UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
Acesso aberto

Ci. e Nat., Santa Maria, v. 43, Ed. Esp. X ERMAC RS, e1, 2021 • https://doi.org/10.5902/2179460X66864
Submissão: 27/07/2021 • Aprovação: 29/09/2021 • Publicação: 05/11/2021

Classificação de risco em redes complexas: o caso da COVID-19 no Rio Grande do Sul

Risk classification in complex networks: the case of COVID-19 in Rio Grande do Sul

Lucas Siviero Sibemberg[I], **Luiz Emilio Allem**[I], **Carlos Hoppen**[I], **Pedro da Silva Peixoto**[II]

[I]Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, RS, Brasil
[II]Universidade de São Paulo, São Paulo, SP, Brasil

RESUMO

Peixoto *et al.* (2020) apresentaram uma metodologia que utiliza dados oriundos da mobilidade de pessoas para classificar municípios de um estado em três zonas de risco (baixo, médio e alto) em relação a uma doença contagiosa transmissível pelo ar. Os autores aplicaram o modelo para avaliar a exposição dos municípios dos estados de São Paulo e Rio de Janeiro à COVID-19, antes que o vírus estivesse amplamente disseminado. Nosso objetivo neste artigo é avaliar como essa metodologia de classificação de risco se aplica no Rio Grande do Sul e como ela se relacionou com a evolução da COVID-19 no estado. A partir dela, obtivemos uma classificação dos municípios do estado em três grupos de risco. Nessa divisão, com raras exceções, municípios mais próximos de Porto Alegre ficaram classificados como alto risco. A região da serra, do litoral e de alguns municípios no oeste do estado ficaram em um risco médio. Os demais municípios foram classificados com risco baixo. Em comparação com os dados oficiais sobre a disseminação da doença no estado verificamos que o risco atribuído foi coerente com a evolução da COVID-19. De um ponto de vista metodológico, encontramos evidências, via clusterização espectral, que dividir os municípios em três grupos é a melhor escolha para os nossos dados.

Palavras-chave: Biomatemática; COVID-19; Dinâmica espacial; Clusterização espectral

ABSTRACT

Peixoto *et al.* (2020) developed a methodology that uses data from the mobility of people to classify municipalities in a state into three risk zones (low, medium and high) in relation to airborne contagious disease. The authors applied the model to the states of São Paulo and Rio de Janeiro, before the COVID-19 virus was present in a large number of municipalities in the state. Our objective in this article is to evaluate how this risk classification methodology applies to Rio Grande do Sul and how it relates to the actual evolution of COVID-19 in the state. Using the methodology, we obtained a classification of the

municipalities in the state into three risk groups. According to this classification, with rare exceptions, municipalities closer to Porto Alegre have been classified as high risk. The mountain region, the coast and some municipalities in the west of the state have been classified as medium risk. The other municipalities have been classified as low risk. In comparison with official data on the spread of the disease in the state, we found that the risk attributed was consistent with the evolution of COVID-19. From a methodological point of view, we found evidence, via spectral clustering, that dividing the municipalities into three groups is the best choice for our data.

Keywords: Phytotherapy; Health; Pharmaceutical preparations; Clinical trial

1 INTRODUÇÃO

O ano de 2020 foi marcado pelo surto global do vírus SARS-CoV-2, que causa a doença COVID-19 em humanos. Em março daquele ano, a Organização Mundial da Saúde (OMS) declarou a COVID-19 como pandemia mundial. A característica principal, e mais preocupante, dessa doença é o fato de ser altamente contagiosa, o que torna difícil barrar a sua propagação e levou governos mundo afora a tomar diversas medidas, desde o fechamento obrigatório de escolas e atividades econômicas até a proibição de viagens entre países. A COVID-19 não demorou muito para chegar, e se espalhar, no Brasil e logo em março de 2020 já havia transmissão comunitária no país (Ministério da Saúde (2020)). No Rio Grande do Sul, em 19 de março, foi declarado estado de calamidade pública (Rio Grande do Sul (2020)). Reconhecendo a gravidade da doença e o desconhecimento sobre seu prognóstico e virulência, muitos cientistas e autoridades de saúde empregaram métodos epidemiológicos para entender o seu comportamento. Os modelos epidemiológicos são amplamente utilizados e há uma vasta literatura relacionada a eles, como por exemplo modelos compartimentais, Brauer (2008). No enfrentamento da COVID-19 os modelos compartimentais também foram de suma importância para compreender a transmissão da doença, como em Linka *et al.* (2020),Senajith *et al.* (2020) e Yoav e Granek (2021). Nessas abordagens, um dos objetivos é entender como a doença se propagaria em uma região no momento

em que os primeiros casos são identificados com base em diversos parâmetros hipotéticos de transmissão e em diversos cenários de redução de mobilidade.

Nessa linha, Peixoto *et al.* (2020) investigaram como dados de mobilidade populacional podem ser empregados para construir uma classificação de risco relacionada à COVID-19. Eles propuseram uma metodologia que foi aplicada para os municípios dos estados de São Paulo e Rio Janeiro, obtendo uma classificação de risco para esses municípios em três categorias, denominadas risco baixo, médio ou alto. No contexto dessa metodologia, a palavra risco está associada ao tempo esperado até que os primeiros casos sejam identificados em um município, isto é, quanto maior o risco, mais rápido a doença tenderá a atingir no município. Para os dados de mobilidade social, foram utilizados dados de deslocamento baseados na posição geográfica de telefones celulares em todo o Brasil.

De forma geral, esse trabalho busca estender essa abordagem a mais um estado, além de avaliar a qualidade da predição nesse caso específico e investigar aspectos que fundamentam essa metodologia. Nosso primeiro objetivo específico foi avaliar essa metodologia no estado do Rio Grande do Sul. O segundo objetivo desse trabalho foi comparar a classificação que o modelo propôs com a evolução da COVID-19 registrada pelos dados oficiais. Um terceiro objetivo foi comparar as técnicas de particionamento utilizadas na metodologia com uma abordagem espectral, onde avaliamos a escolha do número de classes da classificação de risco.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta a metodologia utilizada. Na Seção 3, descrevemos os dados usados neste artigo. A Seção 4 apresenta nossos resultados e utiliza métricas para avaliar a classificação de risco obtida e compará-la com os resultados de Peixoto *et al.* (2020). Concluímos o artigo com as considerações finais.

2 METODOLOGIA

Neste trabalho, aplicamos a abordagem proposta por Peixoto et al. (2020) a dados do estado do Rio Grande do Sul. A seguir apresentamos essa metodologia, que se baseia em um modelo epidemiológico Suscetível - Infectado (SI), ao qual são incorporados dados de mobilidade entre municípios. O modelo SI é uma forma simplificada de descrever a transmissão inicial de doenças em populações e considera que todos os indivíduos podem ser separados em duas categorias, ou compartimentos, os que ainda não contraíram a doença e permanecem vulneráveis (suscetíveis) e os que contraíram a doença e têm a capacidade de transmiti-la (infectados). Modelos SI não prevêem que haja recuperação ou mortalidade, não sendo adequados para previsões de longo prazo. Nesse sentido, tais modelos são indicados para capturar a evolução inicial de uma doença, particularmente quando poucos dados sobre ela são conhecidos, o que é compatível com os objetivos desse trabalho.

A propagação da doença será modelada através de relações de recorrência com duas componentes, uma em que o contágio acontece entre indivíduos de um mesmo município i e outro que acontece entre indivíduos de municípios i e j diferentes. Dentro de cada município i os novos casos são obtidos por um modelo SI tradicional, enquanto que entre os municípios i e j são incorporados os dados de mobilidade. No modelo, o número de novos casos no dia $t+1$ no município i é dado por

$$I_i(t+1) = I_i(t) + (1+r)I_i(t)\left(\frac{N_i - I_i(t)}{N_i}\right) + \left[\sum_{i\neq j}\omega_{ji}(t)I_j(t) - \sum_{i\neq j}\omega_{ij}(t)I_i(t)\right], \qquad (1)$$

Onde r é a taxa de transmissão da doença, s é um parâmetro de correção para os dados de mobilidade, $I_i(t)$ é a quantidade de infectados no dia t e N_i é a

população do município. O modelo considera que a taxa de transmissão da doença e a população de cada município independem do tempo. O termo $\omega_{ij}(t)$ denota a proporção da população de *i* que se desloca para j no tempo t, isto é, a razão $\omega_{ij}(t) = \frac{w_{ij}(t)}{N_i}$ onde $W(t) = (W_{ij}(t))$ é a matriz cuja entrada ij representa a quantidade de pessoas que se deslocaram do município i para o município j no dia t. No contexto de Peixoto *et al.* (2020), os dados que compõem essa matriz são dados anônimos de mobilidade baseados na geolocalização de telefone celulares, disponibilizados pela empresa *In Loco* (o significado desses dados será explicado de forma mais detalhada na Seção 3). O papel de s é corrigir eventuais imprecisões nos dados de mobilidade utilizados. Por um lado, como os dados são anônimos, se uma pessoa faz várias viagens em um mesmo dia, essas viagens são contadas separadamente, valores de s < 1 suporiam que a estimativa está acima do número verdadeiro de viagens. Por outro lado, existem viagens que não são capturadas pelo sistema da *In Loco*, valores de s > 1 suporiam que a estimativa está abaixo do número verdadeiro de viagens. Finalmente, s = 1 significaria que os dados retratam o que de fato aconteceu. Ainda vale notar que, nesse modelo, as variáveis $I_i(t)$ podem assumir valores não inteiros.

