Método de Desagregação utilizando Random Forest: análise da densidade demográfica na Região Metropolitana Vale do Paraíba e Litoral Norte, SP

Análise de Dados Espaciais

Docentes: Dr. Antonio Miguel Monteiro, Dr. Eduardo Celso Gerbi Camargo

2020

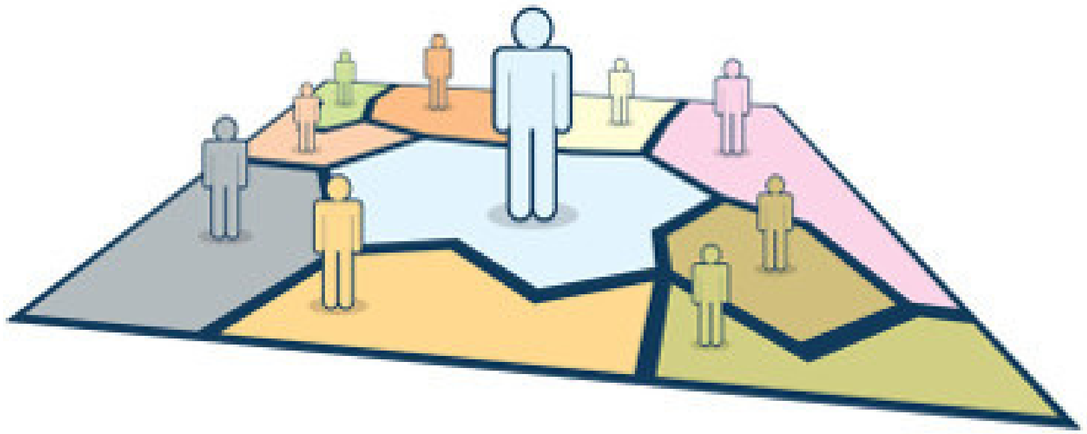
Discente: Diego Moreira Silva

Introdução

- A densidade populacional permite avaliar como a população se distribui em determinado território;

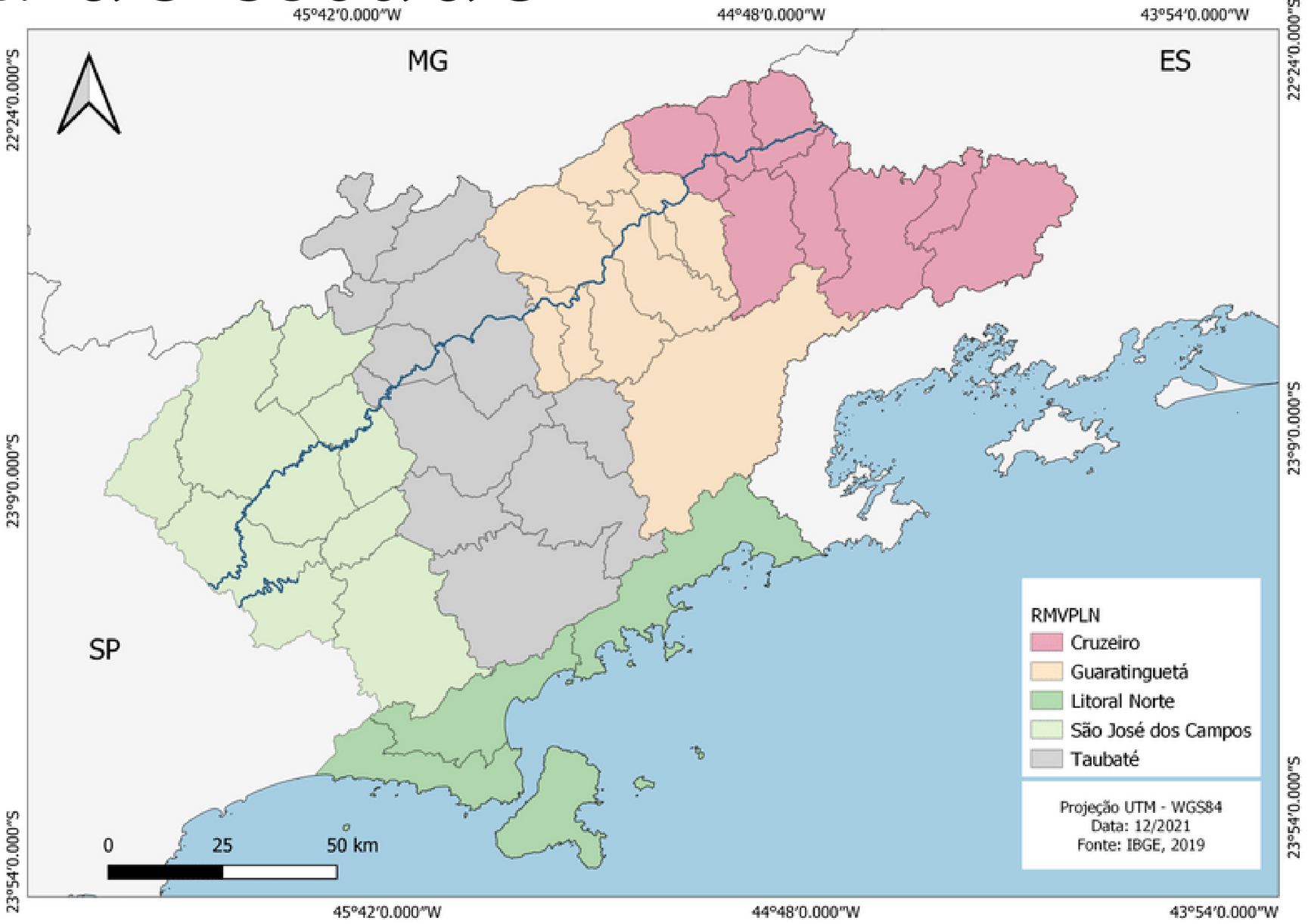
- Fomento para tomada de decisões que beneficiem a sociedade, no contexto do planejamento territorial (ACHEAMPONG, 2019);

- Métodos de desagregação populacional para estimar valores populacionais de áreas maiores para áreas menores.



Fonte: Google imagens

Área de estudo



Dados utilizados

Variáveis explicativas	
Dados utilizados	Tipos
rmvpln_esaccilc_dst011_100m_2015	Distância das bordas de área cultivada de uso e cobertura ESA (2015)
rmvpln_esaccilc_dst040_100m_2015	Distância até a borda de árvores lenhosas de uso e cobertura ESA (2015)
rmvpln_wdpa_dst_cat1_100m_2015	Distância até a reserva natural restrita da IUCN e bordas da área de largura (2015)
rmvpln_srtm_topo_100m	SRTM elevação 2000
rmvpln_osm_dst_roadintersec_100m_2016	Distância até a interseção da estrada principal OSM 2016
rmvpln_osm_dst_waterway_100m_2016	Distância para as principais hidrovias da OSM 2016
rmvpln_viirs_100m_2015	Dados de luzes noturnas 2015
rmvpln_esaccilc_dst190_100m_2015	Distância até as bordas de superfícies artificiais de uso e cobertura ESA (2015)
Variável resposta	
Projeção da população SEADE (2015)	Informação de Cartórios Civis e coleta eventos vitais e registros

