Research, Society and Development, v. 9, n. 10, e5829109059, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i10.9059

A pandemia da COVID-19 no Brasil: uma aplicação do método de clusterização k-means

The COVID-19 pandemic in Brazil: an application of the k-means clustering method

La pandemia de COVID-19 en Brasil: una aplicación del método de agrupamiento de k-medias

Recebido: 05/10/2020 | Revisado: 07/10/2020 | Aceito: 08/10/2020 | Publicado: 09/10/2020

Henrique José de Paula Alves
ORCID: https://orcid.org/0000-0002-0124-3093
Instituto de Pesquisa Econômica Aplicada, Brasil
E-mail: jpahenrique@gmail.com
Felipe Augusto Fernandes
ORCID: https://orcid.org/0000-0003-2552-3433
Universidade Federal de Lavras, Brasil
E-mail: fernandesfelipest@gmail.com
Kelly Pereira de Lima
ORCID: https://orcid.org/0000-0002-2581-8525
Universidade Federal de Lavras, Brasil
E-mail: kelly.lima.88@gmail.com
Ben Dêivide de Oliveira Batista
ORCID: https://orcid.org/0000-0001-7019-8794
Universidade Federal de São João del-Rei, Brasil
E-mail: ben.deivide@ufsj.edu.br
Tales Jesus Fernandes
ORCID: https://orcid.org/0000-0002-1457-9653
Universidade Federal de Lavras, Brasil
E-mail: tales.jfernandes@ufla.br

Resumo

A COVID-19 é uma infecção causada pelo coronavírus SARS-CoV-2, sendo que seus primeiros registros foram na cidade chinesa de Wuhan em dezembro de 2019, e foi considerada pela Organização Mundial da Saúde (OMS) uma pandemia mundial em março de

2020. No Brasil, a COVID-19 se espalhou atingindo as 27 unidades federativas (UFs). Com isso, as tomadas de decisões para diminuir a velocidade de transmissão foram baseadas nas recomendações da OMS, onde a principal é isolamento social. Entretanto, devido a heterogeneidade da população em cada uma das UFs, a pandemia se difundiu de forma distinta. Deste modo, é interessante fazer o agrupamento das UFs por similaridade devido algumas características, e assim, observar as medidas de combate a COVID-19 realizadas em cada um desse grupos. O objetivo deste estudo foi agrupar as UFs usando análise de cluster pelo método não-hierárquico k-means considerando os coeficientes epidemiológicos como incidência, prevalência e letalidade. Os dados foram obtidos do *site* do Ministério da Saúde do Brasil e foi constituído pelas variáveis número de casos e óbitos novos e acumulados nas UFs, além da população em risco. Para análise de cluster a base de dados foi dividida em três períodos cronológicos para os três coeficientes em estudo. Com a análise de cluster foi possível verificar a estratificação da UFs conforme suas similaridades em relação a COVID-19. Assim, a estratificação da incidência, prevalência e letalidade por UFs pode se apresentar como um recurso adicional para sinalizar quais locais e quais medidas deverão ser adotadas e onde essas medidas foram eficazes.

Palavras-chave: Clusters; COVID-19; Coronavírus no Brasil; SARS-CoV-2.

Abstract

COVID-19 is an infection caused by the SARS-CoV-2 coronavirus, its first records were in the Chinese city of Wuhan in December 2019, and was considered by the World Health Organization (WHO) to be a worldwide pandemic in March 2020. In Brazil, COVID-19 spread to 27 states (UFs). As a result, decision-making to decrease the speed of transmission was based on WHO recommendations, where the main one is social isolation. However, due to the heterogeneity of the population in each of the UFs, the pandemic spread differently. Thus, it is interesting to group UFs by similarity due to some characteristics, and thus, observe the measures to combat COVID-19 carried out in each of these groups. The aim of this study was to group UFs using cluster analysis using the non-hierarchical k-means method considering the epidemiological coefficients such as incidence, prevalence, and lethality. The data were obtained from the website of the Ministry of Health of Brazil and consisted of the variables number of cases and new and accumulated deaths in UFs, in addition to the population at risk. For cluster analysis, the database was divided into three chronological periods for the three coefficients under study. With the cluster analysis, it was possible to verify the stratification of UFs according to their similarities in relation to COVID-19. Thus,

the stratification of incidence, prevalence, and lethality by UFs can present itself as an additional resource to signal which places and which measures should be adopted and where these measures were effective.

Keywords: Clusters; COVID-19; Coronavirus in Brazil; SARS-CoV-2.

Resumen

COVID-19 es una infección causada por el coronavirus SARS-CoV-2, sus primeros registros fueron en la ciudad china de Wuhan en diciembre de 2019, y fue considerada por la Organización Mundial de la Salud (OMS) como una pandemia mundial en marzo de 2020 En Brasil, COVID-19 se extendió a 27 estados (UF). Como resultado, la toma de decisiones para disminuir la velocidad de transmisión se basó en las recomendaciones de la OMS, donde la principal es el aislamiento social. Sin embargo, debido a la heterogeneidad de la población en cada una de las UF, la pandemia se propagó de manera diferente. Así, es interesante agrupar las UF por similitud debido a algunas características, y así observar las medidas de combate al COVID-19 llevadas a cabo en cada uno de estos grupos. El objetivo de este estudio fue agrupar UF mediante análisis de conglomerados mediante el método de k-medias no jerárquico considerando los coeficientes epidemiológicos como incidencia, prevalencia y letalidad. Los datos se obtuvieron del sitio web del Ministerio de Salud de Brasil y consistieron en las variables número de casos y muertes nuevas y acumuladas en UF, además de la población en riesgo. Para el análisis de conglomerados, la base de datos se dividió en tres períodos cronológicos para los tres coeficientes en estudio. Con el análisis de conglomerados se pudo verificar la estratificación de las UF según sus similitudes con respecto al COVID-19. Así, la estratificación de incidencia, prevalencia y letalidad por UF puede presentarse como un recurso adicional para señalar qué lugares y qué medidas deben adoptarse y dónde estas medidas fueron efectivas.