Para dividir os municípios no conjunto $M = \{m_1, \dots, m_n\}$ em três grupos de risco, consideram-se diferentes valores de s e, para cada um desses valores, determina-se a evolução da doença aplicando a recorrência 1 com um caso inicial na capital do estado. Cada município i é associado a um vetor vi que, para cada valor de s considerado, identifica o primeiro dia em que o município i contabiliza pelo menos um caso. Esses vetores são divididos em três grupos de risco a partir de um algoritmo de clusterização. Ao final, o risco atribuído a cada município vem do grupo que contém o vetor associado a ele.

Para realizar a clusterização de vetores, um algoritmo muito utilizado é o *k-means* (que pode ser visto com mais detalhes em Hastie *et al.* (2001)). O *k-means* é

um algoritmo que, a partir de uma clusterização inicial, realiza iterativamente dois passos. O primeiro é calcular quem são os centroides de cada *cluster*[1] onde o centroide de um *cluster* é o ponto médio entre os pontos do *cluster*. O segundo é recalcular os clusters, onde cada ponto é designado para o cluster do centroide mais próximo. Esses dois passos são iterados até que não haja mais mudança nos *clusters*, e o algoritmo retorna a última partição obtida. Como o *k-means* é dependente da clusterização inicial, é natural realizar o algoritmo uma quantidade Q de vezes em busca de um particionamento mais estável ou um particionamento que otimize alguma função custo.

O algoritmo para a obter a classificação de risco é apresentado abaixo

Algoritmo 1: Classificação de Risco

Input: Conjuntos $M = \{m_1, m_2, \ldots, m_n\}$ e $N = \{N_{m_1}, \ldots, N_{m_n}\}$. Vetor $S = (s_1, s_2, \ldots, s_K)$ de números positivos. Inteiros positivos T e Q, constante real $r \geq 0$. Matrizes $n \times n$ $(W(t))$, para $1 \leq t \leq T$.

Output: Partição $\mathscr{P} = \{A, B, C\}$ de M, onde os municípios em A, são classificados como alto risco, em B médio e em C baixo.

Defina um vetor

$$v_i = (v_i^1, \ldots, v_i^K) \in \mathbb{R}^K \tag{2}$$

com o valor $T + 1$ em todas entradas, para cada $i \in M$.

for $\ell \in (1, \ldots, K)$ **do**
  Atribua $I_{m_1}^{s_\ell}(0) = 1$ e $I_{m_i}^{s_\ell}(0) = 0, \forall i \in \{2, \ldots, n\}$
  for $0 \leq t \leq T - 1$ **do**
    Para cada $i \in M$, defina $I_i^{s_\ell}(t + 1)$, por meio da equação de recorrência 1.
    Para cada $i \in M$, se $I_i^{s_\ell}(t + 1) \geq 1$ e $v_i^\ell = T + 1$, redefina $v_i^\ell = t + 1$.
  end
end

Separe o conjunto $V = \{v_1, v_2, \ldots, v_n\}$ em três conjuntos de risco C_A, C_B, C_C (alto, médio e baixo, respectivamente) via algoritmo de clusterização *k-means*.

Seja $\mathscr{P}$ a partição tal que $i \in M$ é atribuído ao *cluster* γ se, e somente se, v_i encontra-se em C_γ.

No trabalho de Peixoto *et al.* (2020) foram utilizados os seguintes parâmetros. Os conjuntos *M* e *N* foram constituídos pelos municípios de cada estado e suas populações, respectivamente. As matrizes *W (t)* continham os dados de mobilidade. Para o valor da taxa de transmissão foi utilizado r = 0.4, que é aproximadamente $\frac{R_0}{6}$, onde R_0 = 2.68 é o número efetivo de reprodução da COVID-19, calculado com

[1] Utilizaremos *cluster* como sinônimo os subconjuntos que particionam algum conjunto.

base nos dados sobre a doença em Wuhan, China, e 6 é o tempo médio, em dias, de incubação da doença, conforme Wu *et al.* (2020). Além disso, foi utilizado o vetor

$$S = (0.001, 0.005, 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.5, 3) \quad (3)$$

3 DADOS

Nessa seção, descrevemos os dados utilizados no nosso trabalho. Nós consideramos os n = 167 municípios do Rio Grande do Sul cuja população estimada para 2019 é maior do que 10.000, de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE[2]).

Os municípios do estado que satisfazem esse critério estão marcados no mapa da Figura[3]. O valor de N_i é essa população estimada para o município *i*.

Os dados de mobilidade utilizados neste trabalho foram disponibilizados pela empresa *In Loco*[4] que coletou dados anônimos de localização de usuários de telefones celulares que compõem uma base de dados. No sistema da empresa, é considerado que uma pessoa se deslocou dentro de seu município caso haja pelo menos pelo menos dois registros dela em pontos distantes de pelo menos 450 metros e que uma pessoa se deslocou do município i para o município *j* ($i \neq j$) se há registros da mesma em *i* e *j*, nesta ordem. Esses dados permitem uma estimativa do número de deslocamentos de um município *i* para um outro município *j* no dia *t*, com *i*, *j* ∈ $\{m_1, \ldots, m_n\}$. Assim, construímos a matriz $W(t) = (W_{ij}(t))$ de ordem n, onde $W_{ij}(t)$ é o número de pessoas que se deslocaram do município i para o município j no dia t segundo esses dados.

[2] Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao. Acesso em: 14 ago. 2020.
[3] A lista completa pode ser encontrada em https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados. Acesso em: 26 jun. 2021.
[4] Disponível em: https://www.inloco.com.br/covid-19. Acesso em: 13 out. 2020.

Como o objetivo do modelo é prever uma classificação de risco para a dispersão inicial da doença, foram utilizados dados de mobilidade em uma época de normalidade, isto é, em que não havia redução de mobilidade por conta de medidas voluntárias ou obrigatórias de isolamento social. Portanto, para aplicar esse modelo, utilizamos dados do período anterior a quaisquer medidas de isolamento. Além disso, como o critério de classificação está baseado na ocorrência do primeiro caso, foi necessário escolher um período suficientemente longo para que o modelo atribuísse um primeiro caso a cada município. No nosso caso, foram necessários T = 38 dias para que isso ocorresse. Em virtude disso, utilizamos dados dos 38 dias entre 1º de fevereiro e 14 de março de 2020, que foi o dia em que o isolamento voluntário foi encorajado e os dados de mobilidade mostraram alterações por conta do isolamento social.

No período mencionado acima, a média dos registros do número de deslocamentos diários, no Rio Grande do Sul, foi de aproximadamente 577 mil, que corresponde a aproximadamente 5,8% da população do estado. A partir desses dados, pode-se identificar que ocorreram deslocamentos em todos os n municípios, com cobertura relativamente maior na região metropolitana, no litoral norte e na serra. Se considerarmos a proporção do total de registros de pessoas que se deslocaram em cada município pela população do município, a proporção das regiões metropolitana, litoral e serra, em conjunto, ficou aproximadamente 7,1%, enquanto que no resto do estado foi de 3,7%. Com exceção de três municípios, todos tiveram pelo menos a proporção de 1,0% de registros de deslocamentos diários em relação à população. Para efeitos de comparação, os dados utilizados por Peixoto *et al.* (2020) resultaram em uma média diária de registros de deslocamento de 4,3 milhões para o estado de São Paulo e de 800 mil para o estado do Rio de Janeiro, o que corresponde a aproximadamente 9,5% e 4,8% da população de cada estado, respectivamente.

