UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
DEPARTAMENTO DE ENGENHARIA QUÍMICA

DIEGO DE SANT'ANA COIMBRA PEREIRA

ANÁLISE ESPAÇO-TEMPORAL MULTIVARIADA DA EVOLUÇÃO DA COVID-19 NO BRASIL

Salvador
2021

DIEGO DE SANT'ANA COIMBRA PEREIRA

ANÁLISE ESPAÇO-TEMPORAL MULTIVARIADA DA EVOLUÇÃO DA COVID-19 NO BRASIL

Monografia de Trabalho Final de Curso de Graduação apresentada ao Departamento de Engenharia Química da Escola Politécnica da Universidade Federal da Bahia, como requisito parcial para obtenção do título de Bacharel em Engenharia Química.

Orientador: Prof. Dr. Marcio Luis Ferreira Nascimento

Salvador

2021

Ficha catalográfica fornecida pelo Sistema Universitário de Bibliotecas da UFBA – SIBI/UFBA para ser confeccionada pelo autor.

XXXXX
2021

Pereira, Diego de Sant'ana Coimbra

Análise Espaço-Temporal Multivariada da Evolução da COVID-19 no Brasil / Diego de Sant'ana Coimbra Pereira. 2021.

91f.: il. 30 cm

Orientador: Prof. Dr. Marcio Luis Ferreira Nascimento

Monografia de Trabalho Final de Curso (Graduação) – Escola Politécnica - Universidade Federal da Bahia, Salvador, 2021.

Bibliografia: f. XXX – XXX

1. Palavras Chaves. 2. COVID-19. 3. Análise Multivariada. 4. Análise Espaço-Temporal. 5. Análises Hierárquica e Não Hierárquica. I. Nascimento, Marcio Luis Ferreira. II. Universidade Federal da Bahia; Departamento de Engenharia Química. III. Título.

CDD20. Ed. – xxx

Folha de Aprovação

DIEGO DE SANT'ANA COIMBRA PEREIRA

ANÁLISE ESPAÇO-TEMPORAL MULTIVARIADA DA EVOLUÇÃO DA COVID-19 NO BRASIL

Monografia de Trabalho Final de Curso de Graduação apresentada ao Departamento de Engenharia Química da Escola Politécnica da Universidade Federal da Bahia, como requisito parcial para obtenção do título de Bacharel em Engenharia Química

BANCA EXAMINADORA

Prof. Dr. Marcio Luis Ferreira Nascimento (Orientador)
Departamento de Engenharia Química – Escola Politécnica da UFBA

Prof. Dr. Marcus Vinícius Americano da Costa Filho
Departamento de Engenharia Química – Escola Politécnica da UFBA

Prof. Dr. Ângelo Conrado Loula
Departamento de Ciências Exatas – UEFS

Salvador / BA, 26 de novembro de 2021

Agradecimentos

Começo os meus agradecimentos com os meus pais, Rogério e Cláudia, que nunca me deixaram passar por maiores dificuldades e sempre me ensinaram que a educação e a dedicação libertam e transformam. Obrigado por toda a rede de apoio e paciência, sem vocês eu não seria nada. Amo vocês!

Agradecimentos especiais, também, à dona Sônia, minha amada avó, e à dona Sandra, minha tia e segunda mãe, que sempre me forneceram carinho, suporte e lições que levarei para a vida. Além delas, aproveito para dedicar este trabalho e agradecer imensamente ao meu eterno e muito amado avô Vivaldo. Que do Céu o senhor possa me enxergar e ver o quanto tenho orgulho de ser seu neto e saiba que jamais te esquecerei!

Ainda para a minha família, agradeço ao meu tio Marcos e à tia Marília, a quem tenho muito apreço e orgulho! Aos meus avô e avó, Jair e Neusa, e às minhas tias, Alessandra e Érica, obrigado pelo amor, carinho, cuidado e educação que pude ter com vocês! Levarei para sempre comigo, aonde eu estiver!

Aos meus irmãos, Felipe, Gabriel e Rodrigo, agradeço todos os momentos que passamos juntos e todo o suporte e amizade que temos. Amo vocês!

À minha amada namorada, Arianne, com quem descobri os verdadeiros significados de amar e cuidar. Obrigado por todo o suporte e paciência ao longo de todos esses anos! Te amo muito! Aproveito para estender os agradecimentos à sua família, que agora também é minha: meus queridos sogro e sogra, Reinaldo e Solange, e tia Maura. Obrigado por tudo!

Aos meus grandes e valiosos amigos: Matheus, Bruno, Felipe, Pedro, Thales, Bernardo, Caio (s), Matheus (Azeite), Rafael, Isaias, Clara, Carol, Stephanie, Vinicius, Cleice, Sarinha, Dandara, Erick e à minha querida turma 2016.1. Levarei vocês para sempre comigo! Obrigado por tudo!

Ao meu professor orientador Márcio. Prof., obrigado por toda a paciência, confiança e pelo compartilhamento de tanto conhecimento para a execução deste trabalho. Conseguimos!

PEREIRA, Diego de Sant'ana Coimbra. **Análise Espaço-Temporal Multivariada da Evolução da COVID-19 no Brasil**. 2021. Monografia de Trabalho Final de Curso de Graduação (Bacharel em Engenharia Química) – Escola Politécnica, Universidade Federal da Bahia, Brasil.

Resumo

A COVID-19, doença relacionada à infecção respiratória causada pelo coronavírus SARS-CoV-2, surgiu em dezembro de 2019 e foi considerada como pandemia pela Organização Mundial de Saúde (OMS) em março de 2020. No Brasil, desde março de 2020 até maio de 2021, mais de 16 milhões de casos haviam sido diagnosticados, com mais de 460 mil mortes, trazendo fortes repercussões políticas, sociais e econômicas. Com o objetivo de analisar de forma abrangente e quantitativa o impacto da COVID-19 no Brasil desde seu início, além de avaliar a efetividade das políticas públicas adotadas e o impacto da aplicação das vacinas, o presente trabalho elencou, para todas as 27 Unidades Federativas (UF, a saber: 26 estados mais o Distrito Federal), 18 variáveis de diferentes categorias para algumas semanas epidemiológicas do período analisado, sendo utilizado dados oficiais de portais do governo brasileiro: DataSUS, TCU, IBGE, IPEA, ANS e Painel COVID-19, atualizado pelo Ministério da Saúde. Inicialmente, para avaliar a configuração de grupos entre as UF no país, alocadas por similaridade, aplicou-se as técnicas de análise de agrupamentos hierárquica (dendrograma) e não-hierárquica (*K-Médias*). Ademais, utilizou-se a análise fatorial por componentes principais para a criação de um indicador capaz de ordenar todas as UF a partir das 18 variáveis, chamado de Índice COVID-19 (IC19), que possibilitou verificar a mudança espaço-temporal da doença para as UF nos diferentes momentos da pandemia. Como resultados principais das técnicas de agrupamento, percebeu-se a indicação de 3 a 5 grupos de acordo com a análise hierárquica e de 4 a 6 grupos para a análise não-hierárquica, tendo sido possível verificar padrões e características específicas para cada um deles em cada semana epidemiológica. Para a análise fatorial, os estados de São Paulo (SP), Rio de Janeiro (RJ) e Minas Gerais (MG) apresentaram os maiores índices IC19, enquanto Amapá (AP), Acre (AC) e Maranhão (MA) apresentaram os menores. Tal proposta demonstrou a capacidade de auxiliar gestores de órgãos públicos quanto a tomadas de decisão, dentre elas a priorização de recursos.

Palavras-chave: COVID-19, Análise Multivariada, Análise Espaço-Temporal, Análise Hierárquica e Não Hierárquica, Analise Fatorial.

PEREIRA, Diego de Sant'ana Coimbra. **A Multivariate Analysis on Spatiotemporal Evolution of COVID-19 in Brazil**. 2021. Undergraduate Project Report (Bachelor's Degree in Chemical Engineering) – Polytechnic School, Federal University of Bahia, Brazil.

Abstract

COVID-19, disease related to the respiratory infection caused by the SARS-CoV-2 coronavirus, emerged in December 2019 and it was considered an epidemic by the World Health Organization (WHO) in March 2020. In Brazil, between March 2020 and May 2021, over 16 million cases were confirmed, with over 460,000 deaths, bringing relevant political, social and economic consequences. With the purpose of analyzing broadly and numerically the COVID-19 impact in Brazil since the beginning, besides evaluating the effectiveness of the government guidelines and the impact of vaccines, this work brought up, for all the 27 Federative Units (FU, i.e. 26 states and the Federal District), 18 variables from different categories for a few epidemiological weeks of the given time period, using official data from the government's database: DataSUS, TCU, IBGE, IPEA, ANS and Painel COVID-19, updated by the Ministry of Health. Initially, in order to evaluate the groups formed between all the FU, allocated by similarity, hierarchical (dendrogram) and non-hierarchical (K-Means) clustering techniques have been applied. Additionally, a principal component factor analysis was used to create an index capable of ordering the FU by the set of 18 variables, called COVID-19 Index (IC19), which made possible to verify the space-time changes of the disease for the FU in different pandemic moments. As main results of the clustering methods, it was noticed a classification between 3 and 5 groups for the hierarchical method and 4 to 6 groups for the non-hierarchical, allowing patterns and specific characteristics to be seen for each one in all epidemiological weeks. For the factor analysis, the states of São Paulo (SP), Rio de Janeiro (RJ) and Minas Gerais (MG) had the highest IC19 scores, while Amapá (AP), Acre (AC) and Maranhão (MA) had the lowest. This proposal has successfully shown the capability of helping public managers to make decisions, *e.g.* prioritizing resources.

Keywords: COVID-19, Multivariate Analysis, Spatiotemporal Analysis, Hierarchical and Non-Hierarchical Analysis; Factor Analysis.

Lista de Figuras

Lista de Tabelas

LISTA DE ABREVIATURAS

ANS	Agencia Nacional de Saúde Suplementar
COVID-19	*Corona Virus Disease – 2019*
IBGE	Instituto Brasileiro de Geografia e Estatística
IPEA	Instituto de Pesquisa Econômica Aplicada
K-Means	*K*-médias
KMO	Estatística Kaiser-Meyer-Olkin
OMS	Organização Mundial da Saúde
SARS-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus* 2
SE	Semana Epidemiológica
SUS	Sistema Único de Saúde
TCU	Tribunal de Contas da União
UF	Unidade Federativa
UTI	Unidade de Terapia Intensiva
WHO	*World Health Organization*

Lista de Símbolos

X_{ij}	Variável X referente ao parâmetro j da observação i
Z_{ij}	Variável Z padronizada referente ao parâmetro j da observação i
$\mathbf{X}$	Matriz de dados
I	Matriz identidade
ρ	Matriz de correlações
d_{ik}	Distância euclidiana d entre observações i e k referente a um ou mais parâmetros
i, j, k, p, t	Índices: i (observações, objetos ou lições); j (variáveis, parâmetros ou características); k, p, t (índices auxiliares)
l, c	Índices para linhas e colunas
λ	Autovalor
ν	Autovetor
H_0	Hipótese nula
H_1	Hipótese alternativa
K	Número de grupos selecionado para rotina de K-médias

SUMÁRIO

1. INTRODUÇÃO

Em dezembro de 2019, órgãos de saúde chineses reportaram grupos de pacientes com pneumonia cuja causa ainda era desconhecida, relacionada localmente a um mercado em Wuhan, na província de Hubei. Especificamente em 31 de dezembro daquele ano, o Centro de Controle e Prevenção de Doenças da China, com auxílio das autoridades locais, conduziu uma investigação epidemiológica sobre o surto, sendo descoberto um novo tipo de coronavírus designado SARS-CoV-2 (ZHU *et al.*, 2020).

A COVID-19, doença relacionada ao coronavírus, consiste em uma infecção respiratória aguda grave que se tornou uma pandemia e emergiu como um dos principais problemas de saúde, economia e desafios geográficos e sociais neste século (ANÔNIMO, 2020). O surto foi declarado emergência de saúde pública pela Organização Mundial da Saúde (OMS) em 30 de janeiro de 2020 e, no dia 11 de março, a ONU declarou pandemia. Em termos gerais, os sintomas clínicos mais comumente relatados a esta doença são: tosse seca, febre, dispneia (falta de ar), fadiga, ageusia (perda ou redução do sentido do paladar), anosmia (perda ou redução do olfato) ou quaisquer de suas combinações (SOUZA *et al.*, 2020[a]).

Apesar de representar uma doença inédita e perigosa, tendo sido presenciado aumentos expressivos em termos de quantidade de casos e de óbitos por todo o mundo, a COVID-19 segue, em geral, um padrão definido e modelável, já visto em outros surtos de doenças infecciosas, letais ou não (NASCIMENTO, 2020*b*). É de se esperar, portanto, que as ferramentas matemáticas à disposição da comunidade científica possam auxiliar governos e agentes públicos quanto ao direcionamento efetivo de recursos e priorização de atuação dentro das diversas regiões de um país.

Em termos históricos, um dos primeiros matemáticos a fazer uso deste artifício foi o suíço Leonhard Euler (1707 – 1783, Figura 1) que, no estudo de crescimento de populações, estabeleceu que “o crescimento de uma população, em um determinado instante, é proporcional ao número de indivíduos existentes naquele instante”, sendo visto, portanto, um crescimento exponencial (EULER, 1760). É importante salientar, obviamente, que este crescimento ocorre sob uma ótica teórica e em ambiente controlado, tendo em vista que se faz necessária a presença de condições satisfatórias para o desenvolvimento de uma população, como disponibilidade de alimentos, espaço, medicamentos, etc. (NASCIMENTO, 2020*b*). Além das contribuições de Euler, destaca-se o trabalho pioneiro de um outro matemático suíço, Daniel Bernoulli (1700 – 1782, Figura 2), que utilizou um modelo estatístico para estudo de enfermidades (neste caso, para a varíola, doença infecciosa que se espalhava rapidamente à época) (BERNOULLI, 1760).

Figura 1. Leonhard Paul Euler (1707 - 1783), matemático e físico suíço.

Figura 2. Daniel Bernoulli (1700-1782), matemático e físico suíço.

Em seu trabalho, Bernoulli buscou comparar, de forma matemática, o benefício do incipiente processo de inoculação, isto é, introduzir o patógeno no corpo humano, com os riscos de óbito na pessoa infectada. O processo mencionado chegou a ser utilizado por figuras notórias da época, como o filósofo francês François-Marie Arouet (1694 – 1778, Figura 3), mais conhecido como Voltaire (NASCIMENTO, 2020*b*).

Já a vacina antivariólica foi desenvolvida pelo médico e naturalista inglês Edward Jenner (1749 – 1823, Figura 4), em 1796, após a observação de que as pessoas que tiveram contraído a varíola bovina, variante benigna da mesma doença acometida em humanos, criaram resistência e estavam protegidas da sua forma mais grave (JENNER, 1802). Curiosamente, a palavra “vacina” deriva de *Variolae vaccinae*, literalmente varíola da vaca, sendo Jenner o pioneiro da vacinação em massa (NASCIMENTO, 2020*b*).

No entanto, vale ressaltar que a imunização no Ocidente se deve às observações da aristocrata, escritora, poeta e feminista inglesa Mary Wortley Montagu (1689 – 1762). Ao casar-se com o embaixador britânico e morar na Turquia, percebeu que muitas mulheres turcas não

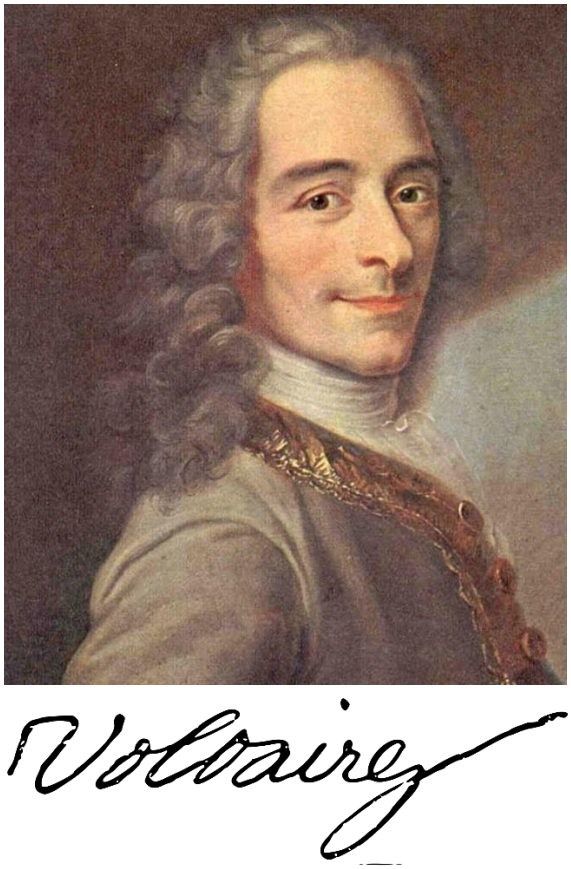

Figura 3. François-Marie Arouet (1694 - 1778), escritor, ensaísta e filósofo iluminista francês mais conhecido como Voltaire.

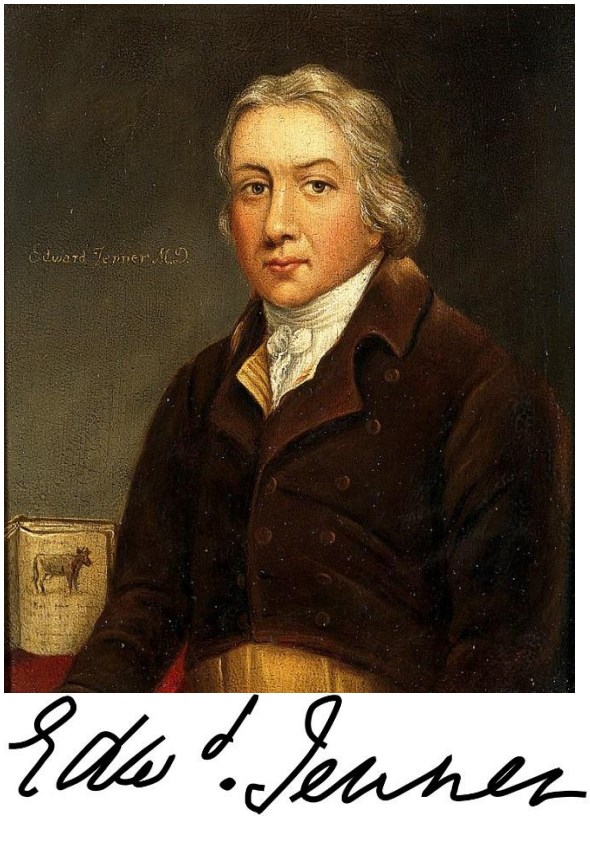

Figura 4. Edward Jenner (1749 - 1823), médico e naturalista inglês.

tinham marcas da doença, muito comum à época. Desde pelo menos o primeiro milênio havia a prática de inocular esta doença em regiões da Ásia Oriental e que se popularizou na época de Lady Montagu. Basicamente, as pessoas inseriam propositadamente o vírus, via pequenas quantidades de pus contendo varíola por meio de pequenos cortes no pulso ou tornozelo. Assim, as pessoas contraiam a doença de forma leve, sendo mais resistentes a novas infecções. Sem temer riscos, Montagu efetuou a prática em seu filho caçula, que sobreviveu ao contágio (WILLETT, 2021). Ao retornar à Grã-Bretanha, defendeu a prática, que não havia sido bem recebida pela sociedade de médicos. Foi através dos registros como os de Montagu que Voltaire tomou conhecimento da prática da inoculação que lhe salvou a vida.

Em relação especificamente ao Brasil, o primeiro caso de COVID-19 foi confirmado em 26 de fevereiro de 2020 por um viajante que voltava do norte da Itália para São Paulo e, no dia 17 de março, foi notificada a primeira morte pela doença na 12ª semana epidemiológica (12ª SE) (OLIVEIRA *et al*., 2020). Em 23 de maio de 2020 (SE 21), o país já ocupava o 3º lugar em número de casos confirmados no mundo (347.398) e em óbitos (22.013), embora estes dados

sejam provavelmente subestimados (SUS, 2020). Apenas três semanas depois, o Brasil atingiu o 2º lugar em termos de infecções por SARS-CoV-2 (828.810) bem como em mortes (41.828) (24ª SE).

A maioria dos 26 governadores estaduais brasileiros, bem como o Distrito Federal, impuseram quarentenas e até mesmo alguns bloqueios para evitar a propagação do vírus em todos as cinco regiões administrativas do Brasil. No entanto, desde o início da pandemia, viu-se quase nenhuma ou pouca liderança por parte do Governo Federal (ANÔNIMO, 2020), que apresentou uma reação de *laissez faire*[1] à epidemia de COVID-19 (CONDE, 2020).

Apesar da dificuldade de se enfrentar um vírus desconhecido e mortal, pesquisas iniciaram em praticamente todo o mundo ainda em 2020. No entanto, um grande entrave foi o negacionismo, tendo como defensores alguns líderes mundiais. A desinformação foi disseminada, inclusive por meios oficiais no Brasil, principalmente com defesas a medicamentos sem nenhuma evidencia científica.[2] O negacionismo foi tão grande que órgãos de imprensa precisaram passar a divulgar os dados da pandemia em seus boletins diários, pois faltava transparência aos dados oficiais. Esta decisão foi tomada pois o governo brasileiro passou a restringir o acesso a informações. [3] Outro aspecto lamentável foi a falta da aplicação de recursos para estudos e pesquisas no país, destacando-se a atuação de redes baianas como a CoVida (https://redecovida.org) e o portal baiano Geocovid (https://portalcovid19.uefs.br). Apesar de trabalharem na linha de frente da pandemia, essas redes não tiveram ofertados recursos, entre eles, para coletar, minerar e divulgar os dados diários de casos e óbitos de COVID-19. Outro destaque é o projeto Mandacaru, de abrangência regional.

---

[1] "E daí? Lamento. Quer que eu faça o que? Eu sou Messias, mas não faço milagre". Exatas palavras do presidente brasileiro de extrema direita Jair Messias Bolsonaro (*n*. 1955) após questionamento de repórteres sobre a perda de 474 vidas num único dia pelo coronavírus na terça-feira, dia 28 de abril de 2020. Até aquele momento haviam sido registrados 5.017 óbitos devido a pandemia.

[2] "Gripezinha", "Resfriadinho" e "Histórico de Atleta". Palavras do presidente brasileiro em célebrer pronunciamento em cadeia nacional no dia 24 de março de 2020. Duas semanas antes, a ONU tinha declarado a doença do coronavírus-19 como pandemia, ou seja, vírus estava circulando em todos os continentes e havia ocorrência de casos oligossintomáticos, o que dificultava a identificação.

[3] "Acabou matéria do Jornal Nacional". Palavras do presidente brasileiro em pronunciamento no dia 5 de junho de 2020. Tal declaração ocorreu sem que ninguém fizesse qualquer menção a nenhum órgão de imprensa específico. O presidente ainda alegou que o atraso se devia à necessidade de apresentar dados mais consolidados, mas não explicou por que, por mais de 70 dias, foi possível consolidar os dados mais cedo. E nem por que os números que eram divulgados às 22h constavam de uma planilha que atualizava dados até as 19h.

Um dos primeiros trabalhos de análise multivariada de dados da pandemia no país foi publicado por Nascimento (2020*a*), considerando dados referentes às SE 14 a SE 32 por meio de um quantitativo de 10 variáveis. Guimarães, Eleuterio e Monteiro-da-Silva (2020) conduziram estudo de análise fatorial em relação ao risco de disseminação e gravidade da COVID-19 no Brasil, enquanto Ferraz *et al.* (2021) construíram um índice de infraestrutura de saúde das UF no país, também utilizando a análise fatorial. Já Souza, Marques e Amorim (2020) realizaram uma análise de agrupamentos hierárquica em relação à vulnerabilidade e à incidência no Nordeste brasileiro, enquanto Alves *et al.* (2020) aplicaram a técnica de agrupamentos não-hierárquica para analisar dados de incidência, prevalência e letalidade para diferentes momentos da pandemia da COVID-19, em 2020. Este trabalho pretende ser mais abrangente, ao estender a análise por mais meses, utilizando mais variáveis e ferramentas.

Com o intuito de visualizar a evolução do quadro da COVID-19 no Brasil ao longo das semanas epidemiológicas, com destaque para número de casos, número de óbitos e índice de fatalidade, construiu-se os gráficos trazidos através das Figuras 5, 6 e 7, respectivamente. Além das informações quantitativas, é possível visualizar, nas respectivas figuras, momentos e frases marcantes relacionadas à postura governamental ao longo da pandemia.

