000
Telefone: (34) 3810-1033 - ppgaig@iciag.ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Agricultura e Informações Geoespaciais				
Defesa de:	Dissertação de Mestrado Acadêmico				
Data:	30/08/2021	Hora de início:	19:00	Hora de encerramento:	21:20
Matrícula do Discente:	31922AIG002				
Nome do Discente:	Bruno Póvoa Rodrigues				
Título do Trabalho:	Reamostragem em redes neurais com aplicação a dados espaciais				
Área de concentração:	Informações geoespaciais e tecnologias aplicadas à produção agrícola				
Linha de pesquisa:	Aplicações e desenvolvimento de métodos em informações espaciais				

Reuniu-se na sala virtual https://conferenciaweb.rnp.br/webconf/ marcelo-tomio-matsuoka a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Agricultura e Informações Geoespaciais, assim composta: : Dr. Enio Tarso de Souza Costa (UFU), Dr. Vinícius Amadeu Stuani Pereira (Universidade Tecnológica Federal do Paraná) e Dr. Marcelo Tomio Matsuoka (UFU) – orientador do candidato.

Iniciando os trabalhos o(a) presidente da mesa, Dr. Marcelo Tomio Matsuoka, apresentou a Comissão Examinadora e o candidato(a), agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

Aprovado(a).

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Marcelo Tomio Matsuoka**, **Professor(a) do Magistério Superior**, em 31/08/2021, às 09:58, conforme horário oficial de Brasília, com fundamento no art. 6º,

§ 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Enio Tarso de Souza Costa**, **Professor(a) do Magistério Superior**, em 31/08/2021, às 14:56, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Vinícius Amadeu Stuani Pereira**, **Usuário Externo**, em 01/09/2021, às 10:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3010849** e o código CRC **C6731E19**.

Referência: Processo nº 23117.057818/2021-25 SEI nº 3010849

Dedico
A minha família, amigos e professores.

AGRADECIMENTOS

Por todo suporte e conforto emocional agradeço aos meus familiares, Nilva, Mauro, Lukas, Christiane, Giuliano, Valéria e Maria.
A minha companheira Elisa, por me auxiliar em todo o desenvolvimento do trabalho e pela paciência nos momentos difíceis, e seus pais Márcia e Aroldo pelo incentivo.
Aos meus amigos Wilton Junior e seus pais Elizeth e Wilton pelo acolhimento em seu lar, Lauren e Luan pelos conselhos, Jessica, Paula, Matheus e Renato M. pelo companheirismo, Cleber e Laura pelo incentivo e Eduardo pelos ensinamentos.
Aos meus professores e colegas de turma do PPGAIG em especial Fernando, Juliano, Renato, Talita e Vinícius.
A todos os membros do Grupo de Pesquisa Controle de Qualidade e Inteligência Computacional em Geodesia, da Universidade Federal de Uberlândia.
Os mais sinceros agradecimentos à Fazenda Juliana, por incentivar a ciência cedendo os dados mais que essenciais para que essa pesquisa fosse realizada, ao Alfredo P. de Oliveira Junior, Gerente Administrativo e Financeiro, que nos recebeu tão solicitamente.
Ao meu orientador Prof. Dr. Marcelo Tomio Matsuoka e coorientador Prof. Dr. Vinicius Francisco Rofatto, sem eles esse estudo não seria possível de ser realizado.
Por fim, agradeço ao Programa de Pós-graduação em Agricultura e Informações Geoespaciais, aos colaboradores que ajudaram direto ou indiretamente a minha formação.
A todos, o meu mais sincero obrigado.

BIOGRAFIA

Bruno Póvoa Rodrigues, nascido em Araguari, Minas Gerais em 1990. Graduado em Engenharia Agronômica pela Universidade Federal de Uberlândia (UFU - 2015). Especialização em Gestão Ambiental e Desenvolvimento Sustentável pelo Centro Universitário Internacional (UNINTER - 2019). Mestrado no Programa de Pós-graduação em Agricultura e Informações Geoespaciais Universidade Federal de Uberlândia, Campus Araras (UFU - 2021).

SUMÁRIO

RESUMO

RODRIGUES, BRUNO PÓVOA**. REAMOSTRAGEM EM REDES NEURAIS COM APLICAÇÃO A DADOS ESPACIAIS.** 2021. Dissertação (Mestrado em Agricultura e Informações Geoespaciais) - Universidade Federal de Uberlândia, Campus Monte Carmelo, Minas Gerais, Brasil[1].

No desenvolvimento de redes neurais artificiais (RNA) o conjunto de dados disponível é dividido em três categorias: treinamento, validação e teste. No entanto, surge aqui um problema importante: como podemos confiar na predição fornecida por uma única RNA? Devido à aleatoriedade relacionada à própria RNA (arquitetura, inicialização e procedimento de treinamento), geralmente, não existe a melhor escolha. Para capturar a aleatoriedade intrínseca à RNA, apresentamos uma abordagem baseada no método Jackknife de reamostragem estatística. O Jackknife clássico consiste em remover uma observação do conjunto de dados disponíveis (n) e usar as ($n - 1$) amostras restantes no processo de estimação. Este processo é repetido para cada observação individual. Ao final, ter-se-á n estimativas advindas de amostras diferentes. No caso de redes neurais, cada observação individual é selecionada para compor o conjunto de teste, enquanto o restante da amostra é destinado para o treinamento da rede. Nesse caso, o número de redes neurais é igual ao tamanho dos dados disponíveis. Entretanto, estendemos a ideia ao replicar esse procedimento por um certo número de vezes. Logo, devido à característica aleatória da rede neural, as predições variam para um mesmo ponto amostral. Consequentemente, podemos descrever a distribuição de cada predição individual. Portanto, o método proposto fornece predições intervalares ao invés da tradicional predição pontual. O método proposto foi aplicado e testado utilizando dados de potencial de hidrogênio (pH), cálcio trocável (Ca^{2+}) e concentração de fósforo (P) obtidos por meio da análise de 118 pontos de solos georreferenciados. Os resultados mostraram que a redução de 60% no conjunto de dados disponível oferece acurácia compatível em relação ao conjunto de dados completo e, portanto, um custo maior de amostragem em campo não seria necessário. O método de reamostragem caracteriza espacialmente os pontos de maior e menor acurácia e incerteza. Na avaliação externa, ou seja, na análise de dados que não participaram da reamostragem, observamos que a taxa de sucesso é maior quando usamos a predição intervalar em vez de usar a predição média. Embora restrinjamos a aplicação em redes neurais, o método proposto pode ser estendido a outras ferramentas estatísticas modernas, tais como Krigagem, Colocação por Mínimos Quadrados, entre outros.

Palavras-chave: Redes Neurais Artificiais, Reamostragem, Delete-1 Jackknife, Análise Espacial, Solo.

Orientador: Prof. Dr. Marcelo Tomio Matsuoka - UFU

ABSTRACT

RODRIGUES, BRUNO PÓVOA. **RESAMPLING IN NEURAL NETWORKS WITH APPLICATION TO SPATIAL ANALYSIS**. 2021. Dissertation (Master's Degree in Agriculture and Geospatial Information) - Federal University of Uberlândia, Campus Monte Carmelo, Minas Gerais, Brazil[2].

In the development of artificial neural networks (ANN), the available dataset is divided into three categories: training, validation and testing. However, an important problem arises here: how can we trust the prediction provided by a single ANN? Due to the randomness related to the ANN itself (architecture, initialization and training procedure), usually, there is no better choice. To capture the intrinsic randomness of RNA, we present an approach based on the Jackknife method of statistical resampling. The classic Jackknife consists in removing an observation from the available dataset (n) and using the (n – 1) remaining samples in the estimation process. This process is repeated for each individual observation. At the end, there will be n estimates from different samples. In the case of neural networks, each individual observation is selected to compose the test set, while the rest of the sample is destined for network training. In this case, the number of neural networks is equal to the size of the available data. However, we extend the idea by replicating this procedure a certain number of times. Therefore, due to the random characteristic of the neural network, predictions vary for the same sampling point. Therefore, due to the random characteristic of the neural network, predictions vary for the same sampling point. Consequently, we can describe the distribution of each individual prediction. Therefore, the proposed method provides interval predictions instead of the traditional point prediction. The proposed method was applied and tested using hydrogen potential (pH), exchangeable calcium (Ca^{2+}) and phosphorus concentration (P) data obtained through the analysis of 118 georeferenced soil points. The results showed that the 60% reduction in the available dataset offers compatible accuracy compared to the full dataset and, therefore, a higher cost of sampling in the field would not be necessary. The resampling method spatially characterizes the points of greater and lesser accuracy and uncertainty. In external validation, i.e., analyzing data that did not participate in the resampling, we observed that the success rate is higher when using interval prediction rather than using average prediction. Although we restrict it to the neural network model, the proposed method can also be extended to other modern statistics tools, such as Kriging, Least Squares Collocation, and so on.

