Journal of Business and Economics, ISSN 2155-7950, USA
August 2023, Volume 14, No. 8, pp. 359-372
DOI: 10.15341/jbe(2155-7950)/08.14.2023/002

http://www.academicstar.us

Groupings of Registered Hospitalization Establishments in TABNET, São Paulo, in the Year 2020, According to the Evolution of Severe Acute Respiratory Syndrome by Covid-19

Antonio Sergio da Silva, Victor de Godoy Terror, Rodrigo Inácio Nogueira, Cesar Augusto Coutinho Ferreira, Fernanda Kelly Marques de Souza Adriano
(Prevent Senior Institute, São Paulo, Brazil)

Abstract: This research identified 151 hospitalization establishments registered in TABNET in the city of São Paulo regarding patients hospitalized with Severe Acute Respiratory Syndrome by Covid-19 in the year 2020 to analyze the mortality rate from SARS. The research applied the k-means method to cluster these establishments. The silhouette coefficient suggested that all individuals were correctly allocated within the six generated clusters. Three stability measures indicated a few 6 clusters to group these establishments. The FOM measure indicated the k-means method to apply in the clustering algorithm. It is concluded that the mortality rate from SARS by Covid-19 in 2020 does not differ from the rates described in the literature for these 151 hospitalization establishments in the city of São Paulo.

Key words: Covid-19, SARS, mortality rate, k-means, clusters

JEL codes: C380, I120

1. Introduction

The global impact of Severe Acute Respiratory Syndrome (SARS) is difficult to estimate: demographic, cultural, economic, and health factors, as well as differences in healthcare systems among countries, may explain some variations in its incidence (Villar, Blanco & Kackmarek, 2016).

The overall mortality rate for SARS approaches 40-50% in all major series described in the literature, although several controlled randomized studies have reported improved survival in selected patients with SARS (Villar, Sulemanji, & Kackmarek, 2014). Hospital mortality from moderate to severe SARS still exceeds 40% (Villar, Blanco & Kackmarek, 2016).

The objective of this research was to cluster the hospitalization establishments registered in the Municipal Health Department of São Paulo regarding the patients' outcomes concerning the diagnosis of SARS-Covid-19 in the year 2020, to assess the mortality rate in the first year of the pandemic in the city of São Paulo. In other words, to determine whether the overall mortality rate for SARS-Covid-19 in the city of São Paulo remained within the range expected in the literature.

Antonio Sergio da Silva, MSc in Administration, Professor, Medical Residency, Prevent Senior Institute; research areas: operations research, multicriteria decision support methods, multivariate data analysis, healthcare. Email: tao281168@gmail.com.

2. Theoretical Framework

2.1 Cluster Analysis

Cluster analysis is a technique that aims to explore information in a set of variables under analysis. Cluster analysis methods are statistical procedures, as the inclusion of new variables or individuals can modify the formation of clusters. This necessitates the development of a new analysis (Fávero & Belfiore, 2017).

The main objective of this technique is to group individuals into clusters in such a way that:

- Individuals within the same cluster are similar in terms of the values of the analyzed variables.
- Conversely, individuals from different clusters are dissimilar.

It is an exploratory or interdependence technique, and its applications do not have a predictive nature for new individuals not present in the initial sample. The modeling needs to be reapplied whenever new individuals are included in the database. Both the inclusion of new individuals and the inclusion of new variables in the database can lead to a complete rearrangement of the group’s individuals (Fávero & Belfiore, 2017).

2.2 Best Practices for Cluster Creation

Algorithms for cluster analysis are based on measures of dissimilarity. These measures allow quantifying the difference between individuals based on the values they present for the set of variables.

The technique segments individuals into internally homogeneous and mutually exclusive groups. It consists of ordering and allocating individuals into these groups. In this way, it is possible to observe how the ordering and allocation of individuals behave within the created groups, aiming to identify a natural structure for them (Fávero & Belfiore, 2017).

3. Methodology

Severe Acute Respiratory Syndrome (SARS) was a severe and sometimes lethal manifestation of COVID-19, observed on all continents and widely publicized in scientific and non-scientific media.

The purpose of this analysis is to group hospitalization facilities into clusters in such a way that:

- Hospitalization facilities within the same cluster are similar in terms of the values of the variables under analysis (evolution of the patients hospitalized in their facilities).
- Conversely, hospitalization facilities from different clusters are dissimilar.

3.1 Selection of Variables for Cluster Formation

The São Paulo Municipal Health Department provides access to TABNET (Ministério da Saúde, Brazil, 2022) for accessing population databases and SUS (Unified Health System) information systems.

On the TABNET (Ministério da Saúde, Brazil, 2022) page menu, under the Severe Acute Respiratory Syndrome section, it is possible to select cases of COVID-19 Severe Acute Respiratory Syndrome, considering various variables of interest.

For this analysis, the following variables were selected:

- **Row**: hospitalization facility (individuals y_i).
- **Column**: evolution (explanatory variables x_{ij} — death, recovery, total hospitalized, and death rate). The death rate was calculated as the ratio of deaths to total hospitalized and represents the mortality rate for SARS in these facilities.

- **Content**: number of cases.
- Available period: 2020.

Given the above:

- Consider the set of individuals y_i for i = {1, ..., n}.
- Each y_i represents a hospitalization facility in the TABNET (Ministério da Saúde, Brazil, 2022) system.
- Objective: describe the relationship of y_i with a set of explanatory variables x_{ij}, where j = {1, ..., p}.
- Each x_{ij} represents {death, recovery, total hospitalized, death rate}.

3.2 Similarity/Dissimilarity Measure

Cluster analysis algorithms are based on measures of dissimilarity. These measures quantify the difference between individuals y_i based on the values they present for the set of variables x_{ij}.

The dissimilarity assessed for a pair of individuals y_i and $y_{i'}$ can be described as:

- $d_{ii'}$, where i and i' are in {1, 2, ..., n}.