Para realizar a comparação entre o risco atribuído pelo modelo e o momento em que o primeiro caso foi registrado em cada município, utilizamos os registros oficiais de casos mantidos pela Secretaria da Saúde do Estado do Rio Grande do Sul (Secretaria Estadual de Saúde (RS) (2020)).

4 RESULTADOS

Essa seção tem como objetivo apresentar a classificação de risco obtida para o estado do Rio Grande do Sul e compará-la com as classificações obtidas para outros estados, assim como com a evolução da COVID-19 segundo os dados oficiais. Por fim, temos o objetivo de comparar a técnica de particionamento descrita na Seção 2 com uma abordagem espectral, onde foi avaliada a escolha do número de classes da classificação de risco.

4.1 Classificação de risco no Rio Grande do Sul

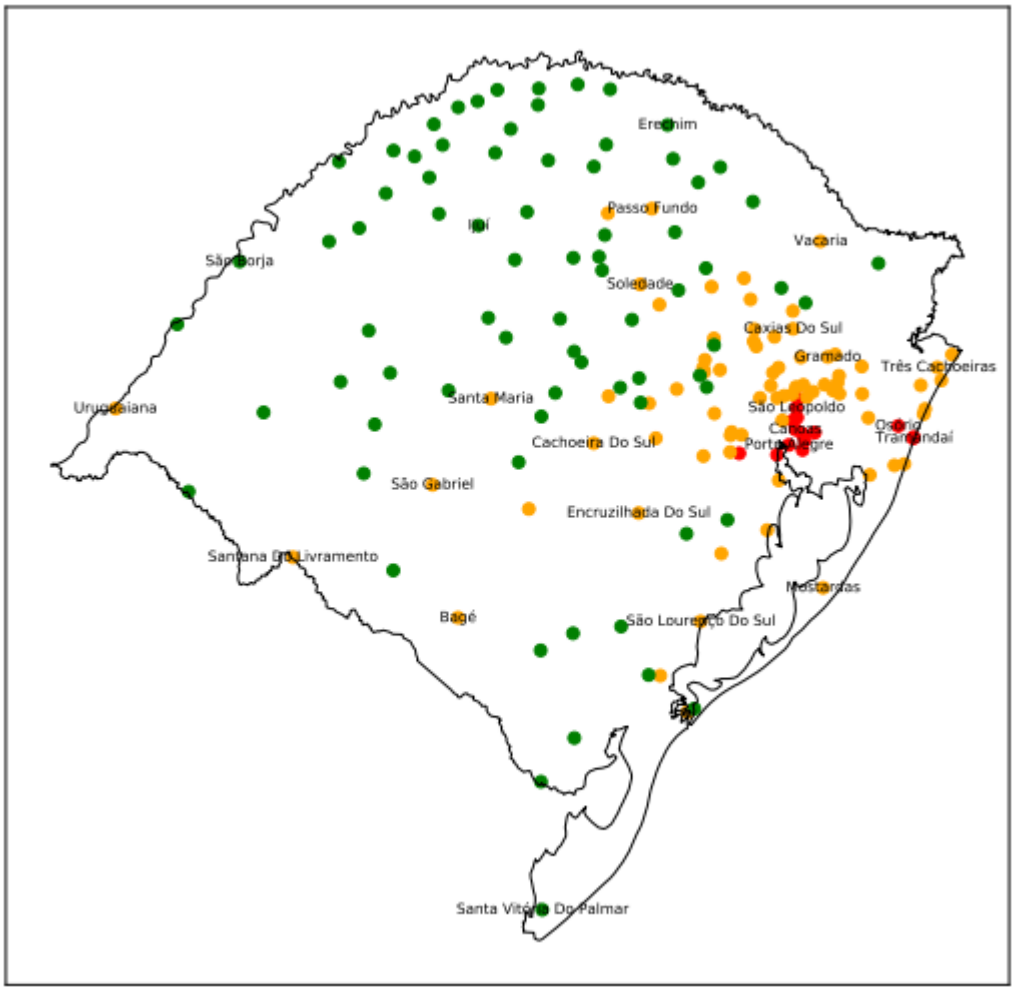
Para realizar o primeiro objetivo desse trabalho, que é definir a classificação de risco para os municípios do Rio Grande do Sul com base na metodologia da Seção 2, aplicamos o Algoritmo 1 para os parâmetros epidemiológicos utilizados por Peixoto *et al.* (2020) e os dados de população e deslocamento descritos na Seção 3. Obtivemos a classificação de risco ilustrada na Figura 1, utilizando $Q = 500$.

Para avaliar como a classificação de risco obtida se relacionou com os dados oficiais sobre a pandemia, definimos um índice que é dado pela proporção de pares (i, j) de municípios, onde o município i, para o qual o modelo atribuiu risco maior, registrou o seu primeiro caso oficial em um tempo inferior ou igual do município j. Primeiro, seja $d = (d_1,....d_n)$ vetor, tal que d_i é o número de dias entre a

data do primeiro caso oficial no estado e o dia que i registrou seu primeiro caso oficial. Definimos a função $f: M \to \{1,2,3\}$, onde

$$f(i) = \begin{cases} 3, se\ i\ \in A \\ 2, se\ i\ \in B \\ 1, se\ i \in C \end{cases}$$

Figura 1 – Divisão dos n = 167 municípios do Rio Grande do Sul em três regiões: risco alto (vermelho), risco médio (laranja) e risco baixo (verde), aplicando a metodologia da Seção 2 aos dados da Seção 3



Fonte: Os autores (2021)

e A, B, C são os clusters de risco alto, médio e baixo, respectivamente. Sejam $C^{RS} = \{(i,j) \in M \times M | f(i) > f(i)\ e\ d_i \leq d_j\}$ e $T^{RS} = \{(i,j) \in M \times M | f(i) > f(i)\}$. Definimos P^{RS} da seguinte forma:

$$p^{RS} = \frac{|C^{RS}|}{|T^{RS|}}$$

Note que, quando P^{RS} está mais próximo de 1, maior é a quantidade de pares de municípios em que o modelo atribuiu risco maior aos municípios que registraram os primeiros casos da doença. Realizando os cálculos para a classificação de risco apresentada na Figura 1 foi obtido P^{RS} = 0;7521. Também é possível definir o índice restrito a cada duas classificações de risco, dessa forma dados $X \neq Y \in \{A, B, C\}, sejam\ C_{X,Y}^{RS} = \{(i,j) \in X \times Y | f(i) > f(j)\ e\ d_i \leq d_j\}\ e\ T_{X,Y}^{RS} = \{i,j\} \in X \times Y | f(i) > f(j)\}$.

