**UNIVERSIDADE FEDERAL DE VIÇOSA**

**PATRÍCIA DE SOUSA ILAMBWETSI**

# PROPOSTA DE UM INTERPOLADOR GEOESTATÍSTICO HÍBRIDO COM APRENDIZADO DE MÁQUINA

**VIÇOSA - MINAS GERAIS**
**2020**

**PATRÍCIA DE SOUSA ILAMBWETSI**

# PROPOSTA DE UM INTERPOLADOR GEOESTATÍSTICO HÍBRIDO COM APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Doctor Scientiae.

Orientador: Gérson Rodrigues dos Santos

Coorientadores: João Marcos Louzada
Paulo César Emiliano

**VIÇOSA - MINAS GERAIS**
**2020**

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa**

T
I27p
2020

Ilambwetsi, Patrícia de Sousa, 1984-
Proposta de um interpolador geoestatístico híbrido com aprendizado de máquina / Patrícia de Sousa Ilambwetsi. – Viçosa, MG, 2020.
96 f. : il. (algumas color.) ; 29 cm.

Orientador: Gerson Rodrigues dos Santos.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise multivariada. 2. Inteligência artificial. 3. *Bertholletia Excelsa*. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.535

**PATRÍCIA DE SOUSA ILAMBWETSI**

# PROPOSTA DE UM INTERPOLADOR GEOESTATÍSTICO HÍBRIDO COM APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Doctor Scientiae.

APROVADA: 23 de novembro de 2020.

Assentimento:

A Deus o autor da vida e
ao meu esposo
companheiro de todas
as horas...

## AGRADECIMENTOS

A Deus, pela promessa; por iluminar os meus pensamentos, guiar os meus caminhos e por tornar tudo possível;

Ao meu esposo Archange, por acreditar, incentivar, compreender, dizer a todo o momento que sou capaz; pelas orações; pela presença; por fazer parte desta conquista;

Aos meus pais, pelo incentivo, fé e presença em todos os momentos;

À minha irmã e ao Be, por acreditarem e incentivarem;

Ao papa André José e a maman Alfreda, pelas orações, pelos incentivos, por acreditarem;

Ao professor Dr. Gérson pela orientação, por acreditar e tornar possível este trabalho;

Ao Jandresson, pela ajuda com a programação do software R;

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para a minha formação;

Aos professores membros da banca, pela contribuição e disponibilidade para a leitura deste trabalho;

Aos meus coorientadores, professores Dr. João Marcos Louzada e Dr. Paulo César Emiliano pelos ensinamentos;

A Rosane, pelos momentos de estudo com muito café e paciência;

Aos amigos do doutorado, pelo incentivo, pela paciência e pelos trabalhos em equipe;

As meninas da república, em especial, a Agnes e a Vanessa, por cada momento compartilhado que fizeram parte dessa história;

A todos os amigos, que de alguma forma, contribuíram para este sonho ser real;

À Universidade Federal de Viçosa, pela oportunidade e pela infraestrutura disponibilizada;

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Enfim, a todos que fazem parte da minha vida e torcem pelo meu sucesso, o meu

***"MUITO OBRIGADA"!***

"Ao Senhor Jesus seja toda a honra, toda a glória e todo o louvor..."

## RESUMO

ILAMBWETSI, Patrícia de Sousa, D.Sc., Universidade Federal de Viçosa, novembro de 2020. P**roposta de um interpolador geoestatístico híbrido com aprendizado de máquina.** Orientador: Gérson Rodrigues dos Santos. Coorientadores: João Marcos Louzada e Paulo César Emiliano.

A krigagem tem sido um método univariado muito utilizado na literatura para interpolação de dados. Entretanto, apresenta a desvantagem de ser computacionalmente inviável para modelar o estimador de semivariograma em grandes conjuntos de dados e descartar variáveis importantes no estudo pela presença do efeito pepita puro. Para solucionar essas desvantagens e melhorar a capacidade de predição desse interpolador, apresenta-se nesse trabalho, um estudo que envolve a metodologia da Geoestatística com aprendizado de máquina para implementar, computacionalmente, um interpolador híbrido capaz de modelar, em uma abordagem multivariada, a influência da variabilidade espacial de todas as variáveis presentes no estudo na predição da variabilidade espacial da variável de interesse, sem a restrição ao número de variáveis e ao tamanho do conjunto de dados. E, para fins de comparação, foi realizada via coeficiente erro quadrático médio (EQM) e coeficiente de determinação (R2) uma análise para verificar o desempenho do interpolador implementado. Para isso, foram coletadas amostras do solo de 50m×30m em todas as linhas da região do estudo e amostras da produção média das castanheiras, no período 2007 a 2015. As análises estatísticas e geoestatísticas foram realizadas no ambiente computacional do software R e todos os pontos foram georreferenciados. Como resultado, obteve-se não só um aprimoramento do ajuste do modelo implementado e uma redução significativa para erro quadrático médio, bem como, o detalhamento do grau de importância de cada atributo do solo para predizer a variabilidade espacial da produção média das Castanheiras-da-amazônia.



**Palavras-chave:** Random Forest. FRK. Inteligência Artificial. Bertholletia excelsa. Análise Multivariada.

## ABSTRACT

ILAMBWETSI, Patrícia de Sousa, D.Sc., Universidade Federal de Viçosa, November, 2020. P**roposal for a hybrid geostatistical interpolator with machine learning.** Advisor: Gérson Rodrigues dos Santos. Co-advisors: João Marcos Louzada and Paulo César Emiliano.

A kriging has been a univariate method widely used in the literature for data interpolation; however, it presents a disadvantage of being computationally unfeasible to model the semivariogram estimator in large data sets and to discard important variables in the study due to the presence of the pure nugget effect. In order to solve these disadvantages and improve the interpolator's predictive capacity, this research presents a study involving the geostatistics methodology with machine learning to implement, computationally, a hybrid interpolator capable of defining, in a multivariate approach, the degree of importance of each variable under study to predicting the spatial variability of the interest's variable, without restriction on the number of variables and the size of the data set. And, for comparison purposes, an analysis was performed through mean square error coefficient (EQM) and determination coefficient (R2) to verify the performance of the implemented interpolator. For that, samples of soil of 50 × 30m were collected in all lines of the study region and samples of the average production of chestnut trees in the period 2007 to 2015. Statistical and geostatistical analyzes were performed in the computational environment of the R software and all points were georeferenced. As a result, a perfect fit of the model was obtained and a significant reduction for mean squared error when using the implemented hybrid interpolator, as also, the degree of importance of each soil attribute to predict the spatial variability of the average production of Chestnuts of the Amazon.



**Keywords:** Random Forest. FRK. Random Forest. Artificial intelligence. Bertholletia excelsa. Multivariate Analysis.

## LISTA DE FIGURAS

### INTRODUÇÃO GERAL



### CAPÍTULO 1

**CAPÍTULO 2**

## LISTA DE TABELAS

### CAPÍTULO 1



### CAPÍTULO 2

**SUMÁRIO**

# INTRODUÇÃO GERAL

Na década de 1950 na África do Sul, o engenheiro de minas Daniel G. Krige, ao trabalhar com dados de mineração, sentiu a necessidade de levar em consideração as distâncias entre as amostras para estudar as variâncias estimadas. Dessa forma, a Geoestatística surgiu para avaliar a variabilidade e a dependência espacial das variáveis.

A Geoestatística é um ramo da Estatística Espacial que utiliza o conceito de variáveis regionalizadas na avaliação de variabilidade espacial (Matheron, 1971; Ferreira et al., 2013). Não se limita apenas em obter um modelo de dependência espacial, mas pretende-se também predizer os valores de pontos nos locais onde não foram amostrados, utilizando-se a teoria do semivariograma e da interpolação por krigagem (Goovaerts, 1997).

O semivariograma é uma ferramenta fundamental para estudar a similaridade entre as amostras vizinhas, de maneira que, as observações mais próximas, geograficamente, apresentam um comportamento mais semelhante entre si do que aquelas separadas por maiores distâncias (Santos et al., 2011). Em termos matemáticos, o semivariograma é a média ao quadrado das diferenças entre os valores de pontos amostrados em uma área estudada, separados pelo vetor distância $h$ (Andriotti, 2013).

Segundo Vieira (2000), o estimador do semivariograma é dado por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \qquad (1)$$

em que: $\hat{\gamma}(h)$ é o valor estimado da semivariância para um vetor de distância h; $Z(x_i)$ é o valor observado da variável no ponto $x_i$; $Z(x_i + h)$ é o valor observado no ponto $x_i + h$; $h$ é o vetor de distância entre os pares de casos amostrados; $N(h)$ é o número de pares de pontos separados entre si por h.

Quando $h$ cresce, o semivariograma aproxima-se da variabilidade total dos dados e, havendo estacionariedade de segunda ordem, o semivariograma expressa o grau de dependência entre os pontos amostrais (Opromolla et al., 2006).

Para uma dada posição fixa $x_i$, dentro de uma área D, cada valor medido da variável em estudo pode ser considerado como uma realização de um conjunto de variável aleatória $Z(x_i)$. Portanto, considere que a função aleatória $Z(x_i)$ apresenta $E\big(Z(x_i)\big) = m(x_i)$, $E\big(Z(x_i + h)\big) = m(x_i + h)$, $Var\big(Z(x_i)\big)$ e $Var\big(Z(x_i + h)\big)$ para os locais $x_i$ e $x_i + h$,

separados por h. Então, a covariância $C(x_i, x_i + h) = E[Z(x_i)Z(x_i + h)] - m(x_i)m(x_i + h)$, o variograma $2\gamma(x_i, x_i + h) = E[Z(x_i) - Z(x_i + h)]^2$, a variância de $Z(x_i)$ é igual a $Var\big(Z(x_i)\big) = E[Z(x_i)Z(x_i + 0)] - m(x_i)m(x_i + 0) = E[Z^2(x_i)] - m^2(x_i) = C(x_i, x_i)$ e a variância de $Z(x_i + h)$ é dada por $Var\big(Z(x_i + h)\big) = E[Z^2(x_i + h)] - m^2(x_i + h) = C(x_i + h, x_i + h)$

Sob a hipótese de estacionaridade de segunda ordem, temos $E\big(Z(x_i)\big) = m$ para qualquer $x_i$ dentro da área D, então:

$$C(h) = E[Z(x_i)Z(x_i + h)] - m^2 \quad (2)$$

$$Var\big(Z(x_i)\big) = E[Z^2(x_i)] - m^2 = C(0) \quad (3)$$

desenvolvendo o variograma temos:

$$2\gamma(x_i, x_i + h) = 2\gamma(h) = E\big(Z^2(x_i) - 2Z(x_i)Z(x_i + h) + Z^2(x_i + h)\big)$$

somando e subtraindo $2m^2$

$$2\gamma(h) = E(Z^2(x_i) - m^2 - 2Z(x_i)Z(x_i + h) + 2m^2 + Z^2(x_i + h) - m^2)$$

$$2\gamma(h) = E[Z^2(x_i)] - m^2 - 2(E[Z(x_i)Z(x_i + h)] - m^2) + E[Z^2(x_i + h)] - m^2 \quad (4)$$

substituindo (2) e (3) na equação (4), temos:

$$2\gamma(h) = C(0) - 2C(h) + C(0) \quad (5)$$

A partir da simplificação da equação (5), obtêm-se o semivariograma que pode ser escrito em função da matriz de covariância das distâncias, dado por $\gamma(h) = C(0) - C(h)$ e isolando $C(h)$, temos:

$$C(h) = C(0) - \gamma(h)$$

Portanto, se a hipótese de estacionaridade de segunda ordem for satisfeita, o semivariograma $\gamma(h)$ e a covariância $C(h)$ são ferramentas equivalentes para caracterizar a dependência espacial.

O semivariograma pode ser representado por um gráfico das semivariâncias que expressa à variabilidade espacial entre as amostras, sendo uma função que só depende das distâncias $h$ entre os pares de pontos amostrados (Isaaks e Srivastava, 1989).

O semivariograma pode ser construído de duas formas: semivariograma experimental que é obtido através da amostragem; e, semivariograma teórico, que é obtido através do ajuste de modelos teóricos ao semivariograma experimental (Figura 1).

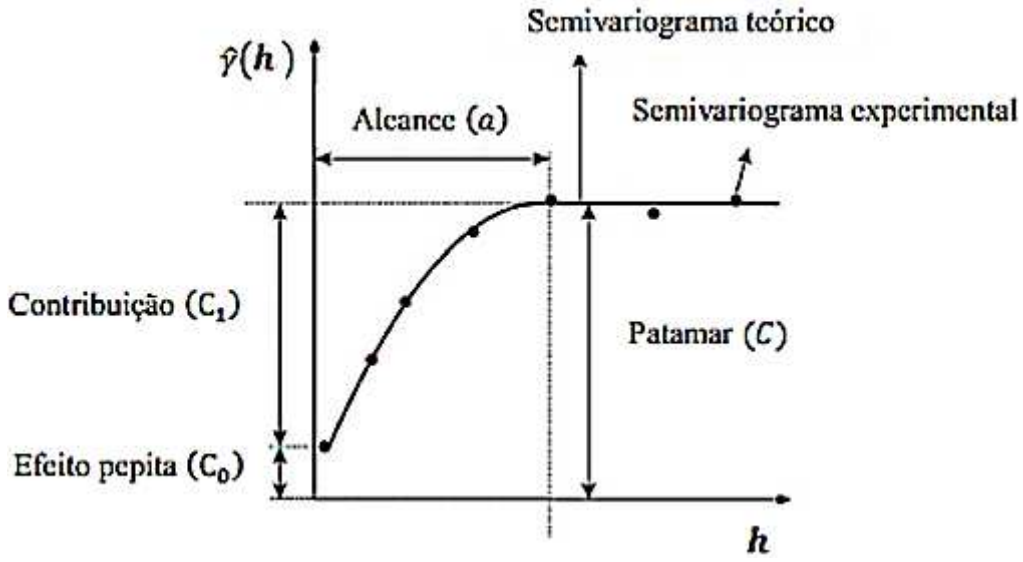


Figura 1: Apresentação gráfica de um semivariograma.
Fonte: Câmara e Medeiros (1998)

O ajuste do modelo teórico ao semivariograma experimental resulta na estimação de alguns parâmetros, conforme a Figura 1, definidos por Isaaks e Srivastava (1989) por:

- Efeito pepita $(C_0)$: é o valor da semivariância para a distância zero;
- Patamar $(C = C_0 + C_1)$: é o valor da variância em que o semivariograma se estabiliza;
- Contribuição $(C_1)$: representa a diferença entre o patamar e o efeito pepita e, refere-se, ao valor total da contribuição da variabilidade da dependência espacial ou, ainda, a variabilidade máxima captada entre os pares de pontos amostrais;
- Alcance $(a)$: é à distância dentro da qual os valores amostrais apresentam-se correlacionadas espacialmente, ou seja, é a distância limite de dependência espacial.

Diferentes modelos teóricos podem ser ajustados a um semivariograma experimental. Contudo, os mais utilizados são os modelos esférico, exponencial e gaussiano, apresentados nas equações (6), (7) e (8), respectivamente.

$$\gamma(\boldsymbol{h}) = \begin{cases} 0 & ,h = 0 \\ C_0 + C_1\left[1{,}5\left(\frac{h}{a}\right) - 0{,}5\left(\frac{h}{a}\right)^3\right] & ,0 < h < a \\ C_0 + C_1 & ,h \geq a \end{cases} \quad (6)$$

$$\gamma(\boldsymbol{h}) = \begin{cases} 0 & ,h = 0 \\ C_0 + C_1\left[1 - exp\left(-\frac{3h}{a}\right)\right] & ,h \neq 0 \end{cases} \quad (7)$$

$$\gamma(\boldsymbol{h}) = \begin{cases} 0 & ,h = 0 \\ C_0 + C_1\left[1 - exp\left(-\frac{3h^2}{a^2}\right)\right] & ,h \neq 0 \end{cases} \quad (8)$$

Yamamoto e Landim (2013) definem que quando o comportamento do semivariograma ocorre mais intensamente numa direção e menos em outra, este é dito ser anisotrópico, caso contrário, é dito ser isotrópico. Rosa (2017) sugere que a anisotropia deve ser tratada por um modelo direcional de semivariograma e, nos casos de isotropia, a variabilidade é simétrica em qualquer direção (Yamamoto e Landim, 2013).

A partir da escolha do modelo teórico para o ajuste do semivariograma, pode-se proceder para a interpolação geoestatística, conhecida como Krigagem. Esse método leva em consideração a dependência espacial existente entre os valores dos pontos amostrados e não amostrados, bem como a distância entre tais pontos. Também, permite a interpolação de valores em qualquer posição da área em estudo, sem tendência e com variância mínima, desde que seja conhecido o semivariograma e que haja dependência espacial entre as amostras (Vieira, 2000).

A Krigagem, em geral, é melhor do que os demais métodos de interpolação numérica, pois sua metodologia é baseada na função do semivariograma que depende da existência ou não do efeito pepita, da amplitude e da anisotropia e, não apenas, na existência da função das distâncias entre pontos (Yamamoto e Landim, 2013).

## 1.1 Krigagem

A krigagem é um método de interpolação que possibilita predizer valores não amostrados $\hat{Z}(s_0)$ em qualquer local $s_0$ por meio das informações dos $n$ pontos vizinhos amostrados $Z(s_i)$ na localização $s_i, i = 1, \ldots, n$. A formulação apresentada pela Equação (9) é

um exemplo de krigagem cujo preditor é uma combinação linear dos valores amostrados, contudo, temos outras krigagens presentes na literatura.

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \alpha_i Z(s_i) \tag{9}$$

De acordo com Yamamoto e Landim (2013), os valores obtidos nos pontos amostrais $Z(s_i)$ são usados na interpolação geoestatística para fornecer uma grade regular que descreve a modelagem da distribuição e a variabilidade espacial da variável de interesse (Figura 2).

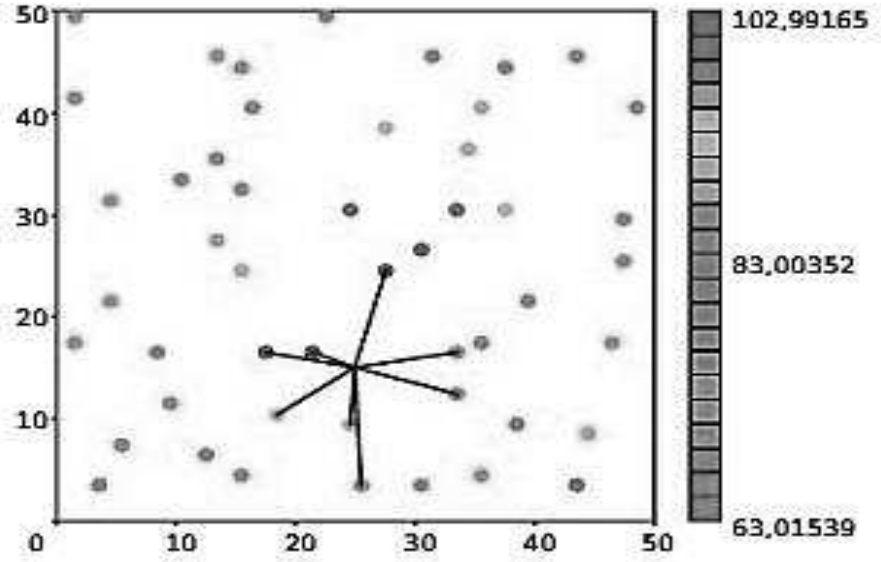


Figura 2: Localização de pontos vizinhos mais próximos para estimativa de pontos não amostrados.
Fonte: Yamamoto e Landim (2013)

A krigagem é considerada o melhor método de predição, pois produz predições não viesadas e com variância mínima (Isaaks e Srivastava, 1989) que, segundo Yamamoto e Landim (2013), a minimização da variância contribui para suavizar a variabilidade espacial da variável em estudo. Além disso, é possível citar alguns detalhes que a diferencia de outros métodos de interpolação tais como a estimação de uma matriz de covariância espacial que determina os pesos atribuídos às diferentes amostras, o tratamento da redundância dos dados, a vizinhança a ser considerada no procedimento de inferência e o erro associado ao valor estimado.

Na literatura são encontrados diversos tipos de krigagem, tais como krigagem simples, krigagem ordinária, krigagem universal, krigagem fatorial, entre outras. Dentre

todas, a krigagem ordinária é a mais utilizada por não exigir o conhecimento sobre a média estacionária (Santos et al., 2011).

### 1.1.1 Krigagem Ordinária

A krigagem ordinária é um preditor linear univariado que se baseia na obtenção de estimativas para um ponto não amostrado $\hat{Z}(s_0)$, obtidas por uma combinação linear ponderada dos pontos amostrados $Z(s_i)$ na localização $s_i$, com i $= 1, \dots,$ n, ou seja

$$\hat{Z}(s_0) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} Z(s_i) \tag{10}$$

em que: n número de pontos amostrados; $\alpha_{\mathrm{i}}$ peso atribuído ao ponto amostrado $Z(s_i)$ na localização $s_i$; $\hat{Z}(s_0)$ valor a ser estimado para a localização espacial $s_0$ em um domínio D.

Muitos são os pesos que podem ser atribuídos aos pontos amostrados $Z(s_i)$. No entanto, para o método de krigagem ordinária os pesos $\alpha_{\mathrm{i}}$ são calculados de tal forma que $\sum_{i=1}^{n} \alpha_i = 1$, para produzir estimativas não tendenciosas $E\left(\hat{\mathrm{Z}}(\mathrm{s}_0)\right) = \mu$ e, com variância mínima.

A média estacionária deve ser estimada e os pesos ótimos são encontrados minimizando, via multiplicador de Lagrange $\delta$, a variância do erro de estimação sob a condição $\sum_{i=1}^{n} \alpha_i = 1$. O vetor dos pesos é determinado a partir do sistema de equações de krigagem ordinária, definido pela matriz de covariâncias das distâncias da seguinte forma (Pasini et al. 2015; Vieira, 2000):

$$\begin{cases} \sum_{i=1}^{n} \alpha_i C(s_i, s_j) - \delta = C(s_i, \mathrm{s}_0), para\ j = 1, \dots, n \\ \sum_{i=1}^{n} \alpha_i = 1 \end{cases}$$

Em termos matriciais, o sistema de krigagem pode ser escrito da forma: $A\alpha = B$ e a solução para determinar os pesos pode ser vista na Equação (11).

$$\alpha = A^{-1}B \tag{11}$$

sendo:

$$A = \begin{bmatrix} C(s_1, s_1) & \dots & C(s_1, s_n) & 1 \\ \dots & \ddots & \dots & \vdots \\ C(s_n, s_1) & \dots & C(s_n, s_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}; \quad B = \begin{bmatrix} C(s_1, \mathrm{s}_0) \\ \vdots \\ C(s_n, \mathrm{s}_0) \\ 1 \end{bmatrix} \quad \text{e} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \delta \end{bmatrix}$$

em que: $\delta$ é o multiplicador de Lagrange necessário para a minimização da variância dos erros; $C(s_i, s_j)$ são as covariâncias entre as variáveis Z nos pontos amostrais $s_i$ e $s_j$; $C(s_i, \mathrm{s}_0)$ são as covariâncias entre a variável Z no ponto amostrado $s_i$ e a variável Z no ponto não amostrado $\mathrm{s}_0$; $\alpha$ é a matriz dos pesos de krigagem a ser estimada; $A^{-1}$ é a matriz inversa da matriz A com as covariância entre as localidades vizinhança de um ponto amostrado, determinada pelo modelo de semivariograma; $B$ é a matriz com as covariâncias entre as localidades vizinhas de um ponto amostrado e o ponto a ser interpolado, também determinada pelo modelo de semivariograma.

A correspondente variância minimizada da krigagem ordinária é dada por (Vieira, 2000):

$$\sigma^2{}_{KO} = C(0) - \sum_{i=1}^{n} \alpha_i C(s_i, \mathrm{s}_0) - \delta \tag{12}$$

Em termos matriciais:

$$\Sigma_{KO} = C(0) - \alpha' B \tag{13}$$

Com o método de interpolação de krigagem é possível obter previsão ótima e erro padrão de predição para dados ruidosos e incompletos de qualquer local de interesse, gerando mapas espacialmente completos. Além disso, a variância da krigagem, Equação (12), fornece informação importante sobre a confiabilidade dos valores interpolados. No entanto, esse procedimento requer a inversão de uma matriz de variâncias e covariâncias espacial $\Sigma_{nxn}$ que determina os pesos atribuídos às diferentes amostras, em que $n$ denota o conjunto de dados (Zhu et. al., 2015).

A inversão dessa matriz de covariância exige o cálculo intenso ou mesmo inviável para grandes conjuntos de dados. Dessa forma, o Fixed Rank Kriging (FRK) foi desenvolvido para reduzir a dimensão dessa matriz, evitando modelos estacionários e isotrópicos de covariância e semivariograma (Zhu et al, 2015; Cressie e Johannesson, 2008).

## 1.2. Fixed Rank Kriging (FRK)

Em 1993, Cressie desenvolveu uma estrutura de previsão e modelagem espacial e/ou espaço-temporal nomeado por Fixed Rank Kriging (FRK) considerado um BLUP espacial (melhor preditor linear espacial não viesado).

O FRK constrói um modelo espacial de efeitos aleatórios (SRE), em um domínio discretizado (BAU – Basic areal unit) para alcançar a redução da dimensão dos dados $[O(n^3) \rightarrow O(n)]$ e para modelar a dependência espacial através de um número fixo e pré-determinado de funções de base (Cressie e Johannesson, 2008; Zhu et. al., 2015).

O preditor de krigagem FRK foi desenvolvido como uma combinação de funções de base, independente do semivariograma, assim como a sua correspondente variância de krigagem. Por esse motivo, é um preditor viável para grandes conjuntos de dados, desde que o número de funções de base seja bem menor que o tamanho da amostra.

O FRK decompõe o processo espacial em dois componentes: um componente de tendência determinística de funções lineares de covariáveis espaciais e, o outro, componente aleatório da variação espacial que depende da utilização de um modelo espacial de efeitos aleatórios (SRE).

Considere um processo espacial $Z(s_0) = Y(s) + \varepsilon$ em que $\{Y(s): s \in D\}$ e D um domínio espacial de interesse. Então, em um conjunto de funções de base, o processo espacial definido por Zammit-Mangion e Cressie (2018) é dado por:

$$Y(.) = \mu(.) + v(.) \tag{14}$$
$$\hat{\mu}(s_0) = t(s)'\hat{\beta}\ ;\ \hat{\beta} = (T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}Z$$
$$v(.) = S(s)'\eta + \varepsilon(s)$$
$$\eta \sim N(0, K);\ \varepsilon(s) \sim N(0, \sigma^2 V)$$
$$\Sigma^{-1} = (\sigma^2 V)^{-1} - (\sigma^2 V)^{-1}S\{K^{-1} + S'(\sigma^2 V)^{-1}S\}^{-1}S'(\sigma^2 V)^{-1}$$

sendo: $T$ matriz de covariáveis conhecidas espacialmente referenciadas; $Z$ matriz de localização espacial; $t(s)$ vetor de covariáveis espacialmente referenciadas; $\hat{\beta}$ vetor de coeficientes de regressão estimados (efeito fixo); $\Sigma^{-1}$ é a matriz inversa de variâncias e covariância; $\sigma^2$ é a variância dos dados; V é a matriz diagonal $nxn$ dos erros de medição conhecidos; $S(s)$ é um vetor de função de base; $K^{-1}$ é a matriz inversa de $K$ qualquer matriz $mxm$ positiva definida em um conjunto de funções de base estimada por meio dos dados

conhecidos em que $m < n$; $\eta$ é um vetor de efeitos aleatórios com estrutura de dependência espacial em K.

