# Overview of Minas Gerais Graphics

A map with blue pins

Description automatically generated

Plot of data collection.

# A diagram of heavy metals and soil properties Description automatically generated

Correlation Plot between all Heavy Metals and Soil Properties. Darker blue colors means stronger more positive correlation, where darker red colors means stronger negative correlations.

A graph with lines and dots

Description automatically generated

Variogram for Arsenic. A **variogram**is an effective tool for describing the behavior of non-stationary spatial random processes. It is used primarily in spatial statistics, it is an “essential step” for analyzing spatial variability (Gómez-Hernández et al., 1999).

The variogram measures dissimilarity in a relatively simple way. Pairs of data points are separated by a specified **lag distance**.

Basically will describe the degree of spatial dependence; it gives a measure of how much two samples taken from the mining area will vary in arsenic percentage depending on the distance between those two samples. Samples taken far apart will vary more than samples taken close to each other.

A graph showing the growth of soil properties

Description automatically generated

Multivariate variogram for all soil properties. It is strictly equivalent to summing univariate variograms. (Wagner, 2003).

A diagram of a number of dots

Description automatically generated with medium confidence

Plot of the distribution of arsenic levels across Minas Gerais region. Pink indicates high levels of arsenic (8.0 mg and above) while blue indicates low levels of arsenic. The larger the dot, the higher the level of arsenic.

Plot made using ggplot package in R (reference Wickham, 2016)

When multivariate data are considered, it is possible to search for spatial structures by computing univariate statistics (e.g., Moran’s Coefficient) on each variable separately. Another alternative is to summarize data by multivariate methods and then detect spatial structures using the output of the analysis.

We show first that there exists spatial autocorrelation between PC1 and PC2

class: krandtest lightkrandtest

Monte-Carlo tests

Call: moran.randtest(x = pca.datum$li, listw = listwdatum)

Number of tests: 2

Adjustment method for multiple comparisons: none

Permutation number: 999

Test Obs Std.Obs Alter Pvalue

1 Axis1 0.1701520 14.29605 greater 0.001

2 Axis2 0.2022094 16.15471 greater 0.001

PCA results are highly spatially structured. Results can be optimized, compared to this two-step procedure, by searching directly for multivariate spatial structures. The multispati function implements a method (Dray, Saïd, and Débias 2008) that search for axes maximizing the product of variance (multivariate aspect) by MC (spatial aspect):

Multivariate Spatial Analysis

Call: adespatial::multispati(dudi = pca.datum, listw = listwdatum,

scannf = FALSE)

Scores from the initial duality diagram:

var cum ratio moran

RS1 6.738016 6.738016 0.3369008 0.1701520

RS2 3.911494 10.649510 0.5324755 0.2022094

Multispati eigenvalues decomposition:

eig var moran

CS1 1.2788871 5.928023 0.2157358

CS2 0.8634495 3.919989 0.2202683

A diagram of a network

Description automatically generated

Plot of PC1 and PC2 from MULTISPATI analysis

A screen shot of a chart

Description automatically generated

Plot of Moran’s I at different threshold values.

Plot helps assessing the spatial autocorrelation of the response variable and the predictors across different distance thresholds. Low Moran’s I and p-values equal or larger than 0.05 indicate that there is no spatial autocorrelation for the given variable and distance threshold. These distance thresholds are the neighborhoods at which the model will check the spatial autocorrelation of the residuals.

A graph of different types of graphs

Description automatically generated with medium confidence

Important residual plots for the nonspatial random forest model. We see from the top plot that the residuals do not seem to be normally distributed. The Residuals versus predictors plot shows that there seems to be some sort of autocorrelation of the residuals. In fact, we see from the plot of the multiscale Moran’s I that there is spatial autocorrelation between the residuals

Results from the R Code are below

Model type

- Fitted with: ranger()

- Response variable: As

Model performance

- R squared (oob): 0.0167187

- R squared (cor(obs, pred)^2): 0.8822978

- Pseudo R squared (cor(obs, pred)):0.9393071

- RMSE (oob): 52.8616

- RMSE: 27.6242

- Normalized RMSE: 3.230466

Model residuals

- Stats:

┌─────────┬────────┬────────┬───────┬────────┬────────┐

│ Min. │ 1st Q. │ Median │ Mean │ 3rd Q. │ Max. │

├─────────┼────────┼────────┼───────┼────────┼──────┤

│ -101.36 │ -4.15 │ -1.89 │ -0.58 │ -0.50 │ 444.51 │

└─────────┴────────┴────────┴───────┴────────┴────────┘

- Normality:

- Shapiro-Wilks W: 0.26

- p-value : 0

- Interpretation : Residuals are not normal

- Spatial autocorrelation:

