ZIP Code Versus Georeference

J.L. Bazán Guzmán [1]†, T.S. de Almeida [2], M.M. Ferreira [3], D.C.F. Guzmán [1], F. Louzada [1], M. Miranda [1], A.L. Mota [1], S. Rangel [4], C.M. Russo [1], L.A. Santos [5], M.O. Santos [2], and F. Toledo [1].

[1] *University of São Paulo, São Carlos, Brazil*
[2] *University at Buffalo SUNY, Buffalo, USA*
[3] *Federal University of Alagoas, Maceió, Brazil*
[4] *São Paulo State University, São José do Rio Preto, Brazil*
[5] *Federal Institute of Paraíba, João Pessoa, Brazil*

(*Communicated to* MIIR *on 28 June 2021*)

Study Group: VI Brazilian Study Group with Industry - Mathematics in Industry Reports (MIIR), Cambridge Open Engage

Communicated by: Jorge Luis Bazán Guzmán

Industrial Partner: StepWise (Statistical Intelligence) www.stepwise.com.br

Presenter: Felipe Marangoni

Team Members: Alex Leal Mota, University of São Paulo, São Carlos, Brazil; Cibele M. Russo, University of São Paulo, São Carlos, Brazil; Daniel Camilo Fuentes Guzman, University of São Paulo, São Carlos, Brazil; Francisco Louzada, University of São Paulo, São Carlos, Brazil; Franklina Toledo, University of São Paulo, São Carlos, Brazil; Jorge Luis Bazán Guzmán, University of São Paulo, São Carlos, Brazil; Lucas Araújo Santos, Federal Institute of Paraíba, João Pessoa, Brazil; Socorro Rangel, São Paulo State University, São José do Rio Preto, Brazil; Maristela Oliveira dos Santos, University of São Paulo, São Carlos, Brazil; Mauricio Ferreira, Federal University of Alagoas, Maceió, Brazil; Milton Miranda, University of São Paulo, São Carlos, Brazil; Thaicia Stona, University at Buffalo SUNY, Buffalo, USA; Leticia Lemos Burla, University of São Paulo, São Carlos, Brazil; Fernanda Yuka Ueno, University of São Paulo, São Carlos, Brazil; Guilherme Ramon Rodrigues da Silva, University of São Paulo, São Carlos, Brazil.

Industrial Sector: Data Analysis; Finance; Social;

Tools: K-Nearest neighbors, Refinement of Clustering using a Mathematical Optimization Model, Polygonal Spatial clustering.

Key Words: Classification; clustering; geospatial data; optimization; credit risk; machine learning; computational geometry, Information Value

MSC2020 Codes: 91G40; 91G70; 62H30; 62H35; 65K10 (see MSC2020).

† Corresponding Author: jlbazan@icmc.usp.br

Summary

When dealing with predictive modeling, focusing on the financial segment, risk management, and credit-granting (Application Scores), different types of attributes are used: Cadastral, Behavioral, Business / Proposal, Credit Bureaux, in addition to Public, Private or Subsidiaries Sources. Within the universe of cadastral attributes, examples such as Age, Income, Education, Profession, and Home or Work Address are often eligible as covariates of great discriminatory power. The Postal Address Code (**Código de Endereçamento Postal CEP in Portuguese**) in Brazil, in particular, has a unique contribution capacity (uncorrelated with most other attributes in general) and reasonably good predictive power (IV - Information Value). CEP is frequently used by truncating its numeric representation, considering the first d digits, for example.

On the other hand, when using five digits, the location is narrowed more, and a smaller number of records will present these values (low representativeness). The question: How to best use it, and what controls should be applied? CEP is not a value between 01000-000 and 99999-999, but it is not a discrete or continuous quantitative variable but a categorical one. What is more, if we consider how it is distributed geographically, we can consider it ordinal. However, their ordering is not direct, increasing, but in a snail shape, making a clear grouping strategy difficult. For this reason, when treated as a nominal category, its stability over time falls considerably, de-calibrating the model and decreasing its useful life. In this report, a preliminary methodology is proposed, aiming to elaborate clustering sets of CEPs by considering the information of clients' defaults over a period of time. Additionally, we tested the number of clusters obtained using the Information Value criterion. Promising solutions are obtained using statistical and optimizing approaches. Additionally, other methodologies are suggested and could be complementary with the principal methodology proposed. Final remarks and suggestions are also included.

1 Introduction

The Postal Address Code (**Código de Endereçamento Postal CEP in Portuguese**) in Brazil, with a structure of 5 (five) digits, was created by the Brazilian company of Posts and Telegraphs, in May 1971. Since 1992 the **CEP** is a numerical set consisting of 8 (eight) digits. The main purpose is to guide and accelerate the routing, treatment, and distribution of correspondence objects by assigning them to localities, public places, post offices, services, public agencies, companies, and buildings.

StepWise (Statistical Intelligence) observed that **CEP** has a unique contribution capacity in predictive modeling in the financial segment, risk management, and credit granting. **CEP** is not correlated with most of the other attributes usually considered and has a high discriminatory power defined by the Information Value indicator (IV)[1].

The proposed challenge, named as ***CEP vs Georeference***, consists of formulating **CEP** groupings using georeferential information and non-default information from a set of clients evaluated between 2016 and 2018 in Brazil. Further details are explained in the following subsection.

The document is organized as follows. In Section 2, we describe the problem, including the definition and the steps of our approach to solving the problem. In section 3, we give Step 1 of our approach, and then we describe the database, and we show some exploratory analysis considering the main variables. In Section 4, we show Step 2 of our approach. We propose a methodology in three steps: K-Nearest neighbors by default, Clustering Visualization, and Refinement considering a Mathematical Optimization Model. Additionally, two methodologies: Polygonal Spatial Clustering and VOroni-Base generated polygons and stability, are discussed. In Section 5, we discuss Model Validation considering Information Value and Complementary variables.

Finally, in the last section, we give some remarks and suggestions.

2 The problem

2.1 Problem Definition

The proposed problem consists of grouping sets of **CEP's** based on their geographical position, using coordinates (latitude and longitude). From a historical mass of data at the level of customers between June 2016 and November 2018, it is required to form georeferential polygons that combine sets of **CEPs** with high similarity within the polygon and high divergence between the polygons taking into account the default information. Whether or not during that period.

Among the conditions required for the solution, initially formulated by StepWise, is that the polygons to be formed cannot intersect each other. Besides, their training must be oriented to the characteristic of interest (default 1 "vs." default 0), and their "representativeness," defined by the proportion of individuals in the cluster concerning the analyzed data, must be at least 3%. However, this condition can be revised, and alternatives can be discussed.

It was also expected to validate the proposed grouping or cluster formation using information about a Score proposed by the company (Credit Score StepWise for Individuals - Eventual Retail).

2.2 Steps of the Methodology

In order to develop a methodology for the problem, we propose the following steps:

- Data set definition (**CEP** vs. Customers) and Exploratory Analysis. In this step, we give some information about the database considered in this study, and we show some exploratory analysis for the main variables in the database.
- Methodology for creating clusters. In this step, we propose different methodologies to create a cluster using geolocation variables by considering information about default.
- Model validation. In this step, we discuss some criteria for the selection of the clusters proposed in the previous step.

3 Step 1: Databases and Exploratory Analysis

The following databases were received.

(1) Main database with important information about geolocation and the customers' failure to pay.

This database has 400,376 registers from 110,820 CEPs in the country. Restricting them to the entries related to the state of São Paulo, there are 333,258 registers from 87,508 CEPs. This database has information about Customers, Time of collection of data, ZIP Code (CEP), Fault, Credit Score, and Georeferencial information.

(2) Database with complementary geolocation information at the postal code level.

This database has 110820 registers and 54 attributes.

(3) Database with economic information at the postal code level.

This database has 110820 registers and seven attributes

The main database was used for the methodology described in Section 4. Additionally, a new database was created, considering the three databases received were unified. This database with 110820 registers and 76 attributes has economic, geolocation, and customer information at the postal code level.

3.1 Insolvency fraction

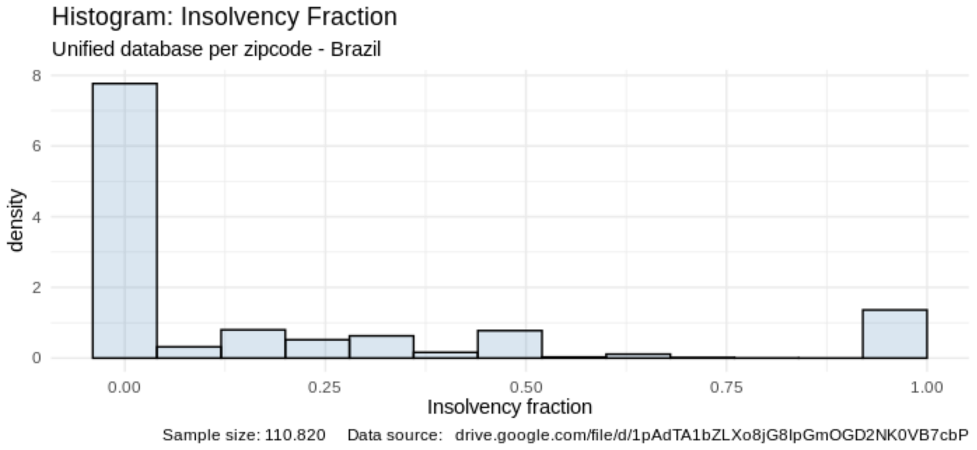
In the following figure, we show the distribution of Insolvency Fraction by ZIP Code for the complete data. Insolvency Fraction is obtained using the fault condition of the consumers with the same ZIP Code.

