Agrupamentos dos estabelecimentos de internação registrados no TABNET, São Paulo, no ano de 2020, de acordo com a evolução da Síndrome Respiratória Aguda Grave por Covid-19

Antonio Sergio da Silva (FMU, BSBr, Prevent Senior), tao281168@gmail.com

Victor de Godoy Terror (Prevent Senior), victor.terror@preventsenior.com.br

Rodrigo Inácio Nogueira (FMU, Prevent Senior), rodrigo.nogueira@preventsenior.com.br

Cesar Augusto Coutinho Ferreira (Prevent Senior), cesar.ferreira@preventsenior.com.br

Fernanda Kelly Marques de Souza Adriano (UNIFESP, Prevent Senior), fernanda.souza@preventsenior.com.br

Resumo

Esta pesquisa identificou 151 estabelecimentos de internação registrados no TABNET, na cidade de São Paulo em relação aos pacientes internados por SRAG por Covid-19 no ano de 2020 para analisar a taxa de óbito por SARG. A pesquisa aplicou o método *k-means* para agrupar estes estabelecimentos. O coeficiente de silhueta sugeriu que todos os indivíduos foram alocados corretamente nos seis *clusters* gerados. Três medidas de estabilidade indicaram um número de 6 *clusters* para agrupar estes estabelecimentos. A medida FOM indicou o método *k-means* para aplicar no algoritmo de agrupamento. Conclui-se que a taxa de óbito por SARG por Covid-19, em 2020, não difere das taxas descritas na literatura para estes 151 estabelecimentos de internação na cidade de São Paulo.

Palavras-chave: Covid-19, SARG, taxa de mortalidade, *k-means*, *clusters*.

1. Introdução

O impacto global da Síndrome Respiratória Aguda Grave (SRAG) é difícil de estimar: fatores demográficos, culturais, econômicos e de saúde, diferenças nos sistemas de saúde entre os diversos países podem explicar algumas diferenças na sua incidência (VILLAR; BLANCO; KACMAREK, 2016).

A taxa de óbitos geral por SRAG aproxima-se de 40-50% em todas as grandes séries descritas na literatura, embora vários estudos randomizados controlados relataram uma melhora na sobrevida em pacientes selecionados com SRAG (VILLAR; SULEMANJI; KACMAREK, 2014). A mortalidade hospitalar por SRAG moderada e grave ainda é superior a 40% (VILLAR; BLANCO; KACMAREK, 2016).

Esta pesquisa teve por objetivo agrupar os estabelecimentos de internação registrados na Secretaria Municipal de Sáude de São Paulo em relação à evolução dos pacientes no que diz respeito ao diagnóstico de SRAG-Covid-19, no ano 2020 para verificar a taxa de óbitos no primeiro ano da pandemia na cidade de São Paulo. Ou seja, verificar se a taxa de óbito geral por SRAG-Covid-19 na cidade de São Paulo se manteve dentro do esperado pela literatura.

2. Fundamentação Teórica

2.1 Análise de agrupamentos

A análise de agrupamentos é uma técnica cujo interesse é explorar informações em um conjunto de variáveis em análise. Os métodos de análise de agrupamento são **procedimentos estatísticos**, uma vez que a inclusão de novas variáveis ou de novos indivíduos pode modificar a formação dos *clusters*. Isto requer, obrigatoriamente, a elaboração de nova análise (FÁVERO; BELFIORE, 2017).

O principal objetivo desta técnica é agrupar os indivíduos em *clusters*, de tal forma que:

- Indivíduos de um mesmo *cluster* sejam semelhantes (**similares**) em relação aos valores das variáveis em análise.
- E em contrapartida, os indivíduos de *clusters* distintos sejam diferentes (**dissimilares**).

Trata-se de uma técnica **exploratória**, ou de **interdependência**, cujas aplicações não apresentam caráter preditivo para novos indivíduos não presentes na amostra inicial. É necessária a reaplicação da modelagem sempre que houver a inclusão de novos indivíduos no banco de dados. Tanto a inclusão de novos indivíduos, quanto a inclusão de novas variáveis no banco de dados podem criar um rearranjo completo dos indivíduos do grupo (FÁVERO; BELFIORE, 2017).

2.2 Boas práticas para a criação de agrupamentos

Os algoritmos para análise de *clusters* se baseiam em medidas de dissimilaridade. Estas medidas permitem quantificar a diferença entre indivíduos com base nos valores apresentados para o conjunto de variáveis.

A técnica segmenta os indivíduos em grupos homogêneos internamente e heterogêneos entre si e mutuamente exclusivos. A técnica consiste em **ordenar** e **alocar** os indivíduos em **grupos**. Desta forma, é possível verificar como se comportam o ordenamento e a alocação desses indivíduos nos grupos criados, buscando uma estrutura natural para eles. (FÁVERO; BELFIORE, 2017).

3 Metodologia

A SRAG foi uma manisfetação grave, por vezes letal, da covid-19, em todos os continentes, amplamente divulgado nos meios de comunicação, sejam científicos ou não.

O propósito desta análise é agrupar os estabelecimentos de internação em *clusters*, de tal forma que:

- Estabelecimentos de internação de um mesmo *cluster* sejam semelhantes (**similares**) em relação aos valores das variáveis em análise (evolução dos pacientes internados em suas dependências).
- E em contrapartida, os estabelecimentos de internação de *clusters* distintos sejam diferentes (**dissimilares**).