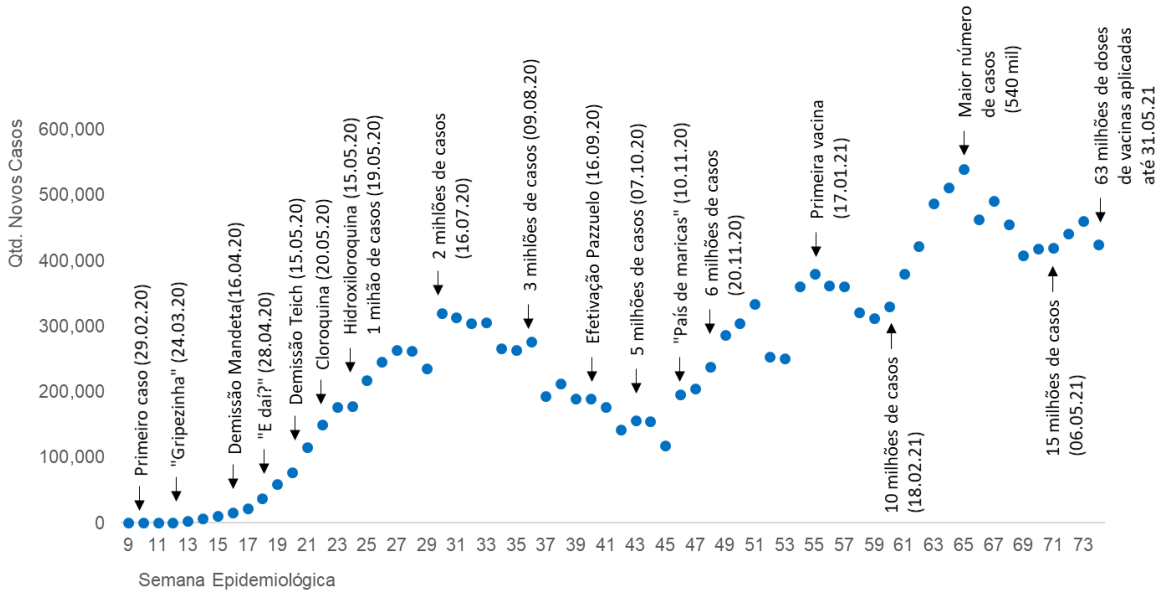

Figura 5. Casos de COVID-19 no Brasil relacionado a alguns marcos epidemiológicos, econômicos, políticos e sociais em função das semanas epidemiológicas (SE) (DATASUS). Entre aspas encontram-se alguns posicionamentos do presidente brasileiro referente à situação.

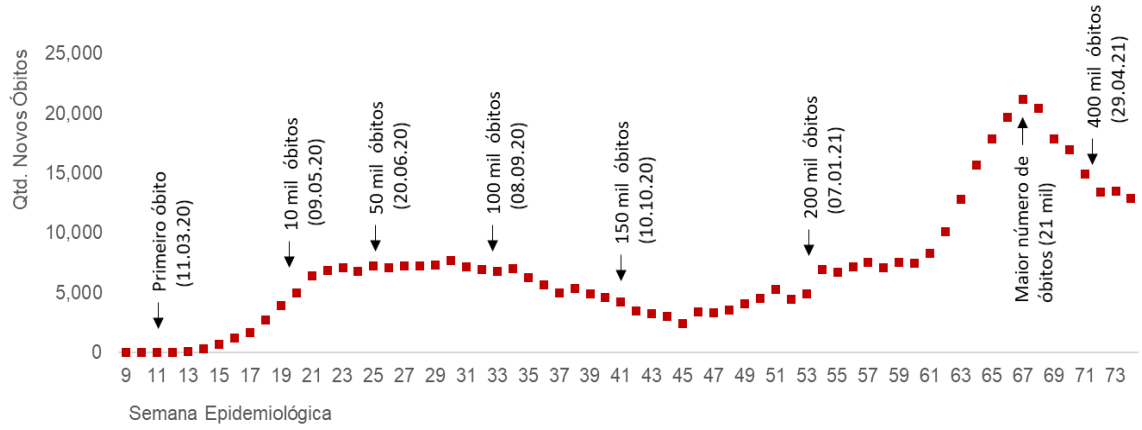

Figura 6. Óbitos de COVID-19 no Brasil relacionado a alguns marcos em termos do número total de perdas humanas em função das semanas epidemiológicas (SE) (DATASUS).

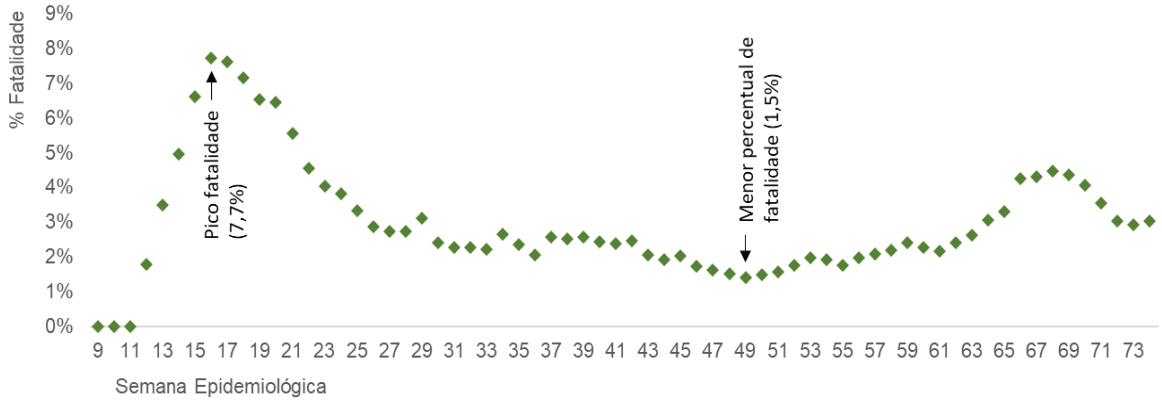

Figura 7. Percentual de fatalidade (*i.e.*, razão entre o número de óbitos pelo número de casos) da COVID-19 no Brasil em função das semanas epidemiológicas (SE) (DATASUS).

1.1. HIPÓTESE

Seria possível, com base na quantidade de dados disponível por meio de boletins semanais, compreender a disseminação da pandemia no país e estabelecer uma ordenação definida por critérios vinculados a determinadas variáveis, definidas por órgãos de saúde como a OMS e o SUS? Quais seriam estas variáveis e se, dentre estas, apenas algumas seriam suficientes? Haveria um padrão na configuração espaço-temporal da disseminação da COVID-19 no Brasil?

Com base nos questionamentos supracitados, formulou-se os objetivos gerais e específicos, com o intuito de aplicar técnicas de análise multivariada e, assim, interpretar matematicamente os dados de diversos parâmetros relacionados à COVID-19 no Brasil.

1.2. OBJETIVOS

Como objetivos principais do presente trabalho, vinculados à hipótese, destacam-se:

- Realizar estudos baseados em técnicas de análises matemáticas multivariadas para auxiliar na compreensão de dados epidemiológicos da COVID-19, com base em variáveis (ou parâmetros) selecionados, usuais para análise espaço-temporal da pandemia no país;
- Fornecer informações baseadas em métodos estatísticos e escolha de variáveis visando à classificação / ordenação de unidades federativas (UF) durante a pandemia, para contribuir quanto à tomada de decisão e alocação de recursos por parte de gestores e agentes públicos.

A partir do supracitado, pretende-se obter os seguintes quesitos específicos:

- Agrupar as UF com base nas variáveis selecionadas para estudo, utilizando a análise hierárquica (Dendrograma) e a não-hierárquica (*K-Médias*);
- Analisar se as variáveis disponíveis seriam suficientes via técnica de Componentes Principais e, eventualmente, avaliar quais destas seriam determinantes.
- Criar o Índice COVID-19 (IC19), parâmetro que resume o comportamento das 17 a 18 variáveis elencadas no estudo e traduz a situação e a evolução temporal da COVID-19 por UF, sendo obtido através da Análise Fatorial por Componentes Principais.

2. FUNDAMENTAÇÃO TEÓRICA

A teoria dos métodos de agrupamento hierárquico e não hierárquico, além da análise fatorial por componentes principais, são apresentadas no presente capítulo, cujas bases foram de fundamental importância para a compreensão e o desenvolvimento do trabalho.

2.1. ANÁLISE DE AGRUPAMENTOS

A análise de agrupamentos pode ser definida como um conjunto de técnicas cujo objetivo é agrupar objetos (ou observações, rotuladas pelo índice *i*) com base nas características possuídas (ou ainda parâmetros ou variáveis, rotuladas pelo índice *j*). A versatilidade quanto à aplicação da ferramenta é notória e, dentre os diversos exemplos, é possível citar estudos de interdependência entre indivíduos de uma determinada população e variáveis como renda familiar e idade, com viés de se analisar eficiência de programas sociais. Também é possível citar análise de desempenho de estudantes em múltiplas disciplinas em instituições de ensino, com o objetivo de avaliar a eficácia da abordagem pedagógica das mesmas. Em linhas gerais, as técnicas de agrupamento podem ser definidas como exploratórias, tendo em vista não possuir o viés preditivo e, em muitos casos, apresentarem a necessidade de reaplicação completa dos cálculos na presença de novas observações na amostra em estudo (FÁVERO e BELFIORE, 2017).

O principal objetivo da ferramenta é gerar agrupamentos (conglomerados, aglomerados ou *clusters*) que apresentem, em relação às observações agrupadas, elevadas homogeneidade interna e heterogeneidade externa aos aglomerados (HAIR *et al.*, 2019). Para que seja mensurada a correta aplicação, faz-se necessário definir os conceitos de distância (ou dissimilaridade) e, também, os métodos de agrupamento comumente implementados (hierárquicos e não-hierárquicos).

2.1.1. Medidas de distância (dissimilaridade)

Para que seja possível avaliar a semelhança entre diferentes observações, é preciso estabelecer uma métrica adequada que forneça numericamente tal interpretação. Dentre as possíveis abordagens, a mais utilizada é a do cálculo de distâncias, tendo em vista o seu entendimento intuitivo de que, quanto maior a medida da distância entre observações em um determinado espaço amostral, maior é a dissimilaridade (ou menor é a semelhança) (HAIR *et al.*, 2019).

Como dito por Fávero e Belfiore (2017), existem diferentes formas de se avaliar a dissimilaridade entre objetos i e k, cada uma dependente de pressupostos e objetivos do condutor do estudo. Em termos de aplicação, a abordagem mais utilizada é a da distância euclidiana, podendo ser vista na Eq. (2.1):

$$d_{ik} = \sqrt{\sum_{t=1}^{p}(X_{it} - X_{kt})^2} \tag{2.1}$$

em que i e k são dois objetos arbitrários, p é a quantidade de variáveis do sistema, X_{it} é o valor da variável X_t para o objeto i e X_{kt} é o valor da mesma variável para k.

Salienta-se, também, a importância da utilização de técnicas de padronização dos dados para que, durante a aplicação dos cálculos das distâncias, todas as p variáveis sejam igualmente importantes para a determinação de d_{ik}, principalmente em estudos que envolvam variáveis de diferentes unidades de medida (MANLY, 2008). Uma abordagem amplamente conhecida é a da padronização *Z*, conforme mostrado na Eq. (2.2):

$$Z_j = \frac{X_j - \bar{X}_j}{\sigma_j} \tag{2.2}$$

onde $\bar{X}_j$ é a média e σ_j é o desvio padrão da variável X_j.

2.1.2. Agrupamento hierárquico

O primeiro dos procedimentos de aglomeração citados é o hierárquico, que pode ser formalmente definido como o processo de aplicação de $n - 1$ "decisões de aglomeração" em série, sendo n o número de observações, visando combinar estas observações em hierarquia construída em uma estrutura denominada dendrograma (HAIR *et al.*, 2019).

As técnicas utilizadas para construção do dendrograma podem ser aglomerativas ou divisivas. Na primeira, todos os objetos começam isolados em grupos simples de um elemento, sendo progressivamente fundidos em *clusters* com base na menor dissimilaridade (ou menor distância) entre si, como visto na Seção 2.2.1, até que todos estejam situados no mesmo *cluster* (MANLY, 2008). Já para a segunda, tem-se a abordagem contrária, tendo em vista a presença inicial de um único grupo contendo todas as observações que é sucessivamente dividido até

atingir grupos de um único elemento (HAIR *et al*, 2019). Para o presente trabalho, a primeira técnica foi a escolhida para implementação, sendo um dos pioneiros da técnica o bioestatístico e entomologista austro-americano Robert Reuven Sokal (1926 - 2012).

Um outro conceito relevante relacionado à temática de agrupamento hierárquico é o de algoritmo de encadeamento. Isto se deve ao fato de que a consideração sobre a forma de se calcular distâncias entre objetos individuais e *clusters* desempenha um papel decisivo para a obtenção do resultado final, não sendo tão direto como o cálculo de dissimilaridade entre dois objetos visto anteriormente. Dentre as diversas abordagens, definiu-se para o presente trabalho a aplicação do método do matemático americano Joe Henry Ward Jr. (1926 - 2011) (WARD, 1963), que escolhe quais grupos deve combinar com base na minimização iterativa da função "custo de aglomeração", que representa o aumento da soma de erros quadráticos internos a cada *cluster* ("soma interna" ou "*within-sum"*) dentro do sistema como um todo. As Eqs. (2.3) e (2.4) apresentam os termos mencionados. De acordo com Milligan (1980) e Hair *et al.* (2019), o método de Ward e o algoritmo de agregação por média ("*average linkage*") representam as melhores opções de cálculo disponíveis, devido à estabilidade a à robustez teórica e prática.

$$SS = \sum_{k} \sum_{i \in R_k} (X_i - M_k)^2 \tag{2.3}$$

em que SS representa a soma dos erros quadráticos internos de um sistema aglomerativo, M_k é a média das observações pertencentes a um determinado *cluster* k e X_i é a observação i pertencente a este *cluster*.

O valor final da média de cada cluster M_k é obtida após a realização do cálculo iterativo que busca minimizar o valor de SS por meio da função "custo de aglomeração" Δ_{kt}, mostrada na Eq. (2.4):

$$\Delta_{kt} = \frac{n_k n_t}{n_k + n_t} (M_k - M_t)^2 \tag{2.4}$$

em que Δ é a função "custo de aglomeração" para dois *clusters* k e t, a ser minimizada pelo algoritmo, n_k e n_t são as respectivas quantidades de elementos (ou a cardinalidades) e M_k e M_t são as respectivas médias.

De acordo com Hair *et al.* (2019), os principais benefícios da análise hierárquica são:

- Simplicidade na interpretação do processo de agrupamento e alocação das

observações, através da utilização do dendrograma;

- Elevada aplicabilidade, devido à utilização ampla do conceito de dissimilaridade, o que torna possível agrupar dados numéricos ou categóricos;
- Rapidez na execução do método, sobretudo por conta da relativa simplicidade nos cálculos envolvidos e pelos avanços nos pacotes computacionais disponíveis.

Dentre as limitações, destacam-se a elevada sensibilidade da técnica à presença de *outliers* e a dificuldade em realizar estudos para grandes volumes de dados (HAIR *et al.*, 2019).

Uma das formas mais comuns de se utilizar o agrupamento hierárquico é a de obtenção de uma quantidade razoável e coerente de *clusters*, através da interpretação do dendrograma, fornecendo insumo para a aplicação do agrupamento não-hierárquico na sequência (FÁVERO e BELFIORE, 2017). Isto pode ser visto por meio da construção exemplificada na Figura 8. Na mesma, identificam-se as observações (X_1 a X_5) hipotéticas, as linhas tracejadas logo após cada agrupamento (I a III) e os respectivos saltos de distância desses agrupamentos (S_1 a S_3).

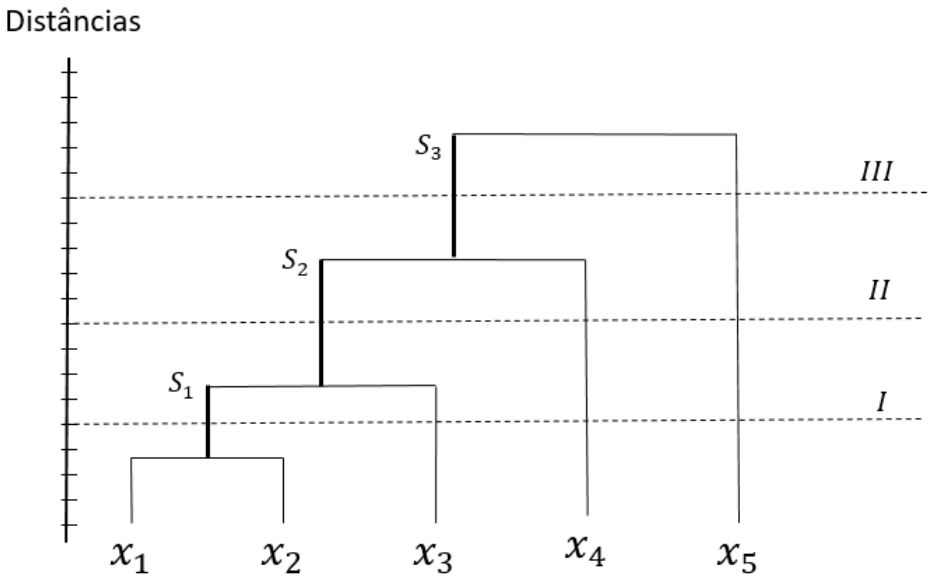

Figura 8. Representação esquemática de um dendrograma, adaptado de Fávero e Belfiore (2017).

A identificação da quantidade de *clusters* é baseada na medida de distância, especificamente nos saltos representados na Figura 8, tendo em vista que uma variação considerável entre um agrupamento e outro indica a importação de observações com características significativamente distintas. A partir da comparação de magnitude dos saltos, o pesquisador é capaz de estimar a linha horizontal tracejada correspondente e, com base na

quantidade de interseções da mesma com as linhas verticais, determina-se o número de *clusters* (FÁVERO e BELFIORE, 2017). Neste exemplo, caso o maior salto fosse entre S_1 e S_2, a linha tracejada escolhida seria a II e, com base no número de interseções verticais, seria possível indicar a presença de três grupos. A definição de grupos pelo dendrograma pode servir de *input* para uso da próxima rotina, denominada não hierárquica (*K*-médias).

2.1.3. Agrupamento não-hierárquico

A técnica de agrupamento não-hierárquico, diferente da técnica mostrada na seção anterior, não utiliza o conceito de dendrograma no seu princípio de funcionamento. Em alternativa, realiza a organização otimizada das observações dentro de um número *K* de *clusters* especificado "*a priori*" pelo pesquisador, isto é, como um dos parâmetros de iniciação (ou *input*) do algoritmo (HAIR *et al.*, 2019). Dentre as diversas opções existentes, o método de *K-Means* (ou "*K*-Médias") é o mais utilizado nos diversos âmbitos da ciência (FÁVERO e BELFIORE, 2019), também sendo implementado no presente trabalho.

A implementação do *K-Médias* pode ser representada por uma sequência de eventos bem definida (JOHNSON e WICHERN, 2007; FÁVERO e BELFIORE, 2019):

- Define-se uma quantidade *K* de *clusters* "*a priori*", conforme supracitado, com as suas respectivas "sementes de agrupamento" ou centróides, que representam a média de um determinado *cluster*, em um dado momento da construção do mesmo. Estas sementes podem ser especificadas pelo condutor do estudo ou observações aleatórias podem ser escolhidas para começar o processo (HAIR *et al.*, 2019);
- Após a especificação das sementes, designa-se cada observação a uma das sementes, tendo como base a menor medida de dissimilaridade (ou seja, proximidade), alocando-a ao *cluster* correspondente. Durante o processo, as observações são realocadas entre os grupos em formação, o que força a repetição do cálculo das sementes;
- A iteração é realizada até o atingimento do estado em que não é mais possível realocar as observações, devido à proximidade das mesmas em relação às respectivas sementes.

Conforme o processo iterativo acontece, as coordenadas das sementes devem ser recalculadas, como dito no segundo passo. Caso um determinado objeto X_j seja inserido no

cluster em análise, então o novo valor da semente pode ser encontrado pela Eq. (2.5).

$$M_{novo} = \frac{n_k M_k + X_j}{n_k + 1} \tag{2.5}$$

em que M_{novo} é o novo valor de média do *cluster* (centroide ou semente), M_k é o valor antigo e n_k é a cardinalidade do *cluster* k .

Para o caso em que X_j seja retirado do *cluster*, então o valor de M_{novo} pode ser encontrado pela Eq. (2.6). A Figura 9 ilustra o processo descrito para um caso hipotético de agrupamento em 2 *clusters*.

$$M_{novo} = \frac{n_k M_k - X_j}{n_k - 1} \tag{2.6}$$

Figura 9. Ilustração do algoritmo da técnica de *K-Médias*. (A) Observações de um determinado sistema. (B) Definido K = 2, escolhem-se aleatoriamente duas sementes para servir como iniciadoras do processo. (C) Com base na dissimilaridade, as observações são agrupadas nestes dois grupos. As sementes são então recalculadas no processo iterativo até as distâncias estabilizarem, devido às elevadas proximidades encontradas.

Além da definição "*a priori*", por parte do pesquisador, acerca da quantidade de grupos

a serem criados na análise não-hierárquica, existem alguns procedimentos que visam fornecer, numericamente ou graficamente, a indicação de um número ótimo de grupos. Dentre as opções, destacam-se o "método do cotovelo" ("*elbow method*") e a "pontuação de silhueta" (*"silhouette score"*).

A primeira forma consiste na análise do gráfico $SS \times K$, sendo SS a soma dos erros quadráticos internos de um sistema aglomerativo (mostrado na Seção 2.1.2 por meio da Eq. (2.3)) e K o número de *clusters*. O número ótimo é encontrado por meio do ponto de inflexão da curva, que se assemelha a uma inflexão que a literatura denomina de cotovelo, interpretado como aquele que fornece proporcionalmente a maior redução da variabilidade interna de cada *cluster* (KETCHEN e SHOOK, 1996; CHEN, SHI e WONG, 2021; HOZUMI *et al.*, 2021).

Quanto a segunda forma, define-se o número ótimo de *clusters* por meio da pontuação de silhueta (PS), através da Eq. (2.7).

$$PS = \frac{D_w - D_i}{\max(D_w, D_i)} \tag{2.7}$$

em que D_i é a distância média de cada observação à semente de aglomeração do grupo ao qual pertence e D_w é a distância média de cada observação à semente do grupo mais próximo.

O valor de PS diz numericamente, portanto, o quanto cada ponto dentro de um grupo está afastado dos pontos de um outro grupo vizinho. Este parâmetro varia de 0 a 1, em que quanto mais próximo da unidade, mais bem alocados estão os objetos nos respectivos *clusters*, sendo o oposto para valores próximos de zero (ROUSSEEUW, 1987; CHEN, SHI e WONG, 2021; BEHPOUR *et al.*, 2021). Logo, o número ótimo poderá ser visto em um gráfico $PS \times K$ como o ponto de máximo da curva.

A Figura 10 resume a interpretação do método do cotovelo e da pontuação de silhueta, com a identificação dos respectivos pontos ótimos por meio de um exemplo arbitrário. Salienta-se que os dois métodos podem fornecer resultados diferentes.

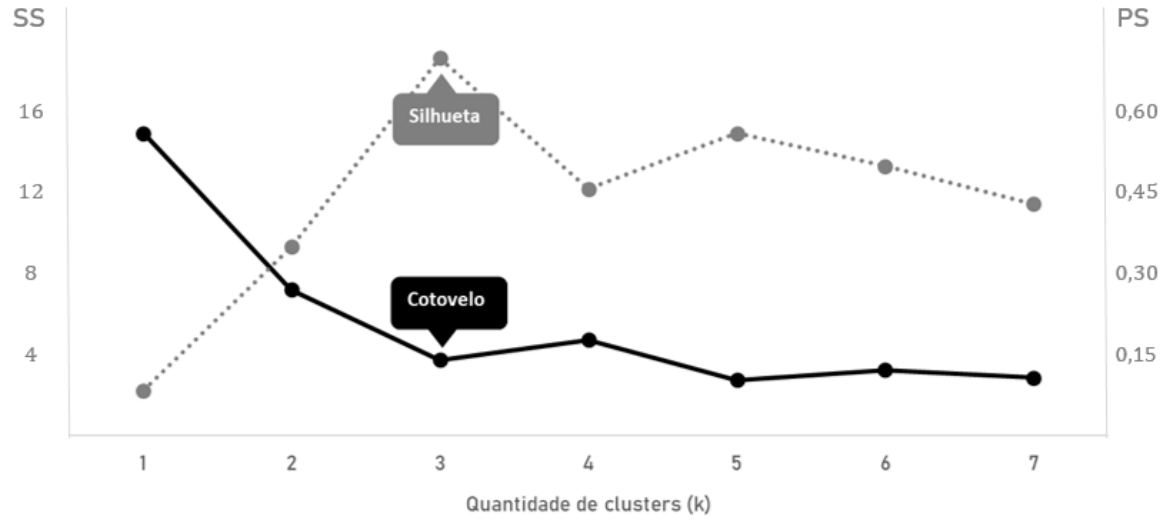

Figura 10. Comparação teórica das curvas de $SS \times K$, para o método do cotovelo, e $PS \times K$, para o método da silhueta, representando os respectivos números ótimos de *clusters* de acordo com cada uma.