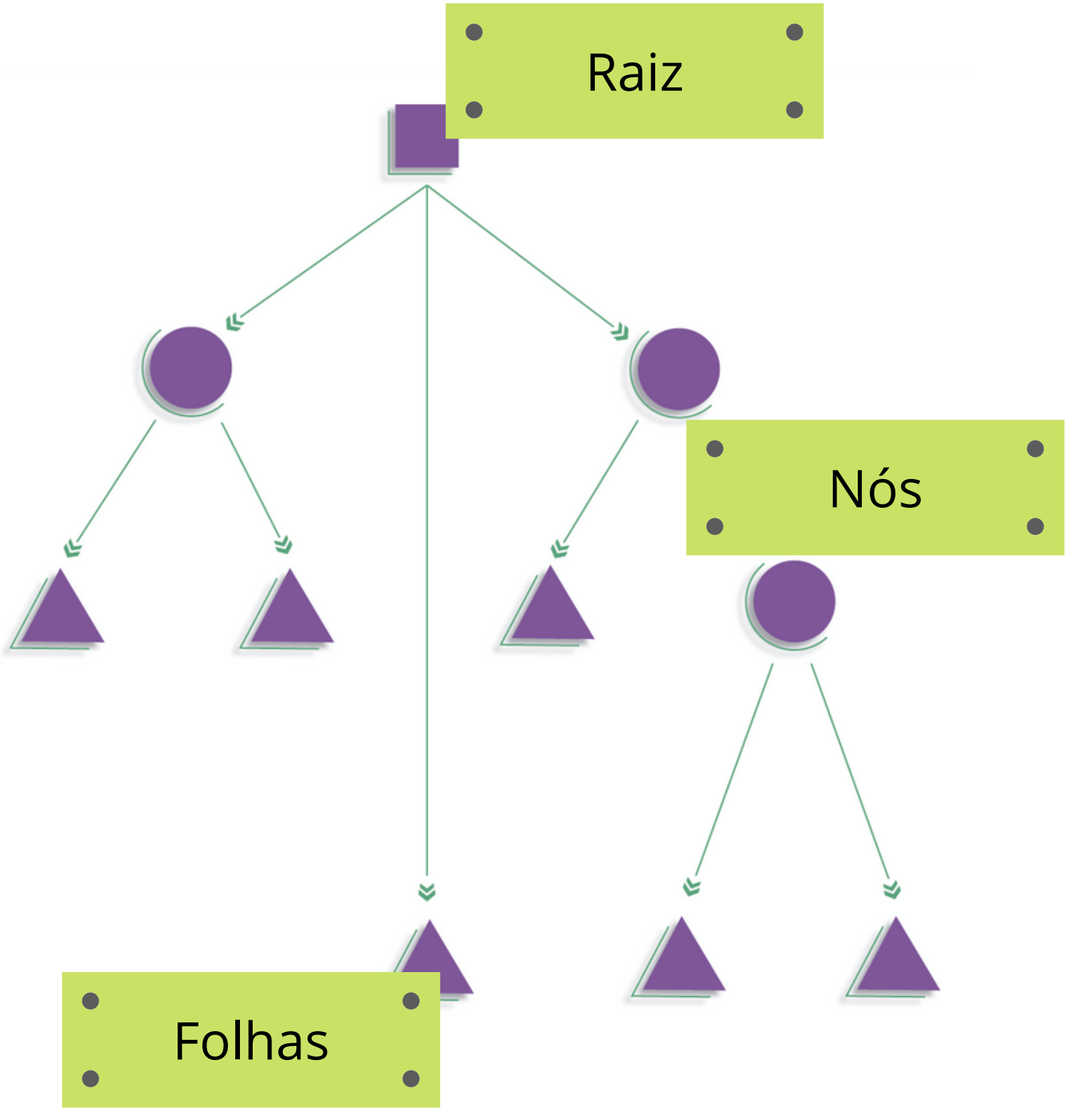
Perguntas

1. Dentre as variáveis explicativas utilizadas, quais demonstram ser boas para o modelo de predição?

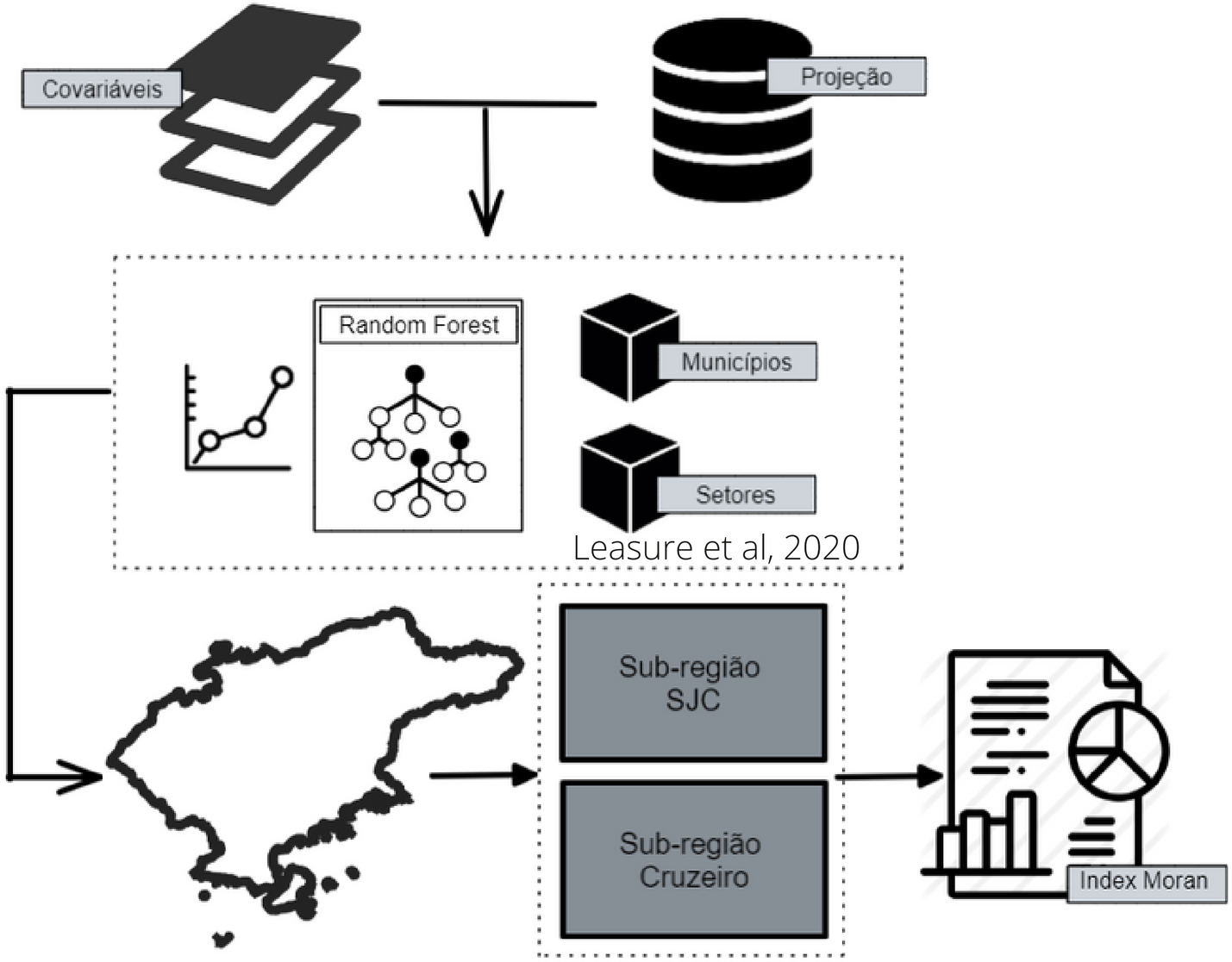
2. Utilizando as métricas de análise espacial do Moran Global e Local, quais padrões foram encontrados na RMVPLN ?