3.1 Seleção das variáveis para a formação dos agrupamentos

A Secretaria Municipal de Sáude de São Paulo disponibiliza o TABNET para o acesso às bases de dados de população e dos sistemas de informações do SUS.

No menu da página do TABNET, no item, Síndrome Respiratória Aguda Grave, é possível selecionar os casos de COVID 19 SÍNDROME RESPIRATÓRIA AGUDA GRAVE, considerando diversas variáveis de interesse.

Para esta análise foram selecionadas:

- **linha**: estabelecimento de internação (*indivíduos* y_i).
- **coluna**: evolução (*variáveis explicativas* x_{ij} - óbito, cura, total de internados e taxa de óbito). A taxa de óbito foi calculada pela razão entre o número de óbitos e o total de internados e traduz a taxa de óbitos por SRAG nestes estabelecimentos.
- **conteúdo**: número de casos.
- **período disponível**: 2020.

Ante o exposto:

- Considere o conjunto de indivíduos y_i para $i = \{1, \ldots, n\}$.
- Cada $y_i \in$ {estabelecimentos de internação no sistema TABNET}.
- Objetivo: descrever o relacionamento de y_i com um conjunto de variáveis explicativas x_{ij} , com $j = \{1, \ldots, p\}$.
- Cada $x_{ij} \in$ {óbito, cura, total de internados, taxa de óbitos}.

3.2 Medida de similaridade/dissimilaridade

Algoritmos de análise de agrupamentos se baseiam em **medidas de dissimilaridade**. Estas medidas quantificam a diferença entre os indivíduos y_i com base nos valores apresentados para o conjunto de variáveis x_{ij}.

A dissimilaridade avaliada para um par de indivíduos y_i e $y_{i'}$ pode ser descrita como:

- $d_{ii'}$, com i e $i' \in \{1, 2, \ldots, n\}$.

As medidas de dissimilaridade atendem às propriedades:

- $d_{ii'} \geq 0$, com $d_{ii'} = 0$ se i = i';
- $d_{ii'} = d_{i'i}$ para todo i, i' $\in$ {1, 2, …, n} (simetria);
- $d_{ii'} \leq d_{ik} + d_{i'k}$, para todo k $\in$ {1, 2, …, n} (desigualdade triangular).

3.3 Algoritmo oportuno para o contexto

Cada indivíduo y_i deve ser alocado a um *cluster* k (k $\in$ {1, 2, …, K}) segundo um codificador $k = C(y_i)$.

O objetivo consiste em identificar um codificador **ótimo**, de tal sorte que seja possível criar, o máximo possível, *clusters* homogêneos internamente e heterogêneos entre si.

Os algoritmos *não hierárquicos* se baseiam em realocações sucessivas dos indivíduos y_i aos *clusters* para criá-los internamente mais homogêneos.

Esta pesquisa fez uso do algoritmo ***k-means***. Este se aplica quando as variáveis x_{ij} são quantitativas e a dissimilaridade é baseada na distância Euclideana (menor distância entre dois indivíduos no espaço):

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

Nesta pesquisa, para remover o **efeito de escala**, a dissimilaridade foi ponderada na etapa de padronização das variáveis x_{ij} pelo **z-escore**: média 0 e variância 1.

3.4 Formação dos agrupamentos

Indivíduos similares (y_i) nas características definidas (x_{ij}) que ocupam o mesmo espaço formam um padrão de alta densidade no *cluster* formado.

Medidas de densidades avaliam quão próximos estão os indivíduos dentro do *cluster*. Uma variação menor dentro do *cluster* indica uma boa densidade, ou seja, um bom agrupamento. Os diferentes índices para avaliar a densidade dos *clusters* são baseados em medidas de distância, como as distâncias médias/medianas entre os indivíduos (KASSAMBARA, 2017).

Entre os *clusters* formados deve haver espaços vazios que separam os indivíduos dissimilares.

Medidas de separação avaliam o quão bem um *cluster* está separado dos demais. Os índices usados como medidas de separação incluem as distâncias entre os centros dos *clusters* e as distâncias mínimas em pares entre indivíduos em diferentes *clusters* (KASSAMBARA, 2017).

3.5 Algoritmo de agrupamento

Nesta pesquisa, consideramos o método de **particionamento** para o algoritmo de agrupamento. Um método de particionamento constrói ***k clusters***. Isto é, classifica os dados em **k grupos**, que juntos satisfazem os requisitos de uma partição: cada grupo deve conter pelo menos um indivíduo e cada indivíduo deve pertencer a exatamente um grupo (KAUFMAN; ROUSSEEUW, 2009).

Convém enfatizar que **k** é definido pelo pesquisador. O algoritmo construirá uma partição com quantos *clusters* se desejar. Entretanto, nem todos os valores de **k** criam agrupamentos **naturais**. Assim, se recomenda executar o algoritmo várias vezes com diferentes valores de **k** e selecionar aquele **k** para o qual certas características ou gráficos parecem melhores, ou manter o agrupamento que parece dar origem à interpretação mais significativa (KAUFMAN; ROUSSEEUW, 2009).

5 Resultados

5.1 Análise exploratória dos dados

Esta pesquisa apresenta o número de casos (**contéudo**) por estabelecimento de internação (**linha**) e por evolução (**coluna**) da página COVID 19 Síndrome Respiratória Aguda Grave (SGAG) no sistema TABNET.