Definimos $P_{X,Y}^{RS}$ como a razão

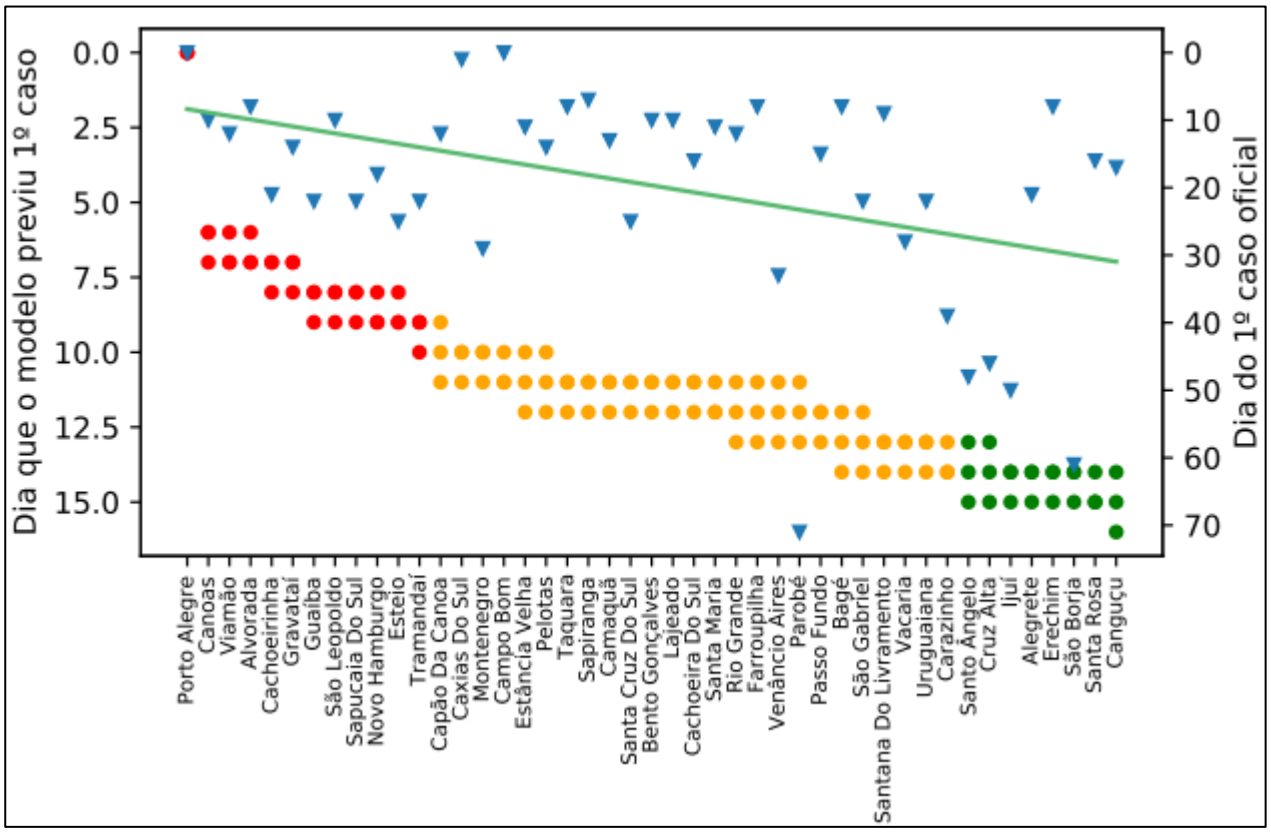
$$P_{X,Y}^{RS} = \frac{|C_{X,Y}^{RS}|}{|T_{X,Y}^{RS}|}$$

Assim, $P_{A,B}^{RS}$ é a proporção de pares (*i,j*) de municípios onde além do modelo atribuir risco alto a *i* e médio a *j*, *i* registrou o primeiro caso oficial antes de *j*. Nós obtivemos $P_{A,B}^{RS} = 0{,}7662, P_{A,C}^{RS} = 0{,}9248\ e\ P_{B,C}^{RS} = 0{,}7182$. Esses resultados revelam uma correlação positiva entre o risco atribuído ao município e o tempo que ele levou para se contaminar.

Uma segunda comparação que fizemos foi relacionar o valor de v_i^l (definido na equação 2) para municípios com mais de 50 mil habitantes para alguns valores do vetor *S*, definido na equação 3, e o tempo que levaram para confirmar um caso pelos dados oficiais, dado pelo vetor *d*. Note que como não havia consenso científico sobre o valor de R_0 e devido às limitações do modelo utilizado, estamos novamente interessados na ordem relativa entre os municípios com relação ao risco atribuído e a data de registro do primeiro caso, e não aos valores específicos de v_i^l. Essa comparação pode ser vista no gráfico na Figura 2. Nós utilizamos o

método dos mínimos quadrados para investigar se os valores v_i^l ficaram proporcionais aos valores d_i.

Figura 2 – Os pontos se referem aos valores de v_i^l para l tal que $s_l \in \{0.5,1,1.6\}$, isto é, o número de dias até que o modelo previsse o primeiro caso no município i para os três valores de s_l, os triângulos se referem ao dia d_i em que o município i registrou o primeiro caso oficial de COVID-19. A reta é uma aproximação dos triângulos por meio do método dos mínimos quadrados. As cores se referem ao nível de risco do município de acordo com a modelagem. Consideramos apenas os municípios com mais de 50 mil habitantes



Fonte: Os autores (2021)

Mais precisamente, considerando $M_0 = \{i \in M | N_i \geq 50.000\}$ seja $u = (u_1, \ldots, u_{|M_0|}$ o vetor tal que $u_j = \frac{1}{K}\sum_{l=1}^{K} v_{mj}^l$, para cada $m_j \in M_0$, e o ordenamento $(r_1, \ldots, r_{|M_0|})$ definido recursivamente da seguinte forma:

$i)$ $r_1 = 1$. Redefine $u_{r_1} = T + 1$.

$ii)$ $r_l = argmin\{u_i | m_i \in M_0\}$. Se existe $j \neq i$ tal que $u_i = u_j$ é escolhido *i* tal que N_{m_i} é máximo. Redefine $u_{r_l} = T + 1$.

Considerando o conjunto de dados $\mathbb{X} = \{(r_i, d_{m_i})\} | m_i \in M_0\}$, a reta que melhor aproxima os dados de $\mathbb{X}$ pelo método dos mínimos quadrados (veja Hastie *et al.* (2001)), em verde na Figura 2. Isso reforça a hipótese de que a evolução da doença está relacionada com o nível de risco atribuído pelo modelo.

Para efeito de comparação, calculamos os índices correspondentes para as classificações de risco obtidas para os estados de São Paulo e Rio de Janeiro por Peixoto *et al.* (2020), que podem ser vistas na Figura 3 e na Figura 4, respectivamente.

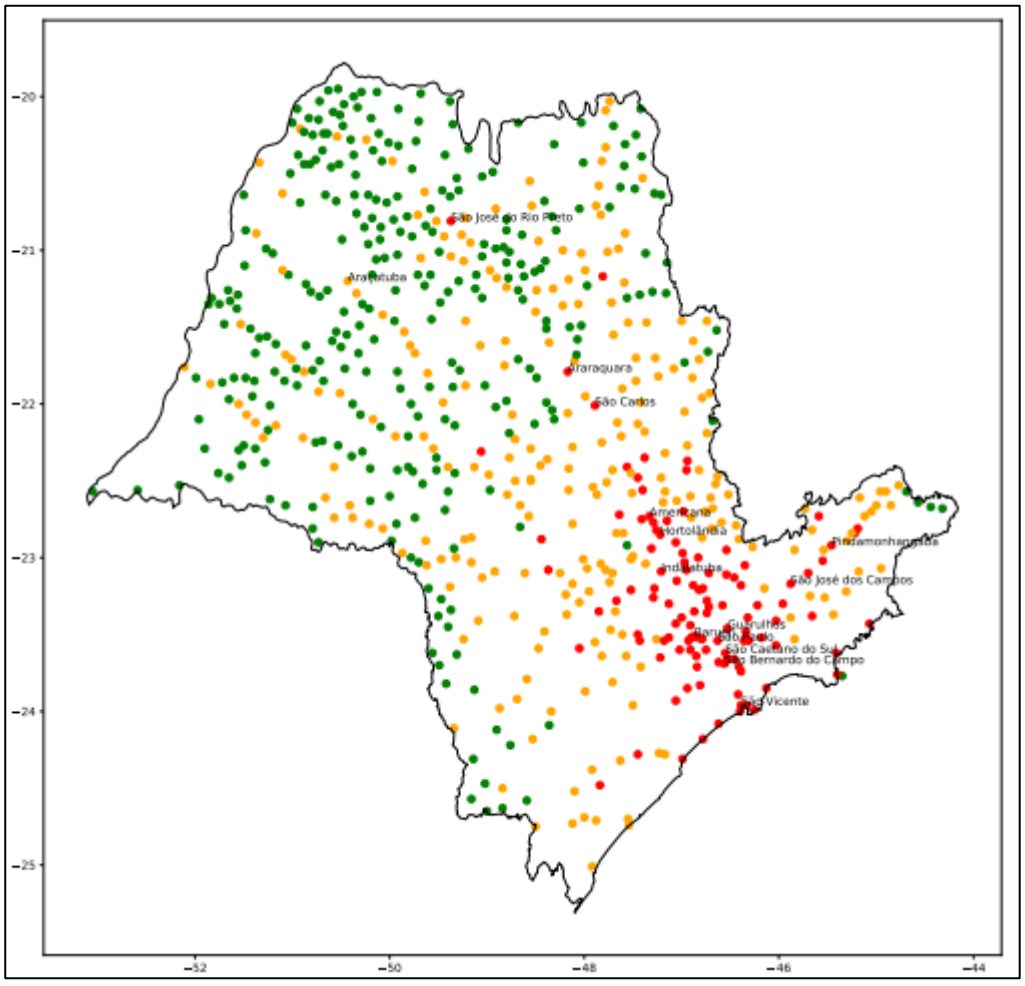
No cálculo dos índices P e $P_{\alpha,\beta}$ para São Paulo o resultado obtido foi de $P^{SP} = 0{,}8562, P^{SP}_{A,B} = 0{,}8535, P^{SP}_{A,C} = 0{,}9750$ e $P^{SP}_{B,C} = 0{,}8032$. Já para o Rio de Janeiro, os valores foram de $P^{RJ} = 0{,}8536, P^{RJ}_{A,B} = 0{,}7764, P^{RJ}_{A,C} = 0{,}9898$ *e* $P^{RJ}_{B,C} = 0{,}8359$.