Ao substituir a Equação (14) nos termos da Equação (11), podemos escrever:

$[A]^{-1} = (T\Sigma^{-1}T)^{-1}$ e $[B] = T'\Sigma^{-1}Z$

Assim, o preditor de krigagem FRK é dado por:

$$\hat{Z}_v(s) = t(s)'(T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}Z + S(s)'\eta + \varepsilon(s) \quad (15)$$

em que: T matriz de covariáveis conhecidas espacialmente referenciadas; Z matriz de localização espacial; $t(.)$ vetor de covariáveis espacialmente referenciadas, $\Sigma^{-1}$ é a matriz inversa de variâncias e covariância; $S(s)$ é um vetor de função de base no domínio R; $\eta$ é um vetor de efeitos aleatórios com estrutura de dependência espacial em K; K qualquer matriz mxm positiva definida em um conjunto de funções de base estimada por meio dos dados conhecidos em que $m < n$; n tamanho do conjunto de dado.

Essa combinação de uma matriz $K_{mxm}$ positiva definida com um conjunto de funções de base produz uma família flexível de funções de covariância (Nguyen, 2009). Para qualquer matriz positiva definida $K_{mxm}$ e $\sigma^2 > 0$, pode-se escrever a matriz de variâncias e covariâncias da predição FRK com uma estrutura de covariância igual a $C = SKS'$ e, portanto:

$$\Sigma = SKS' + \sigma^2 V \quad (16)$$

em que: Σ matriz de variâncias e covariâncias de krigagem FRK; $S(.)$ vetor espacial de funções de base; V é a matriz diagonal nxn dos erros de medição conhecidos.

A correspondente variância minimizada de krigagem RFK é dada por:

$$\sigma_{FRK}{}^2(s_0) = S(s_0)'KS(s_0) - S(s_0)'KS'\Sigma^{-1}SKS(s_0) + p'(T'\Sigma^{-1}T)^{-1}p \quad (17)$$

em que $p = t(s_0) - T'\Sigma^{-1}SKS(s_0)$; $T$ matriz de covariáveis conhecidas espacialmente referenciadas; $t(.)$ vetor de covariáveis espacialmente referenciadas; $\Sigma^{-1}$ é a matriz inversa de

variâncias e covariância; $S(.)$ é um vetor de função de base no domínio $R$; $K$ qualquer matriz $mxm$ positiva definida em um conjunto de funções de base estimada por meio dos dados conhecidos em que $m < n$.

Conforme a previsão da localização $s_0$varia nas Equações (15) e (17) ao longo do domínio espacial discretizado de interesse são gerados um mapa de predição de krigagem e um mapa da variância de krigagem (Cressie e Johannesson, 2008) representado pela Figura 3.

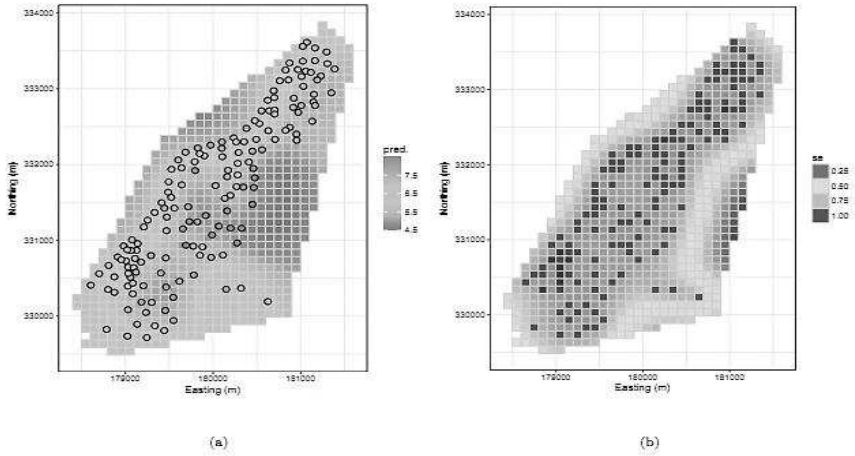


Figura 3: [a] Predição dos dados pelo FRK; [b] Variância de previsão pelo FRK.
Fonte: Zammit-Mangion e Cressie, 2018.

Cressie e Johannesson (2008) definem que a não exigência da ortogonalidade das funções de base, $S_1, \ldots, S_r$, torna a sua escolha irrestrita, porém, pré-determinadas. Além disso, recomendam que as funções de base apresentem múltiplas soluções para garantir que o modelo da função de covariância da Equação (16) capture múltipla escala de variação.

Zammit-Mangion e Cressie (2018) descrevem que a consideração de um domínio discretizado no FRK permite combinar várias observações com suporte diferente e fazer a distinção entre o erro de medição e a variação de escala na resolução da unidade base de área, o que leva a uma melhor quantificação da incerteza na predição. Para demonstrar, a Figura 4 representa a localização dos dados e o seu respectivo domínio discretizado construído com base no domínio inicial.

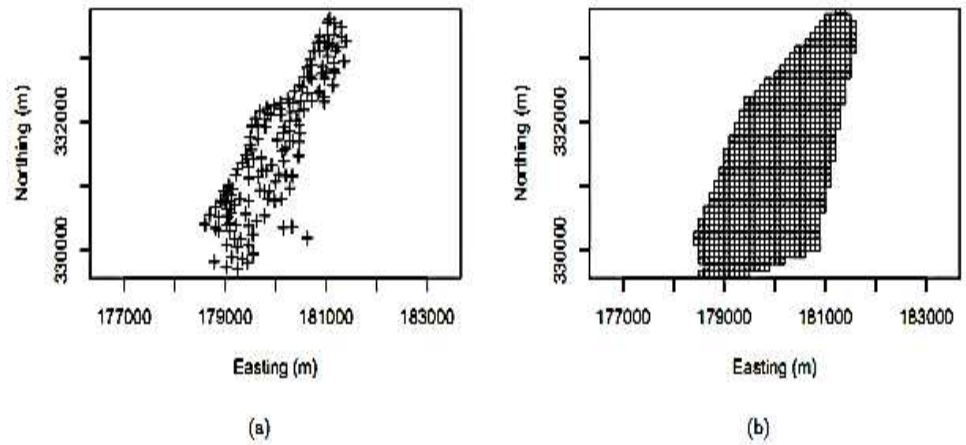


Figura 4: [a] Localização dos dados; [b] Unidades de área construída pelo FRK sobre o conjunto de dados.
Fonte: Zammit-Mangion e Cressie, 2018.

Em geral, o método de interpolação FRK permite maneiras alternativas de calcular o sistema de equações de krigagem determinado pela Equação (12), envolvendo a inversão apenas de matrizes $mxm$ em que $m < n$, independente do semivariograma, em que n é o tamanho amostral.

A estimação dos parâmetros é feita via algoritmo EM (expectation-maximization) que produz estimadores de máxima verossimilhança. É uma estratégia de maximização baseada na distribuição do que está faltando dado o que foi observado (Faria, 2011).

Outros métodos de predição que, atualmente, vêm ganhando destaque na literatura são os algoritmos de Aprendizado de Máquina (Machine Learning, em inglês).

O Aprendizado de máquina é uma área que estuda a capacidade para o aprendizado computacional da Inteligência Artificial cujo objetivo é desenvolver algoritmos capazes de adquirir conhecimentos de forma automática e fazer predições sobre os dados (Brink e Richards, 2014).

Na literatura, os algoritmos desenvolvidos pelo aprendizado de máquina que estão sendo mais utilizados para predições de dados são: Random Forest; Vector Machines; Extra-Tree (Hastie et al., 2008; James et al., 2013; Carvalho, 2014).

## 1.3 Random Forest

Breiman et al. (1984) desenvolveram um algoritmo nomeado por Random Forest (RF), que utiliza um conjunto de árvores de decisão com a metodologia de partição binária recursiva para classificar ou predizer valores de uma variável resposta.

O algoritmo RF é um classificador/preditor que consiste em um conjunto de árvores de decisão $\{h(x, \theta_k), k = 1, \dots K\}$ geradas dentro de um mesmo espaço, em que $\theta_1, .., \theta_k$ são vetores aleatórios, independentes e identicamente distribuídos. Usando um conjunto de treinamento, um vetor $\theta_k$ é gerado e uma $k -$ ésima árvore é cultivada, resultando em um classificador/preditor de árvore $h(x, \theta_k)$ que assume valores qualitativos ou quantitativos para a classe correta no vetor de entrada x (Breiman, 2001).

A Figura 5 representa um RF em um conjunto de cinco árvores de decisão.

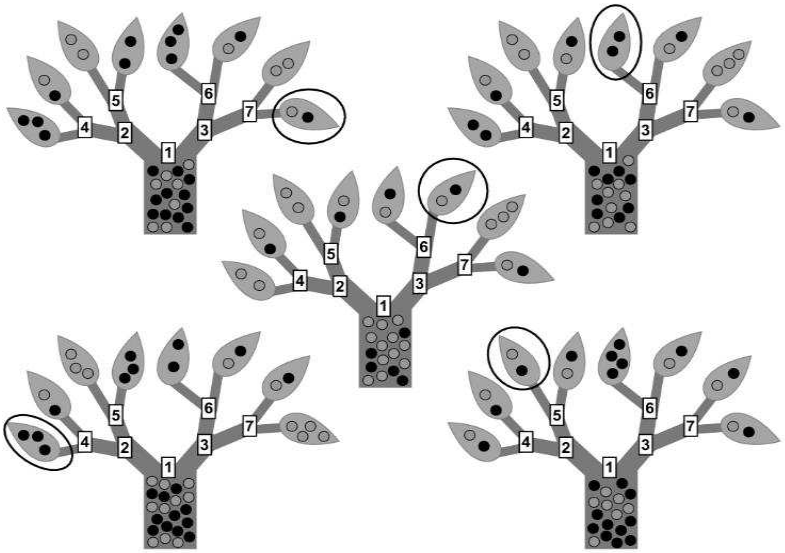

Figura 5*:* Representação de um Random Forest.
Fonte: Baker et al., 2015.

Observa-se que cada árvore de decisão contém um conjunto de treinamento diferente, com objetos que pertencem ao conjunto preto ou cinza. Para cada nó, representado por um número, as cinco árvores sofreram uma partição binária baseada em um atributo diferente e as folhas representam o momento em que não mais ocorrerão as partições.

A Figura 6 ilustra uma árvore de decisão e seu respectivo espaço de interesse bidimensional com repartição binária

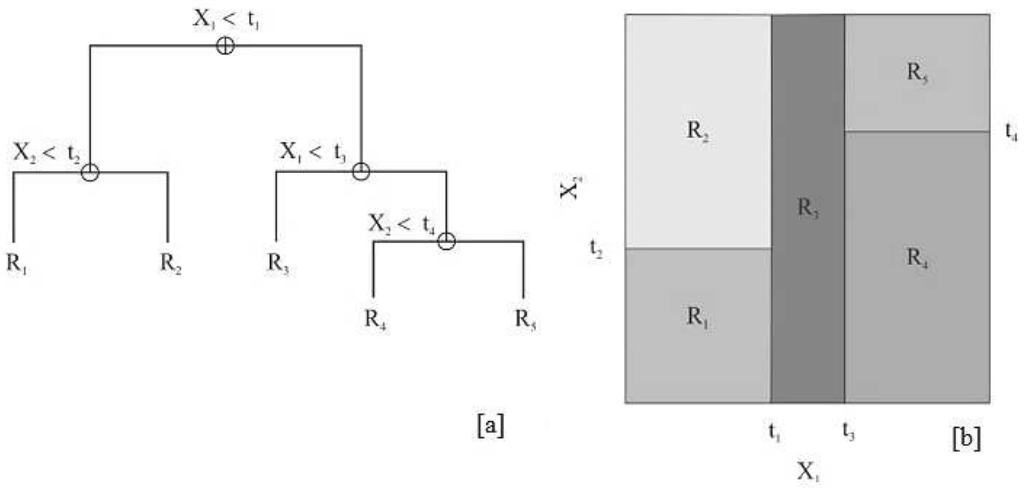


Figura 6: [a] Árvore de decisão com quatro nós de decisão e cinco nós terminais; [b] Espaço bidimensional com uma repartição binária recursiva correspondente.
Fonte: Adaptado de Viana (2019)

Observa-se que para cada nó de decisão condicionado à $X_1 < t_1; X_2 < t_2; X_1 < t_3; X_2 < t_4$, a árvore de decisão sofreu uma repartição binária baseado em um atributo $t_s$ diferente, resultando em cinco nós terminais determinado pela média dos valores da variável resposta do conjunto de treinamento dentro das sub-regiões $R_1; R_2; R_3; R_4$.

O critério de repartição binária, que divide o espaço amostral particionado em $j$ regiões distintas, $R_1, \ldots R_j$, em duas sub-regiões: $R_1(j,s): \{X: X < t_s\}$; $R_2(j,s): \{X: X \geq t_s\}$ condicionado ao atributo $t_s$, é usado para que o processo de minimização da soma dos quadrados dos erros de predição (Equação 19) passa a ser viável computacionalmente (James et al, 2013, Viana 2019).

$$\text{SQR} = \min_{j,s} \left[ \sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \right] \qquad (18)$$

em que: $y_i$ são as observações do conjunto de treinamento para um vetor de entrada x; $\hat{y}_{R_1}$ é a resposta média para as observações em treinamento em $R_1(j,s)$; $\hat{y}_{R_2}$ é a resposta média para as observações em treinamento em $R_2(j,s)$.

O algoritmo RF é caracterizado como regressão se a variável resposta é quantitativa ou, de classificação, se é qualitativa (Breiman, 2001; James et al, 2013).

### 1.3.1 Random Forest para Regressão

Random Forest para regressão representa uma combinação de várias árvores aleatórias, independentes e com mesma distribuição, em que o resultado final é a média geral de todos os resultados gerados para uma variável reposta quantitativa. (Junior et al., 2016).

Essas árvores são construídas inicialmente por um único nó que se divide em prováveis resultados e, cada resultado, se ramifica em outros nós, gerando outras possibilidades até que uma condição de parada seja atingida e não mais ocorram essas partições. Então, um modelo de regressão simples poderá ser aplicado (Breiman, 2001).

Segundo Breiman (2001) previsões realizadas por modelos de árvores também sofrem com tendenciosidade e erros de variância. Para contornar esse problema, o processo da construção de um Random Forest para regressão consiste em fazer um bootstraps ou bagging das variáveis do modelo (feature bagging) para diminuir a correlação entre as árvores.

Bagging é um meta-algoritmo criado para melhorar a predição e a regressão dos modelos de acordo com a estabilidade e a precisão dos preditores (Lopes et al., 2017).

Essa medida de aleatoriedade faz com que as árvores sejam diferentes e, portanto, diminui a correlação entre elas. A correlação pequena tende a diminuir o erro de predição o que torna mais preciso a floresta aleatória construída (Oshiro, 2013).

Random Forest para regressão aplica a técnica bagging para produzir $K$ amostras aleatórias do conjunto de treinamento para cada $k$ árvore aleatória. Essa técnica de amostragem com reposição constrói novos subconjuntos de treinamento a partir do conjunto de treinamento inicial (Oshiro, 2013).

A Figura 7 representa a construção de um Random Forest com a técnica de bagging aplicada.

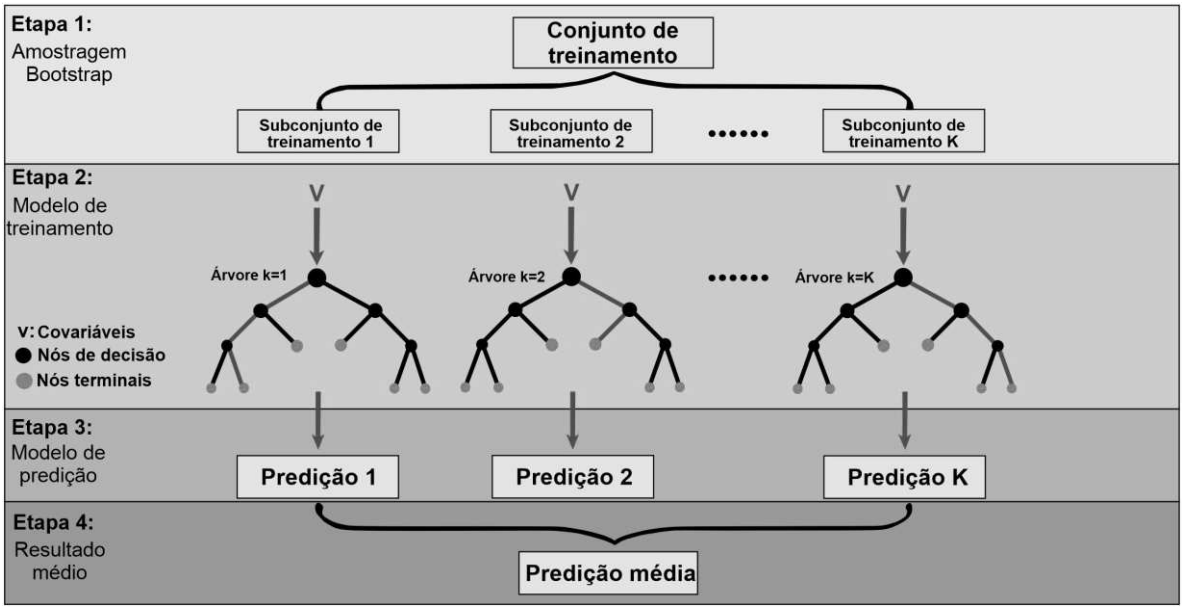


Figura 7: Random Forest com amostragem Bootstraps para predição.
Fonte: Adaptada de He, et al (2016)

Na construção de um Random Forest (Figura 7), considere um conjunto de treinamento inicial e $\theta_k$ $\{\theta_k : k = 1, \dots, K\}$ subconjuntos de amostras bootstrap. A técnica de bagging seleciona ao acaso amostras aleatórias com substituição para um subconjunto de treinamento $\theta_k$ e ajusta-o a $k$ −ésima árvore, fazendo isso, repetidamente. Dessa forma, as previsões para amostras não vistas em x, podem ser obtidas calculando a média das previsões de todas as árvores de regressão individuais k em x, por meio do preditor de regressão Random Forest, dado por (Breiman, 2001):

$$\hat{h}(x) = \frac{1}{K}\sum_{k=1}^{K} h(x\,, \theta_k) \tag{19}$$

em que: $K$ é no número total de árvores; $h(x, \theta_k)$ representa a previsão de uma árvore $k$ de um subconjunto $\theta_k$ para o vetor de entrada $x$, $\theta_k$ subconjunto de amostras bootstrap; $\hat{h}(x)$ média de todas as previsões para amostras não vista em $x$.

A técnica de bagging, em Random Forest, utiliza 2/3 da amostra do conjunto de treinamento inicial para testar o modelo e o restante 1/3 para validar, conhecidos como dados out-of-bag (OOB). Essa técnica reduz a variância, pois o viés produzido é análogo ao modelo original (Breiman, 1996).

De acordo com Liaw e Wiener (2002), as estimativas das incertezas das previsões OOB de todas as árvores podem ser obtidas pelo cálculo do erro quadrado médio ($EQM_{OOB}$), representado por:

$$EQM_{OOB} = \frac{1}{K}\sum_{i=1}^{k}(h_k - \hat{h}_k^{\,OOB})^2 \tag{20}$$

em que $h_k$ é a resposta de uma árvore $k$ para um vetor de entrada $x$; $\hat{h}_k^{\,OOB}$ é a média de todas as predições na técnica OOB (bagging); K o número de árvores.

O percentual da variabilidade explicada pelo modelo é dado por (Liaw e Wiener, 2002):

$$Var_{exp} = R^2 = 1 - \frac{\mathrm{EQM_{OOB}}}{\hat{\sigma}^2{}_{\mathrm{h}}} \tag{21}$$

em que $\hat{\sigma}^2{}_{\mathrm{h}}$ é a variância total das árvores preditoras.

O algoritmo Random Forest fornece o grau importância de cada covariável para predizer o valor da variável resposta. O grau de importância é determinada pela estimativa do erro quadrado médio $EQM_{OOB}$, que indica uma maior importância à medida que o valor dessa estimativa aumenta conforme os dados OOB para a covariável são permutados, enquanto todos os outros são deixados inalterados (Liaw e Wiener 2002). Desta forma, o algoritmo Random Forest além de apresentar a vantagem de identificar o grau de importância de cada covariável para predizer a variável resposta, consegue também, modelar conjuntos de dados extensos sem a restrição ao número de covariáveis.

Diante do exposto, no intuito de melhorar o desempenho dos interpoladores geoestatísticos de krigagem: Krigagem Ordinária (KO) e Fixed Rank Kriging (FRK), o presente trabalho, propôs a junção computacional do algoritmo Random Forest para regressão nos interpoladores em questão.

Os interpoladores híbridos implementados dispõem de uma estrutura capaz de gerar mapas de predição em uma análise multivariada, pois considera o processo de empilhamento de rasters para todas as covariáveis, concatenando-os em uma mesma resolução para extração dos pontos que serão utilizados na interpolação, analisando assim, a influência simultânea de todas as variáveis envolvidas no estudo.

A Figura 8 ilustra o processo de empilhamento de rasters de todas as covariáveis em estudo concatenando-os em uma mesma resolução.

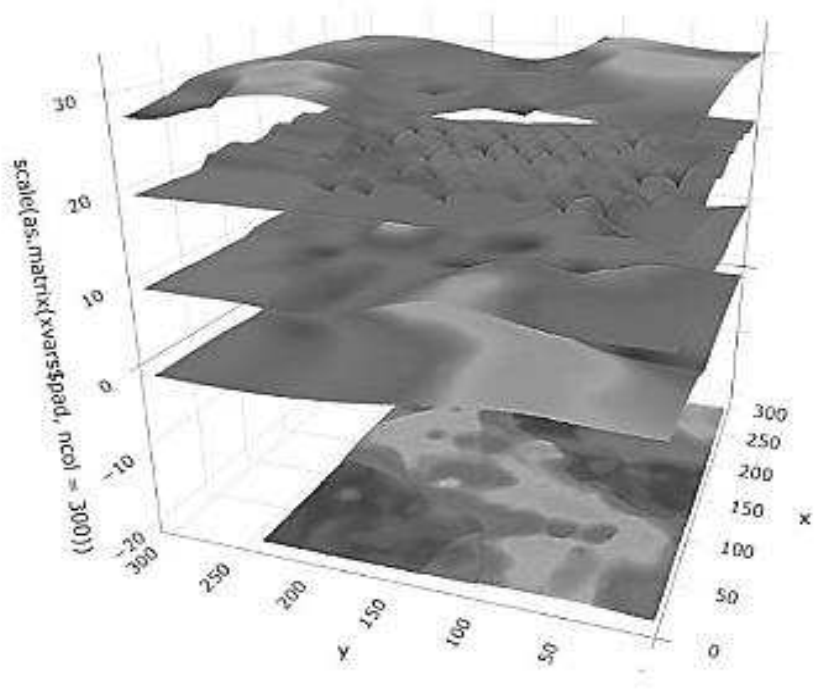


Figura 8: Empilhamento de rasters de todas as covariáveis em estudo concatenando-os em uma mesma resolução

Os interpoladores híbridos implementados, em uma abordagem multivariada, serão capazes de avaliar o comportamento de todas as covariáveis simultaneamente em relação a variável resposta, permitindo assim, uma análise estatística mais robusta, já que os modelos geoestatísticos existentes não conseguem alcançar esse resultado por apresentarem apenas procedimentos univariados. Além disso, o interpolador implementado pelo FRK com Random Forest será capaz, também, de expressar o que está acontecendo em termos de um conjunto reduzido de dimensões.

Para comparar a eficiência dos modelos de predição tem sido comumente utilizado na literatura algumas medidas de erro de predição e ajuste do modelo, tais como: erro médio (EM); erro quadrático médio (EQM); raiz do erro quadrático médio (RMSE); desvio médio absoluto (DMA), coeficiente de determinação $R^2$, entre outras.

Como critério de comparação do desempenho entre os preditores de krigagem, já existentes na literatura, com os híbridos implementados por esse trabalho, será avaliado o coeficiente do erro quadrático médio (EQM) e o coeficiente de determinação ($R^2$), dado por:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (22)$$

$$R^2 = 1 - \frac{EQM}{\hat{\sigma}^2} \quad (23)$$

em que: $y_i$ valor observado na localização $i$; $\hat{y}_i$ valor predito; $n$ número de observações da amostra em estudo; $\hat{\sigma}^2$ estimativa da variância total dos dados.

Valores menores do EQM indicam modelos propostos com melhores desempenhos, pois medem o desvio de suas predições em relação aos valores reais. O coeficiente de determinação avalia a qualidade de ajuste do modelo com variação de escala entre 0% e100%, em que valores próximos de 100% indicam melhores ajustes, ou seja, o modelo está conseguindo explicar bem toda a variabilidade dos dados de resposta ao redor de sua média.

Toda a parte inovadora da metodologia foi produzida por meio do software livre R (R Development Core Team, 2017), em que as análises geoestatísticas foram realizadas através dos pacotes: geoR, desenvolvido por Ribeiro Júnior e Diggle (2001); FRK, desenvolvido por Cressie (1993); Random Forest, desenvolvido por Breiman et al. (1984); INLA, desenvolvido por Rue et al. (2009); splancs, desenvolvido por Bivand et al. (2017); gstat, desenvolvido por Pebesma et al. (1998).

# OBJETIVOS E CONTRIBUIÇÃO

A presente pesquisa, de uma forma geral, objetiva-se elaborar computacionalmente um interpolador geoestatístico híbrido dotado de uma abordagem multivariada. Para isso, foram estudados alguns temas considerados relevantes na área de Geoestatística e do Aprendizado de Máquina que necessitavam de aprofundamento. Objetiva-se também, aplicar os interpoladores híbridos no estudo da variabilidade espacial dos atributos do solo e da produção média das Castanheiras-da-amazônia em um município no estado do Amapá/Brasil. E, para fins de comparação quanto ao desempenho dos interpoladores será utilizado o resultado via erro quadrático médio (EQM) e o coeficiente de determinação ($R^2$).

Inicialmente, será proposto um interpolador híbrido constituído pela junção da Krigagem Ordinária com Random Forest para regressão, denominado por Random Forest Ordinary Kriging (RFOK), no intuito de verificar se esse interpolador implementado computacionalmente resultará em predições mais acuradas da krigagem ordinária (KO). Entretanto, esse interpolador apresenta a desvantagem não só de excluir variáveis consideradas importantes no estudo pela presença do efeito pepita puro, bem como, dispõe de uma metodologia inviável para modelar grandes conjuntos de dados. Para contornar essa desvantagem, posteriormente, será proposto um interpolador híbrido pela junção do Fixed Rank Kriging (FRK) com Random Forest para regressão, intitulado por Random Forest Kriging (RFK). Como hipótese, espera-se que as desvantagens citadas acima sejam superadas e que o grau de importância de cada variável seja determinado.

O presente estudo visa contribuir com desafios, tais como: solucionar problemas de predição envolvendo conjuntos de dados geoestatísticos extensos; analisar o comportamento de grandes números de covariáveis simultaneamente, sem a perda de informações importantes para o desenvolvimento do estudo, conforme a decisão do especialista; direcionar amostragens mais eficientes e minimizar prejuízos de ordem social, ecológica e econômica.

## ESTRUTURAÇÃO DO TRABALHO

Esta pesquisa está estruturada em dois capítulos e um texto de conclusão final, em que serão exploradas a parte teórica, computacional e a aplicação da Geoestatística com Aprendizado de Máquina.

Na Introdução Geral, foi apresentada uma breve revisão, sobre a krigagem geoestatística e o algoritmo Random Forest do aprendizado de máquina, necessárias para o entendimento do novo método proposto computacionalmente.

No capítulo 1, será apresentado o primeiro interpolador híbrido implementado computacionamente no estudo da variabilidade espacial dos atributos do solo e da produção média das Castanheira-da-amazônia amostrados no sul da reserva extrativista do Rio Cajari, na Zona Rural do município de Laranjal do Jari, no Estado do Amapá, Brasil. Inicialmente, será considerada a elaboração e análise de semivariogramas com interpolação univariada por Krigagem Ordinária e, para fins de comparação, será aplicado na variável produção o interpolador híbrido multivariado RFOK. O desempenho dos interpoladores será aferido pelo método EQM.