The dissimilarity measures adhere to the following properties:

- $d_{ii'} \geq 0$, with dii' = 0 if i = i'.
- $d_{ii'} = d_{i'i}$ for all i, i' in {1, 2, ..., n} (symmetry).
- $d_{ii'} \leq d_{ik} + d_{i'k}$ for all k in {1, 2, ..., n} (triangular inequality).

3.3 Appropriate Algorithm for the Context

Each individual y_i should be assigned to a cluster k (k in {1, 2, ..., K}) according to a coder k = C(y_i).

The objective is to identify an **optimal coder** in order to create internally homogeneous and externally heterogeneous clusters as much as possible.

Non-hierarchical algorithms rely on successive reallocations of individuals y_i to clusters to create internally homogeneous clusters.

This research utilized the **k-means** algorithm, which is applicable when the variables x_{ij} are quantitative and dissimilarity is based on Euclidean distance (the shortest distance between two individuals in space).

$$d_{ii'} = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \| x_i - x_{i'} \|^2$$

In this research, to remove the scale effect, dissimilarity was weighted in the standardization step of variables x_{ij} using **z-scores**: mean 0 and variance 1.

3.4 Cluster Formation

Similar individuals (y_i) in defined characteristics (x_{ij}) occupying the same space form a high-density pattern within the formed cluster.

Density measures evaluate how close individuals are within the cluster. A smaller variation within the cluster indicates good density, i.e., good clustering. Different indices for assessing cluster density are based on distance measures, such as average/median distances between individuals (Kassambara, 2017).

There should be empty spaces between the formed clusters, separating dissimilar individuals.

Separation measures assess how well a cluster is separated from others. Indices used as separation measures include distances between cluster centers and minimum distances in pairwise comparisons between individuals from different clusters (Kassambara, 2017).

3.5 Clustering Algorithm

In this research, a **partitioning method** was considered for the clustering algorithm. A partitioning method constructs **k clusters**, meaning it classifies the data into **k groups** that together satisfy the requirements of a partition: each group must contain at least one individual, and each individual must belong to exactly one group (Kaufman & Rousseeuw, 2009).

It is important to emphasize that **k** is defined by the researcher. The algorithm will build a partition with as many clusters as desired. However, not all values of **k** will create natural clusters. Therefore, it is recommended to run the algorithm multiple times with different values of **k** and select the value of **k** for which certain characteristics or graphs appear better or choose the clustering that seems to yield the most meaningful interpretation (Kaufman & Rousseeuw, 2009).

4. Results

The data were analyzed using the R software (R Core Team, 2021), with the RStudio (RStudio Team, 2021) IDE.

4.1 Exploratory Data Analysis

This research presents the number of cases (content) per hospitalization facility (row) and per evolution (column) from the COVID-19 Severe Acute Respiratory Syndrome (SARS) page in the TABNET (Ministério da Saúde, Brazil, 2022) system.

All hospitalization facilities y_i that recorded a total of at least 100 cases of SARS in their facilities in 2020 were selected, resulting in a total of 151 facilities.

Within TABNET (Ministério da Saúde, Brazil, 2022), the evolution object consists of five variables x_{ij}: recovery, death, death due to other causes, unknown, and not informed. The selected variables for this analysis were recovery, death, and total cases. The death rate variable is the ratio between the death and total cases variables.

The variables death due to other causes, unknown, and not informed were excluded from the analysis as they were less discriminative, but they were absorbed into the total hospitalizations variable.

It is observed that:

- All variables xij are quantitative (int and dbl).
- There were 151 selected hospitalization facilities (yi).
- There are no missing values in the selected data.

These pieces of information can be visualized in Figure 1.

4.1.1 Correlation Between Variables

Figure 2 shows a very high correlation between the variables total and cure (*"cura"*) (0.97) and total and death (0.82). However, the total variable was kept in the study as it encompasses the variables (no_info (*"sem_info"*), ignored ("*ignorado"*), other_deaths *("óbito_outros"*).