Palabras clave: Clústeres; COVID-19; Coronavirus en Brasil; SARS-CoV-2.

1. Introdução

O governo chinês surpreendeu o mundo ao anunciar, em dezembro de 2019, que em Wuhan (Hubei, China) descobria-se um novo coronavírus, denominado SARS-CoV-2. Com isso, a Organização Mundial de Saúde (OMS), alertada pela China, declarou que a infecção causada em humanos, denominada COVID-19, surgia como uma emergência em saúde pública mundial (Nassiri, 2020; Wang et al., 2020).

De acordo com Kodinariya & Makwana (2013), Nassiri (2020) e Velavan & Meyer (2020), esse vírus é definido com RNA de fita simples possuindo um alto potencial de mutações favorecendo a infectividade e a virulência, sendo que, o principal mecanismo de entrada desse vírus nas células dos hospedeiros se dá por meio das células epiteliais do trato respiratório superior (Khailany et al. 2020; Letko et al. 2020). Assim, esse tipo de infecção pode levar a sinais e sintomas como febre, tosse seca, cefaleia, fadiga, dispneia, falta de ar, calafrios e artralgia, sendo que esses sinais e sintomas surgem em média entre 5 e 6 dias após a incubação (Salathé et al 2020).

Os danos causados a saúde pública ficam evidentes à medida que inexistem vacinas e medicações específicas para o tratamento (Khailany et al., 2020; Nassiri, 2020; Salathé et al., 2020; Velavan & Meyer, 2020). Outro fato importante é que muitos pacientes em estado grave precisam de assistência hospitalar e respiradores mecânicos. O SARS-CoV-2 tem alto potencial de transmissão, o que leva a necessidade de distanciamento social e, por muitas vezes, até mesmo isolamento social total, conhecido como “lock down” (Salathé et al., 2020). Tais características fizeram com que a COVID-19 fosse rapidamente considerada uma pandemia pela OMS.

No Brasil essa pandemia se alastrou rapidamente atingindo todas as 27 unidades da federação (UF). Esse fato ocorreu devido aos desafios quanto as condições de vulnerabilidade social, de moradia e saneamentos precários, além de superpopulação domiciliar. Sendo assim, seguir as recomendações da OMS de isolamento social parcial, e até mesmo total em algumas situações e de higienização, que se mostrou tão eficaz em outros países, se torna impossível para milhões de brasileiros (Werneck & Carvalho, 2020).

Em 15 de setembro de 2020 o Brasil apresentava 4.497.434 casos, totalizando 135.857 mortes e 3.789.139 curados de contaminação da COVID-19, com isso, o país se tornou o terceiro em número de casos e mortes segundo um rankeamento feito pela OMS com relação aos demais países. O Brasil ainda possui 572.438 casos suspeitos ou sem confirmação, o que mostra a importância da realização de testes rápidos e eficientes da COVID-19, outra carência no Brasil. Diante do exposto, fica evidente que a situação no Brasil é preocupante de modo que ações de combate ao rápido contágio dessa pandemia devam ser tomadas imediatamente.

Medidas isoladas de combate a COVID-19 podem e devem ser implementadas pelos órgãos responsáveis em cada uma das UF’s brasileiras. Entretanto, devido a heterogeneidade populacional de cada uma das UF’s brasileiras, número de leitos de unidades de terapia intensiva (UTI) disponíveis para atendimento emergencial além de

condições de vulnerabilidade sócio econômica, é interessante agrupar essas UF's por similaridade devido a algumas características e então, uma vez observado tais agrupamentos, direcionar medidas de combate a COVID-19 plausíveis a cada um desses grupos.

Na literatura, existem técnicas de agrupamentos que consideram múltiplas variáveis como condicionantes, conhecida como análise agrupamentos para dados multivariados (Fávero & Belfiore 2019; Ferreira 2018), cujo objetivo é dividir as observações em grupos que são homogêneos internamente e heterogêneos externamente. Nessa linha de pensamento, James & Menzies (2020) propuseram um método baseado em análise de cluster para analisar a evolução de séries temporais multivariadas e aplicá-lo a COVID-19, onde os autores justificam o uso dessa técnica afirmando que a cada dia cientistas e analistas dividem os países, estados e até mesmo municípios em grupos de acordo com seus casos e contagens de mortes. Como consequência, mudanças em ambas as quantidades ao longo do tempo são esperadas e esse tipo de análise pode ajudar a destacar as políticas públicas mais e menos significativas para minimizar a taxa de mortalidade da COVID-19 de um país.

Iritani et al. (2020) também aplicaram essa técnica em 47 prefeituras do Japão entre os dias 16 de janeiro e 9 de maio, onde essas prefeituras foram agrupadas de acordos com o tempo de ocupação de hospitais/instalações por pacientes que necessitavam de um tempo longo de internação e também instalações não médicas e médicas de bem estar e mal estar com morbidade e mortalidade. Zarikas et al. (2020) também investigaram estratégias de agrupamentos na Índia.

Além disso, Maciel et al. (2020) destacam a importância de analisar fatores associados a óbitos em indivíduos de ambos os sexos e diferentes faixas etárias no estado do Espírito Santo, no Brasil, em hospitais públicos e privados. Guimarães et al (2020), também estudaram a estratificação do risco para a predição de disseminação e gravidade COVID-19 no Brasil utilizando o método k-means com três clusters.

Diante do exposto, o objetivo do trabalho é agrupar as UF's brasileiras utilizando o método não-hierárquico k-means, considerando como variáveis os coeficientes de incidência, prevalência e letalidade, que são importantes medidas epidemiológicas de saúde pública, permitindo assim visualizar a evolução temporal da COVID-19 de cada UF brasileira através desses coeficientes. Justifica-se a importância desses coeficientes pois eles medem, respectivamente, o risco de ocorrência de um indivíduo ficar doente (novos casos), a probabilidade da população estar doente no período estudado (casos acumulados) e

também a severidade da doença (mortes acumuladas) e optou-se por dividir o estudo em três períodos: o inicial de contágio da doença, o período intermediário de contágio da doença e o período de transição que é o estágio que os números de casos e óbitos começaram a declinar.

