UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE ENGENHARIA DE COMPUTAÇÃO

LUCAS KAMINSKI DE FREITAS

FERRAMENTAS PARA CONTEXTUALIZAÇÃO GEOGRÁFICA DE *OUTLIERS* EM CONJUNTOS DE DADOS MULTIDIMENSIONAIS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2021

LUCAS KAMINSKI DE FREITAS

FERRAMENTAS PARA CONTEXTUALIZAÇÃO GEOGRÁFICA DE *OUTLIERS* EM CONJUNTOS DE DADOS MULTIDIMENSIONAIS

TOOLS FOR GEOGRAPHICAL CONTEXTUALIZATION OF OUTLIERS IN MULTI-DIMENSIONAL DATA SETS

Proposta de Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Luiz Celso Gomes Junior
DAINF - Departamento Acadêmico de Informática -UTFPR

CURITIBA
2021

LUCAS KAMINSKI DE FREITAS

FERRAMENTAS PARA CONTEXTUALIZAÇÃO GEOGRÁFICA DE *OUTLIERS* EM CONJUNTOS DE DADOS MULTIDIMENSIONAIS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 07 de dezembro de 2021

_____

Prof. Dr. Luiz Celso Gomes Junior
Universidade Tecnológica Federal do Paraná

_____

Prof. Dr. José Antonio Buiar
Universidade Tecnológica Federal do Paraná

_____

Felipe Marx Benghi
Externo

CURITIBA

2021

RESUMO

FREITAS, Lucas. Ferramentas para Contextualização Geográfica de *Outliers* em Conjuntos de Dados Multidimensionais. 2021. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

A análise, contextualização e compreensão de *outliers* em *datasets* complexos, com muitos atributos heterogêneos, apresenta grandes desafios. Para o especialista realizando a análise, nem sempre é trivial identificar quais dados e atributos são relevantes para o problema em questão, mesmo com a utilização de técnicas de visualização de dados. Este problema é ainda mais desafiador em *datasets* que demandam a interpretação geográfica de *outliers*, como por exemplo: (i) dados meteorológicos; (ii) dados de censos demográficos; (iii) dados sócio-econômicos de diversos municípios. Este trabalho tem como objetivo propor ferramentas que simplifiquem a tarefa de interpretação e contextualização geográfica de *outliers*, através de visualizações criadas com o auxílio de algoritmos de *Outlying Aspect Mining*. Com essas ferramentas, pretende-se propiciar análises mais precisas, diretas e eficientes, permitindo que o especialista compreenda e contextualize os *outliers* com mais facilidade, sob uma perspectiva geográfica. Como caso de teste, serão utilizados os dados públicos de vacinação contra a Covid-19 no Brasil, disponibilizados pelo OpenDataSus.

Palavras-chave: Ciência de dados, *Outlier*, OAM, *Outlying Aspect Mining*, Visualização de dados.

ABSTRACT

FREITAS, Lucas. Tools for Geographic Contextualization of Outliers in Multidimensional Datasets. 2021. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

Analyzing, contextualizing and understanding outliers in complex datasets, with many heterogeneous attributes, presents big challenges. For the specialist performing the analysis, it is not always trivial to identify which attributes are relevant to the problem at hand, even with the usage of data visualization techniques. This problem is even more challenging in datasets that demand the geographic interpretation of outliers, such as: (i) meteorological data; (ii) demographic census data; (iii) socio-economic data from several cities. The present work proposes tools for simplifying the task of geographic contextualization and interpretation of outliers, through visualizations generated with the help of Outlying Aspect Mining algorithms. With these tools, it is expected that more accurate, direct and efficient analyses are possible, allowing the specialist to understand and contextualize outliers more easily, from a geographic perspective. As a test case, public data on vaccination against Covid-19 in Brazil, made available by OpenDataSus, will be used.
Keywords: Data science, Outlier, OAM, Outlying Aspect Mining, Data visualization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OAM	*Outlying Aspect Mining*
LOF	*Local Outlier Factor*
PCP	*Parallel Coordinates Plot*

SUMÁRIO

1 INTRODUÇÃO

Tem-se observado, nas últimas décadas, uma rápida evolução das tecnologias de coleta e de armazenamento de dados. Com isso, tem sido possível a criação de conjuntos de dados cada vez maiores e mais complexos. Infelizmente, as tecnologias de análise de dados não têm evoluído na mesma velocidade. Ao tentar obter uma maior compreensão sobre um conjunto de dados multidimensional, um especialista se depara com uma enorme quantidade de dados, muitas vezes heterogêneos e indevidamente organizados, e as ferramentas que possui a sua disposição são limitadas (KEIM et al., 2010).

Como exemplo, podemos citar o contexto de vacinação contra a Covid-19 no Brasil. Os dados públicos, disponibilizados pelo OpenDataSus[1], são bastante volumosos: no dia 27/06/2021, continha cerca de 88 milhões de registros, e mais de 30 colunas. Além disso, para que seja possível realizar algumas análises, é necessário cruzar estes dados com outras fontes (como por exemplo dados do IBGE), o que aumenta ainda mais a complexidade do cenário.

Dentre as dificuldades encontradas por um especialista ao lidar com conjuntos de dados multidimensionais, pode-se destacar a análise de *outliers*. Existem diversos algoritmos e ferramentas disponíveis na literatura para o problema de detecção de *outliers*. No entanto, a tarefa de compreender, contextualizar e interpretar os *outliers* detectados ainda é um ramo do conhecimento pouco explorado (VINH et al., 2016). Algumas técnicas foram propostas, dentre as quais podemos destacar os algoritmos de *Outlying Aspect Mining*, ou OAM, que buscam identificar os atributos mais relevantes do conjunto de dados em questão para a compreensão de um determinado *outlier*. No entanto, existem poucos estudos referentes a testes dessas técnicas em casos reais (VINH et al., 2016).

Ao lidar com grandes volumes de dados, é interessante que o especialista tenha à sua disposição algumas ferramentas visuais. Para análises de baixa dimensionalidade, técnicas como *box plots* e *scatterplots* são comumente utilizadas. Para dados multidimensionais, porém, não é trivial gerar visualizações adequadas. Gráficos de coordenadas paralelas, propostos por Inselberg (1985), são uma opção interessante, porém podem facilmente sobrecarregar o usuário com informações excessivas.

Em alguns casos, podem ser necessárias análises geográficas dos dados. No contexto da meteorologia, por exemplo, identificar *outliers* e contextualizá-los geograficamente pode ser útil para se prever alterações climáticas abruptas, ou até desastres naturais (EDSALL, 2003). No caso de dados demográficos e socio-econômicos, como dados de acesso à educação e de mortalidade infantil, a contextualização geográfica é fundamental para possibilitar análises de desigualdade social, através da comparação de regiões próximas, como por exemplo bairros vizinhos.

[1] https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao

Com este trabalho, pretende-se contribuir para este problema de contextualização de *outliers* em dados multidimensionais. Por se tratar de um problema bastante complexo, decidiu-se por focar em contextualizações de *outliers* em conjuntos de dados que apresentem um ou mais atributos de natureza geográfica, como municípios, países, ou pontos de interesse (PoI). Para esse fim, usa-se um algoritmo de OAM, e verifica-se sua utilidade para a geração de visualizações interativas que permitam ajudar um especialista na contextualização geográfica dos *outliers* em questão.

Como caso de teste, foi escolhido utilizar os dados públicos de vacinação contra a Covid-19 no Brasil (citado acima), por se tratar de um tema atual e bastante relevante, além de ser um conjunto de dados bastante rico. A proposta é a de se analisar municípios que apresentaram estratégias de vacinação destoantes com relação ao restante do país, através de uma contextualização geográfica, ou seja, de uma comparação destes municípios com seus vizinhos, ou com padrões regionais.

1.1 Objetivos

O objetivo deste trabalho é propor ferramentas visuais interativas capazes de auxiliar analistas na contextualização geográfica de *outliers* em conjuntos de dados multidimensionais.

Como objetivos secundários, pretende-se:

- Testar a utilidade de OAM em dados reais.
- Contribuir para a área de interpretabilidade de *outliers* em conjuntos de dados multidimensionais de forma geral.

1.2 Requisitos das Ferramentas

As operações realizadas pelas ferramentas propostas serão:

- A detecção de *outliers*;
- Aplicação de um algoritmo de OAM aos *outliers* detectados, para que seja possível identificar os atributos mais relevantes para a análise de cada um;
- Geração de visualizações interativas, levando em consideração os principais atributos identificados pelo algoritmo de OAM, com o objetivo de facilitar para o especialista a contextualização geográfica dos *outliers*.

1.3 Estrutura do Documento

No capítulo 2, são apresentadas as bases teóricas utilizadas para o desenvolvimento do trabalho. No capítulo 3, é apresentado o conjunto de dados usado ao longo do desenvolvimento deste trabalho. São descritos o processo de obtenção dos dados, uma análise exploratória realizada, e alguns filtros, agrupamentos e processamentos aplicados

sobre os dados. Já no capítulo 4, é descrito o desenvolvimento do trabalho em si, com a solução proposta para o problema, e uma descrição dos testes realizados. Por fim, o capítulo 5 apresenta as considerações finais do trabalho, e sugestões para trabalhos futuros.