Random Forest??

- Técnica de machine learning proposta por Breiman (2001), composta por um conjunto determinado de árvores de decisão;
- Constrói um conjunto de treinamento para cada árvore;
- Seleção de subconjuntos de atributos para cada preditor;
- Lida com a classificação (variáveis categóricas) e modelos de regressão (variáveis contínuas).



Metodologia



Fonte: autor.

Argumentos do RF

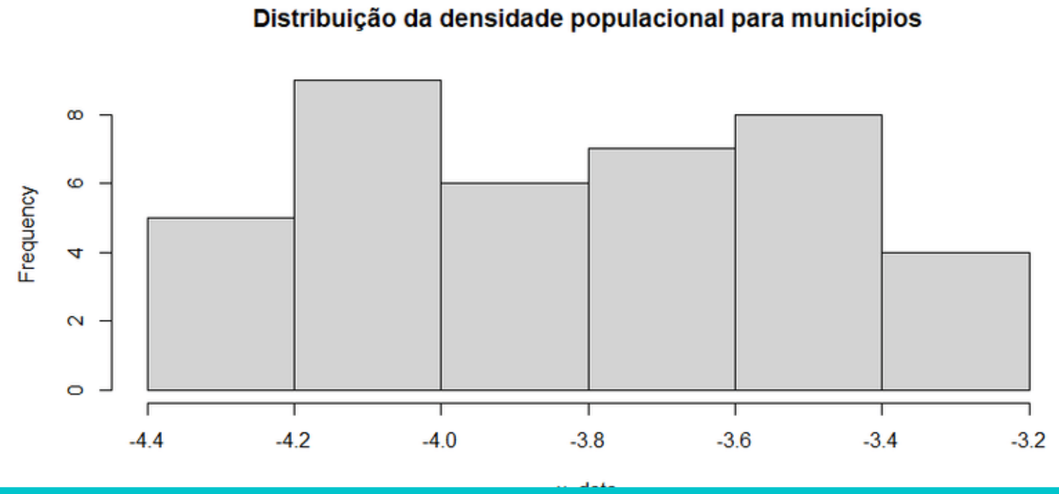
```
popfit <- tuneRF(x=x_data,
                 y=y_data,
                 plot=TRUE,
                 mtryStart = length(x_data)/2, # valor inicial de mtry*
                 ntreeTry=901, # número de árvores determinada par o modelo
                 improve=0.0001,
                 stepFactor=1.5,
                 trace=TRUE,
                 doBest=TRUE,
                 nodesize=length(y_data)/150,  # quantidade mínima de nós
                 na.action=na.omit,
                 importance=TRUE,
                 sampsize=length(y_data), # tamanho da amostra para calculo do OOB
                 replace=TRUE)
```

mtry: nº de variáveis aleatórias como candidatas em cada divisão do nó.

OOB (out-of-bag): previsão de erro do modelo.

Desagregação do RF

y_data = população total / área (m²)



Cálculo dos pesos:

$$weight_i = \frac{exp(predicted_i)}{\sum_{i=1}^{I_j} exp(predicted_i)}$$

onde:

$predicted_i$ **é a predição para os setores censitários;**

I_j **é o total de setores no município *j*.**

Redistribuição para o nível dos setores:

$$population_i = total_j \times weight_i$$

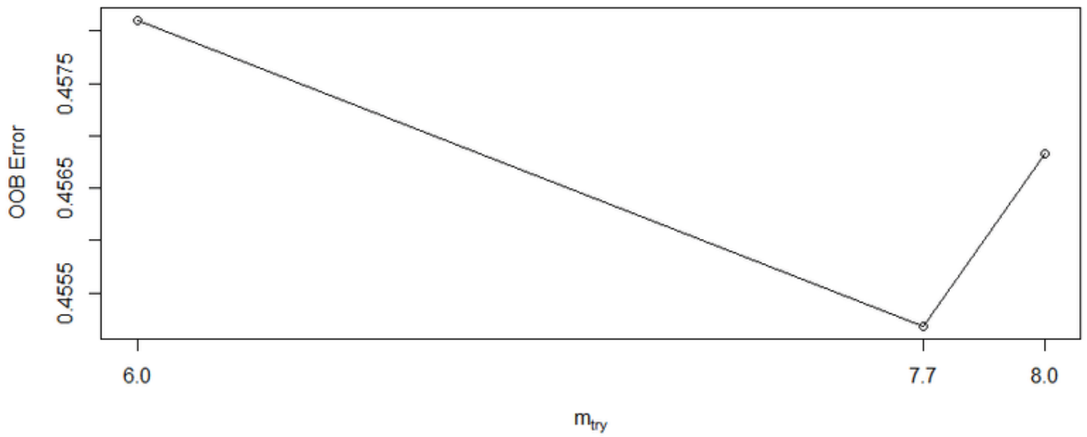
em que:

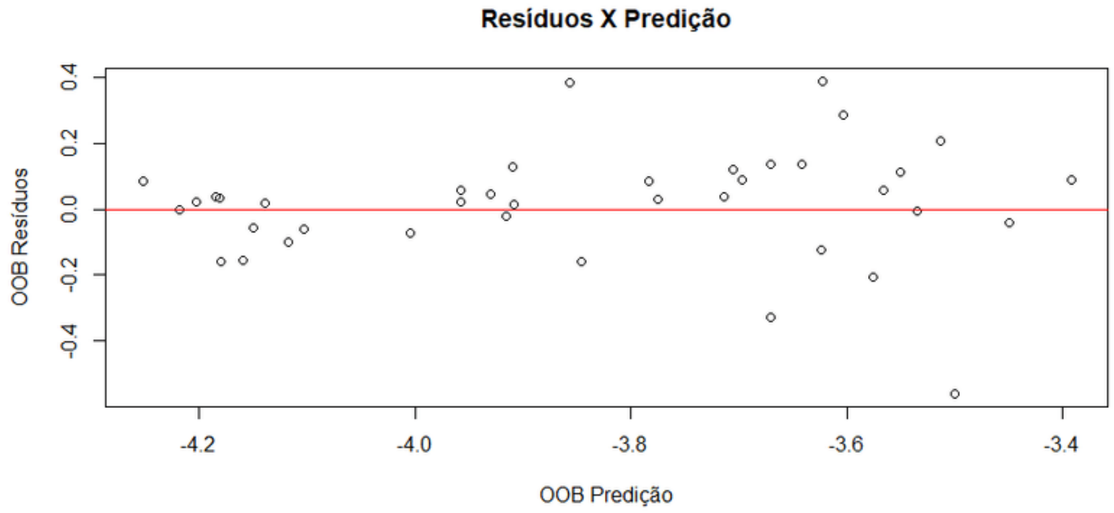
$total_i$ **é o total para cada município *j*;**