Keywords: Artificial Neural Networks, Resampling, Delete-1 Jackknife, Spatial Analysis, Soil.

Advisor: Prof. Dr. Marcelo Tomio Matsuoka - UFU

1. INTRODUÇÃO

Uma das tarefas mais importantes no desenvolvimento de uma rede neural é o particionamento dos dados disponíveis (WU et al., 2012). Geralmente, é comum entre pesquisadores dividir os dados aleatoriamente e uniformemente em três categorias: treinamento, validação e teste (MAIER et al., 2010). O conjunto de treinamento é frequentemente usado para estimar os parâmetros desconhecidos da rede (por exemplo, pesos e tendências). Pode-se usar também uma subdivisão do conjunto de treinamento, criando um conjunto de validação, valendo-se como critério de parada do treinamento. Por fim, os dados de teste são utilizados para verificar a eficiência da rede sob condições reais de aplicação.

Este método de divisão é conhecido como Hold-out (YADAV; SHUKLA, 2016) e é amplamente adotado. Entretanto, existem limitações significativas para sua aplicação (ZIGGAH et al. 2019): (i) os resultados produzidos são baseados em uma divisão não controlada; (ii) a divisão inadequada do conjunto de dados pode ter um efeito adverso no desempenho do modelo; (iii) inadequado em situações de baixa densidade de dados (dados insuficientes e/ou esparsos).

Como alternativa às limitações do método Hold-out, a validação cruzada K-fold tem sido recomendada (BURMAN, 1989; REITERMANOVÁ, 2010). Neste caso, os dados são separados em *K* subconjuntos (K-fold) de tamanho aproximadamente igual, de modo que cada subconjunto estará apenas uma única vez no conjunto de teste. Embora K-fold apresente vantagens, ainda não está claro como escolher os subconjuntos, e em alguns casos os subconjuntos não possuem tamanhos iguais, o que não garante uma versão balanceada de validação cruzada. Além disso, tanto o K-fold quanto o Hold-out fornecem predições pontuais. Portanto, a divisão de dados ainda é um desafio. Aqui, surge uma questão importante: como podemos confiar na predição fornecida por uma única RNA?

Dentro do contexto da modelagem de rede neural, a melhor escolha é aquela em que o conjunto de teste é o menor possível. Em outras palavras, se o conjunto de treinamento estiver próximo do tamanho total da amostra, mais acurada será a rede neural. Ainda assim, as incertezas relacionadas à própria rede (arquitetura, inicialização e procedimento de aprendizagem) ainda são questionáveis. Logo, não existe a melhor escolha (PAN, 1998). À luz desta questão, apresentamos uma abordagem baseada no método Jackknife de reamostragem estatística (QUENOUILLE, 1949). O método proposto foi testado utilizando dados espaciais de elementos químicos do solo. Também investigamos se a abordagem proposta é capaz de fornecer boas predições em condições de baixa densidade de dados.

2. MATERIAL E MÉTODOS

2.1. Dados disponíveis

Os dados utilizados para o experimento foram coletados na área agrícola da Fazenda Juliana, localizada no município de Monte Carmelo-MG, latitude de -18°42’30.45’’ e longitude -47°32’49.74’’, área esta cultivada com cafeeiro arábica (*Coffea arábica L.*), possuindo aproximadamente 98,6 hectares.

O conjunto de dados utilizado é composto por atributos químicos do solo coletados na profundidade 0 - 20 cm: potencial de hidrogênio (pH) − medido em água por pHmetro −; cálcio trocável (Ca^{2+}) − determinados por método KCl 1 mol L^{-1} − e concentração de fósforo (P) – determinado por método resina de troca −, quantificados por meio de análise de amostras de solo conforme descrito no Manual Brasileiro de Métodos de Análise de Solo (TEIXEIRA et al., 2017) de 118 pontos georreferenciados (Figura 1).

As coordenadas foram fornecidas no sistema *Universal Transversa de Mercator* (UTM). Estes atributos químicos desempenham um papel importante na fertilidade, na vida e nas práticas de manejo do solo e possuem variabilidades diferentes, conforme observado na Tabela 1, o que torna um problema interessante para a aplicação do método proposto.

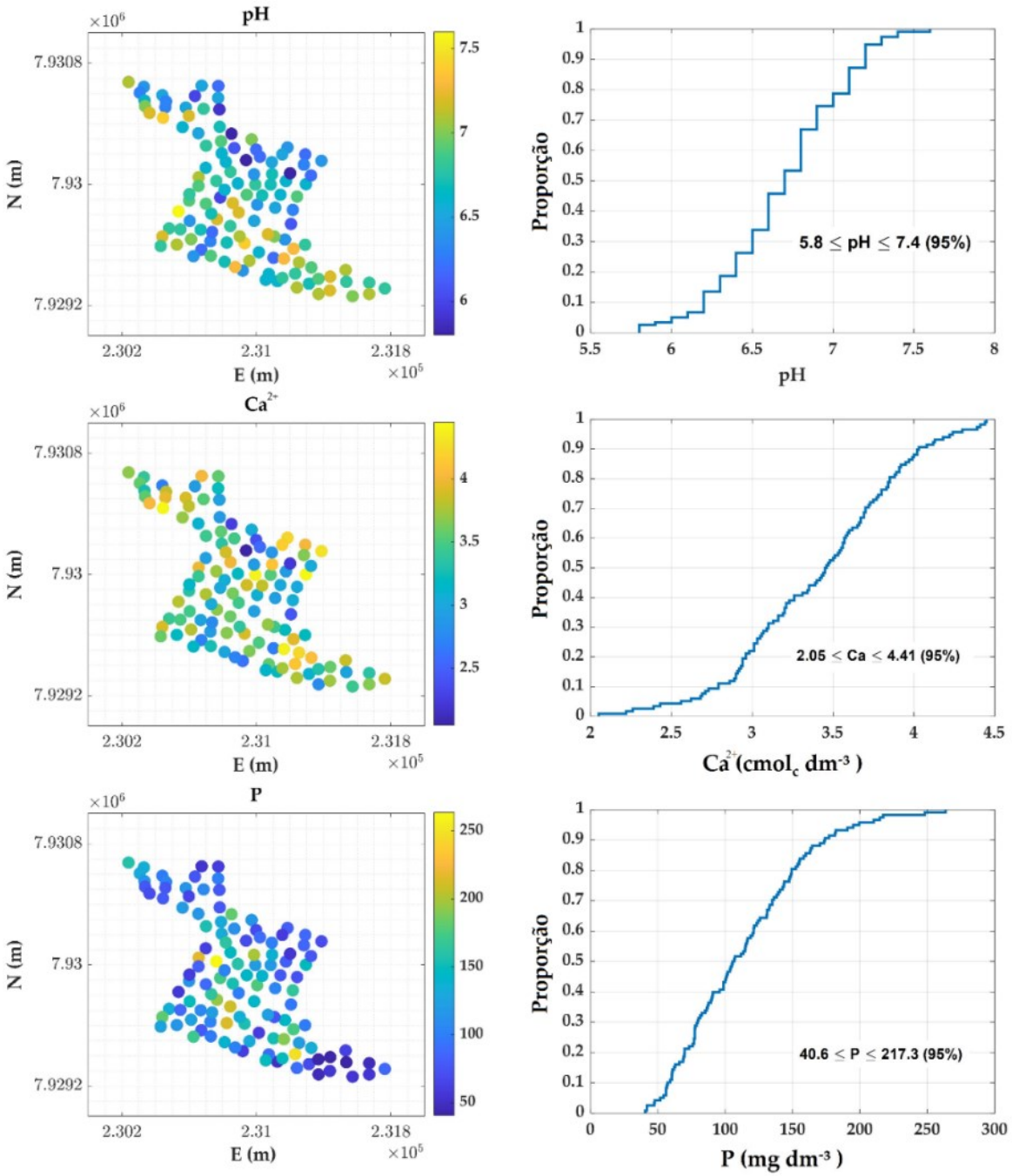

FIGURA 1. Conjuntos de dados dos atributos químicos do solo distribuídos espacialmente segundo as coordenadas *North* e *East* (N e E – UTM).