We observe (1) that the insolvency fraction by ZIP Code has a distribution with a concentration in Non-Fault and Fault.

3.2 Score

Now we show the distribution of the Score variable for the complete data and São Paulo.

We observe (2) that the population of São Paulo has a similar performance to Brazil's. Thus São Paulo database will be selected for the following analysis.

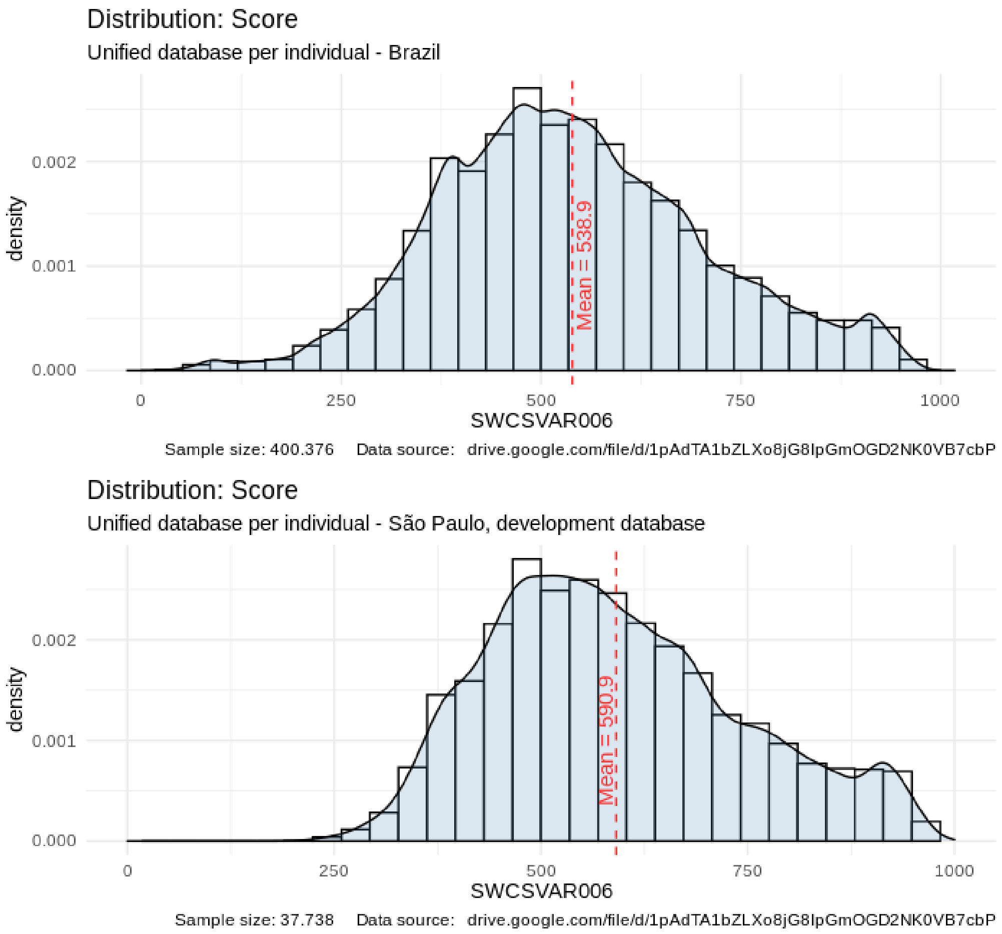


Histogram: Insolvency Fraction

Unified database per zipcode - Brazil

Sample size: 110.820 Data source: drive.google.com/file/d/1pAdTA1bZLXo8jG8IpGmOGD2NK0VB7cbP

Figure 1. Distribution of Insolvency Fraction by ZIP Code for the complete database



Distribution: Score

Unified database per individual - Brazil

Sample size: 400.376 Data source: drive.google.com/file/d/1pAdTA1bZLXo8jG8IpGmOGD2NK0VB7cbP

Distribution: Score

Unified database per individual - São Paulo, development database

Sample size: 37.738 Data source: drive.google.com/file/d/1pAdTA1bZLXo8jG8IpGmOGD2NK0VB7cbP

Figure 2. Distribution of the Score variable for the complete data and São Paulo

4 Step 2. Methodology for creating Clusters

4.1 K-Nearest neighbors by Default

In this proposal, an unsupervised algorithm to find clustering sets is developed, aiming to select groups of addresses by their latitude and longitude values, considering the neighbors' observed default. This method adapts the well-known K-Nearest Neighbors method (see, for instance, [5]). The procedure for creating groups from N observations starts with N clusters. In each step, given an integer K, the K nearest neighbors of each cluster are identified, and the observed default points to the new cluster classification that the original cluster will merge into. For the specific application, the user must set the population representativeness that each cluster must contain, the number K of neighbors that must be selected, and the distance criterion ε to decide whether the two clusters must merge in each step or not. A basic proposal is described below.

Algorithm

(1) Define the N initial clusters as the N observations;

(2) For each cluster, check its representativeness. If the cluster i is representative, do not merge it with any other cluster. Else, find its nearest K neighbors;

(3) Among the selected neighbors, select the one with the nearest default: this will be the candidate cluster to merge;

(4) Compute the default difference of the two clusters. If it is smaller than ε, merge them into a new cluster.

(5) Repeat Steps 2-4 until all the clusters are representative or when a stop criterion is reached.

4.1.1 *Application*

For the CEP vs. Georeference challenge, 27536 distinct CEPs from São Paulo city were initially considered. 1711 CEPs were excluded from the analysis since the non-default observations were 0, remaining distinct 25825 CEPs.

As suggested by Stepwise, only the training dataset was considered for the clustering procedure, remaining 8220 distinct CEPs to create the clustering sets.

Representativeness was set to 3%, which means that each cluster ideally has at least 3% of the total population of São Paulo city.

The stop criterion for the algorithm was set to 10 iterations. The values $\varepsilon = 0.4$ and $K = 10$ led to the best information value, with 33 clusters (see Figure 3), when compared to the other configurations considered ($p \in \{3, 5, 7, 10\}$ and $\varepsilon \in \{0.2, 0.3, 0.4, 0.5\}$). This algorithm was developed in R package [7]. Others solutions will be explained later.

By considering the algorithm described above and implemented in R, different solutions could be proposed if some algorithms' criteria are modified.

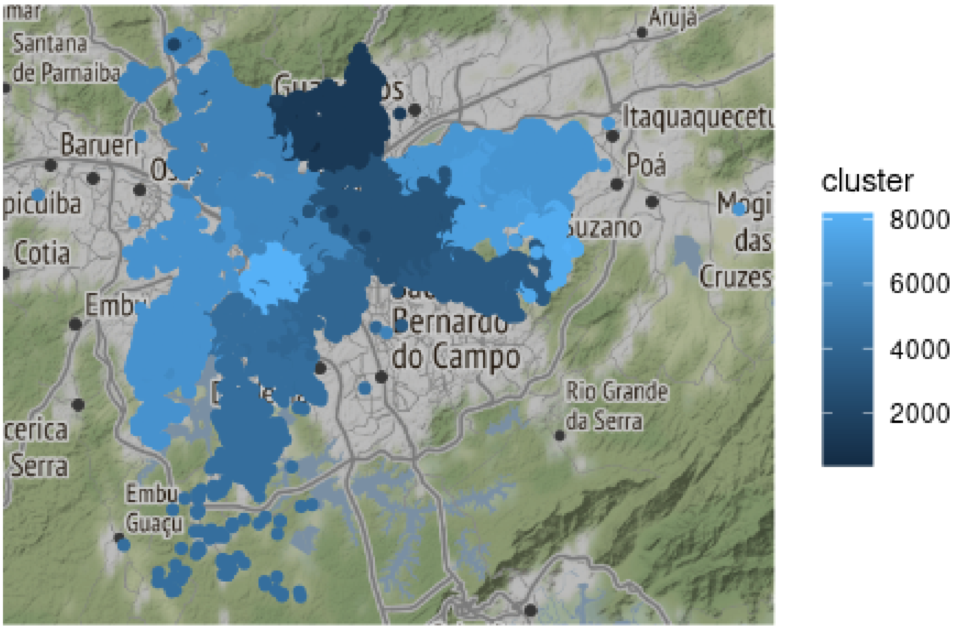


Figure 3. Nearest neighbors by default, with 33 clusters obtained

4.2 Clustering Visualization

As a proposal for visualizing the clusters, we consider the division of clusters by CEPs on the training dataset obtained in the previous section, and we consider the following geometries:

- If the cluster has only one CEP, the cluster will be represented by only this point.
- In clusters with two CEPs, the line segment joining these two CEPs will represent this cluster.
- Finally, in clusters with three or more CEPs, the convex hull of this set of CEPs was considered.

In all cases, we identify a CEP geometrically as a point through its latitude and longitude. The algorithm for visualization (4) was developed using shapely and folium packages in Python.