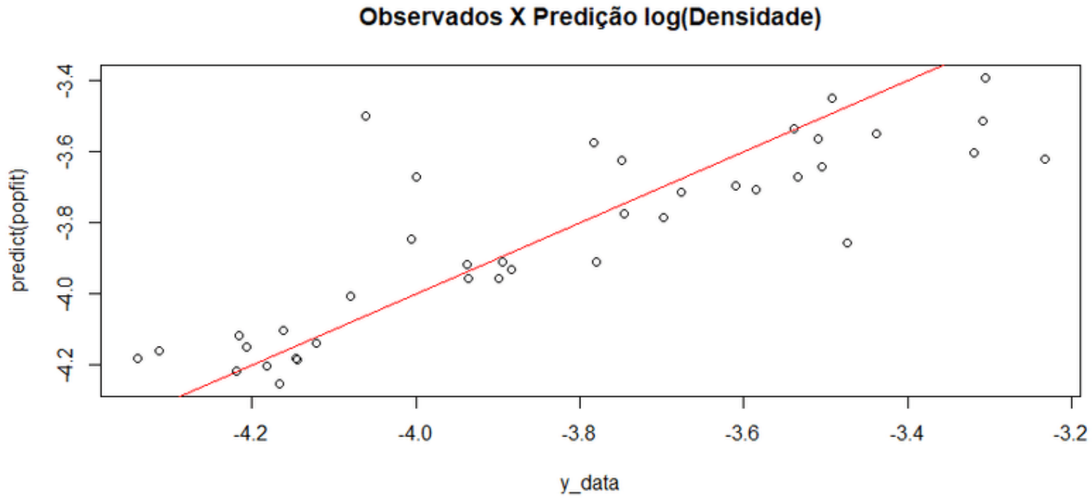
$weight_i$ **é a proporção da população que vive no setor *i*.**

Resultados

```
Call:
 randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1],
replace = TRUE, sampsize = ..4, nodesize = ..1, importance = TRUE,
na.action = ..2)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 0.02894021
                    % Var explained: 70.95
```

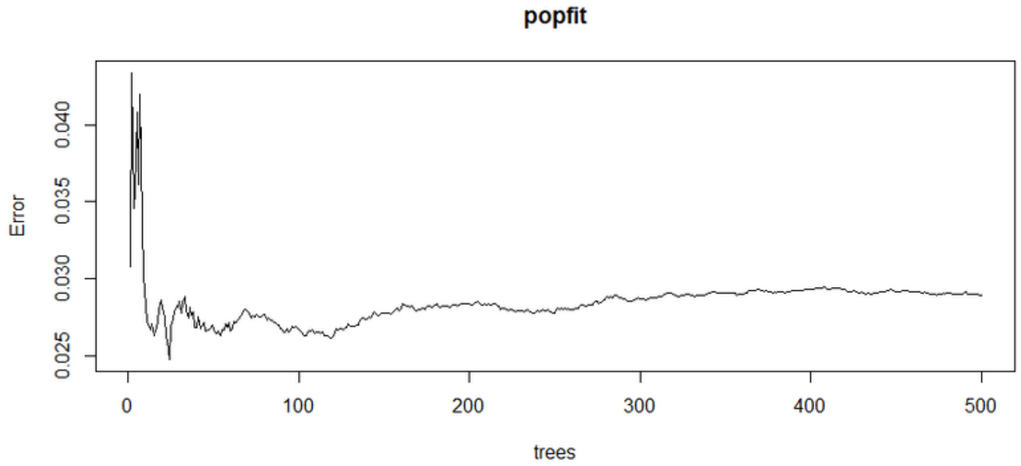


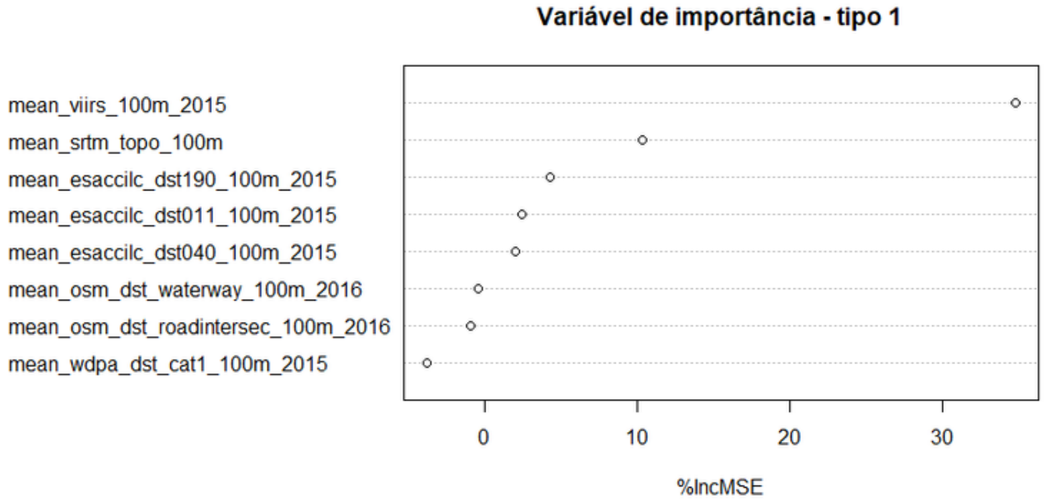


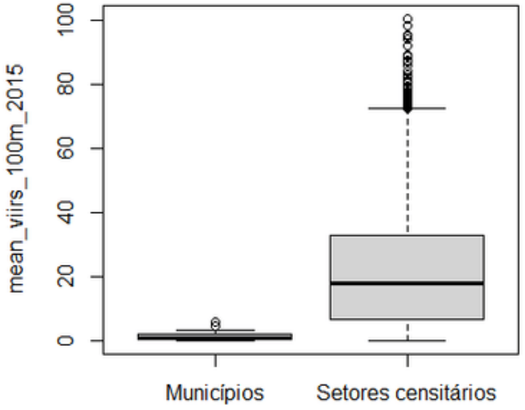


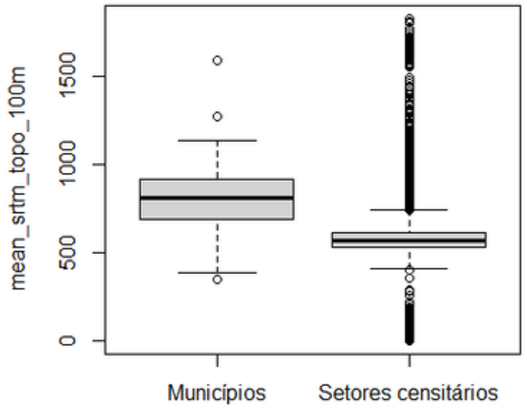
Fonte: autor.

Resultados









Fonte: autor.