No capítulo 2, para evitar os modelos estacionários, isotrópicos de covariância e semivariogramas será apresentado o segundo interpolador híbrido implementado computacionamente no estudo da variabilidade espacial dos atributos do solo e da produção média das Castanheira-da-amazônia amostrados no sul da reserva extrativista do Rio Cajari, na Zona Rural do município de Laranjal do Jari, no Estado do Amapá, Brasil. Inicialmente, será considerada a interpolação univariada por Fixed Rank Kriging (FRK) e, para fins de comparação, será aplicado na variável de interesse produção o interpolador híbrido RFK dentro de uma abordagem multivariada. O desempenho dos interpoladores será aferido pelo método EQM e $R^2$.

Finalmente, nas conclusões gerais, serão retomados os principais resultados das análises e exposta uma reflexão sobre o trabalho desenvolvido.

## REFÊRÊNCIAS BIBLIOGRÁFICAS


ANDRIOTTI, J.L.S. **Fundamentos de Estatística e Geoestatística**. Editora Unisino: São Leopoldo. 2013.

BAKER, P. T; CAUDILL, S; HODGE, H. A; TALUKDER, D; CAPANO, C; CORNISH, N. J. Multivariate classification with random forest for gravitational wave searchers of black hole binary coalescence. **Physical Review Journals**. v.91(6), p.1-26, 2015.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN. R.; STONE, C. **Classification and Regression Trees**. Taylor & Francis, 1984. 368p.

BREIMAN, L. **Bagging predictors**. Machine learning, v.24 (2), p.123-140, 1996. http://dx.doi.org/10.1023/A:1018054314350

BREIMAN, L. **Random Forest**. Machine Learning, v.45 (1), p.5-32, 2001. Disponível em: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>. Acesso em: 05 abr. 2018.

CAMARGO, E.C.G. **Geoestatística: Fundamento e Aplicações**. In: Camara, G.; Medeiros, J. S. Geoprocessamento em Projetos Ambientais. 2ª ed. São José dos Campos: INPE, 1998.

CARVALHO, H. M. **Aprendizado de máquina voltado para mineração de dados: árvores de decisão**. Monografia (Graduação). Departamento de Engenharia de Software. Universidade de Brasília. Brasilia, 2014.

CRESSIE, N. **Statistics for Spatial Data**. Revised edn. New York, 1993.

CRESSIE, N.; JOAHANNESSON, G. **Fixed rank kriging for very large spatial data sets.** Journal of the Royal Statistical Society. Statistical Methodology, v.70 (1), pp. 209–226, 2008.

FARIA, V. B. **Estimação de máxima verossimilhança via algoritmo EM**. Dissertação (Mestrado). Departamento de Estatística da Universidade Federal de Juiz de Fora. Juiz de Fora, 2011.

FERREIRA, Í. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas.
**Revista Brasileira de Cartografia (Online)**, v. 65(5), p. 831/03-842, 2013.

GOOVAERTS, P. **Geostatistics for natural resources evaluation**. New York: Oxford University Press, 1997. 476 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. Springer Series in Statistics. Second Edition. California, 2008. 744p.

HE, X.; CHANEY, N. W.; SCHLEISS, M.; SHEFFIELD, J. Spatial downscaling of precipitation using adaptable random forests. **Water Resources. Research**, v.52, 2016. doi:10.1002/2016WR019034.

ISAAKS, E. H.; SRIVASTAVA, M. **An introduction to applied geostatistics.** New York: Oxford University Press: 600 p. 1989.

JAMES, G; WITTEN, D; HASTIE, T; TIBSHIRANI, R. **An Introduction to Statistical Learning with applications in R**. Springer New York Heidelberg Dordrecht London, 2013. 441p.

JUNIOR, W. C.; CALDERANO FILHO, CHAGAS, C. da S.; B.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa agropecuária Brasileira, Brasília, v.51 (9), p.1428-1437, 2016.** DOI: 10.1590/S0100-204X2016000900041

KRIGE, D. G. A statistical approach to some basic mine evalution problems on the witwatersrand. **Journal of the Chemical, Metallurgical and Mining Society of South Africa**, v. 52, p. 151-163, 1951.

KATZFUSS, M.; CRESSIE, N. **Tutorial on Fixed Rank Kriging (FRK) of CO2 data**. Technical Report No. 858. Department of Statistics the Ohio State University 1958 Neil Avenue Columbus, OH 43210-1247.

LIAW, A.; WIENER, M. **Classification and regression by random forest**. R News, v.2, p.18-22, 2002.

LOPES, T. D.; GOEDTEL, A.; PALACIOS, R. H. C.; GODOY, W. F. **Aplicação do Algoritmo Random Forest como classificador de padrões de falhas em rolamentos de motores de indução**. XIII Simpósio Brasileiro de Automação Inteligente. Porto Alegre, 2017.

MATHERON, G. **The Theory of Regionalized Variables and Its Applications**. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, n° 5. Fontainebleau: Ecole Nationale Supérieure des Mines de Paris, 1971.

McBRATNEY, A.G.; WEBSTER, A.G. **Choosing functions for semi-variograms and fitting them to sampling estimates**. Journal of Soil Science, v.37, p.617-39, 1986.

OPROMOLLA, P.A.; DALBEN, I.; CARDIM, M. **Análise geoestatística de casos de hanseníase no Estado de São Paulo, 1991-2002**. Revista Saúde Pública. 2006;40(5):907-13.

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica**. Dissertação (Mestrado). Departamento de Bioinfomática. Ribeirão Preto, 2013.

PASINI, M. P. B; LÚCIO, A. D; FRONZA, D; WEBER, L. S. Krigagem ordinária e inverso da distância ponderada aplicados na espacialização da população da mosca-do-figo. **Revista Brasileira de Ciências Agrárias, v.10(3), p.452-459, 2015.**

RAMIREZ, J. E. G. **Variabilidade espacial do parâmetro geomecânico rqd no depósito mineral de animais – peru.** Tese (Doutorado). Pontifícia Universidade Católica do Rio de Janeiro – PUC–RIO. Rio de Janeiro, 2009.

ROSA, L. M. F. **Estudo sobre a influência de afirmações populares na Geoestatística Clássica**. Tese (Doutorado). Departamento de Estatística da Universidade Federal de Viçosa. Viçosa, 2017.

SANTOS, G. R.; OLIVEIRA, M. S.; LOUZADA, J. M.; SANTOS, A. M. R. T. Krigagem Simples versus Krigagem universal: qual o preditor mais preciso? **Energia na Agricultura**, v.26, n°2, p.49-55, 2011.

VIANA, R. S. M. **O uso da geoestatística espaço-temporal e aprendizagem de máquina na predição da temperatura máxima do ar**. Tese (Doutorado). Departamento de Estatística da Universidade Federal de Viçosa. Viçosa, 2019.

VIEIRA, S. R. **Geoestatística em estudos de variabilidade espacial do solo**. In. NOVAES, R. F.; ALVAREZ V., V. H.; SCHAEFER, C. E G. R. Tópicos em ciências do solo. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, v.1. p.2-54, 2000.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. São Paulo: Editora Oficina de Letras; 2013.

ZHU, Y.; KANG, E. L.; BO, Y.; CHENG, Q. T.; CHENG, J.; HE, Y. **A Robust Fixed Rank Kriging Method for Improving the Spatial Completeness and Accuracy of Satellite SST Products**. IEEE Transactions on Geoscience and Remote Sensing, v.53(9), p.5021-5035, 2015.

# CAPÍTULO 1

# MODELAGEM MULTIVARIADA DA VARIABILIDADE DO SOLO DA AMAZÔNIA BRASILEIRA USANDO GEOESTATÍSTICA E APRENDIZAGEM DE MÁQUINA

## RESUMO

O estudo da variabilidade espacial dos atributos do solo relacionada com a produção média das Castanheiras-da-amazônia é fundamental para direcionar amostragens mais eficientes, que considerem as variações edáficas. Isso é importante para aumentar a confiabilidade dos estudos sobre as relações entre o solo e a vegetação da Amazônia. A Geoestatística é a metodologia utilizada para este tipo de estudo, uma vez que considera as características estruturais e aleatórias de uma variável espacialmente distribuída aplicada na teoria do cálculo da semivariância e da interpolação. No intuito de melhorar a capacidade de predição do interpolador de krigagem ordinária para a produção média das castanheiras-da-amazônia na região do Amapá, Amazônia oriental, e assim, fornecer subsídios mais confiáveis para o manejo florestal, a manutenção e a ampliação dessa produtividade na região em estudo, o presente artigo, propôs a implementação computacional de um interpolador híbrido do algoritmo Random Forest para a Regressão com a krigagem ordinária e, para fins de comparação, quanto ao desempenho desses interpoladores, será utilizado o coeficiente Erro Quadrático Médio (EQM). Para isso, foram coletadas amostras do solo de 50m×30m em todas as linhas da região do estudo e amostras da produção média das castanheiras no período 2007 a 2015. As análises estatísticas e geoestatísticas foram realizadas no ambiente computacional do software R e todos os pontos foram georreferenciados. Como resultado obteve-se uma redução significativa do erro quadrático médio para o interpolador proposto e, de uma forma multivariada, o ranqueamento de cada atributo solo quanto ao grau de importância em relação à produção das Castanheiras-da-amazônia.

Palavras-chave: Krigagem. Random Forest. Análise Multivariada. Castanheiras-da-amazônia.

## ABSTRACT

The study of the spatial variability of the soil attributes related to the production of the Chestnuts of the Amazon is fundamental to direct more efficient sampling, that consider the edaphic variations. That is important to increase the reliability of the studies about the relations between the soil and the Amazonian vegetation. The Geostatics is the methodology utilized for this type of study, since it considers the structural and random characteristics of a spatially distributed variable applied in the theory of the calculation of semivariance and interpolation. In the sense of improving the capacity of prediction of the ordinary kriging interpolator for the production of the Amazonian chestnuts in the region of Amapa, eastern Amazon, and thus, providing more reliable subsidies for the forest management, the maintenance and enlargement of this productivity in the region studies, the present article, proposes the computational implementation of a hybrid interpolator of the Random Forest algorithm for the Regression with the ordinary kriging and, for comparison purposes, regarding the performance of this interpolator, the MSE error measurement will be used. For this, samples of the 50mx30m soil were collected in all of the lines of the study region and samples of the medium production of the chestnuts in the period of 2007 to 2015. The statistical and geostatistical analysis were realized in the software R computational environment and all points were georeferenced. As result, there was a significant reduction in the mean square error for the proposed interpolator and, in a multivaried form, the discrimination for each soil attribute as to the degree of importance in relation to the production of Chestnuts of the Amazon.


**Keywords:** Kriging. Random Forest. Multivariate Analysis. Chestnuts of the Amazon.

# 1. INTRODUÇÃO

A Bertholletia excelsa, popularmente conhecida como Castanheira-da-amazônia é considerada uma das árvores mais nobres do bioma Amazônia pela importância social, ecológica e econômica para a região (Guerreiro et al., 2017). Relacionar os atributos do solo à produção é importante para entender os fatores que afetam a capacidade produtiva das castanheiras, para subsidiar modelos de predição de safra e políticas públicas, como a do preço mínimo dos produtos da sociobiodiversidade.

O estudo da variabilidade espacial dos atributos do solo e da produção dessa espécie pode direcionar amostragens mais eficientes, contribuir para o manejo dos castanhais naturais e também com os programas de plantio e ou enriquecimento da espécie. A Geoestatística é a metodologia utilizada para este tipo de estudo, uma vez que considera as características estruturais e aleatórias de uma variável espacialmente distribuída, para obter um modelo de dependência espacial, capaz de predizer valores de pontos nos locais onde não foram amostrados, empregando a teoria do cálculo da semivariância e da interpolação (Moolman e Van Huyssteen, 1989).

A interpolação espacial é o processo de gerar mapas representativos da realidade a partir de um conjunto de amostras. Esse processo é aplicado em várias áreas, tais como na Mineralogia, Hidrogeologia, Meteorologia (Matias et al., 2004) e, especialmente, em sistemas de produção de monoculturas (Machado et al., 2007; Lima et al., 2013) e sistemas agroflorestais (Campos et al., 2013; Oliveira et al., 2013).

O método de interpolação espacial muito utilizado na Geoestatística é a krigagem ordinária (KO), por não ser necessário o conhecimento sobre a média estacionária. No intuito de melhorar a capacidade de predição do interpolador KO, foi realizada computacionalmente a junção dessa krigagem com o algoritmo Random Forest para regressão, intitulado por Random Forest Ordinary Kriging (RFOK). Para análise do desempenho desse interpolador implementado computacionalmente foi verificado o coeficiente erro quadrático médio (EQM).

Diante do exposto, o estudo foi desenvolvido em quatro etapas: (1) Análise descritiva dos atributos do solo e da produção média das Castanheiras-da-amazônia; (2) Estudo da variabilidade espacial dos atributos do solo e da produção média das Castanheiras-da-amazônia, considerando-se a elaboração e análise univariada de semivariogramas com interpolação por krigagem ordinária; (3) Estudo da variabilidade espacial da produção média

das castanheiras-da-amazônia, considerando-se uma análise multivariada pelo interpolador híbrido RFOK; e, (4) Comparação do desempenho dos interpoladores: KO e RFOK, pela estatística do erro quadrático médio (EQM) e o ranqueamento de cada atributo solo quanto ao grau de importância em relação à produção das Castanheiras-da-amazônia

## 2. MATERIAL E MÉTODOS

Visando alcançar os objetivos dessa proposta serão apresentados os conceitos fundamentais para compreensão dos resultados.

### 2.1. Descrição da Área do Estudo

A área do estudo compreende uma parcela (300m x 300m) do bioma Amazônia, situada em um castanhal natural no sul da reserva extrativista do Rio Cajari, na Zona Rural do município de Laranjal do Jari, no Estado do Amapá, Brasil. É delimitada pelas latitudes: 9937400S; 9937900S e pelas longitudes: 354400W; 354850W, como mostra a Figura 1.

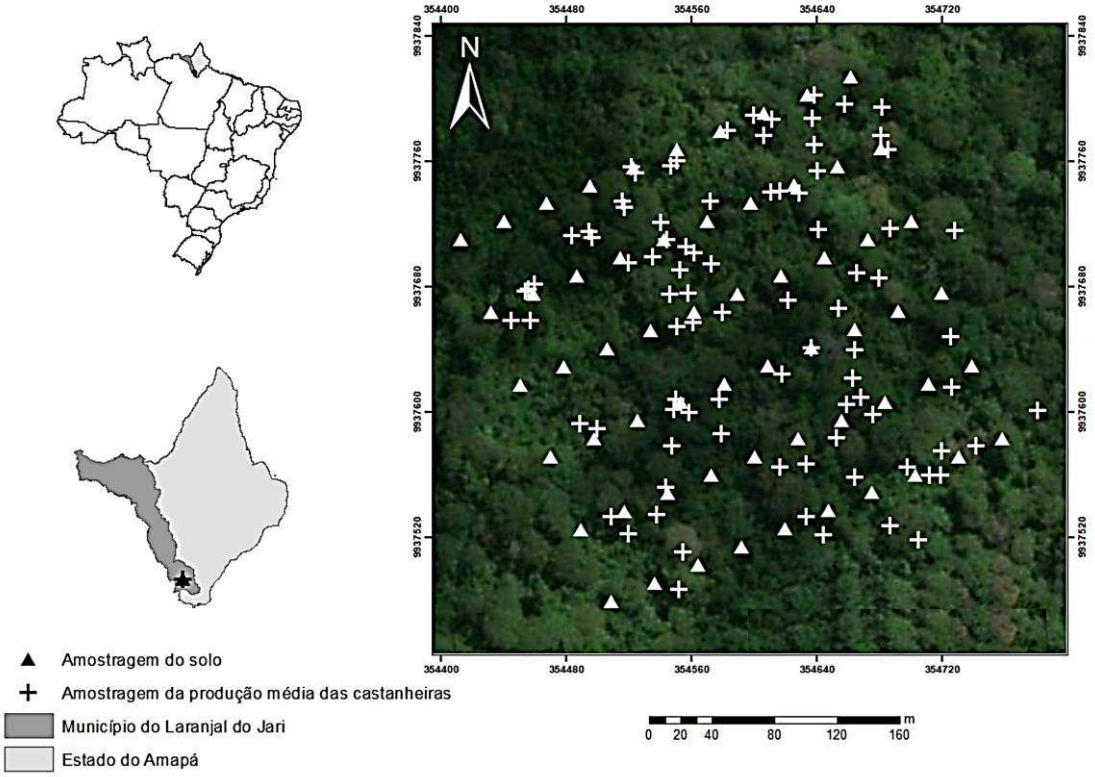


Figura 1: Representação dos pontos amostrais para amostra do solo e da produção média das castanhas-da-amazônia na reserva extrativista do Rio Cajari, município de Laranjal do Jari, no sul do Estado do Amapá, Brasil.

As amostras do solo e da produção média das castanheiras deste banco de dados foram georreferenciados ao sistema geodésico Datum SIRGAS 2000 e sistema de coordenada UTM (Universal Transverse Mercator Coordinate System) no fuso 22S.

Na área de estudo ocorre vegetação de mata nativa, com presença de agregados de castanheiras, que apresentam elevada densidade natural na região. A vegetação é classificada como Floresta Ombrófila Densa Submontana com dossel emergente (Ibge, 2012). O tipo de solo predominante é o Argissolo Vermelho-amarelo (Ibge, 2015). O clima da região é o tropical de monções com duas estações bem definidas, período chuvoso e período menos chuvoso (Alvares et al., 2014).

A área amostral foi dividida em seis transectos de 300 m, equidistantes em 50m, para orientar a amostragem do solo (Figura 1).

## 2.2. Coleta, preparação e análise físico-química das amostras de solo e da produção média das castanheiras

Ao lado de cada transecto foram coletadas amostras a cada 30 m, totalizando 10 amostras por transecto e 60 amostras no total, formando uma grade regular quadrática, com distâncias entre amostras de 30 m x 50 m. Todos os pontos foram georreferenciados no sistema de coordenadas UTM no fuso 22S.

Para a determinação granulométrica e química de atributos do solo, a coleta de amostras foi realizada com auxílio de um trado holandês, a uma profundidade de 0-20 cm. O preparo das amostras para obtenção de terra fina seca ao ar (TFSA) e realização das análises, assim como o cálculo das variáveis derivadas daquelas realmente medidas, foi realizado conforme as orientações de Nogueira e Souza (2005) e Embrapa (2007).

Foram quantificados 27 atributos que são considerados indicadores físicos e químicos da qualidade de solo (Gomes e Flizola, 2006), tais como: Potencial Hidrogeniônico (pH); Carbono (C); Matéria Orgânica (mo); Nitrogénio (N); Fósforo (P); Potássio (K); Sódio (Na); Cálcio (Ca); Magnésio (Mg); Alumínio (Al), Acidez Potencial (H+Al); Soma de Base (SB); Trocas de Cátions Efetiva (t); Trocas de Cátion a pH 7 (Tc); Saturação por Base (V%); Saturação por Alumínio (m%); Ferro (Fe); Zinco (Zn); Manganês (Mn); Cobre (Cu); Selênio (Sel), Areia Grossa (AreiaG); Areia Fina (AreiaF); Areia Total (AreiaT); Silte; Argila e Classificação Textural.

No local do estudo, também foram quantificadas 82 amostras da produção média das castanheiras no período 2007 a 2015, sendo cada castanheira, também georreferenciadas. As castanheiras adultas, com diâmetro maior que o diâmetro da menor castanheira produtiva, que não produziram frutos no período, foi considerado na análise com produção zero. Foram excluídas as jovens não reprodutivas, com diâmetros menores que o diâmetro da menor castanheria produtiva.

Nas subseções que seguem apresenta-se uma breve introdução dos fundamentos teóricos que foram necessários para a implementação do interpolador híbrido proposto por esse trabalho.

## 2.3. Krigagem Ordinária

O estudo do semivariograma em uma análise geoestatística é fundamental para descrever quantitativamente a variabilidade espacial de uma variável em função da distância $h$ (Isaaks e Srivastava, 1989; Vieira, 2000). Quando $h$ cresce, o semivariograma aproxima-se da variabilidade total dos dados e, havendo estacionariedade de segunda ordem, o semivariograma expressa o grau de dependência espacial entre os pontos amostrais (Opromolla et al., 2006).

O semivariograma, em termos matemáticos, é definido pela média do quadrado das diferenças entre todos os pares de pontos amostrados em uma área estudada, separados por um vetor h (Andriotti, 2013).

O estimador do semivariograma é dado por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \qquad (1)$$

em que: $\hat{\gamma}(h)$ é o valor estimado da função de semivariância para uma distância h;; $Z(x_i)$ é o valor observado da variável no ponto $(x_i)$; $Z(x_i + h)$ é o valor observado no ponto $(x_i + h)$; $N(h)$ é o número de pares de pontos separados entre si por uma distância h; $h$ distância entre os pares de pontos amostrados.

O semivariograma pode ser construído de duas formas: semivariograma experimental que é obtido através da amostragem; e, semivariograma teórico, que é obtido através do ajuste

de modelos teóricos tais como exponencial, esférico, gaussiano, dentre outros, ao semivariograma experimental.

O ajuste do modelo teórico ao semivariograma resulta na estimação de alguns parâmetros: efeito pepita $(C_0)$, contribuição $(C_1)$, patamar $(C = C_0 + C_1)$ e alcance $(a)$.

O efeito pepita é o valor da função na origem do semivariograma para distâncias menores do que a menor distância entre as amostras. A contribuição representa a diferença entre o patamar e o efeito pepita e, refere-se ao percentual da variabilidade explicada. O patamar é o valor em que o semivariograma se estabiliza e, é aproximadamente, igual à variância dos dados. O alcance é a distância limite de dependência espacial (Isaaks e Srivastava, 1989).

A partir do ajuste do modelo teórico ao semivariograma pode-se proceder à interpolação geoestatística conhecida como krigagem (Pires e Strieder, 2006).

A krigagem leva em consideração a dependência espacial existente entre os valores dos pontos amostrados e não amostrados, bem como a distância entre tais pontos. Além disso, permite a interpolação de valores em qualquer posição da área em estudo, sem tendência e com variância mínima, desde que seja conhecido o semivariograma e que haja dependência espacial entre as amostras (Isaaks e Srivastava, 1989; Vieira, 2000).

Na literatura são encontrados diversos tipos de krigagem, com suas especificidades, tais como: krigagem simples, krigagem universal, krigagem fatorial, entre outras. Dentre elas, a krigagem ordinária é a mais utilizada pelo fato de não ser necessário conhecer a média estacionária, podendo ser estimada (Isaaks e Srivastava, 1989).

O método de krigagem ordinária baseia-se na obtenção de estimativas para pontos não amostrados a partir da combinação linear ponderada dos valores das variáveis em pontos amostrados $Z_i$, com $i = 1, 2, \dots, n$, dado por (Vieira, 2000):

$$Z(s) = \sum_{i=1}^{n} \alpha_i Z_i + \varepsilon(s) \qquad (2)$$

em que: n é o número de pontos amostrados; $Z_i$ é o valor do ponto amostrado; $Z(s)$ é o valor a ser predito para a localização espacial s (variável interpolada); $\alpha_i$ peso atribuído ao ponto amostrado $t_i$ tal que $\sum_{i=1}^{n} \alpha_i = 1$.

No método de krigagem ordinária os pesos são calculados de tal forma que o $\sum_{i=1}^{n} \alpha_i = 1$ para produzir estimativas não tendenciosas e com variância mínima. Além disso,

são alteráveis de acordo com a variabilidade espacial expressa no semivariograma e determinado por (Pasini et al. 2015; Vieira, 2000):

$$\hat{\alpha} = A^{-1}B \quad (3)$$

em que: $\hat{\alpha}$ é a matriz dos pesos de krigagem a ser estimada; $A^{-1}$ é a matriz inversa da matriz A com as covariância entre as localidades vizinhança de um ponto amostrado, determinada pelo modelo de semivariograma; $B$ é a matriz com as covariâncias entre as localidades vizinhas de um ponto amostrado e o ponto a ser interpolado, também determinada pelo modelo de semivariograma.

Por meio da krigagem é possível conhecer a variância de krigagem, isto é, a precisão associada a cada predição (CAMARGO, 1998).

Em termos matriciais, a variância de krigagem é dada por:

$$\Sigma_{KO} = C(0) - \alpha' B \quad (4)$$

em que: $B$ é a matriz com as covariâncias entre as localidades vizinhas de um ponto amostrado e o ponto a ser interpolado; $\alpha'$ é a matriz transposta da matriz dos pesos de krigagem a ser estimada, $C(0) = \sigma^2$ é a covariância de um ponto na localização zero dado pela variância dos dados.

## 2.4. Random Forest

Aprendizagem de máquina (Machine Learning, em inglês) é uma área que estuda a capacidade para o aprendizado computacional de Inteligência Artificial (Brink e Richards, 2014). Arthur Samuel (1959) definiu que o aprendizado de máquina deveria explorar o estudo e a construção de algoritmos que poderiam aprender com seus erros e descobertas para fazer previsões sobre os dados.

Os algoritmos mais utilizados no aprendizado de máquina são: Random Forest, Support Vector Machines e Extra-Tree (Hastie et al., 2008; James et al., 2013; Carvalho, 2014).

Random Forest (RF) é um algoritmo de árvore de decisão desenvolvido por Breiman et al. (1984) que apresenta uma metodologia de partição binária recursiva (Breiman, 2001;

James et al., 2013;). É caracterizada como árvore de regressão, quando a variável resposta é quantitativa, ou de classificação, quando a variável resposta é qualitativa.

O algoritmo de árvore de decisão particiona o espaço de interesse em um conjunto de sub-regiões que utilizam um conjunto de treinamento supervisionado para a classificação ou previsão dos dados (James et al., 2013).

Random Forest é um classificador que consiste em um conjunto de árvores de decisão $\{h(x, \theta_k), k = 1, \dots L\}$, geradas dentro de um mesmo objeto, em que $\theta_1, .., \theta_k$ são vetores aleatórios, independentes e identicamente distribuídos. Para $k$-árvores, um vetor $\theta_k$ é gerado e uma árvore é cultivada usando esse conjunto de treinamento, resultando em um classificador de árvore $h(x, \theta_k)$ que assume valores numéricos para a classe correta no vetor de entrada x (Breiman, 2001).

Segundo Junior et al. (2016), Random Forest para regressão representa uma combinação de várias árvores aleatórias, independentes e com mesma distribuição, em que o resultado final é a média geral de todos os resultados gerados. Essas árvores são construídas inicialmente por um único nó que se divide em prováveis resultados e, cada resultado se ramifica em outros nós, gerando outras possibilidades (Breiman, 2001).

O processo de construção de uma floresta aleatória resulta em fazer um bootstraps ou bagging das variáveis do modelo (feature bagging) para diminuir a correlação entre as árvores. Bagging é um meta-algoritmo para melhorar a previsão e a regressão dos modelos de acordo com a estabilidade e a precisão dos preditores (Lopes et al., 2017). Essa medida de aleatoriedade faz com que as árvores sejam diferentes e, portanto, diminui a correlação entre elas. A baixa correlação tende a reduzir o erro de predição o que torna mais preciso a floresta aleatória construída (Oshiro, 2013).

Random Forest aplica a técnica bagging para produzir amostras aleatórias de conjuntos de treinamento para cada árvore aleatória. Essa técnica de amostragem com reposição constrói novos conjuntos de treinamento a partir do conjunto de treinamento inicial (Oshiro, 2013).

Considere um conjunto de treinamento $\theta_k$ e $x$ amostras bootstrap; a técnica de bagging seleciona ao acaso uma amostra aleatória com substituição para o conjunto de treinamento e ajusta as árvores a essa amostra, fazendo isso repetidamente. Dessa forma, as previsões para amostras não vistas em x podem ser feitas calculando a média das previsões de todas as árvores de regressão individuais k em x.

A equação a seguir descreve a função de predição dada pelo bagging

$$\hat{h}(x) = \frac{1}{K}\sum_{k=1}^{K} h(x\,, \theta_k) \tag{5}$$

em que $K$ é no número total de árvores; $h(x, \theta_k)$ representa a resposta (previsão) de uma árvore $k$ para um vetor de entrada $x$, $\theta_k$ representa os parâmetros desta árvore; $\hat{h}(x)$ média de todas as previsões para amostras não vista em $x$.

A técnica de bagging utiliza 2/3 da amostra no conjunto de treinamento para testar o modelo e o restante 1/3 para validar, conhecidos como dados out-of-bag (OOB). Essa técnica reduz a variância, pois o viés produzido é análogo ao modelo original (Breiman, 1996).