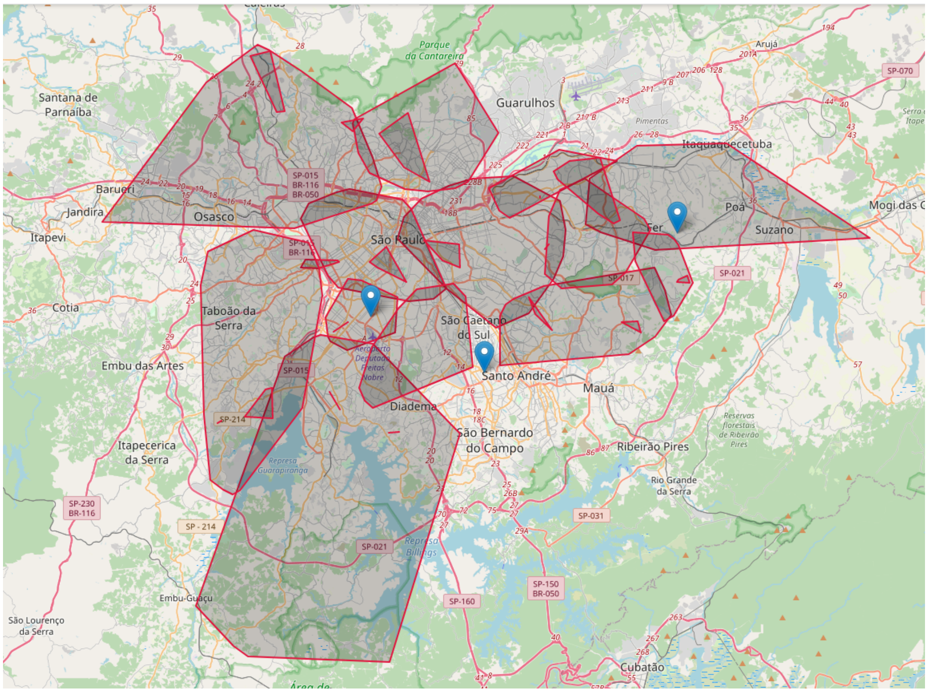


Figure 4. Visualization of the polygons obtained by the clustering by default method, with 33 clusters obtained (preliminary results)

4.3 Refinement of Clustering using a Mathematical Optimization Model

Several situations can be more widely studied if represented through models that capture their main elements. A mathematical optimization model involves the representation of a problem or situation through a set of mathematical relationships such as equations, inequalities, logical dependencies, and functions (*e.g.* [20, 21, 19]). There are several reasons for building optimization models, among which we can highlight: increase the degree of understanding of the situation studied; analyze the situation and propose solutions that are not apparent; test various scenarios that would not otherwise be possible, or recommended [9, 8]. Also, there are several general-purpose systems capable of finding the optimal solution, or at least a feasible solution with a certificate of quality in a reasonable amount of computational resources.

In this section, a mathematical optimization model to represent the problem addressed in challenge ***CEP vs. Georeference*** is proposed. It is based on models given in the literature for location problems [10, 11, 8, 13].

Suppose there are N customers (*e.g.*: persons, **CEP** code, centroids) and K groups (clusters). The indexes, parameters, and decision variables necessary to represent the problem are defined in what follows.

Indexes and Sets

- $i, j \in \mathcal{N} = \{1...N\}$ - customers or clients
- $k \in \mathcal{K} = \{1...K\}$ - groups

Parameters

- s_i: Client status i (1 default, 0 if not in default)
- b_i: customer i default level
- p: percentage of minimum representativeness to be reached by each group
- $dInd_{ij}$: difference between defaults of two customers (abs(b_i - b_j)).

Decision Variables

- $x_{ik} = 1$ if client i is allocated to group k, $= 0$ c.c.
- $y_k = 1$ if group k is created, $= 0$ c.c.
- $z =$ maximum default

Several criteria can be used to guide the decision process. In what follows, we suggest some. The decision-maker might solve the problem considering a multiobjective point of view [12] or using a mono objective approach [10]. In the latter case, the criterion that better represent the proposed scenario should be chosen.

Optimization Criteria

- Minimize the total distance between pairs of customers in the same group:

$$\textstyle\sum_{k,i,j} d_{ij} x_{ik} x_{jk}. \tag{4.1}$$

- Minimize the total number of groups:

$$\textstyle\sum_k y_k. \tag{4.2}$$

- Minimize the difference in total defaults between groups:

$$z. \tag{4.3}$$

- Minimize the number of groups considering that two customers may be in the same group if they are geografically close and if the default is similar:

$$\textstyle\sum_{k,i,j} d_{ij} x_{ik} x_{jk} + \sum_{k,i,j} dInd_{ij} x_{ik} x_{jk}. \tag{4.4}$$

A set of constraints should be considered when solving this problem. Below we define some of them.

Constraints

- Each customer is assigned to exactly one group:

$$\textstyle\sum_k x_{ik} = 1, \;\; i = 1...N. \tag{4.5}$$

- Customer i is only assigned to group k if it is created:

$$x_{ik} \leq y_k, \;\; i = 1...N, k = 1...K. \tag{4.6}$$

- Cluster representativeness. The number of defaulting customers in a group must be a percentage (p) of the database:

$$\textstyle\sum_i s_i \cdot x_{ik} \geq p \; N y_k, \;\; k = 1...K. \tag{4.7}$$

- Balance between clusters:

$$\textstyle\sum_i b_i \cdot x_{ik} \leq z, \;\; k = 1...K. \tag{4.8}$$

For this preliminary study, a mono-objective approach using as the optimization criteria (4.1) is employed. Also, only a subset the constraints defined above will be considered, constraints (4.5), (4.6), (4.7). A summary of the mathematical optimization model (OTM1) is given in (4.9).

Model Summary - OTM1

$$\begin{array}{ll} \textit{minimize} & \sum_{k,i,j} d_{ij} x_{ik} x_{jk} \\ \text{subject to:} & \end{array}$$

$$\begin{array}{rl} \sum_k x_{ik} = 1, & i = 1...N \\ x_{ik} \leq y_k & i = 1...N; k = 1...K \\ \sum_i s_i \cdot x_{ik} \geq p\ N\ y_k, & k = 1...K \\ x_{i,k} \in \{0,1\} & i = 1...N; k = 1...K \\ y_k \in \{0,1\} & k = 1...K, \end{array} \tag{4.9}$$

4.3.1 *Application*

The mathematical optimization model was implemented using the Python programming Language and solved using the general-purpose solver Gurobi [14].

The model (OTM1) was used to verify if it was possible to refine the solution generated applying the proposal described in Section 4.1. Figure 5 shows a solution with 86 clusters defined using the technique ”K-Nearest neighbors by Default”. This instance of the model OTM1 has 314330 quadratic objective terms and 7482 binary variables. The constraint's matrix has 14878 rows, 14792 columns, and 358362 nonzeros. A time limit of 30 seconds was defined to solve it. A feasible solution with an optimality gap of 99.82% was found with 14 clusters, as shown in Figure 6.

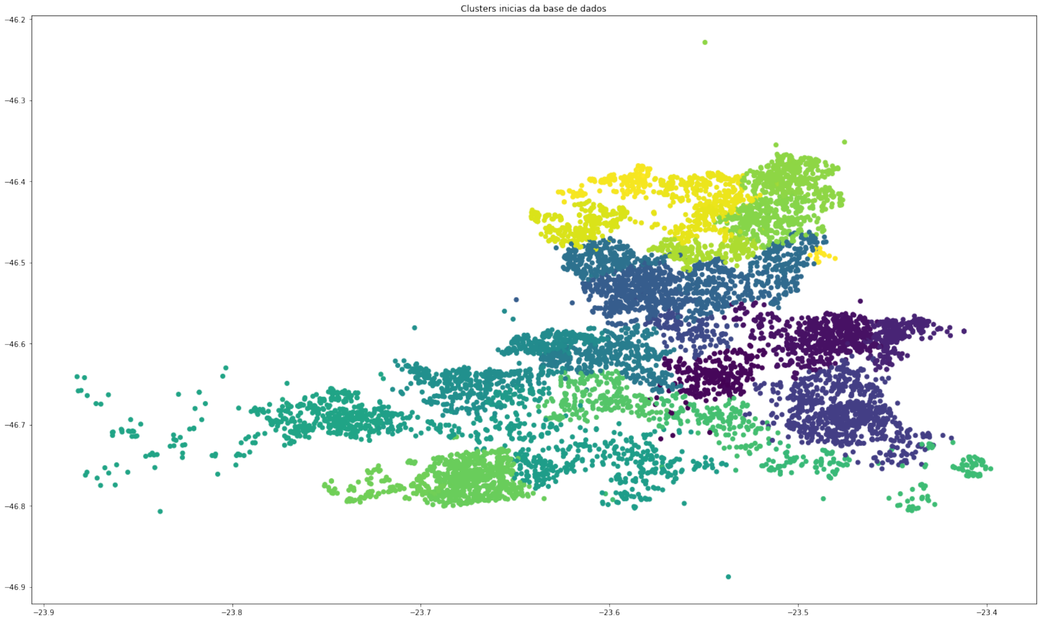


Figure 5. The 86 Clusters defined by the ”K-Nearest neighbors by Default”