2. Material e Métodos

Os dados em estudo são disponibilizados pelo Ministério da Saúde do Brasil (Brasil 2020), em que, a base de dados obtida contém a informação das variáveis população, o número de casos novos e acumulados, e também o número de novas mortes e acumuladas de COVID-19 para cada uma das 27 unidades da federação (UF's) brasileiras. Com isso, as informações contidas foram separadas em três períodos seguindo uma ordem cronológica de contágio da doença como fase inicial (04-02-2020 a 01-04-2020), fase intermediária (02-04-2020 a 01-07-2020) e fase de transição da doença (02-07-2020 a 15-09-2020).

A justificativa para a escolha desses três períodos se deu pelos seguintes motivos: primeiro, no contágio inicial da doença espera-se um aumento de novos casos e óbitos acumulados em todas as UF's brasileiras. No segundo tem-se uma época de frio, o que pode gravar a pandemia em quase todas as UF's brasileiras, neste momento especulou-se ser esperado ocorrer o pico de infecção da COVID-19 no Brasil. No terceiro período esperava-se que a maioria das pessoas já produzissem anticorpos ocorrendo então um período de estabilização e queda da infecção por COVID-19 (Fernandes et al., 2020; Werneck & Carvalho, 2020).

Posteriormente, calculou-se os coeficientes de incidência, prevalência e letalidade das UF's brasileiras pertencentes a cada um dos grupos em cada período considerado e, finalmente, utilizou-se esses coeficientes calculados como variáveis para obter os clusters das UF's brasileiras em cada período estudado, o que possibilita entender como ocorreu o avanço da COVID-19 em cada cluster encontrado. A metodologia aqui utilizada é de natureza qualitativa, como é apresentado em Pereira et al (2018).

2.1 A análise de clusters não-hierárquica

A análise de clusters não-hierárquico tem como objetivo classificar observações de um "dataset" de forma que suas semelhanças sejam alocadas um mesmo grupo, portanto, aquelas que pertencem a diferentes grupos são consideradas dissimilares (Everitt et al., 2011;

Ferreira, 2018; Fávero & Belfiore, 2019). A semelhança entre as observações é quantificada por meio de uma métrica de proximidade que pode ser, por exemplo, a distância quadrática euclidiana entre observações x_i e o centroide $\bar{x}$ do cluster.

2.1.1 A distância Euclidiana quadrática

Trata-se de uma distância ou métrica que tem como intuito medir a distância entre dois pontos localizados em um espaço dimensional utilizando o teorema de Pitágoras. Segundo Fávero & Belfiore (2019), sua expressão é dada por:

$$d^2(x,y) = \sum_{i=1}^{p} (x_i - \bar{x})^2,$$

em que p representa o número de observações dentro de um cluster k, x_i representa a i-ésima observação dentro de um cluster com i=1, 2, ..., p e $\bar{x}$ representa o centroide do cluster k. Ainda de acordo com os autores supracitados, essa distância é comumente utilizada quando a variável em estudo apresenta pouca dispersão.

2.1.2 O método k-means

Esse método tem algumas condições como a informação “a priori” do número de clusters k e as observações são agrupadas nesses k clusters utilizando uma função objetivo a critério. Apresentamos uma sequência lógica de passos baseado em Johnson & Wichern (2002) e Fávero & Belfiore (2019).

1. Defina inicialmente o número de clusters bem como seus respectivos centroides $\bar{x}_j$, j = 1, 2, ..., k. Dessa forma, o principal objetivo é dividir as observações x_i, i = 1, 2, ..., n do conjunto de dados em k clusters, k = 1, 2, ..., onde as observações x_i dentro de cada cluster k são mais próximas umas das outras quando comparadas a qualquer outra observação x_i que pertença a um outro cluster k. As observações x_i serão alocadas arbitrariamente nesses k clusters para que seus respectivos centroides $\bar{x}_j$ possam ser calculados;
2. O próximo passo é verificar se uma determinada observação x_i está mais perto de um outro centroide $\bar{x}_j$, j = 1, 2, ..., k, utilizando a distância Euclidiana quadrática entre

pontos como em 2.1.1 e, caso positivo, realocá-la nesse outro cluster k. Neste momento, algum cluster k acaba de perder essa observação x_i para algum outro cluster k e, portanto, os centroides $\bar{x}_j$ do cluster k que a recebe e do outro cluster k que a perde devem ser recalculados;

3. O passo 2 deve ser repetido até que não exista mais possibilidade de realocação de alguma observação x_i em algum outro cluster.

Cada centroide $\bar{x}_j$ deve ser recalculado sempre que uma observação x_i for incluída ou excluída em um novo centroide $\bar{x}_j$ no respectivo cluster k, baseado nas seguintes expressões:

$\bar{x}_{new} = \frac{N\bar{x}_j + x_i}{N+1}$, se a observação x_i é incluída no cluster k;