De acordo com Liaw e Wiener (2002), as estimativas das incertezas das previsões OOB de todas as árvores podem ser obtidas pelo cálculo do erro quadrado médio ($EQM_{OOB}$), representado por:

$$EQM_{OOB} = \frac{1}{K}\sum_{i=1}^{k} (h_k - \hat{h}_k^{\,OOB})^2 \tag{6}$$

em que $h_k$ é a resposta de uma árvore $k$ para um vetor de entrada $x$; $\hat{h}_k^{\,OOB}$ é a média de todas as previsões na técnica OOB (bagging); K o número de árvores.

O percentual da variabilidade explicada pelo modelo é dado por:

$$1 - \frac{\mathrm{EQM_{OOB}}}{\widehat{\sigma}^2{}_{\mathrm{h}}} \tag{7}$$

em que $\widehat{\sigma}^2{}_{\mathrm{h}}$ é a variância total das árvores preditoras.

Liaw e Wiener (2002) explicam que no Random Forest o grau importância de cada covariável para predizer a variabilidade da variável resposta é determinado pelo coeficiente do erro quadrado médio $EQM_{OOB}$ quando se permutam aleatoriamente os n valores de uma variável ao manter fixo os valores das demais variáveis. Segundo Morais (2017) essa medida captura a sensibilidade do modelo para cada preditor, quanto mais sensível, mais importante será a variável para o desempenho do modelo.

## 2.5. Método Proposto

Utilizando os conhecimentos teóricos e computacionais sobre a krigagem e o algoritmo Random Forest, o presente artigo, propôs a implementação computacional de um interpolador híbrido constituído pela junção da Krigagem Ordinária com o algoritmo Random Forest para regressão, intitulado por Random Forest Ordinary Kriging (RFOK) (Figura 2).

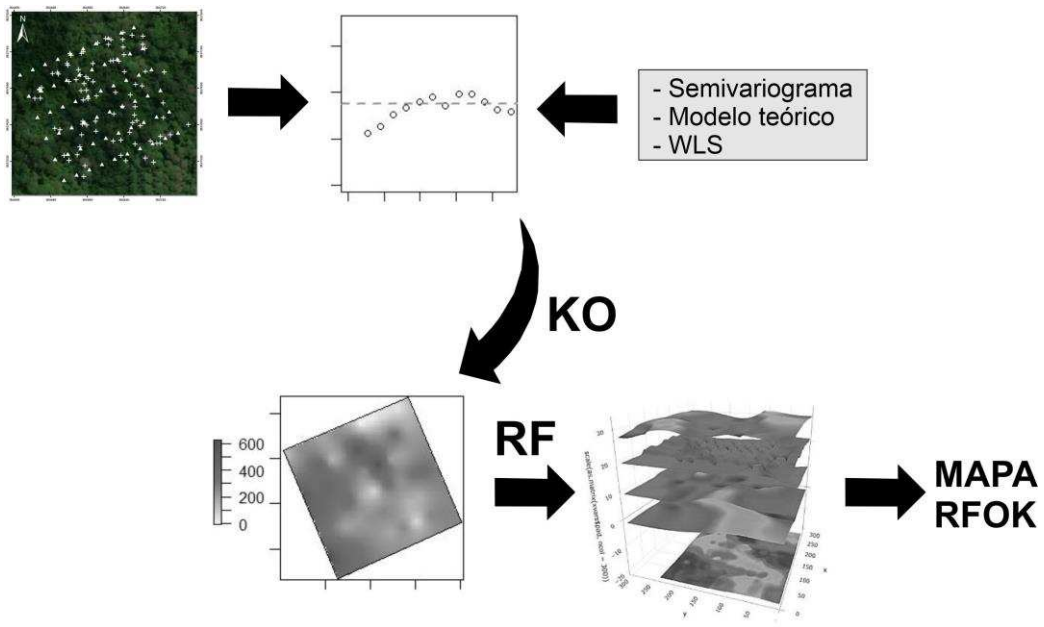


Figura 2: Representação do interpolador multivariado Random Forest Ordinary Kriging (RFOK)

A Figura 2 representa a construção do RFOK dada por uma adaptação na implementação computacional do método de dependência espacial univariado da krigagem ordinária ao algoritmo Random Forest (RF), um método não espacial e independente. Primeiramente, a metodologia do semivariograma é empregada no processo para determinar as variáveis dependentes do modelo e gerar o mapa de predição por KO. Após, o RF utiliza o empilhamento dos rasters gerados no KO para extração dos pontos do conjunto de treinamento da metodologia de árvore e, em uma abordagem multivariada, gerar o mapa final de predição.

O interpolador híbrido RFOK apresenta a restrição de utilizar a análise do semivariograma para selecionar as variáveis espacialmente dependentes, acarretando na exclusão de variáveis importantes para estudo pela presença do efeito pepita puro.

No entanto, exibe uma capacidade de melhorar a predição pelo fato de dispor de uma estrutura capaz de gerar mapas de predição em uma análise multivariada, pois considera o

processo de empilhamento de rasters para todas as variáveis em estudo concatenando-os em uma mesma resolução para extração dos pontos que serão utilizados na interpolação.

O RFOK apresenta também a habilidade de capturar a influência simultânea da variabilidade espacial das variáveis presentes no estudo e determinar o grau de importância de cada variável para predizer a variabilidade espacial da variável de interesse.

O intuito dessa proposição se dá pela busca de uma melhoria no desempenho do interpolador de Krigagem Ordinária (KO) para predizer uma variável de interesse, mais especificamente, analisar simultaneamente a relação dos atributos do solo com a variabilidade espacial da produção média das Castanheiras-da-amazônia.

Para análise das variáveis em estudo, atributos do solo e produção média das Castanheiras-da-amazônia, foi considerado o método de interpolação de krigagem ordinária (KO) com ajuste automático de um modelo teórico à semivariâncias, calculadas pelo método dos mínimos quadrados ponderados (Weighted Least Squares -WLS) (Melo et al., 2005). No ajuste automático do modelo téorico às semivariâncias foi considerado o modelo que apresentou o maior $R^2$.

No Random Forest a escolha adotada para definicão dos principais parâmetros recomendados por Liaw e Wiener (2002) foi a padrão/automática do pacote randomForest do software R que definiu: mtry= 1/3 do número total de covariáveis (número de covariáveis utilizadas em cada árvore), ntree=500 (número de árvores construídas pelo algoritmo) e o nodesize = 2 (número mínimo de observações em cada nó terminal).

Para fins de comparação, foi aplicado na variáveil produção o interpolador implementado RFOK e para avaliar a sua precisão foi utilizado à estatística do erro quadrático médio (EQM).

As análises estatísticas e geoestatísticas foram produzidas no ambiente computacional R (R Development Core Team, 2017).

Toda a parte inovadora da metodologia foi produzida por meio do software livre R (R Development Core Team, 2017), em que as análises estatísticas e geoestatísticas foram realizadas através dos pacotes: geoR, desenvolvido por Ribeiro Júnior e Diggle (2001); Random Forest, desenvolvido por Breiman et al. (1984); INLA, desenvolvido por Rue et al. (2009); splancs, desenvolvido por Bivand et al. (2017);

# 3. RESULTADOS E DISCUSSÃO

Buscando alcançar os objetivos propostos, utilizou-se dos dados dos atributos do solo e da produção média das Castanheiras-da-amazônia amostrados no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil

## 3.1. Análise Descritiva dos atributos do solo e da produção média das Castanheiras-da-amazônia

Os resultados da estatística descritiva para as propriedades físico-químicas do solo e para a produção média das Castanheiras-da-amazônia são apresentados na Tabela 1.

Tabela 1: Análise descritiva dos atributos do solo e da produção média das Castanheiras-da-amazônia no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

| | pH | C | mo | N | P | K | Na | Ca | Mg | Al | H+Al | SB | t | Tc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mín** | 4,7 | 3,5 | 6,1 | 0,5 | 2,0 | 10,0 | 1,0 | 0,3 | 0,1 | 0,0 | 0,9 | 0,6 | 0,8 | 2,1 |
| **méd** | 5,7 | 8,3 | 14,2 | 0,8 | 3,5 | 19,3 | 3,0 | 1,2 | 0,5 | 0,1 | 2,0 | 1,8 | 1,9 | 3,8 |
| **máx** | 6,5 | 14,7 | 25,4 | 1,1 | 10,0 | 47,0 | 31,0 | 2,3 | 1,1 | 0,6 | 4,1 | 3,2 | 3,2 | 6,6 |
| **sd** | 0,1 | 4,6 | 13,6 | 0,0 | 2,1 | 54,3 | 15,2 | 0,3 | 0,1 | 0,01 | 0,3 | 0,5 | 0,4 | 0,8 |
| **CV** | 6,2 | 25,9 | 25,9 | 16,8 | 41,9 | 38,1 | 131,2 | 41,3 | 43,0 | 217,3 | 26,9 | 37,3 | 33,1 | 23,6 |

| | V | m | Fe | Zn | Mn | Cu | Sel | AreiaG | AreiaF | AreiaT | Silte | Argila | Produção |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mín** | 22,8 | **0,0** | 38,0 | 0,2 | 129,6 | 0,9 | 149,8 | 510,0 | 80,8 | 659,3 | 51,7 | 113,5 | 4,0 |
| **méd** | 46,6 | 4,3 | 95,1 | 0,8 | 218,3 | 1,8 | 401,4 | 612,6 | 125,3 | 737,9 | 84,0 | 178,1 | 110,3 |
| **máx** | 68,3 | 50,7 | 153,0 | 3,7 | 351,2 | 4,4 | 755,2 | 735,3 | 166,3 | 821,7 | 147,2 | 244,5 | 636,0 |
| **var** | 125,0 | 92,9 | 667,0 | 0,3 | 2182,8 | 0,4 | 15891,7 | 2546,3 | 361,6 | 1236,6 | 261,2 | 878,8 | 8939,9 |
| **CV** | 23,9 | 222,1 | 27,1 | 72,8 | 21,4 | 32,7 | 31,4 | 8,2 | 15,2 | 2,6 | 19,2 | 9,0 | 85,7 |

Legenda: pH: potencial hidrogeniônico; C: carbono (g/kg); mo: matéria orgânica (g/kg); N: nitrogénio (g/kg); P: fósforo (mg/dm$^3$); K: potássio (mg/dm$^3$); Na: sódio (mg/dm$^3$); Ca: cálcio(cmolc/dm$^3$); Mg: magnésio (cmolc/dm$^3$); Al: alumínio (cmolc/dm$^3$); H+Al: Acidez Potencial (cmolc/dm$^3$); SB: soma de base (cmolc/dm$^3$); t: trocas de cátions efetiva (cmolc/dm$^3$); Tc: trocas de cátion a pH 7 (cmolc/dm$^3$); V: saturação por base (%); m: saturação por alumínio (%); Fe: ferro (mg/dm$^3$); Zn: zinco (mg/dm$^3$); Mn: manganês (mg/dm$^3$); Cu: cobre (mg/dm$^3$); Sel: selênio (µg/kg); AreiaG: areia grossa (g/ kg); AreiaF: areia fina (g/ kg); AreiaT: areia total (g/ kg); Silte (g/ kg); Argila (g/ kg ); produção (UA); CV: coeficiente de variação (%); mín: mínimo; méd: média; máx: máximo; var: variância.

O Solo em estudo apresenta uma classificação textural moderadamente grossa, sendo composto, em grande parte por areia (74%) e, em menor parte, por argila (17%), denominado franco arenoso pelo triângulo simplificado da Embrapa (Embrapa, 2006), o que corrobora para os valores encontrados na Tabela 1 para areia e argila. Solos arenosos apresentam boa drenagem e capacidade de retenção de água (Embrapa, 2003). Os teores de areia fina quando

são muito maiores do que os de areia grossa contribuem para o aumento da disponibilidade de água no perfil do solo, o que não corrobora para os resultados do solo em questão. Centeno et al. (2017) destacaram que esse tipo de solo oferece as maiores deficiências de matéria orgânica (mo) e fósforo (P) pelo processo de lixiviação, o que foi confirmado na Tabela 1 com valores médios de 14,2 g/kg e 3,5 mg/dm$^3$, respectivamente. Os valores baixos de matéria orgânica (mo < 15 g/kg) desfavorecem o solo em estudo, pois a variável em questão exerce um papel fundamental na melhoria da fertilidade e, também, no aumento da produtividade da planta. Os valores baixos de fósforo ($2 \leq p \leq 10$ mg/dm$^3$) em solos mais arenosos estão diretamente relacionados com baixos teores de argila ($113{,}5 \leq argila \leq 244{,}5$ g/kg), que foi considerado baixo a médio pelo Guia Prático para Interpretação de Resultados de Análises de Solo (Embrapa, 2015).

Por outro lado, Junior (2016) destaca que altos teores de matéria orgânica no solo, assim como, elevados teores de óxido de ferro e de argila podem inibir a absorção de selênio pelas plantas e animais, pois absorvem ou ligam o selênio ao solo. Nesse caso, o solo em estudo é favorecido pela quantidade de selênio que varia em torno de 401,4 µg/kg e por baixa concentração de matéria orgânica e argila (Tabela 1). Segundo Whanger (2004) doses diárias de 100-200µg/kg de selênio podem inibir danos genéticos e o desenvolvimento de câncer em humanos. Também incluem defesa contra estresse oxidativo, regulação da ação dos hormônios da tireóide e regulação do potencial redox da vitamina C e de outras moléculas. No entanto, o valor do Limite Superior Tolerável de Ingestão (UL) de 400 µg/dia foi fixado devido ao risco de selenose (Institute of Medicine, 2000). Assim, correlacionar a ocorrência de selênio no solo com aqueles disponíveis nas amêndoas (castanhas) é fundamental para que se amplie o conhecimento dos fatores responsáveis pela disponibilização do selênio nas castanhas.

Os nutrientes como selênio e manganês são importantes para a nutrição e o crescimento das Castanhas-da-amazônia, o que favorece o cultivo dessa espécie no solo em estudo por apresentar altos valores para esses nutrientes. Como sugerido por Malavolta (1980) o solo é considerado deficiente para selênio quando o teor for inferior a 200µg/kg e, segundo RAIJ et al. (1997) o solo é considerado não deficiente para o manganês quando os teores extraídos com DTPA estão acima de 5,0 mg/dm$^3$.

A variabilidade das variáveis em estudo, medida pelo coeficiente de variação (CV), baseada nos limites propostos por Warrick e Nielsen (1980), para classificação dos atributos do solo, que consideram: variabilidade baixa (CV < 12%); média (12% < CV < 60%) e alta

(CV > 60%), se mostraram altas apenas para o zinco, sódio, alumínio e a produção; baixa para o pH, areia e argila; as outras variáveis mantiveram-se em torno de uma variabilidade média (Tabela 1).

### 3.2. Variabilidade espacial dos atributos do solo e da produção média das Castanheiras-da-amazônia considerando-se a elaboração e análise univariada dos semivariogramas com interpolação por Krigagem Ordinária

Após a análise exploratória, as variáveis em estudo foram submetidas à análise geoestatística, com o objetivo de verificar a dependência espacial dos dados. No presente estudo, não foi realizado o teste de normalidade para as variáveis, pois segundo Isaaks e Srivastava (1989), a normalidade não é significativa para análise geoestatística.

Ao construir os semivariogramas foram eliminadas do estudo as variáveis que apresentaram o efeito pepita puro, ou seja, variáveis que são espacialmente independentes e torna-se improvável determinar os componentes do semivariograma. Segundo Machado et al. (2010) o efeito pepita puro mostra a variabilidade espacial não explicada, devido a erros de medição ou microvariabilidade não detectada.

Desta forma, foram selecionadas para o estudo apenas as variáveis: Acidez Potencial (H+Al), Zinco (Zn), Cobre (Cu), Selênio (Sel), AreiaTotal, Argila e Produção.

Os semivariogramas e os mapas de predição por krigagem ordinária (KO) para as variáveis selecionadas estão representadas nas Figuras 3 e 4. Também foram plotadas nos mapas as coordenadas georreferenciadas da produção média das castanheiras, visando comparar a sua ocorrência com a concentração dos atributos no solo.

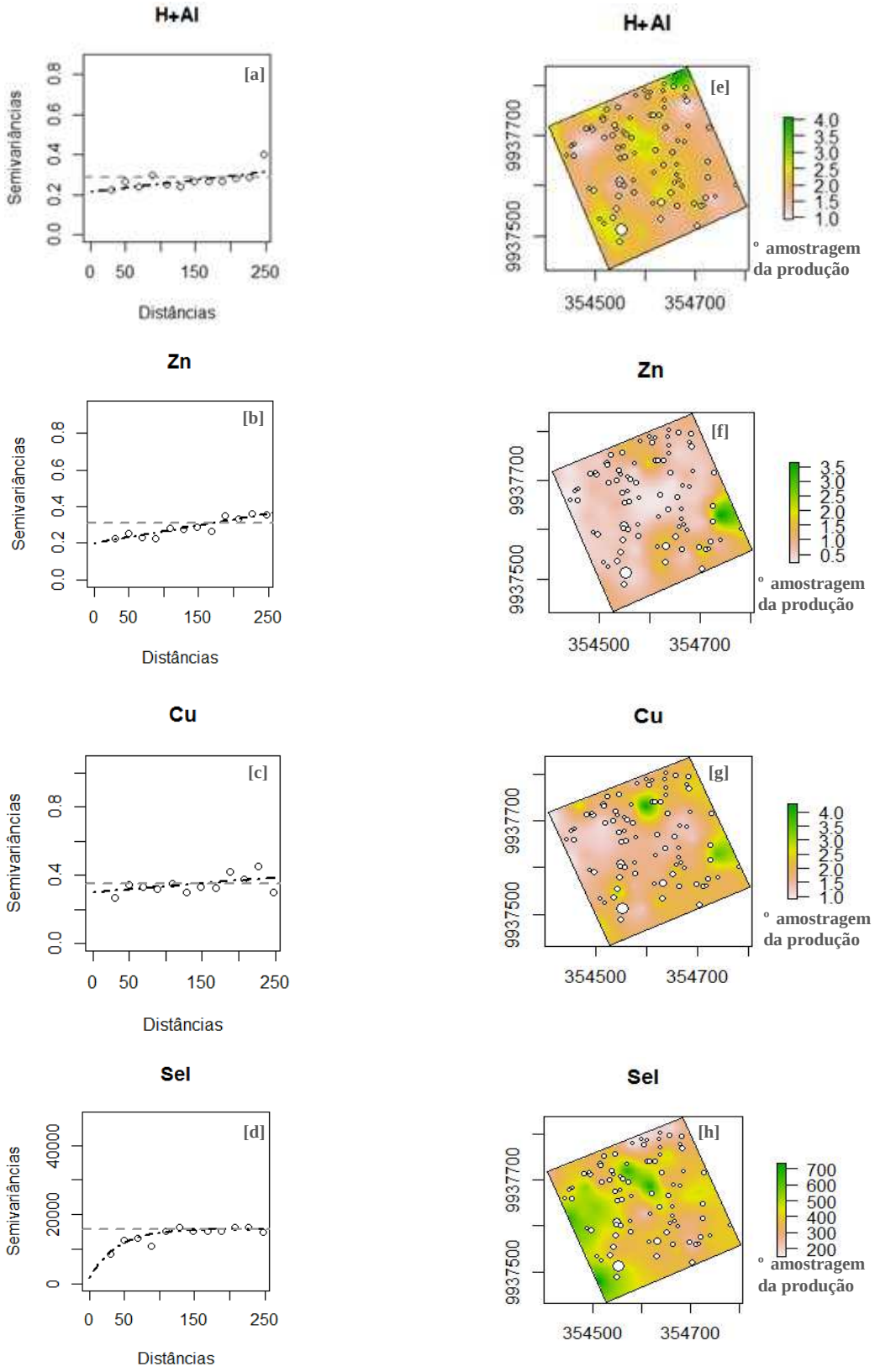


Figura 3: [a-d] Semivariogramas experimental e teórico; [e-h] mapas da predição por krigagem ordinária para os atributos do solo: Acidez Potencial (H+Al), zinco (Zn), cobre (Cu), selênio (Sel), com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil

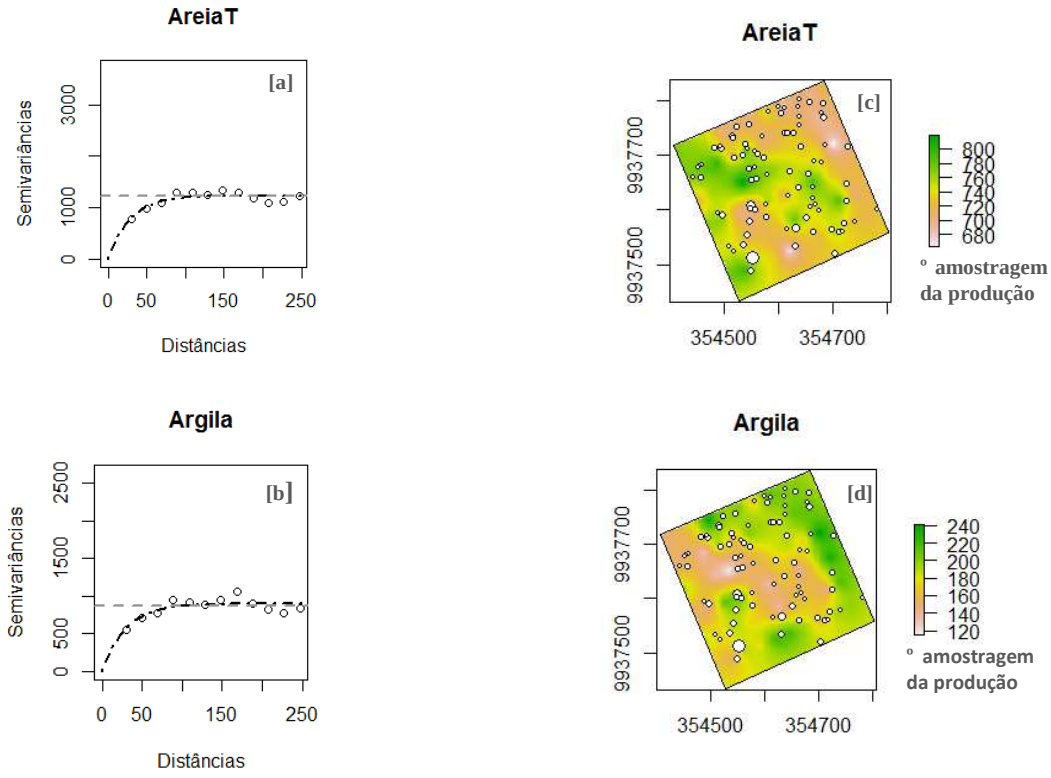


Figura 4: [a-b] Semivariogramas experimental e teórico; [c-d] mapas da predição por krigagem ordinária para os atributos do solo: areia total (AreiaT) e argila, com amostragem da produção média das castanheiras nativas no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

As áreas com altas concentrações de areia foram identificadas no mapa de krigagem mais ao noroeste e sudoeste da região em estudo que apresenta uma concentração baixa de argila (argila < 160g/kg), o que corrobora para um solo mais arenoso, ou seja, um solo textural moderadamente grosso, que normalmente apresenta um índice alto de erodibilidade, de boa drenagem e de capacidade de retenção de água (Embrapa, 2003) (Figuras 4c, 4d). Também foram observadas nessa área altas concentrações de selênio (Sel > 300µg/kg) (Figura 3h), que segundo os autores Malavolta (1980), Freitas et al. (2008) e Faria (2009), esse valor é indicativo de um solo fértil para esse componente e, consequentemente, produz menores teores de zinco (Zn < 1,0 mg/dm$^3$) e de cobre (Cu < 1,5mg/dm3) (Figuras 3f, 3g). O zinco pode ser afetado pelo pH do solo, sendo mais disponível em solos mais ácidos, composto por Acidez Potencial (Figura 3e). Fageria (2000) e Fraige et al. (2007) indicaram que concentrações de zinco superiores a 25mg/dm$^3$ podem ser tóxicos para as plantas o que não corrobora para o solo em estudo pois apresentou baixas concentrações desse nutriente (Figura 3f), os autores indicaram também que os solos mais arenosos contribuem para a lixiviação do cobre, o que pode estar influenciando nos baixos valores encontrados (Figura 3g). Segundo

Guerreiro et al. (2017), a variabilidade observada (valores altos e baixos) indica uma plasticidade da espécie em crescer em solos com variações nas concentrações dos nutrientes.

Diante do exposto, foi plotado o semivariograma e o mapa de krigagem ordinária (KO) do campo amostral para a variável produção, para predizer as possíveis áreas de maiores concentrações dessa produção na região em estudo (Figura 5).

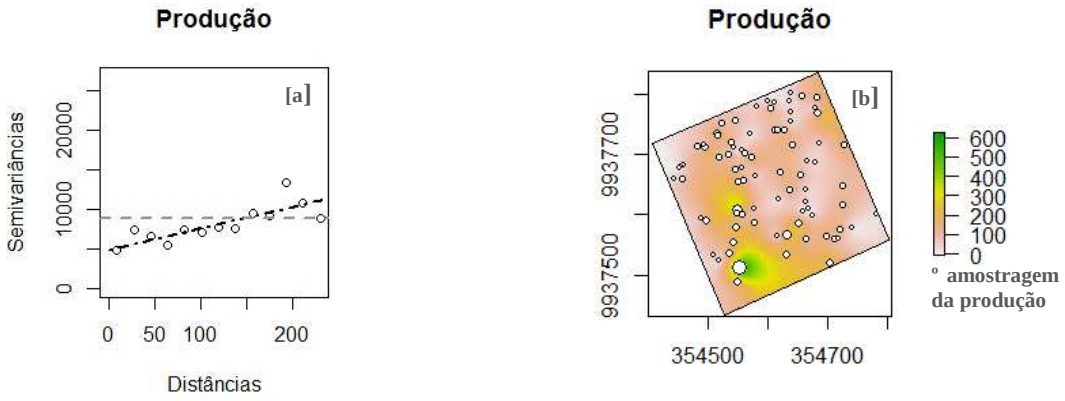


Figura 5: [a] Semivariogramas experimental e teórico; [b] mapas da predição por krigagem ordinária para a produção média das Castanheiras-da-amazônia, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

A Figura 5b representa uma indicação de que em média os valores mais altos da produção das Castanheiras-da-amazônia encontram-se mais ao sudoeste da região em estudo. Esse resultado é a confirmação dos estudos de Fernandes e Alencar (1993), Muller (1995) e Espírito-Santo et al. (2005), que em seus trabalhos indicaram maiores concentrações para a produção de castanha em solos com classificação textural arenosa.

A Figura 4g também contribuiu para confirmar o resultado da Figura 5b, pois a região sudoeste no mapa de krigagem apresentou um solo arenoso com predominância do selênio que é o nutriente mais abundante na amêndoa da Castanheira-da-amazônia segundo vários autores na literatura.

### 3.3. Variabilidade espacial da produção média das Castanheiras-da-amazônia, considerando-se a interpolação multivariada pelo híbrido Random Forest Ordinary Kriging (RFOK)

Em busca de resultados mais precisos sobre a variabilidade espacial da produção de castanhas-da-amazônia, o presente artigo, propôs o interpolador híbrido Random Forest Ordinary Kriging (RFOK) dado pela fusão computacional da KO com o Random Forest para regressão.

O interpolador híbrido RFKO apresenta a desvantagem de empregar a análise do semivariograma para selecionar as variáveis espacialmente dependentes, no entanto, dispõe de uma estrutura capaz de gerar mapas de predição em uma análise multivariada, que considera simultaneamente a influência dos nutrientes do solo para predizer a variabilidade da variável produção.

Esta abordagem não é vista no mapa de Krigagem Ordinária, pois o método de KO considera as predições para as variáveis individualmente (Figura 6a).

A seguir será apresentada uma breve discussão sobre os resultados obtidos para a produção média das castanheiras pelo método de interpolação KO e RFOK. Também, será apresentado o grau de importância de cada atributo do solo para o estudo da variabilidade espacial da produção das castanheiras (Tabela 2).