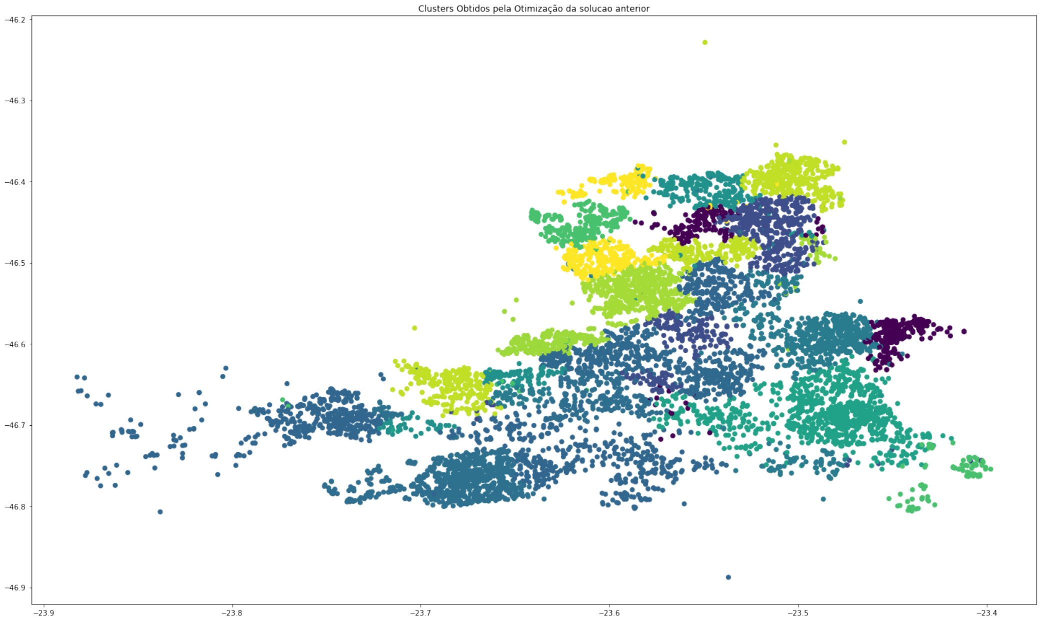


Figure 6. A solution with 14 Clusters obtained using Model OTM1 to refine the solution with 86 clusters

4.4 Polygonal Spatial Clustering

In this section, we present the result of the methodology based on the agglutination of ZIP Code. This approach groups the IDs together and then uses the Bayesian classifier to assign classifications to the ZIP Codes.

We can see in Figures 7 and 8 that the distribution of customers' defaults is practically uniform throughout the municipality of São Paulo. This feature makes spatial clustering more difficult, as there is no well-defined spatial pattern. Thus, it is not possible to build homogeneous regions. However, it is possible to clustering ID by ID and then uses the Bayes classifier to determine spatial polygons. These results are shown in Figures 9, 10, 11, and 12.

We can notice that the *dbscan* method for spatial clusters showed overall results with few groups but without concentration of groups in regions. This characteristic makes division into groups more difficult, as in this case, we will determine the classification of a region, or a new observation, based on the mode of the clusters observed in this region or closer to the new observation. On the other hand, the *k-means* method has a low number of groups, as does *dbscan*, but we can now separate it into two very distinct regions. The first encompasses the entire east and part of the north and is characterized by high customer default. The second region encompasses the western, southern, and part of the northern region and is characterized by the dispersion of defaults, even though it is present throughout the occupied territory, with medium and high concentrations.