Primeiramente, notamos que, assim como para o Rio Grande do Sul, os índices sugerem que a evolução inicial da doença está relacionada com a classificação de risco obtida pelo modelo.

Quando se compara os valores de SP e RJ com os do RS, nota-se que os números deste último foram relativamente mais baixos. Levantamos dois fatores que podem ter influenciado nisso. O primeiro ponto é que, em nossa abordagem, consideramos apenas os municípios com mais de 10 mil habitantes, enquanto que as avaliações dos outros dois estados foram feitas para todos os municípios. A presença de municípios pequenos, tipicamente classificados com baixo risco, pode ter influenciado positivamente no valor dos índices. O segundo ponto a se destacar é que apenas quatro municípios do Rio Grande do Sul (Porto Alegre, Caxias do Sul, Campo Bom e Sapiranga) tiveram o primeiro registro de caso oficial antes do dia 19 de março, quando o governo do estado instituiu medidas obrigatórias de isolamento social, como o fechamento de escolas e estabelecimentos comerciais (Rio Grande do Sul (2020)). Isso ocasionou uma mudança de mobilidade, que não foi considerada no modelo. Nos estados de São Paulo e Rio de Janeiro, a COVID-19

se espalhou de forma mais ampla antes que medidas sanitárias fossem colocadas em prática, de forma que os dados de mobilidade podem ter sido mais condizentes com o padrão de deslocamentos até a identificação dos primeiros casos.

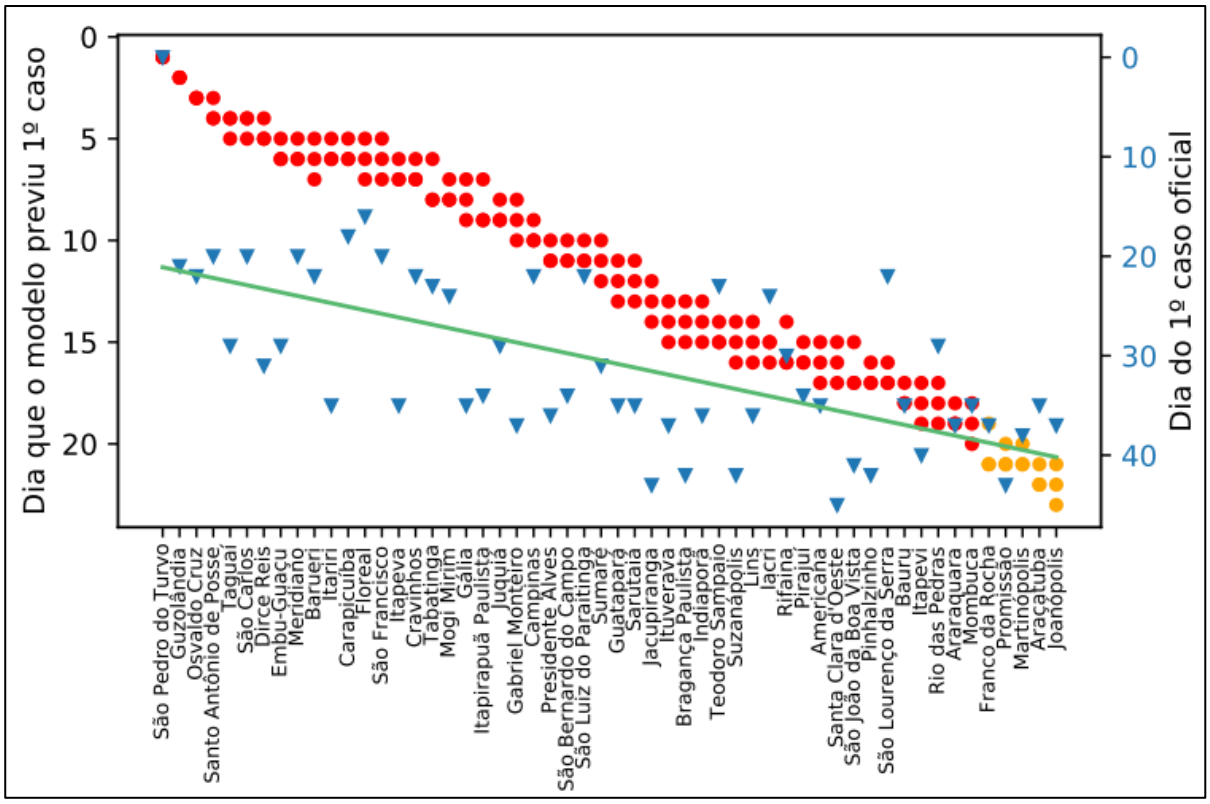
Figura 3 – Divisão dos municípios de São Paulo em três regiões: risco alto (vermelho), risco médio (laranja) e risco baixo (verde), obtida em Peixoto *et al.* (2020)



Fonte: Peixoto *et. al.* (2020)

Quanto a comparação com a reta obtida pelo método dos mínimos quadrados para São Paulo e Rio de Janeiro, relacionada aos valores v_i^l, percebemos comportamentos semelhantes ao observado para o Rio Grande do Sul. O caso de São Paulo, para os municípios com mais de 150 mil habitantes, está retratado na Figura 5 e para os municípios com mais de 50 mil habitantes do Rio de Janeiro, a Figura 6 ilustra essa comparação.

Figura 5 – Os pontos se referem aos valores de v_i^l para l tal que $s_l \in \{0.5,1,1.6\}$, isto é, o número de dias até que o modelo previsse o primeiro caso no município i para os três valores de s_l, os triângulos se referem ao dia d_i em que o município i registrou o primeiro caso oficial de COVID-19. A reta é uma aproximação dos triângulos por meio do método dos mínimos quadrados. As cores se referem ao nível de risco do município de acordo com a modelagem. Consideramos apenas os municípios com mais de 150 mil habitantes

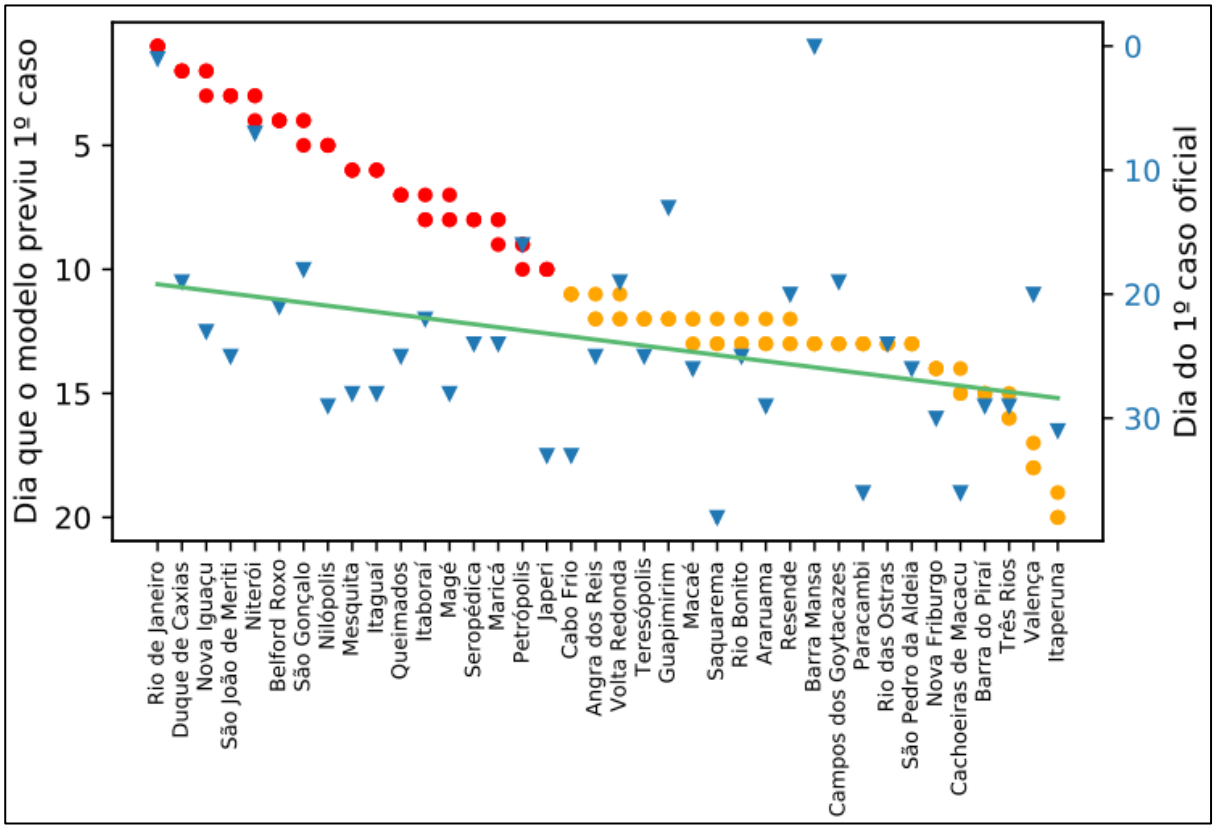


Fonte: Os autores (2021)