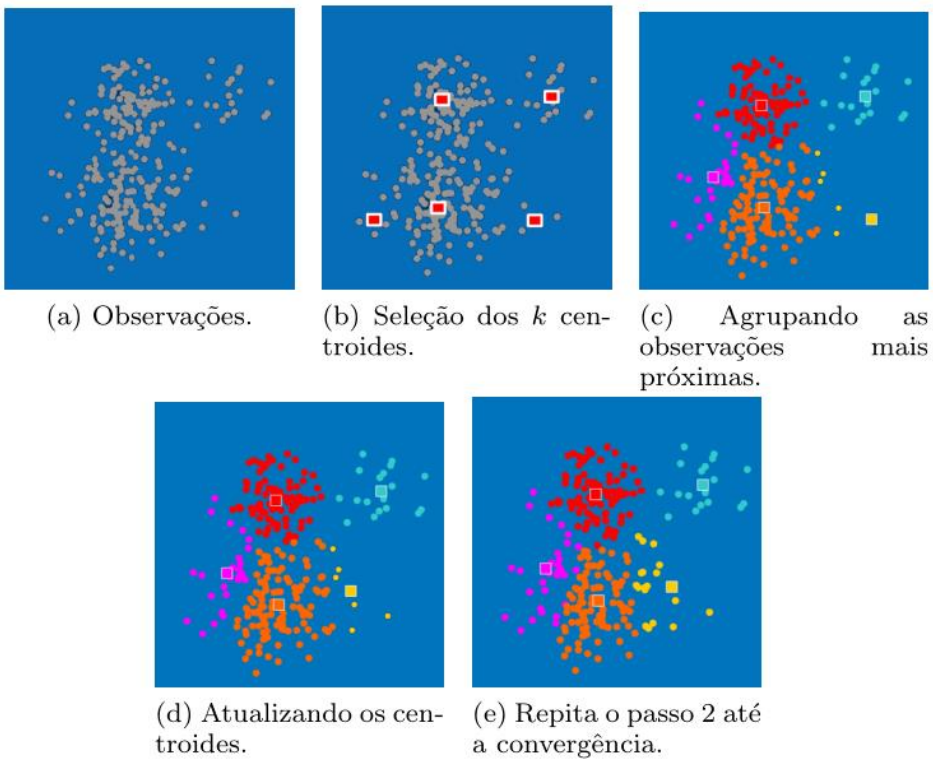
ou

$\bar{x}_{new} = \frac{N\bar{x}_j - x_i}{N-1}$, se a observação x_i é excluída do cluster k;

em que N se refere ao número de observações no cluster, $\bar{x}_j$ se refere ao centroide de realocação dessa observação e x_i se refere i-ésima observação a ser incluída em um novo cluster k. A escolha "a priori" do número de clusters k é necessário para evitar grande esforço computacional, pois ao considerarmos todos os possíveis agrupamentos podemos obter um número muito grande de possibilidades o que demanda muito esforço computacional.

A Figura 1 representa o algoritmo do método *k*-means, passos lógicos, sendo que, os pontos representam as observações e os quadrados representam os centroides dos clusters.

Figura 1. Em (a) e (b) tem-se o passo 1, em (c) e (d) o passo 2, e o passo 3 é ilustrado em (e).



(a) Observações. (b) Seleção dos k centroides. (c) Agrupando as observações mais próximas. (d) Atualizando os centroides. (e) Repita o passo 2 até a convergência.

Fonte: Dos próprios autores.

Uma vez realizado esse procedimento, é importante definir se os clusters encontrados fazem sentido, isto é, se a variabilidade dentro dos clusters é realmente menor que a variabilidade entre os clusters. Esse fato é equivalente a testar as seguintes hipóteses:

$$\begin{cases} H_0: \text{a variável tem a mesma média em todos os grupos formados} \\ H_1: \text{a variável tem médias diferentes em todos os grupos formados.} \end{cases}$$

Alguns autores mencionam uma estatística de teste que sob a hipótese nula H_0 segue uma distribuição F com k-1 graus de liberdade no numerador e n-k graus de liberdade no denominador, sendo n o número de observações. Esse método seria usado na análise de variância e, portanto, exige normalidade multivariada dos dados, o que nem sempre é garantido (Kodinariya & Makwana, 2013; Fávero & Belfiore, 2019). Aqui, utilizou-se a função “NbClust'” presente no pacote “NbClust” no software estatístico R (R Core Team, 2020), como apresentado em Charrad et al. (2015), que determina, entre os vários critérios

existentes, o número de clusters k ótimo. Utilizou-se as medidas de avaliação de saúde pública conhecidas como coeficientes de incidência, prevalência e letalidade.

2.2 Medidas epidemiológicas diagnósticas

2.2.1 O Coeficiente de Incidência

O coeficiente de incidência CI expressa o risco de ocorrência de novos casos de uma doença em uma população durante um período de tempo. Sua expressão é dada por:

$$CI = \frac{|NC|}{|POP|} \times 100,$$

em que $|NC|$ expressa o número de novos casos e $|POP|$ a população em risco.

2.2.2 O Coeficiente de Prevalência

O coeficiente de prevalência CP expressa a probabilidade da população estar doente durante um período de tempo. Sua expressão é dada por:

$$CP = \frac{|CC|}{|POP|} \times 100,$$

em que $|CC|$ expressa o número de casos acumulados e $|POP|$ a população em risco.

2.2.3 O Coeficiente de Letalidade

O coeficiente de incidência CL mede a severidade de uma doença em uma população durante um período de tempo. Sua expressão é dada por:

$$CI = \frac{|DC|}{|POP|} \times 100,$$

em que $|DC|$ expressa o número de mortes e $|POP|$ a população em risco.

3. Resultados e Discussão

3.1 Determinando o número de clusters “a priori”

Com relação ao número de clusters ou grupos a ser considerado em cada período, adotou-se o critério proposto por Ratkowsky & Lance (1978), de acordo com o pacote “NbClust” presente no R (Charrad et al., 2015; R Core Team, 2020). Então, pelos critérios da função “NbClust” disponível no pacote supracitado, tem-se que k=3 como sendo o número de clusters ótimo adotado “a priori”.

Posterior a adoção do número de clusters, tem-se a distribuição espacial das médias das medidas epidemiológicas de cada um desses clusters para cada período. Uma vez determinado o número de clusters nos períodos adotados é apresentado nas Figuras 2, 3 e 4, a distribuição espacial das médias das medidas epidemiológicas diagnósticas (seção 2.2) de cada um desses clusters em cada estágio de contaminação.