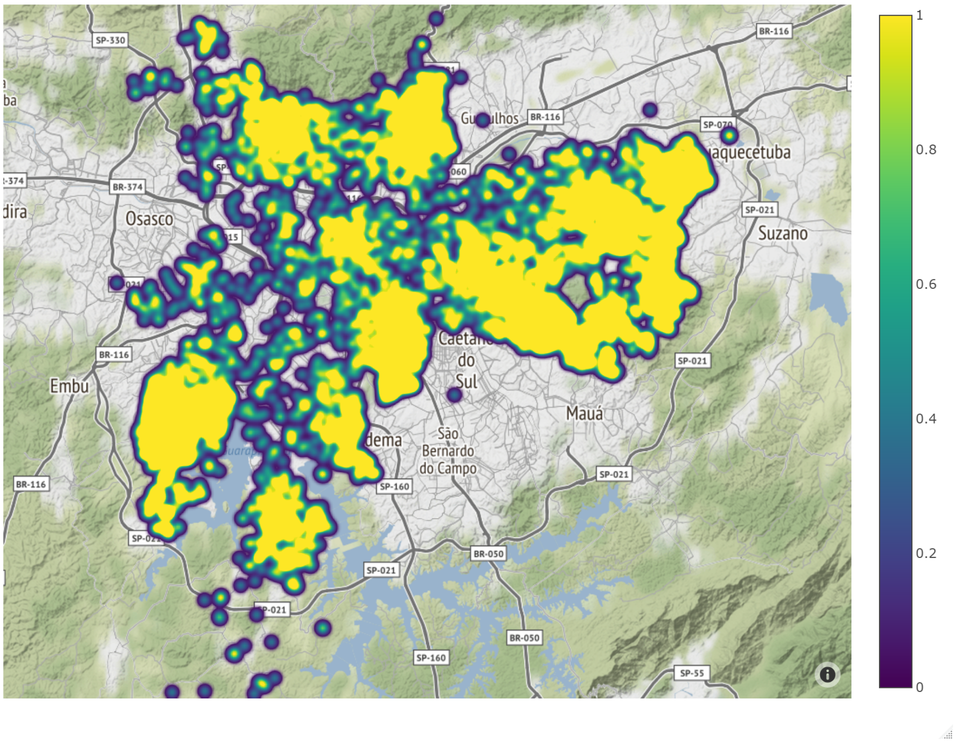


Figure 7. Spatial distribution of Defaulters customers

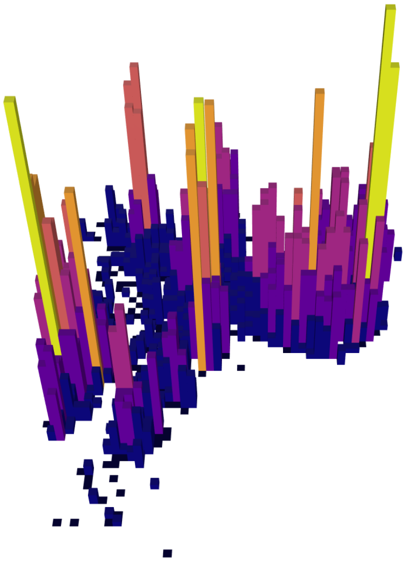


Figure 8. Bar graph of the spatial distribution of defaulting customers

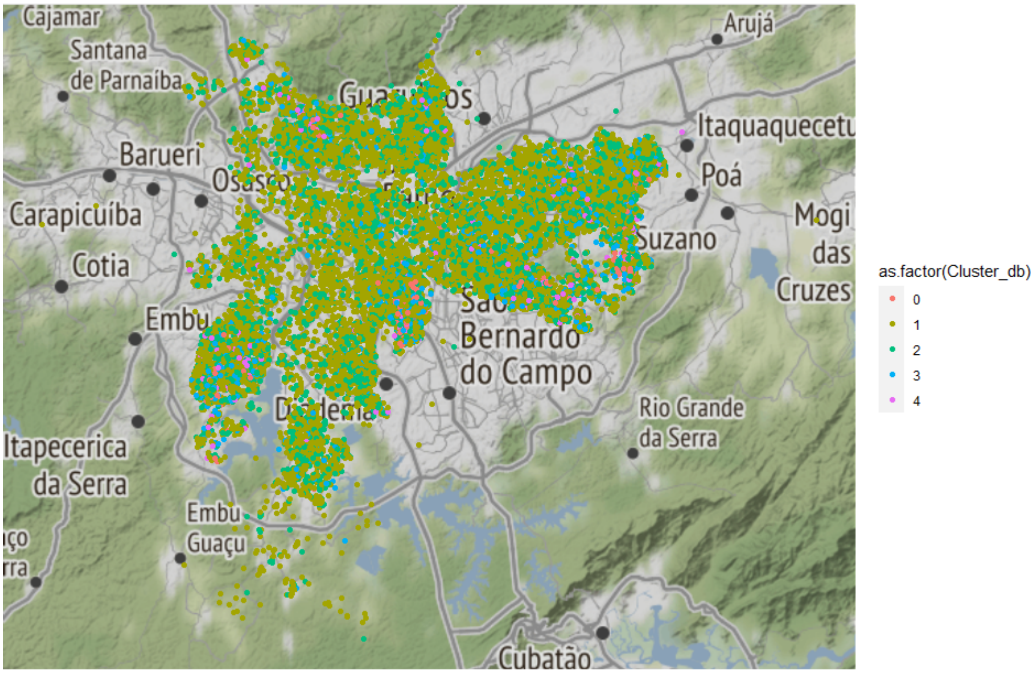


Figure 9. Cluster of ID obtained by the dbscan method

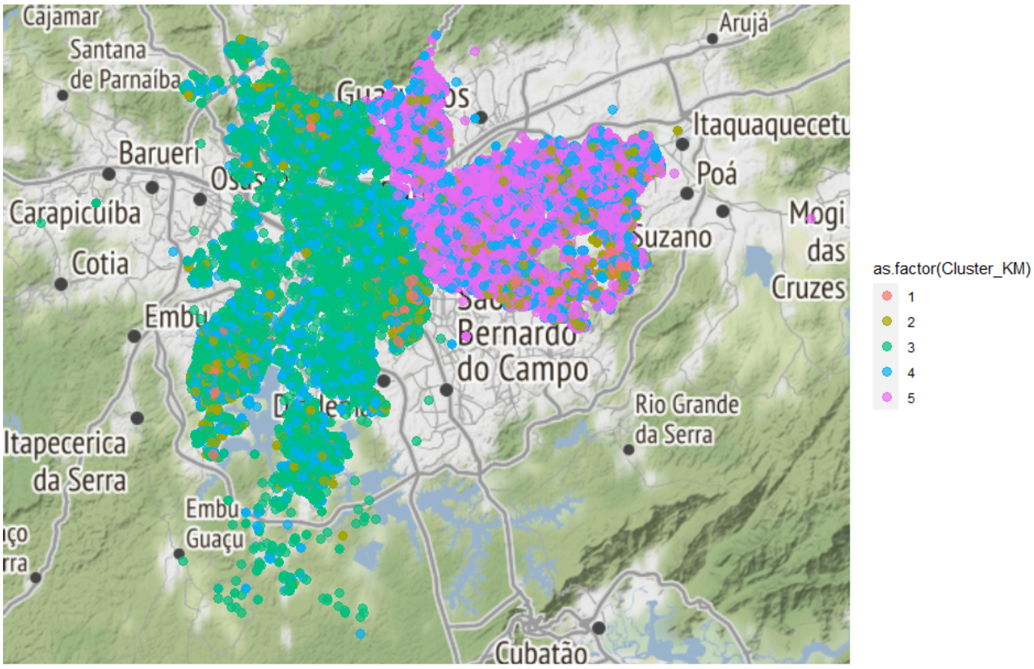


Figure 10. Cluster of ID obtained by the kmeans method

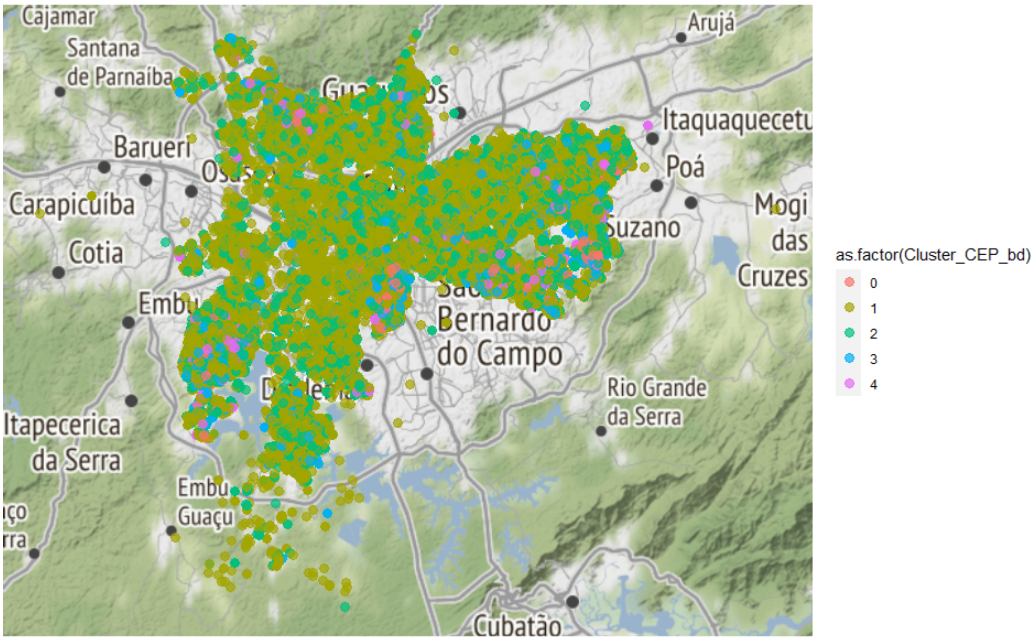


Figure 11. Cluster of ZIP Code obtained by the dbscan method

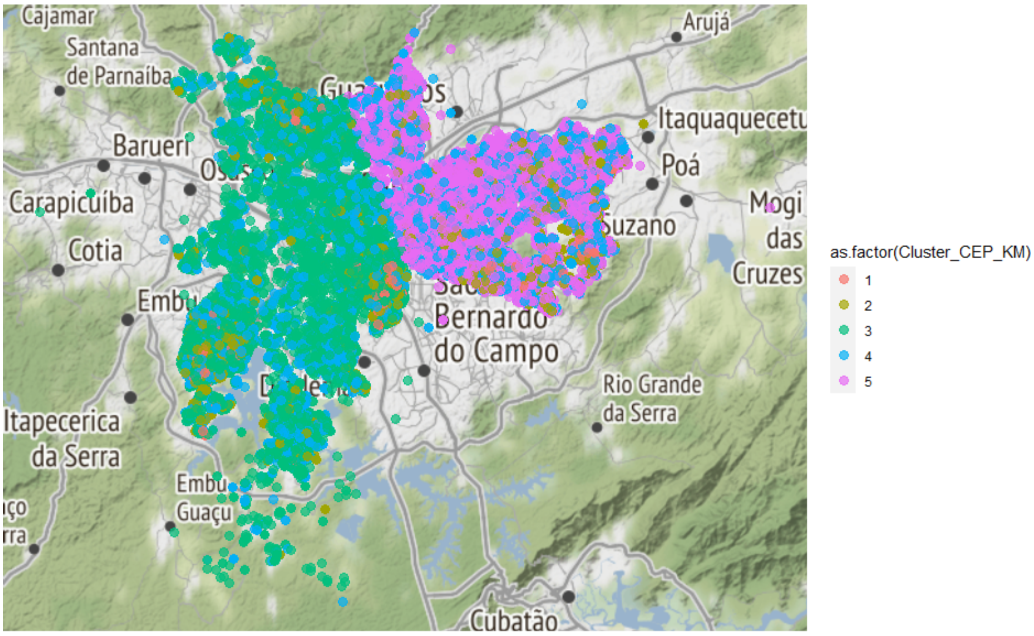


Figure 12. Cluster of ZIP Code obtained by the kmeans method

5 Step 3: Model Validation

An essential issue in our study is deciding how to select the best proposal for the formation of clusters meeting the company's requirements. For this, different approaches can be considered.

5.1 Using Information Value

In this study, we used the criteria Weight of evidence (WOE) in order to obtain the value of information (IV), which are effective techniques to perform the transformation and selection of variables and are widely used in credit scores to measure the separation between good customers it is bad for having several advantages such as dealing with missing values and outliers [16]. For example, the WOE transformation is based on the logarithmic value of the distributions. Therefore, it is aligned with the output function of a logistic regression without the need for dummy variables [17]. Besides, the IV value can be used to select variables quickly.

To calculate both WOE and IV below, we present definitions and an algorithm scheme that allows one to calculate it [18] easily.

$$\boldsymbol{WOE} = \log\left(\frac{\text{Distribution of goob}}{\text{Distribution of bad}}\right)$$

$$\boldsymbol{IV} = \sum\left([\text{Distribution Goob} - \text{Distribution Bad}] * WOE\right)$$

Precisely, to compute the Information Value (IV) for our problem, a table is created with the following values. We will consider that K is the total number of clusters.

- Step 1: Fix the number of clusters K and group the database from them and total number of clients.
- Step 2: Set n_k the number of clients in each K.
- Step 4: Set $s0_k$ the number of clients with default value equal to 0 ('good payers') in cluster, $k = 1...K$
- Step 4: Set $s0_k$ the number of clients with default value in each cluster equal to 0 ('good payers') in cluster, $k = 1...K$
- Step 5: $p1_k = \dfrac{s1_k}{n_k}$, $k = 1...K$
- Step 6: Calculate the distribution of customers in each group, $p2_k = \dfrac{n_k}{\sum_k (n_k)}$, $k = 1...K$
- Step 7: Calculate distribution of bad payers, $p3_k = \dfrac{s1_k}{\sum_k (s1_k)}$, $k = 1...K$
- Step 8: Calculate distribution of good payers, $p4_k = \dfrac{s0_k}{\sum_k (s0_k)}$, $k = 1...K$
- Step 9: Calculate Weight of Evidence (WOE), $EW_k = \ln \dfrac{p3_k}{p4_k}$
- Step 10: Calculate difference between the distribution of good and bad payers, $d1_k = p3_k - p4_k$, $k = 1...K$
- Step 11: $d2_k = 10 \cdot d1_k$, $k = 1...K$

[ht]

Table 1. The Information Value for obtained clustering sets

Solution Technique	Clusters	***IV***
DBSCAN	2	0.0150
K-means	2	0.0140
KNN by Default 1	19	0.0152
KNN by Default 2	33	0.0883
KNN by Default 3	55	0.1030
KNN by Default + Optimization 1	17	0.0135
KNN by Default + Optimization 2	19	0.0152
KNN by Default + Optimization 3	33	0.0100

The information value is then computed according to (5.1).

$$IV = \sum_k d2_k. \tag{5.1}$$

The Information Value was computed for several solutions obtained applying the techniques described in Section 4. The results are presented in Table 1 .

Table 1 shows that it is possible to test different proposals using IV. Preliminary results suggest that a low number of clusters have lower IV values, and a high number of clusters is associated with a higher value of IV, but it is no true in general. The additional analysis must be considered. The best solutions at the moment are obtained considering between 33 and 55 clusters. It is essential to notice that these two last clustering sets do not respect the representativeness criterion since they were obtained when the algorithm reached ten iterations.

5.2 Dimensionality reduction in Complementary information

Initially, a study of dimensionality reduction was developed, with leveraged methods of applied Mathematics and data mining [2]. The data considered correspond to 76 geospatial and economic features for the different ZIP codes of Brazil. Specifically, we consider the following strategies: Selection of features through Dimensionality Reduction: Principal Component Analysis, Correlation Heatmap; Selection of a new coordinate system at the appropriate level of CEP resolution. Some of **R** [7] tools adopted were packages *FactoMineR* and *factoextra.*

5.2.1 *Principal Component Analysis for Complementary Attributes*

A subset of the raw database containing 76 geospatial and economic features was selected; 46 quantitative features were processed with PCA and clustered correlations heatmap.