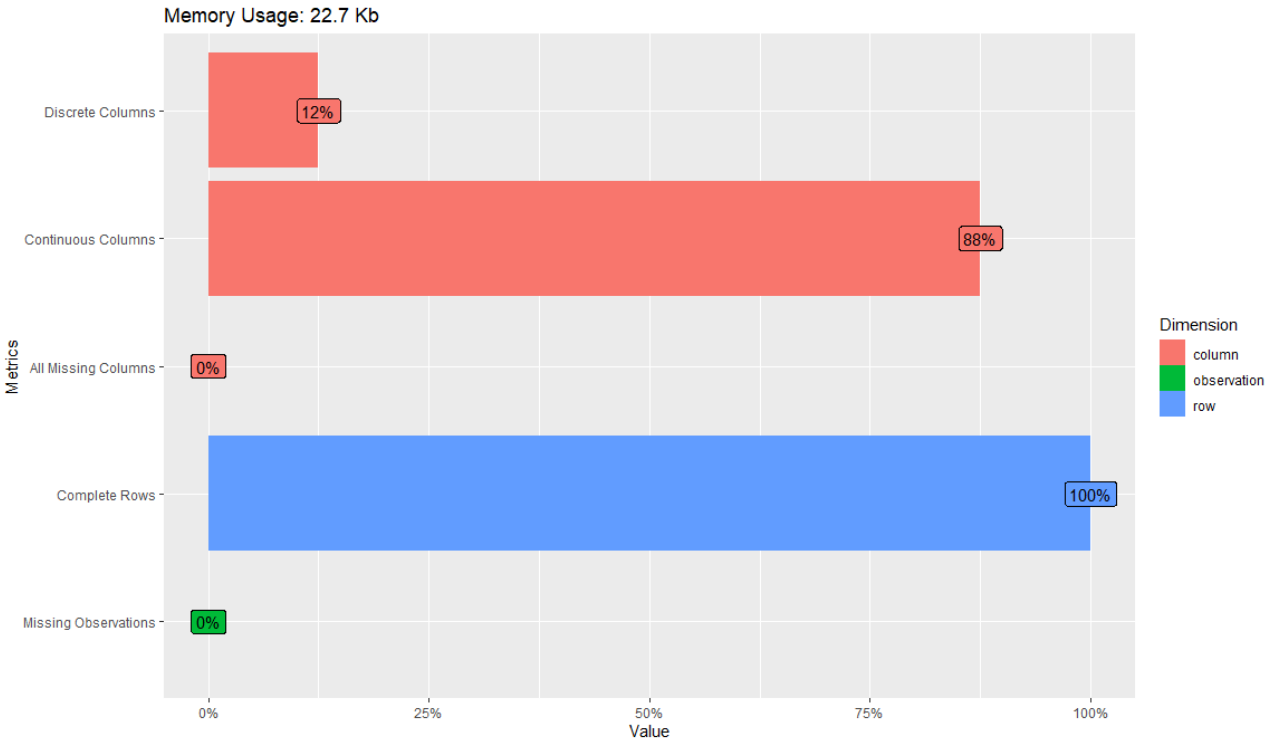
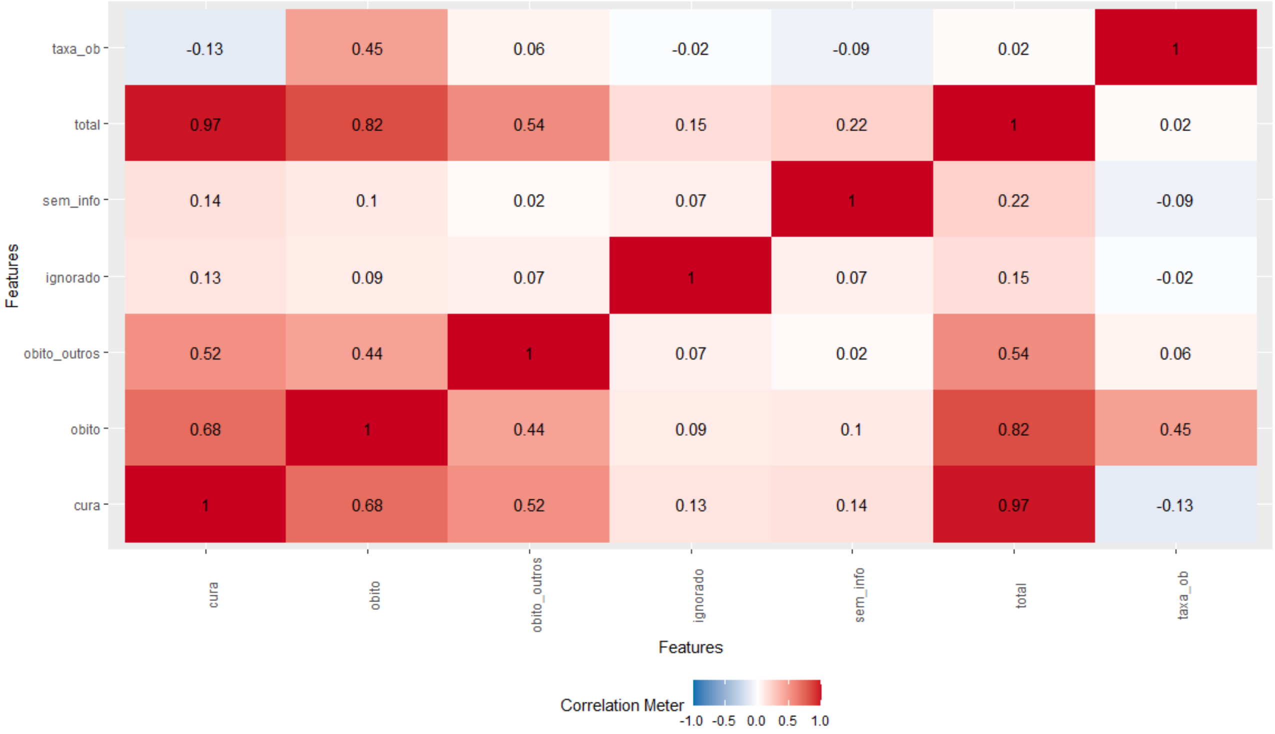