2 FUNDAMENTOS E TRABALHOS RELACIONADOS

2.1 Detecção de *Outliers*

Dentro da estatística e da ciência de dados, *outliers* (ou anomalias) são definidos por Hawkins (1980) como observações que se destacam significativamente das demais, a ponto de ser possível supor que tenham sido geradas a partir de um mecanismo diferente.

A ocorrência de *outliers* em um conjunto de dados pode ser resultado de vários processos distintos. Pode ser, por exemplo, resultado de uma falha em algum sensor, ou de um erro humano. Também é possível que um *outlier* represente uma ação maliciosa, como ataques a uma rede de computadores, ou fraudes de cartões de crédito. Além disso, também podem resultar de falhas no funcionamento de equipamentos (como motores), ou até mesmo indicar, em exames médicos, problemas de saúde de um paciente. Em alguns casos, podem ser ocorrências completamente naturais, não causadas por uma ação externa: em um conjunto de dados contendo características físicas de um grupo de pessoas, pode haver indivíduos com altura ou peso destoantes do restante. Identificar *outliers* e compreender sua natureza são etapas fundamentais para se obter uma maior compreensão sobre o conjunto de dados, posteriormente auxiliando na tomada de decisões (WANG; BAH; HAMMAD, 2019; CHANDOLA; BANERJEE; KUMAR, 2009).

O processo de detecção de *outliers* apresenta diversos desafios. Determinar critérios claros para distinguir uma amostra normal de um *outlier* pode ser uma tarefa consideravelmente difícil, dependendo do conjunto de dados. A presença de ruído nos dados também pode dificultar a tarefa de identificar amostras que realmente sejam *outliers*. Além disso, é possível que o comportamento dos dados se altere com o tempo, e portanto os critérios de determinação de *outliers* também podem variar (CHANDOLA; BANERJEE; KUMAR, 2009).

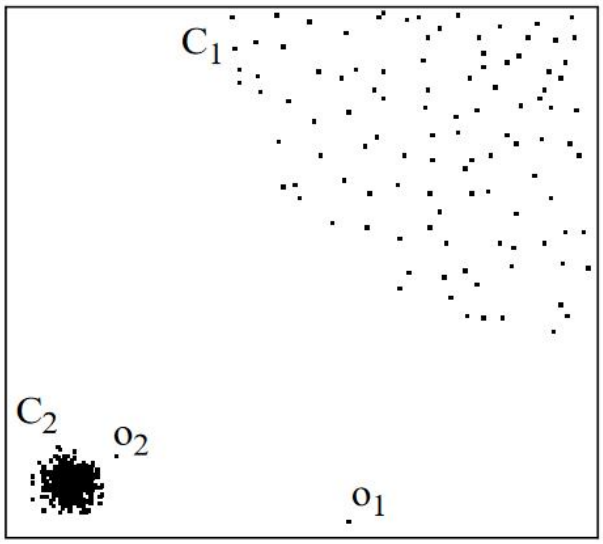
Diversas abordagens e técnicas foram propostas para o problema de detecção de *outliers* em um espaço multidimensional definido pelas variáveis das observações. Uma ideia bastante comum nestas abordagens é a de que amostras normais tendem a estar agrupadas em *clusters*, ou seja, estão a uma distância curta de suas amostras vizinhas. *Outliers*, por sua vez, não pertencem a nenhum *cluster*, e tendem a estar distantes deles. Essa ideia permitiu o desenvolvimento de diversos algoritmos de detecção de *outliers* baseados no cálculo da distância das amostras de um conjunto de dados a suas amostras vizinhas: quando essa distância for muito grande (de acordo com parâmetros pré-estabelecidos), essa amostra é classificada como um *outlier* (BREUNIG et al., 2000).

2.1.1 *Local Outlier Factor*

Breunig et al. (2000) cita dois problemas nestes algoritmos baseados em distância. Primeiro, estes algoritmos costumam funcionar de forma binária, ou seja, apenas classificam as amostras em "*outliers*" ou "não-*outliers*", e portanto não identificam o quanto uma amostra é diferente das demais.

O segundo problema nestes algoritmos, segundo Breunig et al. (2000), é que eles não são capazes de detectar corretamente *outliers* locais, que ocorrem nos casos em que existam *clusters* com densidades diferentes no conjunto de dados. A figura 1 ilustra um caso em que existem *clusters* com densidades diferentes: C_1 e C_2. O ponto o_2 é um *outlier* com relação ao cluster C_2, mas podem existir pontos em C_1 cuja distância a seu vizinho mais próximo é maior do que a distância entre o_2 e C_2. Nesse caso, algoritmos de detecção de *outliers* baseados apenas na distância entre os pontos não identificariam o_2 como *outlier* sem também identificar diversos pontos de C_1 como *outliers*.

Figura 1 – Conjunto de dados contendo dois *clusters* de densidades diferentes, C_1 e C_2



Fonte: (BREUNIG et al., 2000)

Para resolver estes problemas, Breunig et al. (2000) propôs o algoritmo *Local Outlier Factor*, ou LOF, que fornece um "coeficiente de anomalia" (*outlier factor*) para cada amostra no conjunto de dados, de acordo com o quanto cada uma delas se distancia de suas k vizinhas mais próximas. O algoritmo será detalhado a seguir.

Inicialmente, definimos a k-distância de uma amostra p como sendo a distância entre p e uma outra amostra q, denotada por $d(p,q)$, de forma que:

- No mínimo k amostras estejam a uma distância de p que seja menor ou igual a $d(p,q)$.
- No máximo $k-1$ amostras estejam a uma distância de p que seja menor do que $d(p,q)$.

Em termos mais simples, pode-se dizer que a k-distância de p é a distância de p a seu k-ésimo vizinho mais próximo. Com isso, pode-se definir a *reachability density*, ou RD, de uma amostra p com respeito a outra amostra o, de acordo com a equação 1.

$$RD_k(p,o) = max(k(p), d(p,o)) \tag{1}$$

Onde $k(p)$ é a k-distância de p e $d(p,o)$ é a distância entre p e o. Basicamente, a RD é a distância entre as duas amostras, exceto se o estiver a uma distância menor do que a k-distância de p. Nesse caso, RD é exatamente a k-distância de p.

A partir da RD, é possível calcular a densidade local de um ponto p com relação a seus k vizinhos próximos, denominada *local reachability density* ou LRD, de acordo com a equação 2:

$$LRD_k(p) = 1/\left(\frac{\sum_{o \in N_k(p)} RD_k(p,o)}{|N_k(p)|}\right) \tag{2}$$

Onde $N_k(p)$ é o conjunto de amostras que está a uma distância menor ou igual a $k(p)$ da amostra p. Dessa forma, a LRD é o inverso da média das RD de p a todas as amostras contidas em $N_k(p)$. Intuitivamente, a LRD pode ser considerada como sendo a densidade local de uma amostra, de acordo com as suas vizinhas. Quanto mais densa for a vizinhança de p, maior será a sua LRD.

O LOF é, por fim, calculado a partir da LRD de uma amostra e da LRD de seus vizinhos, conforme a equação 3:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{LRD_k(o)}{LRD_k(p)}}{|N_k(p)|} \tag{3}$$

Intuitivamente, pode-se dizer que o LOF compara a densidade local de uma amostra com a densidade local de suas vizinhas. No interior de um cluster, todas as amostras tendem a apresentar uma densidade local bastante parecida, então o LOF de todas elas estará próximo a 1. Já um *outlier* terá uma densidade local baixa, e terá um LOF alto quando comparado a seus vizinhos *inliers*.

2.1.2 *Isolation Forest*

O algoritmo *Isolation Forest* proposto por Liu, Ting e Zhou (2008), parte da ideia de criar árvores binárias através de divisões aleatórias no conjunto de dados. Para cada nó da árvore, são escolhidos aleatoriamente uma *feature* e um ponto de corte, e são gerados os nós filhos com o resultado dessa divisão.

As divisões continuam sendo executadas até que todas as amostras do conjunto tenham sido isoladas, cada uma em um nó folha da árvore. Espera-se que amostras *outliers*

sejam isoladas com menos divisões, e portanto considera-se que a proximidade de cada nó folha com relação à raiz da árvore seja o fator de anormalidade de cada amostra.

Como este algoritmo apresenta elementos aleatórios em sua execução, é necessário que o procedimento seja executado diversas vezes, e que seja calculada a média do fator de anormalidade de cada amostra. Experimentalmente, Liu, Ting e Zhou (2008) demonstram que 100 execuções costuma ser um bom valor para que o resultado do algoritmo apresente convergência do resultado.

Além disso, o algoritmo trabalha com sub-amostragem. Para cada árvore gerada, são consideradas apenas N amostras do conjunto de dados. Isso permite que essa técnica apresente um bom desempenho, independente do tamanho total do conjunto. Liu, Ting e Zhou (2008) recomendam o valor $N = 256$, determinado experimentalmente.