Foram selecionados todos os estabelecimentos de internação y_i que registraram um total de pelo menos 100 casos de SRAG, em 2020, em suas dependências, o que retornou um total de 151 estabelecimentos.

No TABNET, no objeto evolução constam cinco variáveis x_{ij}: *cura, óbito, óbito por outras causas, ignorado e sem informação*. As variáveis selecionadas para esta análise foram **cura**,

óbito e **total de casos**. A variável **taxa de óbito** é a razão entre as variáveis **óbito** e **total de casos**.

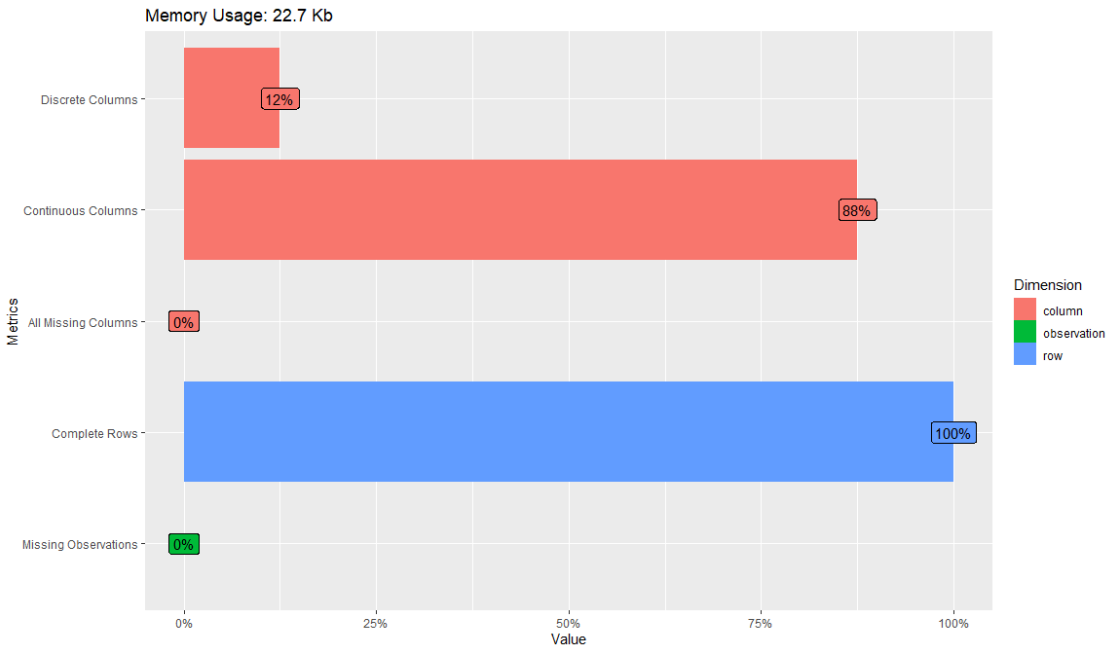
As variáveis **óbito por outras causas**, **ignorado** e **sem informação** foram excluídas da análise por serem pouco discriminativas, mas foram absorvidas na variável total de internações.

Observa-se que:

- Todas as variáveis x_{ij} são quantitativas (*int* e *dbl*);
- Houve 151 estabelecimentos de internação selecionados (y_i);
- Não há *missing values* nos dados selecionados.

Estas informações podem ser visualizadas figura 1.

Figura 1 – Tratamento inicial da base de dados

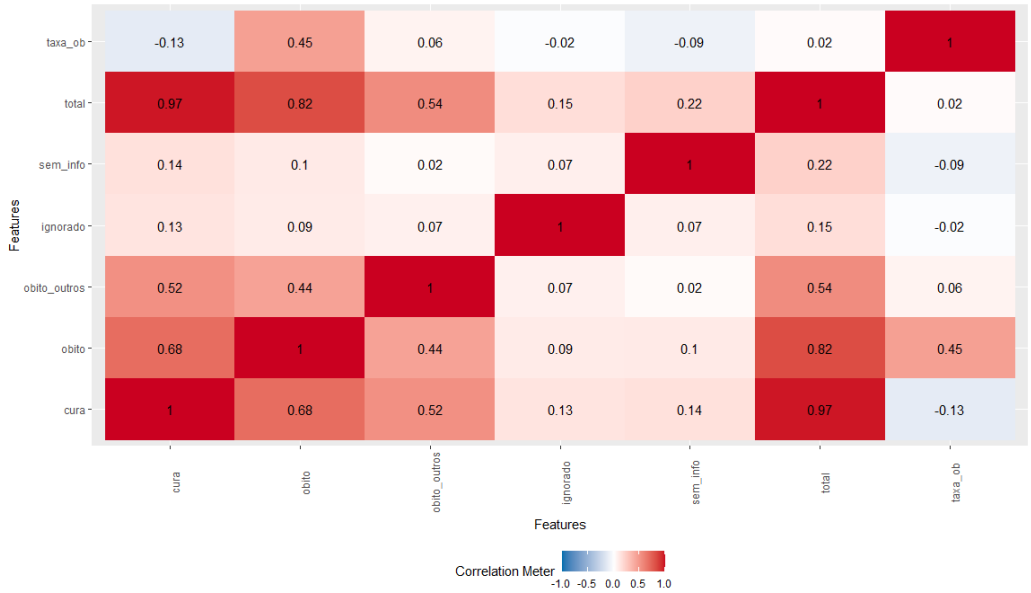


Fonte: os autores

a) Correlação entre as variáveis

Observa-se na figura 2 uma correlação muito alta entre as variáveis total e cura (0.97) e total e óbito (0.82). No entanto, a variável total foi mantida na pesquisa por absorver as variáveis (sem_info, ignorado, obitos_outros).