As UF’s brasileiras são representadas nos mapas obtidos na seção 3.1 pelo código de classificação segundo o Instituto Brasileiro de Geografia e Estatística (IBGE): 11 – Rondônia (RO), 12 – Acre (AC), 13 – Amazonas (AM), 14 – Roraima (RR), 15 – Pará (PA), 16 – Amapá (AP), 17 – Tocantins (TO), 21 – Maranhão (MA), 22 – Piauí (PI), 23 – Ceará (CE), 24 – Rio Grande do Norte (RN), 25 – Paraíba (PB), 26 – Pernambuco (PE), 27 – Alagoas (AL), 28 – Sergipe (SE), 29 – Bahia (BA), 31 – Minas Gerais (MG), 32 – Espírito Santo (ES), 33 – Rio de Janeiro (RJ), 35 – São Paulo (SP), 41 – Paraná (PR), 42 – Santa Catarina (SC), 43 – Rio Grande do Sul (RS), 50 – Mato Grosso do Sul (MS), 51 – Mato Grosso (MT), 52 – Goiás (GO), 53 – Distrito Federal (DF).

3.2 Descrição dos clusters em cada período considerado

Uma vez que as UF’s estão agrupadas pelo método k-means em cada um dos três períodos supracitados, optou-se em calcular as médias dos seus coeficientes e usou-se essas medidas para representar resumidamente a situação do surgimento de novos casos, dos casos acumulados e também dos óbitos em cada um desses grupos.

A justificativa para a escolha da média como medida, ocorre uma vez que a mesma é uma boa medida para representar os dados quando esses não apresentam outliers, o que é esperado quando se utiliza técnicas de agrupamentos. Cabe salientar que os resultados apresentados aqui não tornam os clusters obtidos mais ou menos importantes uns em relação a

outros. O que deseja-se aqui é verificar onde as medidas de combate à COVID-19 foram mais eficazes, em média. É esperado que nas UF's formadoras desses clusters essas medidas foram mais eficazes para os clusters com menores riscos.

A Figura 2 apresenta as médias dos coeficientes de incidência, prevalência e letalidade no período inicial de contágio da COVID-19 no Brasil, onde essas médias foram avaliadas para cada 100 habitantes. Destaca-se o cluster 2, formado apenas pelo Distrito Federal (53) como sendo o de maior risco para o surgimento de novos casos (incidência) (Figura 2 (a)), bem como maior probabilidade da população estar doente e maior severidade da COVID-19 (prevalência) (Figura 2 (b)) e óbitos (Figura 2 (c)). O cluster 1, formado por São Paulo (35), Santa Catarina (42), Rio de Janeiro (33), Espírito Santo (32), além do Acre (12), Amazonas (13), Roraima (14) e Ceará (23) vem logo em seguida em risco de novos casos (Figura 2 (a)), probabilidade da população estar doente e severidade da doença (Figura 2 (b)) e óbitos (Figura 2 (c)). O cluster 3, formado pelas demais UF's apresenta os menores riscos, as menores probabilidades e severidades.

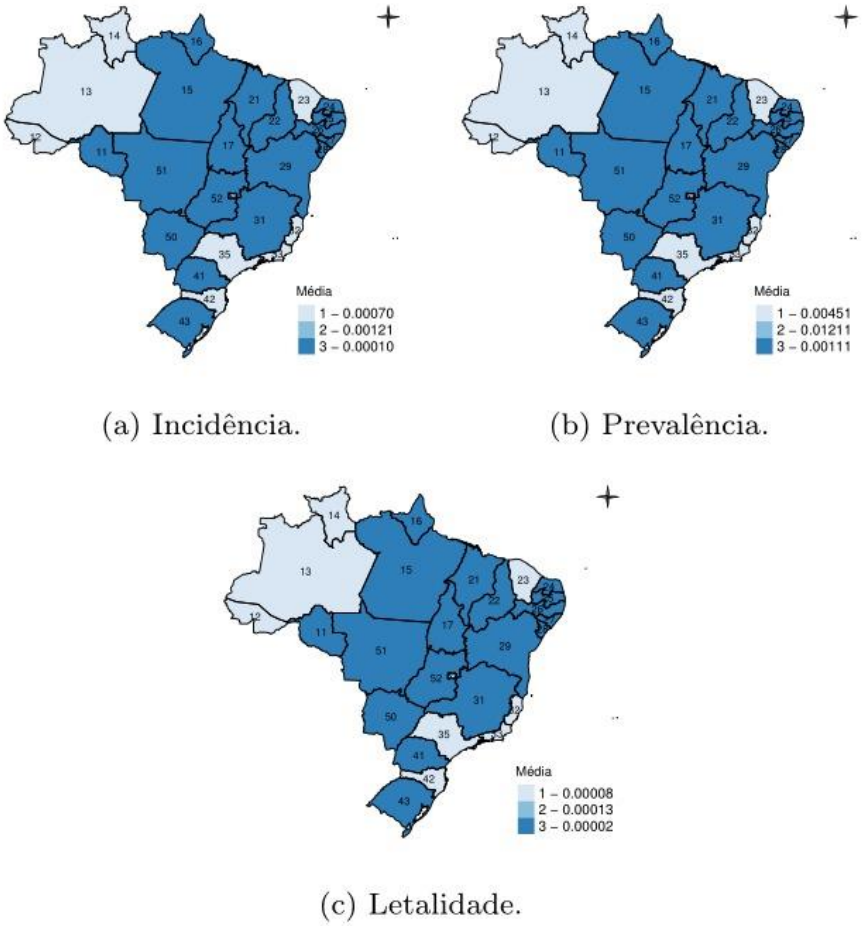
Pela Figura 2 pode-se observar que os resultados evidenciaram, assim como no trabalho de Stier et al. (2020), que a evolução inicial da pandemia de COVID-19 ocorreu de forma distinta nos estados brasileiros. Isso deve-se ao fato de que, no caso das grandes metrópoles e centros urbanos, a disseminação do vírus é potencializada por diversos fatores entre eles cita-se o uso de transporte público, a concentração de pessoas e alta densidade populacional nas periferias público.

De acordo com Fundação Oswaldo Cruz (2020a), observou-se que os primeiros casos aconteceram nas duas grandes metrópoles brasileira: São Paulo e Rio de Janeiro, ambas localiza-se os principais aeroportos do Brasil, além disso, está vinculada as maiores economia do país. Em seguida é contabilizado em Brasília e em outras metrópoles ao redor do país, com grande vinculação aérea às principais cidades e conectadas também a outros países. Desta forma, o vírus se espalhou por todas as regiões do país, onde notou-se que na região sul um maior espalhamento da pandemia em direção ao interior (cidades de médio e pequeno porte), já na região Nordeste, o maior número de casos concentrou-se no litoral e capitais.