2.1.3 Z-score

Na estatística, o *z-score*, também chamado de *standard score*, ou escore padrão, é uma medida do quanto um determinado valor se destaca dos demais, de acordo com a média e o desvio padrão do conjunto sendo observado. A fórmula para o cálculo do *z-score* é apresentada na equação 4, onde x é o valor da amostra, $\overline{x}$ é a média de todas as amostras, e σ é o desvio padrão.

$$z = \frac{x - \overline{x}}{\sigma} \tag{4}$$

Ou seja, o *z-score* indica a quantos desvios padrão uma determinada amostra se destaca da média. Por se tratar de uma métrica bastante simples e intuitiva, o *z-score* é utilizado em muitos casos. Porém, trata-se de uma métrica que só pode ser utilizada em contextos unidimensionais. Dessa forma, em dados multidimensionais, essa métrica pode ser aplicada apenas para um atributo das amostras por vez (PECK; OLSEN; DEVORE, 2015).

2.2 Explicabilidade de *Outliers*

Em conjuntos de dados muito complexos, com muitos atributos, a interpretação dos *outliers* pode não ser trivial. Um algoritmo como o LOF é capaz de identificar que uma amostra se destaca das demais, mas não evidencia quais são as características (i.e. atributos, *features*) dessa amostra que a tornam diferente. Dessa forma, entender por que uma amostra foi classificada como *outlier* muitas vezes é um desafio a parte (SAMARIYA; MA; ARYAL, 2020).

Esse problema é conhecido como *explicabilidade de outliers*, *interpretação de outliers* ou *outlying aspect mining*, OAM. Uma definição mais formal do problema é: dado um conjunto de dados D, formado por um conjunto de atributos (dimensões) $C = c_1, c_2 ... c_n$,

e um *outlier* $q \in D$, busca-se encontrar quais são os subespaços de C nos quais a amostra q mais se destaca das demais (VINH et al., 2016).

As técnicas de explicabilidade de *outliers*, segundo Vinh et al. (2016), podem ser separadas em duas categorias: seleção de atributos e pontuação-e-procura (*score-and-search*).

Técnicas baseadas em *seleção de atributos*, como o método proposto por Micenková et al. (2013), modelam o problema de explicabilidade como um problema de seleção de *features* para classificação. A amostra q (o *outlier*) é identificada como pertencendo a uma classe denominada "positiva", e todo o restante do conjunto de dados é considerada como sendo uma outra classe, denominada "negativa". Com isso, é possível usar métodos já existentes na literatura para determinar o conjunto de atributos que mais evidenciam a separação entre essas duas classes.

A principal vantagem desse método é a eficiência computacional. Algoritmos baseados em seleção de atributos não costumam são muito custosos. Dentre as desvantagens, pode-se citar o enorme desbalanço entre as classes (pois a classe positiva é composta apenas por uma amostra, o *outlier*). Micenková et al. (2013) propõem métodos para equilibrá-las, através de *oversampling* da classe positiva e *subsampling* da classe negativa, mas Vinh et al. (2016) argumentam que em alguns casos essas abordagens podem não funcionar adequadamente para todos os subespaços. Além disso, essa técnica não identifica uma lista dos principais subespaços, apenas o principal, o que pode prejudicar a interpretação do *outlier*, e não trata aspectos combinatórios das variáveis.

Nos algoritmos de *pontuação-e-procura*, é usada uma função ω que retorna um grau de anormalidade para uma determinada amostra no conjunto de dados, ou seja, o quanto essa amostra se destaca das demais. Essa função ω é então calculada para o *outlier* q em todos os subespaços possíveis, separadamente. Os subespaços em que ω resultar nos maiores valores são considerados como mais relevantes para a explicabilidade do *outlier* (VINH et al., 2016).

A abordagem de pontuação-e-procura apresenta duas grandes dificuldades: o custo computacional, pois para conjuntos de dados com muitos atributos, a quantidade de subespaços que devem ser avaliados é enorme; e a dificuldade de se encontrar uma função ω que seja possível de se utilizar em subespaços de diferentes dimensionalidades, ou seja, que não privilegie subespaços com dimensionalidade maior ou menor (VINH et al., 2016; MICENKOVÁ et al., 2013).

O algoritmo LOF, apesar de ter sido proposto como um algoritmo de detecção de *outliers*, pode ser usado como função ω no contexto de OAM. Vinh et al. (2016) mencionam que, como este algoritmo retorna um valor próximo de 1 para registros que estejam no centro de um cluster, independente da dimensionalidade dos dados, isso faz com que o LOF seja uma escolha bastante razoável.

Vinh et al. (2016) propõem o algoritmo *isolation path*, ou iPath, que por sua vez

é baseado no algoritmo *isolation forest*, descrito na seção anterior. A principal diferença é que o iPath foi projetado para o contexto de OAM, e portanto busca apenas a pontuação referente ao *outlier* que está sendo analisado. Dessa forma, pode-se descartar os nós da árvore que não contêm o *outlier*. No entanto, é necessário executar o algoritmo para cada subespaço do conjunto de dados, de forma a identificar quais são os subespaços mais relevantes.

2.3 Visualização de dados

Técnicas de visualização de dados (e.g. gráficos) são fundamentais para a análise de dados e, mais especificamente, de *outliers*. Através de visualizações apropriadas, o ser humano consegue processar facilmente uma grande quantidade de informações, e identificar padrões nos dados de forma intuitiva. No entanto, para conjuntos de dados multidimensionais, gerar boas visualizações não é uma tarefa trivial. A seguir apresentaremos algumas técnicas de visualização que serão usadas ao longo deste trabalho.

2.3.1 PCP

Gráficos de coordenadas paralelas, também conhecidos como PCP, foram propostos por Inselberg (1985) como uma forma de se representar graficamente dados multidimensionais. Cada dimensão (*feature*, no contexto de ciência de dados) é representada por um eixo, geralmente vertical. Cada ponto no espaço n-dimensional em questão (cada registro no conjunto de dados) é representado por uma polilinha que cruza cada eixo no ponto referente ao valor do registro na dimensão em questão. (HEINRICH; WEISKOPF, 2013).

A figura 2 mostra um exemplo de PCP criado a partir do conjunto de dados IRIS[1]. Essa figura também ilustra uma técnica comumente utilizada na criação de PCPs: a utilização de cores para representar *clusters* distintos. Essa abordagem é bastante útil quando queremos comparar dados de *clusters* diferentes.

Também é possível usar cores para outras finalidades. Benghi (2020), por exemplo, propõe a utilização de um gradiente de cores para se representar a evolução temporal dos dados de um conjunto de dados multidimensional.

2.3.2 Mapas coropléticos

Mapas coropléticos são uma forma bastante popular de se exibir dados geográficos de maneira concisa e intuitiva. Tratam-se de mapas (geralmente mapas políticos, i.e. mapas que mostram as bordas de municípios, estados e países) em que são utilizadas cores para se representar um determinado valor associado a cada unidade geográfica. O mapa da

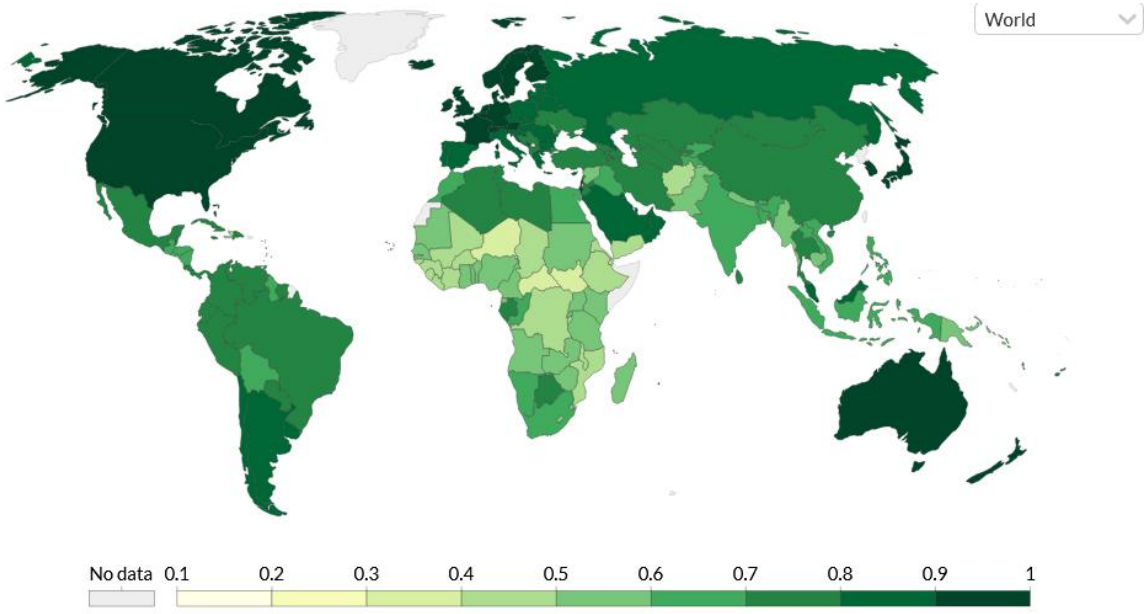
[1] *Dataset* com dados de 50 amostras de flores do gênero *Iris*. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Iris>

Figura 2 – Gráfico de coordenadas paralelas gerado a partir do conjunto de dados IRIS

Fonte: (PLOTLY, 2021)

figura 3, por exemplo, exibe o IDH de cada país usando uma escala de cores que vai de branco a verde escuro.