De acordo com Hair *et al.* (2019), como vantagens dos métodos de agrupamento não-hierárquico, destacam-se:

- Resultados menos suscetíveis a *outliers* e ao tipo de cálculo da dissimilaridade escolhido;
- Capacidade de analisar volumes elevados de dados.

Já como limitação, pontua-se a dependência de uma boa especificação das "sementes" para obtenção de resultados superiores aos obtidos pela técnica hierárquica (HAIR *et al.*, 2019).

2.2. ANÁLISE FATORIAL POR COMPONENTES PRINCIPAIS

Dentre as diversas técnicas de análise exploratória, a análise fatorial se destaca quando existem, dentro de um sistema, muitas variáveis com coeficientes de correlação relativamente altos e se busca criar novos preditores, chamados de fatores, que reduzam a dimensionalidade do sistema sem perdas expressivas de informação. Tais coeficientes são atribuídos ao matemático e estatístico inglês Karl Pearson (1857 - 1936) (PEARSON, 1896). Para a determinação dos preditores citados, a abordagem dos componentes principais é a mais utilizada dentro da análise fatorial, partindo do conceito de fatores como combinações lineares das variáveis originais e, portanto, não correlacionados entre si (FÁVERO e BELFIORE, 2017).

Como dito por Fávero e Belfiore (2017), a análise fatorial por componentes principais tem quatro objetivos destacáveis:

- Redução de dimensionalidade ou redução estrutural do sistema;
- Verificação da validade de constructos;
- Extração de fatores para posterior uso em outras técnicas multivariadas que necessitam ausência de multicolinearidade (por exemplo, modelos de regressão);
- Construção de ordenamentos (*rankings*).

É interessante pontuar a semelhança entre a análise de agrupamentos e a análise fatorial, em relação à avaliação da estrutura do sistema. Enquanto a primeira realiza a agregação de observações sob o viés de distância, a segunda busca agregar variáveis por meio de padrões de variação (ou correlação) (HAIR *et al.*, 2019).

2.2.1. Conceito de Correlação Linear de Pearson e Fator

Com o intuito de aplicar satisfatoriamente a técnica de componentes principais, faz-se necessário compreender, primeiramente, o conceito de correlação linear de Pearson. Para dois parâmetros hipotéticos X_j e X_k, este é calculado de acordo com a Eq. (2.8).

$$\rho_{jk} = \frac{\sum_{t=1}^{n}\left(X_{jt} - \bar{X}_j\right)\left(X_{kt} - \bar{X}_k\right)}{\sqrt{\sum_{t=1}^{n}\left(X_{jt} - X_j\right)^2}\sqrt{\sum_{t=1}^{n}(X_{kt} - X_k)^2}} \tag{2.8}$$

em que $\bar{X}_j$ e $\bar{X}_k$ representam as médias de X_j e X_k, respectivamente.

Utilizando este conceito, torna-se possível construir a matriz de correlações $\boldsymbol{\rho}$, que apresenta as correlações lineares de cada par para um sistema de n variáveis conforme representado na Eq. (2.9).

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix} \tag{2.9}$$

A correlação de Pearson é, então, um indicador da relação linear entre dois parâmetros métricos, variando de −1 a 1. A obtenção de um valor próximo a esses extremos indica uma forte relação linear entre as variáveis analisadas, seja de forma inversa ($\rho_{jk} \approx -1$) ou direta

($\rho_{jk} \approx 1$), o que favorece a extração de um único fator. Já um valor de correlação próximo a zero ($\rho_{jk} \approx 0$) leva à constatação de uma relação linear muito fraca ou inexistente entre as variáveis, que tende a causar a extração de dois ou mais fatores (FÁVERO e BELFIORE, 2017).

Em relação especificamente aos fatores, estes podem ser entendidos como representações de dimensões latentes, capazes de explicar o comportamento de variáveis originais. Como forma de exemplificar o conceito, seja um sistema hipotético formado por três parâmetros (X_j, X_k e X_t), em que X_j e X_k são altamente correlacionados e X_t não apresenta correlação linear com as outros dois. Há, então, duas situações relacionando tais variáveis.

a) Relação entre X_j e X_k com $\rho_{jk} \approx 1$: linear.

b) Relação entre X_k e X_t com $\rho_{kt} \approx 0$. Mesmo comportamento para o par X_j e X_t: não-linear.

Portanto, devido à alta correlação entre X_j e X_k, ambas podem ser agrupadas em uma única variável nova (ou fator), apto para explicar o comportamento conjunto sem perdas expressivas de informação. Já X_t, não relacionado linearmente com os outros dois parâmetros, mantém-se em um fator separado, causando a redução de três para duas variáveis no sistema.

2.2.2. Testes de Adequação da Análise Fatorial: Estatística KMO e Teste de Bartlett

Como forma de avaliar a adequação da análise fatorial para o sistema em estudo, dois testes são aplicados: a estatística Kaiser-Meyer-Olkin (KMO) e o teste de esfericidade de Bartlett.

A estatística KMO, desenvolvida inicialmente pelo psicólogo americano Henry Felix Kaiser (1927 - 1992) (KAISER, 1970), é calculada de acordo com a Eq. (2.10), em que *l* e *c* representam as linhas e as colunas da matriz de correlações ρ, respectivamente, e φ é o coeficiente de correlação parcial entre duas variáveis *j* e *k* fixando-se *t*, exemplificado para um sistema hipotético e dado pela Eq. (2.11).

$$KMO = \frac{\sum_{l=1}^{n}\sum_{c=1}^{n}\rho_{lc}^{2}}{\sum_{l=1}^{n}\sum_{c=1}^{n}\rho_{lc}^{2} + \sum_{l=1}^{n}\sum_{c=1}^{n}\varphi_{lc}^{2}} \qquad (2.10)$$

$$\varphi_{jk,t} = \frac{\rho_{jk} - \rho_{jt}.\rho_{jt}}{\sqrt{\left(1 - \rho_{jt}^2\right).\left(1 - \rho_{kt}^2\right)}} \quad (2.11)$$

A KMO fornece informações de variância, calculando a proporção de variância comum a todas as variáveis na amostra em análise, com resultado dentro do intervalo de 0 a 1. Quanto mais próximo de 1, maior é o grau de compartilhamento de variância (*i.e.*, elevadas correlações de Pearson), enquanto valores próximos a 0, causados por baixas correlações de Pearson, podem indicar a inadequação da análise fatorial. Já os coeficientes de correlação parcial φ representam um ajuste ao cálculo de correlação de Pearson, através do cálculo da correlação entre duas variáveis desconsiderando os efeitos das demais variáveis presentes no sistema. Para que a análise fatorial seja considerada válida, os valores de φ devem ser baixos, já que este fato implica elevado percentual de variância compartilhado e a ausência de um ou mais parâmetros prejudica a qualidade da extração dos fatores (FÁVERO e BELFIORE, 2017).

A forma amplamente aceita de avaliação da adequabilidade, conforme resultado obtido para KMO, está resumida na Tabela 1.

Tabela 1. Relação entre a KMO e a adequação da análise fatorial.

Estatística KMO	**Adequação da análise fatorial**
$KMO > 0{,}90$	Excelente
$0{,}80 < KMO < 0{,}90$	Ótima
$0{,}70 < KMO < 0{,}80$	Média
$0{,}60 < KMO < 0{,}70$	Razoável
$0{,}50 < KMO < 0{,}60$	Má
$KMO < 0{,}50$	Inaceitável

Fonte: (FÁVERO E BELFIORE, 2017)

Um teste adicional é o de esfericidade desenvolvido pelo estatístico inglês Maurice Stevenson Bartlett (1910 – 2002) (BARTLETT, 1954), que consiste em comparar a matriz de correlações ρ com uma matriz identidade I de mesma dimensão por meio de um teste de hipóteses, como mostrado abaixo.

$$H_0: \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix} = I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$H_1: \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix} \neq I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Caso não haja evidência estatística sobre a diferença entre as duas matrizes, falha-se em rejeitar a hipótese nula e considera-se a análise fatorial como inadequada, tendo em vista que, estatisticamente, as correlações de Pearson entre as variáveis serão iguais a 0. O teste de esfericidade utiliza a curva χ^2 ("*chi*-quadrado"), calculado por meio da Eq. (2.12) (FÁVERO E BELFIORE, 2017).

$$\chi^2_{Bartlett} = -\left[(n-1) - \left(\frac{2k+5}{6}\right)\right] \ln|D| \tag{2.12}$$

tendo $\frac{k(k-1)}{2}$ graus de liberdade, n o tamanho amostral e k variáveis. O termo D representa o determinante da matriz ρ.

Logo, para rejeitar a hipótese nula, deve-se verificar se o valor calculado para $\chi^2_{Bartlett}$ é maior do que o valor crítico χ^2_C, para um dado grau de liberdade e nível de significância. De acordo com Fávero e Belfiore (2017), o teste de Bartlett é preferível frente à estatística KMO, devido ao primeiro associar o resultado a uma distribuição de probabilidades e a um teste de hipóteses, respaldando a tomada de decisão, enquanto o segundo representa um coeficiente calculado sem associar aos artifícios mencionados.

2.2.3. Definição dos Fatores por Componentes Principais

Tendo em vista que os fatores podem ser entendidos como combinações lineares das variáveis originais (ou como um "agrupamento de variáveis"), em um sistema de k variáveis será possível obter, portanto, k fatores, expressos de forma geral pela Eq. (2.13) (FÁVERO e BELFIORE, 2017).

$$F_{kj} = s_{1k}X_{1j} + s_{2k}X_{2j} + \cdots + s_{kk}X_{kj} \quad (2.13)$$

sendo s os *scores* fatoriais, obtidos a partir dos autovalores e dos autovetores da matriz ρ.

Estes parâmetros são amplamente utilizados no campo da Álgebra Linear e, de maneira formal, define-se autovetor como o vetor v não nulo de uma transformação linear, associado a um escalar λ, este chamado de autovalor, conforme mostra a Eq. (2.14). O autovetor v também tem a característica de não alterar a sua direção quando a transformação linear é aplicada. Para o cálculo dos mesmos, utiliza-se a equação característica de X, como mostra a Eq. (2.15) (ANTON e RORRES, 2012).

$$Xv = \lambda v \quad (2.14)$$

$$\det(\lambda \mathbf{I} - \mathbf{X}) = 0 \quad (2.15)$$

onde **I** é a matriz identidade, de mesma dimensão de **X**.

Para o sistema em análise, a Eq. (2.15) pode ser particularizada de acordo com a Eq. (2.16), com o autovalor λ elevado ao quadrado a título de convenção.

$$\det\left(\lambda^2 \mathbf{I} - \boldsymbol{\rho}\right) = 0 \quad (2.16)$$

A partir da Eq. (2.16), determina-se os k autovalores λ^2, relacionados aos seus respectivos fatores. Em seguida, é possível calcular os autovetores correspondentes, mediante a Eq. (2.17).

$$\begin{pmatrix} \lambda_k^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_k^2 - 1 & \cdots & -\rho_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_k^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{1k} \\ v_{2k} \\ \cdots \\ v_{kk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix} \quad (2.17)$$

Com os autovalores e autovetores determinados para o sistema, calcula-se os *scores* dos fatores de acordo com a Eq. (2.18).

$$S_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \cdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{2k}}{\sqrt{\lambda_k^2}} \\ \cdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix} \quad (2.18)$$

Por fim, a partir das informações encontradas nas equações supracitadas, é possível obter os fatores de forma padronizada, como mostrado na Eq. (2.19) por meio da manipulação das Eqs. (2.13) e (2.18), em que Z_{kj} representa o valor padronizado da variável X_j de acordo com a pontuação *Z*.

$$F_{kj} = \frac{v_{1k}}{\sqrt{\lambda_k^2}} Z_{1j} + \frac{v_{2k}}{\sqrt{\lambda_k^2}} Z_{2j} + \cdots + \frac{v_{kk}}{\sqrt{\lambda_k^2}} Z_{kj} \quad (2.19)$$

É importante ressaltar que, devido à padronização dos *scores* pelos autovalores, vista na Eq. (2.18), as variáveis que constituem o sistema e que fazem parte da Eq. (2.13) devem ser também padronizadas, de acordo com o procedimento de pontuação *Z* mostrado na Eq. (2.2). Dessa forma, ao final do processo, obtém-se fatores ortogonais entre si, com correlações de Pearson iguais a 0 (FÁVERO e BELFIORE, 2017).

O critério de escolha da quantidade de fatores que serão contabilizados na análise fatorial é diretamente relacionado aos autovalores, que representam o percentual da variância compartilhada pelas variáveis para a formação dos componentes principais respectivos. Logo, fatores que tenham autovalores pequenos representam, também, um percentual pequeno da variância compartilhada do sistema e, para autovalores menores que 1, isso significa que o fator em questão não explica nem ao menos o comportamento da variável original. A escolha de fatores com autovalores maiores que 1 é uma prática comum e é chamada de critério da raiz latente (FÁVERO e BELFIORE, 2017; MANLY, 2008).

2.2.4. Cargas Fatoriais e Comunalidade

Um outro conceito importante na análise fatorial por componentes principais é o de

cargas fatoriais, que pode ser entendido como as correlações lineares de Pearson existentes entre as variáveis originais e os fatores obtidos de acordo com os métodos mencionados anteriormente, como mostrado na Tabela 2 (FÁVERO E BELFIORE, 2017).

Tabela 2. Cargas fatoriais para cada par variável-fator.

Variável | Fator	F_1	F_2	...	F_k
X_1	c_{11}	c_{12}	...	c_{1k}
X_2	c_{21}	c_{22}	...	c_{2k}
...	...	...	...	...
X_k	c_{k1}	c_{k2}	...	c_{kk}

Fonte: (FÁVERO E BELFIORE, 2017)

Ao se aplicar o critério da raiz latente, a quantidade de fatores extraídos será menor que k, o que faz com que nem toda a variância da variável X_j seja vista no novo sistema obtido, já que uma parte desta será "perdida" com os fatores descartados na priorização. A parte da variância "retida" entre os fatores selecionados é chamada de comunalidade, obtida mediante a Eq. (2.20) (MANLY, 2008).

$$c_{k1}^2 + c_{k2}^2 + \cdots = comunalidade\ X_k \tag{2.20}$$

A comunalidade tem, como principal objetivo, avaliar a inclusão de determinada variável na análise fatorial, tendo em vista que valores relativamente baixos para este parâmetro podem indicar que a variável em questão não está sendo explicada por nenhum dos novos fatores extraídos. Além disso, a somatória dos quadrados das cargas fatoriais, fixando-se os fatores, será igual ao quadrado do autovalor λ_k relacionado, devido justamente ao conceito do percentual de variância compartilhada, sendo mostrada na Eq. (2.21) (FÁVERO e BELFIORE, 2017).

$$c_{1k}^2 + c_{2k}^2 + \cdots + c_{kk}^2 = \lambda_k^2 \tag{2.21}$$

2.2.5. Rotação de Fatores

De forma geral, é desejável que as cargas fatoriais sejam próximas de zero ou muito diferentes de zero, com o intuito de simplificar a interpretação acerca da constituição dos fatores

obtidos. Isso se deve ao fato de que, para cargas próximas de zero, diz-se que a variável X_j se relaciona fracamente com um determinado fator F, enquanto um valor próximo da unidade, em módulo, significa que esta mesma variável relaciona-se fortemente com F (MANLY, 2008).

O processo de redistribuição das cargas entre os fatores é chamado de rotação, tendo por objetivo obter a situação descrita acima. Portanto, o sistema passa a apresentar um novo conjunto de cargas c', juntamente com novos autovalores λ' e, consequentemente, *escores* rotacionados s' diferentes dos originais. Como a quantidade de fatores extraídos não muda, a comunalidade para cada variável X_j permanece a mesma, assim como os testes de adequação KMO e Bartlett, já que a matriz de correlações ρ não sofre alterações (FÁVERO e BELFIORE, 2017).

Um dos métodos de rotação mais utilizados é o Varimax, desenvolvido por Kaiser (1958), sendo do tipo ortogonal (fatores rotacionados não têm correlação entre si). O objetivo deste procedimento é, como o nome diz, maximizar a variância compartilhada em fatores com autovalores relativamente baixos, podendo ser aplicado com ou sem a normalização das cargas, embora a primeira opção seja preferida (MANLY, 2008).

3. METODOLOGIA

Os métodos utilizados na construção deste trabalho, através dos estudos, referências e considerações realizadas, estão descritos no presente capítulo. São apresentados e detalhados a definição das variáveis e o levantamento dos dados, além da definição dos recortes temporais selecionados para estratificação da análise e avaliação da evolução da COVID-19 no Brasil ao longo das semanas epidemiológicas.

3.1. DEFINIÇÃO DAS VARIÁVEIS E LEVANTAMENTO DE DADOS

Com o intuito de construir um índice do panorama geral da COVID-19 por cada uma das unidades da federação (UF) no Brasil, relacionadas aos 26 estados mais o Distrito Federal, foram levantadas variáveis capazes de fornecer informações em diversos níveis e contextos, sendo representadas na Tabela 3. Tais variáveis são consideradas relevantes para a análise da COVID-19 no país, muitas citadas também em literatura publicada, como nos trabalhos de: Bezerra *et al.* (2020); Guimarães, Eleutério e Monteiro-da-Silva (2020) e Nascimento (2020*a*).

Tabela 3. Descrição das variáveis X_j em termos de categorias, características ou variáveis da base de dados utilizada e as referidas atualizações. A última variável, X_{18} (relacionada às doses aplicadas pelas vacinas), foi contabilizada a partir da SE 56 (janeiro/21).

Variáveis	Descrição	Categoria	Base de dados	Periodicidade de atualização	Última atualização
X_1	Quantidade de novos casos	COVID-19	Ministério da Saúde	Semanal	2021
X_2	Quantidade de novos casos por 100 mil habitantes	COVID-19	Ministério da Saúde	Semanal	2021
X_3	Quantidade de novos óbitos	COVID-19	Ministério da Saúde	Semanal	2021
X_4	Quantidade de novos óbitos por 100 mil habitantes	COVID-19	Ministério da Saúde	Semanal	2021
X_5	População	Demografia	TCU	Fixo	2020
X_6	Densidade populacional	Demografia	IBGE	Fixo	2020

Tabela 3. (Continuação)

Variáveis	Descrição	Categoria	Base de dados	Periodicidade de atualização	Última atualização
X_7	Quantidade de leitos de UTI (SUS)	Infraestrutura de Saúde	Data SUS	Mensal	2021
X_8	Quantidade de leitos de UTI (Não SUS)	Infraestrutura de Saúde	Data SUS	Mensal	2021
X_9	Quantidade de leitos de UTI COVID-19	Infraestrutura de Saúde	Data SUS	Mensal	2021
X_{10}	% Dependência do SUS	Social	ANS	Mensal	2021
X_{11}	Quantidade de profissionais da saúde	Infraestrutura de Saúde	Data SUS	Mensal	2021
X_{12}	Profissionais da saúde por 100 mil habitantes	Infraestrutura de Saúde	Data SUS	Mensal	2021
X_{13}	Número médio de moradores por domicílio	Social	IBGE	Fixo	2019
X_{14}	% pessoas com 60 anos ou mais	Fatores de risco	IBGE	Fixo	2019
X_{15}	% pessoas com 18 anos ou mais diagnosticadas	Fatores de risco	IBGE	Fixo	2019
X_{16}	% pessoas com 18 anos ou mais diagnosticadas	Fatores de risco	IBGE	Fixo	2019

Tabela 3. (Continuação)

Variáveis	Descrição	Categoria	Base de dados	Periodicidade de atualização	Última atualização
X_{17}	Índice de Desenvolvimento Humano	Social	IPEA	Fixo	2017
X_{18}	Doses aplicadas das vacinas	COVID-19	Ministério da Saúde	Semanal	2021

Todas as variáveis apresentadas na Tabela 3 foram construídas a partir de bases de dados oficiais de diferentes órgãos e entidades governamentais brasileiros, sendo: Ministério da Saúde (X_1 a X_4, além de X_{18}), Tribunal de Contas da União (TCU, X_5), Instituto Brasileiro de Geografia e Estatística (IBGE, X_6 e X_{13} a X_{16}), DATASUS (X_7 a X_9, além de X_{11} e X_{12}), Agência Nacional de Saúde Suplementar (ANS, X_{10}) e Instituto de Pesquisa Econômica Aplicada (IPEA, X_{17}). Os portais oficiais utilizados, com os respectivos endereços eletrônicos, assim como a base de dados compilada, podem ser vistos no Apêndice A do presente trabalho.

As variáveis "Densidade populacional" (X_6) e "Porcentagem (%) Dependência SUS" (X_{10}) foram calculadas conforme as Eqs. (3.1), relacionando população à área (hab/km²), e (3.2), relacionando a quantidade de pessoas com plano de saúde privado ("QTD", extraído da base da ANS) à população (%)**.**

$$X_6 = \frac{População}{Área} = \frac{X_5}{Área} \tag{3.1}$$

$$X_{10} = 1 - \frac{Qtd.pessoas\ com\ plano\ de\ saúde\ privado\ (QTD)}{População} = 1 - \frac{QTD}{X_5} \tag{3.2}$$

Além disso, para os parâmetros relacionados à estrutura de saúde de cada estado, foram feitas as seguintes considerações:

- X_7 e X_8: Leitos de UTI Adulto Tipos 1, 2 e 3, que se referem a leitos de unidades de terapia intensiva para pacientes com idade superior a 15 anos. Também foram considerados os leitos de UTI Pediátrica Tipos 1, 2 e 3, que se referem a leitos de unidades de terapia intensiva para pacientes com idade entre

29 dias e 15 anos e, por fim, leitos de UTI Neonatal, utilizados para o atendimento de pacientes recém-nascidos, com idade até 28 dias, em estado grave (MINISTÉRIO DA EDUCAÇÃO, 2016);

- X_{11}: Quantidade de médicos, enfermeiros, técnicos de enfermagem e auxiliares de enfermagem, de acordo com as bases extraídas do portal do DATASUS.

A análise efetuada pode ser aplicada a qualquer uma das semanas epidemiológicas entre março de 2020 e maio de 2021. Para simplificar o presente estudo, para as análises de agrupamentos foram escolhidas algumas SE de forma aleatória representando cada mês dentro do recorte temporal, utilizando-se a distribuição uniforme de probabilidade. Para a análise fatorial, a abordagem de estudo foi semelhante, tendo sido estratificadas 4 (quatro) SE específicas, como será detalhado nas seções seguintes. Ademais, a variável X_{18} (doses aplicadas de vacinas) teve a sua contabilização iniciada na SE 56, referente à terceira semana de janeiro/21, o que implica a ausência da mesma em relação à construção das análises referentes aos meses de 2020.

Em suma, a abordagem visou expandir a análise tradicional do panorama da COVID-19 nas UF, trazendo não apenas os números de casos e óbitos, mas também informações relacionadas a demais aspectos sociais, estruturais e econômicos de cada unidade federativa.

3.2. APLICAÇÃO DAS TÉCNICAS DE AGRUPAMENTO

3.2.1. Agrupamento hierárquico

Para a construção do dendrograma, aplicou-se o método de Ward, conforme explicitado na Seção 2.1.2, tendo sido utilizada a ferramenta computacional Python, através da biblioteca “SciPy” (SCIPY, 2021). O cálculo das distâncias (dissimilaridades), para o método mencionado, pode ser visto na Eq. (3.3), oriunda das Eqs. (2.3) e (2.4).

$$d(u,v) = \sqrt{\frac{n_v + n_s}{T} d(s,v)^2 + \frac{n_v + n_w}{T} d(v,w)^2 - \frac{n_v}{T} d(s,w)^2} \tag{3.3}$$

em que $d(u,v)$ representa a distância do grupo u (formado pelos grupos s e w) ao grupo v

ainda não agregado e T é a soma das cardinalidades n dos grupos exceto u.