Figure 1 Initial Data Processing

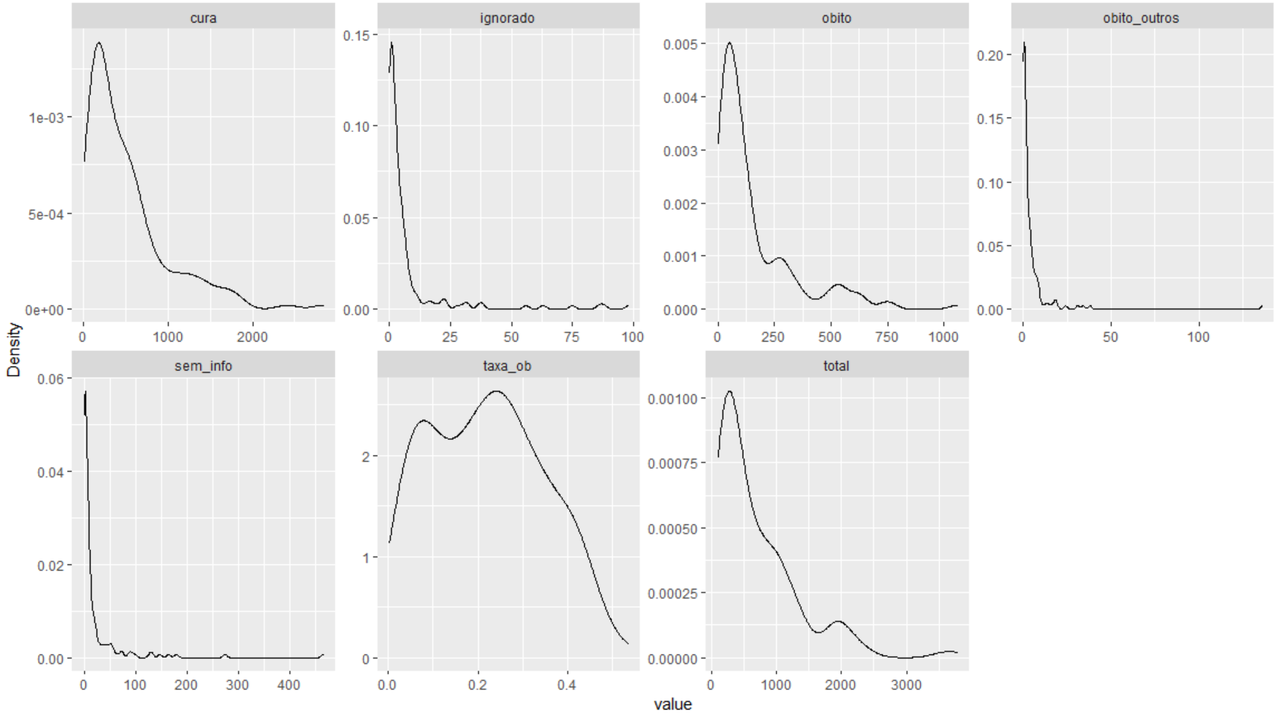


	cura	obito	obito_outros	ignorado	sem_info	total	taxa_ob
taxa_ob	-0.13	0.45	0.06	-0.02	-0.09	0.02	1
total	0.97	0.82	0.54	0.15	0.22	1	0.02
sem_info	0.14	0.1	0.02	0.07	1	0.22	-0.09
ignorado	0.13	0.09	0.07	1	0.07	0.15	-0.02
obito_outros	0.52	0.44	1	0.07	0.02	0.54	0.06
obito	0.68	1	0.44	0.09	0.1	0.82	0.45
cura	1	0.68	0.52	0.13	0.14	0.97	-0.13

Figure 2 Correlation Between Variables

4.1.2 Distribution of Quantitative Variables

A strong positive skewness (to the right) is observed for all variables, except for the mortality rate variable, which exhibits a bimodal pattern (Figure 3).



Legend: *cura* (cure), *ignorado* (ignored), *obito* (death), *obito_outros* (other_deaths), *sem_info* (no_info), *taxa_ob* (death rate).

Figure 3 Distribution of Variables

4.1.3 Summary of the Variable of Interest (Death Rate)

In 2020, a total of 109,961 cases of COVID-19 severe acute respiratory syndrome (SARS) were recorded in TABNET (Ministério da Saúde, Brazil, 2022), of which 24,581 cases resulted in death (death rate = 22.35%).

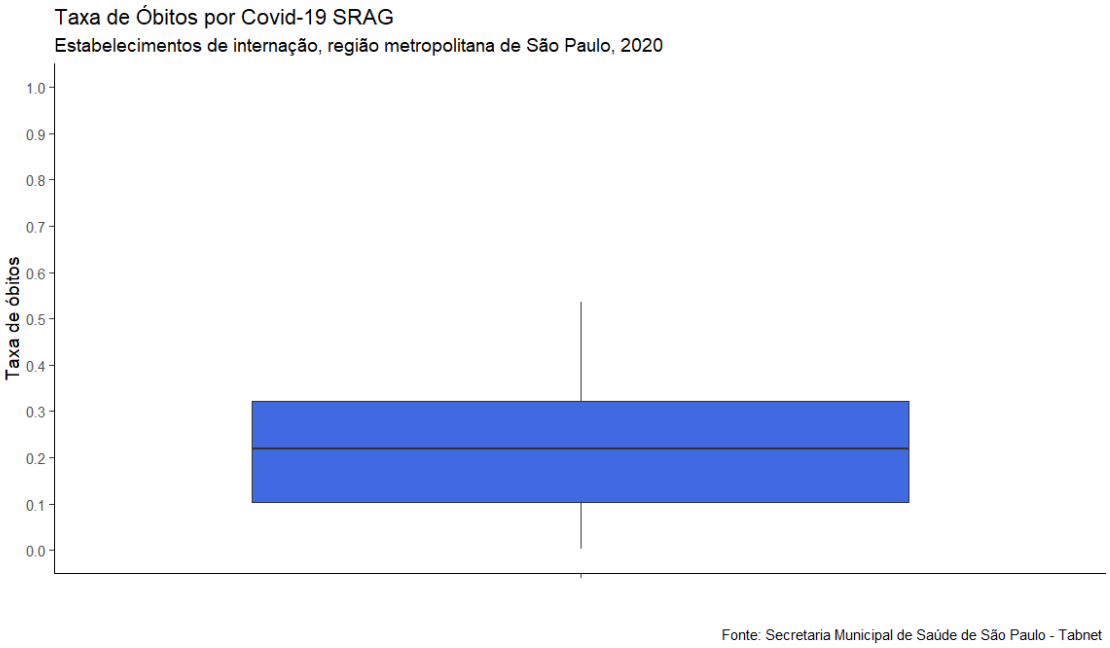
Among the 151 selected hospitalization facilities for this research, the average SARS death rate was found to be 21.91% (standard deviation of 12.8%).

In Figure 4, the central line represents the median of the dataset. The median value was 21.82%. In other words, in 2020, 50% of these 151 hospitalization facilities in the metropolitan region of São Paulo reported a SARS death rate lower than 21.82%.

Note that 25% of the hospitalization facilities recorded a SARS death rate equal to or greater than 32.15%.

The maximum recorded death rate in 2020 was 53.53% and is not considered an outlier.

These findings can be visualized in the boxplot in Figure 4.



Legend: *Taxa de Óbitos por Covid-19 SRAG* (COVID-19 SARS Death Rate.), *Estabelecimentos de internação, região metropolitana de São Paulo* (Hospitalization facilities, São Paulo metropolitan region), *Taxa de Óbitos* (Death Rate), *Fonte: Secretaria Municipal de Saúde de São Paulo* (Source: Municipal Health Secretary of São Paulo).