Figura 3 – Mapa de IDH



Fonte: (DATA, 2017)

2.4 Trabalhos Relacionados

Foram encontrados poucos trabalhos referentes a aplicações de algoritmos de OAM a casos reais, de forma a testar sua utilidade. Um exemplo é o trabalho de Benghi (2020), que experimenta o algoritmo iPath em dados de sensores de locomotivas, com o objetivo de verificar se o algoritmo ajuda na identificação das causas de falhas nestas locomotivas. A partir do resultado do algoritmo, o trabalho propõe algumas visualizações de forma a contextualizar temporalmente as falhas em questão.

Boukela et al. (2021) estudam a utilidade de técnicas de OAM a um contexto de segurança e confiabilidade de dispositivos IoT (internet das coisas). O algoritmo usado é do tipo pontuação-e-procura, usando um LOF modificado como função ω. Como caso de teste, são usados dois conjuntos de dados reais: um referente a vários sensores (temperatura, umidade, luz e tensão elétrica) localizados no Intel Berkeley Research Lab, e outro referente a conexões e acessos a uma rede de dados.

Com relação a visualizações de dados geográficos, pode-se citar o trabalho de Edsall (2003), que explora a utilidade de gráficos PCP para essa questão. No entanto, a proposta é mais abrangente do que a deste trabalho, não se restringindo à análise de *outliers*, e está mais focado na utilidade de gráficos PCP.

Como mencionado anteriormente, foram encontrados poucos trabalhos que abordam a aplicação de OAM a conjuntos de dados reais. O único trabalho encontrado até o momento que usa OAM especificamente com o intuito de auxiliar na geração de visualizações significativas é o trabalho de Benghi (2020). Com relação à contextualização geográfica de *outliers*, porém, não foram encontrados estudos que utilizem OAM.

3 ORIGEM E PROCESSAMENTO DOS DADOS

O conjunto de dados referente à vacinação contra a Covid-19 no Brasil foi obtido a partir do DATASUS (DATASUS, 2021) no dia 27/06/2021. Este é o principal conjunto de dados utilizado para o desenvolvimento do trabalho, contendo um registro para cada aplicação de vacina, desde o começo do ano. Dentre as informações disponíveis neste conjunto de dados, pode-se destacar as informações (anônimas) do cidadão que recebeu a vacina, como idade, grupo prioritário, município de nascimento; e as informações referentes à aplicação em si, como o nome do estabelecimento, nome da vacina, data de aplicação, município etc.

Neste trabalho, as informações usadas serão: data de aplicação da vacina, município, grupo prioritário da pessoa vacinada e número da dose (i.e. primeira dose ou segunda dose). Com estas informações, pretende-se fazer uma análise a respeito da estratégia de vacinação de cada município no contexto dos grupos prioritários, de forma a identificar possíveis municípios que adotaram estratégias diferentes dos demais.

Além disso, também foram utilizados dois conjuntos de dados auxiliares, para complementar as informações do conjunto principal. O primeiro também foi obtido no DATASUS (DATASUS, 2020), e apresenta dados da população de cada município do Brasil. Esses dados são baseados no censo de 2010, corrigidos de forma estimada para o ano de 2020. Já o segundo conjunto de dados é fornecido pelo IBGE (PRADO, 2021), e apresenta mais algumas informações sobre os municípios. No contexto deste trabalho, este último conjunto de dados foi utilizado apenas para fornecer os nomes dos municípios, uma vez que os dados do DATASUS os identificam apenas pelo código do IBGE; e a latitude e a longitude de cada município.

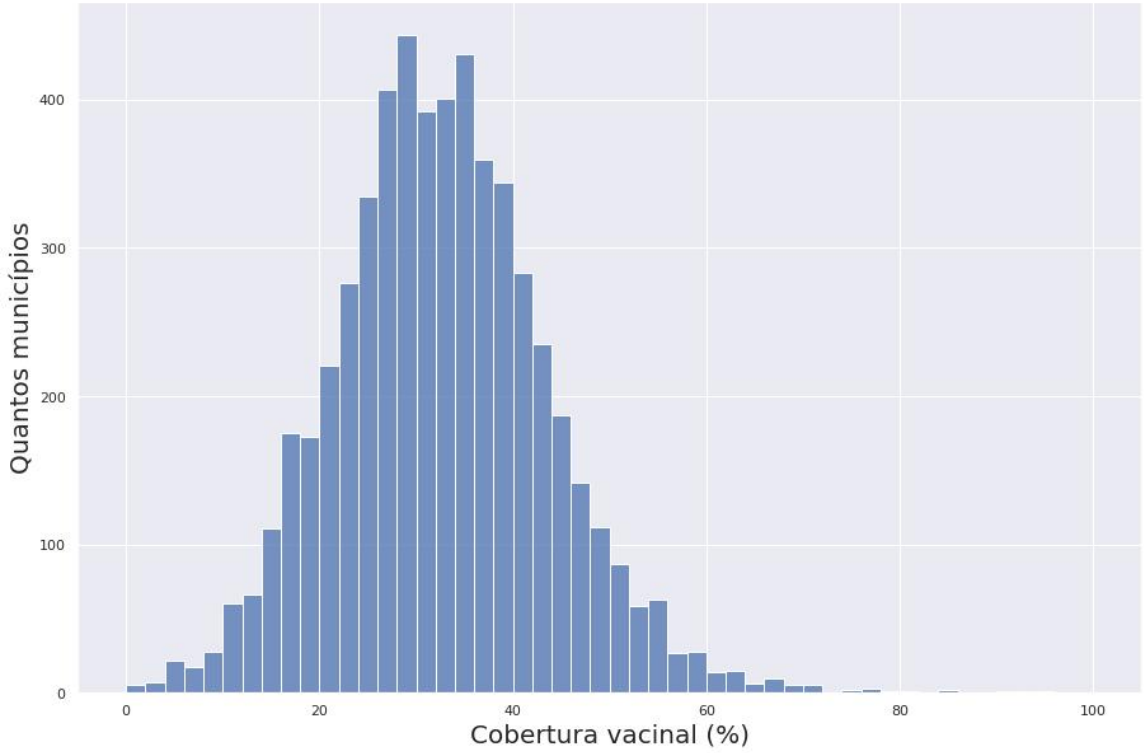
3.1 Análise Exploratória

Com o objetivo de se obter uma maior compreensão a respeito de possíveis padrões e tendências presentes no conjunto de dados de vacinação, foi realizada uma análise exploratória.

Inicialmente, foi gerado um histograma referente à cobertura vacinal em cada município do Brasil no dia 27/06/2021 (último dia disponível). Foram considerados apenas registros de primeira dose. Para fins de normalização, o total de vacinas aplicadas em cada município foi dividido por sua população total. Apesar de não ser a população exata, podendo gerar pequenos erros (pois os dados de população são apenas estimados), essa normalização ajuda a equilibrar as ordens de grandeza da população de cada município. O histograma é apresentado na figura 4.

Pode-se perceber, pelo histograma, que há uma grande variação na cobertura

Figura 4 – Histograma de cobertura vacinal por município (dia 27/06/2021)



Fonte: Autoria própria

vacinal entre os municípios. Apesar de a maior parte destes apresentar cerca de 30-40% da população vacinada com a primeira dose no dia 27/06/2021, alguns poucos já estavam em quase 80%, enquanto alguns ainda estão próximos a zero. Dos poucos municípios que apresentavam um percentual superior a 70%, todos são municípios pequenos: com exceção de Jardim do Seridó (RN), que tem uma população estimada de cerca de 12 mil habitantes, todos os demais municípios têm menos de 5 mil habitantes. Por isso, é possível que esses percentuais altos sejam decorrentes de erros na estimativa da população, como mencionado anteriormente, pois estes erros são mais evidentes em municípios menores.

Em seguida, foi feita uma comparação entre as regiões do Brasil no que diz respeito à velocidade de avanço na vacinação. Para este fim, foi partido do pressuposto que todas as regiões seguiram a vacinação de grupos prioritários na mesma ordem, definida no plano nacional de vacinação (BRASIL, 2021). Dessa forma, se uma região iniciou a vacinação de um grupo prioritário antes das outras, foi suposto que esta região estava mais adiantada do que a média nacional.

Como são muitos grupos prioritários (no conjunto de dados existem 87), a análise foi limitada apenas aos maiores (ou seja, os grupos com mais registros de vacinação), apresentados na tabela 1.