Indicadores de regimes espaciais

Moran Global

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n} z_i^{\,2}}$$

Sub-região São José dos Campos

```
        Moran I test under randomisation

data:  OA.Census$predicted_pop
weights: listw

Moran I statistic standard deviate = 54.494, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.6822347546     -0.0004006410      0.0001569221
```

Índice de Moran: 0.6822

p-valor: ~ 0

Sub-região Cruzeiro

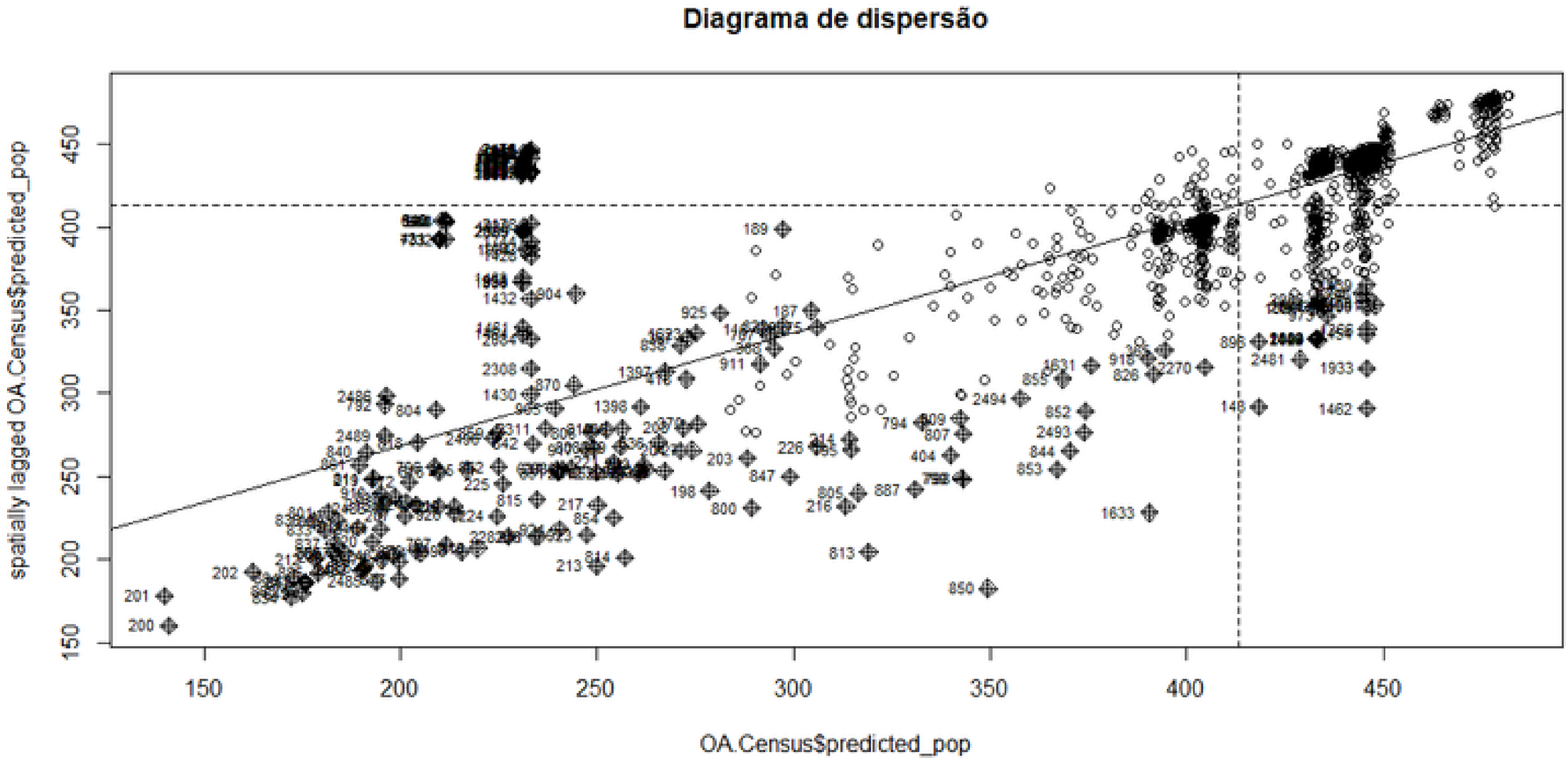
```
        Moran I test under randomisation

data:  OA.Census$predicted_pop
weights: listw

Moran I statistic standard deviate = 24.395, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.889076695      -0.003278689       0.001338055
```

Índice de Moran: 0.8891

p-valor: 9.66064 e-132

Dispersão de Moran

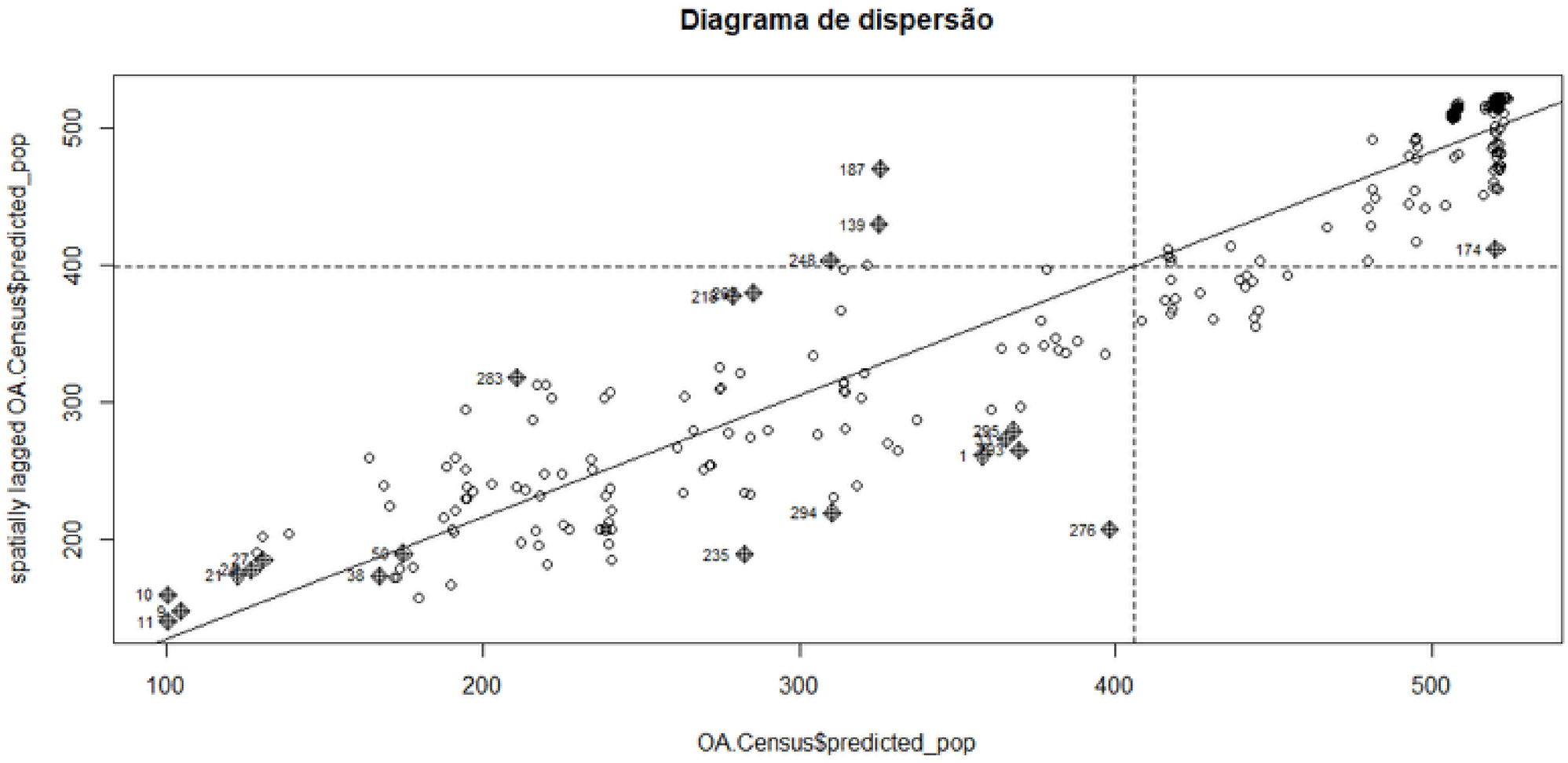
Sub-região São José dos Campos



Fonte: autor.