TABELA 1. Caracterização do conjunto de dados original.

Atributo	Min	Max	Med	DP	CV (%)
pH	5,8	7,6	6,7	0,4	5,6
Ca^{2+}	2,1	4,5	3,4	0,5	14,9
P	40,6	190,7	104,8	38,2	36,5

pH = potencial de hidrogênio; Ca^{2+} = cálcio trocável ($cmol_c\ dm^{-3}$); P = conteúdo de fósforo ($mg\ dm^{-3}$); DP = desvio padrão; CV = coeficiente de variação.

2.2. Desenvolvimento da rede neural espacial

Uma rede neural *feedforward* foi desenvolvida utilizando *backpropagation* (retropropagação como procedimento de aprendizado) com as coordenadas UTM como conjunto de dados de entrada e os atributos químicos do solo (pH, Ca^{2+} e P) como alvos. A funcionalidade de linha de comando da caixa de ferramentas de Rede Neural do Matlab (R2019b) foi usada para treinar e validar a RNA. Após vários testes, a arquitetura de rede neural ideal encontrada foi definida por três camadas ocultas, com a primeira camada composta por três neurônios, a segunda por quatorze e a última por um único neurônio, ou seja, [3 14 1] (Figura 2), esta arquitetura também foi encontrada em Cagliari et al. (2011). Como os dados de entrada são compostos apenas por coordenadas UTM (E, N), a rede neural foi denominada rede neural espacial.

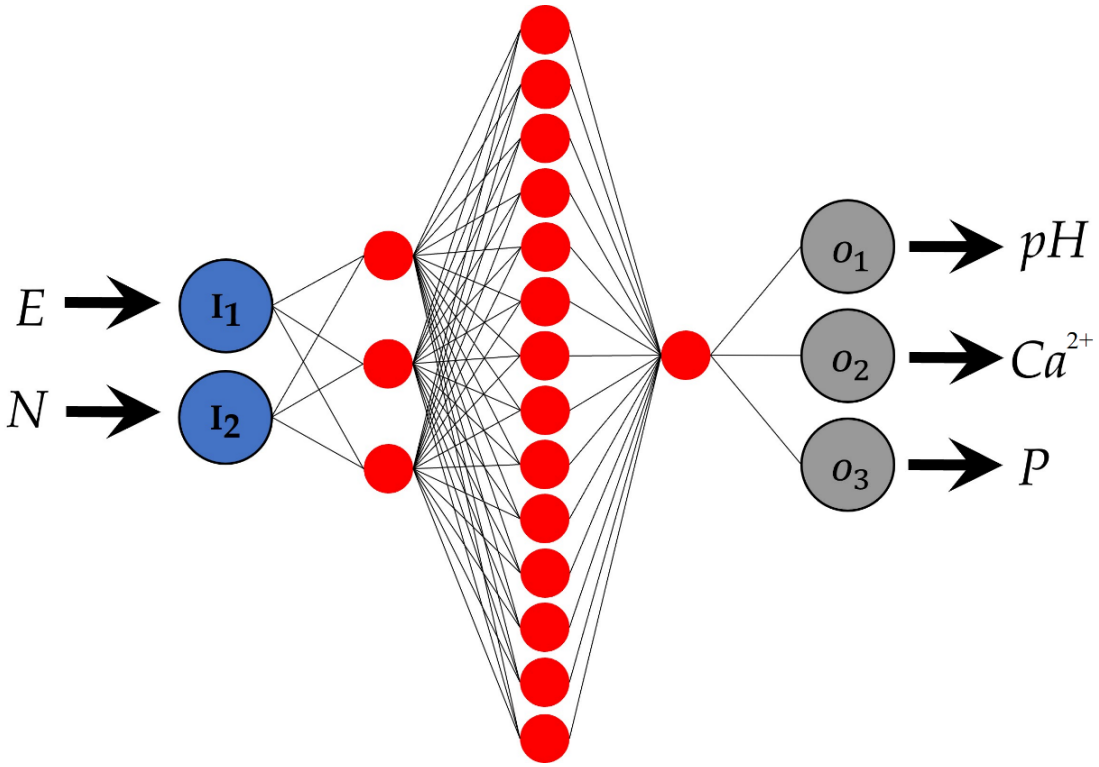

FIGURA 2. Topologia da rede neural *feedforward* para predição de atributos químicos do solo.

Normalizar o conjunto de dados para acelerar o aprendizado está entre as melhores práticas para treinar uma rede neural e assim obter uma convergência mais rápida (HUANG et al., 2020). Portanto, os valores de entradas e saídas foram normalizados da seguinte forma (equação 1):

$$y_n = 2 \times \left(\frac{y - y_{min}}{y_{max} - y_{min}}\right) - 1 \quad (1)$$

sendo: y_n as entradas ou saídas foram padronizadas no intervalo [-1,1], y as entradas originais ou dados de destino, y_{min} e y_{max} os valores mínimo e máximo obtidos a partir das entradas ou metas originais, respectivamente. Neste caso, a função normalizada fornece valores de entrada/alvos entre [-1,1].

A função de transferência log-sigmóide (função de ativação) foi aplicada nas camadas intermediárias (camadas ocultas) para calcular os dados de saída a partir da camada de entrada da rede. A função logística (denotada por φ) adota valores entre 0 e 1, conforme a seguir (equação 2):

$$\varphi(x_i) = \frac{1}{1 + e^{-x_i}} \quad (2)$$

no qual x_i é um valor proveniente da combinação linear das entradas, pesos e polarização para um determinado neurônio, ou seja (equação 3):

$$x_i = w_{1i} \times I_1 + w_{2i} \times I_2 + \cdots + w_{ni} \times I_n + \theta_i \quad (3)$$

com w sendo os parâmetros de pesos da rede, I os dados de entrada e θ o parâmetro de polarização para dados de amostra de entrada 1, …, n.

Durante a fase de aprendizagem, é habitual e conveniente minimizar o erro quadrático médio (MSE) entre os alvos (t_i) e saída estimada (o_i), conforme equação 4 (HAYKIN, 1999):

$$\chi^2 = \sum_{i=1}^{m} (t_i - o_i)^2 \rightarrow min \quad (4)$$

Os pesos e bias são parâmetros desconhecidos da rede neural e foram estimados minimizando a função apresentada na equação (4) a fim de produzir o resultado desejado, ou seja, os atributos químicos. Aqui, o algoritmo de Levenberg-Marquardt (LM), também conhecido como método de *Damped Least-Squares* (DLS), foi empregado para estimar estes parâmetros desconhecidos (LEVENBERG, 1944; MADSEN et al., 2004; MARQUARDT, 1963; GAVIN, 2020) e o algoritmo Nguyen-Widrow foi usado para inicializá-los (NGUYEN; WIDROW, 1990).

2.3. Extensão do método de reamostragem Delete-1 Jackknife em redes neurais

Em geral, o número de combinações possíveis para validação cruzada independente (*independent crossvalidation*), denotado por C_n^d, e o número de repetições de cada ponto de predição (R_n^d) são representados respectivamente por (equação 5 e 6):

$$C_n^d = \binom{n}{d} = \frac{n!}{(n-d)!d!} \tag{5}$$

$$R_n^d = C_n^d \times \left(\frac{d}{n}\right) \tag{6}$$

no qual: n é o tamanho da amostra disponível (ou tamanho do conjunto de dados em mãos), d é o número de pontos que são removidos da amostra disponível para validação. Isso significa que validamos o subconjunto de tamanho d e treinamos o restante $(n-d)$ observações de cada vez. O tamanho do subconjunto d é selecionado a partir de todas as observações sem substituição. Por exemplo, se $n = 10$ e $d = 2$, nós teríamos $C_{10}^2 = 45$ subconjuntos de validação cruzada independentes, com cada saída predita $R_{10}^2 = 9$ vezes. Este método é conhecido como Jackknife-d (EFRON, 1980; 1992; WANG E YU, 2020).

O Delete-1 Jackknife é um caso especial ao considerar $d = 1$. Neste caso, cada uma das observações é utilizada para validar o desempenho da rede neural treinada pelas demais $(n-1)$ observações. No contexto de redes neurais, o método Delete-1 Jackknife é uma versão balanceada de validação cruzada, já na linguagem de aprendizado de máquina, o método é comumente referido como procedimento de *leave-one-out* (EFRON, 1982). Por exemplo, se $n = 10$, a aplicação do método Delete-1 Jackknife forneceria 10 validações cruzadas, com cada observação predita apenas uma vez, ou seja, $C_{10}^1 = 10$ e $R_{10}^1 = 1$.