Figura 2 – Correlação entre as variáveis

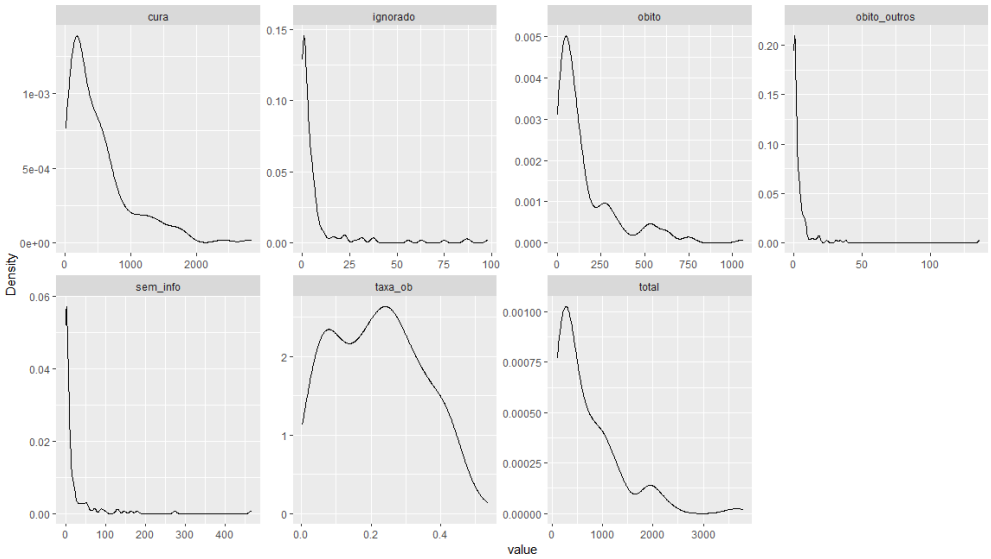


Fonte: os autores

b) Distribuição das variáveis quantitativas

Observa-se forte assimetria positiva (à direita) para todas as variáveis, exceto para a variável taxa de óbito, cujo padrão é bimodal (figura 3)

Figura 3 – Distribuição das variáveis



Fonte: os autores

c) Resumo sumário da variável de interesse (taxa de óbito)

Em 2020, foram registradas 109.961 casos de COVID 19 SRAG no TABNET, dos quais, 24.581 casos evoluíram para óbito (taxa de óbito = 22.35%).

Dos 151 estabelecimentos de internação selecionados para esta pesquisa, se verificou que a média da taxa de óbitos por SRAG foi de 21,91% (desvio-padrão de 12.8 %).

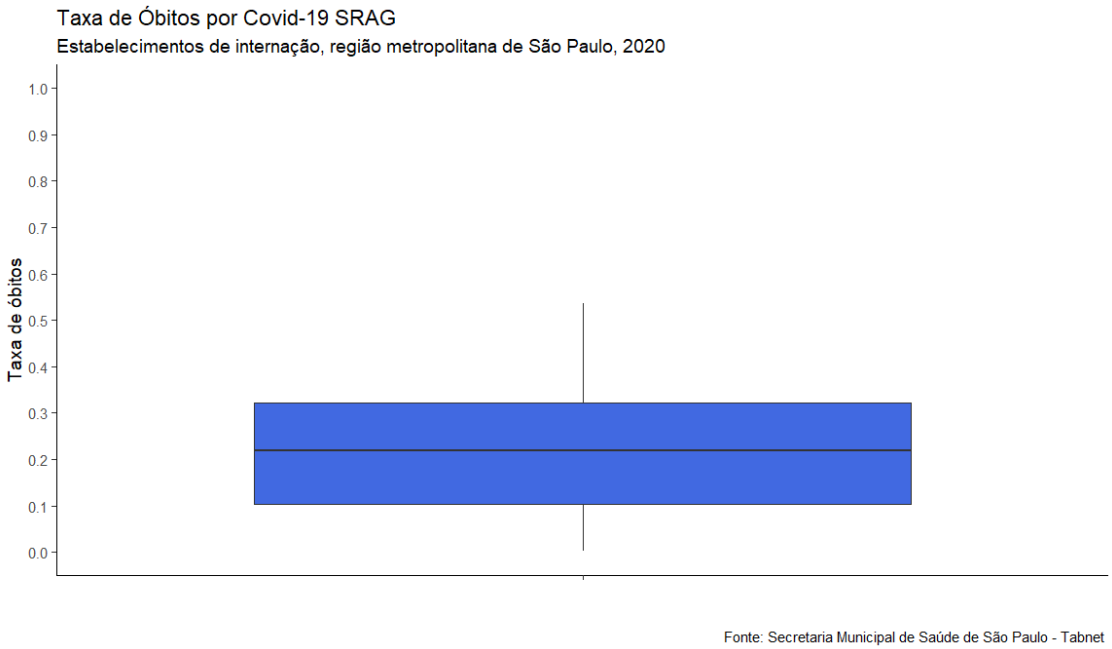
Na figura 4, a linha central marca a mediana do banco de dados. O valor da mediana foi de 21,82%. Dito de outra maneira, em 2020, 50% destes 151 estabelecimentos de internação da região metropolitana de São Paulo registraram uma taxa de óbito por SRAG inferior a 21.82%.