The selected subset requires 7 dimensions to explain $\sim 66\%$ of the variability, which indicates that feature selection by dimensionality reduction might be viable with linearization methods. Next we search for the correlated feature clusters with a clustered heatmap of correlation. Similar results were found for the subset "city = Sao Paulo", development sample:

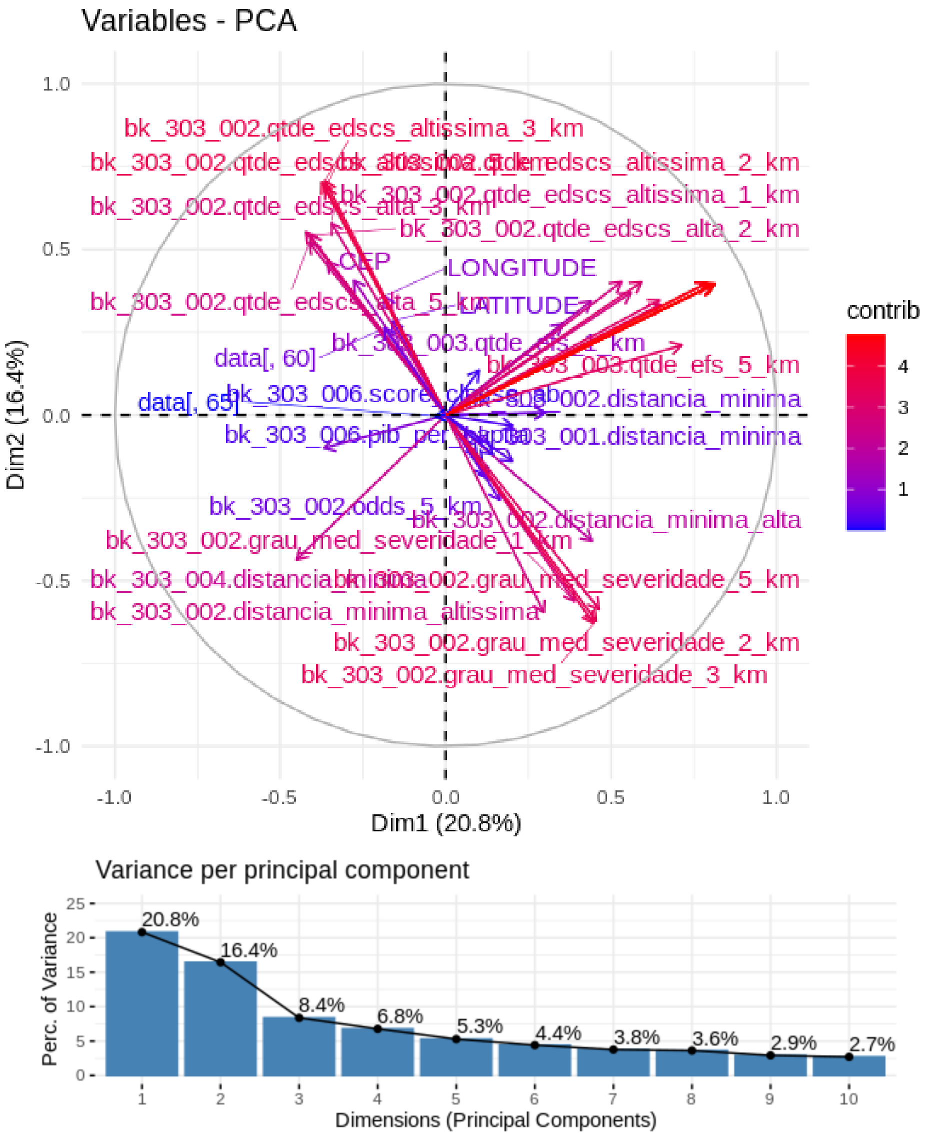


Figure 13. Principal component analysis for 46 geospatial and economic features

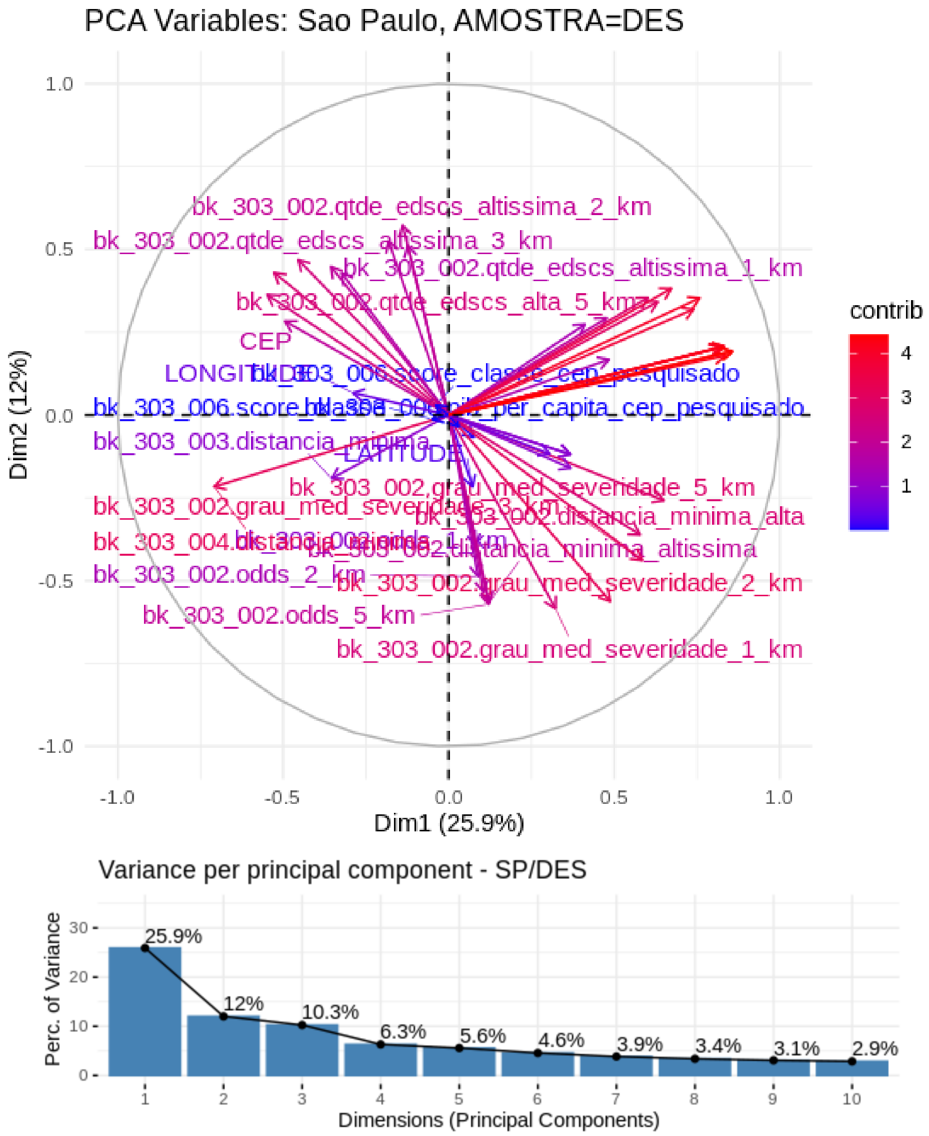


Figure 14. Principal component analysis for Sao Paulo subset

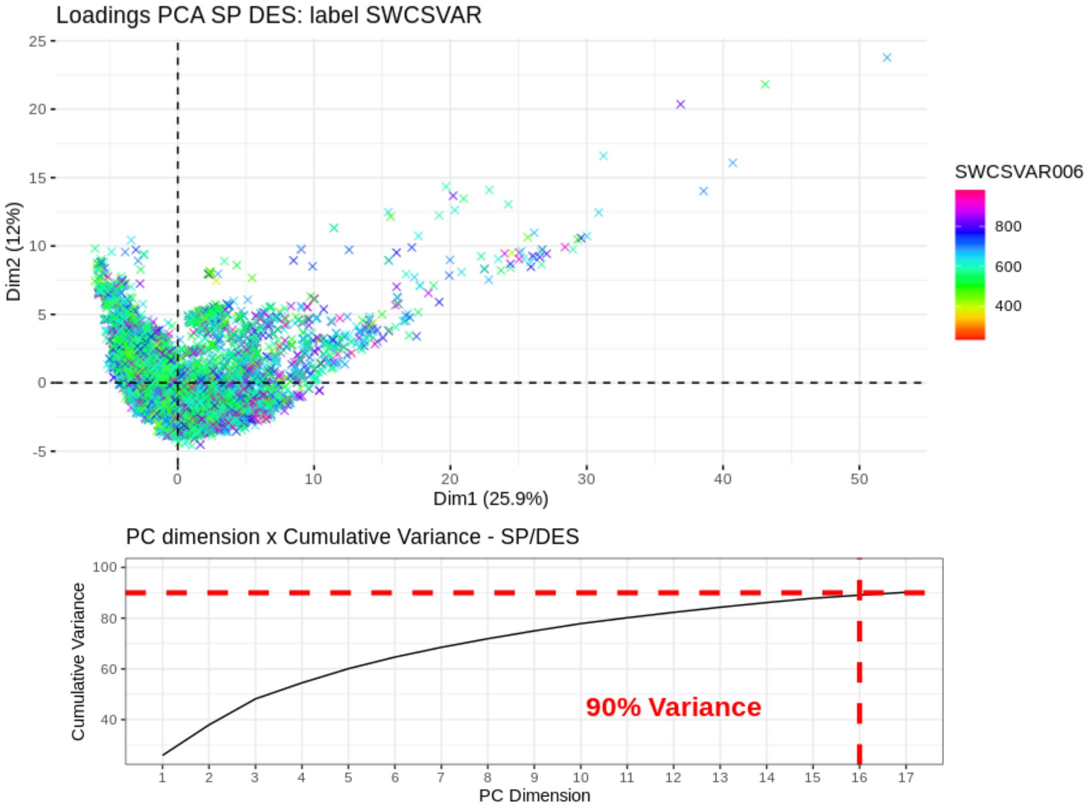


Figure 15. PCA loadings for Sao Paulo subset and PC cumulative variance

5.2.2 *Clustered Correlogram Heatmap for Complementaries Atributes*

It is possible to visualize seven main blocks of correlated features in the heatmap. Each block can be parametrized by one best representative feature, e.g. *"mean income"* parametrizes the block of economic features.