Por outro lado, estendemos o método Delete-1 Jackknife, ao desenvolver várias repetições a fim de capturar a aleatoriedade relacionada à própria rede (arquitetura, inicialização e procedimento de aprendizagem). Chamamos esse método *Delete-1 Jackknife Trial* (ou simplesmente Jack-1T). Portanto, em vez de ter apenas um único preditor de rede neural particular, teremos centenas ou mesmo milhares deles, isto nos dá a oportunidade de fazer uma predição de intervalo em vez de pontual, ou seja, podemos descrever sua distribuição empírica. Para o mesmo exemplo, se $n = 10$ e o número de repetições (denotado por t) é $t = 500$, logo teremos $10 \times 500 = 5000$ preditores de rede neural, sendo 500 predições de cada amostra com uma resposta pontual a cada teste. Neste caso, o número de preditores de rede neural será igual a $n \times t$.

A eficiência do método de reamostragem Jackknife depende da escolha do número de grupos que serão excluídos do processo de treinamento da rede neural (ou seja, d). Em geral, quanto mais observações no conjunto de treinamento, melhor será a rede em termos de aprendizagem, o que justifica a ideia do Jack-1T, conceito semelhante pode ser encontrado em Miller (1974) e Pan (1998).

2.4. Resultados de Jack-1T

A estimativa mais provável (valor esperado) de um dado de saída não se baseia apenas em uma única predição, mas sim em uma base de intervalos. Para isso, classificamos os N valores preditos de um ponto em ordem estritamente crescente. Os valores preditos classificados fornecem uma função de distribuição empírica para cada ponto de saída (denotado por G). Então, para uma probabilidade de intervalo estipulada (denotada por α) somos capazes de calcular os percentis desejados p como a seguir (equação 7):

$$p = G_{[(1-\alpha)\times N]} \tag{7}$$

onde o valor encontrado entre os colchetes $[(1-\alpha)\times N]$ deve ser arredondado para o inteiro mais próximo. Esse valor indica a posição do elemento selecionado para uma dada probabilidade α. Isso pode ser feito para qualquer α em paralelo. De fato, um intervalo de confiança das predições também pode ser construído. Este procedimento é muito semelhante ao encontrado para o cálculo de valores críticos em testes estatísticos para detecção de outliers (ROFATTO et al., 2020).

Os métodos de reamostragem apresentados também nos permitem medir o desempenho de predição para todo o conjunto de amostra. Isso se deve ao fato de que todos os pontos de amostragem são replicados no conjunto de teste. A raiz quadrada média dos erros (*RMSE*) no conjunto de teste foi usado para medir a acurácia da predição de cada ponto "i" como a seguir (equação 8):

$$RMSE_{(i)} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(T_k - \hat{O}_k\right)^2}, \quad \mathrm{i} = 1, \ldots, n \tag{8}$$

em que $\hat{O}_k$ são os N valores preditos para cada ponto i e T_k são os N valores reais de saída. O $RMSE$ tem sido usado como um parâmetro estatístico padrão para medir o desempenho do modelo em várias aplicações. O $RMSE$ indica a que distância as respostas de saída da rede

neural estão das saídas reais (alvos) para cada i. Observe que $RMSE$ é calculado para cada um dos n pontos de amostra disponíveis. Embora essa expressão represente os *RMSE*s para uma única saída, ela pode ser aplicada ao caso em que mais de uma saída esteja disponível. Outras estatísticas também podem ser calculadas para cada ponto predito, como média, desvio padrão (incerteza), quartis, erro relativo, coeficiente de variação e assim por diante. Consequentemente, mapas de acurácia e incerteza também podem ser fornecidos.

2.5. Experimentos

Inicialmente, utilizamos 100% dos dados disponíveis (118 pontos) para a aplicação do método de reamostragem Jack-1T (Figura 1). Nesta etapa, a validação cruzada baseou-se apenas no conjunto de testes, que consistiu na avaliação interna do procedimento de reamostragem. O número de testes foi adotado como $t = 100$, que forneceu 11800 preditores de rede neural para o caso em que usamos 100% do conjunto de dados disponível. O tamanho do conjunto de validação foi fixado tomando dez pontos aleatoriamente para cada repetição. O objetivo do conjunto de validação é melhorar a generalização da aprendizagem da rede. Quando a rede começa a se ajustar ao conjunto de treinamento, o desempenho no conjunto de validação geralmente começa a diminuir. Quando o erro de validação aumenta para um número especificado de épocas, o treinamento é interrompido e os pesos e tendências no mínimo do erro de validação são retornados. Este método para evitar o ajuste excessivo/insuficiente é chamado de parada precoce (REITERMANOVÁ, 2010).

O processo de aprendizagem foi configurado para parar quando o desempenho da rede no conjunto de validação falhar em melhorar ou permanecer o mesmo por 6 épocas (iterações). Logo o conjunto de validação tem efeito sobre o treinamento da rede com a possibilidade de interrompê-lo, sendo assim, ele foi considerado parte do processo de treinamento.

Ainda nesta etapa, investigamos até que ponto o Jack-1T é capaz de fornecer uma boa predição em condições nas quais as amostras são menores do que o conjunto de dados original. Para este efeito, o conjunto de dados original ($n = 118$ pontos) foi intencional e aleatoriamente reduzido a ~40% (47 pontos), valor este encontrado após testes empíricos com diferentes reduções do conjunto de dados original. O número de testes foi considerado como $t = 100$, que forneceu 4700 preditores de rede neural para o caso em que usamos 40% do conjunto de dados disponível. Assim, foram fornecidas 100 predições para cada ponto de teste da amostra, o que permitiu uma avaliação interna em relação ao conjunto original de dados disponível.

O conjunto restante (71 pontos) foi utilizado na avaliação externa. Nesse caso, esse conjunto não participou do procedimento de reamostragem Jack-1T e, portanto, o que garantiu

uma avaliação imparcial do método proposto. Em outras palavras, 4700 preditores neurais gerados a partir da reamostragem de 40% dos dados (47 pontos) foram aplicados a cada um dos 71 pontos (60% do tamanho dos dados disponíveis). Consequentemente, cada ponto de amostra foi predito 4700 vezes, o que permitiu a predição de um intervalo em vez da predição pontual clássica.

As análises foram realizadas a partir de duas perspectivas: (i) quantitativa, comparando as estatísticas descritivas entre os atributos originais e os preditos usando a rede neural baseada em Jack-1T; (ii) qualitativa, comparando a classe à qual cada predição individual pertence em relação à sua classe original a partir de tabelas de classificação de atributos químicos do solo.

3. RESULTADOS E DISCUSSÃO

3.1. Avaliação Quantitativa

Não houve diferença significativa entre as predições de diferentes tamanhos de amostra e o conjunto de dados original em termos de avaliação interna (Figura 3). Os valores máximo e mínimo (barras Figura 3) para as predições são o máximo dos valores do percentil 97,5 (α = 0,975 calculado a partir da Equação 7) e o mínimo dos valores do percentil 2,5 (α = 0,025 calculado a partir da Equação 7), respectivamente. Esses percentis correspondem ao limite inferior (α = 0,025) e superior (α = 0,975) do intervalo de confiança de 95%, respectivamente. Isso significa que as predições que estão fora desse intervalo de confiança foram consideradas outliers e não foram consideradas nas soluções.

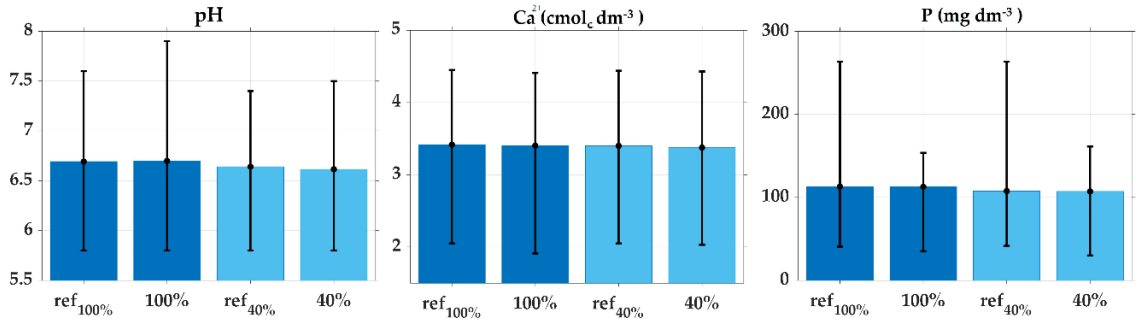

FIGURA 3. Conjunto de dados original (ref) e predições baseadas em rede neural para 118 (100%) e 47 (40%) pontos de amostragem de pH, Ca^{2+} e P, colunas representam as médias e barras mínimo e máximo.