De um ponto de vista qualitativo, podemos perceber que na divisão do Rio Grande do Sul a região de maior risco foi a região metropolitana de Porto Alegre (região que reúne 34 municípios do estado) junto com dois municípios mais distantes: Osório e Tramandaí. Vale notar que Osório não é um município muito populoso (é apenas o 46º mais populoso do estado), mas possui um fluxo elevado de pessoas, visto que é um ponto de ligação entre o litoral gaúcho e as outras regiões do estado, além de ser um ponto de passagem da maior estrada que liga os municípios da região metropolitana com os outros estados do Brasil. Tramandaí é um município litorâneo mais distante de Porto Alegre, porém apresentou um alto fluxo de pessoas com Porto Alegre e Osório. Apesar de ser o menos povoado

dentre os municípios de alto risco, Eldorado do Sul (52º mais populoso do estado), que está no limite da região metropolitana de Porto Alegre, apresentou um alto fluxo com a região metropolitana de Porto Alegre, o que justifica o município ter ficado em uma região de alto risco. Ainda de forma qualitativa percebe-se que alguns municípios importantes do Rio Grande do Sul, por exemplo Caxias do Sul, Pelotas e Santa Maria, não ficaram em uma região de alto risco, diferentemente do que aconteceu no estado de São Paulo retratado na Figura 3, onde municípios importantes, mas distantes da capital, também ficaram em um grupo de alto risco.

Figura 6 – Os pontos se referem aos valores de v_i^l para l tal que $s_l \in \{0.5,1,1.6\}$, isto é, o número de dias até que o modelo previsse o primeiro caso no município i para os três valores de s_l, os triângulos se referem ao dia di em que o município i registrou o primeiro caso oficial de COVID-19. A reta é uma aproximação dos triângulos por meio do método dos mínimos quadrados. As cores se referem ao nível de risco do município de acordo com a modelagem. Consideramos apenas os municípios com mais de 50 mil habitantes



Fonte: Os autores (2021)

4.2 Uma abordagem espectral

Em termos gerais, o problema de clusterização em um conjunto $\mathfrak{x} = \{v_1, v_2, \dots, v_n\}$ consiste em encontrar uma partição $P = C_1 \cup C_2 \cup \dots \cup C_k$ do conjunto $\mathfrak{x}$, onde $C_i \cap C_j = \emptyset \; \forall_i \neq j \in \{1, \dots, k\}$e cada C_i é não vazio, tal que P otimiza alguma medida de qualidade, para um valor de k de clusters, que pode ser pré-definido ou escolhido a partir de algum critério.

O último passo do Algoritmo 1 é um exemplo de problema de clusterização, onde queremos agrupar *n* vetores em k = 3 *clusters*, que foi resolvido via *k-means*. Porém, é importante observar que há algumas limitações para o *k-means*, como por exemplo a dependência das condições iniciais e o embasamento em um critério puramente geométrico, sendo bem sucedido quando os *clusters* correspondem a uma região convexa. Assim é natural investigar alternativas que possam levar a clusterizações melhores.

Diversas referências no artigo de von Luxburg (2007) trazem evidências de que a clusterização espectral possui vantagens em relação ao *k-means*, em particular a clusterização espectral se revelou muito adequada para conjuntos em que os *clusters* não formam regiões convexas. Nesse artigo, utilizaremos o algoritmo espectral de Ng *et al.* (2001) para agruparmos vetores $\mathfrak{x} = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^n$ em um número pré-definido *k* de subconjuntos, que utiliza autovetores ortogonais associados a uma matriz simétrica.

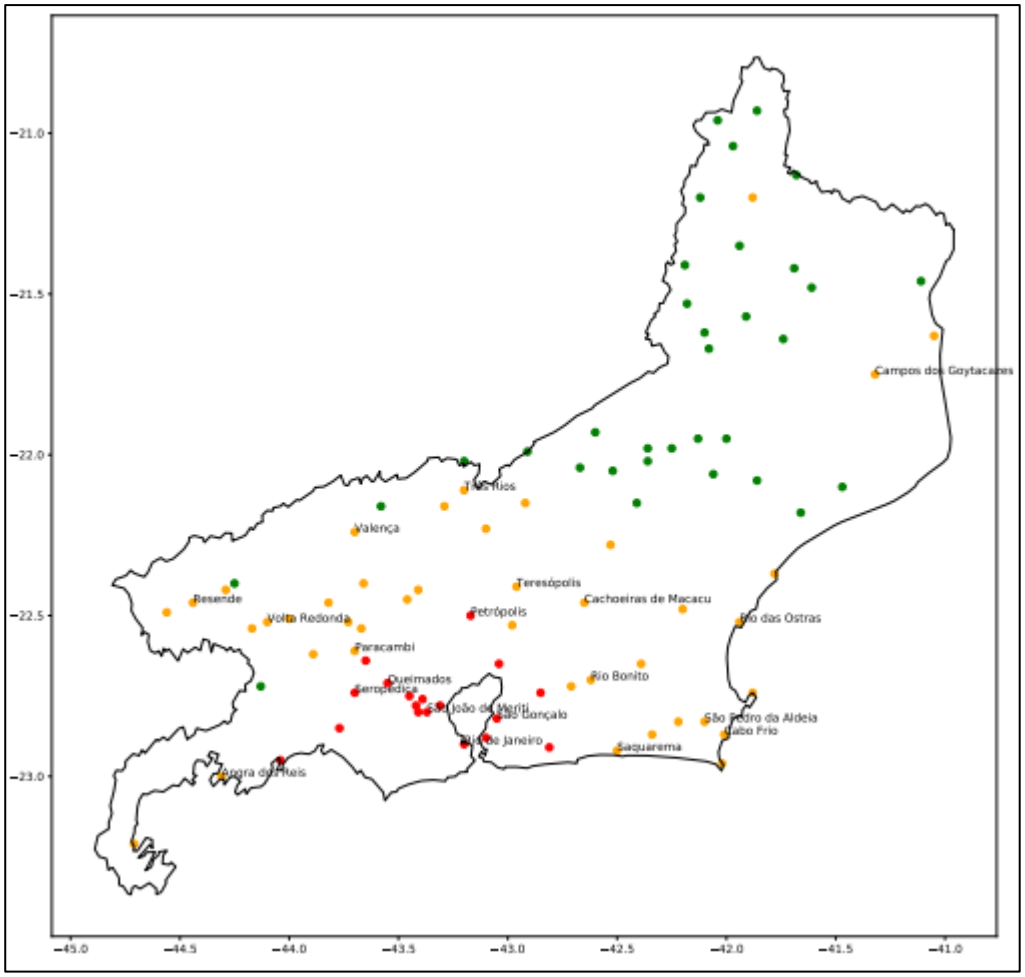
(1) Defina a matriz de afinidade $A = \left(a_{ij}\right) \in \mathbb{R}^{n \times n}$ definida por $a_{ij} = exp\left(-\| x_i - x_j \|^2 / 2\sigma^2\right) \; se \; i \neq j, e \; a_{ii} = 0$

(2) Construa a matriz simétrica $\mathcal{L} = D^{-1/2} D L^{-1/2}, onde \; L = D - A \; e \; D = \left(d_{ij}\right), para \; d_{ij} = \sum_{l=1}^{n} a_{il}, se \; i = j, e \; 0$, caso contrário.

