# Voronoi Data Linkage: Extracting data from polygons to points

**Lucas da Cunha Godoy** [1]
**Luís Gustavo Silva e Silva** [2]
**Douglas R. Mesquita Azevedo** [3]

## 1 Introduction

According to [Jacox & Samet, 2007], a spatial join finds all pairs of two datasets of multi-dimensional objects in Euclidean space satisfying a specific relation of their spatial components, such as intersection. Usually this kind of technique is used either to extract data from points to polygons, i.e., count the number of events spatially represented by points and its coordinates within each polygon belonging to a polygons' set or combining data from different sets of polygons. When the interest is extract information from polygons to points, the most common approach consists in simply consider that points will inherit the features from the polygons which they are contained in. Although this method is reasonable, it has some drawbacks. Besides the fact that it wastes a lot of information from the polygons, it will not preserve any spatial structure present on the polygons' features. Our proposal is using the voronoi tesselation and polygons' intersection aiming not to lose information from the polygons' features. That is the reason why we call it Voronoi Data Linkage. In the next sections we will define the method formally and show how does it take advantage from all the data, incorporating the spatial structure. Furthermore, we will apply our method and the first technique mentioned, which we will call naive approach, in brazilian electoral data. Moreover, we will show that the features obtained by our method give a better performance in the predictions.

## 2 Methods

### 2.1 Voronoi Tesselation

The Voronoi Tesselation is a method with application in several science fields, such as Physics and Spatial Statistics ([Bock *et al.* , 2010, Thäle *et al.* , 2016]).

Consider a space $X$, where a distance $d$ can be computed for every pair of elements contained within this space. Let $d$ be a given distance measure, $K$ a set of indexes and

[1]DEST - UFMG. e-mail: *lucasdac.godoy@gmail.com*
[2]Datalive
[3]DEST - UFMG

$\{P_k\}_{k \in K}$ a list of points observed within $X$. A Voronoi Cell $\{R_k\}$ associated with the point $P_k$ is defined as a polygon such as for every point $x \in R_k$ the distance $d$ between $x$ and $P_k$ is smaller than the distance between $x$ and $P_j : \forall j \neq k$.

## 2.2 Voronoi Data Linkage

Suppose we have a set of points $Y = \{Y_1, ..., Y_{n_y}\}$ contained within a region $A$. In addition, let $Z = \{Z_1, ..., Z_{n_z}\}$ be a set of polygons which consist in a partition of $A$. Also, let $V = \{V_1, ..., V_{n_y}\}$ be the Voronoi cells associated with the points $Y$. Moreover, each polygon $Z_k$ has a associated vector of continuous variables $\boldsymbol{X}_k = \{X_{k,1}, ..., X_{k,p}\}$. We are interested in obtain the vector of variables $\boldsymbol{X}^*$ associated with points $Y$. Both the naive and the voronoi data linkage approaches are based on the strategy to obtain one variable of $\boldsymbol{X}$ each time. The first one will consider that the variable $X_{k,1}$ associated with the point $Y_k$ will be

$$X^*{}_{k,1} = \{X_{j,1} : Y_k \subset Z_j\}.$$

Then we have

$$\begin{aligned} E[X^*{}_{k,1}] &= E[\{X_{j,1} : Y_k \subset Z_j\}], \\ Var[X^*{}_{k,1}] &= Var[\{X_{j,1} : Y_k \subset Z_j\}]. \end{aligned} \tag{1}$$

On the other hand, our method will depend of one additional variable which is the proportion of voronoi cell $k$ covered by the polygon $j$. This variable is defined as

$$p_{j,k} = \frac{Area(Z_j \cap V_k)}{Area(V_k)}.$$

Now, the Voronoi Data Linkage is given by

$$X^*{}_{k,1} = \sum_{i=1}^{n_z} p_{i,k} X_{i,1}.$$

Hence, we can obtain the expectation and the variance of this variable easily as follows

$$\begin{aligned} E[X^*{}_{k,1}] &= \sum_{i=1}^{n_z} p_{i,k} E[X_{i,1}] \\ Var[X^*{}_{k,1}] &= \sum_{i=1}^{n_z} p_{i,k}^2 Var[X_{i,1}] + 2 \sum_{i<j} \sum p_{j,k} p_{i,k} Cov(X_{j,1}, X_{i,1}). \end{aligned} \tag{2}$$

From the equations 1 and 2, we can conclude that our method is able to incorporate

any covariance structure present on data. Furthermore, when the data is independent and identical distributed, our method to estimate such variables always has a variance smaller or equal than the variance of the estimates obtained through the naive approach.