O uso de limites inferior e superior é sugerido por Maldaner et al. (2020). Independentemente do tamanho dos dados, observamos que as predições para pH e Ca^{2+} mostraram uma variabilidade ligeiramente maior do que os dados originais (ref). Por outro lado,

os valores máximos das predições para P foram subestimados. Isso significa que seria necessário desenvolver uma arquitetura de rede neural específica para a variável P. No entanto, não estamos preocupados com a questão de melhorar a arquitetura, por exemplo, adicionando novos parâmetros (mais camadas ocultas e/ou neurônios), ou ainda ajustando os hiperparâmetros envolvidos na etapa de treinamento da rede. Em vez disso, tentamos ajustar o intervalo de predição em relação ao intervalo do conjunto de dados original. Na verdade, nesse caso, a variabilidade foi melhor descrita ajustando o limite superior aos valores máximos em vez dos valores do percentil 97,5 (Figura 4).

Em termos de avaliação interna (Figura 5), 95% dos *RMSE*s ficaram abaixo de ~0,7 (pH), ~1 $cmol_c\ dm^{-3}$ (Ca^{2+}) e ~100 mg dm^{-3} (P).

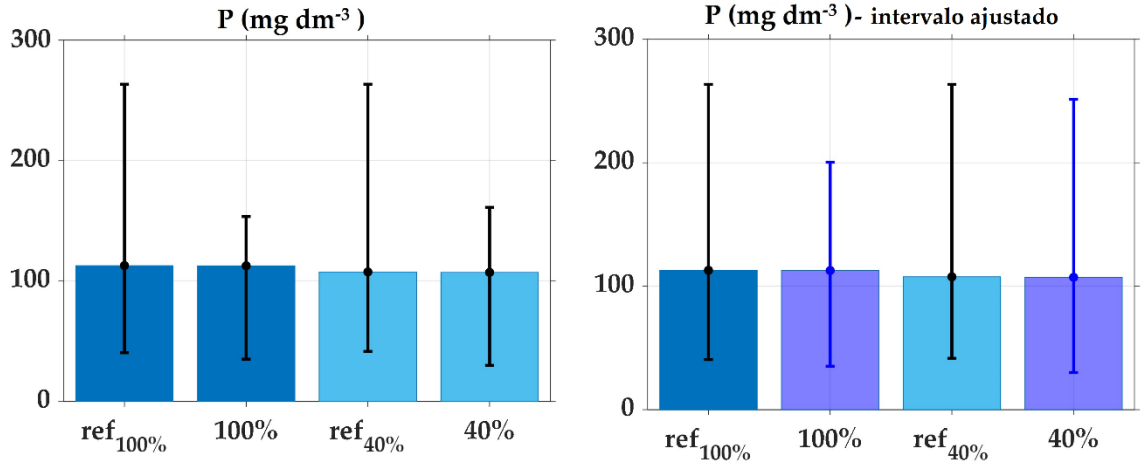

FIGURA 4. Predições do limite superior com base em α = 0,975 (esquerda) e limite superior dado pelos valores máximos preditos (direita) para P.

Em termos de avaliação externa (Figura 6), 90% dos *RMSE*s estavam abaixo de ~0,7 (pH), ~0,9 $cmol_c\ dm^{-3}$ (Ca^{2+}) e ~70 mg dm^{-3} (P). Estes resultados são consistentes com os da avaliação interna. Isso significa que o modelo de rede neural foi capaz de generalizar os resultados. Consequentemente, o uso de 47 pontos de amostragem para gerar um modelo de predição baseado em redes neurais é suficiente e um custo maior de amostragem em campo não seria necessário.

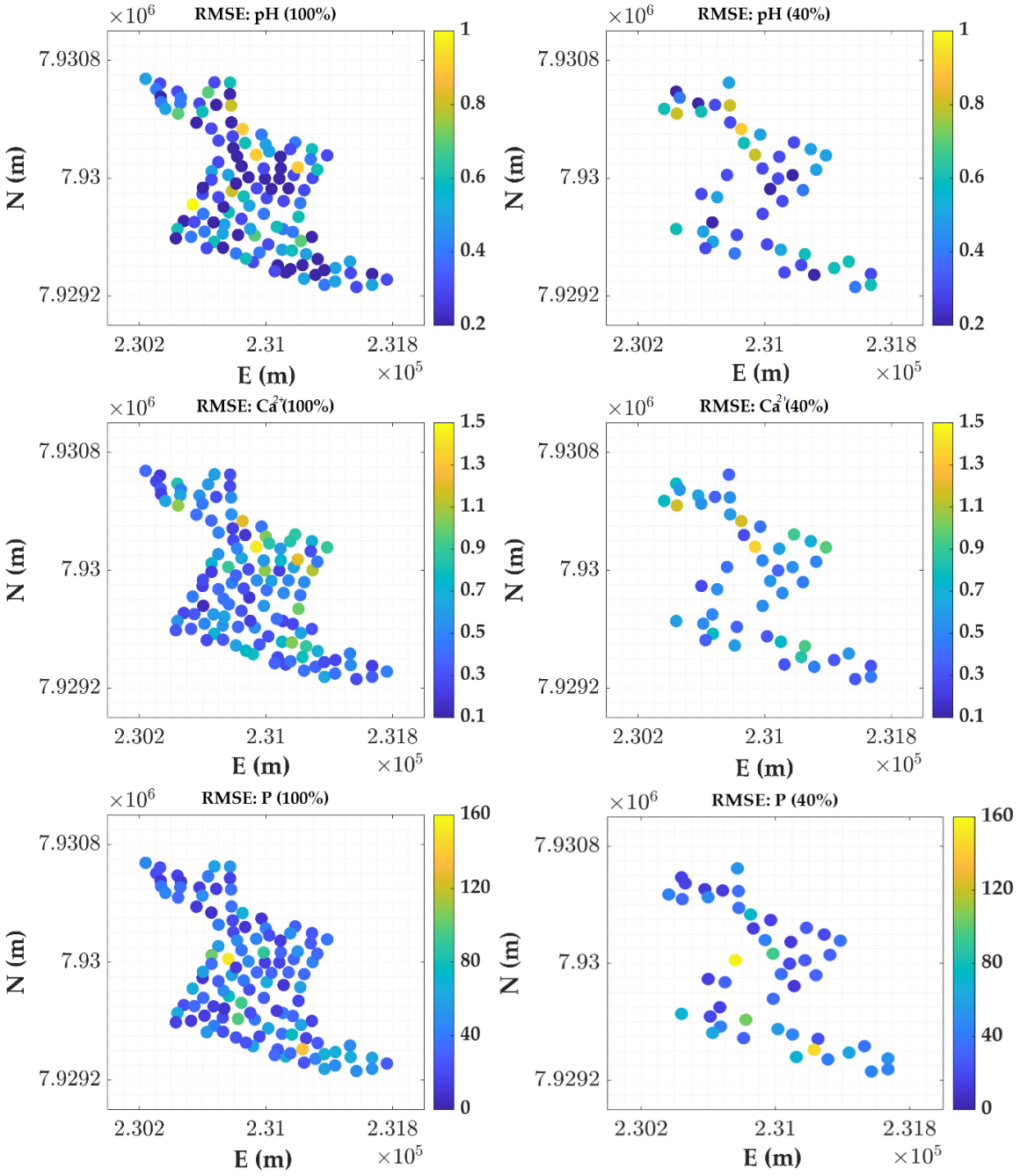

FIGURA 5. Distribuição espacial dos *RMSE*s dos pontos de amostragem de pH, Ca^{2+} ($cmol_c$ dm^{-3}) e P (mg dm^{-3}) para 118 pontos (esquerda: 100% disponível) e 47 pontos (direita: 40% disponível).