Observe que 25% dos estabelecimentos de internação registraram uma taxa de óbitos por SRAG igual ou maior a 32.15%.

A taxa máxima de óbito registrada em 2020 foi de 53.53% e não se configura como *outiler*.

Estas informações podem ser visualizadas no *boxplot* da figura 4.

Figura 4 – Boxplot da taxa de óbitos



Fonte da figura: Elaborado pelos autores

5.2 Método *K-means*

O algoritmo *K-means* é o algoritmo de agrupamento particional mais utilizado baseado no critério da soma dos quadrados (KASSAMBARA, 2017; MACQUEEN, 1967). Trata-se de um algoritmo simples, fácil de implementação e implementado em quase todos os *softwares* de mineração de dados. Ademais, é muito versátil, considerando os aspectos inicialização, distância, função, critério de finalização.

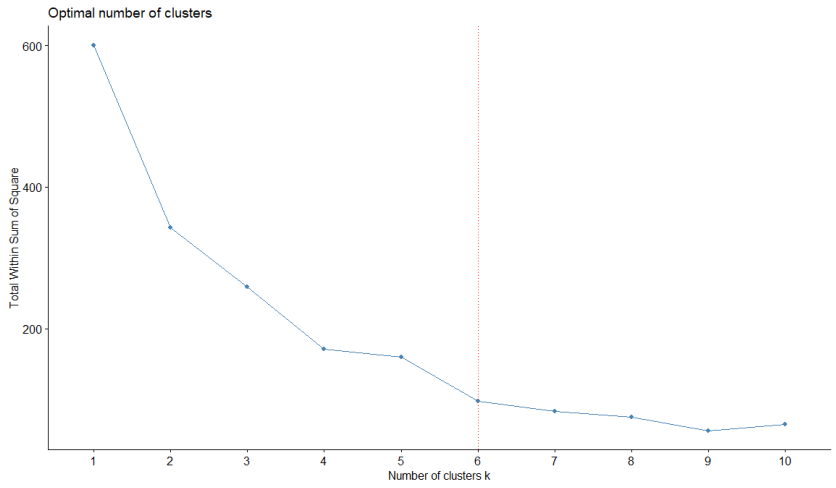
a) Determinação do número de *clusters*

O método *k-means* é um procedimento não hierárquico para agrupar indivíduos, cujo número inicial de *clusters* é definido pelo pesquisador (FÁVERO; BELFIORE, 2017; KASSAMBARA, 2017).

A análise visual a partir de um gráfico de linhas de duas dimensões: os *K Clusters* (*Number of clusters k*) e a Soma dos Quadrados dos Erros de Predição (*Total Within Sum of Square - SSE*) sugere um número inicial de *clusters* para ser usado. O SSE resulta na variância e desvio padrão (inércia) dos dados da base utilizada. Desta forma pode-se visualizar o valor de quão próximo os dados estão uns dos outros. Quanto menor for o número de *clusters*, maior será o valor dessa inércia (KASSAMBARA, 2017).

A figura 5 exibe o gráfico com o número de *clusters* selecionado (linha tracejada vertical vermelha).

Figura 5 – Número de clusters.



Fonte: elaborado pelos autores

O gráfico acima representa a variação dentro dos *clusters*. Nota-se que a partir do quinto *cluster*, o valor do SSE não tem maiores variações. Aumentar o número de *clusters* indica que esta variação se tornaria cada vez menor. Sendo assim, a quantidade de cinco ou seis *clusters* parece ser um número bem interessante para se aplicar nesse conjunto de dados.

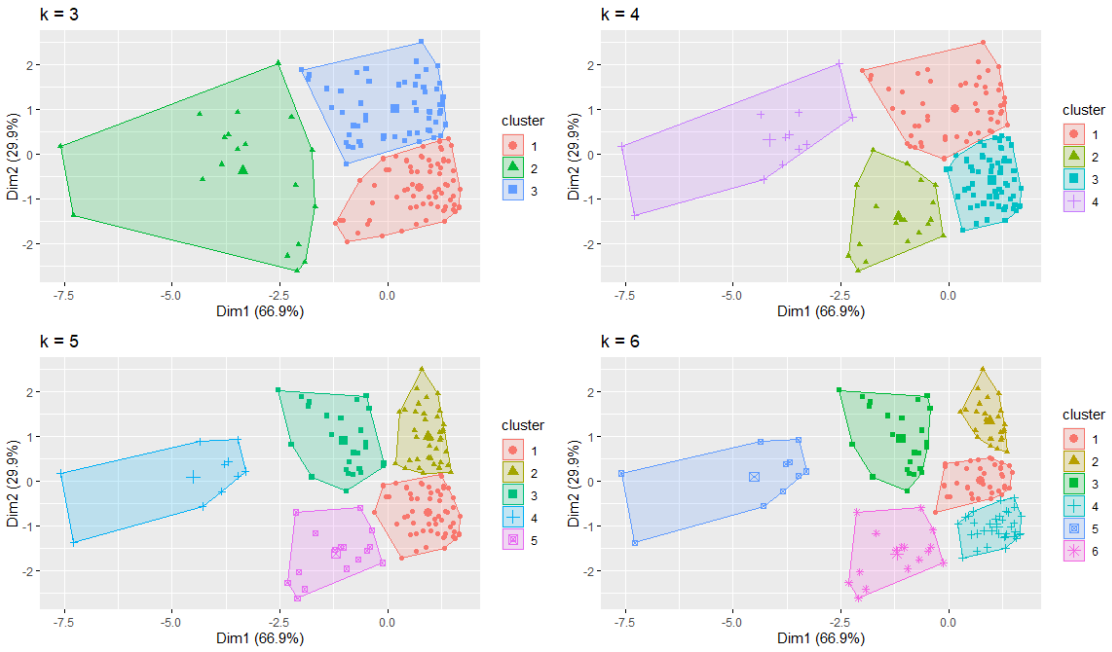
b) Visualização gráfica dos *clusters*

O algoritmo para determinar a alocação das observações em cada conglomerado é denominado *nearest centroid sorting*. O *K-means* usa a distância euclidiana como critério de distância para formar os grupos.