# 3 Application: Brazilian Elections Data

In Brazil, electoral data has four main aggregation levels, they are: the state, municipality, electoral zone and electoral section. The lower level, which is the electoral sections, does not have any administrative division, i.e., it is just an address and not an area. Considering that socio-demographic variables can be useful to explain the election outcome, we are interested in aggregate such kind of data, provided by the Instituto Brasileiro de Geografia e Estatistica (IBGE) in small areas called census sectors. The data used here is from the second round of the President election in the year of 2014 at the city of São Paulo. The socio-demographic data is from the 2010 IBGE census. The variables extracted from the census sectors to the electoral sections are population, average income, household density, illiteracy rate, proportion of white people, proportion of women, proportion of people with age under 25 years old, proportion of people with age between 25 and 40, proportion of people with age between 40 and 55, and the proportion of people with age over 55 years old. The outcome of interest, present on electoral section data, is a binary variable which is 1 if the right wing party got more votes in such place and 0 otherwise. Note that, when we are considering averages or rates we must use the population, in order to transform it into a summation, when using the Voronoi Data Linkage.

# 4 Results

## 4.1 Exploratory Data Analysis

In the figure 1, we have the average income's natural logarithm plus one for the census sectors from IBGE. Clearer values indicate a higher average income. Note that, this variable seems to have a well defined spatial structure.

Figure 2 shows that in this case, where the data clearly has an autocorrelation spatial structure, the variability of the variable obtained using the Voronoi Data Linkage is similar to the naive approach.

In figures 3 and 4, it is easy to see that in the case where the variable from which we are extracting information does not have a spatial structure of covariance, our method presents a smaller variance than the other one.

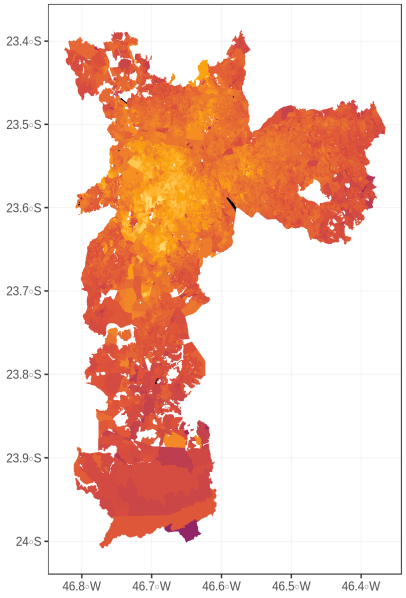


Figura 1: Average income for the IBGE data.

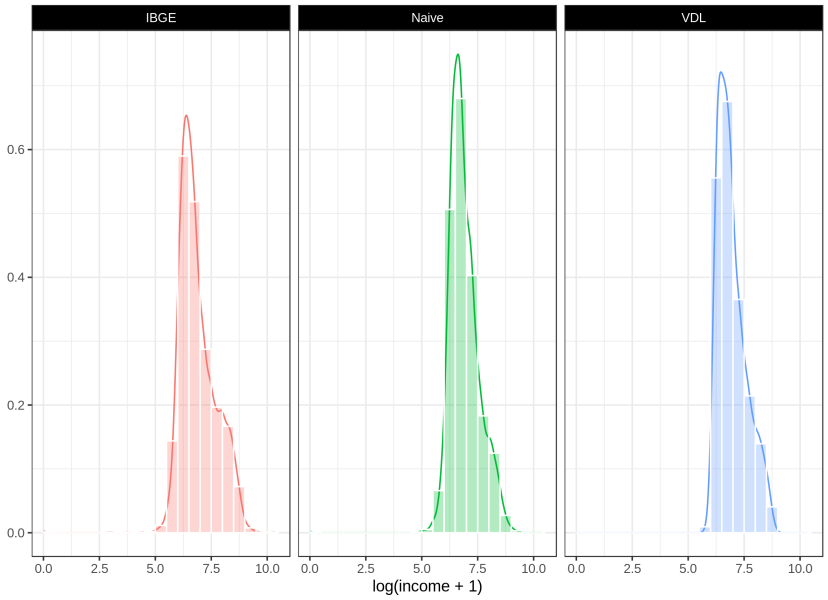


Figura 2: Average income for each dataset.

## 4.2 Predictions

For the predictions we used logistic regression and the Spatial Cross-Validation described in [Lovelace *et al.* , n.d.] with 5 folds and 100 repetitions. As we can see in figure 5, the model using the covariates obtained by our method over-performed the model using the covariates extracted using the naive approach with respect to the area under the curve.

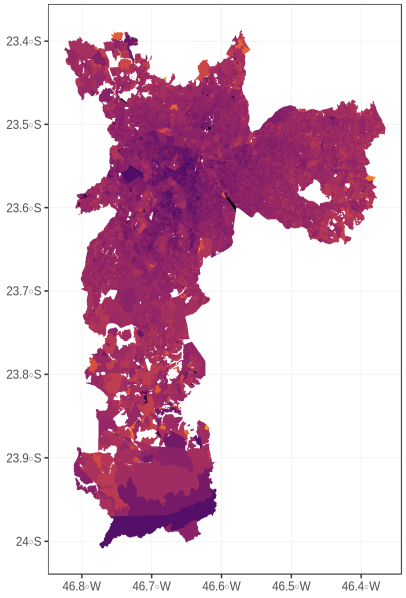


Figura 3: Household density for the IBGE data.