Após a obtenção dos resultados, utilizou-se a biblioteca “Plotly” através da função “Choropleth” (MAPBOX, 2021), também do Python, para geração de gráficos com visualização geográfica acerca da composição dos agrupamentos obtidos com o método. A Figura 11 mostra, por meio de um fluxograma, a metodologia aplicada para coleta e interpretação de resultados, em relação à análise hierárquica.

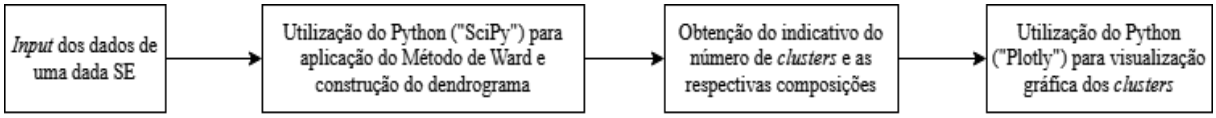

Figura 11. Fluxograma referente à metodologia de trabalho aplicada para obtenção de resultados da análise hierárquica, com auxílio de bibliotecas do Python (“SciPy” e “Plotly”).

3.2.2. Agrupamento não-hierárquico

Como explicitado na Seção 2.1.3, a técnica de *K-Médias* foi aplicada para o agrupamento não-hierárquico. Para isso, a biblioteca “Sklearn” foi aplicada através da função “KMeans” (SKLEARN, 2021), tendo sido escolhido o método de inicialização aleatório para um dado *random state* declarado arbitrariamente (neste caso, escolheu-se o estado 42).

Em seguida, as técnicas de *elbow method* e *silhouette score* foram aplicadas, com o objetivo de obter um indicativo de uma quantidade plausível de grupos, ou seja, o valor de *K*. Para o presente trabalho, a técnica do *elbow method* foi escolhida como base para utilização do valor de *K* para o algoritmo de *K-Médias*, como será discutido na Seção 4.2. Portanto, os resultados obtidos para o *silhouette score* foram aplicados como um viés comparativo em relação ao *elbow method*, não tendo prosseguido para visualização de composição dos *clusters* e obtenção do resultado final.

Por fim, aplicou-se a função “Choropleth”, assim como feito para o agrupamento hierárquico, com o intuito de visualizar as características geográficas de cada grupo. A Figura 12 resume os processos supracitados.

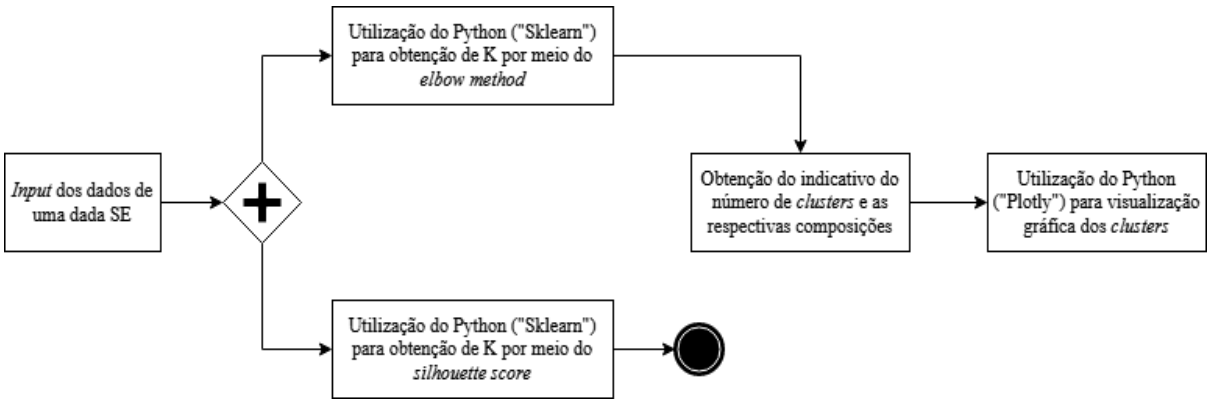

Figura 12. Fluxograma referente à metodologia de trabalho aplicada para obtenção de resultados da análise não hierárquica, com auxílio de bibliotecas do Python (“Sklearn” e “Plotly”).

3.3. APLICAÇÃO DA ANÁLISE FATORIAL

Para a aplicação da análise fatorial, o *software* IBM SPSS ® foi utilizado. De início, foi feita a padronização das variáveis (como foi mostrado na Seção 2.1.1 por meio da Eq. (2.1)) e, em seguida, realizou-se os testes de adequabilidade (KMO e esfericidade de Bartlett) para todas as SE analisadas.

Após os cálculos, abordados na Seção 2.2, terem sido computados pelo programa, obteve-se os fatores para cada UF atrelados aos seus respectivos autovalores, autovetores e cargas fatoriais. Com esses dados, tornou-se possível realizar o ordenamento (*ranking*) das UF com base na construção do IC19. Para isso, utilizou-se como critério a soma dos fatores rotacionados, ponderados pelos respectivos autovalores extraídos segundo o critério da raiz latente (vide Seção 2.2.3), tendo em vista que estes indicam a proporcionalidade da variância explicada (FÁVERO e BELFIORE, 2017). A Eq. (3.3) indica este conceito para a construção do IC19, objetivo do presente trabalho.

$$IC19_j = \sum_{i=1}^{n} \%VAR_k \times F_{kj} \qquad (3.4)$$

em que $IC19_j$ é o índice COVID-19 para uma determinada UF j, F_k é o fator k extraído de acordo com o método da raiz latente e $\%VAR_k$ é o percentual da variância do sistema explicada pelo fator k.

Por fim, para que fosse possível realizar a interpretação do resultado e a posterior comparação entre as UF de maneira intuitiva, aplicou-se a padronização de mínimos e máximos, como mostra a Eq. (3.4). Este artifício transforma um conjunto de medidas limitando-

as no intervalo de 0 a 1, sendo 0 o menor valor e 1 o maior.

$$Z_{IC19j} = \frac{IC19_j - \min(IC19)}{\max(IC19) - \min(IC19)} \quad (3.5)$$

em que Z_{IC19j} é a medida do índice padronizada e $\min(IC19)$ e $\max(IC19)$ são os valores mínimos e máximos do índice entre todas as UF, em uma dada semana epidemiológica, respectivamente.

Ao final, assim como nas técnicas de agrupamento, foi utilizado o "Choropleth" para visualização gráfica do IC19 em cada UF do Brasil, nas SE selecionadas para estudo. O fluxo para aplicação da técnica pode ser visto na Figura 13.

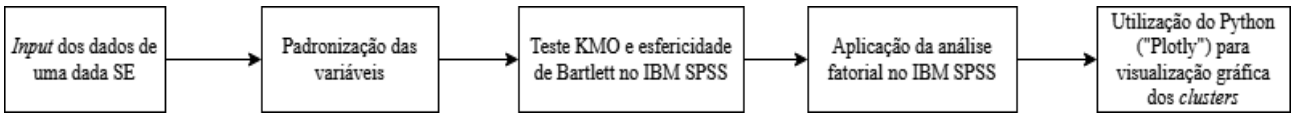

Figura 13. Fluxograma referente à metodologia de trabalho aplicada para obtenção de resultados da análise fatorial, com auxílio do *software* IBM SPSS.

4. RESULTADOS E DISCUSSÕES

Nos itens a seguir são apresentados e discutidos os resultados vinculados às análises multivariadas demonstradas no capítulo anterior: agrupamentos hierárquico e não hierárquico, além da análise fatorial, vinculados aos parâmetros da Tabela 3. Observou-se agrupamentos espaço-temporais, vinculados à pandemia, distribuídos entre todos os estados brasileiros entre os anos 2020 e 2021.

4.1. AGRUPAMENTO HIERÁRQUICO

Para avaliação da semelhança entre as UF em relação às variáveis mostradas na Tabela 3, construiu-se inicialmente a análise de agrupamento hierárquico através do dendrograma para cada mês, considerando 17 ou 18 variáveis (dependendo da SE). Como mencionado na Seção 3.1, SE aleatórias foram escolhidas para representar cada mês, embora, como também dito anteriormente, seja possível realizar a análise semana a semana, ou ainda uma análise da média mensal. É importante salientar que também foi possível construir dendrogramas e análises utilizando a média das SE constituintes de cada mês, como mostrado na Figura 14 e na Tabela 4. A primeira mostra o dendrograma obtido para alguns meses, refletindo os estágios da pandemia no Brasil e no mundo. Já a segunda, utilizando-se da interpretação das distâncias euclidianas e dos respectivos saltos após cada aglomeração, conforme visto na Seção 2.1.2, mostra o número de agrupamentos indicado pela técnica.

A mesma Tabela 4 mostrou que a análise da média no mês, quando comparada a uma semana especifica daquele mês, usando as mesmas técnicas e condições (análise hierárquica), apresentou algumas diferenças no número de *clusters*. De fato, houve alguma coincidência nos números de *clusters* da análise hierárquica (10 em 15 situações). Isto mostra algo da sensibilidade das análises frente ao uso de valores médios, ilustrado quando se comparam as Figuras 14 e 16 e 15 e 17.

Um mapa de calor também foi construído para a representação dos *clusters* obtidos na análise hierárquica, conforme ilustrado na Tabela 5. Tais mapas permitem mostrar a variabilidade dos agrupamentos em cada semana epidemiológica, conforme indicado. Em linhas gerais, os resultados mostraram que alguns estados migraram de um determinado agrupamento para outro em função do tempo e das condições pandêmicas, representadas pelos valores dos parâmetros (ou variáveis). Com a alocação das UF nos grupos apontados pela análise hierárquica, tornou-se possível visualizar a distribuição espaço-temporal no território brasileiro para alguns meses

selecionados, evidenciado através da Figura 15.

Tabela 4. Número de *clusters* indicado pela análise hierárquica, para cada período analisado, a saber: uma SE especifica ou ainda sobre a média de dados daquele mês.

SE	Período (mês/ano)	Número de *clusters* para SE	Número de *clusters* para a média mensal
13	Março/2020	5	5
18	Abril/2020	5	5
22	Maio/2020	5	5
25	Junho/2020	5	3
30	Julho/2020	3	3
32	Agosto/2020	5	3
36	Setembro/2020	3	3
41	Outubro/2020	3	3
45	Novembro/2020	3	3
53	Dezembro/2020	3	3
56	Janeiro/2021	4	3
61	Fevereiro/2021	3	3
64	Março/2021	3	3
67	Abril/2021	4	5
74	Maio/2021	5	3

Um mapa de calor também foi construído para a representação dos *clusters* obtidos na análise hierárquica, conforme ilustrado na Tabela 5. Tais mapas permitem mostrar a variabilidade dos agrupamentos em cada semana epidemiológica, conforme indicado. Em linhas gerais, os resultados mostraram que alguns estados migraram de um determinado agrupamento para outro em função do tempo e das condições pandêmicas, representadas pelos valores dos parâmetros (ou variáveis). Com a alocação das UF nos grupos apontados pela análise hierárquica, tornou-se possível visualizar a distribuição espaço-temporal no território brasileiro para alguns meses selecionados, evidenciado através da Figura 15.

Ademais, realizou-se um estudo específico para algumas semanas epidemiológicas, selecionadas com base em momentos relevantes da pandemia: a SE 30 (julho/20) representa o primeiro pico em relação a número de casos (320.155); a SE 45 (novembro/20) diz respeito à melhora significativa em relação aos números de casos e de óbitos quando comparado ao pico

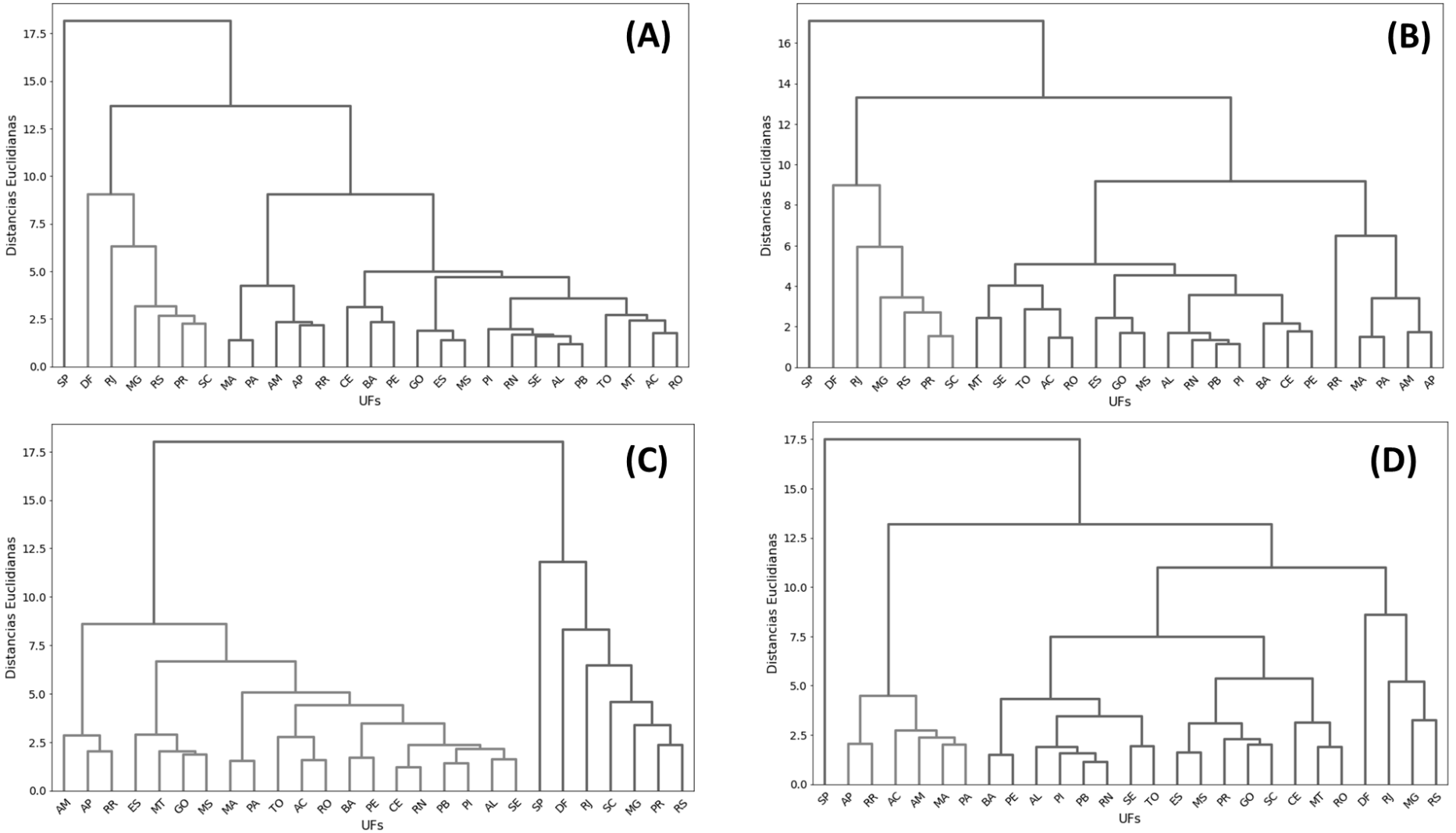

Figura 14. Dendrogramas obtidos por meio da análise hierárquica, construídos a partir dos dados referentes as médias dos meses: (A) março/2020, (B) julho/2020, (C) novembro/2020 e (D) abril/2021. Tais resultados ilustram a formação de três a cinco agrupamentos e estão relacionadas às distribuições espaço-temporais da Figura 15.

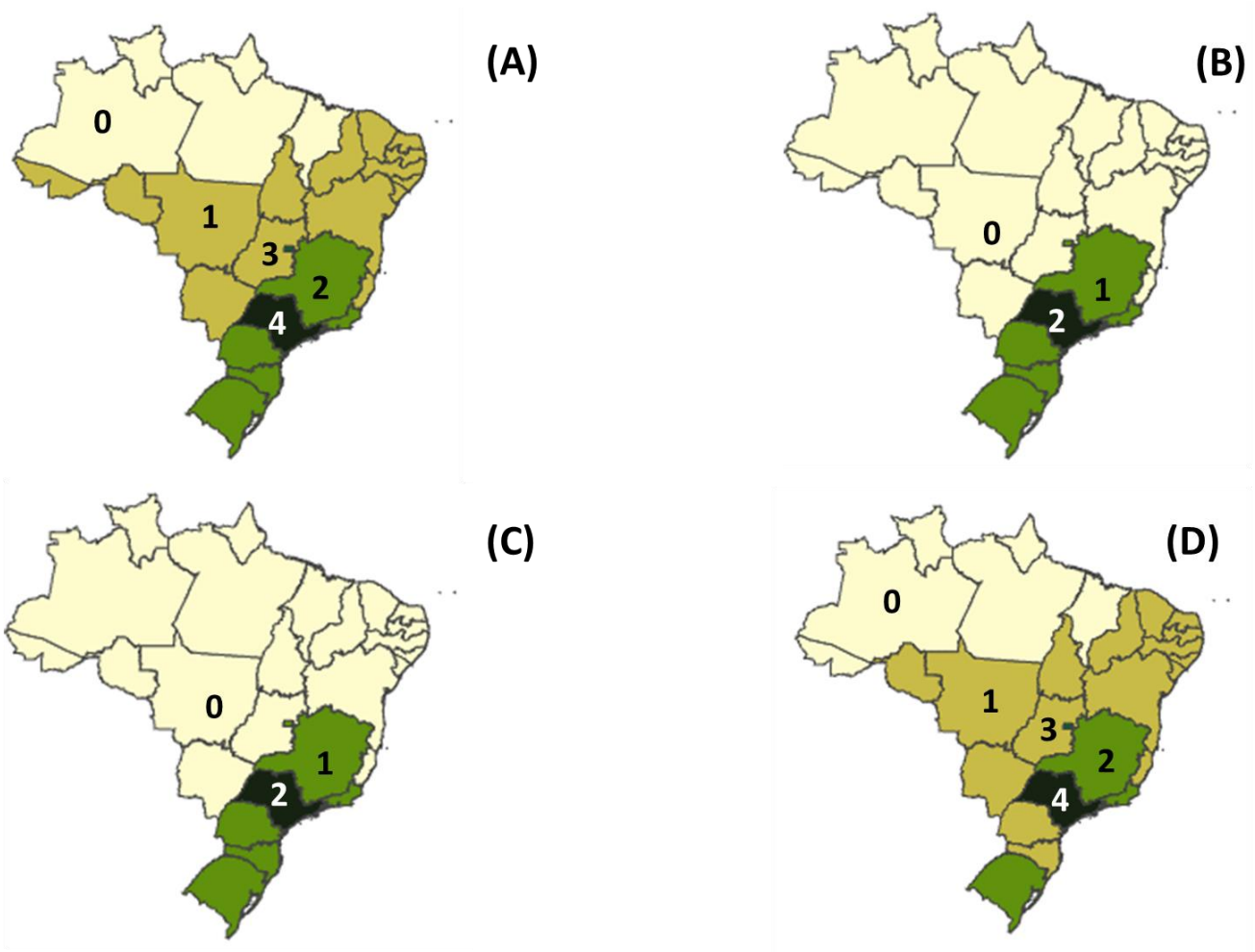

Figura 15. Distribuição espaço-temporal dos *clusters* 0, 1 e 2 obtidos das médias dos meses de (A) março/2020, (B) julho/2020, (C) novembro/2020 e (D) abril/2021 utilizando da técnica de análise hierárquica. Em (A) e (B), embora com diferentes parâmetros, os agrupamentos resultaram semelhantes, com destaque para SP, que permaneceu em *cluster* único, separada das demais UF.

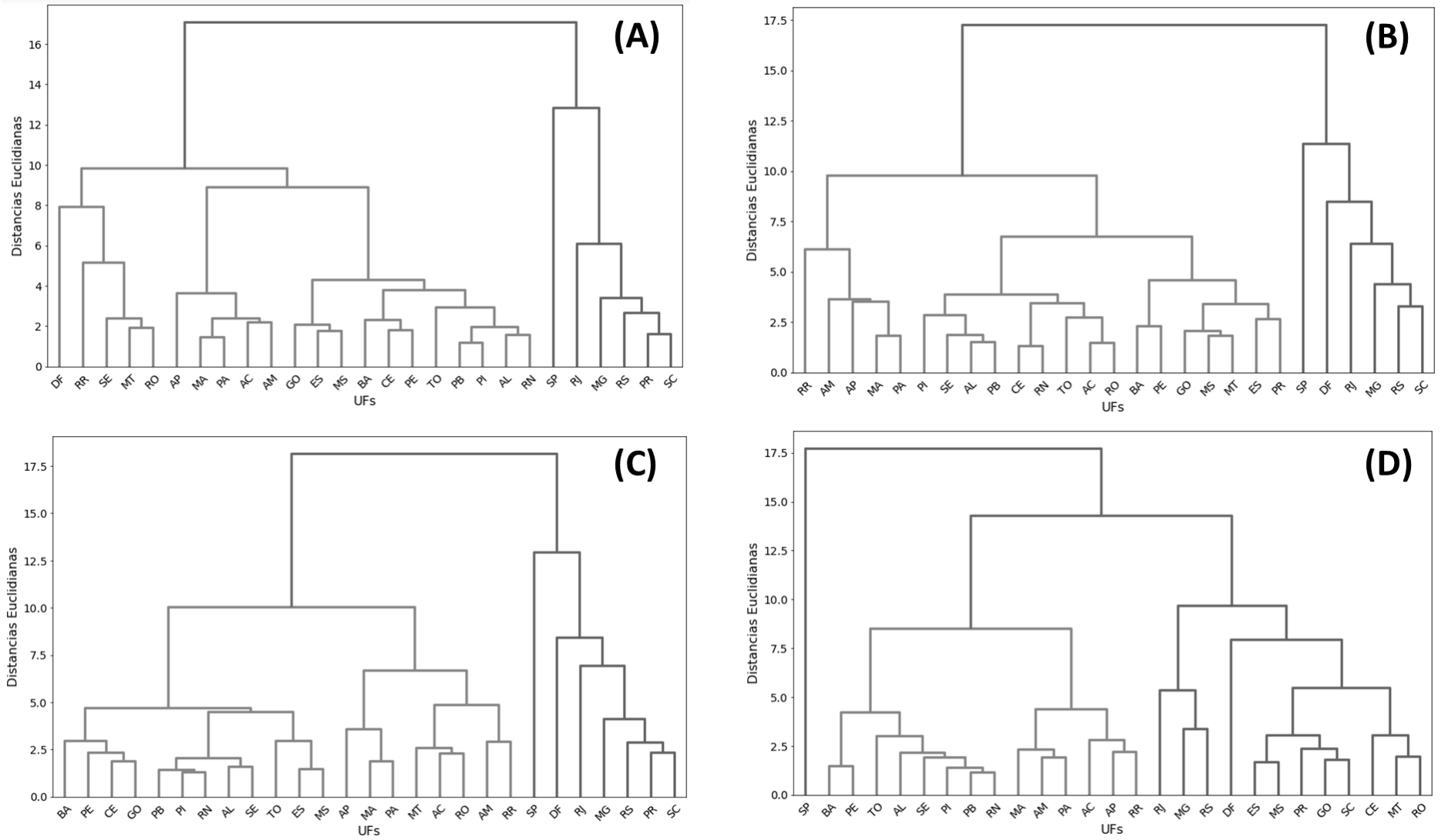

Figura 16. Dendrogramas obtidos por meio da análise hierárquica, construídos a partir dos dados referentes a: (A) SE 30, (B) SE 45, (C) SE 61 e (D) SE 67. Tais resultados ilustram a formação de três a cinco agrupamentos. Tais dendrogramas estão relacionadas às distribuições espaço-temporais da Figura 17.