Dispersão de Moran

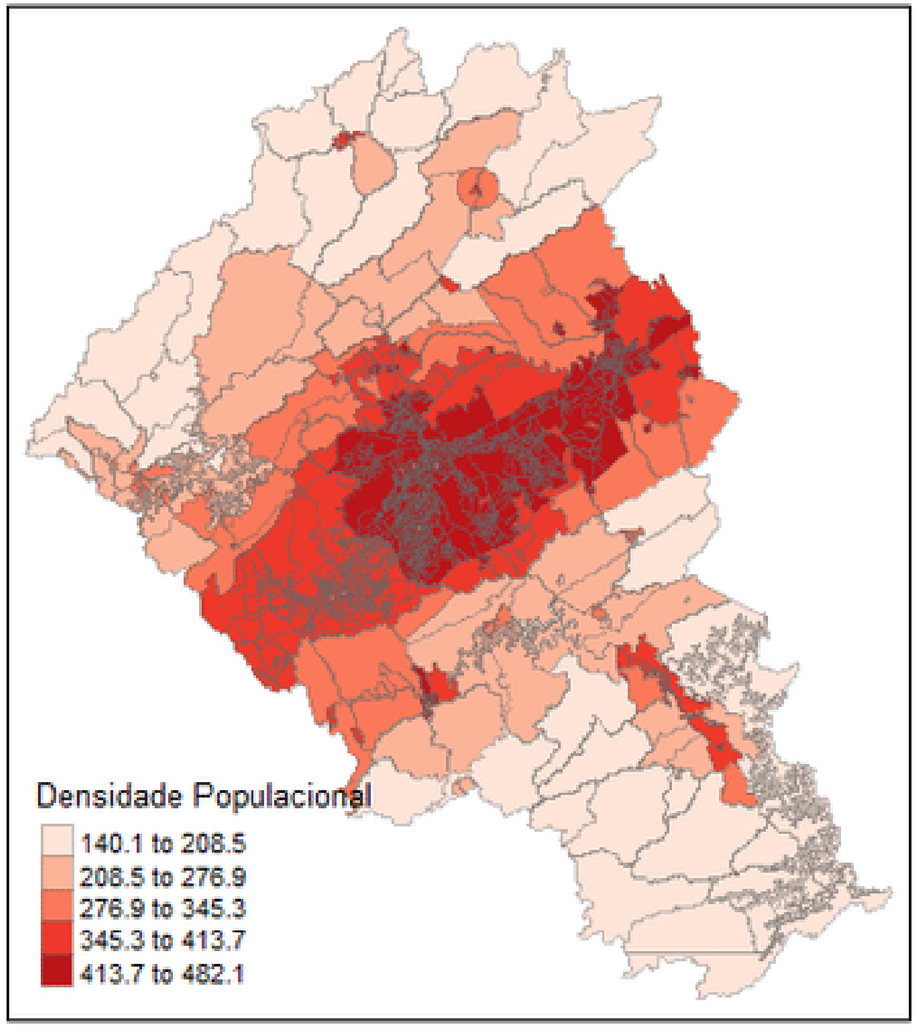
Sub-região Cruzeiro



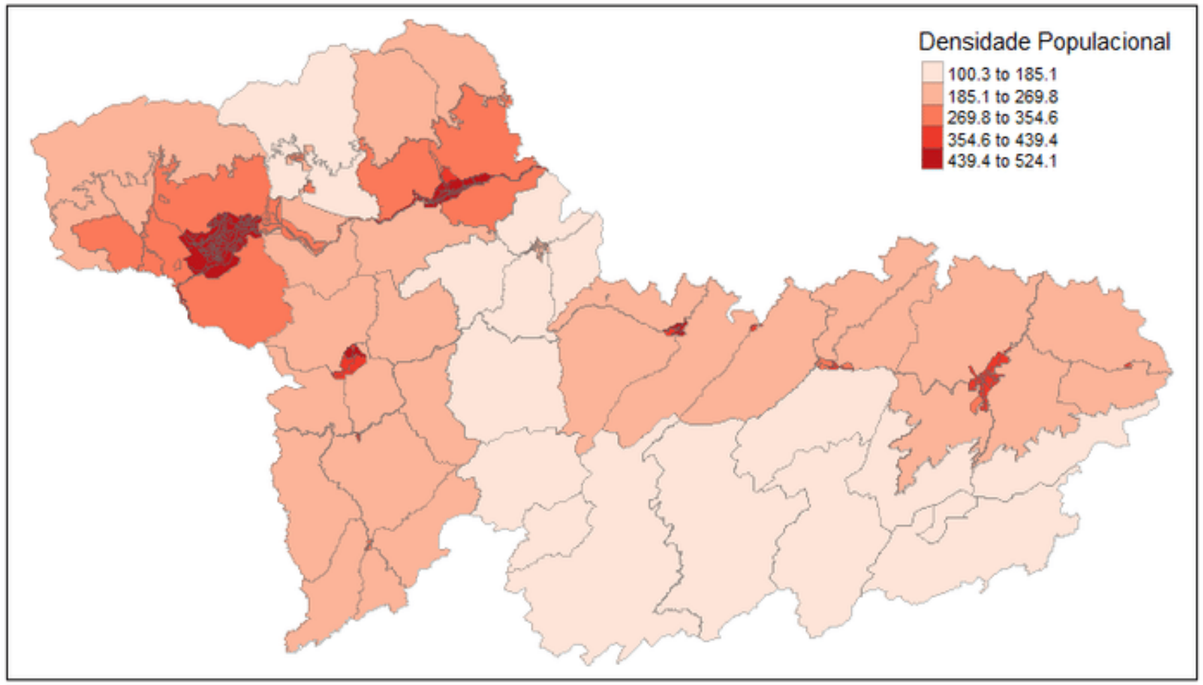
Fonte: autor.