O mapa de predição por KO e RFOK para a produção média das Castanheiras-da-amazônia está representado pela Figura 6.

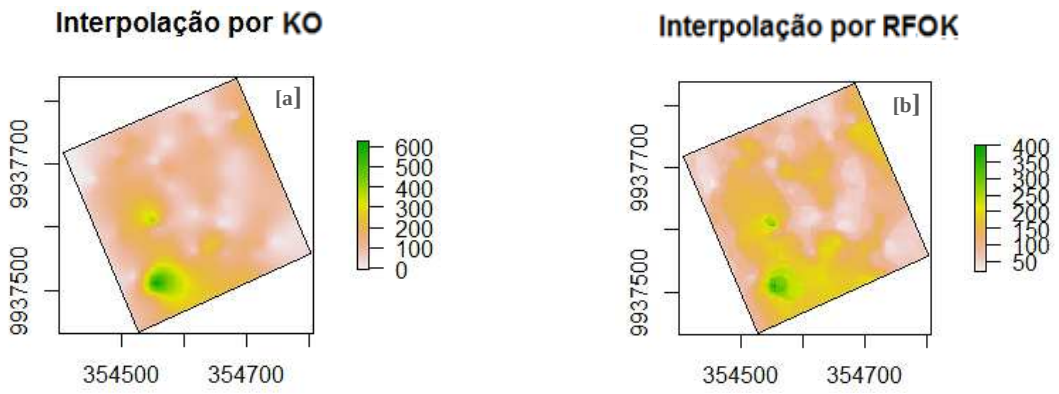


Figura 6: [a] Interpolação por Krigagem Ordinária (KO); [b] interpolação por Random Forest Ordinary Kriging (RFOK) para a produção média das Castanheiras-da-amazônia, na região em estudo no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

Pela Figura 6, percebe-se alguns pontos importantes tais como: ao observar a legenda dos valores interpolados fica evidente uma diferença nos valores; isso ocorreu porque uma única castanheira produziu algo em torno de 650UA (unidade de medida), o que pode ser considerado um outlier em relação aos demais valores representados no mapa da Figura 6a. O método de RFOK suavizou essa produção, o que gerou um destaque nas produções das demais regiões, conforme o mapa da Figura 6b. Isso aconteceu porque a krigagem é um tipo de média e, por isso, sofre uma tendência nas interpolações na presença de outliers e, como o RFOK ignorou essa medida discrepante, as demais regiões tiveram um destaque maior, contribuindo assim, para maiores detalhes sobre a produção média nessa região.

Tabela 2: Grau da importância dos atributos do solo em relação à produção das Castanheiras-da-amazônia

| Atributos do Solo | Grau de Importância |
|---|---|
| Selênio | 58449 |
| Zinco | 56761 |
| Acidez Potencial | 53508 |
| Argila | 49163 |
| Areia Total | 39113 |
| Cobre | 34893 |

A Tabela 2 é a confirmação de todos os resultados anteriores. Mesmo que a região em estudo tenha demonstrado concentração baixa de zinco no solo (Zn < 1,0mg/$dm^3$), esse nutriente foi detectado com o grau de importância para a produção das castanheiras bem próximo do selênio, o que corrobora para os resultados de vários autores tais como Filho et al., 1982, Suhet e Neptune (1987), Fraige et al. 2007, que indicam a importância do zinco para o crescimento da planta. Além disso, a argila apresentou uma importância maior do que a areia, o que era o esperado por vários autores na literatura, pelo fato de estar diretamente relacionado com disponibilidade de água e a adsorção dos nutrientes no solo.

### 3.4. Comparação do desempenho dos interpolares: Krigagem Ordinária (KO) e o híbrido Random Forest Ordinary Kriging (RFOK)

Para comparar a precisão da predição do interpolador RFOK em relação ao interpolador de KO foi calculado a estatística do erro quadrático médio (EQM) (Tabela 3).

Tabela 3: Erro quadrático médio para os interpoladores: KO e RFOK

| | KO | RFOK |
|---|---|---|
| EQM | 8407,66 | 4307,57 |

Os resultados obtidos para o critério de comparação do erro quadrático médio (EQM) (Tabela 3), demostraram que os menores erros foram obtidos para o interpolador RFOK, que reduziu quase 50% dos erros indicando uma melhor eficiência para o interpolador híbrido.

# 4. CONCLUSÕES

Os resultados mostraram que o tipo de solo textural arenoso com presença moderada de argila, com destaque para os elementos selênio e zinco podem estar mais fortemente associados às altas produções de frutos das Castanheiras-da-amazônia na região em estudo.

O interpolador RFOK conseguiu uma análise multivariada para a produção das Castanheiras-da-amazônia de forma satisfatória ao discriminar o percentual da importância de cada atributo do solo simultaneamente.

O método também ignorou o outlier presente na amostragem da produção suavizando o mapa de predição e gerando um destaque para as demais regiões. Além disso, uma maior acurácia foi identificada nos resultados obtidos por esse interpolador, uma vez que conseguiu explicar de forma mais precisa a variabilidade espacial da produção, em função da variabilidade espacial dos nutrientes presentes no solo, apresentando uma redução de quase 50% do coeficiente do erro quadrático médio (EQM).

Finalmente, com os resultados apresentados, conclui-se que o RFOK apresentou uma menor diferença entre o valor real e o valor predito na interpolação das variáveis, o que irá contribuir para subsídios mais confiáveis no que se refere ao manejo florestal, à manutenção e a ampliação da produção das castanheiras na região em estudo.

## REFERÊNCIAS BIBLIOGRÁFICAS

ALVARES, C. A; STAPE, J. L; SENTELHAS, P. C; GOLÇALVES, M. J. L; SPAROVEK, G. ‘Köppen's climate classification map for Brazil’**, Meteorologische Zeitschrift**, vol. 22, no. 6, pp. 711–728, 2013.

BAKER, P. T; CAUDILL, S; HODGE, H. A; TALUKDER, D; CAPANO, C; CORNISH, N. J. Multivariate classification with random forest for gravitational wave searchers of black hole binary coalescence. **Physical Review Journals**. v.91(6), p.1-26, 2015.

BREIMAN, L. Bagging predictors. **Machine learning**, v.24 (2), p.123-140, 1996. http://dx.doi.org/10.1023/A:1018054314350

BREIMAN, L. Random Forest. **Machine Learning**, v.45 (1), p.5-32, 2001. Disponível em: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>. Acesso em: 05 abr. 2018.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN. R.; STONE, C. **Classification and Regression Trees**. Taylor & Francis, 1984. 368p.

BRINK, H.; RICHARDS, J. **Real World Machine Learning**. [S.l.]: Manning Publications C.O, 2014.

CAMPOS, M. C. C; SOARES, M. D. R; SANTOS, L. A. C; OLIVEIRA, I. A; AQUINO, E. A. Spatial variability of physical attributes in Alfissol under agroforestry, Humaitá region, Amazonas state, Brazil. **Rev. de Ciências Agrária**. v.56, p.149-159, 2013.

CARVALHO, H. M. **Aprendizado de máquina voltado para mineração de dados: árvores de decisão**. Monografia (Graduação). Departamento de Engenharia de Software. Universidade de Brasília. Brasília, 2014.

CENTENO, L. N; GUEVARA, M. D. F; CECCONELLO, S. T; SOUSA, R. O. D; TIMM, L.C. Textura do solo: Conceitos e aplicações em solos arenosos. **Revista Brasileira de Engenharia e Sustentabilidade**, v.4, n.1, p.31-37, 2017.

EMBRAPA– Empresa Brasileira de Pesquisa Agropecuária. **Solos**. Revista Embrapa Algodão Sistemas de Produção, 3 ISSN 1678-8710, Versão Eletrônica, 2003. Disponível em: https://www.ft.unicamp.br/~sandro/.../Solos%20-%20EMBRAPA%20ALGODÃO.doc. Acesso em: 25 nov. 2019.

EMBRAPA– Empresa Brasileira de Pesquisa Agropecuária. **Guia Prático para Interpretação de Resultados de Análises de Solo**. Documentos 206. Embrapa Tabuleiros Costeiros Aracaju, ISSN 1678-1953, 2015.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Solos. **Sistema Brasileiro de Classificação dos Solos**. 2.ed. Rio de Janeiro: Embrapa Solos, 2006. 306p.

ESPIRITO-SANTO, F. D. B; SHIMABUKURO, Y. E; OLIVEIRA, L. E; ARAGÃO, F. C; MACHADO, E. L. M. Análise da composição florística e fitossociológica da Floresta

Nacional do Tapajós com o apoio geográfico de imagens de satélites. **Acta Amaz**, v.35, n.2, p.155-173, 2005.

FARIA, L. A. **Levantamento sobre selênio em solos e plantas do estado de São Paulo e sua aplicação em plantas forrageiras**. Dissertação (mestrado). Universidade de São Paulo. Pirassununga, 2009

FERNANDES, N. P; ALENCAR, J. C. Desenvolvimento de árvores nativas em ensaios de espécies - Castanha-da-amazônia (Bertholletia excelsa H.B.K.), dez anos após o plantio. **Acta Amaz**, v.23, n.2, p.191-198, 1993.

FREITAS, S. C; GONÇALVES, E. B; ANTONIASSI, R; FELBERG, I. OLIVEIRA, S. P. Meta-análise do teor de selênio em Castanha-da-amazônia. **Brazilian Journal of Food Technology**, v. 11, n. 1, p. 54-62, 2008.

FILHO, M. P. B; FAGERIA, N. K; CARVALHO, J. R. P. Fonte de zinco e modos de aplicação sobre a produção de arroz em solos de serrado. **Pesquisa agropecuária brasileira**, Brasília, 17(12):1713-1719, 1982.

FAGERIA, N. K. Níveis adequados e tóxicos de zinco na produção de arroz, feijão, milho, soja, e trigo em solo de cerrado. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.4(3), p.390-395, 2000.

FRAIGE, K; CRESPILHO, N; REZENDE, M. O. O. Determinação de zinco em solo utilizando calorimetria. **Química Nova**, v.30 (3), p.588-591, 2007.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. Springer Series in Statistics. Second Edition. California, 2008. 744p.

IBGE. 2012. **Manual técnico da vegetação basileira. Coordenação de Recursos Naturais e Estudos Ambientais**. Série Manuais Técnicos em Geociências 1, 2. Ed. revista e ampliada. IBGE, Rio de Janeiro.

IBGE. 2015. **Manual técnico de pedologia. Coordenação de Recursos Naturais e Estudos Ambientais**. Série Manuais Técnicos em Geociências 1, 3. ed. IBGE, Rio de Janeiro.

GUERREIRO, Q. L. M; JÚNIOR, R. C. O; SANTOS, G. R; RUIVO, M. L. P; BELDINI, T. P; CARVALHO, E. J. M; SILVA, K. E; GUEDES, M. C; SANTOS, P. R. B. Spatial variability of soil physical and chemical aspects in a Brazil nut tree stand in the Brazilian Amazon. **African Journal of Agricultural Research**. v.12(4), p.237-250, 2017.

GOMES, M. A. F; FILIZOLA, H. F. **Indicadores físicos e químicos de qualidade de solo de interesse agrícola**. EMBRAPA Meio Ambiente. Jaguariúna, 2006.

ISAAKS, E. H; SRIVASTAVA, R. M. **An introduction to applied geostatistcs**. New York: Oxford University Press, p.561, 1989.

INSTITUTE OF MEDICINE (ESTADOS UNIDOS). **Dietary reference intakes for vitamin C, vitamin E, selenium and carotenoids: a report of the panel on dietary**

**antioxidants and related compounds.**Washington: National Academy, 2000. 506 p. Disponível em: <http://nap.edu/openbook/0309069351/html/R1.html>. Acesso em: 25 nov. 2018.

JAMES, G; WITTEN, D; HASTIE, T; TIBSHIRANI, R. **An Introduction to Statistical Learning with applications in R**. Springer New York Heidelberg Dordrecht London, 2013. 426p.

JUNIOR, W. C.; CALDERANO FILHO, CHAGAS, C. da S.; B.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas**. Pesq. agropec. Bras. Brasília**, v.51 (9), p.1428-1437, 2016. DOI: 10.1590/S0100-204X2016000900041

JUNIOR, E. C. DA S. **Selênio na Castanha-da-amazônia (Bertholletia excelsa) em solos da região amazônica brasileira**. Dissertação (Mestrado). Universidade Federal de Lavras. Lavras, 2016.

LIAW, A.; WIENER, M. **Classification and regression by random forest**. R News, v.2, p.18-22, 2002.

LIMA, J. S. S; SILVA, A. S; SILVA, J. M. Variabilidade espacial de atributos químicos de um Latossolo Vermelho-Amarelo cultivado em plantio direto. **Rev. Ciências Agronômica**. v.44(1), p.16-23, 2013.

LOPES, T. D.; GOEDTEL, A.; PALACIOS, R. H. C.; GODOY, W. F. **Aplicação do Algoritmo Random Forest como classificador de padrões de falhas em rolamentos de motores de indução**. XIII Simpósio Brasileiro de Automação Inteligente. Porto Alegre, 2017.

MACHADO, L. O; LANA, A. M. Q; LANA, R. M. Q; GUIMARÃES, E. C; FERREIRA, C. V. Variabilidade Espacial de atributos químicos do solo em áreas sob sistema plantio convencional. **Revista Brasileira de Ciências do Solo**, v.31, p. 591-599, 2007.

McBRATNEY, A.G.; WEBSTER, A.G. Choosing functions for semi-variograms and fitting them to sampling estimates. Journal of Soil Science, v.37, p.617-39, 1986.

MALAVOLTA, E. **Elementos de nutrição mineral de plantas**. Ed. Agronômica Ceres, p.211-212, 1980.

MATÍAS, J. M. et al. Comparison of Kriging and Neural Networks With Application to the Exploitation of a Slate Mine. **Mathematical Geology**, v. 36, n. 4, p. 463–486, 2004.

MELLO, J. M. DE; BATISTIA, J. L. F.; RIBEIRO JUNIOR, P. J.; OLIVEIRA, M. S. Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de Eucaliptus grandis. **Scientia Florestalis**, v.1, n.69, p.25-37, 2005.

MOOLMAN, J. H; VAN HUYSSTEEN, L. A geostatistical analysis of the penetrometer soil strength of a deep ploughed soil. **Soil Tillage Research**. v.15(1-2), p.11-24, 1989.

MORAIS, R. L. **Uso de árvores aleatórias para classificação sensorial de arroz cozido**. Monografia (Graduação). Departamento de Estatística da Universidade de Brasília. Brasília, 2017.

MULLER, C. H; FIGUEIRÊDO, F. J. C; KATO, A. K; CARVALHO, J. E. U; STEIN, R. L. B; SILVA, A. B. **A cultura da Castanha-da-amazônia**. Belém: EMBRAPA, p.65, 1995.

NOGUEIRA, A. R. A; SOUZA, G. B. **Manual de laboratório: solo, água, nutrição vegetal, nutrição animal e alimentos**. São Carlos: Embrapa Pecuária Sudeste. p.334, 2005.

OLIVEIRA, I. A; CAMPOS, M. C. C; SOARES; M. D. R; AQUINO, R. E; MARQUES JÚNIOR, J; NASCIMENTO, E. P. Variabilidade espacial de atributos físicos em um Cambissolo Háplico, sob diferentes usos na região Sul do Amazonas. **Rev. Bras. de Ciências do Solo**. v.37, p.1103-1112, 2013.

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica**. Dissertação (Mestrado). Departamento de Bioinfomática. Ribeirão Preto, 2013.

PASINI, M. P. B; LÚCIO, A. D; FRONZA, D; WEBER, L. S. Krigagem ordinária e inverso da distância ponderada aplicados na espacialização da população da mosca-do-figo. **Revista Brasileira de Ciências Agrárias, v.10(3), p.452-459, 2015.**

PIRES, C. A. F.; STRIEDER, A. J. Modelagem Geoestatística de dados geofísicos, aplicada a pesquisa de Au no prospecto volta grande (Complexo Intrusivo Lavras do Sul, RS, BRASIL). **Revista Geomática**, v.1(1), 2006.

RAIJ, B. VaN.; QUAGGIO, J.A.; CANTARELLA, H.; FERREIRA, M.E.; LOPES, A.S.; BATAGLIA, O.C. **Análise química do solo para fins de fertilidade**. Campinas: Fundação Cargill, 1987. 170p.

R-DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2017.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM JOURNAL**, **Research and Development**, v.3, p. 210-219, 1959.

SARAIVA, G. S.; BONOMO, R.; DE SOUZA, J. M. Avaliação de interpoladores geoestatísticos e determinísticos da evapotranspiração de referência diária para o estado do Espírito Santo. **Revista Agroambiente On-line**, v.11(1), p.21-30, 2017.

SUHET, A. R; NEPTUNE, A. M. L. Efeito do ferro e do zinco e da natureza de três tipos de solo na produção de matéria seca e na composição química do feijoeiro (Phaseolus Vulgaris l.). **Scielo** v.36, 1979

VIEIRA, S. R. **Geoestatística em estudos de variabilidade espacial do solo.** In. NOVAES, R. F.; ALVAREZ V., V. H.; SCHAEFER, C. E G. R. Tópicos em ciências do solo. Viçosa, MG. **Sociedade Brasileira de Ciência do Solo,** v.1. p.2-54, 2000.

WHANGER, P. D. Selenium and its relationship to cancer: an update. **British Journal of Nutrition**, Cambridge, v. 91, n. 1, p. 11-28, 2004.

WARRICK, A.W; NIELSEN, D.R. **Spatial variability of soil physical properties in the field**. In: HILLEL, D. (Ed.). Applications of soil physics. New York: Academic Press, p. 319-44, 1980.

# CAPÍTULO 2

## RANDOM FOREST KRIGING (RFK): UMA PROPOSTA DE UM PREDITOR MULTIVARIADO GEOESTATÍSTICO


### RESUMO

A metodologia de uma abordagem geoestatística robusta tem sido uma das técnicas mais eficientes para auxiliar o estudo dos fatores que tem afetado a capacidade produtiva das Castanheiras-da-amazônia na atividade extrativista no bioma Amazônia. Relacionar a variabilidade espacial dos atributos do solo com a produção dessa espécie pode direcionar amostragens mais eficientes, com programas de plantio para o enriquecimento da espécie e subsidiar modelos de previsão de safra e políticas públicas. O método de interpolação espacial mais utilizado na geoestatística é a krigagem, entretanto, apresenta a desvantagem de ser computacionalmente inviável para modelar o estimador de semivariograma em grandes conjuntos de dados e, além disso, descartar variáveis importantes no estudo pela presença do efeito pepita puro. Dessa forma, para contornar essa desvantagem, foi desenvolvido na literatura o interpolador de Fixed Rank Kriging (FRK) que utiliza a teoria do modelo espacial de efeitos aleatórios (SRE) para modelar a dependência espacial através de um número fixo e pré-determinado de funções de base, independente da teoria do semivariograma. Para melhorar a capacidade de predição, como proposta de trabalho, foi implementado um híbrido do FRK com o algoritmo Random Forest para regressão em uma abordagem multivariada e, para fins de comparação, foi realizada uma análise via erro quadrático médio (EQM) e coeficiente de determinação ($R^2$) para o desempenho do interpolador implementado com a krigagem já existente. Para isso, foram coletadas amostras do solo de 50m×30m em todas as linhas da região do estudo e amostras da produção média das castanheiras no período 2007 a 2015. As análises estatísticas e geoestatísticas foram realizadas no ambiente computacional do software R e todos os pontos foram georreferenciados. Como resultado obteve-se um ajuste perfeito do modelo e uma redução significativa para erro quadrático médio ao utilizar-se o interpolador híbrido implementado, como também, o grau de importância de cada atributo do solo para predizer a variabilidade espacial da produção média das Castanheiras-da-amazônia.


**Palavras-chave:** Fixed Rank Kriging (FRK). Random Forest. Interpolador Híbrido. Castanheiras-da-amazônia.

## ABSTRACT

The methodology of a more robust geostatistical approach has been one of the most efficient techniques to assist the study of the factors that have affected the productive capacity of Chestnuts of the Amazon in the extractive activity in the Amazon biome. Relating the spatial variability of soil attributes to the production of this species can lead to more efficient sampling, with planting programs to enrich the species and subsidize crop forecasting models and public policies. The spatial interpolation method most used in geostatistics is kriging, however, it has a disadvantage of being computationally unfeasible to model the semivariogram estimator in large data sets and, furthermore, to discard important variables in the study due to the presence of the pure nugget effect. Therefore, to circumvent this disadvantage, the Fixed Rank Kriging (FRK) interpolator was developed in the literature, which uses the theory of the spatial model of random effects (SRE) to model the spatial dependence through a fixed and predetermined number of functions basis, independent of the semivariogram theory. To improve the prediction capacity, as a work proposal, a FRK hybrid with the Random Forest algorithm was implemented for regression in a multivariate approach and, for comparison purposes, an analysis was carried out via mean square error (EQM) and determination coefficient (R2) for the performance of the interpolator implemented associated with the existing kriging. For that, samples of soil of 50m × 30m were collected in all lines of the study region and samples of the average production of chestnut trees in the period 2007 to 2015. Statistical and geostatistical analyzes were performed in the computational environment of the R software and all points were georeferenced. As a result, a perfect fit of the model was obtained and a significant reduction for mean squared error when using the implemented hybrid interpolator, as also, the degree of importance of each soil attribute to predict the spatial variability of the average production of Chestnuts of the Amazon.

**Keywords:** Fixed Rank Kriging (FRK). Random Forest. Hibrid Interpolator. Chestnuts of the Amazon.

# 1. INTRODUÇÃO

Um dos grandes desafios encontrados na atividade extrativista no bioma Amazônia tem sido construir diretrizes técnicas para auxiliar no estudo dos fatores que afetam a capacidade produtiva das Castanheiras-da-amazônia, árvores nativas, que segundo Guerreiro et al (2017), são consideradas nobres pela importância em âmbito ecológico, econômico e social para a região.

Umas das técnicas consideradas eficientes para construir essas diretrizes têm sido o estudo da variabilidade espacial dos atributos do solo relacionado com a produção das Castanheiras-da-amazônia. Esse tipo de estudo é capaz de direcionar amostragens que minimizam as possibilidades de erros visando projetos de plantio para o enriquecimento da espécie e subsídios de modelos de previsão de safra e políticas públicas para determinar um preço mínimo dos produtos que possam agradar produtores e consumidores em diversas regiões.

Para a execução desse estudo é necessário empregar a metodologia da Geoestatística, uma vez que utiliza modelos de dependência espacial capaz de estimar valores de pontos nos locais onde não foram amostrados por meio do cálculo da semivariância e da interpolação espacial com o objetivo de gerar mapas representativos da realidade (Moolman e Van Huyssteen, 1989).

Na Geoestatística existem vários métodos univariados de interpolação espacial, tais como krigagem ordinária, krigagem simples, krigagem universal, krigagem fatorial, krigagem indicatriz, entre outras. Dentre todos os métodos, quando a média estacionária dos dados for desconhecida é recomendado por Santos et a.l (2011) utilizar a krigagem ordinária. Porém, esse método apresenta a desvantagem de ser computacionalmente inviável para modelar o estimador do semivariograma em grandes conjuntos de dados e descartar variáveis que podem ser consideradas importantes para o estudo pela presença do efeito pepita puro. Para mais detalhes observar Yamamoto e Landim, 2013.

Para contornar essa situação, Cressie e Johannesson (2008) desenvolveram o Fixed Rank Kriging (FRK) que utiliza a teoria do modelo espacial de efeitos aleatórios (SRE) para modelar a dependência espacial através de um número fixo e pré-determinado de funções de base, independente da teoria do semivariograma. Esse preditor apresenta uma abordagem univariada em que cada variável, em estudo, será apresentada em uma análise individualizada.

Assim, para melhorar a capacidade de predição do FRK foi desenvolvido computacionalmente, como proposta do trabalho, o interpolador híbrido do FRK com o algoritmo Random Forest para regressão, denominado por Random Forest Kriging (RFK), utilizando uma abordagem multivariada. E, para fins de comparação, quanto ao desempenho desses interpoladores será empregada o coeficiente do erro quadrático médio (EQM) e o coeficiente de determinação do modelo ($R^2$).

Diante do exposto, espera-se que o método de interpolação multivariado híbrido desenvolvido em uma abordagem estatística mais robusta possa contribuir para melhor descrever a importância simultânea de cada atributo do solo para predizer à variabilidade espacial da produção média das castanheiras-da-amazônia.

O estudo foi desenvolvido em três etapas: (1) Estudo da variabilidade espacial dos atributos do solo e da produção média das castanheiras-da-amazônia, considerando-se a elaboração de mapas de predição em uma análise univariada pelo interpolador FRK; (2) Estudo da variabilidade espacial da produção média das castanheiras-da-amazônia e do grau de importância de cada atributo do solo para predizer essa variabilidade, considerando-se a interpolação multivariada pelo RFK; e (3) Comparação do desempenho dos interpoladores FRK e RFK para a variável de interesse produção média das castanheiras-da-amazônia, considerando-se o coeficiente do erro quadrático médio (EQM) e o coeficiente de determinação do modelo $R^2$.

## 2. MATERIAL E MÉTODOS

Buscando alcançar os objetivos dessa proposta serão apresentados os conceitos mais relevantes para compreensão dos resultados.

### 2.1. Descrição da Área de Estudo

A região do estudo está localizada no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil, delimitado pelas latitudes: 9937400S; 9937900S e pelas longitudes: 354400W; 354850W (Figura 1).

Essa área apresenta uma vegetação local de mata nativa com agregados de castanheiras que é classificada como Floresta Ombrófila Densa Submontana com dossel emergente e o solo Argissolo Vermelho-amarelo prevalecente (Ibge, 2012; 2015). O clima da

região é tropical de monções com duas estações bem determinadas, período com chuvas intensas e outro com menos intensidade (Alvares et al., 2013).

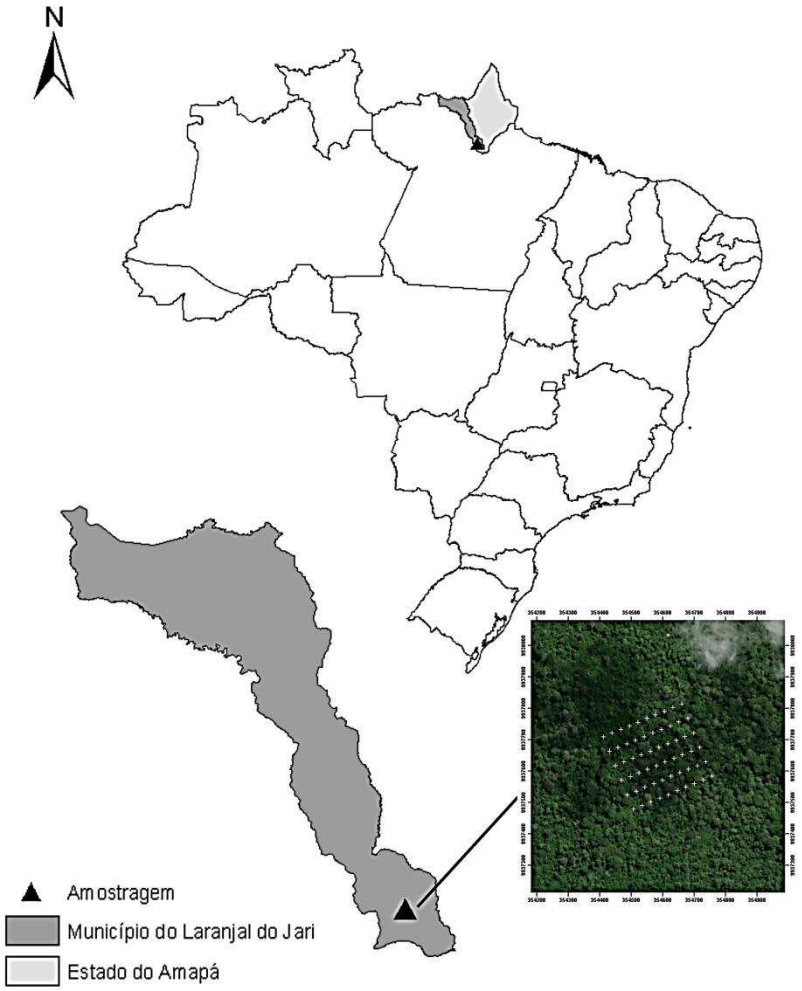


Figura1: Área de localização da reserva extrativista do Rio Cajari no município de Laranjal do Jari no sul do Estado do Amapá, Brasil.