De acordo com Souza et al. (2020), no dia 28 de março 2020 que corresponde a décima terceira semana epidemiológica, dez estados já somavam 110 óbitos: Santa Catariana, Rio Grande do Sul, Paraná, Rio de Janeiro, São Paulo, Goiás, Amazonas, Ceará, Pernambuco e Piauí. As maiores taxas foram observadas nos estados do Piauí, Pernambuco e São Paulo. A maior letalidade foi observada no Piauí, Rondônia e Alagoas. Já Fernandes et al. (2020),

observaram que os estados mais afetados na fase inicial do COVID-19 foram AM, PA, CE, PE, SP e RJ, no entanto, foi observado casos e óbitos em números absolutos.

Figura 2. Médias dos coeficientes de incidência, prevalência e letalidade na fase inicial de contaminação da doença por cada 100 habitantes (02-04-2020 a 04-01-2020) para cada cluster. O número presente em cada UF nos mapas são referentes a classificação de acordo com o IBGE.



(a) Incidência. (b) Prevalência.

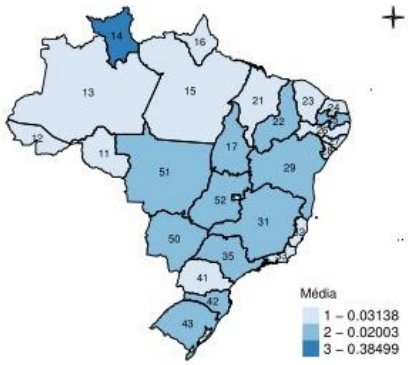
(c) Letalidade.

Fonte: Dos próprios autores.

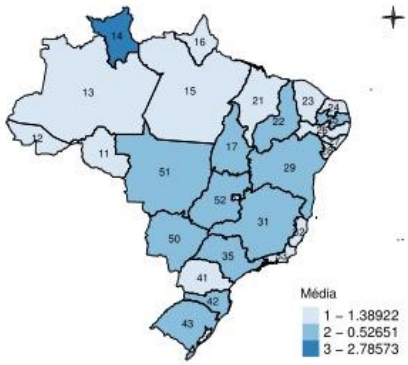
Com relação a taxa de casos por milhões de habitantes até dia 16 de abril, pode-se destacar três estados da região Sudeste (SP, RJ e ES) e uma unidade do Centro-Oeste (DF), quatro do Norte (AM, AP, RR e AC) e dois do Nordeste (CE e PE). Essa lista tem a inclusão dos locais com maior Produto Interno Bruto (PIB) entre as unidades federativas destaca-se SP, RJ e DF. Já com relação ao índice de incidência os maiores valores foram por Amapá, Amazonas, Ceará, São Paulo, Roraima, Rio de Janeiro e Distrito federal (Fundação Oswaldo Cruz, 2020b).

No período considerado com "pico" da doença no Brasil, ou seja, uma ascensão do número de casos novos e acumulados e número de óbitos novo e acumulados, observa-se uma mudança nos clusters levando em consideração o período inicial (Figura 3), agora nota-se o cluster 3 (Figura 3 (c)) é formado somente pelo estado de Roraima. Nesse caso, o estado supracitado apresenta um maior risco para o surgimento de novos óbitos. O cluster 2 (Figura 3 (b)) é dado pelos estados de menor risco para surgimento de novos casos (Figura 3 (a)), porém é o terceiro maior risco para os casos acumulados e óbitos (Figura 3 (b) e (c)). Já o cluster 1 (Figura 3 (a)) formado pelos estados de segundo maior risco para o surgimento de novos caso e casos acumulados, e risco elevado para óbitos (Figura 3 (c)).

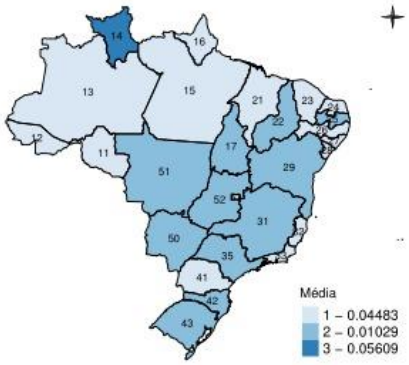
Figura 3. Médias dos coeficientes de incidência, prevalência e letalidade na fase inicial de contaminação da doença por cada 100 habitantes (04-02-2020 a 01-07-2020) para cada cluster. O número presente em cada UF nos mapas são referentes a classificação de acordo com o IBGE.



(a) Incidência.



(b) Prevalência.



(c) Letalidade.

Fonte: Dos próprios autores.

Na Figura 3, destaca-se que se trata de um período de muito frio em algumas UF's situadas principalmente no sudeste e sul brasileiros, o que pode ser um indicativo para o aumento desse risco de novo casos nesse cluster. Com relação ao cluster 1 é formado, em grande maioria, pelas UF's de menor condição socioeconômicas do Brasil, com exceção do Paraná (41), Rio de Janeiro (33) e Espírito Santo (32), o que pode indicar que neste cluster as medidas de isolamento social parcial ou total e de higienização não foram eficazes e/ou podem não ter sido respeitadas, destacando-se a cidade do Rio de Janeiro (33) pela presença de favelas e comunidades carentes.

Conforme Guimarães et al. (2020), de todas as 27 unidades federativas com o maior potencial para disseminação e gravidade por COVID até dia 8 de maio, somente cinco adotaram o chamado "lockdown" como uma estratégia de controle da expansão dos casos e do consegue colapso do sistema de saúde. Entre elas cita-se Maranhão, Pará, Ceará (Fortaleza), Bahia (Salvador) e Rio de Janeiro (Niterói). Contudo, as outras unidades, como Rio de Janeiro, São Paulo, Distrito Federal, Minas Gerais, Bahia, Pernambuco e Rio Grande do Norte apresentaram um elevado potencial de disseminação da doença e adotaram medidas restritiva considerando diversos aspectos e estágio da pandemia em cada uma se encontrava.