┌──────────┬───────────┬───────── ──────────────────────────────┐

│ Distance │ Moran's I │ P value │ Interpretation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 0.0 │ 0.014 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 175.0 │ 0.430 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 351.0 │ 0.378 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 526.0 │ 0.282 │ 0.000 │ Positive spatial correlation │

└──────────┴───────────┴─────────┴──────────────────────────────┘

Variable importance:

┌─────────────┬─────────────┐

│ Variable │ Importance │

├─────────────┼─────────────┤

│ Clay │ 0.526 │

├─────────────┼─────────────┤

│ Ca2+ │ 0.518 │

├─────────────┼─────────────┤

│ H+Al │ 0.495 │

├─────────────┼─────────────┤

│ SB │ 0.452 │

├─────────────┼─────────────┤

│ Silt │ 0.425 │

├─────────────┼─────────────┤

│ pH\_KCl │ 0.403 │

├─────────────┼─────────────┤

│ SOM │ 0.401 │

├─────────────┼─────────────┤

│ V\_(%) │ 0.385 │

├─────────────┼─────────────┤

│ PREM │ 0.385 │

├─────────────┼─────────────┤

│ pH\_H2O │ 0.384 │

├─────────────┼─────────────┤

│ CEC │ 0.357 │

├─────────────┼─────────────┤

│ ECEC │ 0.351 │

├─────────────┼─────────────┤

│ Coarse\_sand │ 0.333 │

├─────────────┼─────────────┤

│ Fine\_sand │ 0.315 │

├─────────────┼─────────────┤

│ Mg2+ │ 0.314 │

├─────────────┼─────────────┤

│ m\_(%) │ 0.282 │

├─────────────┼─────────────┤

│ Latitude │ 0.280 │

├─────────────┼─────────────┤

│ PM │ 0.240 │

├─────────────┼─────────────┤

│ Al3+ │ 0.238 │

├─────────────┼─────────────┤

│ Longitude │ 0.230 │

├─────────────┼─────────────┤

│ K+ │ 0.191 │

└─────────────┴─────────────┘

The main take away is the OOB (out of bag) R-square is low (1.67%). Indiciates a very weak predictive importance. All random forest fit using spatialRF package which uses ranger package as wrapper.

A graph with red dots

Description automatically generated

Nonspatial RF variable importance.

Variable importance represents the increase in mean error (computed on the out-of-bag data) across trees when a predictor is permuted. Values lower than zero would indicate that the variable performs worse than a random one.

A screenshot of a computer

Description automatically generated

The package [randomForestExplainer](https://github.com/ModelOriented/randomForestExplainer) (Ishwaran et. al, 2010) offers a couple of interesting options to deepen our understanding on variable importance scores. The first one is [measure\_importance()](https://rdrr.io/pkg/randomForestExplainer/man/measure_importance.html), which analyzes the forest to find out the average minimum tree depth at which each variable can be found (mean\_min\_depth), the number of nodes in which a variable was selected to make a split (no\_of\_nodes), the number of times the variable was selected as the first one to start a tree (times\_a\_root), and the probability of a variable to be in more nodes than what it would be expected by chance (p\_value).

A graph showing a line and a curve

Description automatically generated with medium confidence

A graph with lines and dots

Description automatically generated

Plot of model transferability across spatial folds

Plot of median contribution across 30 different spatial folds

The importance computed by random forest on the out-of-bag data by permutating each predictor (as computed by [rf()](https://blasbenito.github.io/spatialRF/reference/rf.html)) and the contribution of each predictor to model transferability (as computed by [rf\_importance()](https://blasbenito.github.io/spatialRF/reference/rf_importance.html)) show practically no correlation, indicating that the importance measures do not capture different aspects of the effect of the variables on the model results.

We see that each variable ends up contributing only slight increases to R-Squared across the spatial folds. In fact, we see that most variables seem to decrease the R-squared values.

Using the spatial folds for cross-validation we get

Spatial evaluation

- Training fraction: 0.75

- Spatial folds: 30

Metric Median MAD Minimum Maximum

r.squared 0.006 0.009 0 0.168

A screenshot of a graph

Description automatically generated

Essentially, works well in training in specific spatial region but poor when that is applied to a different region. Regions should not be considered the same, there is a spatial element that needs to be accounted for.

A screenshot of a graph

Description automatically generated

Plot of different spatial fold cross validations

A screenshot of a graph

Description automatically generated

Residual plot for spatial random forest.