## 2.2. Descrição dos Dados Amostrados

O castanhal nativo presente nessa região insere-se na configuração de uma pequena parcela do bioma Amazônia dividida em seis transectos de 300 m, equidistantes em 50m, para orientar a amostragem, totalizando 60 amostras do solo, georreferenciadas ao sistema geodésico Datum SIRGAS 2000 e sistema de coordenada UTM (Universal Transverse Mercator Coordinate System) no fuso 22S (Figura2).

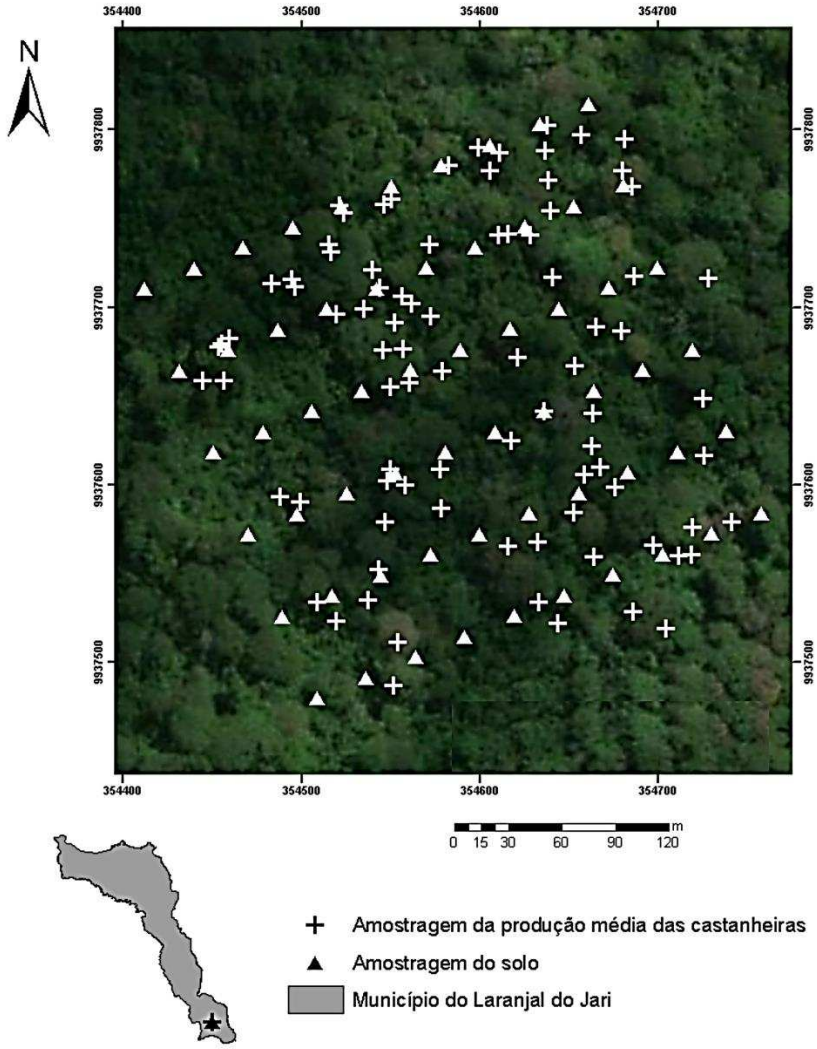


Figura 2: Amostragem do solo e da produção média do castanhal nativo da reserva extrativista do Rio Cajari, município de Laranjal do Jari, no sul do Estado do Amapá, Brasil.

A uma profundidade de 0-20 cm foram coletadas amostras de solo com o auxílio de um trado holandês. Os cálculos para a determinação granulométrica e química dos atributos do solo, que são os indicadores físicos e químicos da qualidade do solo citados por Gomes e Flizola (2006), e a obtenção de terra fina seca ao ar (TFSA) foram obtidos conforme as recomendações de Nogueira e Souza (2005) e Embrapa (2007).

Os atributos do solo quantificados foram: potencial hidrogeniônico (pH); carbono (C)(g/kg); matéria orgânica (mo)(g/kg); nitrogénio (N)(g/kg); fósforo (P)(mg/dm3); potássio (K)(mg/dm3); sódio (Na)(mg/dm3); cálcio (Ca)(cmolc/dm3); magnésio (Mg)(cmolc/dm3); alumínio (AL)(cmolc/dm3); Acidez Potencial (H+AL)(cmolc/dm3); soma de base (SB)(cmolc/dm3); trocas de cátions efetiva (t)(cmolc/dm3); trocas de cátion a pH 7

(Tc)(cmolc/dm3); saturação por base (V)(%); saturação por alumínio (m)(%); ferro (Fe)(mg/dm3); zinco (Zn)(mg/dm3); manganês (Mn)(mg/dm3); cobre (Cu)(mg/dm3); selênio (Sel)(µg/kg); areia grossa (AreiaG)(g/kg); areia fina (AreiaF)(g/ kg); areia total (AreiaT)(g kg/1); silte (g/kg) e argila (g/kg).

Os pontos coletados para a produção média das castanheiras (Producao) (UA), referente ao período de 2007 a 2015, foram também, georrefenciadas no sistema de coordenadas UTM no fuso 22S, SIRGAS 2000, para 82 amostras (Figura 2). Para coleta foi considerado como produção zero as castanheiras adultas que apresentaram o diâmetro maior que o diâmetro da menor castanheira produtiva que não produziram frutos no período; foram excluídas as castanheiras jovens não reprodutivas com diâmetros menores que o diâmetro da menor castanheria produtiva.

## 2.3. Proposição do Método

Em busca de um preditor que apresente um melhor desempenho para a execução da interpolação de uma variável de interesse, mais especificamente, analisar simultaneamente a relação dos atributos do solo com a variabilidade espacial da produção média das Castanheiras-da-amazônia, o presente artigo propôs a implementação de um interpolador híbrido, que é a junção do Fixed Rank Kriging (FRK) com o algoritmo Random Forest para regressão, denominado de Random Forest Kriging (RFK) (Figura3).

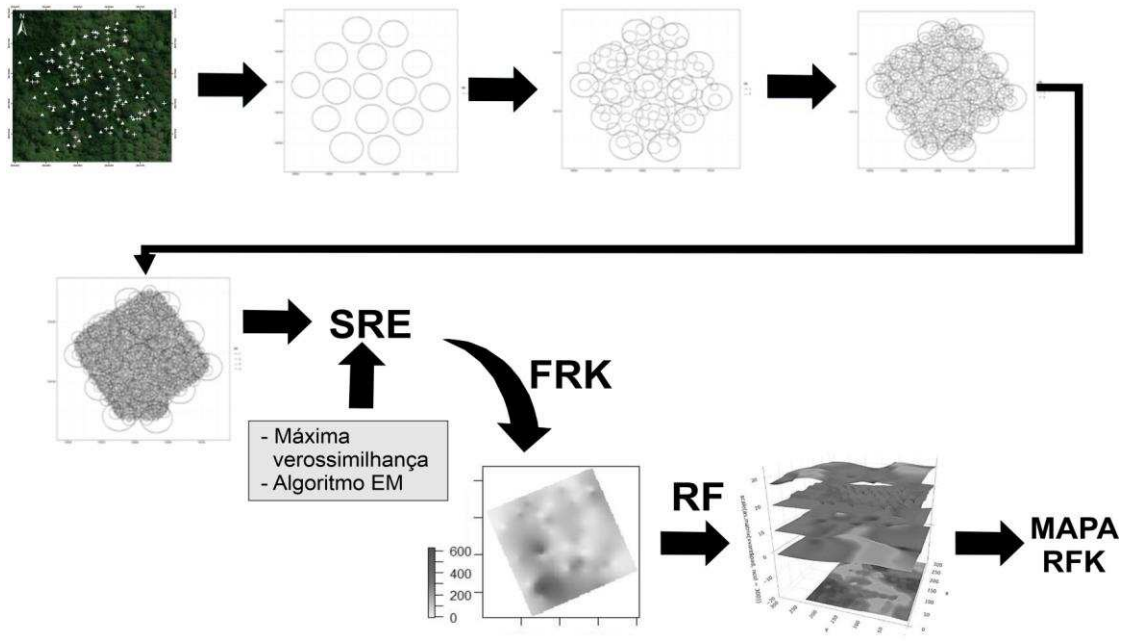


Figura 3: Representação do Random Forest Kriging (RFK)

A Figura 3 representa a construção do método computacionalmente implementado por essa pesquisa, o RFK, que se refere a uma adaptação do método FRK ao algoritmo Random Forest (RF). Inicialmente, o FRK utiliza o modelo SRE para construção das unidades básicas de área em uma interpolação univariada de cada variável. Em seguida, o RF, um método não espacial e independente, emprega o empilhamento dos rasters gerados no FRK para extração dos pontos que serão utilizados no conjunto de treinamento da metodologia de árvore e, em uma abordagem multivariada, gera o mapa final de predição.

A proposição desse novo método de interpolação, implementado computacionalmente, embasou-se pela desvantagem de alguns métodos de interpolação geoestatísticos presentes na literatura descartarem variáveis consideradas importantes no estudo devido à presença do efeito pepita puro, como também, pela busca de um interpolador que utilize uma abordagem multivariada para avaliar o comportamento da variabilidade espacial de múltiplas variáveis envolvidas no estudo sem a restrição do tamanho do conjunto de dados.

Adotando o entendimento da metodologia geoestatística espacial do interpolador FRK e do algoritmo Random Forest para regressão foi possível, computacionalmente, realizar a junção desses dois métodos e construir um interpolador híbrido que apresenta uma série de vantagens, tais como: não depende da teoria do semivariograma para a construção do modelo; não exclui variáveis que podem ser consideradas importantes para o estudo pela presença do efeito pepita; não apresenta restrições quanto ao tamanho do conjunto de dados. Além disso, apresenta uma capacidade melhor de predição pelo fato de dispor de uma estrutura adequada para gerar mapas de predição em uma análise multivariada, considerando um processo de empilhamento de rasters para todas as variáveis em estudo, concatenando-os em uma mesma resolução para extração dos pontos que serão utilizados na interpolação (Figura 3).

Outro aspecto interessante desse interpolador híbrido é a capacidade de capturar a influência simultânea da variabilidade espacial das variáveis presentes no estudo e determinar o grau de importância de cada variável para predizer a variabilidade espacial da variável de interesse.

Nas subseções que seguem apresenta-se uma breve introdução dos fundamentos teóricos que foram necessários para a implementação do interpolador híbrido proposto por esse trabalho.

### 2.3.1 Fixed Rank Kriging (FRK)

Com o método de interpolação de krigagem é possível obter predição ótima e erro padrão de predição para dados ruidosos e incompletos de qualquer local de interesse, gerando mapas espacialmente completos. No entanto, esse procedimento requer a inversão da matriz de covariância espacial, $\Sigma_{nxn}$, que determina os pesos atribuídos às diferentes amostras, em que $n$ denota o conjunto de dados (Zhu et. al., 2015).

A inversão dessa matriz de covariância exige o cálculo muito intenso ou mesmo inviável para grandes conjuntos de dados. Dessa forma, o Fixed Rank Kriging (FRK) foi desenvolvido para reduzir a dimensão dessa matriz evitando modelos estacionários e isotrópicos de covariância e semivariograma (Zhu et al, 2015; Cressie e Johannesson, 2008).

O FRK é uma estrutura de previsão e modelagem espacial e/ou espaço-temporal, considerada um BLUP espacial (melhor previsão espacial linear não tendencioso), que utiliza o modelo espacial de efeitos aleatórios (SRE) em um domínio discretizado de unidade básica de área para alcançar a redução da dimensão dos dados modelando a dependência espacial por meio de um número fixo e pré-determinado de funções de base (Cressie, 2008; Zhu et. al., 2015).

O FRK decompõe o processo espacial em um componente de tendência deterministica de funções lineares de covariáveis espaciais e um componente aleatório da variação espacial determinado por um modelo espacial de efeitos aleatórios (SRE). O processo espacial é dado por (Zammit-Mangion e Cressie, 2018):

$$\mathrm{Y}(.) = \mu(.) + \mathrm{v}(.) \qquad [1]$$

$$\hat{\mu}(\mathrm{s}_0) = t(\mathrm{s}_0)'\hat{\beta}\ ;\ \hat{\beta} = (T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}Z$$

$$\mathrm{v}(.) = S(.)'\eta + \varepsilon(.)$$

$$\eta \sim N(0, K);\ \varepsilon(.) \sim N(0, \sigma^2 V)$$

$$\Sigma^{-1} = (\sigma^2 V)^{-1} - (\sigma^2 V)^{-1} S\{K^{-1} + S'(\sigma^2 V)^{-1} S\}^{-1} S'(\sigma^2 V)^{-1}$$

em que $\mathrm{Y}(.)$ é um processo espacial de valores; $\mathrm{s}_0$ vetor de localização espacial; $\mathrm{t}(.)$ vetor de covariáveis espacialmente referenciadas; $\hat{\alpha}$ vetor de coeficientes de regressão estimados (efeito fixo); $T$ matriz de covariáveis conhecidas espacialmente referenciadas; $Z$ matriz de localização espacial; $\Sigma^{-1}$ inversa da matriz de variâncias e covariâncias; $S(.)$ vetor espacial de funções de base; $\eta(.)$ é um vetor espacial de efeitos aleatório; K matriz finita positiva $mxm$; V é a matriz diagonal $nxn$.

Essa combinação de uma matriz $K_{mxm}$ positiva definida com um conjunto de funções de base produz uma família flexível de funções de covariância (Nguyen, 2009). Para qualquer matriz positiva definida $\mathrm{K_{mxm}}$ e $\sigma^2 > 0$, podemos escrever a função de covariância como $\mathrm{C} = \mathrm{SKS}'$ e, portanto, $\Sigma = \mathrm{SKS}' + \sigma^2\mathrm{V}$.

Desta forma, a escolha da função de covariância $C = SKS'$ permite maneiras alternativas de calcular as equações de krigagem envolvendo a inversão apenas de matrizes $mxm$ em que $m < n$, independente do semivariograma.

A estimação dos parâmetros desse modelo é feita via algoritmo EM e a variância de krigagem é dada por:

$$\begin{aligned}\sigma^2{}_{FRK}(s_0) = {} & S(s_0)'KS(s_0) - S(s_0)'KS'\Sigma^{-1}SKS(s_0) \\ & + \left(t(s_0) - T'\Sigma^{-1}SKS(s_0)\right)'(T'\Sigma^{-1}T)^{-1}\left(t(s_0)\right. \\ & \left. - T'\Sigma^{-1}SKS(s_0)\right) \qquad [2]\end{aligned}$$

O domínio espacial discretizado é conhecido como uma unidade base de área, que permite gerar mapas espacialmente completos de predições conforme ocorre a variação da previsão de localização nas equações do preditor de krigagem e da variância de krigagem (Cressie e Johannesson, 2008) e, admite-se também, combinar várias observações fazendo distinção entre o erro de medição e a variação de escala desse elemento, o que leva a uma melhor quantificação da incerteza (Zammit-Mangion e Cressie, 2018).

### 2.3.2. Random Forest para Regressão

Random Forest (RF) é um algoritmo de aprendizado de máquina baseado em árvores de decisão, desenvolvido por Breiman et al. (1984), capaz de fazer predições sobre os dados. Esse algoritmo apresenta uma metodologia de partição binária que particiona o espaço de interesse em um conjunto de sub-regiões, que utilizam um conjunto de treinamento supervisionado para a classificação quando a variável resposta é qualitativa, ou para a regressão quando a variável resposta é quantitativa (Breiman, 2001; James et al., 2013).

Random Forest para regressão é uma combinação de várias árvores aleatórias, independentes e com mesma distribuição, em que o resultado final é a média de todos os possíveis resultados (Junior et al. 2016). Em seu desenvolvimento, o RF utiliza a técnica bagging para produzir amostras aleatórias de conjuntos de treinamento para cada árvore

aleatória. Essa técnica de amostragem, com reposição, constrói novos conjuntos de treinamento a partir do conjunto de treinamento inicial, fazendo com que as árvores sejam diferentes e, portanto, diminui a correlação entre elas. A correlação pequena tende a diminuir o erro de predição o que torna mais preciso a floresta aleatória (Oshiro, 2013).

A função de predição bagging é dada por:

$$\hat{h}(x) = \frac{1}{K}\sum_{k=1}^{K} h(x, \theta_k) \tag{3}$$

em que $K$ é no número total de árvores; $h(x, \theta_k)$ representa a resposta (previsão) de uma árvore $k$ para um vetor de entrada $x$, $\theta_k$ representa os parâmetros desta árvore; $\hat{h}(x)$ média de todas as previsões para amostras não vista em $x$.

Segundo Breiman (1996) a técnica de bagging utiliza 2/3 da amostra no conjunto de treinamento para testar o modelo e o restante 1/3 para validar, conhecidos como dados out-of-bag (OOB) o que reduz a variância, pois, o viés produzido é equivalente ao modelo original.

Liaw e Wiener (2002) dizem que as estimativas das incertezas das predições OOB para todas as árvores podem ser obtidas pelo cálculo do erro quadrado médio ($EQM_{OOB}$), dado por:

$$EQM_{OOB} = \frac{1}{K}\sum_{k=1}^{K} (h_k - \hat{h}_k^{\,OOB})^2 \tag{4}$$

em que $h_k$ é a resposta de uma árvore $k$ para um vetor de entrada $x$; $\hat{h}_k^{\,OOB}$ é a média de todas as previsões na técnica OOB (bagging); K o número de árvores.

O percentual da variabilidade explicada pelo modelo é representado por:

$$1 - \frac{\mathrm{EQM_{OOB}}}{\widehat{\sigma}^2{}_{\mathrm{h}}} \tag{5}$$

em que $\widehat{\sigma}^2{}_{\mathrm{h}}$ é a variância total das árvores preditoras.

O grau de importância de cada variável para predizer a variabilidade espacial da variável resposta no Random Forest é determinado pelo coeficiente do erro quadrado médio

$EQM_{OOB}$, quando se permutam aleatoriamente os n valores de uma variável, ao manter-se fixo os valores das demais variáveis (Liaw e Wiener, 2002).

Primeiramente, de forma univariada, foi analisada a variabilidade espacial dos atributos do solo e da produção média das Castanheiras-da-amazônia pelo método de interpolação Fixed Rank Kriging (FRK). Posteriormente, em um processo de empiplamento de rasters foi analisada a variabilidade espacial da produção média das Castanheiras-da-amazônia e o grau de importância de cada atributo do solo para predizer essa variabilidade pelo método de interpolação multivariado, implementado por esse estudo, Random Forest Kriging (RFK). Esse método consiste em uma adaptação conputacional do método espacial dependente univariado de krigagem ordinária com o método independente, não espacial e multivariado Random Forest para regressão.

Para avaliação do desempenho do novo método será realizada a comparação entre o interpolador híbrido RFK com o interpolador FRK via estimativa do erro quadrático médio (EQM) e do coeficiente de determinação ($R^2$) para a variável produção média das Castanheiras-da-amazônia.

Na interpolação por FRK foi adotado, para a construção do modelo SRE e da unidade básica de área, a escolha padrão/automática do pacote FRK do software R para cada variável a ser interpolada.

No Random Forest a escolha adotada para definicão dos principais parâmetros recomendados por Liaw e Wiener (2002) foi a padrão/automática do pacote randomForest do software R que definiu: mtry= 1/3 do número total de covariáveis (número de covariáveis utilizadas em cada árvore), ntree=500 (número de árvores construídas pelo algoritmo) e o nodesize = 9 (número mínimo de observações em cada nó terminal).

Toda a parte inovadora para adaptação dos dois modelos na metodologia proposta foi produzida por meio do software livre R (R Development Core Team, 2017), em que as análises geoestatísticas foram realizadas através dos pacotes: Random Forest, desenvolvido por Breiman et al. (1984); FRK, desenvolvido por Cressie (1993); gstat, desenvolvido por Pebesma et al. (1998); INLA, desenvolvido por Rue et al. (2009); splancs, desenvolvido por Bivand et al. (2017); geoR, desenvolvido por Ribeiro Júnior e Diggle (2001).

## 3. RESULTADOS E DISCUSSÃO

Buscando atingir os objetivos propostos, inicialmente será verificado o comportamento espacial dos dados em estudo por meio dos mapas de predição por interpolação via FRK para cada atributo do solo e para a produção média das castanheiras. Posteriormente, utilizando o interpolador implementado RFK, será apresentado o mapa de predição para a produção média das castanheiras e o grau de importância de cada atributo do solo para o resultado obtido no mapa em questão. Para verificar o desempenho desses interpolares serão aferidos o EQM e o $R^2$.

Como os interpoladores utilizados por esse trabalho não se prendem a teoria das semivariâncias serão consideradas para a construção dos mapas de krigagem todas variáveis amostradas: potencial hidrogeniônico (pH); carbono (C)(g/kg); matéria orgânica (mo)(g/kg); nitrogénio (N)(g/kg); fósforo (P)(mg/$dm^3$); potássio (K)(mg/$dm^3$); sódio (Na)(mg/$dm^3$); cálcio (Ca)(cmolc/$dm^3$); magnésio (Mg)(cmolc/$dm^3$); alumínio (AL)(cmolc/dm3); Acidez Potencial (H+AL)(cmolc/$dm^3$); soma de base (SB)(cmolc/$dm^3$); trocas de cátions efetiva (t)(cmolc/$dm^3$); trocas de cátion a pH 7 (Tc)(cmolc/$dm^3$); saturação por base (V)(%); saturação por alumínio (m)(%); ferro (Fe)(mg/$dm^3$); zinco (Zn)(mg/$dm^3$); manganês (Mn)(mg/$dm^3$); cobre (Cu)(mg/$dm^3$); selênio (Sel)(µg/kg); areia grossa (AreiaG)(g/kg); areia fina (AreiaF)(g/kg); areia total (AareiaT)(g/kg); silte (g/kg); argila (g/kg); produção (UA).

A seguir será apresentada uma breve discussão, com base na literatura, sobre a importância de cada atributo no solo para o crescimento e produção da planta.

### 3.1 Interpolação geoestatística dos atributos do solo e da produção média das Castanheiras-da-amazônia considerando-se o método por Fixed Rank Kriging (FRK)

Foi utilizada à função FRK do software estatístico R para gerar, automaticamente e fundamentados nos dados, as unidades bases de área e as funções de base necessárias para a contrução do modelo SRE e para a execução da interpolação.

Os mapas de predição por FRK para os atributos do solo estão representados pelas Figuras 4–8.

A absorção dos nutrientes pelas plantas é afetada pelo pH do solo (Embrapa, 2002). Os locais amostrados nos mapas pH e H+Al (Figuras 4a; 4b), que apresentaram maiores valores médios para produção, indicaram um solo com classificação de acidez média a boa

($5{,}0 < pH < 5{,}6$) e com baixo valor de acidez potencial ($H + AL < 2{,}5 cmolc/dm^3$), segundo a classificação química e agronômica encontrado em Freire et al. (2000).

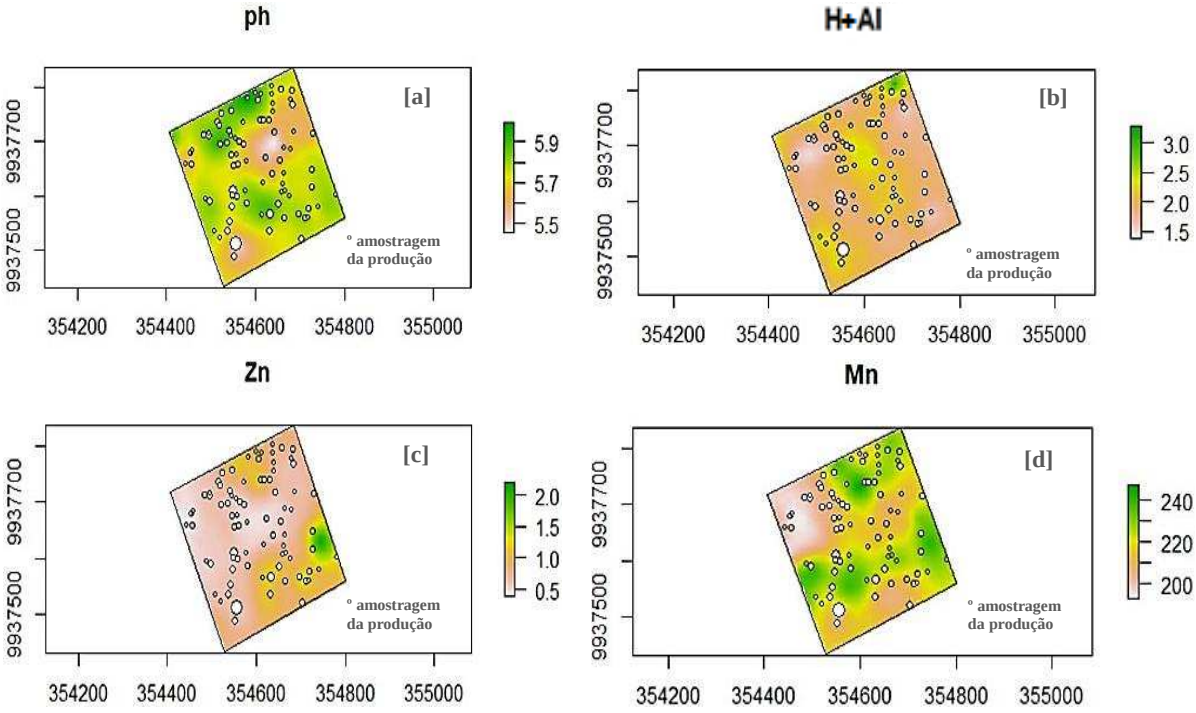


Figura 4: Mapas da distribuição espacial por interpolação FRK para os atributos do solo: [a] potencial hidrogeniônico (pH); [b] acidez potencial (H+AL)(cmolc/dm3); [c] zinco (Zn)(mg/dm3); [d] manganês (Mn)(mg/dm3) com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

O que corrobora para os resultados encontrados nos estudos de Locatelli et al. (2003) e Nicolodi et al. (2008) ao mostrarem que solos com esse tipo de classificação apresentam a maior disponibilidade de micronutrientes para as Castanheiras-da-amazônia.

Solos muito ácidos ($pH < 5{,}0$) podem acarretar vários prejuízos à produção da maioria das culturas, pois o alumínio torna-se tóxico na solução do solo (Kochian et al., 2005; Costa et al., 2017), e pH acima do ideal ($pH > 6{,}0$), induz à deficiência de micronutrientes, sendo os principais: zinco ($Zn < 0{,}9 mg/dm^3$) e manganês ($Mn < 5 mg/dm^3$), como mencionado nos trabalhos de Fageria (2000), Filho (2002) e Nicolodi et al. (2008). Fato que não foi observado nos respectivos mapas da variabilidade espacial do zinco (Zn) e do manganês (Mn) (Figuras 4c; 4d).

A toxicidade do alumínio no solo ($Al > 1 \text{cmolc/dm}^3$) está associada à inibição no crescimento das raízes das plantas em profundidade, o que reduz a absorção de água e nutrientes tais como fósforo: ($P < 8 mg/dm^3$), cálcio ($Ca < 1{,}2\ cmolc/dm^3$) e magnésio ($Mg < 0{,}46\ cmolc/dm^3$) (Filho, 2002; Amorim e Batalha, 2007). Apesar da concentração de alumínio no solo ser inferior a 0,5cmolc/dm3 (Figura 5a), o que contribui para um solo não

tóxico, a deficiência desses nutrientes citados acima exceto para o magnésio, está sendo observada no mapa P e, em alguns pontos, no mapa da Figura 5c. Nesse caso, é recomendado por Costa (2017) usar técnicas de correção para tratar essas deficiências.