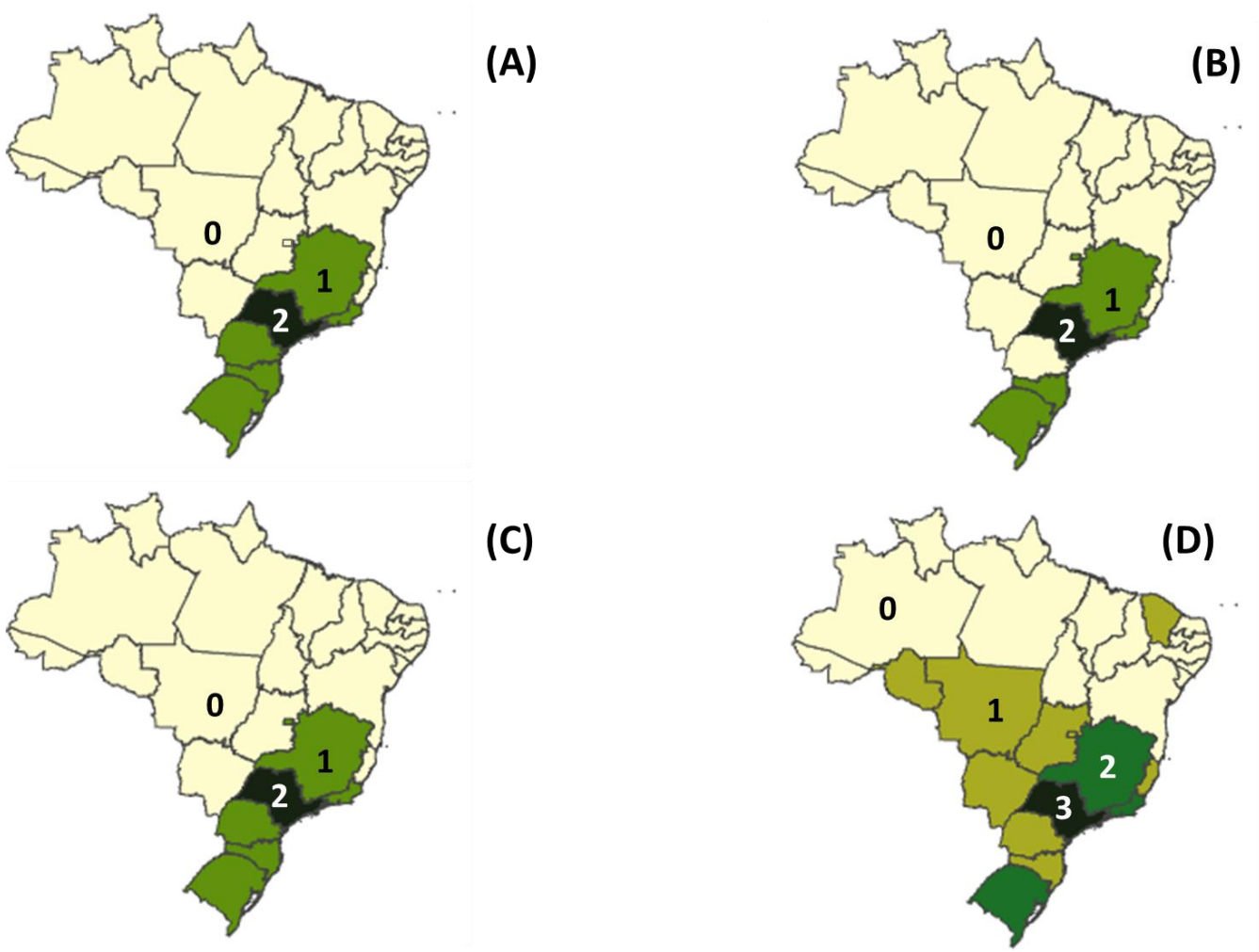

Figura 17. Distribuição espaço-temporal dos *clusters* 0 a 3 obtidos nas semanas (A) SE 30, (B) SE 45, (C) SE 61 e (D) SE 67 utilizando da técnica de análise hierárquica. Em (A) e (C), embora com diferentes parâmetros, os agrupamentos resultaram semelhantes, com destaque para SP, que permaneceu em *cluster* único, separada das demais UF. Tais distribuições estão de acordo com os dendrogramas da Figura 16.

Tabela 5. Identificação numérica do *cluster* (entre 0 e 4) de cada UF ao longo de algumas SE, em forma de mapa de calor, utilizando da técnica de análise hierárquica. Os resultados mostraram que alguns estados migraram de um determinado agrupamento para outro em função do tempo e das condições pandêmicas, representadas pelos valores dos parâmetros (ou variáveis).

UF	SE 13	SE 18	SE 22	SE 25	SE 30	SE 32	SE 36	SE 41	SE 45	SE 53	SE 56	SE 61	SE 64	SE 67	SE 74
AC	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
AL	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
AM	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
AP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BA	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
CE	1	0	0	1	0	1	0	0	0	0	0	0	0	1	1
DF	3	3	3	3	0	3	1	1	1	1	2	1	1	1	3
ES	1	2	2	1	0	1	0	0	0	1	2	0	0	1	1
GO	1	1	2	1	0	1	0	0	0	0	0	0	0	1	1
MA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MG	2	2	2	2	1	2	1	1	1	1	2	1	1	2	2
MS	1	1	2	1	0	1	0	0	0	1	2	0	0	1	1
MT	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
PA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PB	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
PE	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
PI	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
PR	2	2	2	2	1	2	1	1	0	1	2	1	1	1	2
RJ	2	2	4	3	1	2	1	1	1	1	2	1	1	2	2
RN	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
RO	1	0	1	1	0	0	0	0	0	0	0	0	0	1	1
RR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RS	2	2	2	2	1	2	1	1	1	1	2	1	1	2	2
SC	2	2	2	2	1	1	0	1	1	1	2	1	1	1	1
SE	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
SP	4	4	4	4	2	4	2	2	2	2	3	2	2	3	4
TO	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1

da primeira onda; a SE 61 (fevereiro/21) marcou aproximadamente 10% das doses das vacinas aplicadas em relação à data limite considerada no presente trabalho (até maio/21) e, por fim, a SE 67 (abril/21), que representa a pior semana epidemiológica em número de óbitos desde o início da série (21.173), além de representar a data em que cerca de 50% do total das doses das vacinas até maio/21 haviam sido aplicadas. Os dendrogramas e a distribuição territorial dos grupos obtidos podem ser vistos nas Figuras 16 e 17, respectivamente.

Apenas esta escolha de procedimento poderia servir de ferramenta de auxílio à gestão

epidemiológica, separando os estados por agrupamentos semana a semana visando à aplicação de recursos e insumos pelos governos estaduais e federal prioritariamente nas UF pertencentes aos *clusters* mais críticos, com o objetivo de reduzir os impactos da pandemia nas regiões mais necessitadas.

De acordo com os resultados encontrados na análise hierárquica e resumidos na Tabela 4, o dendrograma para os meses da série histórica indicou a presença de 3 a 5 agrupamentos. Com exceção da SE 22 (maio/20), em que RJ e SP ocuparam juntas o *cluster* 4, em todas as outras semanas epidemiológicas SP esteve em um grupo único.

A Figura 14 mostrou, através do dendrograma para diferentes médias de alguns dos meses, que a ordem de agrupamento das UF tem forte relação com o número de pessoas vivendo nas mesmas (isto é, quão populosa a UF é), com o impacto em relação a número de casos e de óbitos e com o grau de desenvolvimento social e infraestrutura em saúde. Isto pode ser visto, primeiramente, com o fato de o estado de São Paulo ter sido alocado em *cluster* único para praticamente todos os períodos, demonstrando a dissimilaridade e a preponderância que a mesma detém em relação ao resto do país na temática da COVID-19. A Figura 16, que mostrou o detalhamento em relação às 4 SE mencionadas anteriormente e que se referem a semanas vinculadas aos mesmos meses da pandemia, quando comparadas à Figura 14, ratifica a afirmação.

Esta singularidade pode ser corroborada com o fato de SP ser o estado mais populoso do Brasil, abrigando mais de 46 milhões de habitantes (cerca de 21,9% da população nacional) e ter melhor infraestrutura de saúde quando comparado às demais UF. E, especificamente em relação à capital paulista, afirma-se que a disseminação inicial de praticamente 80% do coronavírus em território nacional adveio da mesma (NICOLELIS *et al*, 2020), devido à elevada quantidade de residentes e, também, de representar um *hub* econômico do país e de toda a América do Sul. A contribuição inicial do estado de São Paulo para a pandemia pode ser resumida ao fato de a taxa de contaminação nunca ter reduzido abaixo de 30% nos três primeiros meses de 2020, tornando-a a principal cidade brasileira “super-propagadora” da epidemia de SARS-CoV-2, também chamada de *hotspot* pandêmico (ALCÂNTARA *et al.*, 2020). Tal situação explica, também, a rápida disseminação da doença ao redor do mundo por meio de aeroportos internacionais, mas a relevância deste fato deveria ter sido ao menos considerada pelas autoridades sanitárias brasileiras visando mitigar a propagação da doença em território nacional (NICOLELIS *et al*. 2020).

Ademais, nota-se o agrupamento, na análise hierárquica, de UF das regiões Sul, Sudeste e Centro-Oeste, notadamente Rio de Janeiro (RJ), Minas Gerais (MG), Rio Grande do

Sul (RS), Paraná (PR) e Santa Catarina (SC), além do Distrito Federal (DF). Por fim, nos últimos *clusters*, unem-se as UF das regiões Norte e Nordeste. Os resultados encontrados têm relação com outros trabalhos encontrados em literatura, como visto em Ferraz *et al.* (2021), que calcularam um índice de estrutura de saúde relacionado à COVID-19 para 543 microrregiões do país e concluíram que 60% das regiões com os 20 melhores índices de estrutura situavam-se nas regiões Sul, Sudeste ou Centro-Oeste. Além disso, os autores observaram que, das regiões em situação considerada inadequada para enfrentar a pandemia, 60% se localizavam no Norte ou no Nordeste do Brasil. A distribuição geográfica dos grupos encontrados na análise hierárquica, por mês e por semana epidemiológica, foi mostrada nas Figuras 15 e 17, respectivamente.

A Figura 16 exibe o dendrograma para semanas epidemiológicas específicas, como foi explicitado anteriormente. É perceptível que as semanas SE 30 e SE 61 (Fig. 16(A) e Fig.16(C)) resultaram em construções semelhantes, apesar de a segunda contabilizar o parâmetro de doses de vacinas aplicadas, que teve início na SE 56. Apesar de apresentarem macroestrutura de grupos parecida, é possível verificar algumas particularidades, como para DF, que na SE 61 foi alocada no *cluster* 1, fato que pode ser relacionado com a taxa relativamente alta de vacinação referente à sua população. Até a SE 61, havia sido aplicada, nesta UF, uma quantidade de vacinas equivalente à cerca de 4,5% de sua população (136.408 doses para a população de 3.055.149 habitantes), o que representava a 4ª maior taxa do país no período. Outro quesito se refere aos estados do Amazonas (AM) e de Roraima (RR), que detinham cerca de 6,0% e 4,8% de doses aplicadas relativas às suas populações, respectivamente, sendo a maior e a 3ª maior do país até aquela semana. Na estrutura do dendrograma, é possível perceber que, para a SE 30, AM apresentava alta similaridade com o Acre (AC), sendo posteriormente agrupada com RR na SE 61, apesar de ambas terem permanecido no *cluster* 0. A Fig. 16(B), referente à SE 45, mostra uma estratificação semelhante àquela vista na SE 61, com a diferença de que o estado do Paraná foi alocado no *cluster* 0.

Já na Fig. 16(D), que mostra a SE 67, relacionada ao pior quadro da pandemia para todo o período considerado no trabalho, são mostrados agrupamentos significativamente distintos quando comparados às outras três SE elencadas. É possível perceber, por exemplo, o fato de o estado do Ceará (CE) ter sido alocado no *cluster* 1, separado de todas as outras UF do Norte e do Nordeste. Analisando as variáveis do CE para aquela semana, percebeu-se o alto impacto da COVID-19 e a pressão exercida no sistema de saúde do estado quando comparado às demais UF da região: foram 36.185 casos diagnosticados da doença (X_1), com 1.093 óbitos (X_3), que representaram 393,87 casos por 100.000 habitantes (X_2) e 11,90 óbitos por 100.000

habitantes (X_4). À título de comparação, o valor visto para X_1 representou, apenas para o Ceará, 32,8% de todos os casos da região Nordeste na SE 67, enquanto X_3 foi equivalente a 29,3% de todos os óbitos, sendo deveras relevante ao se levar em consideração que a população cearense representa 16% de toda a população nordestina. Acerca desta região, Souza, Marques e Amorim (2020), conduziram estudo de análise hierárquica em relação à vulnerabilidade e à incidência da COVID-19, avaliando dados de densidade populacional, fatores de risco e taxa de urbanização, tendo identificado quatro *clusters*, com os grupos de elevada criticidade localizados em regiões litorâneas e/ou que apresentem alta densidade populacional e elevado percentual de fatores de risco entre os habitantes, como percentual de idosos (60 anos ou mais), incidência de asma, AIDS, tuberculose e afins. No presente trabalho, a região Nordeste foi dividida em 2 grupos pela estrutura do dendrograma, podendo ser associado à quantidade mais elevada de variáveis, que trouxeram outras categorias para a análise, como infraestrutura em saúde, IDHM, doses aplicadas das vacinas, etc.

Já Castro *et al.* (2021), por meio da avaliação de 5 variáveis distintas das 18 descritas no presente trabalho, encontraram a existência de seis *clusters*, com 2 grupos representando os mais críticos em termos de número de mortes por 100 mil habitantes por COVID-19, sendo um formado pelos estados de Pernambuco (PE), Espírito Santo (ES), AP, SP e Sergipe (SE) e o outro por DF, CE, AM, RR e RJ. Bezerra *et al.* (2020), ao agrupar as UF em relação ao índice de infraestrutura em saúde, indicador proposto pelos autores através da aplicação da técnica de análise fatorial em relação a 21 variáveis elencadas, também encontrou seis agrupamentos. A existência de múltiplos resultados a partir de premissas e variáveis distintas demonstra a flexibilidade da técnica de agrupamento hierárquica e, também, ajuda a ratificar o quadro altamente heterogêneo da pandemia no Brasil, com cenários complexos e variáveis surgindo ao longo das semanas epidemiológicas, sendo influenciados por políticas públicas adotadas e, também, pela desigualdade socioeconômica crônica no país (CASTRO *et al.*, 2021; WERNECK e CARVALHO, 2020; SOUZA *et al.*, 2020a).

Quanto à Tabela 5, percebe-se uma outra maneira de enxergar a variabilidade ao usar apenas a análise hierárquica. Ao selecionar algumas semanas epidemiológicas, a saber: 13, 18, 22, 25, 30, 32, 36, 41, 45, 53, 56, 61, 64, 67 e 74, as alocações das UF nos grupos entre 0 e 4 não coincidiram. Isto se explica em parte pelas diferenças dos parâmetros em cada situação. Em particular, houve uma mudança entre *clusters* de semana para semana, considerando um mesmo estado, como por exemplo Acre, Alagoas, Amazonas, Bahia, Ceará, Mato Grosso, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, Rondônia, Sergipe e Tocantins (entre 0 e 1); São Paulo sempre variou entre os *clusters* 2, 3 e 4. Amapá, Maranhão, Pará e Rondônia foram os

únicos a não terem alteradas sua situação durante estas semanas.

4.2. AGRUPAMENTO NÃO-HIERÁRQUICO

Como visto na Seção 2.1.2, o resultado fornecido pela técnica de agrupamento hierárquico pode servir de base para a aplicação dos métodos não-hierárquicos, em especial o *K*-médias. Em adição a este quesito, o presente trabalho também utilizou as ferramentas de estimação do número de agrupamentos fornecidos pelo *elbow method* e pelo *silhouette score*, como foi visto na Seção 2.1.3, com o comparativo desses recursos sintetizado na Tabela 6. A Figura 18 compara os resultados das duas ferramentas para alguns meses selecionados.

Tabela 6. Determinação do número de *clusters* conforme as técnicas não-hierárquicas (*elbow method* e *silhouette score*) para cada período analisado, além das respectivas pontuações de silhueta. Pelo fato de serem critérios diversos, os resultados dos números de agrupamentos não foram similares. Ainda assim, o número de agrupamentos foi considerado pequeno (entre quatro e seis para o *elbow method* e apenas dois para o *silhouette*).

SE	**Período (mês/ano)**	**Número de *clusters elbow method* por SE**	**Número de *clusters silhouette score***	**Pontuação de Silhueta (PS)**
13	Março/2020	4	2	0,536
18	Abril/2020	5	2	0,456
22	Maio/2020	6	2	0,457
25	Junho/2020	5	2	0,429
30	Julho/2020	5	2	0,468
32	Agosto/2020	5	2	0,466
36	Setembro/2020	6	2	0,386
41	Outubro/2020	5	2	0,427
45	Novembro/2020	6	2	0,380
53	Dezembro/2020	5	2	0,384
56	Janeiro/2021	5	2	0,454
61	Fevereiro/2021	5	2	0,387
64	Março/2021	6	2	0,387
67	Abril/2021	5	2	0,446
74	Maio/2021	4	2	0,473

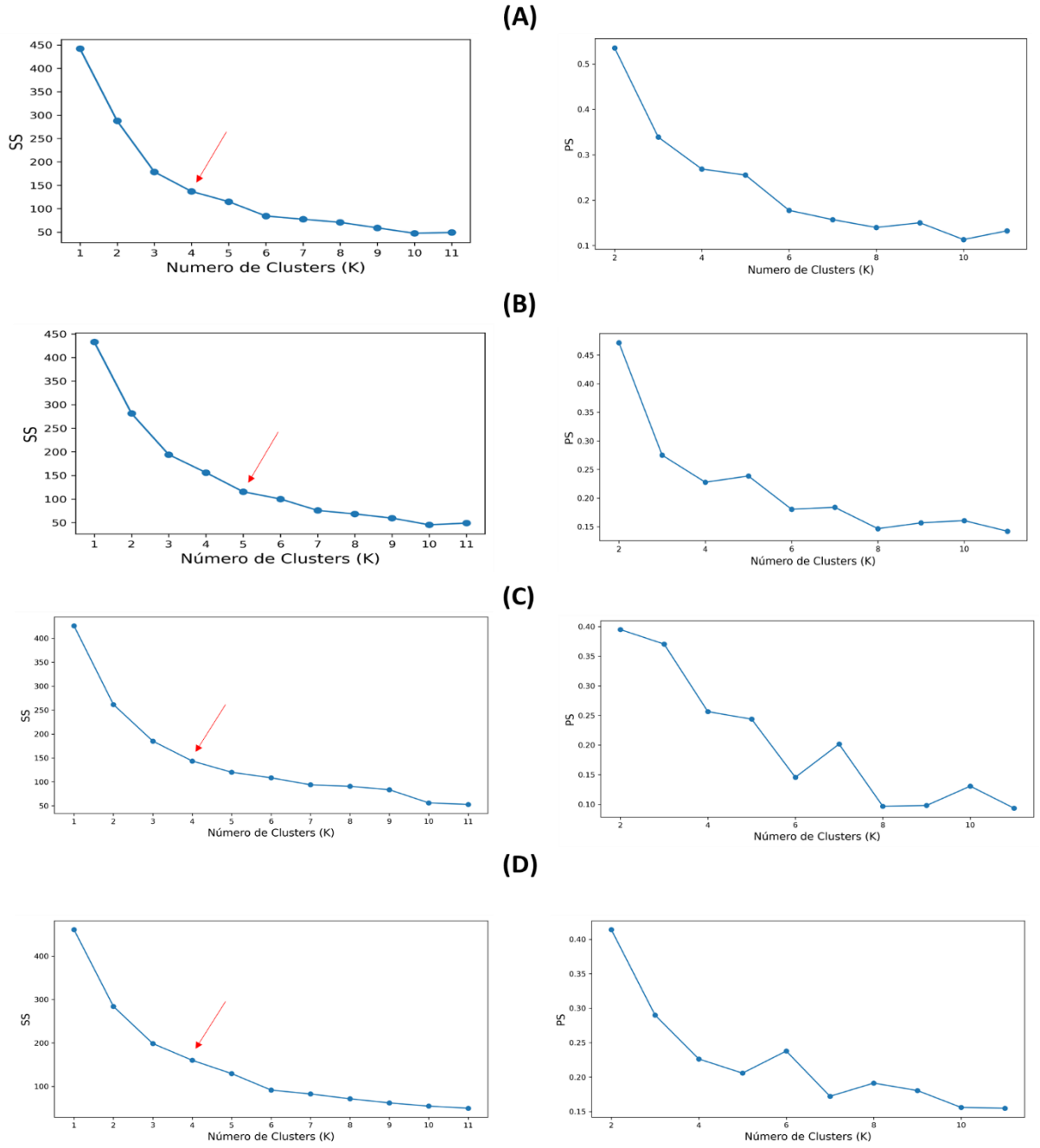

Figura 18. Ilustração da aplicação das técnicas *elbow method* (à esquerda) e o *silhouette score* (à direita) resultando no indicativo do número de *clusters* apenas para (A) março/2020, (B) julho/2020, (C) novembro/2020 e (D) abril/2021.

Tendo como *input* o número de grupos fornecido, conforme ilustração da Figura 18, e resultados apresentados na Tabela 6, foram considerados os resultados do *elbow method* na análise do algoritmo de *K*-médias para cada uma das semanas.

Ao longo de todo o período, a análise indicou a presença de 4 a 6 agrupamentos ao se utilizar o *elbow method*, a depender da SE escolhida. Percebe-se que, diferente da análise hierárquica, que mostrou quantidade de agrupamentos semelhante tanto para estratificação por

SE quanto por médias mensais, na análise não-hierárquica as médias resultaram de 3 a 5 grupos, identificando 4 *clusters* sequencialmente a partir de agosto/20 até o final da série. Apesar disso, a aplicabilidade da técnica pôde ser vista para as duas abordagens, ficando à critério do pesquisador a utilização daquela que mais se adeque aos objetivos do estudo. No presente trabalho, a escolha de interpretação e discussão residiu na estratificação por SE, como mencionado anteriormente.

É perceptível, também, a diferença significativa de indicação da quantidade de agrupamentos entre as técnicas do *elbow method* e do *silhouette score*. Para a segunda, apenas 2 grupos foram indicados para toda a série histórica. Como mencionado na Seção 2.1.3, esta ferramenta utiliza o conceito de Pontuação de Silhueta (PS), trazido por meio da Eq. (2.7), variando de 0 a 1, demonstrando melhor alocação dos grupos no sistema quando PS se aproximar de 1. Observou-se, para quase todas as semanas analisadas, que o valor máximo deste parâmetro esteve abaixo de 0,500 (única exceção vista para SE 13, com 0,536), o que possibilita interpretar a estruturação dos agrupamentos como fraca, sendo recomendável justamente a implementação de métodos adicionais na base de dados (KAUFMAN e ROUSSEEUW, 2005; INEKWE, MAHARAJ e BHATTACHARYA, 2020). Por esta razão, escolheu-se para o presente trabalho a interpretação dos agrupamentos obtidos de acordo com o *elbow method*, sendo possível observar a possibilidade de utilização de diferentes técnicas de estimação do número ótimo de grupos a serem implementados, capaz de serem aplicadas nos mais variados contextos dentro da comunidade científica.

O panorama geral dos *clusters* para cada SE pode ser visto na Tabela 7. Assim como na análise hierárquica, percebeu-se o papel central de SP no que diz respeito à COVID-19 no Brasil ao se realizar a análise não-hierárquica. Em todas as semanas epidemiológicas, o estado esteve em grupo único, seguido por RJ, MG, DF, RS e SC, que se alternaram entre os *clusters* mais elevados. Em contrapartida, os estados do Norte e do Nordeste ocuparam os agrupamentos mais baixos, porém de forma mais regionalizada quando comparado à análise hierárquica, já que o número de grupos indicados pela segunda técnica foi maior. Em geral, os estados do Norte estiveram no *cluster* 0, com exceção de Tocantins (TO) e Rondônia (RO). Já os estados do Nordeste apareceram entre os *clusters* 1 e 2, exceto o MA, que esteve no *cluster* 0 em quase todas as semanas.

Também foi possível analisar a distribuição geográfica dos *clusters* obtidos no estudo, como mostrado na Figura 19, representando os quatro meses previamente selecionados para exemplificação. Além da estratificação mensal, realizou-se a análise para as quatro semanas epidemiológicas selecionadas na Seção 4.1, cuja distribuição geográfica pode ser vista na

Figura 20. Para as SE 30, SE 61 e SE 67, obtiveram-se 5 agrupamentos de acordo com o *elbow method*, enquanto a SE 45 apresentou 6 grupos.

Tabela 7. Identificação numérica do *cluster* (entre 0 e 5) de cada UF ao longo das SE, em forma de mapa de calor, para a análise não-hierárquica. Assim como na análise hierárquica, os resultados mostraram que alguns estados migraram de um determinado agrupamento para outro em função do tempo e das condições pandêmicas, representadas pelos valores dos parâmetros (ou variáveis).