A função fviz_cluster () do pacote fatoextra do R pode ser usada para visualizar facilmente os *clusters k-means*. No gráfico resultante, as observações são representadas por pontos, usando análise de componentes principais (PCA) quando o número de variáveis x_{ij} for maior do que 2 (KASSAMBARA, 2017).

A figura 6 sugere que a formação de seis *clusters* parece ser mais assertiva em alocar os estabelecimentos de internação.

Figura 6 – Segmentação dos estabelecimentos de internação considerando 3, 4 5 ou 6 *clusters*.



Fonte: elaborado pelos autores.

Considerando-se a formação com k = 6 grupos (*clusters*), pode-se perceber que os grupos formados apresentam homogeneidade interna (variabilidade dentro dos grupos), com cada estabelecimento de internação apresentando maior proximidade com outros estabelecimentos do mesmo grupo, do que com estabelecimentos de internação de outros grupos (variabilidade entre os grupos).

c) Validação interna do agrupamento

A validação de agrupamento consiste em projetar o procedimento de avaliação da qualidade dos resultados do algoritmo de agrupamento.

Isso é importante para evitar encontrar padrões em dados aleatórios, bem como na situação em que se deseja a comparação entre dois algoritmos de agrupamento (KASSAMBARA, 2017).

O processo de validação interna usa as informações internas do processo do agrupamento para avaliar a qualidade de uma estrutura de agrupamento, aquém de informações externas. O processo é capaz de estimar o número de **clusters** e o algoritmo de agrupamento apropriado sem nenhum dado externo (KASSAMBARA, 2017).

Medidas internas usam informações intrínsecas nos dados para avaliar a qualidade do agrupamento (KASSAMBARA, 2017).

As medidas internas incluem a **conectividade**, o **coeficiente de silhueta** e o **índice de Dunn**.

A **conectividade** corresponde à medida em que os itens são colocados no mesmo *cluster* que seus vizinhos mais próximos no espaço de dados. A conectividade tem um valor entre 0 e infinito e deve ser minimizada (KASSAMBARA, 2017).

O **coeficiente de silhueta (S_i)** mede a semelhança de um indivíduo y_i com os outros indivíduos em seu próprio *cluster* versus os do *cluster* vizinho (KASSAMBARA, 2017).

Os valores de S_i variam de -1 a +1:

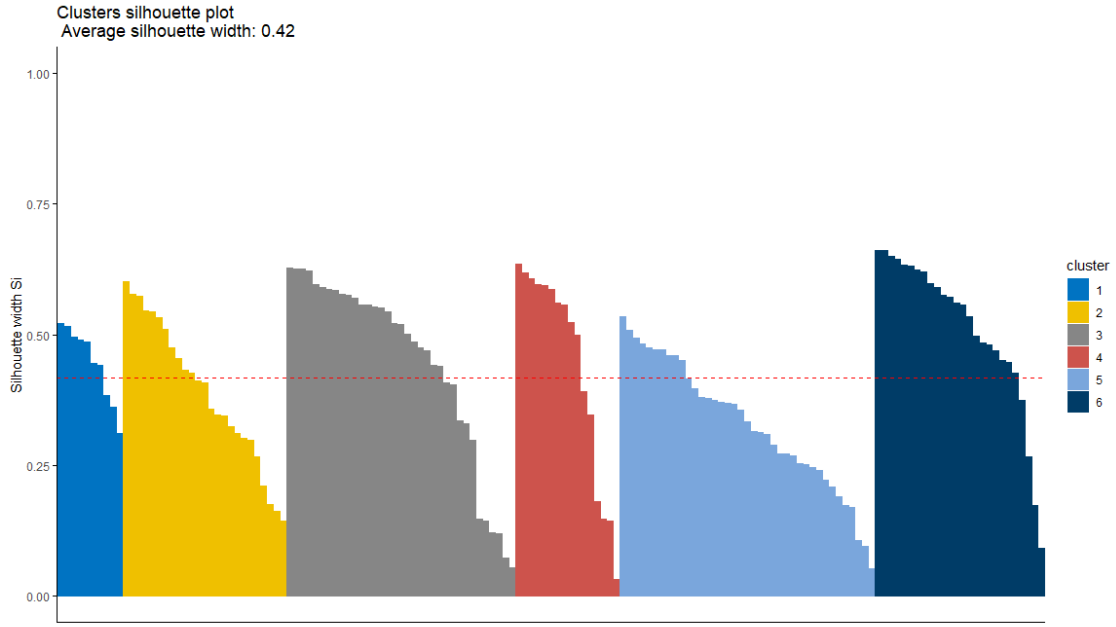
- Um valor de S_i próximo a 1 indica que o indivíduo está bem agrupado. Isto é, o indivíduo y_i é semelhante aos outros indivíduos em seu grupo (KASSAMBARA, 2017).