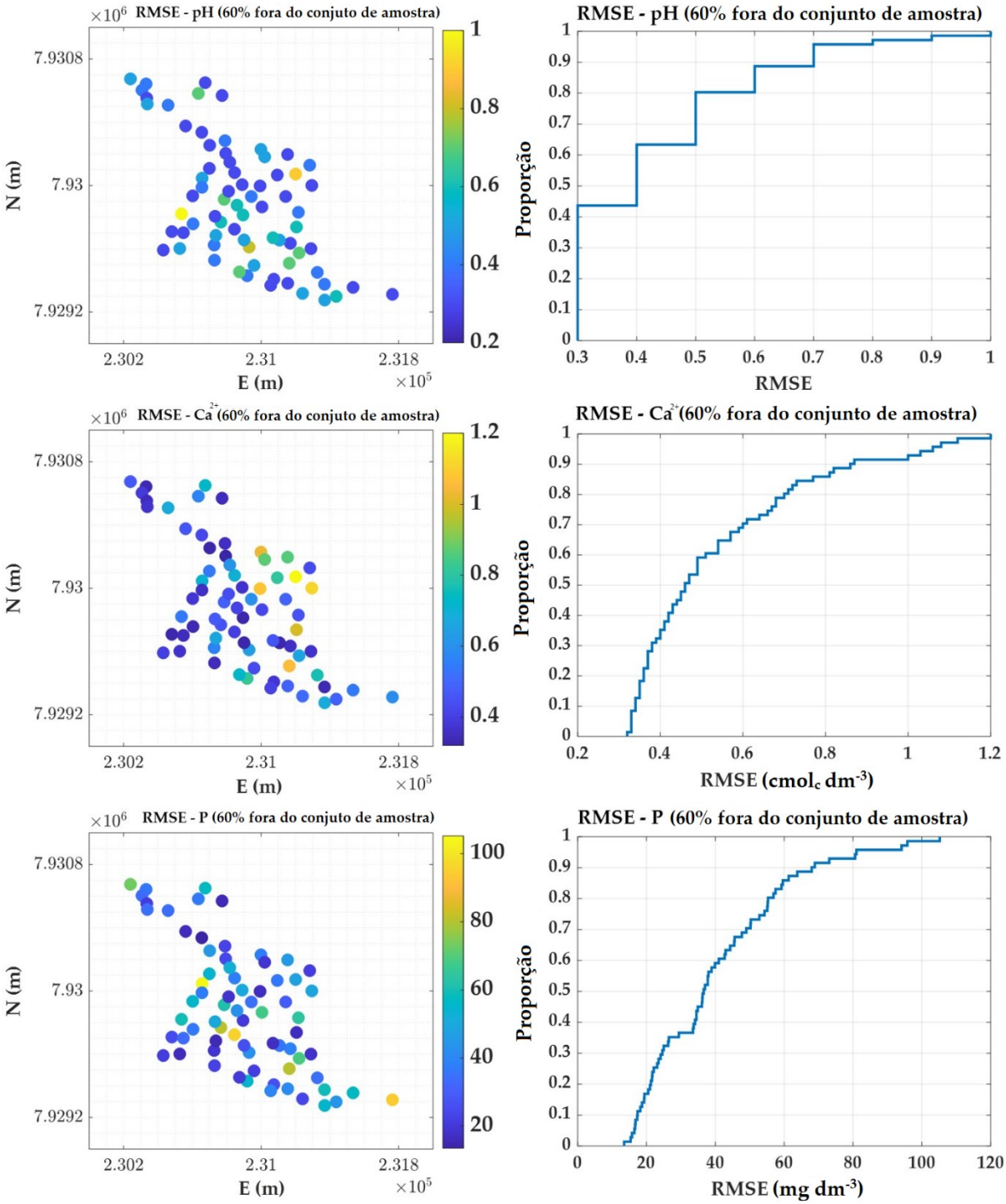

FIGURA 6. Erro dos Mínimos Quadrados (*RMSE*) distribuição espacial (esquerda) e proporção geral do *RMSE* (direita) do pH, Ca^{2+} e P calculado a partir da aplicação de 4700 preditores neurais gerados a partir da reamostragem de 40% dos dados (47 pontos) em 71 de – pontos fora da amostra (60%).

3.2. Avaliação Qualitativa

As predições de pH, Ca^{2+} e P (71 pontos: 60% dos dados que ficaram de fora da amostra) e seus valores originais (valores reais) foram classificados de acordo com a Tabela 2. A taxa de sucesso das predições foi calculada como sendo a razão entre o número de predições

corretamente identificadas para sua classe e o número total de pontos de amostragem. O valor mais provável de cada predição foi dado pelo valor médio, que foi utilizado para classificar o atributo de acordo com a Tabela 2. Lembramos que 4700 preditores neurais gerados a partir da reamostragem de 40% dos dados (47 pontos) foram aplicados a cada um dos 71 pontos (60%) que foram deixados de fora do procedimento de reamostragem (conjunto fora da amostra).

Os resultados são exibidos na Figura 7. As taxas de sucesso foram 72% (pH), 87% (Ca^{2+}) e 89% (P), respectivamente. Todas as médias das predições foram classificadas dentro da mesma classe. Isso significa que a rede neural baseada em reamostragem mostrou não ser sensível a mudanças abruptas, que podem ser interpretadas como outliers. O conjunto de dados original também teve uma predominância de uma mesma classe, embora existam alguns pontos anômalos.

TABELA 2. Classificação do pH em água, Ca^{2+} ($cmol_c\ dm^{-3}$) e P ($mg\ dm^{-3}$) por resina de troca para cultura perene.

Atributo	Classe	Valores
pH	Muito Baixo	< 4,5
	Baixo	4,5 – 5,4
	Adequado	5,5 – 6,0
	Alto	6,1 – 7,0
	Muito Alto	> 7,0
Ca^{2+}	Muito Baixo	≤ 0,40
	Baixo	0,41 – 1,20
	Médio	1,21 – 2,40
	Alto	2,41 – 4,00
	Muito Alto	> 4,00
P	Muito Baixo	≤ 5
	Baixo	6 – 12
	Médio	13 – 30
	Alto	31 – 60
	Muito Alto	> 60

Classificação de pH, Ca^{2+} e P baseada em Alvarez et al. (1999) e Raij et al. (1996), respectivamente.

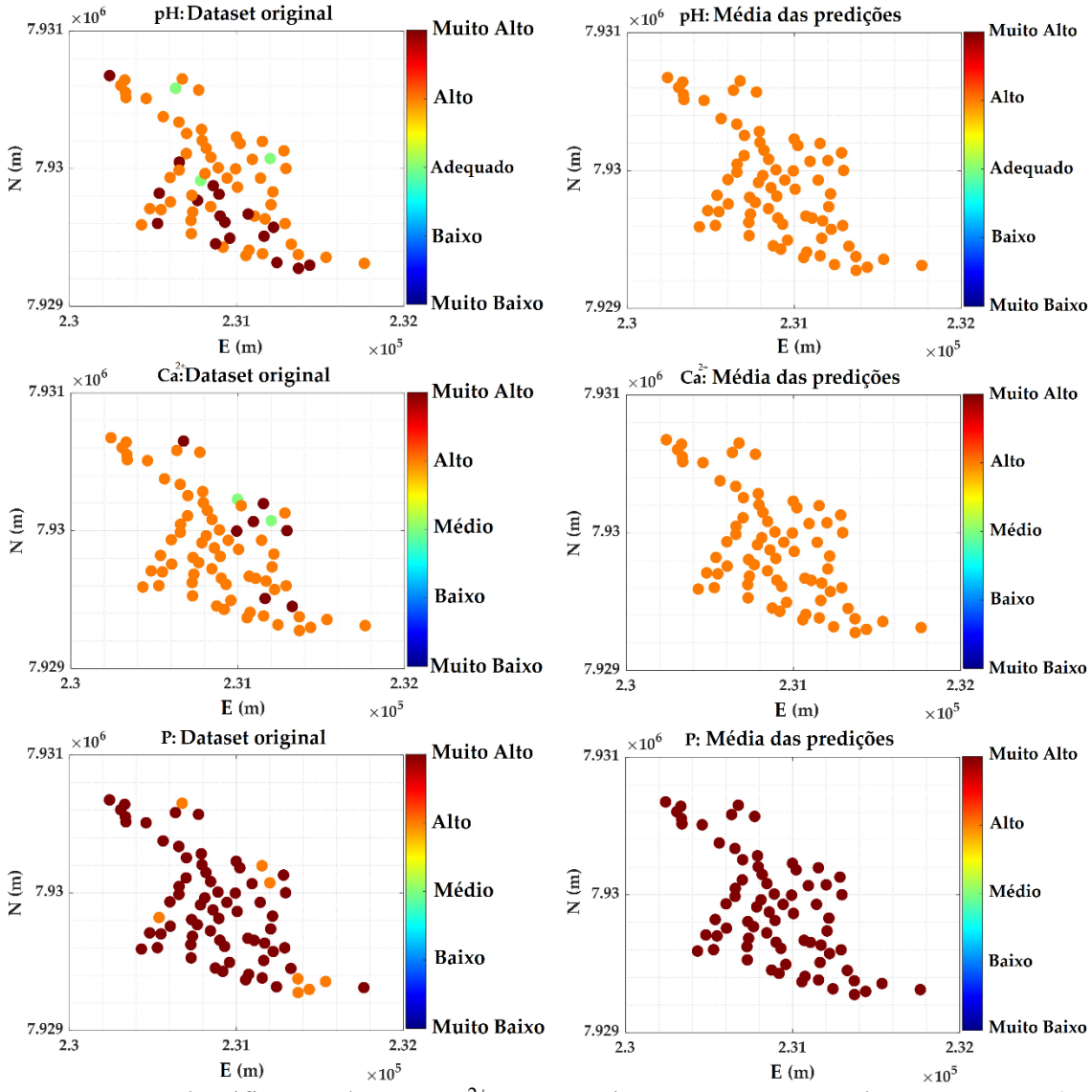

FIGURA 7. Classificação do pH, Ca^{2+} e P com base em 71 pontos de amostragem (60% do conjunto fora da amostra) para o conjunto de dados original (à esquerda) e as predições Jack-1T com base na média (à direita).