(3) Encontre os vetores $z_1, z_2, \dots, z_k \in \mathbb{R}^n$, onde z_i é o autovetor unitário associado ao autovalor $\lambda_i, onde \; \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k$ são os *k* menores autovalores de $\mathcal{L}$. No caso de autovalores repetidos, os autovetores associados a eles devem ser

ortogonais. Construa a matriz $Z = [z_1 z_2 \dots z_k] \in \mathbb{R}^{n \times k}$ colocando esses autovetores nas colunas.

Figura 4 – Divisão dos municípios do Rio de Janeiro em três regiões: risco alto (vermelho), risco médio (laranja) e risco baixo (verde), obtida em Peixoto *et al.* (2020)



Fonte: Peixoto *et al.* (2020)

(4) Normalize cada linha de Z = (Z_{ij}) para obter uma matriz Y = (y_{ij}) com linhas unitárias (isto é $y_{ij} = z_{ij} / \left(\left(\sum_{l=1}^{k} z_{il} \right)^{1/2} \right)$.

(5) Tratando a *i*-ésima linha de Y como um ponto de $y_i \in \mathbb{R}^k$, separe {y_1,..., y_n} em *k clusters* S_1, ..., S_k via *k-means*.

(6) Seja *P* a partição tal que x_i é atribuído ao *cluster* l se, e somente se, y_i encontra-se em S_l.

Em nossa abordagem utilizamos $\mathfrak{x} = \{v_1, \dots v_n\}$, onde v_i são os vetores obtidos pelo Algoritmo 1, para $i \in \{1, \dots, n\}$ e $\sigma = 1/\sqrt{2}$. Note que o passo (5) não é determinístico, uma vez que a aplicação do *k-means* utiliza uma inicialização aleatorizada. Por isso, em aplicações desse algoritmo, tipicamente o passo (5) é realizado um número *Q* de vezes para evitar inicializações que levem a extremos locais.

Utilizando essa abordagem espectral para os nossos dados, obtivemos a classificação de risco retratada na Figura 7. Essa classificação é muito similar à classificação obtida na Figura 1, apenas cinco municípios tiveram sua classificação alterada. Os municípios de Osório e Tramandaí ficaram em risco médio na abordagem espectral, enquanto ficaram em risco alto de acordo com o Algoritmo 1, que utiliza *k-means*. Já os municípios de Guaporé, Veranópolis e Candelária, que anteriormente foram classificados com risco médio, foram classificados com risco baixo nessa nova classificação.

É importante comentar que uma dificuldade típica para a clusterização de dados é escolher o número de clusters. Isso se deve ao fato de que diversas medidas de qualidade utilizadas dependam fortemente do número de classes, de forma que não é possível incorporar esse número como uma indeterminada no problema de otimização.

Neste artigo utilizamos um critério que é comumente empregado, o *eigengap* (von Luxburg (2007)). Segundo esse critério, deve-se escolher *k* tal que os autovalores $\lambda_1, \dots, \lambda_k$ sejam pequenos e haja um salto de λ_k para λ_{k+1}. Esse salto pode ser medido pela diferença ou pela razão entre os autovalores consecutivos. Com base nesse critério, uma escolha possível é aquela que maximiza o *eigengap*. O *eigengap* máximo encontrado em nossa aplicação foi entre os autovalores λ_4 e λ_3, onde a razão entre eles foi consideravelmente maior que a razão entre quaisquer outros dois autovalores consecutivos. Isso sugere que *k* = 3 é a melhor escolha para

a quantidade de clusters que dividem o estado. O gráfico presente na Figura 8 ilustra a sequência das dez primeiras razões entre autovalores consecutivos, evidenciando que k = 3 é a melhor escolha. Os valores dos dez primeiros autovalores foram $\lambda_1 = 0$, $\lambda_2 = 0{,}0133$, $\lambda_3 = 0{,}0201$, $\lambda_4 = 0{,}0473$, $\lambda_5 = 0{,}0581$, $\lambda_6 = 0{,}0875$, $\lambda_7 = 0{,}1353$, $\lambda_8 = 0{,}1674$, $\lambda_9 = 0{,}1916$, $\lambda_{10} = 0{,}2989$.

Em algoritmos que têm um componente aleatorizado, como é o caso deste, um outro critério indicativo da qualidade da escolha de k é a estabilidade das soluções obtidas quando o algoritmo é aplicado um número Q de vezes. Utilizando $Q = 500$, todas partições originadas pelo algoritmo espectral foram idênticas. Ainda nesse aspecto, vale mencionar que na utilização do *k-means* diretamente no conjunto de vetores $\{v_1, \ldots, v_n\}$, com $Q = 500$, a clusterização de *output* foi obtido em apenas 180 vezes. Dessa forma, na nossa aplicação, a clusterização espectral se revelou mais estável que o *k-means*.

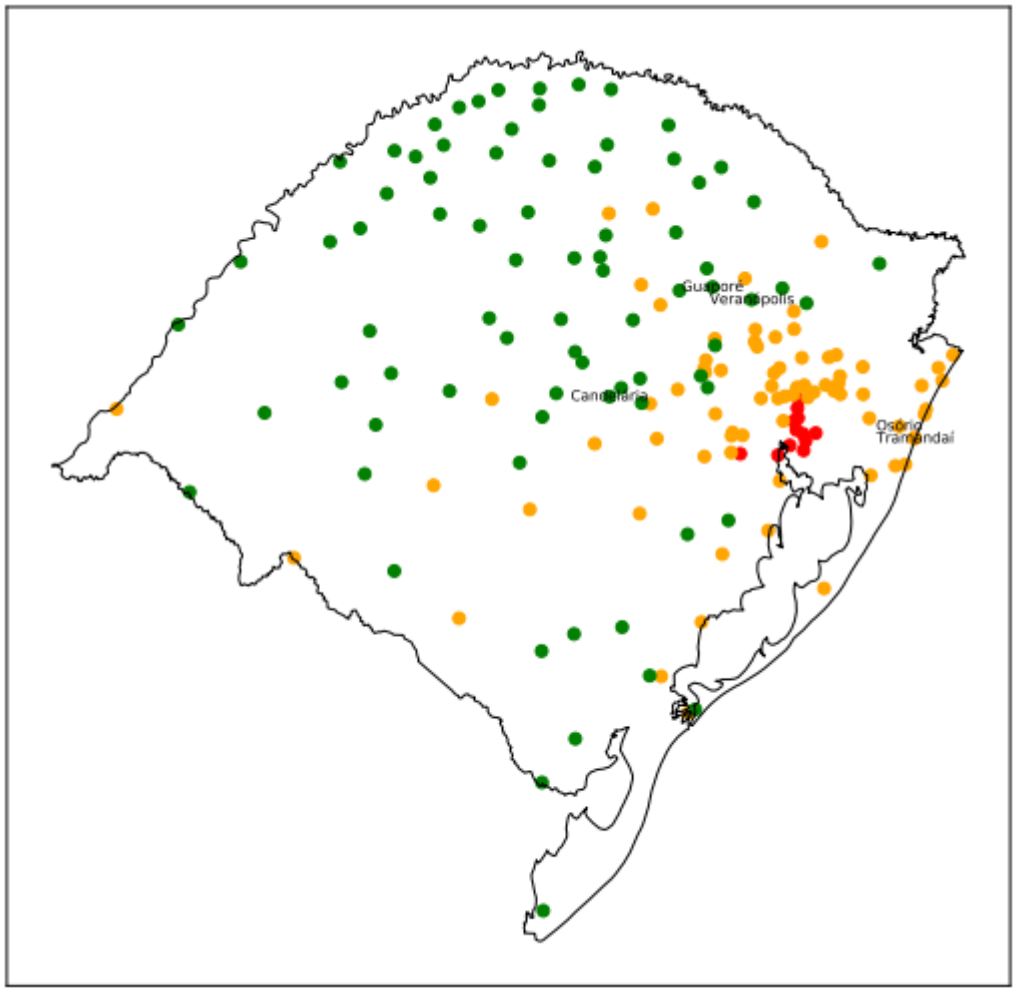
5 CONSIDERAÇÕES FINAIS

Neste trabalho, buscamos avaliar como a metodologia de classificação de risco desenvolvida por Peixoto *et al.* (2020) se aplica ao estado do Rio Grande do Sul e como ela se relaciona com a evolução da COVID-19 no estado. A metodologia se baseia em um modelo SI, que incorpora a mobilidade entre municípios, iniciando com um caso de COVID-19 em Porto Alegre.