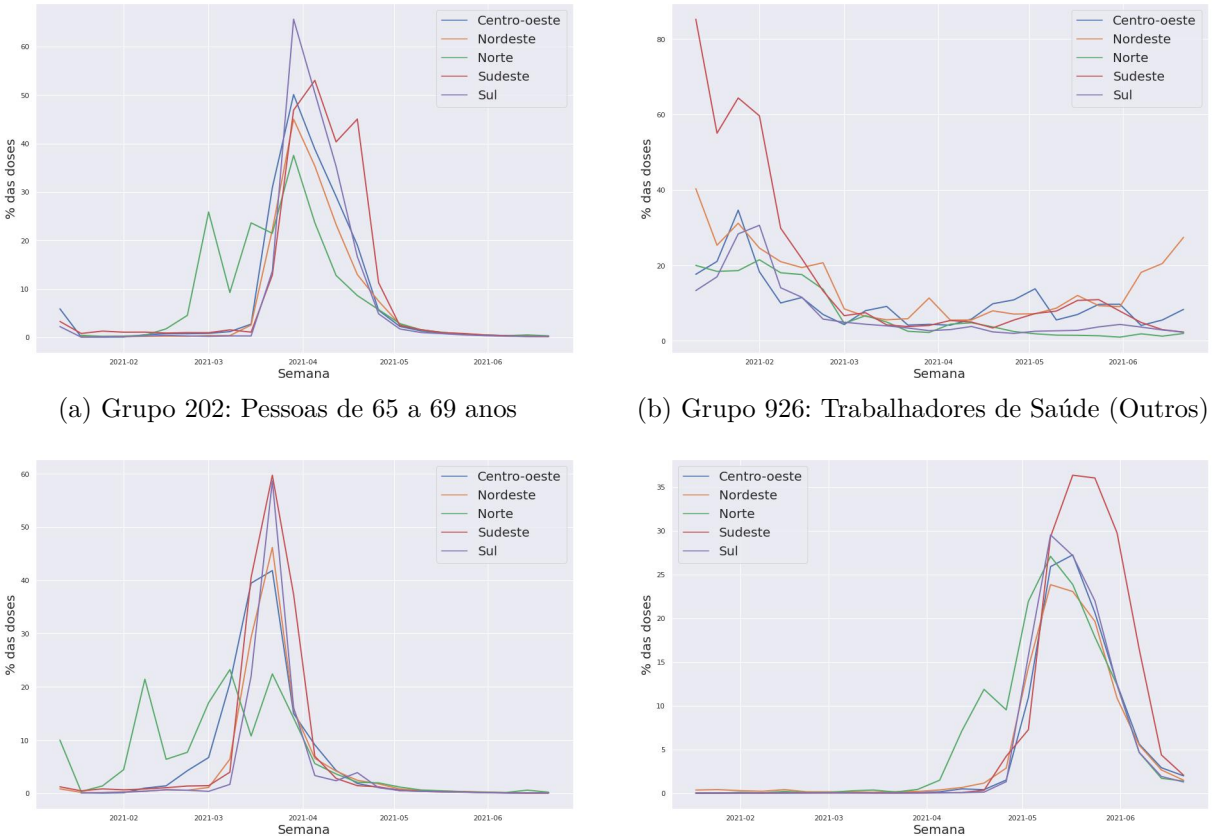
Na figura 5, são mostrados os gráficos referentes a quatro grupos prioritários. Em cada gráfico, há uma linha para cada região. O eixo x representa a passagem do tempo

Tabela 1 – Principais grupos prioritários, ordenados de acordo com a quantidade de registros no conjunto de dados

Código	Categoria	Grupo
201	Faixa Etária	Pessoas de 18 a 64 anos|Pessoas de 60 a 64 anos
202	Faixa Etária	Pessoas de 65 a 69 anos
926	Trabalhadores de Saúde	Outros
203	Faixa Etária	Pessoas de 70 a 74 anos
107	Comorbidades	Hipertensão de difícil controle

(agrupado por semanas), e o eixo y apresenta a porcentagem de aplicações de vacina neste grupo prioritário em relação ao total de aplicações na semana. Por exemplo, se numa determinada semana, o valor do eixo y for 50% numa região, significa que metade das vacinas aplicadas nesta região na semana foram neste grupo prioritário.

Figura 5 – Porcentagem dos grupos prioritários por semana, separados por região



(a) Grupo 202: Pessoas de 65 a 69 anos

(b) Grupo 926: Trabalhadores de Saúde (Outros)

(c) Grupo 203: Pessoas de 70 a 74 anos

(d) Grupo 107: Hipertensão

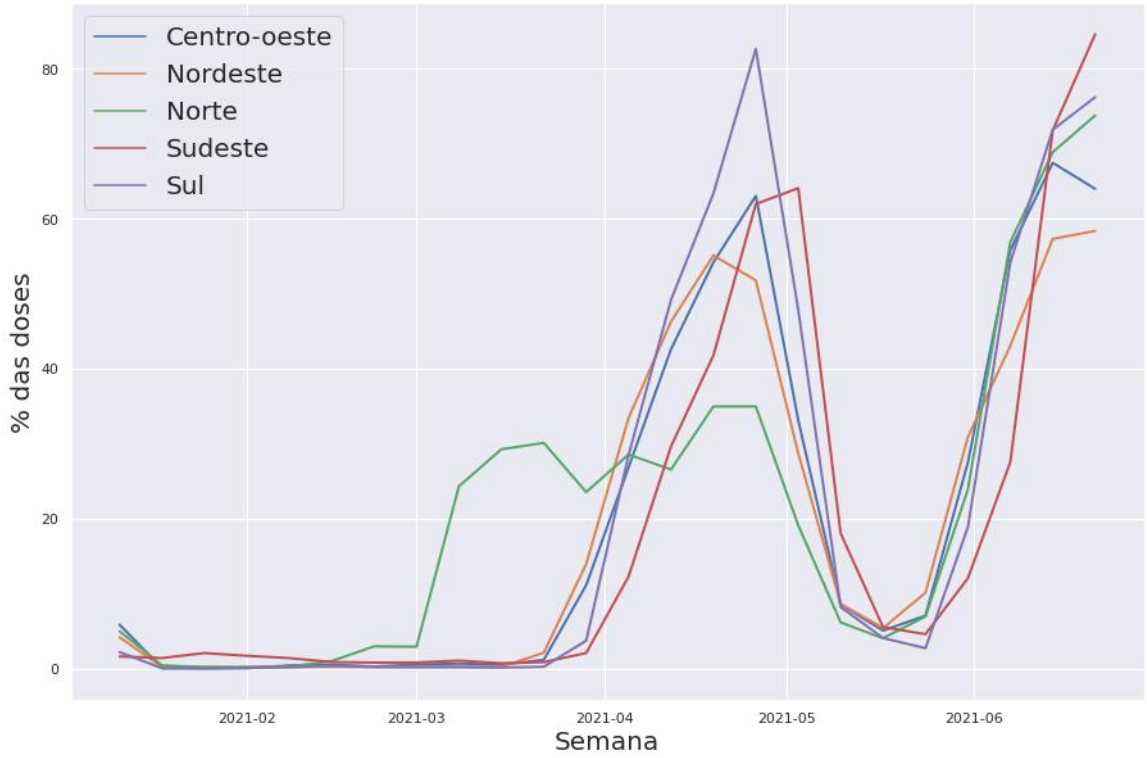
Fonte: Autoria própria

Podemos ver nos grupos 202, 203 e 107 que a região norte costuma começar um pouco mais cedo do que as demais regiões, o que indica que a vacinação tende a estar mais adiantada nesta região do que nas outras.

O grupo 201 (“Pessoas de 18 a 64 anos|Pessoas de 60 a 64 anos”) aparenta ser

um caso diferente dos demais, pois se trata de uma junção de dois grupos diferentes. Na figura 6, é apresentada a progressão temporal desse grupo em cada região, e é fácil notar que o gráfico apresenta dois "picos" muito distintos.

Figura 6 – Porcentagem de vacinas no grupo 201 por semana



Fonte: Autoria própria

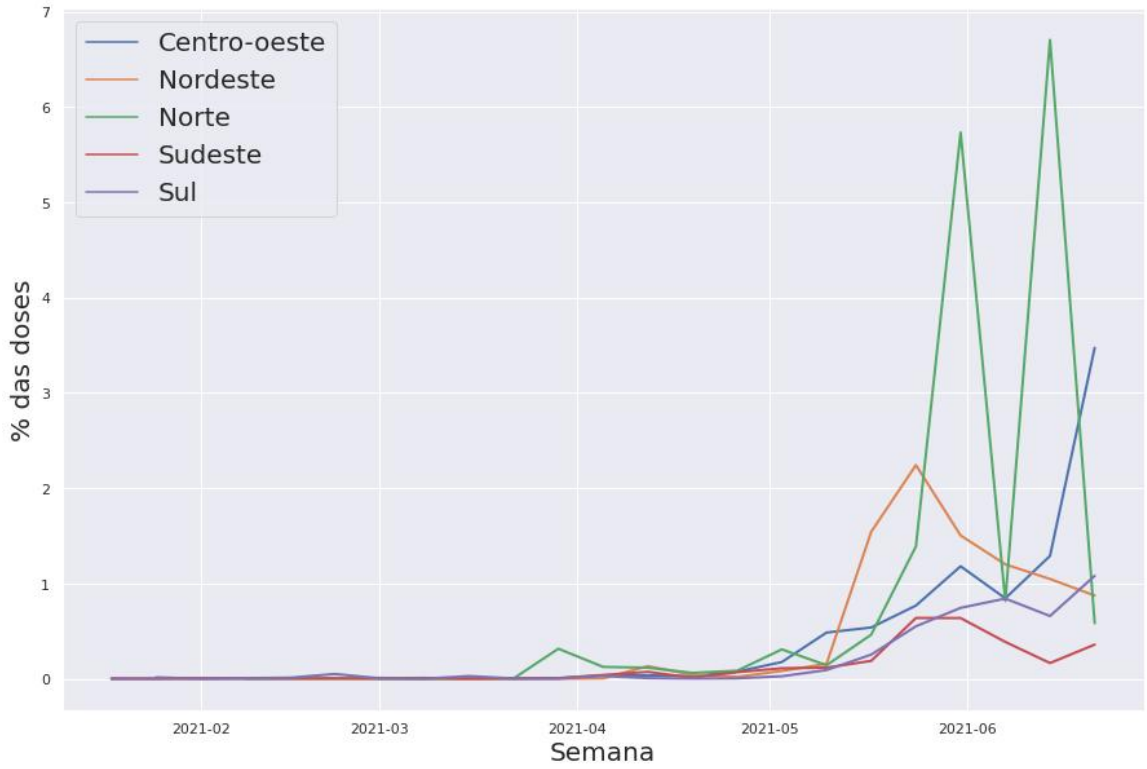
Por fim, foi feita uma análise (também por região) do grupo "Outros Grupos" (código 999999). Na figura 7, podemos ver que especialmente no mês de junho esse grupo esteve consideravelmente presente nos registros da região norte, chegando a representar quase 7% dos registros da região em uma determinada semana. Idealmente, este grupo deveria ter poucos registros, pois aparenta ser uma falha na categorização.