- Um valor de S_i próximo a -1 indica que o indivíduo está mal agrupado e que a atribuição a algum outro agrupamento provavelmente melhoraria os resultados gerais (KASSAMBARA, 2017).

  A próxima saída do R mostra o valor do S_i dos *clusters* formados e podem ser visualizados na figura 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
```

Figura 7 - Valores do S_i para os seis clusters.



Fonte: Elaborado pelos autores.

Abaixo segue a relação dos 10 estabelecimentos (y_i) com os maiores coeficientes S-i (saída do R)

	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

Não se observa estabelecimentos com S_i negativos. Isso significa que todos os indivíduos estão alocados corretamente nos *cluster* gerados.

O **índice de Dunn** identifica *clusters* densos e bem separados. É definida como a razão entre as distâncias mínimas entre os clusters e a distância máxima entre os clusters.

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

O índice de Dunn tem um valor entre zero e infinito e deve ser maximizado. As pontuações de Dunn com alto valor são mais desejáveis. Nesta análise o valor é muito baixo, sugerindo que talvez não seja um bom agrupamento. Seu valor não deve ser interpretado isoladamente.

d) Avaliação simultânea de algoritmos de agrupamento

O pacote R clValid (BROCK *et al.*, 2008) compara simultaneamente vários algoritmos de agrupamento em uma única chamada de função para identificar a melhor abordagem de agrupamento e o número ideal de agrupamentos (KASSAMBARA, 2017).