Performance for the spatial random forest is below:

Model type

- Fitted with: rf\_spatial()

- rf\_spatial() method: mem.moran.sequential

- Response variable: as

Model performance

- R squared (oob): 0.7049443

- R squared (cor(obs, pred)^2): 0.980021

- Pseudo R squared (cor(obs, pred)):0.9899601

- RMSE (oob): 28.95699

- RMSE: 13.0444

- Normalized RMSE: 1.525456

Model residuals

- Stats:

┌────────┬────────┬────────┬──────┬────────┬────────┐

│ Min. │ 1st Q. │ Median │ Mean │ 3rd Q. │ Max. │

├────────┼────────┼────────┼──────┼────────┼────────┤

│ -15.33 │ -1.88 │ -0.89 │ 0.19 │ -0.10 │ 244.69 │

└────────┴────────┴────────┴──────┴────────┴────────┘

- Normality:

- Shapiro-Wilks W: 0.185

- p-value : 0

- Interpretation : Residuals are not normal

- Spatial autocorrelation:

┌──────────┬───────────┬─────────┬──────────────────────────────┐

│ Distance │ Moran's I │ P value │ Interpretation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 0.0 │ 0.010 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 175.0 │ 0.307 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 351.0 │ 0.305 │ 0.000 │ Positive spatial correlation │

├──────────┼───────────┼─────────┼──────────────────────────────┤

│ 526.0 │ 0.257 │ 0.000 │ Positive spatial correlation │

└──────────┴───────────┴─────────┴──────────────────────────────┘

Variable importance:

┌────────────────────────────────────┬─────────────┐

│ Variable │ Importance │

├────────────────────────────────────┼─────────────┤

│ spatial\_predictors (max) │ 19.356 │

├────────────────────────────────────┼─────────────┤

│ silt │ 5.265 │

├────────────────────────────────────┼─────────────┤

│ spatial\_predictors (quantile 0.75) │ 4.613 │

├────────────────────────────────────┼─────────────┤

│ x │ 4.514 │

├────────────────────────────────────┼─────────────┤

│ h\_al │ 3.486 │

├────────────────────────────────────┼─────────────┤

│ clay │ 2.910 │

├────────────────────────────────────┼─────────────┤

│ som │ 2.873 │

├────────────────────────────────────┼─────────────┤

│ ca2 │ 2.783 │

├────────────────────────────────────┼─────────────┤

│ m\_percent │ 2.551 │

├────────────────────────────────────┼─────────────┤

│ spatial\_predictors (median) │ 2.462 │

├────────────────────────────────────┼─────────────┤

│ al3 │ 2.420 │

├────────────────────────────────────┼─────────────┤

│ p\_h\_k\_cl │ 2.128 │

├────────────────────────────────────┼─────────────┤

│ sb\_2 │ 1.993 │

├────────────────────────────────────┼─────────────┤

│ ecec │ 1.523 │

├────────────────────────────────────┼─────────────┤

│ prem │ 1.492 │

├────────────────────────────────────┼─────────────┤

│ mg2 │ 1.327 │

├────────────────────────────────────┼─────────────┤

│ k │ 1.326 │

├────────────────────────────────────┼─────────────┤

│ coarse\_sand │ 1.195 │

├────────────────────────────────────┼─────────────┤

│ spatial\_predictors (quantile 0.25) │ 1.038 │

├────────────────────────────────────┼─────────────┤

│ pm │ 0.985 │

├────────────────────────────────────┼─────────────┤

│ p\_h\_h2o │ 0.705 │

├────────────────────────────────────┼─────────────┤

│ v\_percent │ 0.531 │

├────────────────────────────────────┼─────────────┤

│ y │ 0.367 │

├────────────────────────────────────┼─────────────┤

│ cec │ -0.855 │

├────────────────────────────────────┼─────────────┤

│ fine\_sand │ -1.754 │

├────────────────────────────────────┼─────────────┤

│ spatial\_predictors (min) │ -3.179 │

└────────────────────────────────────┴─────────────┘

A graph with red dots

Description automatically generated

Plot of spatial variable importance. Same interpretation as nonspatial one. Spatial predictors references the Moran eigenvector maps used for spatial weighting.

A screenshot of a computer screen

Description automatically generated

We see from the randomforestExplainer() that most of the importance is with the spatial predictors. Interpretations are same as the nonspatial one. Spatial predictors are named spatial\_predictor\_X\_Y, where X is the neighborhood distance at which the predictor has been generated, and Y is the index of the predictor.

A graph of a graph showing a number of data

Description automatically generated with medium confidence

Response curves for each individual predictor when all others are set at 10th percentile

A graph of a graph showing the growth of a number

Description automatically generated with medium confidence

Response curves for each individual predictor when all others are set at 50th percentile

A graph of a graph of a graph

Description automatically generated with medium confidence

Response curves for each individual predictor when all others are set at 90th percentile

A graph of a number of data

Description automatically generated with medium confidence

Combined view of the response curves so you can see how they differ.