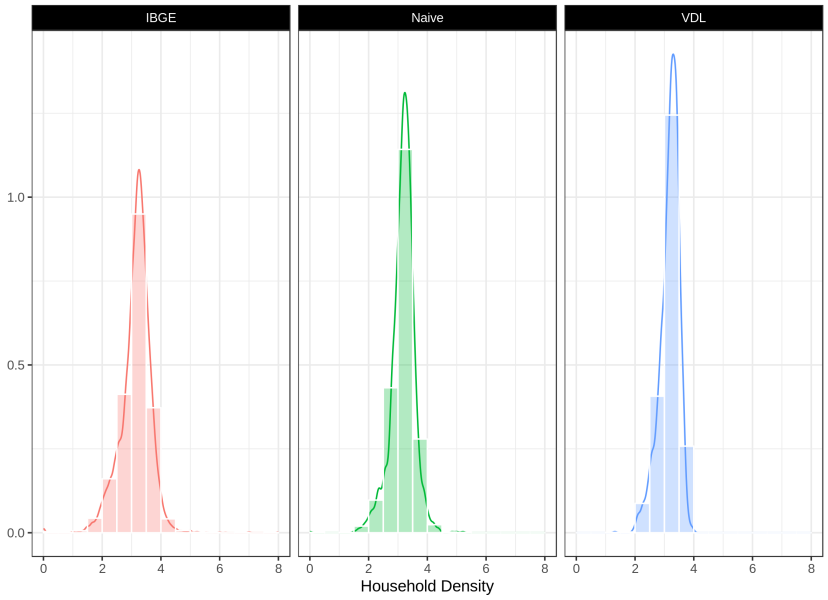


Figura 4: Household density for each dataset.

# 5 Conclusion

The method proposed presented good performance both in the sense statistical analysis and machine learning. The first because it preserves part of the spatial autocorrelation structure contained in the data, if it exists. In addition, when the data is independent and identical distributed the method has a variance smaller than the variance of the naive approach. In the context of machine learning, the cross validated results have showed that our method improves the prediction performance. Given that the approach proposed uses more information than the other one, such behavior already was expected.

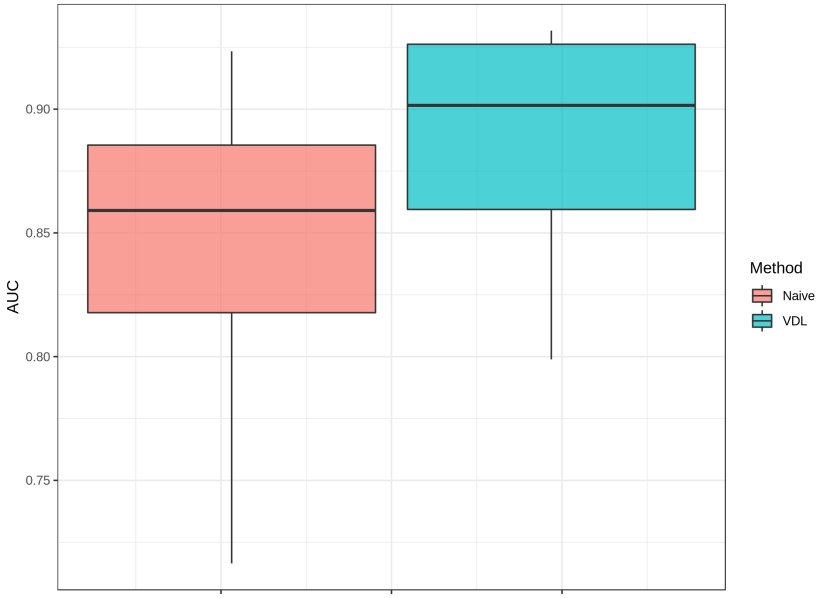


Figura 5: Area under the curve for the predictions using the covariates extracted with different methods.

# Referências

[Bock *et al.* , 2010]Bock, Martin, Tyagi, Amit Kumar, Kreft, Jan-Ulrich, & Alt, Wolfgang. 2010. Generalized voronoi tessellation as a model of two-dimensional cell tissue dynamics. *Bulletin of mathematical biology*, **72**(7), 1696–1731.

[Jacox & Samet, 2007]Jacox, Edwin H, & Samet, Hanan. 2007. Spatial join techniques. *ACM Transactions on Database Systems (TODS)*, **32**(1), 7.

[Lovelace *et al.* , n.d.]Lovelace, Robin, Nowosad, Jakub, & Muenchow, Jannes. Geocomputation with R.

[Thäle *et al.* , 2016]Thäle, Christoph, Yukich, Joseph E, *et al.* . 2016. Asymptotic theory for statistics of the Poisson–Voronoi approximation. *Bernoulli*, **22**(4), 2372–2400.