UF	SE 13	SE 18	SE 22	SE 25	SE 30	SE 32	SE 36	SE 41	SE 45	SE 53	SE 56	SE 61	SE 64	SE 67	SE 74
AC	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
AL	1	1	1	1	1	1	2	1	2	1	2	1	2	1	1
AM	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
AP	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
BA	1	1	1	1	1	1	2	1	4	1	2	1	2	1	1
CE	1	1	1	1	1	1	2	1	2	1	2	1	2	2	1
DF	2	3	3	3	2	2	3	2	3	2	3	2	4	3	2
ES	1	2	2	3	3	3	4	2	4	2	3	1	3	2	2
GO	1	2	2	2	1	1	4	2	4	1	2	1	3	2	1
MA	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
MG	1	2	2	2	3	3	4	3	4	3	3	3	3	4	2
MS	1	2	2	2	1	1	2	2	2	2	2	1	2	2	1
MT	1	1	1	1	2	1	2	2	2	1	2	1	1	2	1
PA	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
PB	1	1	1	1	1	1	2	1	2	1	2	1	2	1	1
PE	1	1	1	1	1	1	2	1	2	1	2	1	2	1	1
PI	1	1	1	1	1	1	2	1	2	1	2	1	2	1	1
PR	1	2	2	2	3	3	4	2	4	2	3	3	3	2	2
RJ	2	3	4	3	3	3	4	3	3	3	3	2	4	4	2
RN	1	1	1	1	1	1	2	1	2	1	2	1	2	1	1
RO	0	1	1	1	2	1	2	1	2	1	0	0	1	2	1
RR	0	0	0	0	2	2	0	0	1	0	1	0	1	0	0
RS	1	2	2	2	3	3	4	3	4	3	3	3	3	4	2
SC	1	2	2	2	3	3	4	2	4	2	3	3	3	2	2
SE	1	1	1	1	2	1	2	1	2	1	2	1	2	1	1
SP	3	4	5	4	4	4	5	4	5	4	4	4	5	5	3
TO	1	1	1	1	1	1	2	1	2	1	2	1	1	1	1

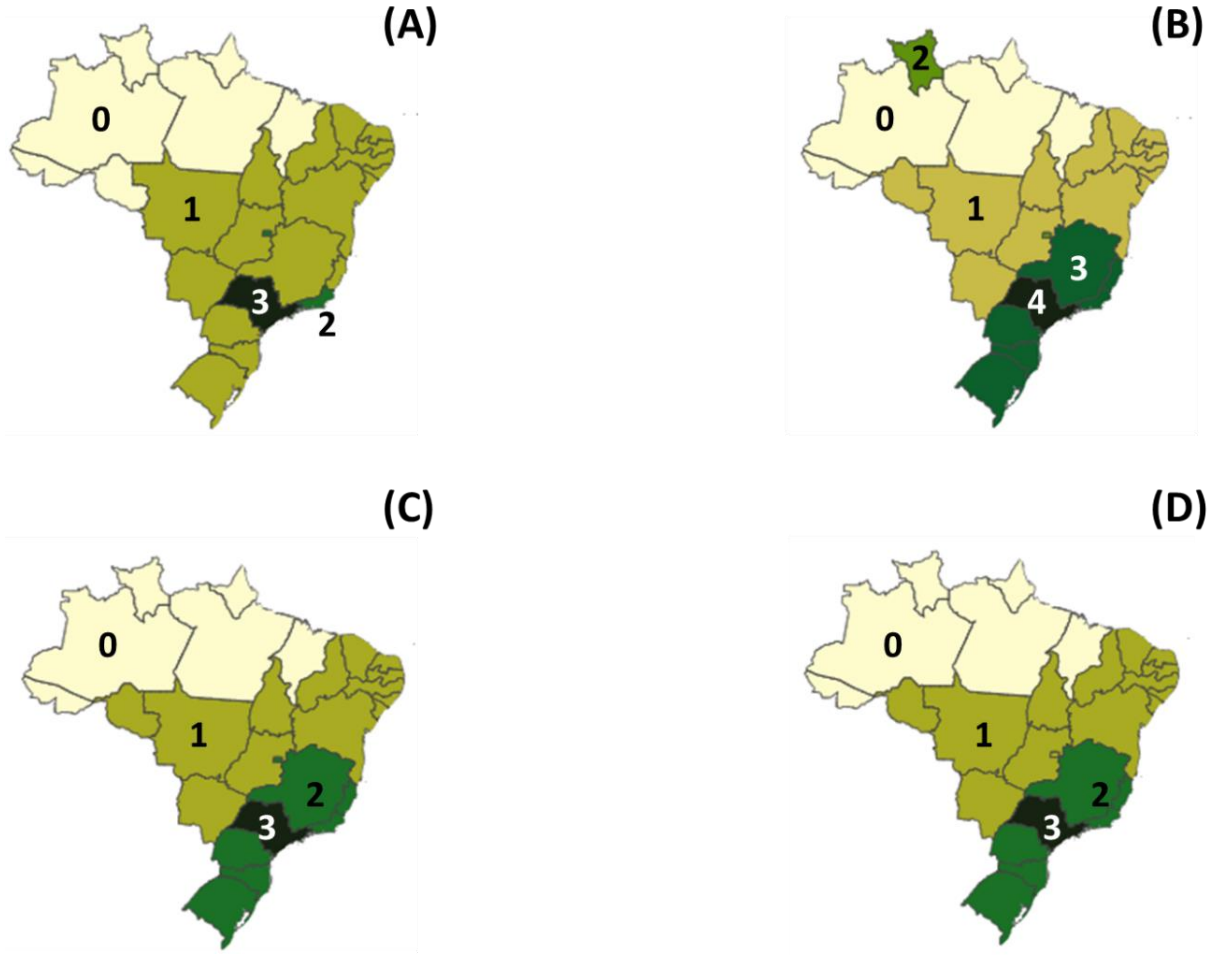

Figura 19. Distribuição espaço-temporal dos *clusters* entre 0 e 4 obtidos a partir das médias dos meses de (A) março/2020, (B) julho/2020, (C) novembro/2020 e (D) abril/2021, utilizando da técnica de análise não-hierárquica. Em (C) e (D), embora com diferentes parâmetros, os agrupamentos resultaram semelhantes, com destaque para SP, que permaneceu em *cluster* único, separada das demais UF.

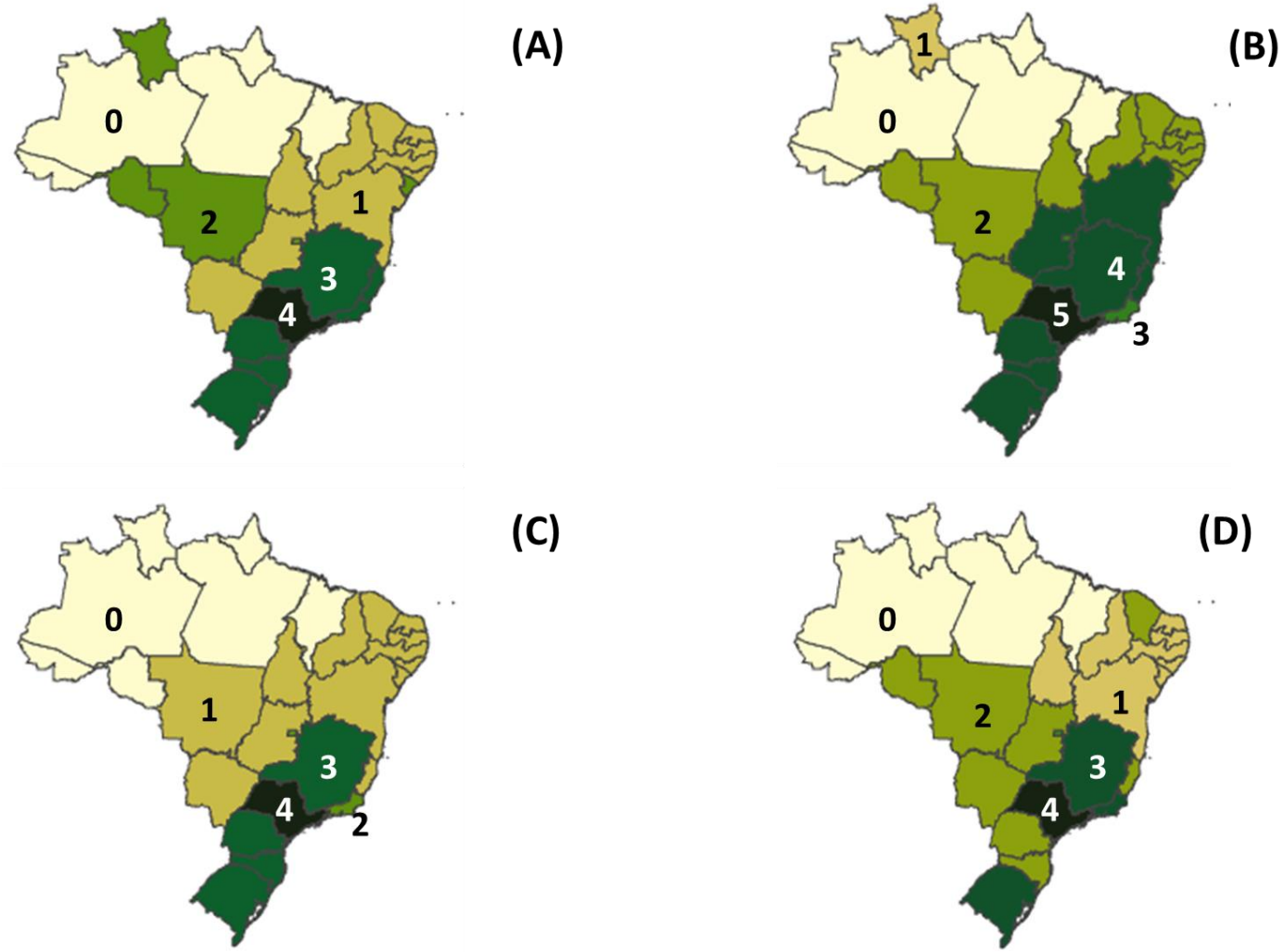

Figura 20. Distribuição espaço-temporal dos *clusters* entre 0 e 5 obtidos para as seguintes semanas epidemiológicas: (A) SE 30, (B) SE 45, (C) SE 61 e (D) SE 67, utilizando da técnica de análise não-hierárquica.

Um caso interessante de se analisar através do *K-Médias* se refere justamente aos estados de Rondônia (RO) e de Tocantins (TO). Como visto nas Figuras 19 e 20, esses estados estiveram, na maior parte do tempo, associados aos estados do Nordeste ou do Centro-Oeste, separados dos demais da sua região. Na Fig. 20(A), por exemplo, referente à SE 30, foi vista inclusão de RO, Roraima (RR), Mato Grosso (MT) e Sergipe (SE) no *cluster* 2. Ao se analisar os dados daquela semana, foi possível perceber o alto impacto da COVID-19 nesses estados quando comparado às demais UF do Norte. RO apresentou X_4 igual a 6,90 óbitos por 100 mil habitantes, sendo comparado ao visto para RR (6,81), MT (7,40) e SE (8,28), valores significativamente maiores daqueles observados para a região na SE 30, que teve um valor médio de X_4 igual a 4,28. O mesmo padrão pôde ser visto para X_2, com RO tendo 359 casos por 100 mil habitantes, enquanto o Norte teve X_2 médio de 264. Já TO foi alocado no *cluster* 1, junto com a maioria do Nordeste, apresentando X_4 igual a 3,27, mais próximo do valor médio desta região (4,03), sendo a mesma tendência observada para X_2.

A Fig. 20(B) mostra a alocação de RO e TO no *cluster* 2, relacionado ao período de melhora significativa no quadro da pandemia no país. Nesta semana, destaque para a indicação de RR em *cluster* único, podendo ser associado ao fato de que, apesar de ter apresentado baixo valor de X_4 (0,48, para uma média da região de 0,82), o estado presenciou uma taxa elevada de transmissão e de aumento do número de casos quando comparado aos demais, com X_2 igual a 292 casos por 100 mil habitantes na semana, contra 105,5 de média. É válido ressaltar, também, o IDH-M (X_{17}) mais elevado do estado frente às UF do Norte, tendo 0,752 contra 0,730 de valor médio, o que demonstra condições socioeconômicas relativamente melhores da UF ao se avaliar o contexto regional.

Na Fig. 20(C), RO foi interpretada pelo *K-Médias* como pertencente ao *cluster* 0. Avaliando a variável X_{18}, já presente na SE 61, foi possível avaliar a taxa de vacinação lenta na UF, com apenas 3,1% de doses aplicadas relativas à população (54.898 doses), sendo apenas a 17ª no *ranking* nacional até aquela semana. Em contraste, TO tinha a taxa de 3,6%, sendo a 9ª maior do país, o que ajuda a justificar a presença da UF no *cluster* 1, juntamente com os estados do Nordeste.

Por fim, a Fig. 20(D), associada à pior semana do histórico, mostra RO no *cluster* 2, fato que se relaciona com o impacto significativo em números de casos (X_1), com 7.011 casos (390,27 casos/100 mil hab.) e número de óbitos (X_3), tendo 266 (14,81 óbitos/100 mil hab.), valores consideravelmente altos para a região Norte-Nordeste. À título de comparação, como mencionado anteriormente na discussão dos resultados da técnica hierárquica, o CE teve 393,87 casos/100 mil hab. e 11,90 óbitos/100 mil hab., valores semelhantes àqueles vistos para RO na

semana. Tanto a análise hierárquica quanto a não-hierárquica indicaram, para o CE, alocação em *cluster* separado daquele referente às UF do Nordeste. Já o TO apresentou situação semelhante em relação ao Nordeste na semana, sendo alocado no mesmo grupo da região.

Os resultados obtidos pela técnica reforçam a heterogeneidade do impacto da COVID-19 ao longo das semanas epidemiológicas, em níveis estadual ou nacional. Alves *et al.* (2020), por exemplo, encontraram um número ótimo de 3 *clusters* para o *K-Médias* ao avaliarem os agrupamentos relacionados às variáveis de coeficientes de incidência, prevalência e letalidade, com a composição alterando substancialmente a depender do período analisado. Já Nascimento (2020*a*), avaliando dados referentes às SE 14 a SE 32 com um quantitativo de 10 variáveis, encontrou 5 grupos com a técnica, sendo mais próximo do número visto no presente trabalho, enquanto Guimarães, Eleuterio e Monteiro-da-Silva (2020) encontraram 3 *clusters* ao agrupar as UF em relação aos riscos de disseminação e de gravidade da COVID-19, construídos a partir de análise fatorial. Assim como a análise hierárquica, a técnica não-hierárquica se mostrou versátil e robusta, mostrando aplicabilidade em diferentes sistemas e considerações de estudo relacionadas à pandemia e a outros campos de pesquisa.

Inekwe, Maharaj e Bhattacharya (2020), por exemplo, compararam a utilização de técnicas de agrupamento hierárquica e não-hierárquica na avaliação das principais variáveis (e, consequentemente, dos *clusters* de países) na emissão de dióxido de carbono, para diferentes modelos empíricos. No trabalho, os autores constataram a formação de 2 a 4 agrupamentos a depender do modelo escolhido e da técnica, verificando através do Índice de Rand Ajustado (RAND, 1971) que as indicações de ambas as técnicas se mostraram compatíveis e robustas, possibilitando ao pesquisador a escolha de uma ou das duas abordagens na condução de análise de agrupamentos, fato que também foi constatado no presente trabalho.

Por fim, salienta-se que os resultados apresentados nas Tabelas 4 e 6 indicaram a consistência das análises hierárquica e não hierárquica, bem como alguma concordância de resultados. Embora os agrupamentos sejam diversos, devido às considerações destas diversas técnicas, elas concordam em si quanto aos agrupamentos finais (em termos de macro composição em relação às UF e regiões), considerando um grande número de observações (todas as UF mais o Distrito Federal) e todos os parâmetros envolvidos, e os números de *clusters* formados variaram entre um mínimo de três (na análise hierárquica) e um máximo de seis (na análise não hierárquica).

As Tabelas 5 e 7 mostram outra maneira de se enxergar a variabilidade das técnicas hierárquicas e não-hierárquicas. Observando algumas semanas selecionadas (13, 18, 22, 25, 30, 32, 36, 41, 45, 53, 56, 61, 64, 67 e 74), a alocação das UF não coincidem necessariamente entre

si. Isto se explica em parte pelas diferenças das análises, em especial: *i*) o número de agrupamentos (entre 0 e 5) não necessariamente coincide; *ii*) ainda que coincidissem no número de agrupamentos, uma envolve aspectos de aglomeração, a outra não; *iii*) há uma mudança entre *clusters* de semana para semana, considerando um mesmo estado. Tomando como exemplo a Tabela 7, é possível notar que Acre, Amapá, Amazonas, Maranhão e Pará encontram-se entre os grupos 0 e 1, Alagoas, Ceará, Sergipe e Tocantins entre os grupos 1 e 2. São Paulo sempre variou entre os *clusters* 3, 4 e 5.

Houve também alguma concordância de resultados das análises não hierárquicas deste trabalho com o estudo prévio de Nascimento (2020*a*). Por exemplo, houve um número pequeno de aglomerados (entre 4 e 6) usando uma rotina um pouco diferente de *K*-médias, baseado na linguagem Python.

As Figuras 14 e 18 ilustram a consistência do uso de técnicas diversas, além da concordância de resultados. A distribuição espaço-temporal obtida neste trabalho encontrou acordo com o trabalho prévio de Nascimento (2020*a*), que analisou um menor número tanto de variáveis quanto de semanas epidemiológicas.

4.3. ANÁLISE FATORIAL: ÍNDICE COVID-19 (IC19)

Inicialmente, para avaliação da adequabilidade da análise fatorial, foram performados os testes KMO e esfericidade de Bartlett, como mostrado na Tabela 8. É possível verificar que os mesmos indicaram a conformidade da aplicação da técnica para todas as SE, tendo em vista a obtenção de resultados, para KMO, superiores a 0,50 (vide Tabela 1) e, para o teste de esfericidade de Bartlett, valores-*p* menores que 0,05 (isto é, nível de significância de 95%), rejeitando-se a hipótese nula, conforme apresentado na Seção 2.2.2. Ademais, indica-se a quantidade de fatores extraídos na análise, de acordo com os critérios da raiz latente (autovalores λ_k^2 maiores que 1), explicitados na Seção 2.2.3.

Com a utilização da técnica para a determinação dos fatores, reduz-se estruturalmente os dados para que seja possível elaborar o Índice da COVID-19 no Brasil (IC19), podendo ser captado o comportamento conjunto e as características das dezoito variáveis previamente apresentadas. A partir do indicador, como proposto pelo presente trabalho, cria-se o *ranking* das UF e a posterior análise de agrupamentos, a ser detalhada nesta seção.

Foram determinados, de acordo com a Tabela 8, entre três e quatro fatores, reduzindo drasticamente o número de variáveis iniciais da Tabela 3 (entre 17 e 18, dependendo da SE). Este é um outro resultado surpreendente, tendo como base os parâmetros (*i.e.*, variáveis), o que

torna possível, por meio desta da redução numérica, uma melhor compreensão da distribuição da pandemia no país. Chama a atenção que, independentemente dos diversos valores, e das diferentes semanas epidemiológicas, a redução dimensional foi considerável.

Tabela 8. Resultados dos testes de adequabilidade da análise fatorial (KMO e Bartlett), além da quantidade de fatores extraídos, por SE. Todas os parâmetros, variáveis, independentemente das semanas, foram reduzidos entre 3 e 4 fatores, simplificando assim a análise multivariada.

SE	**Mês/Ano**	**KMO**	**Valor-*p* (Bartlett)**	**Fatores Extraídos**
13	Março/2020	0,754	0,000	3
14	Abril/2020	0,750	0,000	3
15		0,756	0,000	3
16		0,726	0,000	3
17		0,728	0,000	3
18		0,741	0,000	3
19	Maio/2020	0,694	0,000	3
20		0,733	0,000	3
21		0,730	0,000	4
22		0,726	0,000	4
23	Junho/2020	0,722	0,000	4
24		0,771	0,000	3
25		0,768	0,000	4
26		0,770	0,000	4
27	Julho/2020	0,772	0,000	4
28		0,771	0,000	4
29		0,741	0,000	3
30		0,768	0,000	3
31		0,698	0,000	3
32	Agosto/2020	0,728	0,000	3
33		0,721	0,000	3
34		0,646	0,000	3

Tabela 8. (Continuação)

SE	Mês/Ano	KMO	Valor-*p* (Bartlett)	Fatores Extraídos
35	Agosto/2020	0,646	0,000	3
36	Setembro/2020	0,598	0,000	4
37		0,697	0,000	4
38		0,726	0,000	4
39		0,644	0,000	4
40		0,776	0,000	4
41	Outubro/2020	0,758	0,000	4
42		0,751	0,000	3
43		0,734	0,000	4
44		0,749	0,000	4
45	Novembro/2020	0,723	0,000	4
46		0,686	0,000	4
47		0,775	0,000	4
48		0,711	0,000	4
49	Dezembro/2020	0,704	0,000	4
50		0,734	0,000	4
51		0,681	0,000	4
52		0,720	0,000	4
53		0,736	0,000	4
54	Janeiro/2021	0,720	0,000	4
55		0,739	0,000	4
56		0,683	0,000	4
57		0,625	0,000	4
58	Fevereiro/2021	0,676	0,000	4
59		0,686	0,000	4
60		0,730	0,000	4
61		0,748	0,000	4
62	Março/2021	0,749	0,000	4
63		0,729	0,000	4
64		0,684	0,000	4
65		0,670	0,000	4

Tabela 8. (Continuação)

SE	Mês/Ano	KMO	Valor-*p* (Bartlett)	Fatores Extraídos
66	Março/2021	0,649	0,000	4
67	Abril/2021	0,695	0,000	4
68		0,666	0,000	4
69		0,708	0,000	4
70		0,719	0,000	4
71	Maio/2021	0,730	0,000	3
72		0,719	0,000	3
73		0,758	0,000	4
74		0,779	0,000	4

Após a validação da técnica e a extração dos fatores, calculou-se o IC19 por UF para as semanas epidemiológicas conforme as Eqs. (3.4) e (3.5), sendo resumido na Tabela 9 para as 4 semanas epidemiológicas selecionadas nas Seções 4.1 e 4.2.

Tabela 9. Valores do IC19 para as UF nas semanas epidemiológicas: SE 30, SE 45, SE 61 e SE 67.

UF	SE 30	SE 45	SE 61	SE 67
AC	0,020	0,041	0,080	0,014
AL	0,108	0,094	0,063	0,056
AM	0,031	0,092	0,170	0,012
AP	0,000	0,000	0,000	0,000
BA	0,249	0,327	0,326	0,211
CE	0,190	0,194	0,209	0,234
DF	0,395	0,354	0,251	0,257
ES	0,284	0,357	0,271	0,268
GO	0,240	0,313	0,272	0,240
MA	0,051	0,053	0,018	0,018
MG	0,470	0,587	0,552	0,533
MS	0,192	0,243	0,208	0,200
MT	0,177	0,203	0,192	0,200
PA	0,055	0,074	0,082	0,052
PB	0,154	0,163	0,152	0,115
PE	0,238	0,254	0,191	0,173
PI	0,108	0,155	0,089	0,088
PR	0,355	0,445	0,438	0,361
RJ	0,611	0,618	0,505	0,468
RN	0,138	0,142	0,134	0,111

Tabela 9. (Continuação)

UF	SE 30	SE 45	SE 61	SE 67
RO	0,104	0,109	0,190	0,122
RR	0,044	0,107	0,157	0,005
RS	0,408	0,576	0,529	0,418
SC	0,309	0,481	0,428	0,278
SE	0,158	0,134	0,094	0,099
SP	1,000	1,000	1,000	1,000
TO	0,137	0,170	0,162	0,113

Também foi possível analisar a distribuição espaço-temporal do IC19, conforme pode ser visto na Figura 21. Já os valores das comunalidades extraídas para cada variável, referentes à análise fatorial das 4 semanas epidemiológicas destacadas, são apresentados na Tabela 10.

Tabela 10. Valores das comunalidades das 18 variáveis em estudo nas semanas epidemiológicas: SE 30, SE 45, SE 61 e SE 67.