Figure 4 Boxplot of the Death Rate

4.2 K-means Method

The K-means algorithm is the most used partition-based clustering algorithm based on the sum of squares criterion (Kassambara, 2017; Macqueen, 1967). It is a simple algorithm, easy to implement, and available in almost all data mining software. Moreover, it is highly versatile, considering aspects such as initialization, distance, function, and termination criterion.

4.2.1 Determining the Number of Clusters

The k-means method is a non-hierarchical procedure for clustering individuals, where the initial number of clusters is defined by the researcher (Fávero & Belfiore, 2017; Kassambara, 2017, 2017).

Visual analysis using a two-dimensional line graph of the K Clusters (Number of clusters k) and the Total Within Sum of Squares of Prediction Errors (Total Within Sum of Square — SSE) suggests an initial number of clusters to be used. SSE represents the variance and standard deviation (*inertia*) of the data in the used dataset. Thus, it allows visualizing how close the data points are to each other. A smaller number of clusters leads to a higher value of this inertia (Kassambara, 2017).

Figure 5 displays the graph with the selected number of clusters (red vertical dashed line).

Figure 5 represents the variation within the clusters. It can be observed that from the fifth cluster onwards, the value of SSE does not show significant variations. Increasing the number of clusters indicates that this variation would become smaller and smaller. Therefore, the quantity of five or six clusters appears to be a very interesting number to apply to this dataset.

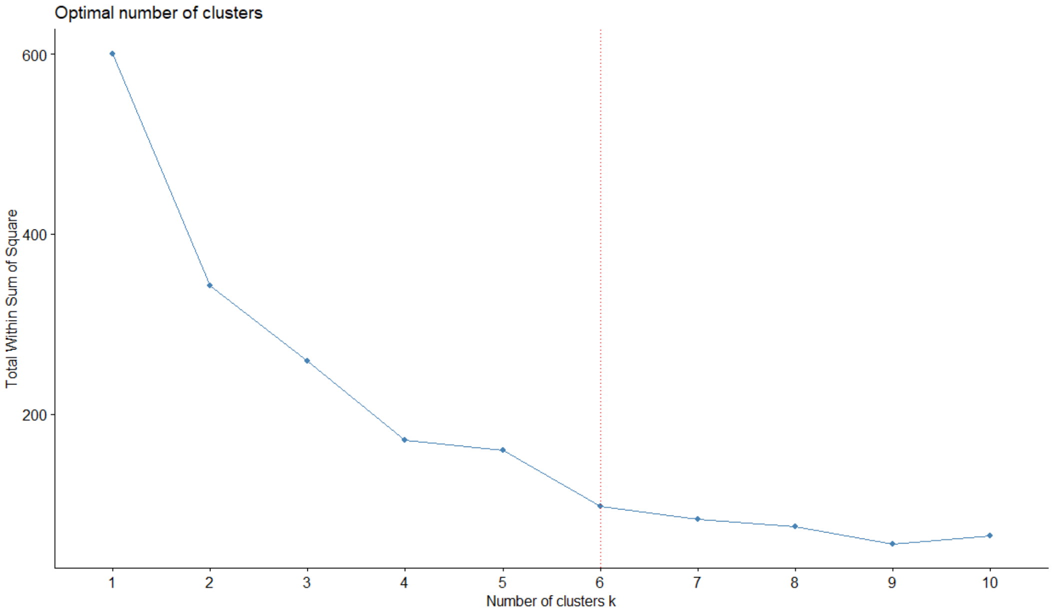


Figure 5 Number of Clusters

4.2.2 Graphical Visualization of the Clusters

The algorithm to determine the allocation of observations in each cluster is called nearest centroid sorting. K-means uses the Euclidean distance as the distance criterion to form the groups.

The *fviz_cluster ()* function from the *factoextra package* in R can be used to easily visualize the K-means clusters. In the resulting graph, the observations are represented by points, using Principal Component Analysis (PCA) when the number of variables “xij” is greater than 2 (Kassambara, 2017).

Figure 6 suggests that the formation of six clusters appears to be more accurate in allocating the healthcare facilities.

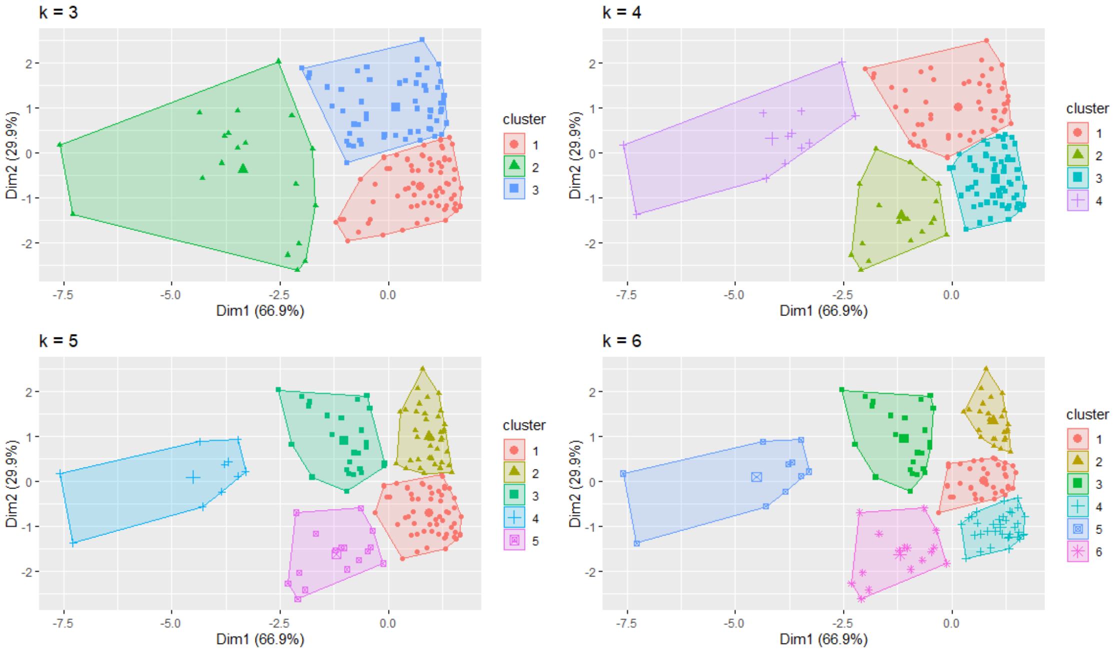


Figure 6 Segmentation of Healthcare Facilities Considering 3, 4, 5, or 6 Clusters

Considering the formation with k = 6 groups (clusters), it can be observed that the formed groups exhibit internal homogeneity (variability within the groups), with each healthcare facility being closer to other facilities within the same group than to facilities from other groups (variability between the groups).