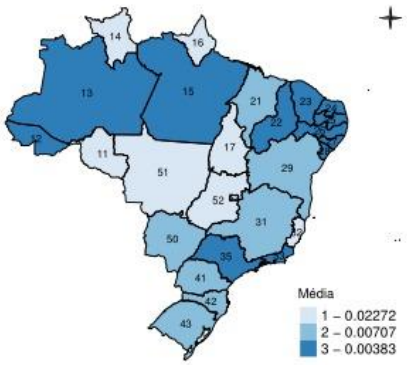
Notou-se que até abril houve crescimento acelerado nos estados Norte-Nordeste onde existe uma lacuna de desenvolvimento por exemplo o Índice de Desenvolvimento Humano (IDH), esse índice reflete um passivo social existente e pode afetar diretamente essas altas taxas de indecência, prevalência e letalidade. (Fundação Oswaldo Cruz, 2020d).

Nos meses de abril e maio de 2020, houve um aumento significativo do número total de óbitos em diversas áreas do Brasil, notadamente nas Regiões Metropolitanas de São Paulo e Campinas, São Luiz do Maranhão, Belém do Pará, Fortaleza no Ceará e Manaus no Amazonas. Assim, apontando com estas cidades como epicentros da transmissão. Quando se observa o período até 19 de maio, dos dez estados com maiores taxas de incidência e mortalidade, nove pertencem às regiões Norte (AM, AP, RR, PA e AC) e Nordeste (CE, PE, MA e SE) e o único do Sudeste passa a ser o Espírito Santo. As UFs que mostram maior potencial de gravidade para a Covid-19 são Pernambuco, Ceará, Maranhão, Rio de Janeiro, São Paulo, Pará e Amazonas. Essas são exatamente as que se destacam, atualmente, com maior número absoluto de óbitos, na seguinte ordem: São Paulo, Rio de Janeiro, Ceará, Pará, Pernambuco, Amazonas e Maranhão (Fundação Oswaldo Cruz, 2020c).

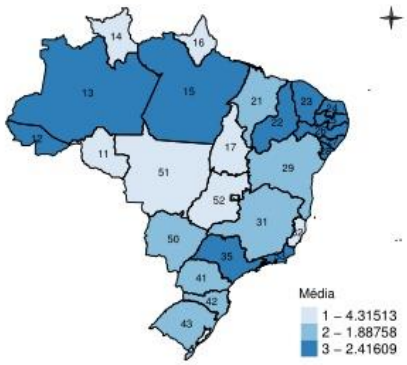
O período considerado "estabilidade", ou seja, a queda infecção da COVID-19 nota-se uma heterogeneidade com relação aos estados, além disso pela Figura 4 pode-se observar que as UF's situadas no centro-oeste e norte brasileiros (Figura 4) apresentaram os maiores riscos

do aparecimento de novos casos, casos acumulados e óbitos (Figura 4 (a), (b) e (c)) onde destacam-se nesse cluster a presença do Distrito Federal (53) e do Espírito Santo (32). O cluster 1 representa os maiores riscos nesse período (Figura 4(a), (b) e (c)), com destaque para Roraima (14) que também apresentou os maiores coeficientes no período de "pico" da COVID-19 (Figura 3).

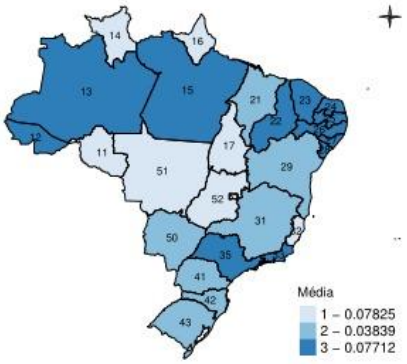
Figura 4. Médias dos coeficientes de incidência, prevalência e letalidade na fase inicial de contaminação da doença por cada 100 habitantes (02-07-2020 a 15-09-2020) para cada cluster. O número presente em cada UF nos mapas são referentes a classificação de acordo com o IBGE.



(a) Incidência.



(b) Prevalência.



(c) Letalidade.

Fonte: Dos próprios autores.

Já o cluster 2 (Figura 4) apresentaram os menores riscos de casos acumulados e óbitos, destaca-se a presença de Minas Gerais (31), Paraná (41), Santa Catarina (42), Rio Grande do Sul (43) e Bahia (29). Com relação ao cluster 3, com a presença de São Paulo (35), Rio de

Janeiro (33), Ceará (23) apresentaram os menores riscos de novos casos, porém apresentam segundo maior risco para casos acumulados e óbitos.

A resposta à pandemia da COVID-19 pode ser dividida em quatros fases: contenção, mitigação, supressão e recuperação (Werneck & Carvalho, 2020). Estas fases se sobrepõem em regiões de profunda heterogeneidade no território, como por exemplo no Brasil, onde cada unidades federativas tem sua característica específica. Assim, pode-se acontecer que algumas áreas ainda se encontram em fase de contenção da doença, outras, já mostram uma flexibilização do isolamento social adotado e reestruturando a sociedade e a economia. Como pode ser visto, a evolução de alguns estados na fase de estabilização.

4. Conclusão

A técnica de agrupamentos aqui aplicada se mostrou adequada e pode ser utilizada para indicar as localidades de alguma região onde as medidas de combate a COVID-19 estão sendo eficazes e como ela se deu no espaço durante o passar do tempo. Na literatura existem outras técnicas que possibilitam a mesma conclusão. Entretanto, essa técnica de agrupamentos é muito utilizada em análises estatísticas. Com isso, a estratificação de risco pode se apresentar como um recurso adicional para sinalizar os locais que medidas poderão ser adotadas, bem como para planejamento de pandemias futuras.

Agradecimentos

Felipe Augusto Fernandes agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a bolsa de doutorado.