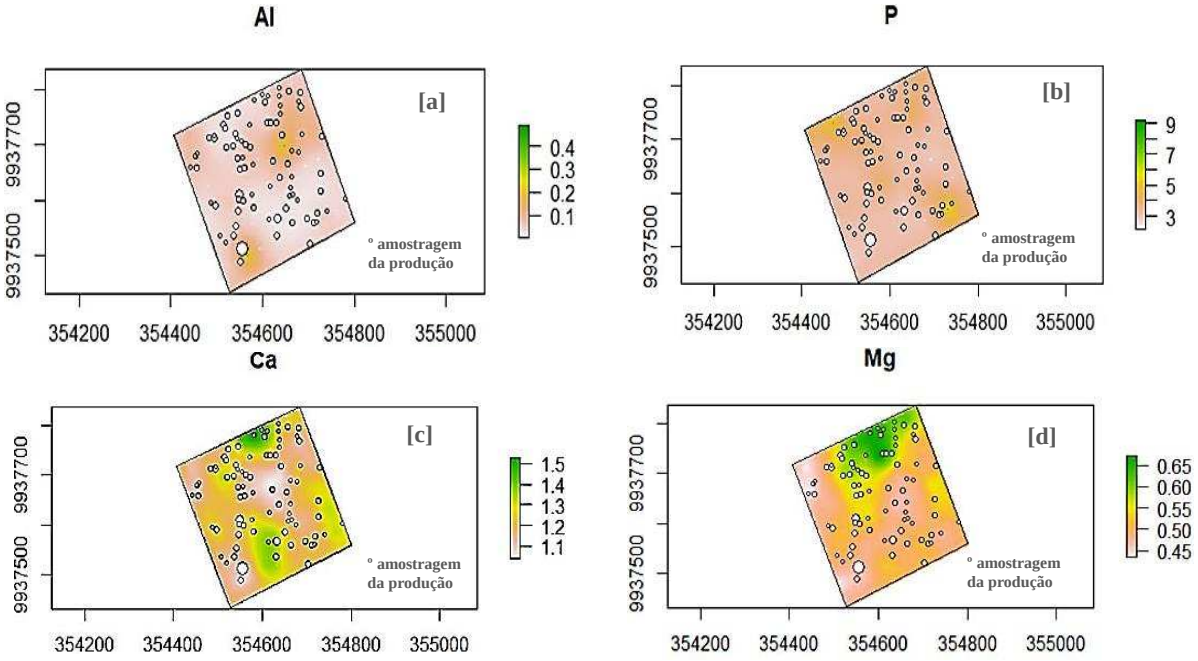


Figura 5: Mapas da distribuição espacial por interpolação FRK para os atributos do solo: [a] alumínio (Al)(cmolc/dm3); [b] fósforo (P)(mg/dm3); [c] cálcio (Ca)(cmolc/dm3); [d] magnésio (Mg)(cmolc/dm3) com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

Além disso, percebe-se na Figura 5a que as maiores produções em média das castanheiras ocorreram em regiões com baixíssimas taxas de alumínio ($Al < 0{,}2$ cmolc/ $dm^3$), resultados já mencionados por Anghinoni e Nicolodi (2004) e Nicolodi et al. (2008) em seus estudos.

Ao comparar os resultados obtidos no mapa da variabilidade espacial da saturação por alumínio ($m$) com os estudos de Lima (2004), pode-se dizer que o solo em questão apresenta um baixo valor de saturação ($m < 20\%$) (Figura 6a), o que confirma a não toxidade do solo. Resultado já esperado pelo fato de o alumínio (Al) encontrar-se em proporção baixíssima na área do estudo (Figura 5a).

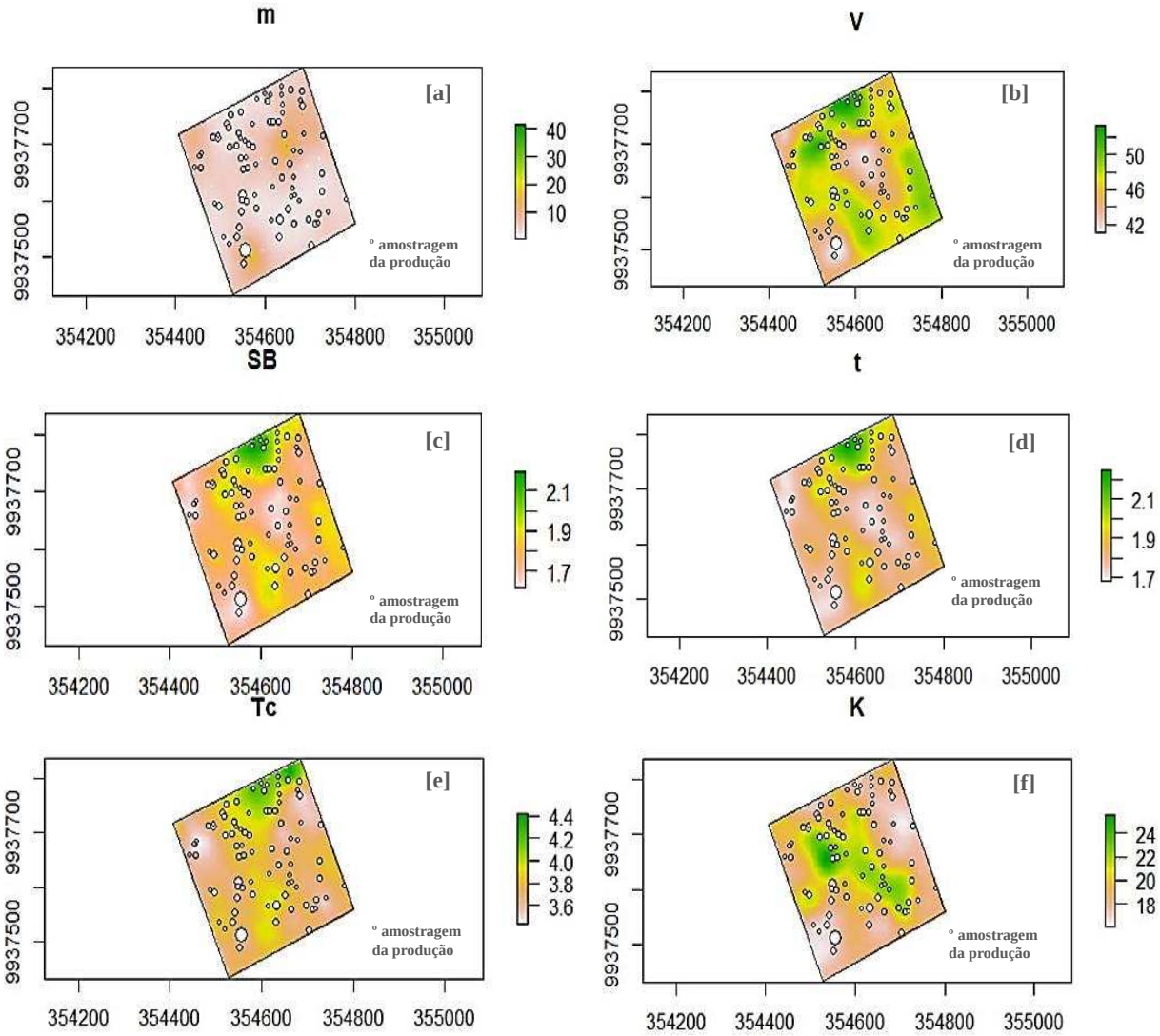


Figura 6: Mapas da distribuição espacial por interpolação FRK para os atributos do solo: [a] saturação por alumínio (m)(%); [b] saturação por base (V)(%); [c] soma de base (SB)(cmolc/dm3); [d] trocas de cátions efetiva (t)(cmolc/dm3); [e] trocas de cátion a pH 7 (Tc)(cmolc/dm3), [f] potássio (k)(mg/dm3), com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

O índice geral da fertilidade do solo representada pela saturação por base ($V\%$), sendo a base considerada a soma dos nutrientes como cálcio, magnésio, potássio e sódio, indica que o solo em estudo apresenta de médio a baixa fertilidade mapa ($V < 50\%$) em alguns pontos do mapa (Figura 6b), ocorrência que pode ser confirmada também, pelos baixos valores da soma de base ($SB < 2{,}2\ cmolc/dm^3$) (Figura 6c) (Corrêa, 2013). Esse tipo de situação é muito comum em solos arenosos e, se faz necessário, o uso de técnicas de manejo adequadas para o enriquecimento desse solo, pela deficiência de alguns nutrientes que compõem essa base (Alvarez, et al., 1999, Ronquim 2010).

Ronquim (2010) explica que se a maior parte da capacidade de troca de cátions do solo (CTC – capacidade de armazenar nutrientes) estiver composta por cátions essenciais tais

como cálcio, magnésio e potássio, em seus valores considerados de médio-bom, será favorável para a nutrição, equilíbrio estrutural e enzimático das plantas. No entanto, de acordo com Lopes (2004), o valor extremamente baixo da CTC efetiva (capacidade do solo em reter cátions próxima ao valor do pH natural) ($t < 2{,}00 cmolc/dm^3$) e da CTC potencial (capacidade do solo em reter cátions próxima ao valor do pH 7) ($Tc < 4{,}30\ cmolc/dm^3$), observada em maior parte da área nos mapas (Figuras 6d; 6e), reflete que o solo em estudo, nas condições naturais ácidas, apresenta pequena capacidade de reter os cátions por ação provocada por intemperismo e lixiviação. Fato que é testificado pelos baixos valores de referência desses cátions observados em vários pontos nos mapas da variabilidade espacial Ca ($Ca < 1{,}2\ cmolc/dm^3$), Mg ($Mg <\ 0{,}46\ cmolc/dm^3$) e K ($K < 40 mg/dm^3$) (Figuras 5c; 5d ; 6f). Portanto, nessas áreas de baixos valores, é recomendado por Ronquim (2010) fazer adubações e calagens em pequenas quantidades, de forma fracionada, para evitar grandes perdas por lixiviação.

Esses problemas de armazenamento de nutrientes encontrados no solo também foram verificados nos estudos de Falesi et al. (1967), Valente et al. (1997) e Locatelli et al. (2003), que definiram o solo Argissolo Vermelho-amarelo, como é o caso do solo em estudo, como: solos profundos, altamente intemperizados, ácidos de baixa fertilização natural e, em certos casos, apresentam alta saturação por alumínio.

O solo em estudo, Argissolo Vermelho-amarelo, apresenta uma classificação textural moderadamente grossa, denominado franco arenoso pelo triângulo simplificado da Embrapa (Embrapa, 2006), composto em grande parte por areia (72%) e, em menor parte por argila (18%) e silte (10%), valores que podem ser verificados pelas quantidades desses atributos nos mapas correpondentes (Figuras 7b; 7e; 7f). Os valores considerados pelo Guia Prático para Interpretação de Resultados de Análises de Solo (Embrapa, 2015) como medianos para Argila ($150 - 350 g/kg$) (Figura 7b) favorecem a um solo mais arenoso que está diretamente relacionado com baixos valores de zinco ($Zn < 0{,}9\ mg/dm^3$), fósforo ($P < 12\ mg/dm^3$) e matéria orgânica ($mo < 20 g/kg$), o que podem ser verificados nas Figuras 4c, 5b e 7a. Logo, pode-se comprovar que a deficiência do fósforo no solo em estudo é devido à textura arenosa e não pela toxicidade por Alumínio como observado em outros autores tais como Filho (2002), Silva et al. (2002b), Amorim e Batalha (2007).

Outro ponto importante se dá pelo fato de que o solo em estudo apresenta em sua textura uma quantidade maior de areia grossa do que fina, acarretando em uma maior

porosidade que influencia na permeabilidade da água, que pode resultar na remoção desses nutrientes essenciais para a planta (Figura 7c; 7d) (Ribeiro et al., 2007).

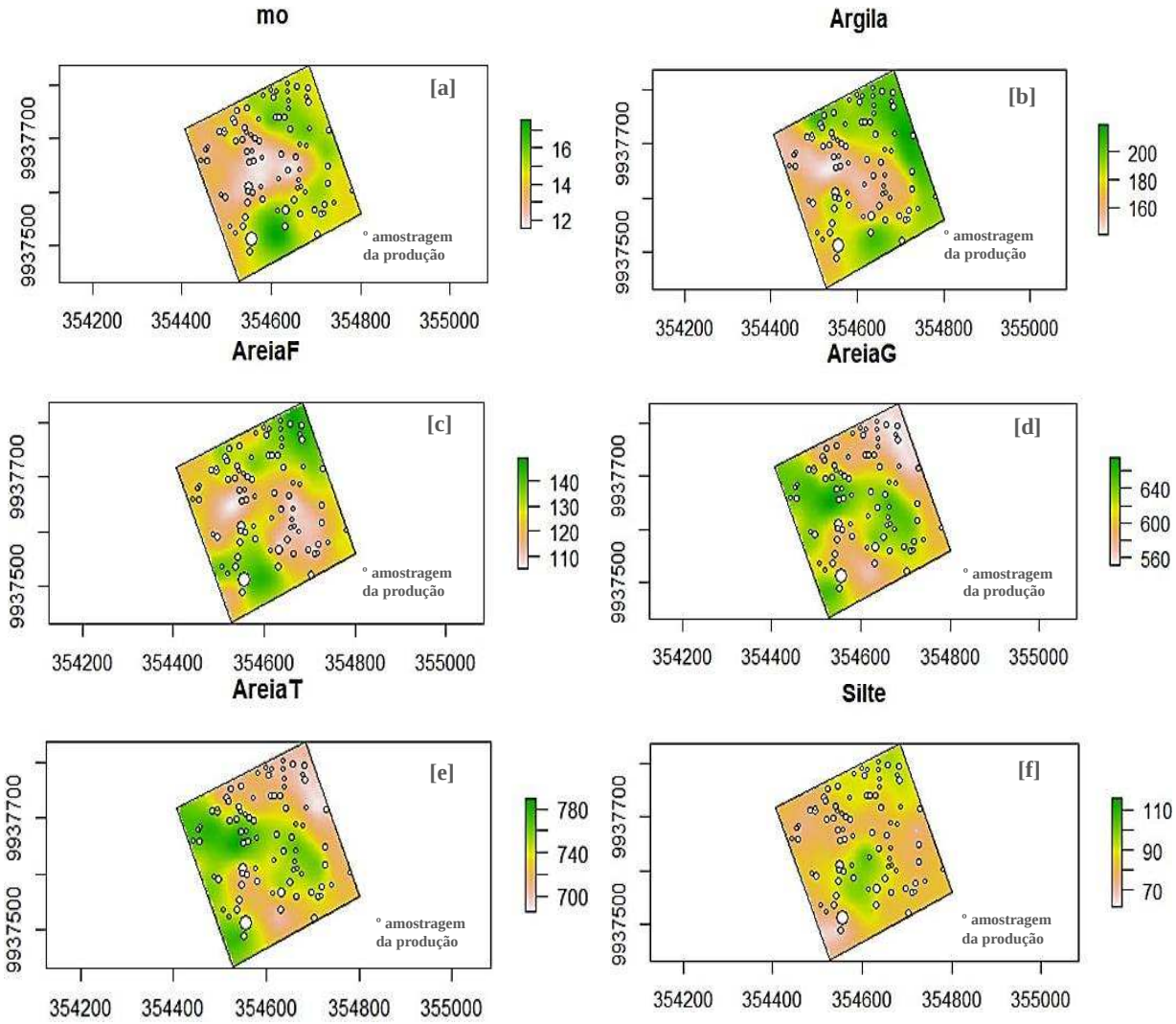


Figura 7: Mapas da distribuição espacial por interpolação FRK para os atributos do solo: [a] matéria orgânica (mo)(g/kg); [b] argila (Argila)(g/kg); [c] areia fina (AreiaF)(g/kg); [d] areia grossa (AreiaG)(g/kg); [e] areia total (AreiaT)(g/kg); [f] silte (Silte)(g/kg) com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

Teores baixos de matéria orgânica ($mo < 20g/kg$) presente no solo em estudo (Figura 7a) estão favorecendo para os baixos valores de carbono ($C < 11{,}6g/kg$) e nitrogênio ($N < 1{,}0\ g/kg$) (Figuras 8a; 8b), o que corrobora para os estudos de Centeno et al. (2017) e Costa Junior (2008) que definem a matéria orgânica como uma reserva desses nutrientes. Entretanto, percebe-se que a textura e a deficiência de matéria orgânica no solo em questão não tem afetado a absorção do cobre ($Cu > 1{,}3mg/dm^3$) e do ferro ($Fe > 31\ mg/dm^3$) (Figuras 8c; 8d), o que é favorável, pois exercem um papel importante na fotossíntese, respiração, redução e fixação de nitrogênio (Marchetti e Barp, 2015).

Vale a pena ressaltar que o intemperismo, muito presente no solo arenoso, favorece a presença do ferro, o que tem justificado os altos valores no mapa Fe.

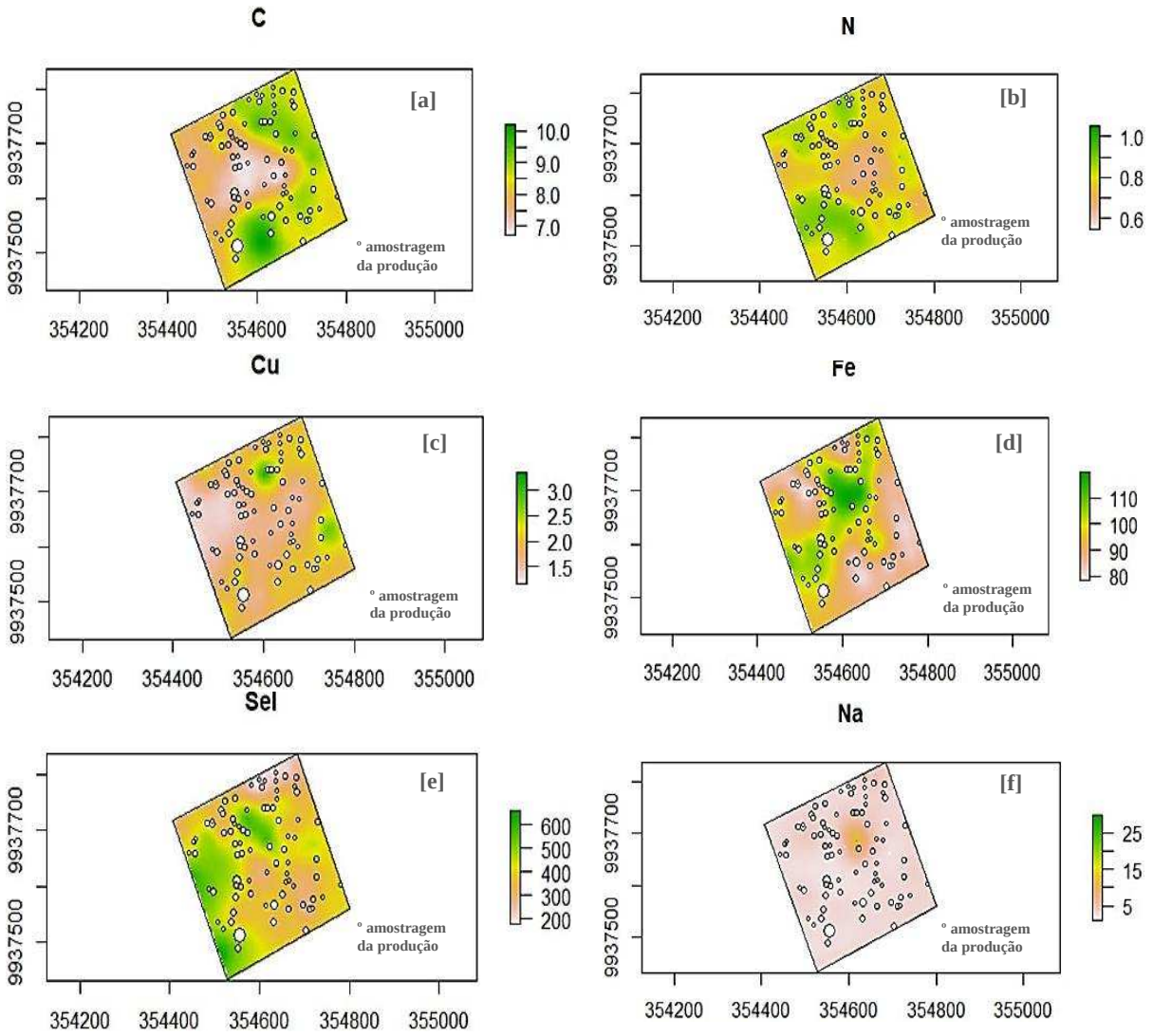


Figura 8: Mapas da distribuição espacial por interpolação FRK para os atributos do solo: [a] carbono (C)(g/kg); [b] Nitrogênio (N)(g/kg); [c] cobre (Cu)(mg/dm3); [d] ferro (Fe)(mg/dm3); [e] selênio (Sel)(µg/kgcom); [f] sódio (Na)(mg/dm3); com amostragem da produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

De acordo com Camargo et al. (2002), Correa (2013) e Gama (2018) a deficiência de carbono, nitrogênio, potássio, ferro e fósforo são limitantes para a cultura da Castanheira-da-amazônia. Esses nutrientes exercem um papel fundamental na melhoria da fertilidade, no aumento da produtividade da planta, está diretamente relacionada com as características químicas, físicas e biológicas do solo e com o aumento da capacidade de acúmulo de água e da fotossíntese. Gama (2018) mostrou que a aplicação de nitrogênio e, para alguns casos associado ao biocarvão, contribuiu para o aumento desses nutrientes no solo e na planta. No entanto, Hirel et al. (2011) e Dempster (2013) sugeriram um cuidado especial no manejo no

nitrogênio, pelos riscos de contaminação do lençol freático, por ser lixiviado facilmente nos solos.

Apesar da desvantagem que o baixo teor da matéria orgânica está trazendo para o solo em questão, vale ressaltar que Junior (2016) em seu trabalho, destacou que esse valor baixo, assim como, baixos teores de argila podem ajudar na absorção do selênio pelas plantas.

O manganês e o selênio são importantes, especificamente, para a nutrição e o crescimento das castanheiras-da-amazônia, o que favorece o cultivo dessa espécie no solo em estudo, por apresentar altos valores para esses nutrientes ($Mn > 12\, mg/dm^3$) e ($Sel > 200\mu g/kg$) (Figura 4d; 8e) (Freitas et al., 2008; Faria, 2009).

Apesar do sódio não ser um nutriente, a presença pode sanilizar o solo e acarretar vários problemas, tais como: diminuição da disponibilidade de nutrientes, diminuição do potencial da água no solo, dispersão de partículas, encrostamento e compactação do solo, aumento da resistência à penetração de raízes, toxicidade de íons específicos (inibição competitiva), tornando-o infértil. Como o solo em questão apresenta baixíssimos valores de sódio ($Na < 10 mg/dm^3$) (Figura 8f), conclui-se que esse fato se dá pela textura arenosa do solo (Franco 2013), entretanto, essa textura tem contribuído para os outros problemas encontrados no estudo sobre a deficiência e armazenamento dos nutrientes.

No geral, solos mais arenosos como identificado no estudo requer um cuidado especial na adubação ao focar em mais parcelamentos e realizar manejos que visem o aumento da CTC do solo, como por exemplo, aplicar matéria orgânica tais como esterco, composto orgânico, palhas de braquiária e casca de café, o que irá contribuir para solucionar esses problemas encontrados no solo.

O mapa de predição por FRK para a produção média das Castanheiras-da-amazônia está representado pela Figura 9. Esse mapa descreve o comportamento da variabilidade espacial da produção das castanheiras no solo estudado, assim como, os pontos de amostragem plotados.

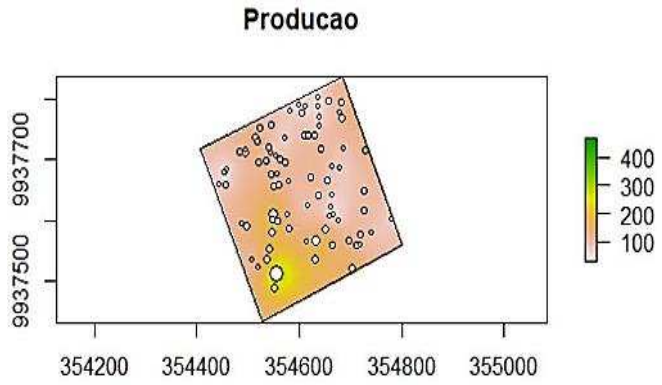


Figura 9: Mapa da distribuição espacial por interpolação FRK para produção média das castanheiras nativas, no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil.

No mapa da produção (Figura 9), a heterogeneidade indica que em média os valores mais altos da produção das Castanheiras-da-amazônia encontram-se mais ao sudoeste da região do estudo. Resultados que corrobora para os estudos de Fernandes e Alencar (1993), Muller (1995) e Espírito-Santo et al. (2005), que mostraram as maiores concentrações para a produção dessas castanhas em solos com textura mais arenosa (Figura 7e).

A predominância de selênio nessa região (Figura 8e) confirma o valor nutricional encontrado no fruto da castanheira que é o nutriente mais abundante nesse tipo de espécie.

> "É comumente divulgado que o consumo de uma castanha por dia ajuda a combater doenças cardiovasculares, diabetes do tipo 2, câncer e obesidade. Um estudo da Universidade de Otago, na Nova Zelândia, comprovou que a ingestão diária de duas castanhas eleva em cerca de 65% o teor de selênio no sangue, mas alertou sobre os problemas com a toxidade desse mineral. Outro estudo da USP, que ganhou o prêmio Jovem Cientista do CNPq em 2015, também comprovou efeitos benéficos do selênio, nesse caso, contra o mal de Alzheimer. De uma maneira geral, os nutricionistas brasileiros recomendam a ingestão de uma única castanha por dia, mas também alertam que quantidades elevadas do mineral podem provocar intoxicação por selênio, ou selenose, causadora de perda de cabelo, fadiga, fraqueza das unhas, lesões na pele e problemas gastrointestinais". Embrapa (2016)

Com a interpolação FRK foi possível estudar, de forma univariada, o comportamento da variabilidade espacial de cada variável em estudo. Entretanto, um dos desafios encontrados está exatamente em obter um resultado que possa dizer o quanto o comportamento da variabilidade espacial de cada atributo do solo está influenciando, diretamente, nos resultados

da variabilidade espacial da produção, de forma a predizer essa variabilidade o mais perto da realidade possível e, assim, resolver a carência nutricional do solo em pontos específicos para obter um melhor desenvolvimento dessa espécie de maneira rápida, eficiente e sem prejuízos.

## 3.2. Interpolação geoestatística para produção média das Castanheiras-da-amazônia considerando-se o método multivariado por Random Forest Kriging (RFK)

Em busca de um preditor que apresente um melhor desempenho e consiga resolver o desafio acima citado, o presente artigo, propôs o interpolador híbrido Random Forest Kriging (RFK) dado pela fusão computacional do FRK com o Random Forest para regressão.

A seguir será apresentada uma breve discussão sobre os resultados obtidos para a produção média das castanheiras pelo método de interpolação RFK e FRK. Também, será apresentado o grau de importância de cada atributo do solo para o estudo da variabilidade espacial da produção das castanheiras.

O mapa de predição por RFK e FRK para a produção média das Castanheiras-da-amazônia está representado pela Figura 10.

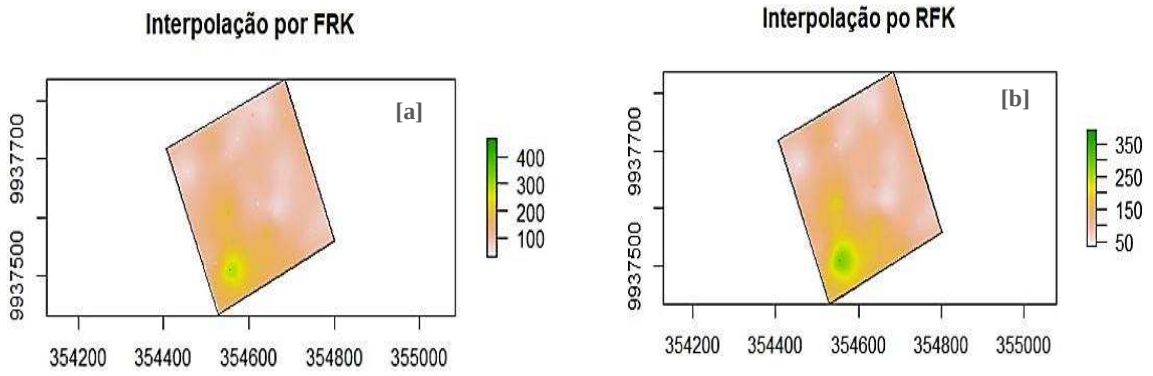


Figura 10: Interpolação da produção média das Castanheiras-da-amazônia, na região em estudo no sul da reserva extrativista do Rio Cajari, Zona Rural do município de Laranjal do Jari, Amapá, Brasil: [a] pelo híbrido Random Forest Kriging (RFK); [b] pelo Fixed Rank Kriging (FRK).