4.2.3 Internal Validation of Clustering

Internal validation of clustering involves evaluating the quality of the clustering algorithm's results by using intrinsic information from the clustering process, without relying on external data. This is important to avoid finding patterns in random data and to compare different clustering algorithms (Kassambara, 2017).

Internal validation methods utilize intrinsic information within the data to assess the quality of the clustering. These methods include *connectivity*, *silhouette coefficient*, and *Dunn index* (Kassambara, 2017).

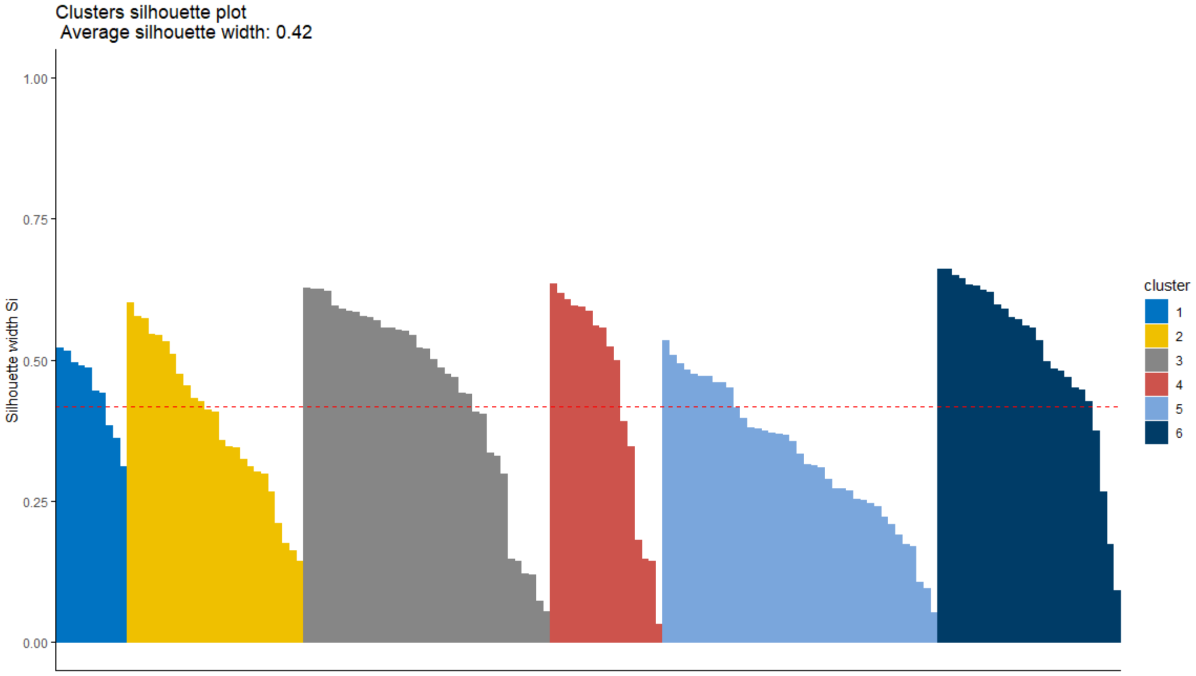
Connectivity measures the extent to which items are placed in the same cluster as their nearest neighbors in the data space. Connectivity values range from 0 to infinity and should be minimized (Kassambara, 2017).

The *silhouette coefficient* (S_i) measures the similarity of an individual yi with other individuals in its own cluster compared to those in the neighboring cluster (Kassambara, 2017). The S_i values range from -1 to +1:

- A value close to 1 indicates that the individual is well-clustered, meaning that yi is like other individuals in its group (Kassambara, 2017).
- A value close to -1 indicates that the individual is poorly clustered and that assigning it to another cluster would likely improve the overall results (Kassambara, 2017).

The following R output displays the values of S_i for the formed clusters, which can be visualized in Figure 7.

```
## cluster     size       ave.sil. width
##      1        10           0.44
##      2        25           0.39
##      3        35           0.45
##      4        16           0.44
##      5        39           0.33
##      6        26           0.51
```



Clusters silhouette plot
Average silhouette width: 0.42

Figure 7 S_i Values for the Six Clusters

Below is the list of the top 10 establishments (y_i) with the highest S-i coefficients (R output).

```
##          cluster     neighbor          sil_width
## 6          1            2              0.5209484
## 4          1            2              0.5155121
## 7          1            2              0.4949843
## 5          1            2              0.4906690
## 3          1            4              0.4861582
## 8          1            2              0.4450594
## 2          1            2              0.4415461
## 1          1            4              0.3840257
## 10         1            2              0.3605553
## 12         1            2              0.3104518
```

No establishments with negative S-i coefficients are observed. This means that all individuals are correctly allocated in the generated clusters.

The **Dunn index** identifies dense and well-separated clusters. It is defined as the ratio between the minimum distances between clusters and the maximum distance between clusters. In this study, the Dunn index computed by R resulted in the following outcome:

```
## [1] 0.005515741
```

The Dunn index has a value between zero and infinity and should be maximized. Higher Dunn scores are more desirable. In this analysis, the value is very low, suggesting that it may not be a good clustering. Its value should not be interpreted in isolation.