Referências

Brasil (2020). Ministério da Saúde. *COVID-19 no Brasil*. Recuperado de https://susanalitico .saude.gov.br/extensions/covid-19_html/covid-19_html.html

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2015). *Determining the best number of clusters in a data set*. Package 'NbClust'. Recuperado de http://cran.rediris.es/ web/packages/NbClust/NbClust.pdf

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis,* (5th ed.), John Wiley.

Fávero, L. P., & Belfiore, P. (2019). *Data Science for Business and Decision Making*. Academic Press, Cambridge, MA, USA.

Fernandes, F. A., Alves, H. J. P., Fernandes T. J., & Muniz. J. A. (2020). Panorama da fase inicial do crescimento dos números de casos e óbitos causados pela COVID-19 no Brasil. *Research, Society and Development*, 9(10), 1-19. DOI: http://dx.doi.org/10.33448/rsd-v9i10.8560

Ferreira, D. F. (2018). *Estatística Multivariada*, (3a ed.), 624. Editora UFLA, Universidade Federal de Lavras.

Fundação Oswaldo Cruz. (2020a). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. *Monitora COVID-19*. Nota Técnica 1. Recuperado de https://bigdata-covid19.icict.fiocruz.br

Fundação Oswaldo Cruz. (2020b). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. *Monitora COVID-19*. Nota Técnica 2. Recuperado de https://bigdata-covid19.icict.fiocruz.br

Fundação Oswaldo Cruz. (2020c). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. *Monitora COVID-19*. Nota Técnica 11.

Fundação Oswaldo Cruz. (2020d). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. *Monitora COVID-19*. Nota Técnica 3. Recuperado de https://bigdata-covid19.icict.fiocruz.br

Guimarães, R. M., Eleuterio, T. D. A., & Monteiro-da-Silva, J. H. C. (2020). Estratificação de risco para predição de disseminação e gravidade da Covid-19 no Brasil. *Revista Brasileira De Estudos De População*, 37,.1-17. DOI: http://dx.doi.org/10.20947/s0102-3098a0122

Iritani, O., Okuno, T., Hama, D., Kane, A., Kodera, K., Morigaki, K., Terai, T., Maeno, N., & Morimoto, S. (2020). Clusters of covid-19 in long-term care hospitals and facilities in japan from 16 january to 9 may 2020. *Geriatrics & gerontology international*, 20(7), 715-719. DOI: 10.1111/ggi.13973

James, N., & Menzies, M. (2020). Cluster-based dual evolution for multivariate time series: Analyzing covid-19. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30. DOI: https://doi.org/10.1063/5.0013156

Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice hall Upper Saddle River, NJ, Upper Saddle River, 5.

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel sars-cov-2. *Gene reports*, 19, 1-6. DOI: https://doi.org/10.1016/j.genrep.2020.100682

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 90-95.

Kumar, S. (2020). Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis. *Annals of Data* Science, 7(3), 417-425. DOI: https://doi.org/10.1007/s40745-020-00289-7

Letko, M., Marzi, A., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature Microbiology*, 5, 562-569. DOI: https://doi.org/10.1038/s41564-020-0688-y

Maciel, E. L., Jabor, P., Goncalves Júnior, E., Tristão-Sá, R., Lima, R. C. D., Reis-Santos, B., Lira, P., Bussinguer, E. C. A., & Zandonade, E. (2020). Fatores associados ao óbito hospitalar por covid-19 no Espírito Santo. *Epidemiologia e Serviços de Saúde*, 29(4), 1-11. DOI: 10.5123/S1679-49742020000400022

Nassiri, R. (2020). Perspective on Wuhan viral pneumonia. *Advances in Public Health, Community and Tropical Medicine*, 2, 1-3.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka R. (2018). *Metodologia da pesquisa científica. [e-book]*. Santa Maria. Ed. UAB/NTE/UFSM. Recuperado de https://repositorio.ufsm.br/bitstream/handle/1/15824/Lic_Computacao_Metodologia-Pesquisa-Cientifica.pdf?sequence=1.

R Core Team. R: *a language and environment for statistical computing*. Vienna, 2020. Recuperado de https://www.Rproject.org/

Ratkowsky, D., & Lance, G. (1978). Criterion for determining the number of groups in a classification. *Australian Computer Journal*, 10(3), 115-117.

Salathé, M., Althaus, C.L.b., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., Zwahlen, M., Senti, G., Battegay, M., Wilder-Smith, A., Eckerle I., Egger M., & Low N. (2020). Covid-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. *Swiss Medical Weekly*, 150, 1-3. DOI: https://doi.org/10.4414/smw.2020.20225

Souza, C. D. F. D., Paiva, J. P. S. D., Leal, T. C., Silva, L. F. D., & Santos, L. G. (2020). Evolução espaçotemporal da letalidade por COVID-19 no Brasil, 2020. *Jornal Brasileiro de Pneumologia*, 46(4), 1-3. DOI: https://doi.org/10.36416/1806-3756/e20200208

Stier, A., Berman, M., & Bettencourt, L. (2020). COVID-19 attack rate increases with city size. *Mansueto Institute for Urban Innovation Research Paper Forthcoming*. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3564464

Velavan, T. P., & Meyer, C.G. (2020). The COVID-19 epidemic. *Tropical Medicine and International Health*, 25(3), 278-280. DOI: 10.1111/tmi.13383

Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, 395, 470-473. DOI: https://doi.org/10.1016/S0140-6736(20)30185-9

Werneck, G. L., & Carvalho, M. S. (2020). A pandemia de COVID-19 no Brasil: crônica de uma crise sanitária anunciada. *Cadernos de Saúde Pública*, 36(5), 1-4. DOI: 10.1590/0102-311X00068820

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the covid-19 cases dataset. *Data in Brief*, 31, 1-8. DOI: https://doi.org/10.1016/j.dib.2020.105787

Porcentagem de contribuição de cada autor no manuscrito

Henrique José de Paula Alves – 35%

Felipe Augusto Fernandes – 20%

Kelly Pereira de Lima – 20%

Ben Dêivide de Oliveira Batista – 20%

Tales Jesus Fernandes – 5%