Pela Figura 10, vale ressaltar alguns pontos importantes tais como: ao observar a legenda dos valores interpolados percebe-se uma diferença. Essa diferença se dá pela amostragem apresentar uma única castanheira com a produção em torno de 650UA (unidade de medida) que é considerado um outlier em relação aos demais valores. Na presença de outliers a krigagem por ser um tipo de média sofre uma tendência nas interpolações. O método FRK resultou em uma pequena suavização na predição da produção (Figura 10a), mas

ao ser implementado com o Random Forest (Figura 10b), a medida discrepante foi ignorada totalmente e as demais regiões tiveram um destaque maior, contribuindo assim, para maiores detalhes sobre a produção média nessa região.

Tratar variáveis discrepantes tem sido um tema relevante na literatura e abordado por vários autores, tais como Hongxing et al. (2001) para dados espaciais irregularmente distribuídos, Qiao et al. (2013) para dados provenientes de satélites, Appice et al. (2014) para dados de fluxo na mineração, Santos (2016) para dados de altimetria, entre outros, o que confirma os resultados obtidos acima.

O grau de importância de cada atributo do solo para predizer a variabilidade espacial da produção média das castanheiras está representado pela Figura 11.

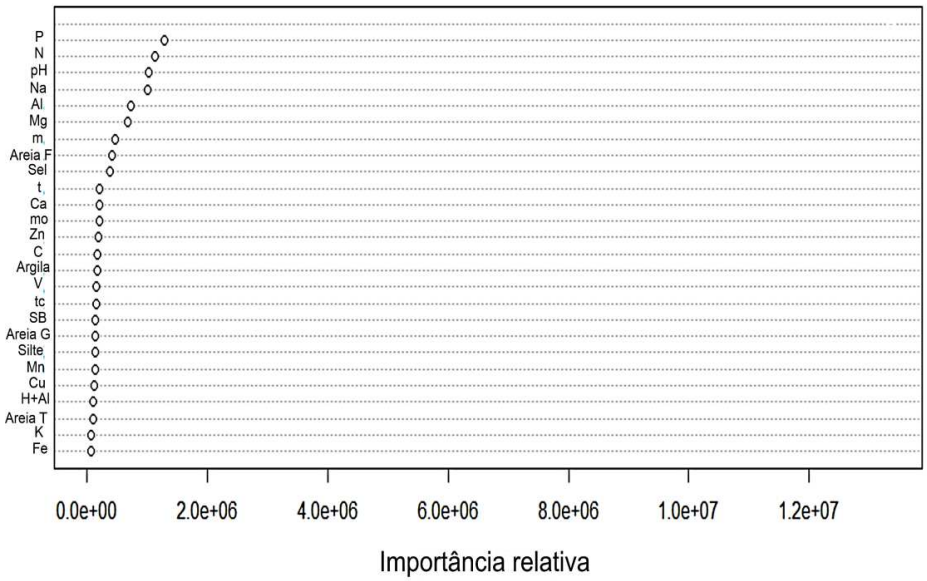


Figura 11: Importância relativa de cada atributo do solo para predizer a variabilidade espacial da variável produção pelo método RFK.

As variáveis mais promissoras, no método RFK, para predizer a variabilidade da produção na região em estudo foram: fósforo, pH, nitrogênio, sódio, alumínio, magnésio, areia fina, saturação por alumínio e selênio (Figura 11).

Os resultados obtidos pela Figura 11 corroboraram para a discussão descrita no presente trabalho sobre a importância desses nutrientes. Apesar de o fósforo e o nitrogênio apresentarem uma deficiência no solo em questão, esses resultados confirmaram os trabalhos dos autores tais como Grant et al. (2001), Meneghin et al. (2008), Santos et al (2008), Pereira (2009), Marchett et.al (2015) entre outros, que mostrou o quanto esses nutrientes são indispensáveis para o crescimento e produção das plantas, por desempenharem um papel

fundamental na fotossíntese, na divisão celular, no transporte de assimilados e na carga genética.

Os valores encontrados para o pH, o alumínio e o sódio estão favoráveis no solo para absorção de nutrientes pela planta, favorecendo o crescimento das raízes profundas, a síntese proteica, a formação de clorofila e o carregamento do floema (Favarin et al., 2013).

Estudos conduzidos por Klein e Libardi (2002), Cássaro et al. (2011), Jonez Fidalski et al. (2013) confirmam que a presença de areia fina em solos arenosos indica menores diâmetros de poros o que contribui para uma menor remoção de partícula e melhor absorção de nutrientes para as plantas. Resultados que podem ser confirmados no mapa Areia Fina (Figura 7c), em que as maiores produções em média das castanheiras foram verificadas nos pontos com maiores proporções de área fina.

### 3.3. Comparação das incertezas de predição dos interpolares: Fixed Rank Kriging (FRK) e o híbrido Random Forest Kriging (RFK)

O erro quadrado médio (EQM) e o coeficiente de determinação ($R^2$) foram estabelecidos nesse estudo para comparar à incerteza de predição dos interpoladores (Tabela 1).

Tabela 1: Erro quadrático médio e coeficiente de determinação para os interpoladores: FRK e RFK

| | FRK | RFK |
|---|---|---|
| EQM | 94,03 | 2,56 |
| $R^2$ | 97,17 | 99,72 |

Nota-se na Tabela 1 que o interpolador RFK apresentou um valor extremamente menor para o EQM, o que indica um melhor desempenho ao ser comparado com o FRK, fato que deve ser explicado pela forma que conseguiu suavizar o ponto discrepante presente no conjunto de dados, promovendo um destaque maior na predição das demais regiões. Resultado que corrobora para o valor do $R^2$, indicando um ajuste perfeito de quase 100%, mostrando que o preditor implementado está conseguindo explicar melhor a variabilidade espacial da produção.

Portanto, a proposta metodológica sobre implementar o algoritmo RF a krigagem FRK, adotando o erro quadrático médio e o coeficiente de determinação do modelo entre os valores observados e preditos, mostrou-se satisfatório e resultou em predições mais acuradas da krigagem. Assim, fica evidente que tratar a exclusão de variáveis importantes e os outliers presente nos dados associado à técnica bagging gerou uma suavização no mapa de predição apresentando-se uma distribuição espacial da variável de interesse mais real.

## 4. CONCLUSÕES

Os resultados mostraram que a deficiência de alguns nutrientes como: fósforo, nitrogênio, matéria orgânica e, outros, é devido à decorrência da textura arenosa presente no solo em estudo, o que contribui para uma forte necessidade de adubação focado em mais parcelamentos, por causa da lixiviação, e de técnicas de correção para melhorar a absorção de nutrientes pelas plantas. No entanto, a presença do selênio em abundância corrobora para que o fruto das Castanheiras-da-amazônia da região em estudo, seja favorável para a saúde humana, no combate a doenças cardiovasculares, diabetes do tipo 2, câncer e obesidade.

A robustez da metodologia proposta por esse trabalho resultou em predições mais acuradas da krigagem, fazendo com que o RFK seja um preditor extremamente apropriado para a predição de uma variável de interesse.

Sabe-se que se as informações relevantes de um estudo forem omitidas, a variância pode vir a ser afetada e o viés nos preditores se fazerem presente, mesmo que os preditores não sejam viesados, o que traria, consequentemente, predições equivocadas a partir deste modelo (Penna e Linhares, 2015).

O RFK conseguiu não só modelar grandes conjuntos de dados, como também, tratar o descarte de variáveis e a presença de outliers, suavizando com maior eficiência o mapa de predição da produção, permitindo assim, um destaque para as demais regiões. E, consequentemente, ao associar-se a metodologia do Random Forest, resultou-se no menor coeficiente do erro quadrático médio (EQM=2,56), fazendo com que esse preditor, dentre os modelos propostos por esse trabalho, seja o mais apropriado para predizer a variabilidade espacial da produção em função da variabilidade espacial dos nutrientes presentes no solo, por apresentar-se o mais acurado. O que pode ser confirmado pela melhora no valor do coeficiente de determinação ($R^2$ =99,72) indicando um ajuste de modelo praticamente perfeito para explicar essa variabilidade.

Além disso, o método de empilhamento de rasters, presente no RFK, corroborou satisfatoriamente na discriminação do grau de importância de cada atributo do solo simultaneamente, o que irá contribuir para subsídios mais confiáveis no que se refere ao manejo florestal, à manutenção e a ampliação da produtividade das castanheiras-da-amazônia na região do estudo.

Em geral, a implementação do Random Forest no FRK melhorou o desempenho do interpolador resultando em uma melhor acurácia e precisão.

## REFERÊNCIAS BIBLIOGRÁFICAS


ALVAREZ V. H.; DIAS, L. E.; RIBEIRO, A. C.; SOUZA, R. B. **Uso de gesso agrícola.** In: RIBEIRO, A. C.; GUIMARÃES, P. T.G.; ALVAREZ, V. H. eds. **Recomendações para o uso de corretivos e fertilizantes em Minas Gerais: 5a aproximação**. Viçosa, MG, Comissão de Fertilidade do Solo do Estado de Minas Gerais, p. 67-78, 1999.

ALVARES, C. A; STAPE, J. L; SENTELHAS, P. C; GOLÇALVES, M. J. L**;** SPAROVEK, G. "Köppen's climate classification map for Brazil," **Meteorologische Zeitschrift**, v.22(6), p. 711–728, 2013.

AMORIM, P. K.; BATALHA, M. A. Soil vegetation relationships in hyperseasonal cerrado, seasonal cerrado, and wet grassland in Emas National Park (Central Brazil). **Acta Oecologica**, v.32(3), p.319-327, 2007. http://dx.doi.org/10.1016/j. actao.2007.06.003.

ANGHINONI, I.; NICOLODI, M. **Estratégias de calagem no sistema plantio direto. In: Reunião Brasileira de Fertilidade do Solo e Nutrição de Plantas**, FERTBIO, 27, Lages, 2004. Anais. Lages, Sociedade Brasileira de Ciência do Solo, 2004. CD-ROM.

APPICE, A., GUCCIONE, P., MALERBA, D., & CIAMPI, A. Dealing. with temporal and spatial correlations to classify outliers in geophysical data streams. **Information Science**, v.285, p.162-180, 2014.

BIVAND, R.; ROWLINDSON, B.; DIGGLE, P.; PETRIS, G.; EGLEN, S. **Spatial and Space-Time Point Pattern Analysis**, 2001. Disponível em: https://rdrr.io/cran/splancs/

BREIMAN, L. Bagging predictors. **Machine learning**, v.24 (2), p.123-140, 1996. http://dx.doi.org/10.1023/A:1018054314350

BREIMAN, L. Random Forest. **Machine Learning**, v.45 (1), p.5-32, 2001. Disponível em: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>. Acesso em: 05 abr. 2018.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN. R.; STONE, C. **Classification and Regression Trees**. Taylor & Francis, 1984. 368p.

CÁSSARO, F. A. M.; BORKOWSKI, A. K.; PIRES, L. F.; COSTA, S. C.; ROSA, J. A. Characterization of a Brazilian clayey soil submitted to conventional and no-tillage management practices using pore size distribution analysis. **Soil & Tillage Research**, v.111, p.175-179, 2011.

CENTENO, L. N; GUEVARA, M. D. F; CECCONELLO, S. T; SOUSA, R. O. D; TIMM, L.C. Textura do solo: Conceitos e aplicações em solos arenosos. **Revista Brasileira de Engenharia e Sustentabilidade**, v.4(1), p.31-37, 2017.

CRESSIE, N. **Statistics for Spatial Data**. Revised edn. New York, 1993.

CRESSIE, Noel; JOHANNESSON, Gardar. Fixed rank kriging for very large spatial data sets. **Journal of the Royal Statistical Society**. v.70 (1), p.209-226, 2008.

CORRÊA, V. M. **Crescimento, aspectos nutricionais e fotossintéticos de plantas jovens de Bertholletia excelsa H.B. submetidas a diferentes tratamentos de fertilização**. Dissertação (Mestrado), Instituto Nacional de Pesquisas da Amazônia - INPA, Manaus, Amazonas. 66p. 2013.

COSTA JUNIOR, C. **Estoque de carbono e nitrogênio e agregação do solo sob diferentes sistemas de manejo agrícola no Cerrado, em Rio Verde (GO)**. Dissertação (Mestrado). Universidade de São Paulo - USP, 2008.

COSTA, M. G.; TONINI, H.; FILHO, P. M. Atributos do Solo Relacionados com a Produção da Castanheira-do-Brasil (Bertholletia excelsa). **Floresta e Ambiente**, v24, 2017. http://dx.doi.org/10.1590/2179-8087.004215 ISSN 2179-8087 (online).

DEMPSTER, D. N. 2013. **Biochar and the Soil Nitrogen Cycle: Unravelling the Interactions**. (Doctoral) Thesis. University of Western Australia. 193p, 213.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Solos. **Revista Embrapa Algodão Sistemas de Produção**, 3 ISSN 1678-8710, Versão Eletrônica, 2003. Disponível em:
https://www.ft.unicamp.br/~sandro/.../Solos%20-%20EMBRAPA%20ALGODÃO.doc

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. **Guia Prático para Interpretação de Resultados de Análises de Solo**. Documentos 206. Embrapa Tabuleiros Costeiros Aracaju, ISSN 1678-1953, 2015.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Solos. **Sistema Brasileiro de Classificação dos Solos**. 2.ed. Rio de Janeiro: Embrapa Solos, 2006. 306p.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. SILVA, R. **Quantidade de selênio nas castanhas-do-brasil varia de acordo com região**. Embrapa (2016)
https://www.embrapa.br/en/busca-de-noticias/-/noticia/11010983/quantidade-de-selenio-nas-castanhas-do-brasil-varia-de-acordo-com-regiao

FAGERIA, N. K. Níveis adequados e tóxicos de zinco na produção de arroz, feijão, milho, soja, e trigo em solo de cerrado. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.4(3), p.390-395, 2000.

FALESI, J.C.; VIEIRA, L.S.; SILVA, B.N.R. da; CRUZ, E. de S.; GUIMARAES, G. de A.; SILVA, R.P. da; LOPES, E. de C. **Solos da Estação Experimental de Porto Velho – T.F. Rondônia.** Belém: IPEAN, 1967. 99p. (IPEAN. Solos da Amazonia, 1).

FARIA, L. A. **Levantamento sobre selênio em solos e plantas do estado de São Paulo e sua aplicação em plantas forrageiras**. Dissertação (mestrado). Universidade de São Paulo. Pirassununga, 2009.

FAVARIN, J. L.; NETO, A. P.; TEZOTTO, T.; MARTINS, P. O.; TEIXEIRA, P. P. C. Correção do magnésio no solo é essencial ao cafeeiro. **Visão Agrícola**, n.12, 2013.

FERNANDES, N. P; ALENCAR, J. C. Desenvolvimento de árvores nativas em ensaios de espécies - Castanha-da-amazônia (Bertholletia excelsa H.B.K.), dez anos após o plantio. **Acta Amazônica**, v.23, n.2, p.191-198, 1993.

FIDALSKI, J.; TORMENA, C. A.; ALVES, S. J.; AULER, P. A. M. Influência das frações de areia na retenção e disponibilidade de água em solos das formações Caiuá e Paranavaí. **Revista Brasileira de Ciências do Solo**, v.37, p.613-621, 2013.

FILHO, M. P. B**. Calagem**. Embrapa, 2002. Disponível em: https:/www.agencia.cnptia.embrapa.br/Repositorio/CONTAG01_87_1311200215104.html

FRANCO, E. M. **Eficiência de equações empíricas utilizadas para determinar lâmina de lixiviação de sais e modelagem da distribuição do sódio no solo**. Tese (Doutorado). Universidade de São Paulo, Piracicaba, 2013.

FREIRE, F. M.; PITTA, G. V. E.; ALVES, V. M. C.; FRANÇA, G. E.; COELHO, A. M. **Fertilidade de Solos**. Embrapa, 2000. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/27337/1/Fertilidade-de-solos-Interpretacao.pdf

FREITAS, S. C; GONÇALVES, E. B; ANTONIASSI, R; FELBERG, I. OLIVEIRA, S. P. Meta-análise do teor de selênio em Castanha-da-amazônia. **Brazilian Journal of Food Technology**, v. 11, n. 1, p. 54-62, 2008.

GAMA, R. T. **Biocarvão e adubação nitrogenada influenciando o crescimento e o estado nutricional de mudas de castanheirasdo-brasil em um latossolo da amazônia central**. Dissertação (Mestrado). Instituto Nacional de Pesquisas da Amazônia - INPA, Manaus, Amazonas. 84p. 2018.

GRANT, C.A.; FLATEN, D. N.; TOMASIEWICZ, D. J.; SHEPPARD, S. C. A importância do fósforo no desenvolvimento inicial da planta. **Informações Agronômicas**, v.95, 2001.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. Springer Series in Statistics. Second Edition. California, 2008. 744p.

HIREL, B.; TÉTU, T.; LEA, P.J.; DUBOIS, F. Improving nitrogen use efficiency in crops for sustainable agriculture. **Sustainability**, v.3, p.1452-1485, 2011.

IBGE. Manual técnico da vegetação basileira. Coordenação de Recursos Naturais e Estudos Ambientais. **Série Manuais Técnicos em Geociências 1, 2**. Ed. revista e ampliada. IBGE, Rio de Janeiro, 2012

IBGE. Manual técnico de pedologia. Coordenação de Recursos Naturais e Estudos Ambientais. **Série Manuais Técnicos em Geociências 1, 3**. ed. IBGE, Rio de Janeiro, 2015.

GUERREIRO, Q. L. M; JÚNIOR, R. C. O; SANTOS, G. R; RUIVO, M. L. P; BELDINI, T. P; CARVALHO, E. J. M; SILVA, K. E; GUEDES, M. C; SANTOS, P. R. B. Spatial variability of soil physical and chemical aspects in a Brazil nut tree stand in the Brazilian Amazon. **African Journal of Agricultural Research**. v.12(4), p.237-250, 2017.

GOMES, M. A. F; FILIZOLA, H. F. **Indicadores físicos e químicos de qualidade de solo de interesse agrícola**. EMBRAPA Meio Ambiente. Jaguariúna, 2006.

HONGXING, L., KENNETCH, C.J., & MORTON, E.O. Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. **International Journal Geographical Information Science,** v.15, p.721-741, 2001.

JAMES, G; WITTEN, D; HASTIE, T; TIBSHIRANI, R. **An Introduction to Statistical Learning with applications in R. Springer New York Heidelberg Dordrecht London**, 2013. 426p.

JUNIOR, E. C. DA S. **Selênio na Castanha-da-amazônia (Bertholletia excelsa) em solos da região amazônica brasileira**. Dissertação (Mestrado). Universidade Federal de Lavras. Lavras, 2016.

JUNIOR, W. C.; CALDERANO FILHO, CHAGAS, C. da S.; B.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa Agropecuária Brasileira**.v.51 (9), p.1428-1437, 2016. DOI: 10.1590/S0100-204X2016000900041.

KLEIN, V. A.; LIBARDI, P. L. Densidade e distribuição do diâmetro dos poros de um Latossolo Vermelho, sob diferentes sistemas de uso e manejo. **Revista Brasileira de Ciências do Solo**, v.26, p.857-867, 2002.

LIAW, A.; WIENER, M. Classification and regression by random forest. **Review News**, v.2, p.18-22, 2002.

LOPES, T. D.; GOEDTEL, A.; PALACIOS, R. H. C.; GODOY, W. F. **Aplicação do Algoritmo Random Forest como classificador de padrões de falhas em rolamentos de motores de indução**. XIII Simpósio Brasileiro de Automação Inteligente. Porto Alegre, 2017.

LOCATELLI, M.; FILHO, E. P. S.; VIEIRA, A. H.; MARTINS, E. P.; PEQUENO, P. L. L. **Castanha-do-Brasil – Opção para solo de baixa fertilidade na Amazônia**. Embrapa, 2003. Disponível em:
http://ainfo.cnpti.embrapa.br/digital/bitstream/item/54332/1/locatelli-2003.pdf.

MARCHETTI, M. M.; BARP, E. A. Efeito rizosfera: a importância de bactérias fixadoras de nitrogênio para o solo/planta–revisão. **Revista de Engenharia e Inovação Tecnológica**, v.4(1), p.61-71, 2015.

MENEGHIN, M. F. S.; RAMOS, M. L. G.; OLIVEIRA, S. A.; RIBEIRO, W. Q. J.; AMABILE, R. F. avaliação da disponibilidade de nitrogênio no solo para o trigo em latossolo vermelho do Distrito Federal. **Revista Brasileira de Ciências do Solo**, v.32, p.1941-1948, 2008.

MOOLMAN, J. H; VAN HUYSSTEEN, L. A geostatistical analysis of the penetrometer soil strength of a deep ploughed soil. **Soil Tillage Research**. v.15(1-2), p.11-24, 1989.

MULLER, C. H; FIGUEIRÊDO, F. J. C; KATO, A. K; CARVALHO, J. E. U; STEIN, R. L. B; SILVA, A. B. **A cultura da Castanha-da-amazônia**. Belém: EMBRAPA, p.65, 1995.

NICOLODI, M.; ANGHINONI, I.; GIANELLO, C. Indicadores da acidez do solo para recomendação de calagem no sistema plantio direto. **Revista Brasileira de Ciência do Solo**, v.32, p.237-247, 2008.

NOGUEIRA, A. R. A; SOUZA, G. B. **Manual de laboratório: solo, água, nutrição vegetal, nutrição animal e alimentos.** São Carlos: Embrapa Pecuária Sudeste. p.334, 2005.

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica**. Dissertação (Mestrado). Departamento de Bioinfomática. Ribeirão Preto, 2013.

PEBESMA, E. J.; WESSELING, C. G. Gstat, a program for geostatistical modelling, prediction and simulation. **Computers & Geosciences**, v.24 (1), p.17–31, 1998.

PENNA, C.; LINHARES, F. Robustez de regressões de crescimento frente à incerteza sobre a especificação do modelo: quão robustos são os regressores para o caso brasileiro? **Estudos Econômicos**, São Paulo, v.45(4), p. 897-925, 2015.

PEREIRA, H. S. Fósforo e potássio exigem manejos diferenciados. **Visão Agrícola**, n.9, p.43-46, 2009.

QIAO, C., HAIBO, H., & HONG, M. Spatial outlier detection based on iterative selforganizing learning model. **Neurocomputing**, v.117, p. 161-172, 2013.

R-DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R. Foundation for Statistical Computing, 2017.

RONQUIM, C. C. **Conceitos de fertilidade do solo e manejo adequado para as regiões tropicais**. Boletim de Pesquisa e Desenvolvimento 8. Embrapa Monitoramento por Satélite Campinas, SP, 2010.

RUE, H.; MARTINO, S.; CHOPIN, N. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". **Journal of the Royal Statistical Society**, Series B, v. 71, p. 31, 2009.

SANTOS, A. M. R. T. **Outliers em variáveis geoespaciais: proposições utilizando geoestatística**. Tese (Doutorado). Departamento de Engenharia Civil. Universidade Federal de Viçosa. Viçosa, 2016.

SANTOS, D. R.; GATIBONI, L. C.; KAMINSKIL, J. Fatores que afetam a disponibilidade do fósforo e o manejo da adubação fosfatada em solos sob sistema plantio direto. **Ciência Rural**, v.38(2), p.576-586, 2008.

SANTOS, G. R.; OLIVEIRA, M. S.; LOUZADA, J. M.; SANTOS, A. M. R. T. Krigagem Simples versus Krigagem universal: qual o preditor mais preciso? **Energia na Agricultura**, v.26(2), p.49-55, 2011.

SILVA, V. R.; REINERT, D. J.; REICHERT, J. M. Fatores controladores da compressibilidade dos solos Argissolo Vermelho-Amarelo distrófico arênico e Latossolo

Vermelho distrófico típico. II. Grau de saturação de água. **Revista Brasileira de Ciências do Solo**, v.26, p.9-16, 2002b.

VALENTE, M. A.; OLIVEIRA JUNIOR, R. C.; SILVA FILHO, E. P. **Caracterização e mapeamento dos solos do campo experimental de Porto Velho**. In: Congresso Brasileiro de Ciência do Solo. Anais.Cd Rom, Rio de Janeiro. 1997.

YAMAMOTO, Jorge K.; LANDIM, Paulo M. B. **Geoestatística Conceitos e Aplicações**. Oficina de Texto, 2013.

# CONCLUSÕES GERAIS

O presente trabalho demonstra um potencial incrível ao associar a krigagem geoestatística com aprendizado de máquina, em estudo de variáveis geoespaciais contínuas, para predizer valores da variável de interesse em qualquer ponto da região do estudo, com uma distribuição espacial o mais real possível, sem a restrição ao tamanho do conjunto de dados e ao número de variáveis.

A utilização dessa metodologia na produção média das castanheiras no estado do Amapá/Brasil relacionado aos atributos do solo é extremante importante devido à extinção das Castanheiras-da-amazônia (Bertholletia excelsa) que tem sido uma preocupação para a região, pois está incluída na Lista Vermelha da União Internacional para Conservação da Natureza (IUCN) como vulnerável pelo desmatamento que ameaça a espécie.

Na busca por interpoladores mais acurados para a predição, foi proposto no primeiro capítulo, o preditor Random Forest Ordinary Kriging (RFOK) que apresentou as desvantagens não só de excluir variáveis consideradas importantes no estudo pela presença do efeito pepita puro, bem como, dispõe de uma metodologia inviável para modelar grandes conjuntos de dados.

Para contornar essas desvantagens, no segundo capítulo, foi proposto o preditor Random Forest Kriging (RFK) que apresentou uma superação nos problemas citados acima, favorecendo a modelagem de grandes conjuntos de dados sem restrição ao número de variáveis.

A robustez da metodologia sobre o preditor RFK, adotando o erro quadrático médio e o coeficiente de determinação do modelo entre os valores observados e preditos, mostrou-se satisfatório e resultou em predições mais acuradas da krigagem, mostrando que esse preditor é extremamente apropriado para a predição de uma variável de interesse. Se compararmos a performance do capítulo 1 com o capítulo 2, percebe-se que a superação do RFK é incrível por apresentar-se o mais acurado, com menor erro quadrático médio ($EQM_{RFK} = 2{,}56$) em relação RFOK ($EQM_{RFOK} = 4307{,}57$), ou seja, apresentou uma menor diferença entre o valor real e o valor predito na interpolação dos dados.

Assim, fica evidente que tratar a exclusão de variáveis importantes e os outliers presente nos dados associado à técnica bagging e o empilhamento de rasters resultaram em uma suavização no mapa de predição, favorecendo a uma distribuição espacial da variável de

interesse mais real e, consequentemente, uma discriminação detalhada sobre cada variável em estudo.

Além disso, o RFK mostrou-se eficiente também para solucionar os desafios propostos no início da pesquisa, tais como: modelar, simultaneamente, o comportamento espacial de múltiplas variáveis, sem a perda de nenhuma informação que seja considerada importante para o desenvolvimento do trabalho; direcionar amostragens com mais eficiências e minimizar prejuízos de ordem social, ecológica e econômica para qualquer região que tenha interesse nesse tipo de estudo, especificamente, para a região do Amapá; e, contribuir para subsídios mais confiáveis envolvendo o manejo florestal, a manutenção e a ampliação da produtividade de qualquer espécie de interesse.

Diante do exposto, pode-se concluir que o preditor RFK apresentou um potencial inédito e extraordinário para contribuir favoravelmente ao conhecimento científico, tecnológico e social, gerando mapas de predição que mais se aproxima da realidade e com o menor erro possível.