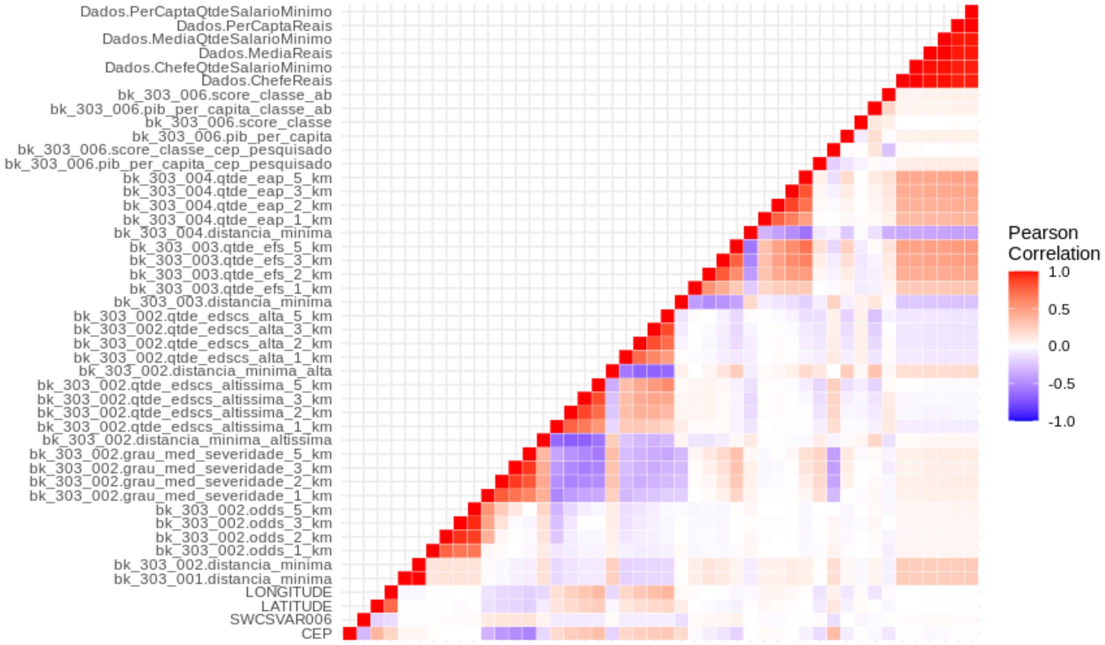


Figure 16. Clustered correlation heatmap forthe 46 variables, complete data

The seven blocks 17 contain, respectively (from left to right on the heatmap) 4, 4, 5, 5, 6, 4, and 6 variables, a total of 34 variables that can be represented by 7 of them, a linear reduction of 27 dimensions. Considering the smallest block of 2 highly correlated variables, "DIST" ($cor = 100\%$), a total of 28 dimensions could be removed.

It is possible to expect a minimal number of 8 representative variables, or a maximum number of 46 - 28 = 18 variables: the ten independent variables plus the eight representative variables.

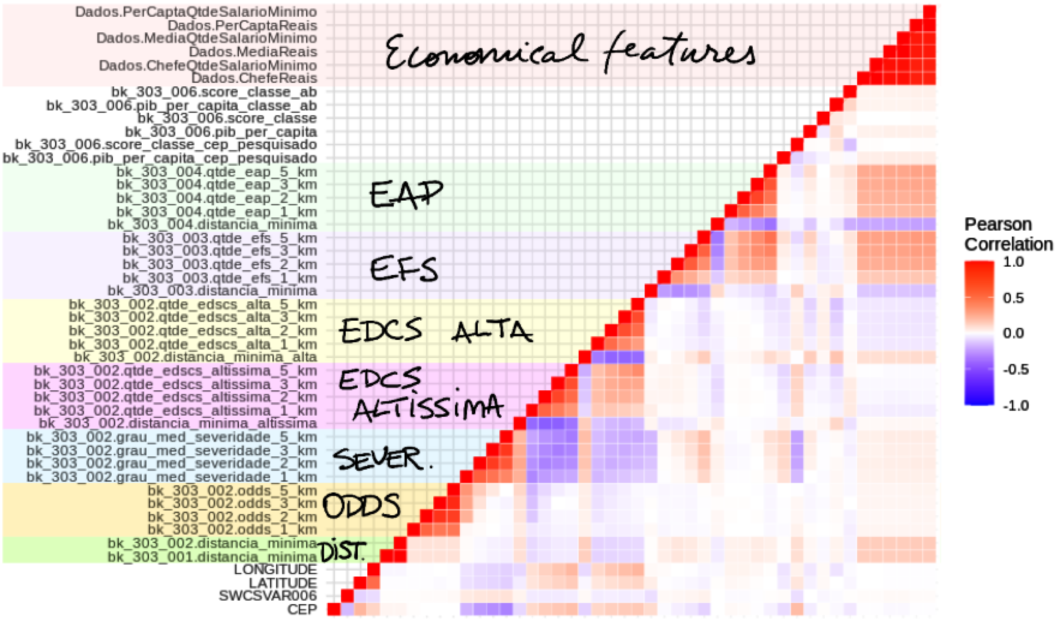


Figure 17. Annotated heatmap with the seven main blocks highlighted.

The cumulative variance plot extracted from the PC analysis confirms the dimensionality found on the heatmap: at least 18 PCs are required to span 90% or more variance observed.

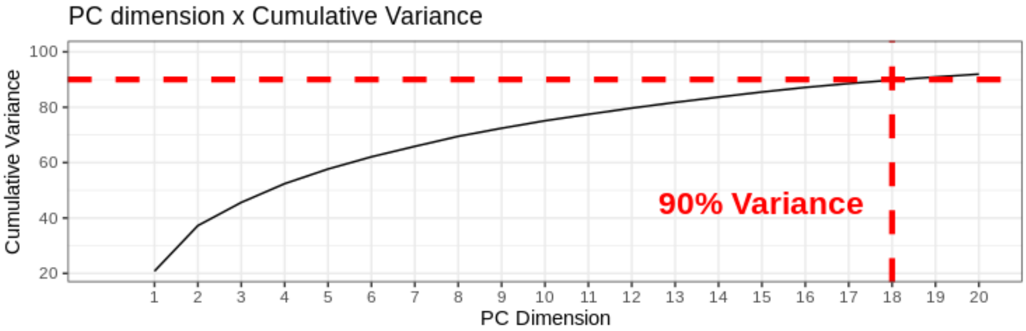


Figure 18. Cumulative variance plot

5.3 Proposed future works: Voronoi-based generated polygons and stability

An alternative solution to overlapping polygons, missing classification, and stability of polygons is to generate a partition of the planar region studied. The method of Voronoi tesselation, or partitions, can be employed as a starting point in this endeavor[2, 3, 6].

The Voronoi tesselation in an ordered, periodic set of points (e.g., a crystal lattice) is a partition whose cells are centered at individual points, with symmetry groups coinciding with the stabilizer groups of the lattice points.

Def. [Voronoi algorithm] Given $M = m_1, \dots, m_k \subset \Re^n$, represent $\Re^n$ as the union of cells V_i, $i = 1, \dots, k$, where V_i is defined by all points $x \in \Re^n$ which are not more distant from m_i than from all other points m_j. The point m_i is called the center of the V_i cell. The family of cells V_{ii} is called a Voronoi partition.

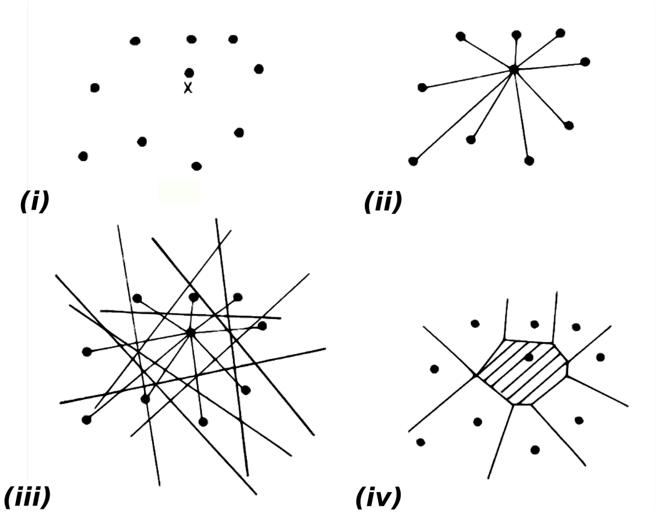


Figure 19. Building Voronoi cells in the plane: for each pair of distances from the reference point (ii), the orthogonal lines crossing the midpoint of distances define an intersection of halfplanes (iii), which is the resulting Voronoi cell (iv). Figure adapted from [3]

In the case of random point data, the stability of such partition can be enhanced with the Voronoi-derived method of centroidal Voronoi tesselation.