Densidade sub-regiões

Sub-região São José dos Campos



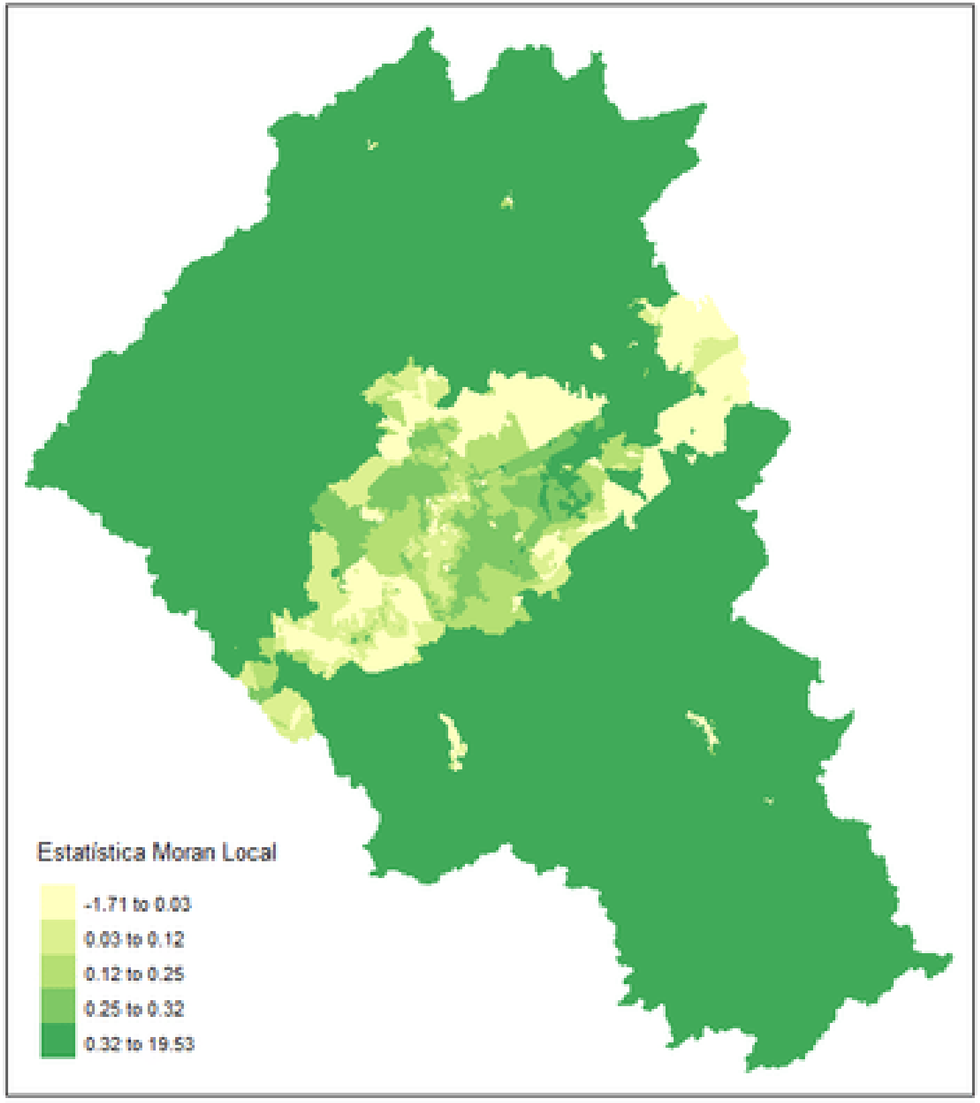
Sub-região Cruzeiro



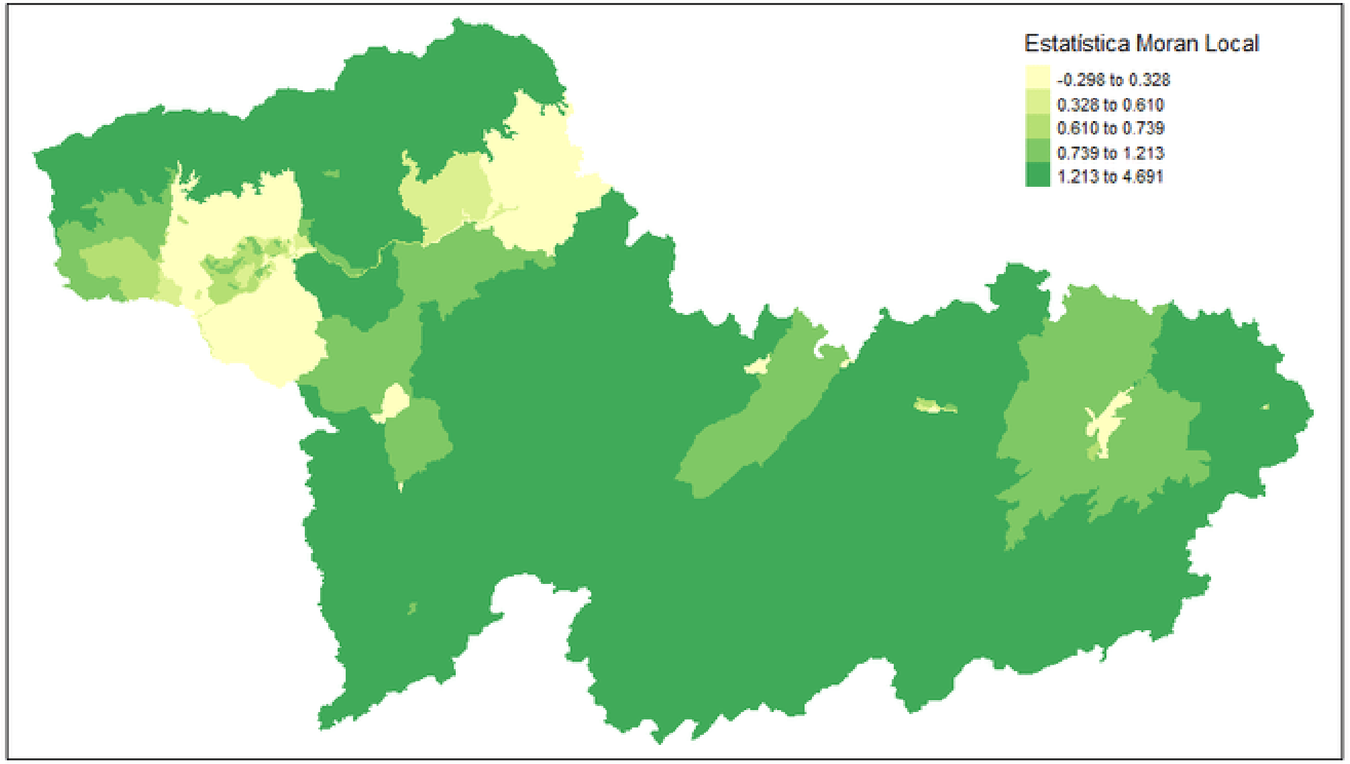
Moran Local

Sub-região São José dos Campos

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}$$



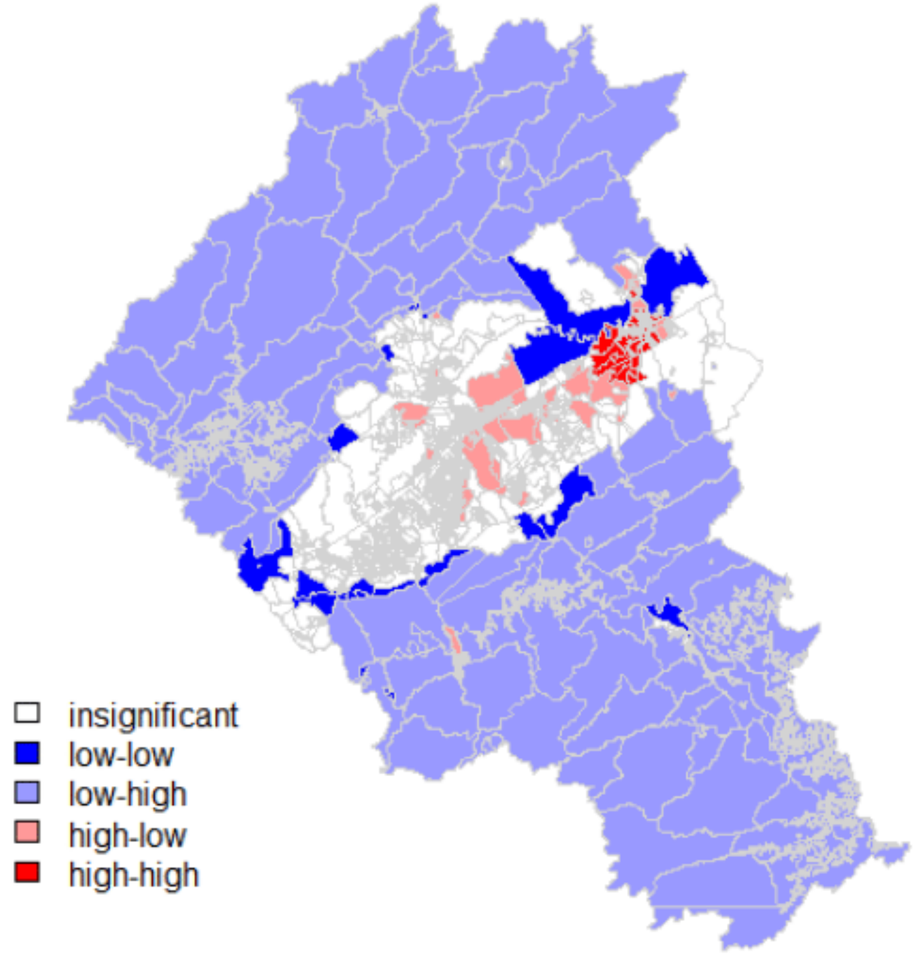
Sub-região Cruzeiro



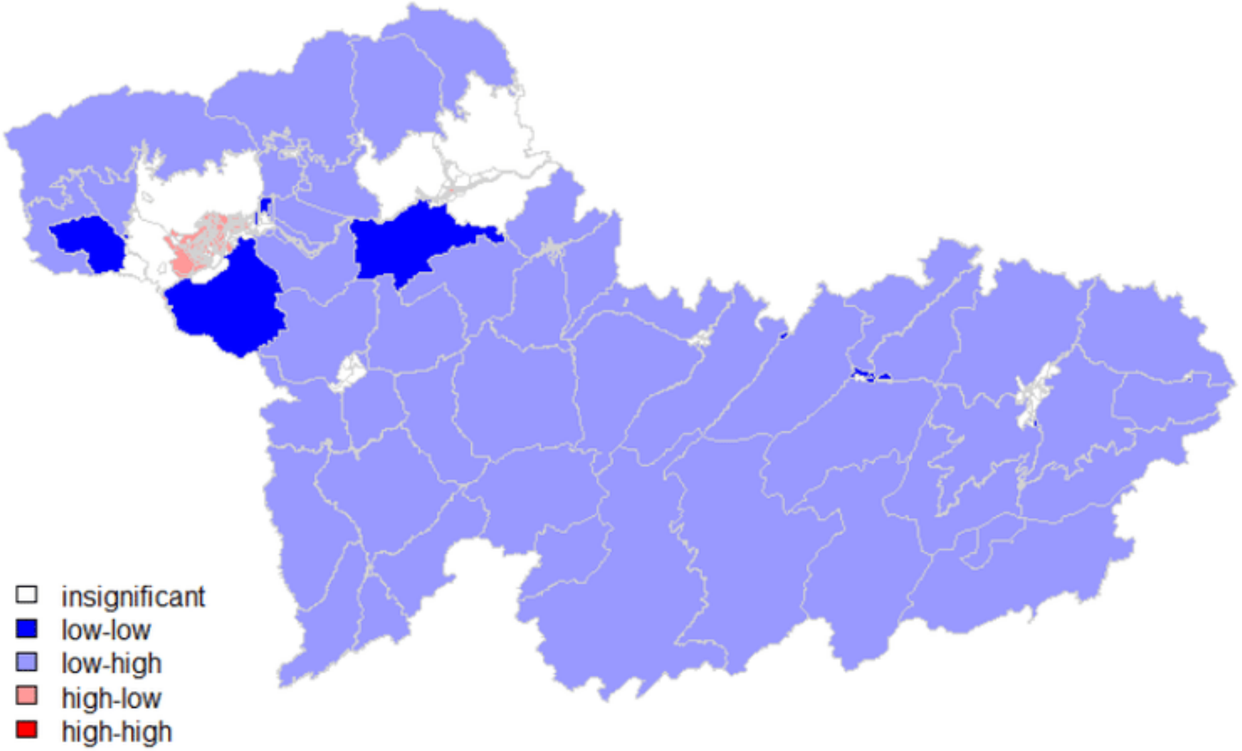
Moran Local

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}$$

Sub-região São José dos Campos



Sub-região Cruzeiro



Considerações

- O Random Forest lida consegue lidar com conjunto de dados grandes e com valores grandes de atributos e devido o caráter aleatório, o RF minimiza os efeitos de sobreajuste;
- Uma das suas principais desvantagens é o custo de tempo de trabalho para o ajuste do modelo a depender da quantidade de dados utilizados;
- O método de desagregação do Random Forest se mostrou útil em lidar com a distribuição da densidade populacional, no entanto, como afirma Leasure et al (2020) , é necessário analisar os métodos de entrada e compreender as relações entre as variáveis explicativas e a variável resposta;
- Como perspectiva futura, utilizar outros algoritmos modelos de regressão espacial, como o spatialRF, pode configurar em uma melhora dos resultados;
- Adotar outras variáveis explicativas para a RMVPLN.

Referências

SORICHETTA, Alessandro et al. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Scientific data, v. 2, n. 1, p. 1-12, 2015.

AMARAL, Silvana et al. Interpoladores Espaciais para Geração de Superfícies de Densidade Populacional na Amazônia Brasileira: problemas e perspectivas. Simpósio Brasileiro de Geoinformática, IV, p. 73-82, 2002.

Leasure DR, Darin E, Tatem AJ. 2020. Small area population estimates using random forest top-down disaggregation: An R tutorial. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00697.

ROKACH, Lior; MAIMON, Oded Z. Data mining with decision trees: theory and applications. World scientific, 2007.

GEORGANOS, Stefanos et al. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International, v. 36, n. 2, p. 121-136, 2021.

WEBSTER, Richard; OLIVER, Margaret A. Geostatistics for environmental scientists. John Wiley & Sons, 2007.

BIVAND, Roger S.; WONG, David WS. Comparing implementations of global and local indicators of spatial association. Test, v. 27, n. 3, p. 716-748, 2018.

Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10, e0107042 (2015).

Obrigado!