4.2.4 Simultaneous Evaluation of Clustering Algorithms

The **clValid** R package (Brock et al., 2008) compares multiple clustering algorithms simultaneously in a single function call to identify the best clustering approach and the optimal number of clusters (Kassambara, 2017). The R output for these values, using the clValid package, resulted in:

```
##
## Clustering Methods:
## hierarchical, kmeans, pam
##
## Cluster sizes:
##   2   3   4   5    6
##
## Validation Measures:
##   2   3   4   5   6
##
## hierarchical method
##            Connectivity    3.6250    7.2829   17.1468   20.0587   22.0587
##            Dunn            0.2086    0.2086    0.0752    0.0942    0.0942
##            Silhouette      0.5830    0.5187    0.3761    0.4242    0.4177
## kmeans   method
##             Connectivity   22.5302   20.6480   35.3583   47.7175   49.7175
##             Dunn            0.0224    0.0235    0.0339    0.0323    0.0323
##             Silhouette      0.5195    0.3834    0.3999    0.4029    0.3946
## pam   method
##            Connectivity   19.7647   25.6599   29.1262   32.7095   34.3560
##             Dunn            0.0183    0.0280    0.0391    0.0235    0.0276
##             Silhouette      0.4005    0.3539    0.3771    0.4426    0.4113
##
```

```
## Optimal Scores:
##
##                      Score        Method        Clusters
## Connectivity         3.6250       hierarchical     2
## Dunn                 0.2086       hierarchical     2
## Silhouette           0.5830       hierarchical     2
```

Considering the clustering methods (*hierarchical, k-means, and PAM*), the internal validation scores (connectivity, Dunn index, and silhouette coefficient) suggest the hierarchical method and the formation of two clusters for this dataset.

4.2.5 Stability Measures

The stability measures evaluate the consistency of a clustering result by comparing it with the clusters obtained after removing each column, one at a time (Kassambara, 2017).

The cluster stability measures include:

- Average Proportion of Non-Overlap (APN)
- Average Distance (AD)
- Average Distance between Means (ADM)
- Figure of Merit (FOM)

APN, AD, and ADM are all based on the cross-tabulation of the original clustering on the complete data with the clustering based on the removal of a single column (Kassambara, 2017).

APN measures the average proportion of individuals not placed in the same cluster when clustering is based on the complete data compared to when clustering is based on data with a single column removed (Kassambara, 2017).

AD measures the average distance between individuals placed in the same cluster in both cases (complete data and column removal) (Kassambara, 2017).

ADM measures the average distance between the cluster centers for individuals placed in the same cluster in both cases (Kassambara, 2017).

FOM measures the average intracluster variation of the excluded column, where clustering is based on the remaining (non-excluded) columns (Kassambara, 2017).

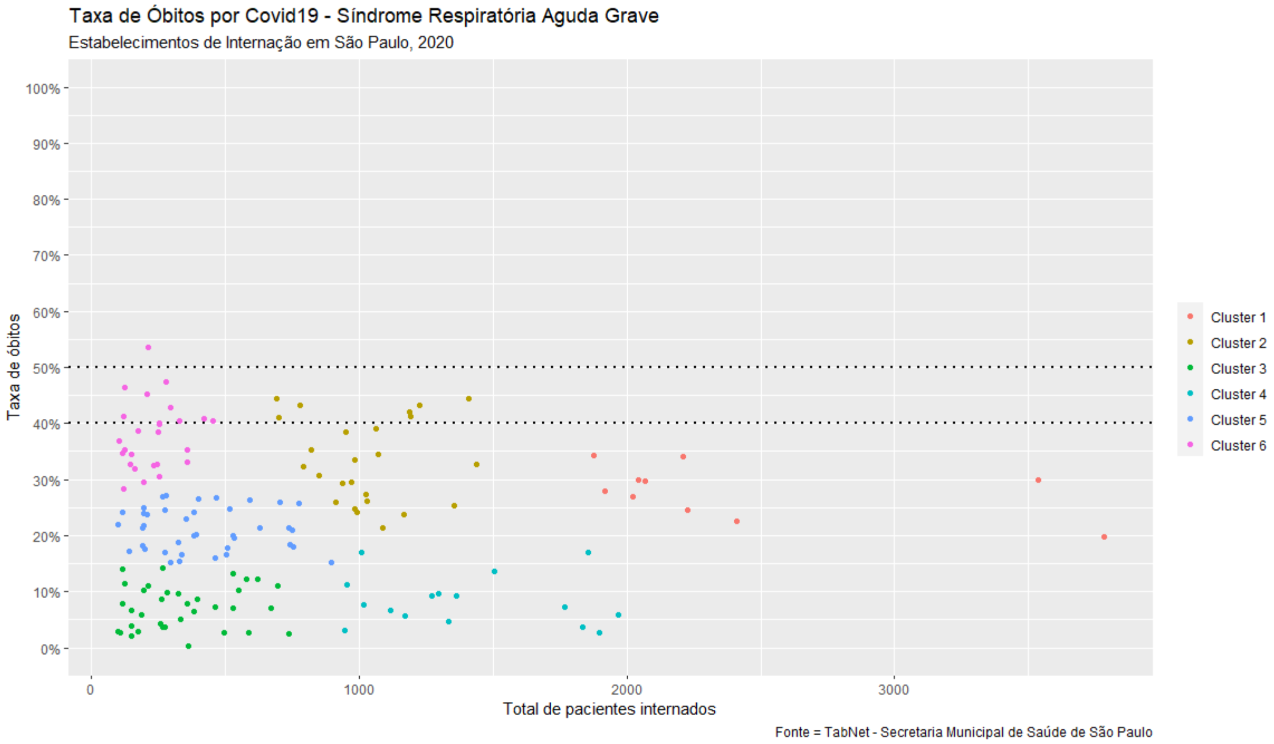
The values of APN, ADM, and FOM range from 0 to 1, with lower values corresponding to highly consistent clustering results. AD has a value between 0 and infinity, and lower values are also preferred (Kassambara, 2017). The R output for these stability measures resulted in:

```
##        Score           Method         Clusters
## APN    0.02119205      hierarchical      2
## AD     1.04508486      pam               6
## ADM    0.26245203      pam               6
## FOM    0.52181358      kmeans            6
```

It is observed that three stability measures suggest a few 6 clusters for this dataset. The FOM measure suggests the k-means method for clustering algorithm application. Therefore, this measure confirms the choice of using the k-means method with 6 clusters in this research to group the 151 identified healthcare facilities in TABNET (Ministério da Saúde, Brazil, 2022).