Através da análise exploratória, foi possível identificar que de um modo geral a vacinação ocorreu de forma semelhante em todas as regiões do Brasil. Apesar de a região norte estar um pouco a frente das demais (por determinação do Ministério da Saúde), a ordem de vacinação dos grupos prioritários foi a mesma no país inteiro, aproximadamente nas mesmas semanas. Dessa forma, espera-se que municípios individuais que adotarem estratégias de vacinação diferentes serão claramente classificados como *outliers*.

3.2 Processamento dos Dados

Para garantir uma maior qualidade nas análises posteriores, optou-se por realizar uma série de processamentos e agrupamentos nos dados. Foram descartados todos os

Figura 7 – Porcentagem de vacinas no grupo “Outros Grupos” por semana



Fonte: Autoria própria

registros de municípios com menos de 100 mil habitantes, pois observou-se que municípios menores tendem a apresentar uma maior variabilidade nos dados, relacionada a problemas de preenchimento e atualização dos dados. De forma semelhante, foram descartados os registros de aplicações de vacina com data anterior ao primeiro dia de março, pois até esse momento os registros eram muito escassos (diversos municípios nem sequer haviam iniciado a vacinação), e isso podia distorcer as análises.

Além disso, como o conjunto de dados utilizado foi obtido no dia 27/06, a aplicação de segundas doses havia recém começado no Brasil inteiro. Dessa forma, optou-se por remover todos os registros referentes à aplicação de segunda dose do conjunto de dados.

Também foi observado que no conjunto de dados existe um número considerável de grupos prioritários (87), e diversos deles apresentam características semelhantes. Por exemplo, existem 29 grupos referentes a “Trabalhadores da Saúde” (médicos, nutricionistas, veterinários, etc), e diversos grupos referentes a faixas etárias específicas (de 70 a 75 anos, de 75 a 80 anos, etc). Por esse motivo, para simplificar as análises futuras, foi decidido agrupá-los em grupos mais abrengentes. Os grupos escolhidos são apresentados na tabela 2. Vale ressaltar que optou-se por manter cada uma das comorbidades em grupos separados, pois considerou-se que algumas comorbidades são consideravelmente mais comuns do que outras.

Por fim, foi decidido utilizar um agrupamento por semana para cada município,

Tabela 2 – Grupos prioritários, após mesclagem de grupos semelhantes

Sem categoria	Forças de Segurança e Salvamento
Anemia Falciforme	Comunidade Quilombola
Neoplasias	Comunidade Ribeirinha
Diabetes Mellitus	Povos Indígenas
Pneumopatias Crônicas Graves	Trabalhadores da Educação
Doença Renal Crônica	Trabalhadores de Saúde
Doenças Cardiovasculares e Cerebrovasculares	Trabalhadores de Transporte
Hipertensão de difícil controle	Pessoas com Deficiência
Indivíduos Transplantados de Órgão Sólido	Pessoas em Situação de Rua
Obesidade Grave (Imc>=40)	Trabalhadores Portuários
Síndrome de Down	Funcionário do Sistema de Privação de Liberdade
Outros Imunossuprimidos	População Privada de Liberdade
Indivíduos Transplantados de Medula Óssea	Trabalhadores Industriais
Cirrose hepática	Trabalhadores de limpeza urbana
Faixa Etária (Não-Idosos)	Gestante
Faixa Etária (Idosos)	Puérpera
Forças Armadas (membros ativos)	Outros Grupos

pois imagina-se que um agrupamento por esse período de tempo seja capaz de reduzir significativamente o impacto de erros humanos (por exemplo, erros de digitação), permitindo uma análise mais precisa sobre o comportamento dos municípios.

4 IMPLEMENTAÇÃO

Durante o desenvolvimento deste trabalho, considerou-se a seguinte situação hipotética: um especialista foi encarregado de analisar os *outliers* de um conjunto de dados com muitos atributos. Os dados desse conjunto, porém, apresentam uma ou mais variáveis de natureza geográfica, como por exemplo o nome ou código de um município, ou de um país, ou até mesmo coordenadas (como latitude e longitude).

Nesse caso, uma contextualização geográfica de cada *outlier* pode ser útil para a análise do especialista. Caso cada registro seja referente a um país, por exemplo, pode ser interessante uma visualização em mapa, que possibilite a comparação de um país *outlier* com seus vizinhos. No caso de os dados apresentarem latitude e longitude, também pode-se comparar o *outlier* com os registros geograficamente mais próximos.

Nesta seção, são propostas algumas ferramentas para auxiliar o especialista na contextualização geográfica dos *outliers*. A seguir, será apresentada a arquitetura das ferramentas propostas.

4.1 Arquitetura Proposta

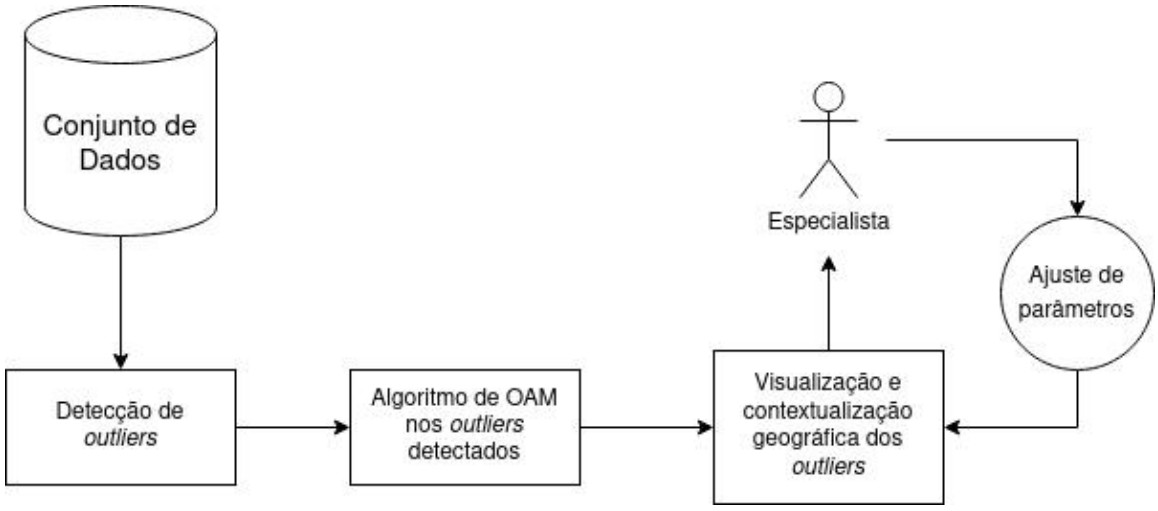
Na figura 8 é exibido um diagrama exemplificando a arquitetura proposta. A primeira etapa é aplicação de um algoritmo de detecção dos *outliers*. Isso pode acontecer de forma automática (por exemplo, diariamente ou semanalmente), sem a necessidade de intervenção do especialista. Em seguida, para cada um dos principais *outliers* detectados, é aplicado um algoritmo de OAM para identificar os subespaços mais relevantes. Por fim, com o resultado do algoritmo, são geradas algumas ferramentas visuais interativas, projetadas para a contextualização geográfica dos *outliers*.

Vale ressaltar que, apesar do foco deste trabalho ser a contextualização geográfica, acredita-se que arquiteturas semelhantes à da figura 8 podem ser úteis para outros problemas de interpretação e contextualização de *outliers*.

4.2 Caso de Uso

A seguir, será demonstrado um caso de uso da arquitetura proposta, usando o conjunto de dados de aplicação de vacinas contra a Covid-19. Para a realização deste trabalho, foi usado um computador com processador Intel Core i5-3330, com 8GB de memória RAM.

Figura 8 – Arquitetura proposta



Fonte: Autoria própria

4.2.1 Detecção de *Outliers*

Inicialmente foi projetado o modelo de detecção de *outliers*. O algoritmo escolhido foi o LOF com parâmetro k = 20, por sua simplicidade e por apresentar bons resultados de acordo com a literatura.

Sobre os dados, como comentado na seção 3.2, optou-se por utilizar um agrupamento por município e por semana. Ou seja, cada registro no conjunto de dados é referente a um município em uma determinada semana. Como *features*, foram selecionados os percentuais de aplicação de vacinas em cada um dos 34 grupos prioritários (listados na tabela 2 da seção 3.2). Por exemplo, se em uma determinada semana e município, todas as aplicações de vacinas foram em um mesmo grupo, então a *feature* referente a este grupo terá valor 100, e todos os demais grupos terão valor 0.