VAR	Descrição	SE 30	SE 45	SE 61	SE 67
X_1	Quantidade de novos casos	0,924	0,825	0,915	0,968
X_2	Quantidade de novos casos por 100 mil habitantes	0,748	0,885	0,938	0,879
X_3	Quantidade de novos óbitos	0,960	0,920	0,971	0,983
X_4	Quantidade de novos óbitos por 100 mil habitantes	0,607	0,565	0,727	0,872
X_5	População	0,986	0,990	0,993	0,992
X_6	Densidade populacional	0,584	0,866	0,909	0,887
X_7	Quantidade de leitos de UTI (SUS)	0,957	0,975	0,971	0,985
X_8	Quantidade de leitos de UTI (Não SUS)	0,918	0,937	0,927	0,917
X_9	Quantidade de leitos de UTI COVID-19	0,987	0,985	0,984	0,991
X_{10}	% Dependência do SUS	0,849	0,889	0,870	0,882

Tabela 10. (Continuação)

VAR	Descrição	SE 30	SE 45	SE 61	SE 67
X_{11}	Quantidade de profissionais da saúde	0,985	0,984	0,984	0,995
X_{12}	Profissionais da saúde por 100 mil habitantes	0,810	0,845	0,869	0,802
X_{13}	Número médio de moradores por domicílio	0,858	0,873	0,898	0,907
X_{14}	% pessoas com 60 anos ou mais	0,887	0,892	0,883	0,924
X_{15}	% pessoas com 18 anos ou mais diagnosticadas com hipertensão	0,802	0,858	0,874	0,889
X_{16}	% pessoas com 18 anos ou mais diagnosticadas com doenças cardiovasculares	0,841	0,881	0,927	0,883
X_{17}	Índice de Desenvolvimento Humano	0,742	0,897	0,904	0,894
X_{18}	Doses aplicadas das vacinas	-	-	0,952	0,973

Ao verificar os resultados obtidos para as 4 semanas epidemiológicas escolhidas, vistos na Tabela 9, percebe-se a presença de informações importantes acerca do quadro da COVID-19 no território brasileiro, por meio das oscilações do IC19 nas UF ao longo dos diferentes momentos da pandemia. Destaca-se inicialmente o estado de São Paulo (SP) com o indicador mais elevado ao longo de todos os períodos analisados, seguido por Rio de Janeiro (RJ) e Minas Gerais (MG). No espectro oposto, os menores valores do IC19 foram obtidos de forma consistente nos estados do Amapá (AP), Acre (AC) e Maranhão (MA).

A análise acerca desta distribuição perpassa inicialmente pelos resultados de comunalidade, vistos na Tabela 10. Conforme mencionado na Seção 2.2.4, a comunalidade de uma variável X_i representa a relevância que a mesma detém na composição de um determinado fator F_k, sendo acentuada quanto maior for a significância deste fator F_k para a construção do *ranking*, mensurada através do módulo do seu autovalor λ_k^2.

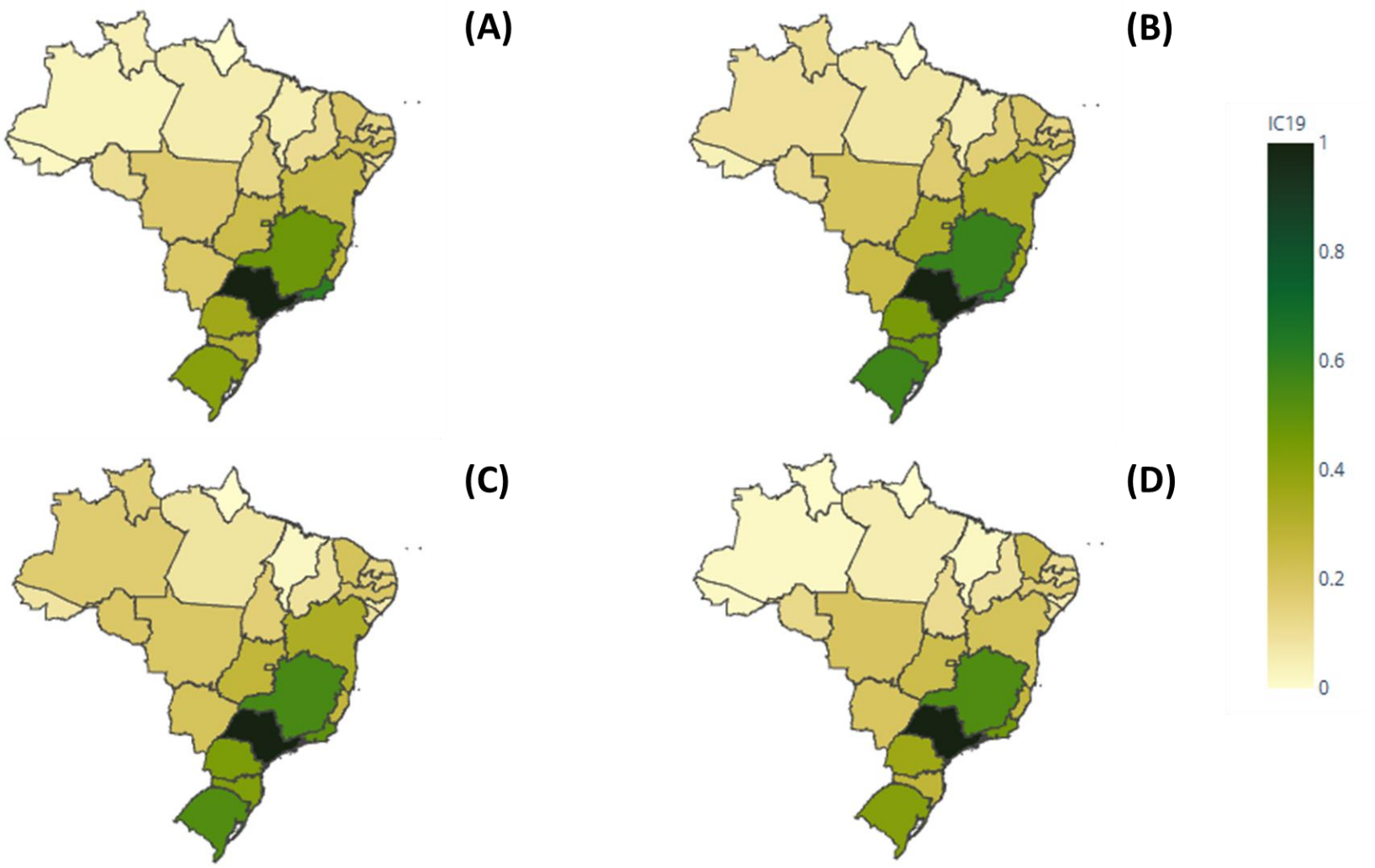

Figura 21. Distribuição espaço-temporal dos valores do IC19 por UF obtidos para as semanas epidemiológicas: (A) SE 30, (B) SE 45, (C) SE 61 e (D) SE 67. É possível perceber variações significativas entre as quatro semanas, com destaque para AM e RR.

Dentre as 17 variáveis estudadas na análise fatorial para o ano de 2020, destacaram-se: X_5 (população), X_7 (quantidade de leitos de UTI SUS), X_9 (quantidade de leitos de UTI COVID-19) e X_{11} (quantidade de profissionais de saúde), com extrações de variância acima de 0,95 para as SE 30 e SE 45. Já para 2021, com as SE 61 e SE 67, levando-se em conta 18 variáveis (introduzido o parâmetro de doses de vacinas aplicadas), destacaram-se X_3 (quantidade de óbitos) e X_{18} (doses de vacinas aplicadas), além das 4 citadas acima. É perceptível que todos os 6 parâmetros citados têm relação direta com a quantidade de pessoas vivendo em uma determinada UF, o que ajuda a explicar os valores mais elevados de IC19 para estados mais populosos, como SP, RJ e MG.

De forma oposta, as variáveis X_4 (óbitos por 100 mil habitantes) e X_6 (densidade populacional) detiveram os menores valores de comunalidade extraída para as SE 30 e SE 45, levando a um menor impacto na composição do indicador. Já para as SE 61 e SE 67, as comunalidades extraídas foram relativamente maiores do que para o grupo de semanas epidemiológicas de 2020, com todas as variáveis acima de 0,70, sendo os menores valores para X_4 e X_{12} (profissionais da saúde por 100 mil habitantes).

Como visto na Tabela 8, para todas as semanas analisadas, houve redução de dimensionalidade de 17 ou 18 variáveis para 3 ou 4 fatores, a depender do período. Esta redução ilustra o potencial de aplicação da ferramenta e dialoga com outros resultados encontrados na literatura, como visto em Bezerra *et al.* (2020), que obteve um conjunto de 3 fatores para o sistema inicial de 21 variáveis relacionadas à infraestrutura de saúde e enfrentamento à COVID-19 no Brasil, e em Guimarães, Eleuterio e Monteiro-da-Silva (2020), que obtiveram 4 fatores para um conjunto de 13 variáveis associadas a categorias populacionais, ocupação territorial e infraestrutura de saúde, com o intuito de agrupar as UF no Brasil pelos riscos de disseminação e gravidade da COVID-19.

Especificamente em relação à composição dos fatores, os principais parâmetros a serem analisados são as cargas fatoriais, vistas na Seção 2.2.4, com os resultados para as 4 SE selecionadas resumidos na Tabela 11. Os termos destacados em vermelho afetaram o fator negativamente, enquanto aqueles em azul contribuíram positivamente. Salienta-se, também, que os fatores indicados se apresentam ordenados por magnitude dos seus respectivos autovalores λ_k^2, isto é, $\lambda_1^2 > \lambda_2^2 > \lambda_3^2 > \lambda_4^2$. Portanto, o fator F_1 foi o mais relevante na construção do indicador IC19, enquanto o fator F_4 foi o menos relevante.

Para definição das variáveis que foram relevantes na construção de determinado fator, é indicado que as mesmas detenham carga fatorial de 0,500, em módulo (MINGOTI, 2005).

Tabela 11. Cargas fatoriais dos fatores F_1 a F_3 (ou F_4, quando o caso), para as semanas epidemiológicas SE 30, SE 45, SE 61 e SE 67.

	SE 30			SE 45				SE 61				SE 67			
VAR	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**
X_1	0,944	0,147	0,106	0,506	0,564	0,002	0,501	0,838	0,376	0,050	0,263	0,942	0,190	0,039	0,206
X_2	-0,266	-0,455	0,685	-0,265	-0,143	0,086	0,887	-0,198	0,078	0,081	0,941	-0,081	0,026	-0,030	0,933
X_3	0,953	0,190	0,126	0,848	0,380	0,232	-0,043	0,953	0,198	0,134	0,070	0,948	0,191	0,162	0,150
X_4	-0,160	-0,304	0,700	-0,208	0,470	0,549	-0,007	0,081	-0,296	0,025	0,795	0,084	0,227	0,356	0,829
X_5	0,963	0,240	-0,041	0,972	0,203	0,045	-0,051	0,970	0,187	0,095	-0,092	0,968	0,203	0,102	-0,050
X_6	0,260	0,214	0,686	0,194	0,036	0,868	-0,272	0,131	0,030	0,879	-0,345	0,119	0,115	0,913	-0,164
X_7	0,933	0,292	0,015	0,954	0,238	0,089	0,014	0,946	0,235	0,141	-0,016	0,956	0,226	0,130	0,044
X_8	0,899	0,226	0,243	0,876	0,131	0,360	-0,149	0,861	0,117	0,377	-0,174	0,855	0,173	0,387	-0,082
X_9	0,964	0,237	0,045	0,957	0,233	0,123	-0,020	0,948	0,249	0,146	-0,038	0,961	0,227	0,122	0,039
X_{10}	-0,566	-0,568	-0,454	-0,542	-0,430	-0,639	-0,045	-0,520	-0,420	-0,651	0,028	-0,529	-0,411	-0,603	-0,264
X_{11}	0,961	0,237	0,073	0,964	0,184	0,145	-0,008	0,953	0,187	0,195	-0,041	0,960	0,193	0,191	0,012
X_{12}	0,215	0,255	0,836	0,203	0,090	0,809	0,376	0,143	0,180	0,848	0,312	0,151	0,108	0,802	0,353
X_{13}	-0,191	-0,883	-0,205	-0,219	-0,868	-0,267	0,001	-0,213	-0,885	-0,264	-0,003	-0,208	-0,872	-0,236	-0,219
X_{14}	0,288	0,896	-0,030	0,320	0,853	0,176	-0,180	0,317	0,828	0,185	-0,251	0,293	0,888	0,204	-0,090
X_{15}	0,201	0,864	-0,123	0,243	0,884	-0,011	-0,131	0,249	0,878	-0,020	-0,202	0,247	0,909	-0,033	-0,005
X_{16}	0,234	0,878	0,126	0,295	0,847	0,169	0,220	0,288	0,879	0,194	0,182	0,297	0,817	0,145	0,326
X_{17}	0,393	0,381	0,665	0,404	0,218	0,707	0,432	0,355	0,258	0,762	0,361	0,366	0,166	0,698	0,495
X_{18}	-	-	-	-	-	-	-	0,923	0,204	0,212	-0,119	0,922	0,306	0,146	-0,093

Portanto, o Fator 1 (F_1), mais relevante na construção do índice, foi definido pelas variáveis X_1 (quantidade de casos), X_3 (quantidade de óbitos), X_5 (população), X_7 (quantidade de leitos de UTI SUS), X_8 (quantidade de leitos de UTI não-SUS), X_9 (quantidade de leitos de UTI COVID-19), X_{10} (% dependência SUS), X_{11} (quantidade de profissionais de saúde) e, para as semanas de 2021, X_{18} (doses aplicadas das vacinas). Desta forma, é possível relacioná-lo principalmente ao impacto da COVID-19 e à infraestrutura e gestão do sistema de saúde das UF.

Para o Fator 2 (F_2), obteve-se: X_{13} (número médio de moradores por domicílio), X_{14} (% pessoas com 60 anos ou mais), X_{15} (% pessoas com 18 anos ou mais com hipertensão) e X_{16} (% pessoas com 18 anos ou mais com doenças cardiovasculares). É possível associar, portanto, a relação primordial deste fator principalmente com aspectos sociais das UF, relacionados à vulnerabilidade e a fatores de risco da população.

Já o Fator 3 (F_3) foi constituído de forma significativa por: X_6 (densidade populacional), X_{10} (% dependência SUS), X_{12} (profissionais de saúde por 100 mil habitantes) e X_{17} (IDHM), sendo relacionado principalmente à infraestrutura de saúde e ao desenvolvimento social.

Por fim, o Fator 4 (F_4), menos relevante na composição do índice, foi caracterizado principalmente por X_2 (casos por 100 mil habitantes) e X_4 (óbitos por 100 mil habitantes), podendo ser descrito principalmente como fator de impacto padronizado em relação à COVID-19.

Por meio dos fatores obtidos, torna-se possível analisar os resultados obtidos para o índice. A concentração de leitos de UTI (SUS, Não-SUS e COVID-19) em SP e em outros estados do Sudeste e do Sul, como RJ, MG e RS, por exemplo, ajuda a explicar o elevado valor do IC19 para esses estados. Esta característica, trazida também por outros pesquisadores, como Junior e Cabral I (2020) e Bezerra *et al.* (2020), mostra a heterogeneidade de estrutura no país e reforça a necessidade de alocação de recursos em estados deficitários que, no presente trabalho, apresentaram os menores resultados de IC19, notadamente na região Norte, como AP, AC e AM e no Nordeste, como MA. Este resultado se mostra em acordo com o estudo de Requia *et al.* (2020), que analisaram o risco de sobrecarga no sistema de saúde de 5.572 municípios brasileiros mediante critérios socioeconômicos, populacionais e estruturais, concluindo que 69% dos mesmos estariam neste cenário devido à COVID-19 e, destes, 52% seriam localizados nas regiões Norte ou Nordeste.

Apesar de o fator F_1 ser composto, também, pelas variáveis quantidade de casos e de óbitos, notadamente maior nas regiões de maior população (Sul e Sudeste) e que contribuíram

para um aumento no IC19, foi visto a grande relevância das variáveis de infraestrutura de saúde, o que contribuiu para os resultados em acordo com outros artigos da literatura, como em Bezerra *et al.* (2020). Neste, os pesquisadores propuseram a criação do Índice de Infraestrutura em Saúde (IIS), por meio de análise fatorial e consequente redução de dimensionalidade de 21 variáveis para 3 fatores, encontrando os maiores valores de IIS para SP, MG e RJ e os menores para AP, RO e AC, em ordens respectivas. Porém, é importante ressaltar que F_1 não é definido unicamente por este quesito, como mencionado, tendo em vista que o impacto da COVID-19 também teve influência no mesmo. Isto pode ser visto, por exemplo, com o episódio do colapso do atendimento da rede hospitalar em Manaus, capital do AM, em janeiro de 2021 (COVID-19, 2021), em que foi presenciado, para a SE 61, um aumento significativo no resultado do IC19 para o estado, mostrado na Tabela 9. Como mencionado anteriormente na análise não-hierárquica, o impacto da pandemia naquela semana, para RO, também foi verificado na análise fatorial, novamente com o aumento do IC19.

Para o fator F_2, é relevante mencionar a carga fatorial fortemente negativa da variável X_{13} (número médio de moradores por domicílio), reduzindo o IC19 para as UF com a maior densidade de pessoas, que mostra o risco de disseminação da doença dentro de uma mesma estrutura familiar e que, consequentemente, tende a expor a UF e o seu sistema de saúde. De acordo com a base da PNAD 2019, utilizada para compor a variável X_{13}, AP (3,8 hab./domicílio), AM (3,6) e RR (3,5) foram os estados com a maior exposição. De forma contrária, RS (2,6) e RJ (2,7) foram as que apresentaram a menor densidade, impactando menos o IC19. SP, por exemplo, tem a densidade de 2,8, uma das menores do Brasil.

Para F_3, observou-se a relevância da variável X_6, com carga fatorial fortemente positiva. Destaque para DF, com densidade de cerca de 530,84 hab./km², que teve aumento no IC19 devido a este quesito, assim como RJ (396,94 hab./km²) e SP (186,49 hab./km²). No espectro oposto, AM (2,70 hab./km²), RR (2,82 hab./km²) e MT (3,90 hab./km²) tiveram os seus IC19 fracamente definidos pela variável. Outra variável mencionável foi X_{10} (% dependência SUS), que se relaciona diretamente com a pressão exercida no sistema de saúde público de cada UF e que teve carga fatorial negativa na formação do índice (ou seja, quanto maior o valor de X_{10}, menor o IC19). A tendência é que esta variável seja menor quanto maior for o desenvolvimento socioeconômico da UF, o que pode ser verificado com os três menores valores obtidos para a mesma (SP (62,3%), RJ (69,4%) e DF (69,9%)) e os três maiores (RR (95,3%), AC (95,3%) e MA (93,6%)).

Por fim, para o fator F_4, menciona-se o impacto padronizado da pandemia em cada estado, por meio das variáveis X_2 e X_4, principalmente para o ano de 2021 (SE 61 e SE 67). As

cargas fatoriais positivas indicaram o aumento do IC19 conforme o aumento do número de casos e de óbitos por 100 mil habitantes, o que justifica o aumento do IC19 para AM e RO na SE 61, por exemplo, que apresentaram o 2º e o 3º maiores valores para X_4 do Brasil, respectivamente. O mesmo quesito foi visto para RR, que apresentou o 2º maior resultado de X_2 e o maior valor de X_4 na semana.

É possível citar, também, a possibilidade de avaliar a relevância e o impacto que a variável X_{18} trouxe para o quadro da COVID-19 no Brasil, por meio das técnicas trazidas no presente trabalho. Para isso, comparou-se os resultados obtidos para as SE 22 e SE 74, referentes às últimas semanas de maio/20 e maio/21, respectivamente. Até a SE 74, haviam sido aplicadas 63.722.478 doses das vacinas, de acordo com dados oficiais do Ministério da Saúde, sendo estratificado na Tabela 12.

Tabela 12. Relação entre a quantidade acumulada de doses de vacinas aplicadas (X_{18}) e o correspondente % de vacinação em relação à população de cada UF até a SE 74. Também é trazido o valor de X_{18} apenas na SE 74. Percebe-se a grande quantidade de vacinas aplicadas em SP, MG e RJ, porém o maior % em relação à população foi visto para RS, MS e ES. O valor da variável, especificamente na SE 74, reflete bem o quadro acumulado até aquele momento.

UF	Quantidade acumulada de doses de vacinas aplicadas	% acumulada em relação à população	Quantidade de doses de vacinas aplicadas apenas na SE 74
AC	197.304	22,1	13.676
AL	993.469	29,6	94.188
AM	1.074.193	25,5	75.423
AP	185.348	21,5	9.874
BA	4.389.283	29,4	269.050
CE	2.072.964	22,6	124.212
DF	853.403	27,9	47.439
ES	1.459.451	35,9	128.364
GO	2.078.382	29,2	151.280
MA	1.789.582	25,2	104.507
MG	6.648.223	31,2	382.886
MS	1.051.524	37,4	90.457
MT	894.016	25,4	58.339
PA	1.851.178	21,3	190.920
PB	1.331.029	33,0	54.412
PE	2.561.420	26,6	177.853
PI	914.785	27,9	53.155
PR	3.666.255	31,8	250.380

Tabela 12. (Continuação)

UF	Quantidade acumulada de doses de vacinas aplicadas	% acumulada em relação à população	Quantidade de doses de vacinas aplicadas apenas na SE 74
RJ	5.796.636	33,4	446.923
RN	1.121.648	31,7	77.494
RO	400.553	22,3	26.128
RR	128.147	20,3	7.156
RS	4.686.770	41,0	307.097
SC	2.137.549	29,5	175.031
SE	563.063	24,3	37.359
SP	14.485.060	31,3	1.142.810
TO	391.243	24,6	21.014

É possível avaliar a grande concentração de doses aplicadas nas regiões Sul e Sudeste, principalmente em SP, MG e RJ, o que pode ser relacionado com a grande quantidade de pessoas vivendo nessas UF. Ao se analisar o valor de percentual acumulado até a SE 74, percebe-se que o resultado se altera, com RS, MS e ES representando os melhores percentuais de aplicação das vacinas em relação às suas respectivas populações. A situação da administração das doses, especificamente na SE 74, reflete bem o cenário acumulado até aquela data, como pode ser visto na Tabela 12. A comparação entre as duas semanas mencionadas pode ser iniciada com o resultado do IC19 para ambas, como mostra a Tabela 13.

Tabela 13. IC19 obtidos para as SE 22 e SE 74. Percebe-se que, para a SE 74, quase todas as UF tiveram redução em seus índices quando comparados aos referentes à SE 22, com exceção de MG, que teve um aumento, e AP e SP, que se mantiveram estáveis.

UF	SE 22	SE 74
AC	0,087	0,002
AL	0,247	0,076
AM	0,105	0,009
AP	0,000	0,000
BA	0,335	0,260
CE	0,346	0,234
DF	0,380	0,229
ES	0,384	0,247
GO	0,293	0,234
MA	0,159	0,040
MG	0,508	0,528

Tabela 13. (Continuação)

UF	SE 22	SE 74
MS	0.266	0.262
MT	0.198	0.175
PA	0.223	0.062
PB	0.253	0.152
PE	0.394	0.217
PI	0.211	0.095
PR	0.400	0.389
RJ	0.820	0.482
RN	0.250	0.125
RO	0.145	0.098
RR	0.023	0.037
RS	0.487	0.397
SC	0.361	0.281
SE	0.231	0.161
SP	1.000	1.000
TO	0.199	0.126

Os resultados obtidos demonstram redução do IC19 ao realizar o comparativo das SE 22 e SE 74, mesmo com a variável X_{18} relacionada positivamente com o fator F_1. Isso se deve à aplicação do procedimento de padronização do índice (vide Seção 3.3, Eq. (3.5)) e, também, pelo fato de SP ter apresentado grande quantidade de doses aplicadas. Consequentemente, o IC19 absoluto de SP (calculado pela Eq. (3.4)), que representou o máximo entre as UF, aumentou (1,58 na SE 22 para 1,93 na SE 74), fazendo com que as demais UF tivessem o seu IC19 relativo diminuído. A Figura 22 ilustra geograficamente a mudança no valor do índice.

A introdução do parâmetro adicional também alterou a estrutura dos *clusters* obtidos para cada semana, como mostrado nas Figuras 23 e 24, referentes às técnicas hierárquica e não-hierárquica, respectivamente. A Fig. 23(B) demonstra que, devido à diferença significativa em relação à quantidade de doses aplicadas, SP passou a ser alocada em *cluster* único, enquanto em Fig. 23(A) a mesma estava dividindo o grupo com RJ. É possível perceber, também, que DF não teve mudança de grupo, associado à sua alta densidade populacional e elevado IDHM. O grupo subsequente, que na SE 22 era formado por ES, GO, MS, PR, SC, MG e RS, passou a contar apenas com RJ, MG, PR e RS na SE 74, sobretudo por conta da influência da nova variável e, também, devido ao agravamento do número de casos e de óbitos, comparativamente.