Como temos à nossa disposição a incerteza de cada ponto de amostragem, também é possível analisar em termos do intervalo de confiança ao invés de apenas considerar a média de cada predição. A distribuição espacial de ambos os desvios-padrão (σ) e sua distribuição para as predições são exibidas na Figura 8. Nesse caso, as taxas de sucesso foram baseadas no intervalo de confiança dado pela média ± EP (Erro Provável). O EP foi definido como sendo ±2σ para pH, ±3σ para Ca^{2+} e ±4,4σ para P, no qual σ é o desvio padrão das predições. Essas escolhas de EP foram baseadas no comportamento de variabilidade descrito pelos resultados dos 4700 preditores neurais gerados a partir da reamostragem de 40% dos dados (47 pontos), como pode ser visto na seção anterior (Figura 4 e 5). A taxa de sucesso ocorre quando o valor do campo original está dentro do intervalo de confiança predito. Para este caso, as taxas de sucesso foram de 92%, 93% e 92% para pH, Ca^{2+} e P, respectivamente.

O intervalo de confiança ficou exatamente entre [6; 7,2] para todas as predições de pH, o que corresponde às classes (Tabela 2) “Adequado” (limite inferior de -2σ) e “Alto” (limite superior de +2σ), respectivamente. Por outro lado, todos os limites superiores das predições de Ca^{2+} (+3σ) foram classificadas como “Muito Alto”, mas os limites inferiores (-3σ) foram classificados como “Médio” por 30 pontos (42%) e os outros 41 pontos (58%) como “Alto”. Os limites superiores (+4,4σ) para P foram classificados como "Muito alto", enquanto os limites inferiores (-4,4σ) foram classificados como “Alto” por 58 pontos (82%), “Médio” por 6 pontos (8%) e “Muito Baixo” por 7 pontos (10%).

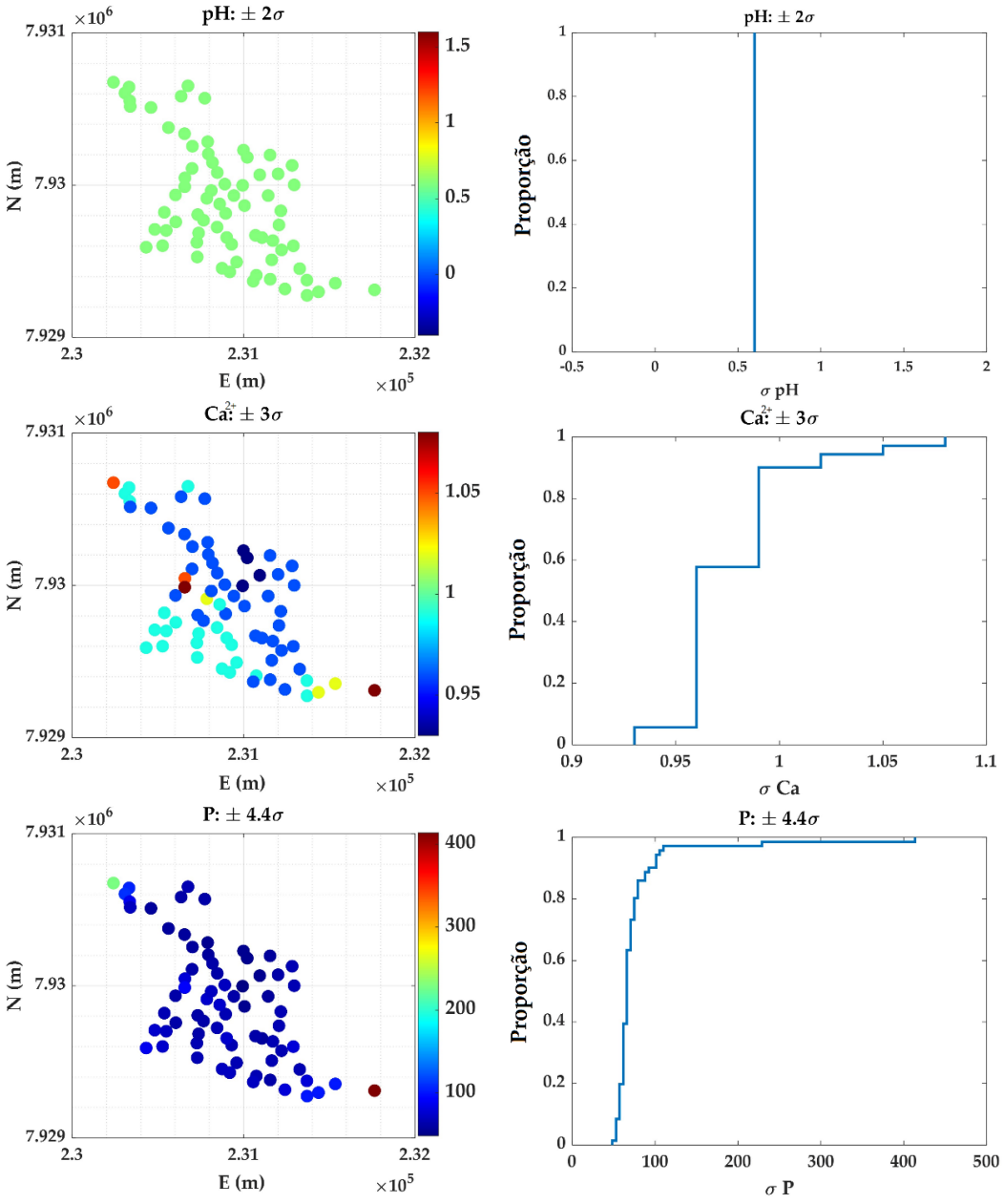

FIGURA 8. Incertezas das predições de pH, Ca^{2+} e P com base em ±2σ; ±3σ e ±4,4σ probabilidade, respectivamente, para os 71 pontos de amostragem (60% do conjunto fora da amostra).

4. CONCLUSÃO

Aqui, apresentamos um método de reamostragem em redes neurais com aplicação em dados espaciais.

O método proposto Jack-1T pode ser usado para a construção de procedimentos inferenciais na análise estatística moderna de dados espaciais. O método substitui as derivações teóricas exigidas na aplicação de métodos tradicionais, reamostrando repetidamente os dados originais e fazendo inferências a partir das novas amostras. A vantagem desta técnica é que ela nos permite determinar a distribuição de probabilidade empírica para o preditor, em vez de simplesmente fornecer uma predição pontual. Consequentemente, o desempenho de predição para cada ponto de amostragem individual está disponível.

Os resultados, baseados na simulação do tamanho dos conjuntos de dados, mostraram que uma redução de 60% nos dados disponíveis oferece acurácia compatível em relação ao conjunto completo e, portanto, reduz o custo de amostragem em campo. Portanto, o método proposto auxilia na tomada de decisão para definir o tamanho da amostra em campo.

O método de reamostragem caracteriza espacialmente os pontos de maior ou menor acurácia e incerteza, diferente dos métodos usuais que apresentam apenas predições pontuais e, além disto, aumenta a taxa de sucesso usando a predição de intervalo em vez de usar a média.