Além disso, também foram usadas como *features*: a semana, para que houvesse uma variável temporal no modelo; e a latitude e a longitude do município, para servirem como variável geográfica. Com isso, no total foram utilizadas 37 *features*. Todas as 37 *features* foram normalizadas de 0 a 100, inclusive o número da semana, a latitude e a longitude.

A tabela 3, com dados fictícios, ilustra o formato dos dados e as *features* selecionadas. As colunas de G1 a G34 representam os 34 grupos prioritários. As três últimas colunas representam a latitude, longitude e a semana.

4.2.2 Aplicação de OAM

Uma das principais dificuldades de se aplicar algoritmos de OAM do tipo pontuação-e-procura é o tempo necessário para a execução dos algoritmos. Dessa forma, foi escolhido o algoritmo *Isolation Path*, ou iPath, por se tratar de um algoritmo com custo computacional

Tabela 3 – *Features* selecionadas para a detecção de *outliers*

	G1	**G2**	**...**	**G34**	**Lat.**	**Lon.**	**Sem.**
Município 1 / Semana 1	0	100	...	0	37	50	10
Município 1 / Semana 2	50	50	...	0	37	50	20
Município 2 / Semana 1	0	100	...	0	85	19	10

relativamente baixo. Foi usada a implementação disponível na biblioteca desenvolvida por Faria e Colli (2021).

Foi escolhido um valor de 256 para o parâmetro de número de amostras, por ser o valor recomendado (VINH et al., 2016). Sobre o parâmetro de número de árvores, porém, apesar do valor recomendado ser 100, foi utilizado um valor de 50 neste trabalho. Isso introduz variabilidade nos subespaços encontrados, mas reduz ainda mais o tempo de execução do algoritmo.

O iPath foi executado sobre exatamente as mesmas *features* usadas para a detecção de *outliers*. Como 37 *features* é um número alto para se calcular todas as combinações de subespaços, optou-se por considerar apenas os subespaços com 3 dimensões ou menos. Isso resultou em um total de 8473 subespaços considerados pelo iPath, e um tempo de execução de aproximadamente 1 hora e meia para cada *outlier* analisado.

O resultado da execução do iPath é uma lista com os principais subespaços, em ordem de melhor pontuação, como exibido na tabela 4. Para algumas das visualizações descritas a seguir, em que foi necessário selecionar um número maior de *features* a partir do resultado do iPath, optou-se por uma abordagem *beam search*: percorrer a lista de subespaços, e ir incorporando as *features* na ordem de sua primeira aparição. Por exemplo, considerando os subespaços exibidos na tabela 4, considerou-se que as 4 principais *features* são A, B, C e D nessa ordem.

Tabela 4 – Exemplo de retorno do iPath

Ordem	**Subespaço**
1º	[A, B]
2º	[A, B, C]
3º	[B, D]

4.3 Ferramentas de visualização

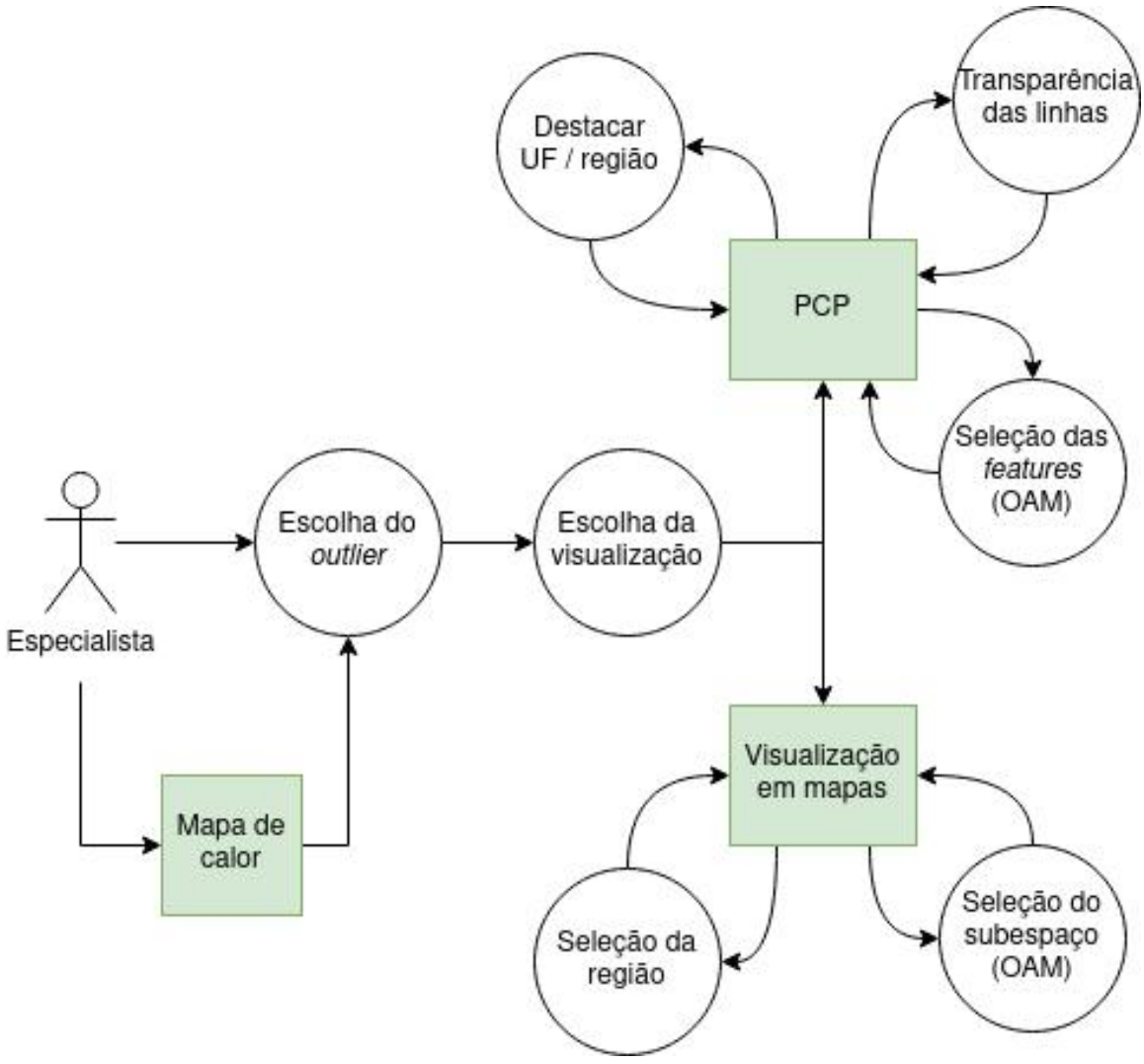
A seguir, serão propostas algumas ferramentas de visualização pensadas para o problema de contextualização geográfica dos *outliers*. Os requisitos básicos das ferramentas propostas são:

- Fornecer uma interface que possibilite uma visão geral dos principais *outliers*, a partir da qual o usuário possa escolher alguns casos para analisar individualmente;

- Tendo escolhido um *outlier* para analisar, deve-se oferecer duas possibilidades diferentes de visualização para o usuário escolher: PCP e visualização em mapas;
- Para cada uma das visualizações disponíveis, deve-se oferecer para o usuário algumas formas de interagir com os gráficos, para facilitar a análise. Nessa etapa, o usuário pode optar por selecionar os subespaços e atributos mais relevantes identificados pelo iPath.

Na figura 9, é apresentado o *workflow* proposto para o analista com as ferramentas em questão.