```
##
## 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
```

Considerando os métodos de agrupamentos (*hierárquico, kmeans e pam*), os escores de validação interna (**conectivdade, índice de Dunn e o coeficiente de silhueta**) sugerem o método hierárquico e a formação de dois *clusters* para esse conjunto de dados.

e) Medidas de estabilidade

As medidas de estabilidade avaliam a consistência de um resultado de um agrupamento comparando-o com os agrupamentos obtidos após a remoção de cada coluna, um de cada vez (KASSAMBARA, 2017).

As medidas de estabilidade de *cluster* incluem:

- A proporção média de não sobreposição (**APN**);
- A distância média (**AD**);
- A distância média entre as médias (**ADM**):
- A figura de mérito (**FOM**).

O **APN**, **AD** e **ADM** são todos baseados no cruzamento tabela de classificação do agrupamento original nos dados completos com o agrupamento baseado na remoção de uma coluna (KASSAMBARA, 2017).

O **APN** mede a proporção média de indivíduos não colocadas no mesmo *cluster* agrupando com base nos dados completos e agrupando com base nos dados com uma única coluna removida (KASSAMBARA, 2017).

O **AD** mede a distância média entre os indivíduos colocados no mesmo *cluster* em ambos os casos (conjunto de dados completo e remoção de uma coluna) (KASSAMBARA, 2017).

O **ADM** mede a distância média entre os centros dos *clusters* para indivíduos colocadas no mesmo *cluster* em ambos os casos (KASSAMBARA, 2017).

O **FOM** mede a variação intracluster média da coluna excluída, em que o agrupamento é baseado nas colunas restantes (não excluídas) (KASSAMBARA, 2017).

Os valores de **APN**, **ADM** e **FOM** variam de 0 a 1, com valor menor correspondendo a resultados de agrupamento altamente consistentes. **AD** tem um valor entre 0 e infinito, e valores menores também são preferidos (KASSAMBARA, 2017).

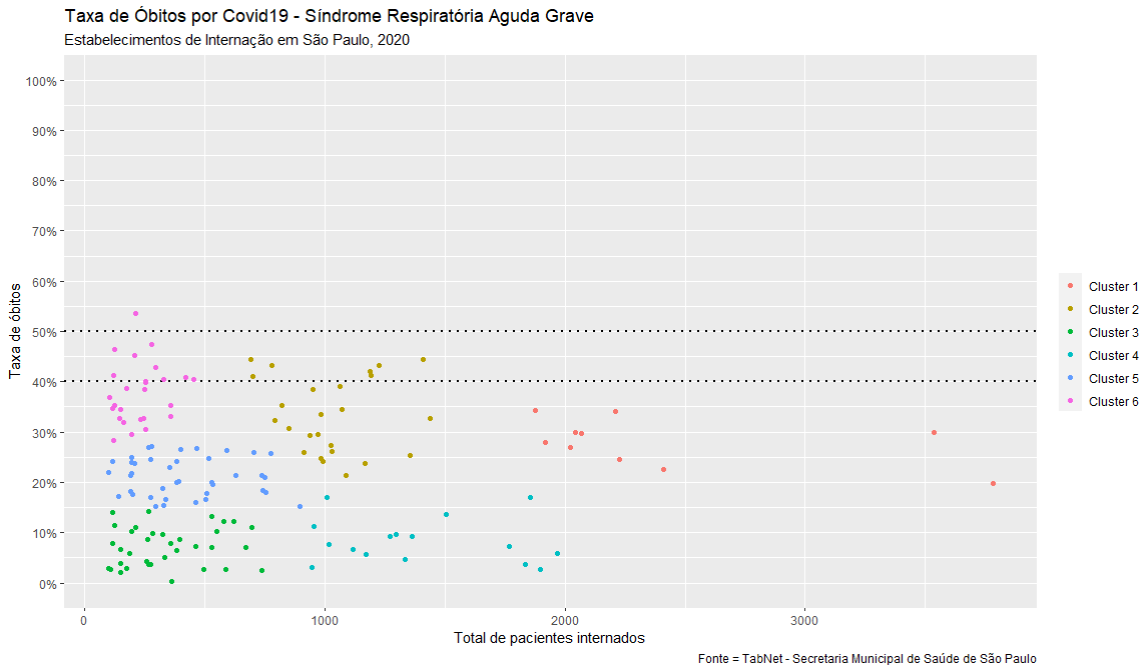
##	Score	Method	Clusters
## APN	0.02119205	hierarchical	2
## AD	1.04508486	pam	6
## ADM	0.26245203	pam	6
## FOM	0.52181358	kmeans	6

Observa-se que três medidas de estabilidade sugerem um número de 6 *clusters* para esse conjunto de dados. A medida **FOM** sugere o método *kmeans* para aplicar no algoritmo de agrupamento. Esta medida, portanto, ratifica a escolha do método *k-means* com 6 *clusters* usado nesta pesquisa para agrupar os 151 estabelecimentos de internação identificados no TABNET.

6 Discussão

Considerando os aspectos técnicos, o algoritmo *k-means* agrupou os 151 estabelecimentos de internação conforme a figura 8.

Figura 8 – Taxa de óbitos por SARG, em 2020, São Paulo.



Fonte da figura: elaborada pelos autores.

Observamos que os estabelecimentos de internação se distribuem em 6 *clusters*, considerando a taxa de óbito em função do número total de internações de pacientes com SRAG em suas dependências, em 2020, como segue:

- *Cluster 1* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 21,2% (3,77%). Há uma baixa variabilidade intracluster, com um coeficiente de variação de 17.80%.
- *Cluster 2* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 37,8% (6,08%). Há uma baixa variabilidade intracluster, com um coeficiente de variação de 16.21%.
- *Cluster 3* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 33,4% (7,49%). Há uma moderada variabilidade intracluster, com um coeficiente de variação de 22.40%.
- *Cluster 4* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 7,21% (4,66%). Há uma alta variabilidade intracluster, com um coeficiente de variação de 53.70%.
- *Cluster 5* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 28.0% (4,66%). Há uma baixa variabilidade intracluster, com um coeficiente de variação de 16.70%
- *Cluster 6* - Este *cluster* representa os estabelecimentos de internação com uma taxa média de óbitos de 8.39% (4.50%). Há uma alta variabilidade intracluster, com um coeficiente de variação de 53.60%

As linhas horizontais pontilhadas em cor preta representam os valores descritos na literatura para a mortalidade geral, de moderada a grave, por SRAG, entre 40 e 50% (VILLAR; BLANCO; KACMAREK, 2016; VILLAR; SULEMANJI; KACMAREK, 2014).

Observa-se, portanto, que os estabelecimentos de internação registrados na Secretaria Municipal de Saúde de São Paulo, em 2020, apresentaram taxas de óbito por SARG condizente com as taxas de óbitos na literatura. Ou seja, não houve um aumento da taxa de óbito por SARG-Covid 19, em São Paulo, além do esperado. A maioria destes 151 estabelecimentos teve uma taxa de óbito abaixo de 40%.

7 Considerações finais

Esta pesquisa identificou 151 estabelecimentos de internação na cidade de São Paulo que registraram pelo menos 100 internações por SARG-Covid 19 no ano de 2020. A pesquisa revelou que a taxa de óbitos nestas instituições permaneceu semelhante à descrita na literatura, entre 40 e 50% para as formas moderada a grave da doença. Apenas 25% dos 151 estabelecimentos de internação registraram uma taxa de óbitos igual ou superior a 32.15% em 2020.

A medida FOM indicou o método *kmeans* com seis *clusters* para aplicar como algoritmo de agrupamento. Não se observou estabelecimentos com S_i negativos. Ou seja, eles foram alocados corretamente nos seis *clusters* gerados.

Uma possível linha de pesquisa seria a comparação dos perfis epidemiológicos e sociodemográficos de cada *cluster* gerado para entender o comportamento da doença.

Referências

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. clValid: An R package for cluster validation. **Journal of Statistical Software**, v. 25, p. 1–22, 2008.

CHIN, W. W. How to write up and report PLS analyses. *Em*: **Handbook of partial least squares**. [s.l.] Springer, 2010. p. 655–690.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata**. [s.l.] Elsevier Brasil, 2017.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, v. 2, n. 1, p. 193–218, 1985.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning**. [s.l.] Sthda, 2017. v. 1

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. [s.l.] John Wiley & Sons, 2009.

MACQUEEN, J. Classification and analysis of multivariate observations. Em: 5th Berkeley Symp. Math. Statist. Probability, 1967, [...]. 1967. p. 281–297.

RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical association**, v. 66, n. 336, p. 846–850, 1971.

ROZEBOOM, W. W. Meehl on metatheory. **Journal of clinical psychology**, v. 61, n. 10, p. 1317–1354, 2005.

VILLAR, J.; BLANCO, J.; KACMAREK, R. M. Current incidence and outcome of the acute respiratory distress syndrome. **Current opinion in critical care**, v. 22, n. 1, p. 1–6, 2016.

VILLAR, J.; SULEMANJI, D.; KACMAREK, R. M. The acute respiratory distress syndrome: incidence and mortality, has it changed? **Current opinion in critical care**, v. 20, n. 1, p. 3–9, 2014.