Com poucas exceções, a região metropolitana de Porto Alegre ficou em uma zona de risco alto, a região da serra, o litoral e alguns municípios a oeste foram classificados com risco médio e os demais municípios em risco baixo. Utilizando critérios baseados em dados oficiais sobre a evolução da epidemia no estado, concluímos que a classificação de risco foi coerente com essa evolução. Acreditamos que a nossa análise teria se beneficiado de dados sobre o início da transmissão comunitária em cada município, ao invés de utilizar os primeiros casos

registrados, porém esses dados não estão disponíveis a nível municipal. Também percebemos que, ao incorporar um algoritmo espectral no último passo da metodologia, substituindo a aplicação direta do algoritmo *k-means*, houve poucas mudanças na classificação de risco obtida. Um aspecto interessante da utilização desse passo espectral foi verificar que critérios usuais para a definição do número de classes corroboraram com a divisão em k = 3 classes definida na metodologia de Peixoto *et al.* (2020).

Figura 7 – Divisão dos n = 167 municípios do Rio Grande do Sul em três regiões: risco alto (vermelho), risco médio (laranja) e risco baixo (verde), de acordo com a metodologia apresentada nessa subseção. Os municípios destacados na figura são os que houve alguma mudança frente a abordagem com utilização do *k-means* apresentada na Seção 2

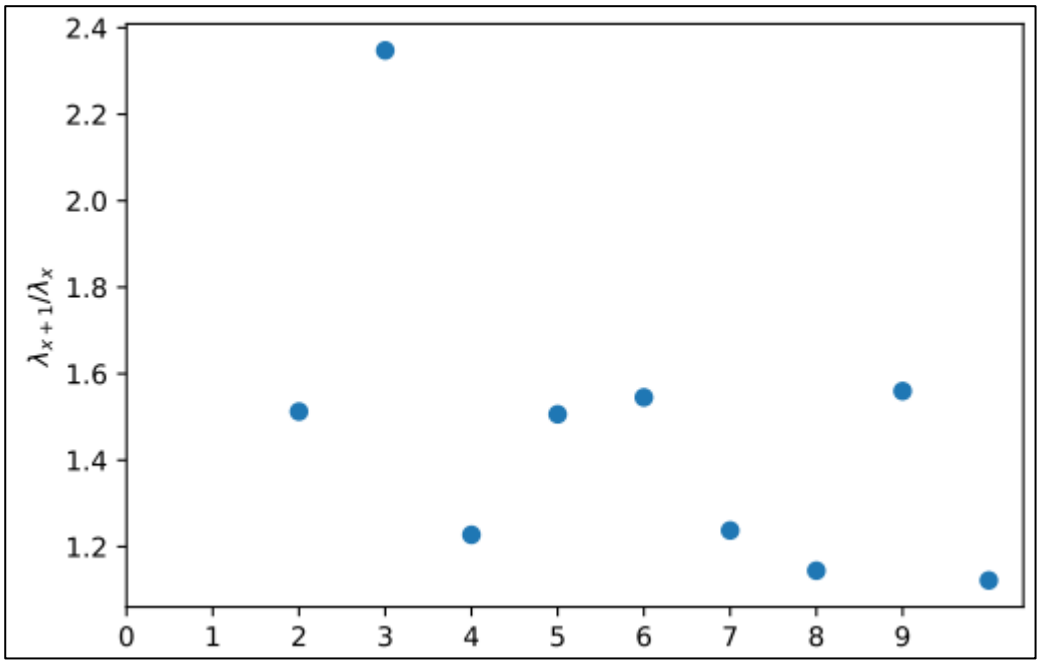


Fonte: Os autores (2021)

Em suma, o modelo utilizado é fácil de ser aplicado, afinal utiliza um único parâmetro epidemiológico da doença que está sendo considerada. Além disso, esses resultados são evidências suplementares de que a metodologia é eficiente para a classificação de risco referente à disseminação inicial da doença.

Para trabalhos futuros, será interessante aplicar a metodologia discutida nesse artigo a todo o Brasil, já que as restrições de deslocamento dentro do país foram muito menos severas do que as restrições de deslocamento entre países. Outra possibilidade é realizar uma classificação de risco baseada em um modelo mais complexo, adicionando novos compartimentos ao mesmo, como por exemplo um modelo Suscetíveis - Expostos - Infectados - Recuperados (SEIR), utilizando mobilidade populacional, para ter resultados mais precisos. Nesse caso, outros valores de k potencialmente levam a resultados melhores, e a clusterização espectral pode fazer mais sentido que o *k-means*.

Figura 8 – Gráfico da sequência dos nove primeiros gaps entre os autovalores



Fonte: Os autores (2021)

REFERÊNCIAS

BRAUER, F. (2008). Compartmental models in epidemiology. **Mathematical Epidemiology**, pp. 19–79.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. (2001). **The elements of statistical learning**, 2º edn. Springer.

LINKA, K., PEIRLINCK, M., COSTABAL, F. S., KUHL, E. (2020). Outbreak dynamics of covid-19 in europe and the effect of travel restrictions. **Computer Methods in Biomechanics and Biomedical Engineering**, pp. 1–8.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. **Statistics and computing**, 17(4), 395–416.

MINISTÉRIO DA SAÚDE (2020). **Portaria 454**, 20 de março de 2020. Diário Oficial da União.

NG, A. Y., JORDAN, M. I., WEISS, Y. (2001). On spectral clustering: Analysis and an algorithm. Proceedings of the 14th International **Conference on Neural Information Processing Systems: Natural and Synthetic**, pp. 849–856.

PEIXOTO, P. S., MARCONDES, D., CLÁUDIA, P., OLIVA, S. M. (2020). Modeling future spread of infections via mobile geolocation data and population dynamics. an application to COVID-19 in Brazil. **PLoS ONE**, 15(7), 1–23.

RIO GRANDE DO SUL (2020). **Decreto nº 55.128**, de 19 de março de 2020. Diário Oficial do Estado do Rio Grande do Sul.

SECRETARIA ESTADUAL DE SAÚDE (RS) (2020). **Painel Coronavírus RS**. Secretaria Estadual de Saúde.

SENAJITH, E. D., RENUKA, S. N., NEELIKA, M. G., JANAKA, S. H. (2020). An epidemiological model to aid decision-making for COVID-19 control in Sri Lanka. **PLoS ONE**, 15(8), 1–10.

WU, J. T., LEUNG, K., LEUNG, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study. **The Lancet**, 395, 689–697.

YOAV, T., GRANEK, R. (2021). Epidemiological model for the inhomogeneous spatial spreading of COVID-19 and other diseases. **PLoS ONE**, 16(2), 1–25.

CONTRIBUIÇÃO DE AUTORIA

1 – Lucas Siviero Sibemberg
Mestrando de Matemática Aplicada, UFRGS

https://orcid.org/0000-0003-3347-5064 - lucas.siviero@ufrgs.br
Contribuição: Curadoria de dados, Software, Validação, Visualização de dados e Escrita – primeira redação

2 – Luiz Emilio Allem:
Doutor em Matemática Aplicada, professor na UFRGS
https://orcid.org/0000-0001-9866-1541 - emilio.allem@ufrgs.br
Contribuição: Curadoria de dados, Software, Validação, Supervisão, Visualização de dados e Escrita – revisão e edição

3 – Carlos Hoppen
Doutor em Matemática Aplicada, professor na UFRGS
https://orcid.org/0000-0003-2358-3221 - choppen@ufrgs.br
Contribuição: Curadoria de dados, Software, Validação, Supervisão, Visualização de dados e Escrita – revisão e edição

4 – Pedro da Silva Peixoto
Doutor em Matemática Aplicada, professor na USP
https://orcid.org/0000-0003-2358-3221 - ppeixoto@usp.br
Contribuição: Curadoria de dados, Software, Validação, Visualização de dados e Escrita – revisão e edição

Como citar este artigo

SIBEMBERG, L.S.; ALLEM,L.E.; HOPPEN, C.; PEIXOTO, P.S. Classificação de risco em redes complexas: o caso da COVID-19 no Rio Grande do Sul. **Ciência e Natura**, Santa Maria, v. 43, Ed. Esp. X ERMAC RS, e1, p. 1-25, 2021. DOI 10.5902/2179460X66864. Disponível em: https://doi.org/10.5902/2179460X66864. Acesso em: 5 Nov. 2021.