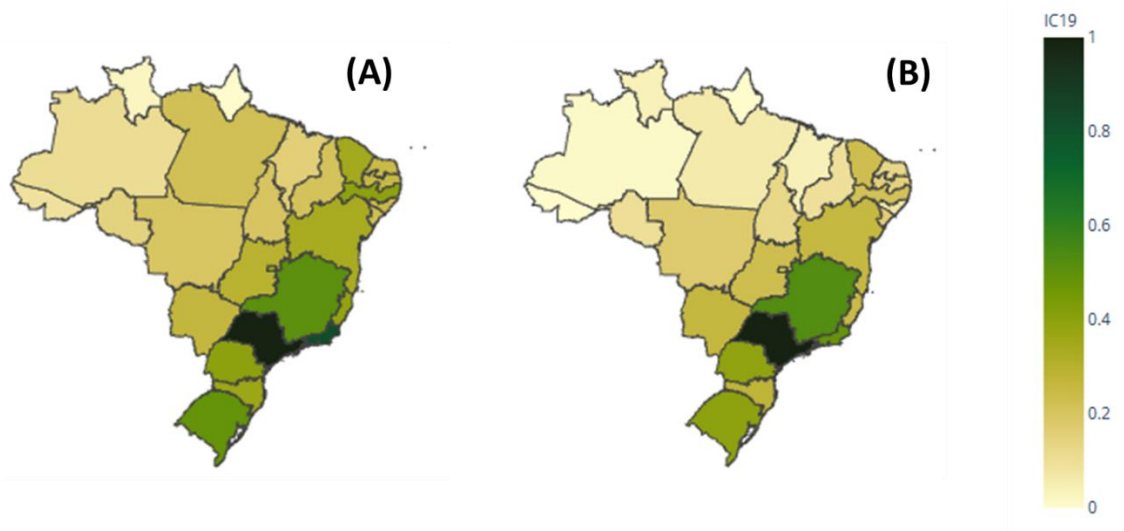

Figura 22. Mudança no valor de IC19 visto na distribuição geográfica do Brasil. Os índices obtidos diminuíram no comparativo entre SE 22 (A) e SE 74 (B), devido ao aumento do valor absoluto para SP. Apesar deste fato, foi possível perceber a mudança no resultado e a influência das doses das vacinas no sistema em estudo.

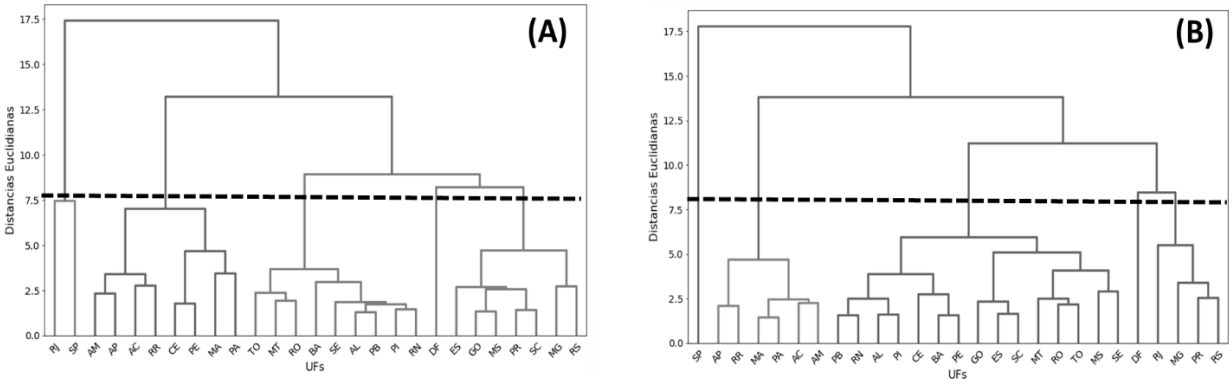

Figura 23. Comparativo dos dendrogramas obtidos para (A) SE 22 e (B) SE 74. O número de *clusters* para as duas semanas foi igual a 5, porém as composições dos grupos foram alteradas, sob influência do novo parâmetro.

As UF do Norte e do Nordeste não apresentaram mudanças significativas, associado à menor quantidade de doses aplicadas em relação às UF de maior população. A Fig. 24(B) mostra a mudança da estrutura de grupos de acordo com *K-Médias*, com destaque para ES, DF e SC que, diferente da análise hierárquica, foram alocadas no 2º *cluster*, junto com MG, RJ e RS, divergindo também da Fig. 24(A), em que DF e RJ foram inseridas em *clusters* únicos. Novamente, o Norte e o Nordeste permaneceram em *clusters* semelhantes.

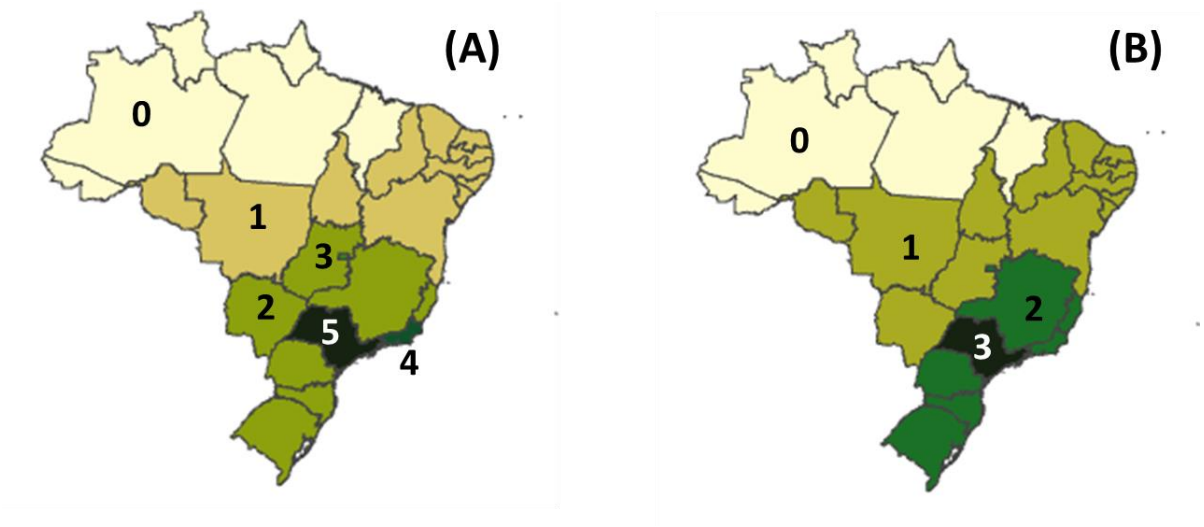

Figura 24. Comparativo dos grupos obtidos para (A) SE 22 e (B) SE 74, geograficamente visualizados. O número de *clusters* para a primeira foi igual a 6, enquanto para a segunda foi 4, como mostrado no mapa. Assim como as demais técnicas, a introdução da nova variável alterou a composição de cada agrupamento.

O estudo das três técnicas em relação à adição do novo parâmetro, portanto, mostrou-se relevante na avaliação de estratégias de alocação de recursos e políticas públicas adotadas em cada UF, refletindo as mudanças espaço-temporais vistas na pandemia no Brasil e no resto do mundo.

Salienta-se que o presente trabalho levou em consideração as duas primeiras ondas relacionadas a pandemia no país, apresentado nas Figuras 5, 6 e 7. Em primeiro lugar, é notável que a disseminação da COVID-19 em solo brasileiro teve um atraso em relação aos primeiros focos mundiais. Referente à primeira onda, ainda em 2020, este tempo poderia ter sido usado para o estabelecimento de políticas públicas para atenuar os impactos. Pode-se também diferenciar as duas ondas da seguinte maneira: o comportamento na primeira onda foi desordenado, com base em certas capitais. Na segunda onda brasileira, houve uma coordenação mais síncrona dos focos pandêmicos, com altas em muitos estados em um mesmo intervalo curto de tempo, que também coincidiu com o surgimento de variantes locais.

É possível verificar, também, que um dos grandes erros do governo brasileiro foi não adotar um controle de distanciamento, bem como uma sistemática de monitoramento testes e, por consequente, de um controle efetivo da taxa de transmissão da doença. Basicamente, o índice levado em conta foi a oferta do número de leitos e a sua ocupação. Isto não evitou a taxa de contágio e visou apenas contornar uma situação de calamidade em um curto período, sem um investimento a médio ou longo prazos. Ao resolver uma situação momentânea, esta política apenas retardou a chegada de novas ondas, impactando no erário público e sem atuar

preventivamente em novos casos.

A conjugação de alguns fatores, como o uso de máscaras, a disseminação de cuidados higiênicos como usar álcool ou lavar as mãos e o distanciamento social auxiliaram, em primeiro momento, no decréscimo dos números de casos e óbitos até o período eleitoral de 2020. Com o fim daquele ano, as férias escolares (de cidades que não adotaram o ensino remoto), o fim do auxílio emergencial, as festividades de fim de ano e o surgimento de novas variantes da doença foram determinantes para o início mais intenso da segunda onda, amainada pela aplicação vacinal nos meses correspondentes.

Algo esclarecedor do comportamento do governo federal pode ser obtido a partir do Boletim Epidemiológico 7, de 6 de abril de 2020, onde foi adotado um modelo de casos da doença no país. Fica bastante claro que a expectativa seria de término da pandemia no inverno (até antes de outubro) considerando três situações: sem distanciamento social, com distanciamentos seletivo e ampliado.

5. CONCLUSÕES

A análise multivariada espaço-temporal da pandemia COVID-19 no país, efetuada entre março de 2020 até maio de 2021, envolvendo mais de 16 milhões de casos diagnosticados e com mais de 460 mil óbitos, apresentou importantes resultados, com fortes repercussões políticas, sociais e econômicas, além dos aspectos de saúde. Tomando como base as 27 Unidades Federativas, foram consideradas 18 variáveis de diferentes categorias para as semanas epidemiológicas do período analisado, utilizando dados oficiais de portais do governo brasileiro como DataSUS, TCU, IBGE, IPEA, ANS e Painel COVID-19, atualizado pelo Ministério da Saúde.

Em termos gerais, análises hierárquicas e não-hierárquicas, baseadas em similaridades, resultaram em pequenos agrupamentos, de 3 a 5 grupos de acordo para a primeira, e de 4 a 6 grupos para a segunda, considerando todos os estados brasileiros (mais o distrito federal). Foi possível perceber alguns padrões e características específicas para cada estado em cada semana epidemiológica, englobando mais de um ano de dados oficiais. Particularmente, este trabalho mostrou que é possível reduzir um problema altamente complexo, envolvendo quase três dezenas de estados e uma população de 213 milhões de habitantes, em poucos grupos.

Um outro resultado importante foi que, em todas as semanas epidemiológicas, foi possível reduzir todas as 18 variáveis (ou parâmetros) em apenas três ou quatro fatores. Particularmente, a criação de um indicador capaz de ordenar todas as UF a partir das 18 variáveis, chamado de Índice COVID-19 (IC19), possibilitou verificar a mudança espaço-temporal da doença para as UF nos diferentes momentos da pandemia. Especificamente para a análise fatorial, os estados de São Paulo (SP), Rio de Janeiro (RJ) e Minas Gerais (MG) apresentaram os maiores índices IC19, enquanto Amapá (AP), Acre (AC) e Maranhão (MA) apresentaram os menores.

Em termos gerais, espera-se que as diversas ferramentas apresentadas possam servir de auxílio à gestão epidemiológica no país, ao contribuir para órgãos públicos quanto a tomadas de decisão, separando os estados por agrupamentos semana a semana, independentemente do número de variáveis, visando a priorização e correta aplicação de recursos e insumos pelos governos estaduais e federal.

6. TRABALHOS FUTUROS

É possível estabelecer uma série de futuros trabalhos com base na proposta deste trabalho. O mais intuitivo seria dar continuidade aos dados de mais semanas epidemiológicas e mais variáveis.

O uso de outras técnicas multivariadas também pode fornecer novos e substanciais resultados. Entre as diversas opções, pode-se destacar o escalonamento multidimensional, modelos de regressão e predição e cálculos de correlações espaciais, como a estatística I de Moran (ALMEIDA, 2012).

A disseminação destas técnicas, muito conhecidas pelas áreas de exatase e tecnologias, em escolas de saúde pública e faculdades de medicina, é outro ponto importante a ser trabalhado.

REFERÊNCIAS

ALCÂNTARA, E. *et al.* Investigating spatiotemporal patterns of the COVID-19 in São Paulo State, Brazil. **Geospatial Health.** v. 15, p. 201 – 209, 2020.

ALMEIDA, E. **Econometria Espacial Aplicada.** 1. Ed. Campinas: Alínea, 2012. 498 p.

ALVES, H. J. P. *et al.* A pandemia da COVID-19 no Brasil: uma aplicação do método de clusterização k-means. **Research, Society and Development.** v. 9, n. 10, e5829109059, 2020.

ANONIMO. COVID-19 in Brazil: "So What?". **Lancet** v. 395, p. 1461-1461, 2020.

ANTON, H., RORRES, C. **Álgebra Linear com Aplicações.** 10. Ed. Porto Alegre: Bookman, 2012. 784 p.

BARTLETT, M.S. A note on the multiplying factors for various χ^2 approximations. **Journal of Statistical Society: Series B** v. 16, p. 296-298, 1954.

BEHPOUR, S. *et al.* Automatic trend detection: Time-biased document clustering. **Knowledge-Based Systems.** v. 220, e106907, 2021.

BERNOULLI, D. Reflexions sur les Avantages de l'Inoculation. **Mercure de France** v. 6, p. 173-190, 1760.

BEZERRA, E. *et al.* Análise espacial das condições de enfrentamento à COVID-19: uma proposta de Índice de Infraestrutura da Saúde do Brasil. **Ciência & Saúde Coletiva.** v. 25, p. 4957-4967, 2020.

CASTRO, M. C. *et al.* Spatiotemporal pattern of COVID-19 spread in Brazil. **Science**, v. 372, p. 821 – 826, 2021.

CHEN, T., SHI, X., WONG, Y. D. A lane-changing risk profile analysis method based on time-series clustering. **Physica A: Statistical Mechanics and its Applications.** v. 565, e125567, 2021.

CONDE, M. Brazil in The Time of Coronavirus. **Geopolítica(s)**, v. 11, p. 239-249, 2020.

COVID-19: Manaus vive colapso com hospitais sem oxigênio, doentes levados a outros estados, cemitérios sem vagas e toque de recolher. **G1**, Manaus, 14 jan. 2021. Disponível em: <https://g1.globo.com/am/amazonas/noticia/2021/01/14/covid-19-manaus-vive-colapso-com-hospitais-sem-oxigenio-doentes-levados-a-outros-estados-cemiterios-sem-vagas-e-toque-de-recolher.ghtml>. Acesso em: 05 jun. 2021.

EULER, L. Recherches Générales sur la Mortalité et la Multiplication du Genre Humain. **Mémoires de l'académie des sciences de Berlin**, v. 16, p. 144-164. 1760.

FÁVERO, L. P., BELFIORE, P. **Manual de Análise de Dados:** Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 1. Ed. Rio de Janeiro: Elsevier, 2017. 1216 p.

FÁVERO, L. P., BELFIORE, P. **Data Science for Business and Decision Making.** Cambridge, MA: Elsevier, 2019. 1227 p.

FERRAZ, D. *et al.* COVID Health Structure Index: The Vulnerability of Brazilian Microregions. **Social Indicators Research**. 2021.

GUIMARÃES, R. M., ELEUTERIO, T. A., MONTEIRO-DA-SILVA, J. H. C. Estratificação de risco para predição de disseminação e gravidade da Covid-19 no Brasil. **Revista Brasileira de Estudos de População.** v. 37, p. 1-17, 2020.

HAIR, J. F. *et al.* **Multivariate Data Analysis.** 10th Edition. Hampshire: Cengage, 2019. 834 p.

HOZUMI, Y. *et al.* UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. **Computers in Biology and Medicine.** v. 131, e104264, 2021.

INEKWE, J., MAHARAJ, E.A., BHATTACHARYA, M. Drivers of carbon dioxide emissions: an empirical investigation using hierarchical and non-hierarchical clustering methods. **Environmental and Ecological Statistics.** v. 27, p. 1-40, 2020.

JENNER, E. **Inquiry into the Causes and Effects of the Variolae Vaccinae... known by the Name of the Cow Pox**. 2nd Edition. Ed. Ashely & Brewer, Springfield, 116 p., 1802.

JOHNSON, R. A., WICHERN, D. W. **Applied Multivariate Statistical Analysis.** 6th Edition. Ed. Pearson Education: Upper Saddle River, 2007. 800 p.

JUNIOR, D. F. C., CABRAL I, L. M. S.; Crescimento dos leitos de UTI no país durante a pandemia de Covid-19: desigualdades entre o público x privado e iniquidades regionais. **Physis: Revista de Saúde Coletiva.** v. 30, e300317, 2020.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v. 23. p. 187-200, 1958.

KAISER, H. F. A second generation little jiffy. **Psychometrika** v. 35, p. 401-415, 1970.

KAUFMAN, L., ROUSSEEUW, P.J. **Finding Groups in Data: An Introduction to Cluster Analysis.** 1st Edition. John Wiley & Sons, New York, 2005. 344 p.

KETCHEN, D. J., SHOOK, C. L. The application of cluster analysis in strategic management research: An analysis and critique. **Strategic Management Journal**. v. 17, p. 441 – 458, 1996.

MANLY, B. F. J. **Métodos Estatísticos Multivariados**: uma introdução. 3. Ed. Porto Alegre: Bookman, 2008. 229 p.

MAPBOX Choropleth Maps in Python. PLOTLY.COM. 2021. Disponível em:

https://plotly.com/python/mapbox-county-choropleth/. Acesso em: 15 abr. 2021

MILLIGAN, G. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. **Psychometrika**, v. 45, p. 325 – 342, 1980.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. 1ª Ed. Belo Horizonte: UFMG, 2005. 297 p.

MINISTÉRIO DA EDUCAÇÃO. **Manual de Conceitos e Nomenclaturas de Leitos Hospitalares.** Brasília, 2016. 24 p.

NASCIMENTO, M. L. F. A Multivariate Analysis on Spatiotemporal Evolution of Covid-19 in Brazil. **Infectious Disease Modelling** v. 5, p. 670-680, 2020*a*.

NASCIMENTO, M. L. F. Matemática das Epidemias e a Fé na Ciência. **Jornal A Tarde.** Salvador, 25 mar. 2020*b*. Disponível em: <https://atarde.uol.com.br/opiniao/noticias/2123996-matematica-das-epidemias-e-a-fe-na-ciencia>. Acesso em: 17 mar. 2021.

NICOLELIS M. A. L., RAIMUNDO R. L. G., PEIXOTO P. S., ANDREAZZI, C. S. How Super-Spreader Cities, Highways, Hospital Bed Availability, and Dengue Fever Influenced the Covid-19 Epidemic In Brazil. **Preprint**. 2020.

OLIVEIRA, W. K., DUARTE, E., FRANÇA, GVA, GARCIA, L. P. Como o Brasil pode deter a COVID-19. **Epidemiologia e Serviços de Saúde** v. 29, e202004, 2020.

PEARSON, K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. **Philosophical Transactions of the Royal Society of London** v. 187, p. 253 – 318, 1896.

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

REQUIA, W. J. *et al.* Risk of the Brazilian health care system over 5572 municipalities to exceed health care capacity due to the 2019 novel coronavirus (COVID-19). **Science of the Total Environment.** v. 730, e139144, 2020.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics.** v. 20, p. 53 – 65, 1987.

SCIPY cluster hierarchy linkage. SCIPY.ORG. 2021. Disponível em: < https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage>. Acesso em: 15 abr. 2021.

SKLEARN cluster KMEANS. SCIKIT-LEARN.ORG. 2021. Disponível em: < https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>. Acesso em: 15 abr. 2021.

SOUZA, M. L., MARQUES, T. V., AMORIM, M. M. Vulnerabilidade e incidência da COVID-19 no nordeste do Brasil através da análise de cluster. **Revista Brasileira de Geografia Médica e da Saúde.** v. 16, p. 232 – 248, 2020.

SOUZA, W. M. *et al.* Epidemiological and Clinical Characteristics of the COVID-19 Epidemic in Brazil. **Nature Human Behaviour** v. 4, 856 – 865, 2020*a*.

SOUZA, C. D. F., PAIVA, J. P. S., LEAL, T. C., SILVA, L. F., SANTOS, L. G. Spatiotemporal evolution of case fatality rates of COVID-19 in Brazil, 2020. **Jornal Brasileiro de Pneumologia**, v. 46, e20200208, 2020*b*.

WARD, J. H. Hierarchical Grouping to Optimize na Objective Function. **Journal of American Statistical Association**. v. 58, p. 236-244, 1963.

WERNECK, G. L, CARVALHO, M. S. A pandemia de COVID-19 no Brasil: crônica de uma crise sanitária anunciada. **Cadernos de Saúde Pública**, v. 36, p. 1-4, 2020.

WILLETT, J. **The Pioneering Life of Mary Wortley Montagu**. Pen & Sword History, London, 2021. 288 p.

World Health Organization. www.WHO.int, (2020).

ZHU, N. *et al.* A novel Coronavirus from Patients with Pneumonia in China, 2019**. The New England Journal of Medicine**. v. 382, p. 727-733, 2020.

APÊNDICES

APÊNDICE A – Base de dados utilizadas na construção do trabalho

A base de dados compilada, contendo as informações de todas as 18 variáveis analisadas para o período completo definido no presente trabalho, pode ser encontrada em: https://github.com/dscp3/Base_de_Dados_TCC.git. A Tabela A1 mostra os endereços eletrônicos que serviram de fonte de informações para cada variável.

Tabela A1. Relação de endereços eletrônicos das 18 variáveis que foram detalhadas na Tabela 3. A variável X_6 (densidade populacional) foi calculada por meio da razão entre população e área de cada UF (Eq. (3.1)), cujas informações constam em 2 *links* distintos, podendo ser visto abaixo. Já a variável X_{10} (% dependência SUS) foi calculada com base na Eq. (3.2).

Variáveis	Descrição	Base de dados	Endereço eletrônico
X_1	Quantidade de novos casos	Ministério da Saúde	https://covid.saude.gov.br/
X_2	Quantidade de novos casos por 100 mil habitantes	Ministério da Saúde	
X_3	Quantidade de novos óbitos	Ministério da Saúde	
X_4	Quantidade de novos óbitos por 100 mil habitantes	Ministério da Saúde	
X_5	População	TCU	PORTARIA Nº PR-254, DE 25 DE AGOSTO DE 2020 - PORTARIA Nº PR-254, DE 25 DE AGOSTO DE 2020 - DOU - Imprensa Nacional (in.gov.br)
X_6	Densidade populacional*	IBGE	PORTARIA Nº PR-254, DE 25 DE AGOSTO DE 2020 - PORTARIA Nº PR-254, DE 25 DE AGOSTO DE 2020 - DOU - Imprensa Nacional (in.gov.br) AR_BR_RG_UF_RGINT_RGIM_MES_MIC_MUN_2020.xls (live.com)
X_7	Quantidade de leitos de UTI (SUS)	Data SUS	TabNet Win32 3.0: CNES - Recursos Físicos - Hospitalar - Leitos Complementares - Brasil (datasus.gov.br)
X_8	Quantidade de leitos de UTI (Não SUS)	Data SUS	
X_9	Quantidade de leitos de UTI COVID-19	Data SUS	

Tabela A1. (Continuação)

Variáveis	Descrição	Base de dados	Endereço eletrônico
X_{10}	% Dependência do SUS*	ANS	https://www.ans.gov.br/images/stories/Materiais_para_pesquisa/Perfil_setor/sala-de-situacao.html
X_{11}	Quantidade de profissionais da saúde	Data SUS	TabNet Win32 3.0: CNES - Recursos Humanos - Profissionais - Indivíduos - segundo CBO 2002 - Brasil (datasus.gov.br)
X_{12}	Profissionais da saúde por 100 mil habitantes	Data SUS	
X_{13}	Número médio de moradores por domicílio	IBGE	Divulgação anual | IBGE
X_{14}	% pessoas com 60 anos ou mais	IBGE	
X_{15}	% pessoas com 18 anos ou mais diagnosticadas com hipertensão	IBGE	
X_{16}	% pessoas com 18 anos ou mais diagnosticadas com doenças cardiovasculares	IBGE	
X_{17}	Índice de Desenvolvimento Humano	IPEA	Radar IDHM: evolução do IDHM e de seus índices componentes no período de 2012 a 2017 (ipea.gov.br)
X_{18}	Doses aplicadas das vacinas	Ministério da Saúde	viz.saude.gov.br