The plots basically show how they would predict arsenic levels and change the arsenic levels predictions. Able to begin to see some trends

A screen shot of a screen

Description automatically generated

Response surface for pH KCL and m%. Basically we see how arsenic changes as both pH KCL and m% change. I chose those variables because of the Spatial PCA analysis, but I can change this to any two variables that you would want. Darker values show higher Arsenic predictions, the dots represent actual observed values

A comparison of a graph

Description automatically generated

Comparison of nonspatial and spatial random forest variable importances

A screenshot of a computer

Description automatically generated

10 most important variables for spatial random forest are all spatial predictors.

A comparison of a map of the same type of distance

Description automatically generated with medium confidence

Look at the eigenvalues from the most important spatial predictors. The one on the left uses spatial\_predictor\_175\_1 which is the most important variable in the spatial random forest. The one on the right uses spatial\_predictor\_0\_1, which is the fourth most important and the first important spatial predictor that did not come from a neighborhood of distance 175.

We can change the two variables. I chose those because they were some of the most important and showed how the eigenvalues change spatially.

A graph showing the difference between the number of objects

Description automatically generated with medium confidence

The spatial predictors are included in the model one by one, in the order of their Moran’s I (spatial predictors with Moran’s I lower than 0 are removed). The selection procedure is performed by the function [select\_spatial\_predictors\_sequential()](https://blasbenito.github.io/spatialRF/reference/select_spatial_predictors_sequential.html), which finds the smaller subset of spatial predictors maximizing the model’s R squared, and minimizing the Moran’s I of the residuals. This is shown in the optimization plot above. (The dots linked by lines represent the selected spatial predictors).

The standard test based on MC is not able to detect the coexistence of positive and negative autocorrelation structures (i.e., it leads to a non-significant test). The moranNP.randtest function allows to decompose the standard MC statistic into two additive parts and thus to test for positive and negative autocorrelation separately (Dray, 2011)

class: krandtest lightkrandtest

Monte-Carlo tests

Call: moranNP.randtest(x = datum$as, listw = listwdatum, nrepet = 999,

alter = "two-sided")

Number of tests: 2

Adjustment method for multiple comparisons: none

Permutation number: 999

Test Obs Std.Obs Alter Pvalue

1 I+ 0.10821110 2.397064 greater 0.017

2 I- -0.01816488 2.927589 less 1.000

Results show that only positive autocorrelation is detected

# References

Gómez-Hernández et al.(1999). geoENV II — Geostatistics for Environmental Applications. Springer Science & Business Media.

Wagner H. H. (2003) Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. Ecology, 84, 1045–1057

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

*Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17.*[*doi:10.18637/jss.v077.i01*](doi:10.18637/jss.v077.i01)

*Blas M. Benito (2021). spatialRF: Easy Spatial Regression with Random Forest. R package version 1.1.0. doi: 10.5281/zenodo.4745208. url:*[*https://blasbenito.github.io/spatialRF/*](https://blasbenito.github.io/spatialRF/)

Hemant Ishwaran and Udaya B. Kogalur and Eiran Z. Gorodeski and Andy J. Minn and Michael S. Lauer (2010) <[doi:10.1198/jasa.2009.tm08622](https://doi.org/10.1198%2Fjasa.2009.tm08622)>, Leo Breiman (2001) <[doi:10.1023/A:1010933404324](https://doi.org/10.1023%2FA%3A1010933404324)>

Bauman, D, T Drouet, S Dray, and J Vleminckx. 2018. “Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors.” Ecography 41: 1638–49.

Borcard, D, and P Legendre. 2002. “All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices.” Ecological Modelling 153: 51–68.

Borcard, D, P Legendre, and P Drapeau. 1992. “Partialling out the spatial component of ecological variation.” Ecology 73 (3): 1045–55.

Dray, S, and A B Dufour. 2007. “The ade4 package: implementing the duality diagram for ecologists.” Journal of Statistical Software 22 (4): 1–20.

Dray, S, P Legendre, and P R Peres-Neto. 2006. “Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM).” Ecological Modelling 196: 483–93.

Dray, S. 2011. “A new perspective about Moran’s coefficient: spatial autocorrelation as a linear regression problem.” Geographical Analysis 43: 127–41.

Dray, S, S Saïd, and F Débias. 2008. “Spatial ordination of vegetation data using a generalization of Wartenberg’s multivariate spatial correlation.” Journal of Vegetation Science 19: 45–56.