[Centroidal Voronoi] The following protocol is known as CVT, Centroidal Voronoi Tesselation:

```
DO
  ○ generate Voronoi
  ○ compute centroids
  ○ update centroids as new cell generators
WHILE   centroid != generator
```

CVT can be computationally demanding, requiring adaptations in order to make the method viable in this context. Suggested R [7] tools: packages *ggvoronoi*, *deldir*.

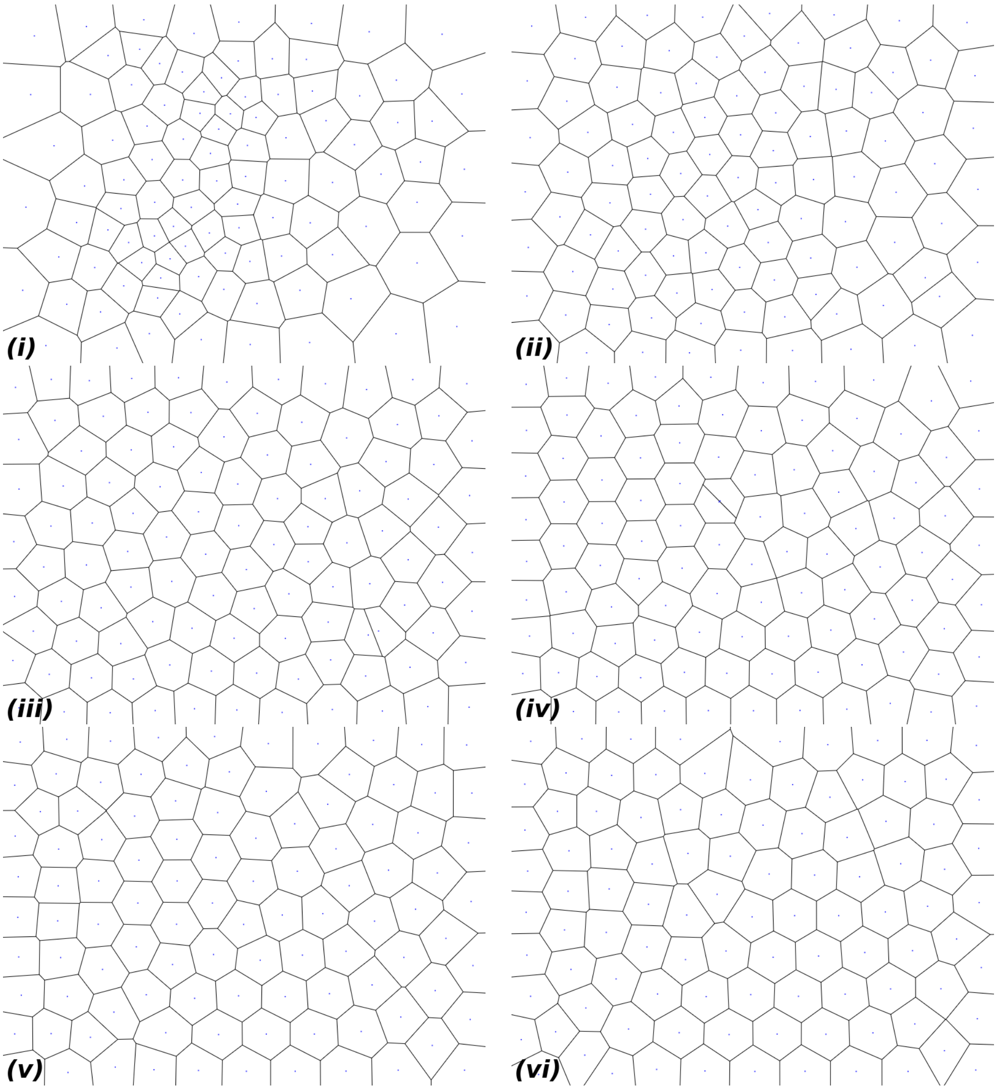


Figure 20. Evolution of CVT for 100 cells. Simulated at [4]

6 Final Remarks and Suggestions

This report develops a methodology to find clustering sets of CEPs using georeferential information and non-default information from a set of clients evaluated between 2016 and 2018 in Brazil.

6.1 Final Remarks

- A main sequential methodology was formulated for the proposed problem. First, clusters were obtained using K-Nearest neighbors by default. Clustering Visualization techniques were then applied to the previous solution, and a Mathematical Optimization Model refinement was proposed.
- The initial experiment, presented in Section 4.3, shows that it is possible to reduce the number of clusters using mathematical models. In our case, we focus on minimizing the distances between clusters. However, a deeper study including the representation of different metrics would enhance the accuracy of the new clusters.
- A second methodology was proposed by considering Polygonal Spatial Clustering.
- Additionally, a Voronoi-based protocol for polygon generation and polygon stability was suggested as a future proposal.
- By considering Information Value Criteria, some of the proposed Cluster solutions were evaluated. Some of the proposed clusterings showed average performance for prediction. These results are preliminaries, and additional analysis could be explored to improve the Information Value obtained.
- Another proposed study considers complementary information suggested in Section 5.2 could be explored in the future development to study the model validation. Studying these other variables by considering the Variables reduction methods shows that it is possible to use this information.
- The results presented in this report are preliminary since the methodology must be reviewed.
- By considering the methodologies here proposed different analyses could be developed and tested.

6.2 Suggestions

- The different methods considered are computationally demanding, so tests were performed using computers with the best computational capacity available.
- The approaches suggested could be automated, and a unique code could be developed for testing different scenarios in Brazil, resulting in a better understanding of the characteristics of the problem for different regions of the country. Even though, analyses for more geographically broad areas, for example, the whole country, would demand higher performance machines.
- Complementary criteria beyond the Information Value could be adopted to test the different proposed clustering methods - e.g., KS Statistics, in a modeling context.
- New analyses must be included to assess the Insolvency Fraction by ZIP Code since its observed distribution is inflated in zeros and ones. Some transformations of this variable

could be explored in future developments, and then the methodology presented here could be replicated to the transformed variable.

- Other techniques as discriminant analysis could also be explored in the future, considering the spatial context.

Acknowledgments
We are grateful to the students Leticia Lemos Burla, Fernanda Yuka Ueno and Guilherme Ramon Rodrigues da Silva of the ICMC/USP for the computational support. This research was funded by FAPESP (grant 2013/07375-0).

References

[1] Marangoni, F. 2021. Workshop CEMEAI: CEP x Georreferência. https://drive.google.com/file/d/1UgBuX_HQ6T8vuVmdKnQUNhGP3eGDvON-/view, Accessed: 2021-23-03.

[2] L. Devroye and L. Gyorfi and G. Lugosi. 1996. A Probabilistic Theory of Pattern Recognition. *Springer*

[3] M. Senechal. 1990. Crystalline Symmetries. *Adam Hilger*

[4] Fortune, S. 2021. Steven Fortune's algorithm to compute Voronoi diagrams Demo 5: Lloyd's relaxation. http://www.raymondhill.net/voronoi/rhill-voronoi-demo5.html, Accessed: 2021-25-03.

[5] N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *10.1080/00031305.1992.10475879. hdl:1813/31637, The American Statistician*, **46 (3)**, 175–185.

[6] Li, H.X. and Yen, V.C. 1995. Fuzzy Sets and Fuzzy Decision-Making. *CRC Press*

[7] R Core Team. 2019. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*

[8] Williams, H.P. 1990. Model Bulding In Mathematical Programming. *John Wiley & Sons*

[9] Taube, M. 2002. Matemática Para Produtividade. *Com Ciência - Revista Eletrônica, SBPC/LBJOR*

[10] Arenales, M. and Armentano, V. and Morabito, R. and Yanasse, H. 2015. Pesquisa Operacional. *Elsevier*

[11] Goldbarg, M.C and HPL Luna. 2000. Pesquisa Operacional. *Editora Campus*

[12] Ehrgott, M. 2005. Multicriteria Optimization. *Springer-Verlag Berlin Heidelberg*

[13] W. Mu and D. TongFirst. 2020. On solving large p-median problems. *Environment and Planning B: Urban Analytics and City Science, 10.1177/2399808319892598*, **47(6)**, 981-996.

[14] Gurobi Optimization. 2021. Build Your Optimization Skills with Python. www.gurobi.com/resource/modeling-examples-using-the-gurobi-python-api\-in-jupyter-notebook

[15] R.A. Howard. 1966. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics, 10.1109/TSSC.1966.300074*, **2 (1)**, 22-26.

[16] Wod, IJ. 1985. Weight of evidence: A brief survey. *Bayesian statistics*, **22 (2)**, 319-331.

[17] Good, IJ. 1960. Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society: Series B (Methodological)*, **2**, 249-270.

[18] Weed, DL. 2005. Weight of evidence: a review of concept and methods. *Risk Analysis: An International Journal*, **25 (6)**, 1545-1557.

[19] Rangel, S. 2012. Introdução à construção de modelos de otimização linear e inteira. 2. ed. *Sociedade Brasileira de Matemática Aplicada e Computacional-SBMAC*

[20] Leao, A.A.S., Toledo, F.M.B., Oliveira, J.F., Carravilla, M.A., Alvarez-Valdés, R. 2020. Irregular packing problems: A review of mathematical models. *European Journal of Operational Research, 10.1016/j.ejor.2019.04.04*, **282 (3)**, 803-822.

[21] Cunha, A. L. and Santos, M. O. 2017. Mathematical Modelling and Solution Approaches for Production Planning in a Chemical Industry. *Pesquisa Operacional* , **37**, 311-331.