5. Discussion

Considering the technical aspects, the k-means algorithm clustered the 151 healthcare facilities as shown in Figure 8.



Legend: Taxa de Óbitos por Covid-19 Síndrome Respiratória Aguda Grave (COVID-19 SARS Death Rate.), Estabelecimentos de internação, em São Paulo (Hospitalization facilities, São Paulo), Taxa de Óbitos (Death Rate), Total de pacientes internados (Total number of hospitalized patients), Fonte: TABNET Secretaria Municipal de Saúde de São Paulo (Source: TABNET Municipal Health Secretary of São Paulo).

Figure 8 COVID-19 SARS Mortality Rate in 2020, São Paulo

We observed that the hospitalization facilities are distributed into 6 clusters, considering the mortality rate in relation to the total number of patients hospitalized with Severe Acute Respiratory Syndrome (SARS) in their premises in 2020, as follows:

- Cluster 1: This cluster represents hospitalization facilities with an average mortality rate of 21.2% (3.77%). There is low intracluster variability, with a coefficient of variation of 17.80%.
- Cluster 2: This cluster represents hospitalization facilities with an average mortality rate of 37.8% (6.08%). There is low intracluster variability, with a coefficient of variation of 16.21%.
- Cluster 3: This cluster represents hospitalization facilities with an average mortality rate of 33.4% (7.49%). There is moderate intracluster variability, with a coefficient of variation of 22.40%.
- Cluster 4: This cluster represents hospitalization facilities with an average mortality rate of 7.21% (4.66%). There is high intracluster variability, with a coefficient of variation of 53.70%.
- Cluster 5: This cluster represents hospitalization facilities with an average mortality rate of 28.0% (4.66%). There is low intracluster variability, with a coefficient of variation of 16.70%.
- Cluster 6: This cluster represents hospitalization facilities with an average mortality rate of 8.39%

(4.50%). There is high intracluster variability, with a coefficient of variation of 53.60%.

The black dashed horizontal lines represent values described in the literature for overall mortality, ranging from moderate to severe, for SARS, between 40% and 50% (Villar, Blanco & Kacmarek, 2016; Villar, Sulemanji & Kacmarek, 2014).

Therefore, it can be observed that the hospitalization facilities registered in the Municipal Health Department of São Paulo in 2020 had mortality rates due to SARS consistent with the rates reported in the literature. In other words, there was no increase in the mortality rate due to SARS-CoV-2 in São Paulo beyond what was expected. The majority of these 151 facilities had a mortality rate below 40%.

6. Conclusion

This study identified 151 hospitalization facilities in the city of São Paulo that recorded at least 100 hospitalizations due to SARS-CoV-2 in 2020. The research revealed that the mortality rate in these institutions remained like what is described in the literature, between 40% and 50% for moderate to severe forms of the disease. Only 25% of the 151 hospitalization facilities recorded a mortality rate equal to or higher than 32.15% in 2020.

The FOM measure indicated the k-means method with six clusters to be applied as the clustering algorithm. No facilities with negative S_i values were observed, indicating that they were correctly allocated to the generated clusters.

A possible line of further research would be to compare the epidemiological and sociodemographic profiles of each generated cluster to understand the behavior of the disease.

Another line of research could include an external validation process of the clusters. This process involves comparing the results of a cluster analysis with an externally known outcome, such as externally provided class labels. It measures the extent to which cluster labels correspond to the externally provided class labels. This approach is used to select the best clustering algorithm for a specific dataset. The Rand index can be explored to assess the correspondence of results between the clusters ((Hubert and Arabie, 1985; Rand, 1971).

References

Brock G., Pihur V. and Datta S. (2008). “clValid: An R package for cluster validation”, *Journal of Statistical Software*, Vol. 25, pp. 1-22.

Fávero L. P. and Belfiore P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata*, Elsevier Brasil.

Hubert L. and Arabie P. (1985). “Comparing partitions”, *Journal of Classification*, Vol. 2, No. 1, pp. 193-218.

Kassambara A. (2017). “Practical guide to cluster analysis in R: Unsupervised machine learning”, *Sthda*, Vol. 1

Kaufman L. and Rousseeuw P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons.

Macqueen J. (1967). “Classification and analysis of multivariate observations”, in: *5th Berkeley Symp. Math. Statist. Probability*, pp. 281-297.

Datasus, Tabnet (Ministério Da Saúde, Brazil, 2022). “Tabulação de Dados Epidemiológicos”, accessed on October 10, 2022, available at: http://www2.datasus.gov.br/DATASUS/index.php?area=0201.

Rand W. M. (1971). “Objective criteria for the evaluation of clustering methods”, *Journal of the American Statistical Association*, Vol. 66, No. 336, pp. 846-850.

R. Core Team (2021). “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria, available online at: https://www.R-project.org/.

RStudio Team (2021). “RStudio: Integrated development environment for R”, *RStudio*, PBC, Boston, MA, available online at: http://www.rstudio.com/.

Villar J., Blanco J. and Kackmarek R. M. (2016). "Current incidence and outcome of the acute respiratory distress syndrome", *Current Opinion in Critical Care*, Vol. 22, No. 1, pp. 1-6.

Villar J., Sulemanji D. and Kackmarek R. M. (2014). "The acute respiratory distress syndrome: Incidence and mortality, has it changed?", *Current Opinion in Critical Care*, Vol. 20, No. 1, pp. 3-9.