Apesar de restringi-lo ao modelo de rede neural e para dados espaciais de atributos químicos do solo, surge como proposta estender o método de reamostragem Jack-1T a outras ferramentas estatísticas modernas, como Krigagem, Método dos Mínimos Quadrados e outros dados espaciais como temperatura, pressão, umidade, altitude, IWV (*Integrated Water Vapor*), PWV (*Precipitable Water Vapor*), gradientes ionosféricos e assim por diante.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVAREZ, V. V. H.; NOVAIS, R. F.; BARROS, N. F.; CANTARUTTI, R. B.; LOPES, A. S. Interpretação dos resultados das análises de solos. In: Ribeiro, A.C.; Guimarães, P.T.G.; Alvarez, V.V.H., eds. Recomendações para o uso de corretivos e fertilizantes em Minas Gerais 5ª Aproximação. **Comissão de Fertilidade do Solo do Estado de Minas Gerais**. Viçosa, MG, BR. 1999. p. 25-32. Acesso em: 05 de janeiro de 2021.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. **Biometrika**. 1989. v.76, n.3: 503-514. DOI: doi.org/10.2307/2336116. Acesso em: 12 de janeiro de 2021.

CAGLIARI, J.; VERONEZ M. R.; ALVES M. E. Remaining phosphorus estimated by pedotransfer function. **Revista Brasileira Ciências Solo**. 2011. v.35, n.1: p. 203-212. DOI: dx.doi.org/10.1590/S0100-06832011000100019. Acesso em: 25 de janeiro de 2021.

EFRON, B. The jackknife, the bootstrap, and other resampling plans. **Technical Report Stanford University**, Stanford, CA, USA. 1980. v.63. Disponível em: <statistics.stanford.edu/sites/g/files/sbiybj6031/f/BIO%2063.pdf> Acesso em: 05 de fevereiro de 2021.

EFRON, B. The Jackknife, the Bootstrap and Other Resampling Plans. **Society for Industrial and Applied Mathematics**, Stanford University, Stanford, CA, USA. 1982. DOI: https://doi.org/10.1137/1.9781611970319. Acesso em: 05 de dezembro de 2020.

EFRON, B. Jackknife-after-bootstrap standard errors and influence functions. **Journal of the Royal Statistical Society: Series B (Methodological)**. 1992. v.54, n.1: p. 83-111. Disponível em: <www.jstor.org/stable/2345949>. Acesso em: 14 de dezembro de 2020.

GAVIN, H. P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. **Duke University: Department of Civil and Environmental Engineering**, Durham, NC, USA. 2020. Disponível em: <https://people.duke.edu/~hpgavin/ce281/lm.pdf> Acesso em: 24 de maio de 2021.

HAYKIN, S. Neural networks A comprehensive foundation. **Prentice Hall**, Hoboken, NJ, USA. 1999. 2ed. Acesso em: 30 de janeiro de 2020.

HUANG, L.; QIN, J.; ZHOU, Y.; ZHU, F.; LIU, L.; SHAO, L. Normalization Techniques in Training DNNs: Methodology, Analysis and Application. **ArXiv, abs/2009. 12836**. 2020. Disponível em: <arXiv:2009.12836>. Acesso em: 09 de março de 2021.

LEVENBERG, K. A method for the solution of certain non-linear problems in least squares. **Quarterly of Applied Mathematics**. 1944. v.2, n.2: 164-168. Disponível em: <www.jstor.org/stable/43633451> Acesso em: 23 de janeiro de 2021.

MADSEN, K.; NIELSEN, H. B.; TINGLEF, O. Methods for nonlinear least squares problems. **Informatics and Mathematical Modelling Technical University of Denmark**, Copenhagen, Denmark. 2004. 2ed. Disponível em: <http://www2.imm.dtu.dk/pubdb/edoc/imm3215.pdf>. Acesso em: 23 de maio de 2021.

MAIER, H. R.; JAIN, A.; DANDY, G. C.; SUDHEER, K. P. Methods used for the development

of neural networks for the prediction of water resource variables in river systems: Current status and future directions. **Environmental Modelling & Software**. 2010. v25, n.8: p. 891-909. DOI: https://doi.org/10.1016/j.envsoft.2010.02.003. Acesso em: 03 de julho de 2021.

MALDANER, L. F.; MOLIN, J. P.; SPEKKEN, M. Methodology to filter out outliers in high spatial density data to improve maps reliability. **Scientia Agricola.** 2020. v.79, n.1: e20200178. DOI: doi.org/10.1590/1678-992X-2020-0178. Acesso em: 03 de junho de 2021.

MARQUARDT, D. "An algorithm for least squares estimation of nonlinear parameters". **Journal of the Society for Industrial and Applied Mathematics**. 1963. V.11: p. 431–441. Society for Industrial & Applied Mathematics (SIAM). Acesso em: 23 de junho de 2021. DOI: http://epubs.siam.org/doi/abs/10.1137/0111030.

MILLER, R. G. The jackknife–a review. **Biometrika**. 1974. v.61, n.1: p. 1-15. DOI: doi.org/10.2307/2334280. Acesso em: 02 de junho de 2021.

NGUYEN, D. and WIDROW, B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. **International Joint Conference on Neural Networks**. 1990. vol.3: p. 21-26. DOI: https://doi.org/10.1109/IJCNN.1990.137819. Acesso em: 30 de janeiro de 2021.

PAN, L. Resampling in neural networks with application to financial time series. **Ph.D. Thesis, The University of Guelph**, ON, CA. 1998. Disponível em: <www.collectionscanada.gc.ca/obj/s4/f2/dsk2/ftp02/NQ47406.pdf> Acesso em: 23 de abril de 2021.

QUENOUILLE, M. H. Approximate tests of correlation in time-series. **Journal of the Royal Statistical Society: Series B (Methodological)**. 1949. v.11, n.1: p. 68-84. Disponível em: <www.jstor.org/stable/2983696> Acesso em: 28 de maio de 2021.

RAIJ, B. V.; CANTARELLA, H.; QUAGGIO, J. A.; FURLANI, A. M. C. Recomendações de adubação e calagem para o Estado de São Paulo, (Boletim técnico, 100). **Instituto Agronómico e Fundação IAC**, Campinas, SP, BR.1996. 2ed. p. 8-9. Acesso em: 30 de abril de 2021.

REITERMANOVÁ, Z. Data splitting. **WDS'10 Proceedings of Contributed Papers, Part I: Mathematics and Computer Sciences.** Charles University, Faculty of Mathematics and Physics, Prague, CZ. 2010. p. 31-36. eds. Disponível em: <www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf> Acesso em: 04 de janeiro de 2021.

ROFATTO, V. F.; MATSUOKA, M. T.; KLEIN, I.; VERONEZ, M. R.; SILVEIRA, L. G. A monte carlo-based outlier diagnosis method for sensitivity analysis. **Remote Sensing**. 2020. v.12, n.5. DOI: doi.org/10.3390/rs12050860. Acesso em: 05 de maio de 2021.

TEIXEIRA P. C.; DONAGEMMA G. K.; FONTANA A. I.; TEIXEIRA W. G. Manual de métodos de análise de solo. In: EMBRAPA (Ed.), **Embrapa Solos**. Brasília, DF. 2017. 573 p. Disponível em: <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1107360/1/Pt4Cap3Analisemineralogicadasfracoesargilaesilte.pdf>. Acesso em: 11 de julho de 2020.

WANG, L. and YU, F. Jackknife resampling parameter estimation method for weighted total least squares. **Communications in Statistics - Theory and Methods**. 2020. v.49, n.23: p.

5810-5828. DOI: https://doi.org/10.1080/03610926.2019.1622725. Acesso em: 11 de julho de 2020.

WU, W.; MAY, R.; DANDY, G. C.; MAIER, H. R. A method for comparing data splitting approaches for developing hydrological ANN models. In: The 6th International Congress on **Environmental Modelling and Software**. 2012. Leipzig, Germany. Disponível em: <https://scholarsarchive.byu.edu/iemssconference/2012/Stream-B/394/>. Acesso em: 02 de julho de 2020.

YADAV, S. and SHUKLA, S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. **2016 IEEE 6th International Conference on Advanced Computing (IACC)**. 2016. p.78-83. DOI: 10.1109/IACC.2016.25. Acesso em: 23 de junho de 2021.

ZIGGAH, Y. Y.; YOUJIAN, H.; TIERRA, A. R.; LAARI, P. B. Coordinate transformation between global and local datums based on artificial neural network with k-fold cross-validation: A case study, Ghana. **Earth Sciences Research Journal**. 2019. v.23, n.1: p. 67-77. DOI: https://doi.org/10.15446/esrj.v23n1.63860. Acesso em: 23 de dezembro de 2020.