Figura 9 – *Workflow* proposto



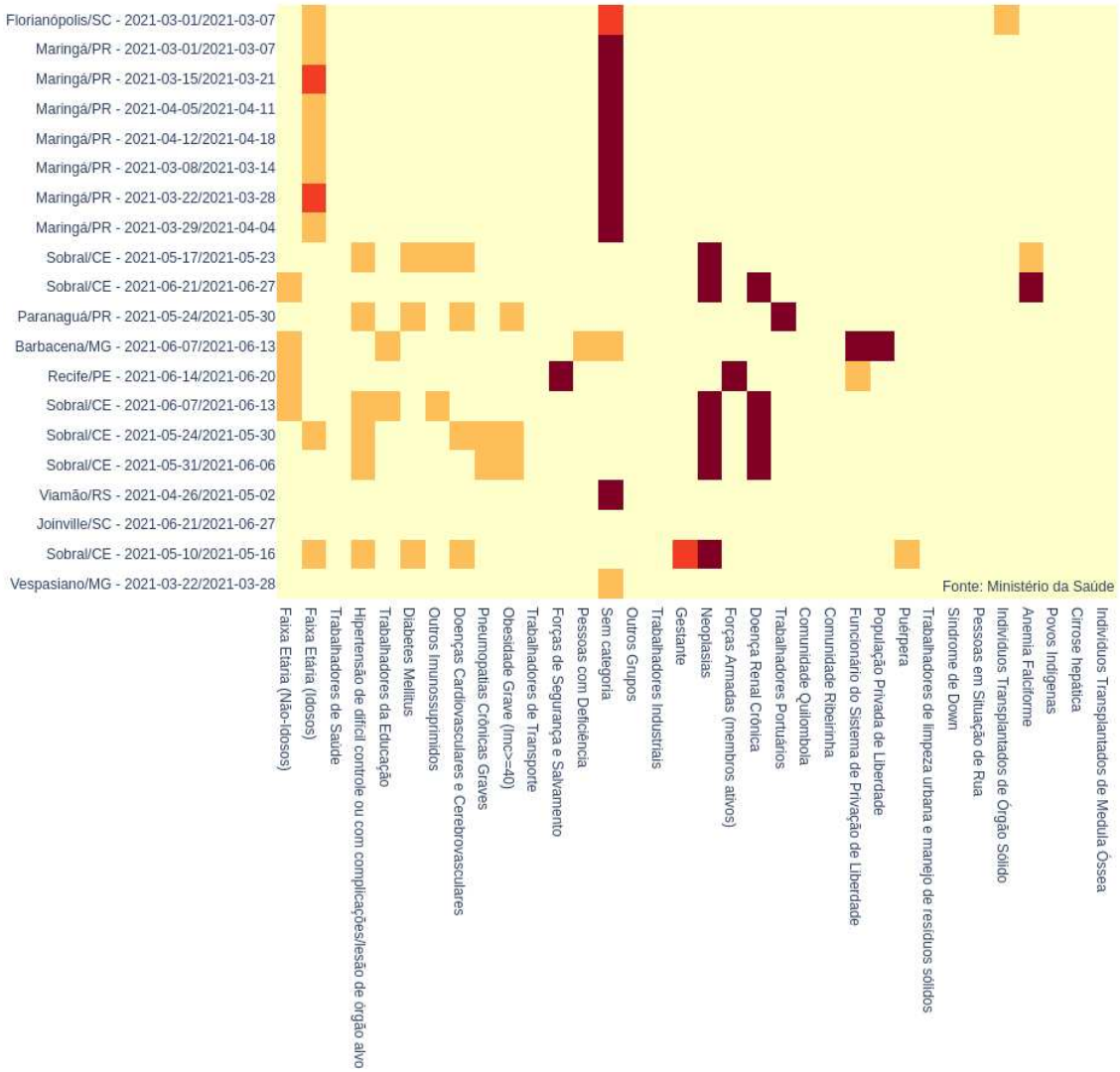
Fonte: Autoria própria

4.3.1 Mapa de calor

Para a primeira etapa, referente à visualização global dos *outliers*, propomos a criação de um mapa de calor baseado em z-score. Para cada semana, foi computado o

z-score separadamente para cada um dos grupos prioritários, comparando os municípios. Com esses valores, foi criado um mapa de calor como o mostrado na figura 10: cada linha representa um registro (município / semana), e cada coluna representa um grupo prioritário. A cor de cada célula representa o z-score: quanto maior o z-score, mais escuro. As linhas foram ordenadas de acordo com o LOF, de forma que os primeiros registros são os mais anômalos.

Figura 10 – Mapa de calor

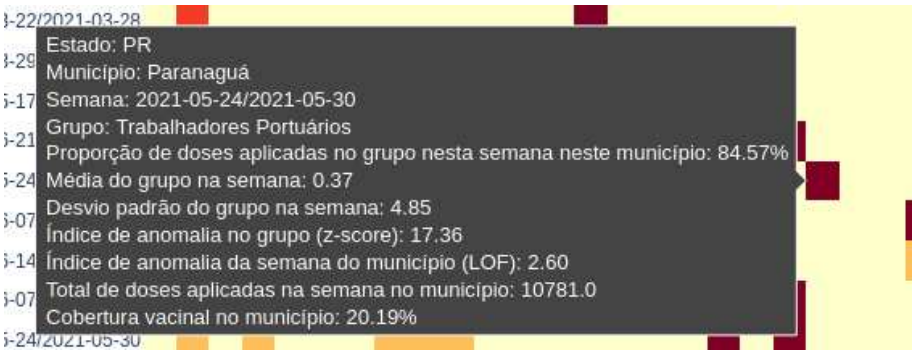


Fonte: Autoria própria

Uma outra funcionalidade presente no mapa de calor é exibida na figura 11: ao posicionar o cursor sobre qualquer célula, é exibido um *hover text* com algumas informações relevantes sobre o município, a semana e o grupo prioritário em questão.

O mapa de calor, por si só, não contribui para a contextualização geográfica dos *outliers*. Porém, essa visualização serve como um bom ponto de partida para a análise do especialista, pois apresenta de forma sintética os principais *outliers* identificados no

Figura 11 – Mapa de calor - *hover text*



Fonte: Autoria própria

conjunto de dados, em ordem de importância. Além disso, com essa técnica, já é possível obter certa compreensão sobre quais grupos prioritários (features) mais contribuíram para o estado anômalo de cada registro.

Seguindo o *workflow* da figura 9, o usuário deve selecionar um *outlier* para analisar. Isso pode ser feito diretamente a partir do mapa de calor. A partir disso, a ferramenta proposta oferece para o especialista duas formas de visualização interativas: gráficos de coordenadas paralelas (PCP) e visualizações em mapas. Essas visualizações serão descritas em mais detalhes a seguir.

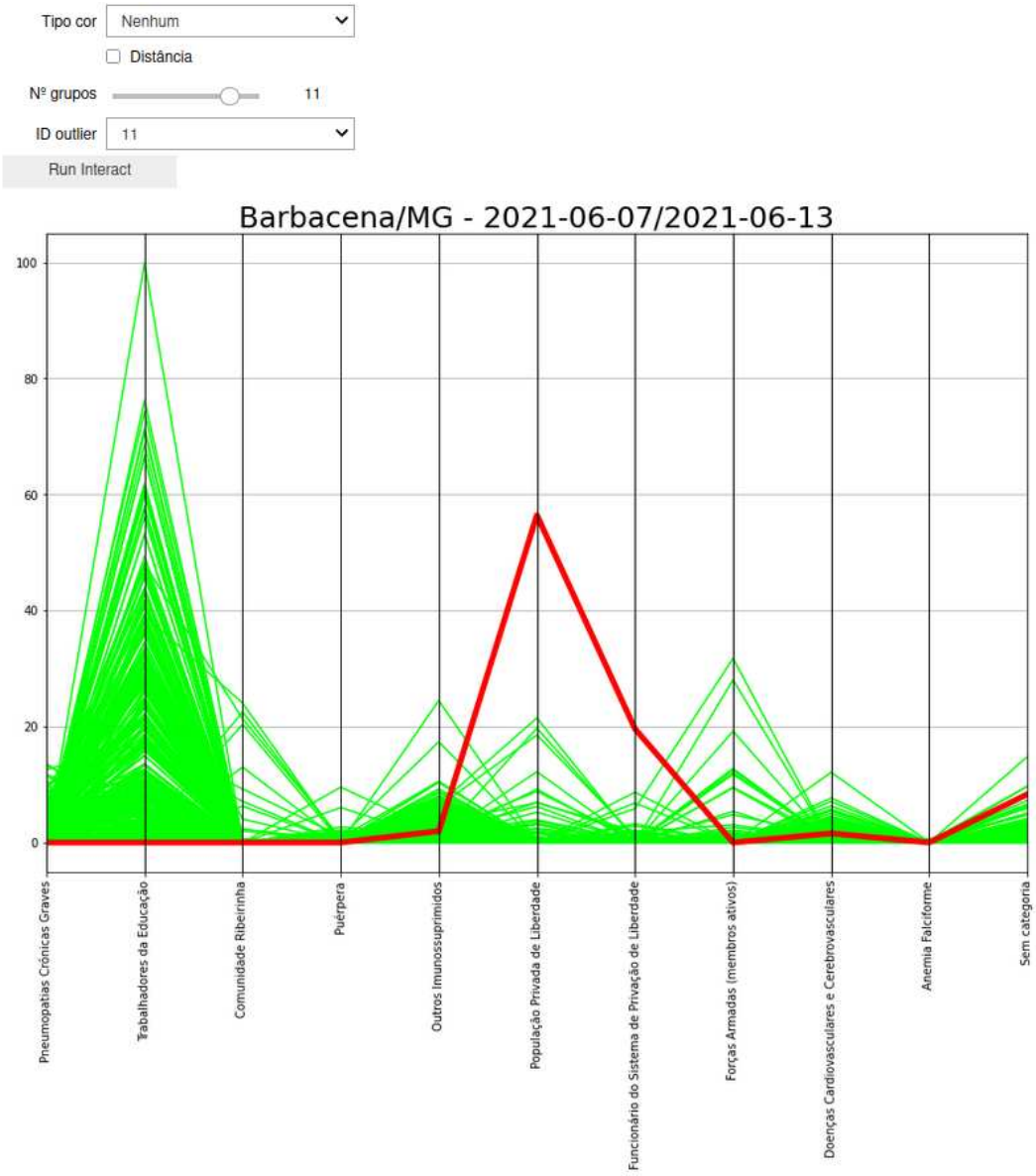
4.3.2 Visualização em PCP

A visualização em PCP foi desenvolvida a partir de ferramentas já contidas na biblioteca *pandas*[1], em python. A interface básica é exibida na figura 12, com um gráfico de exemplo. Nela, cada município é representado por uma linha verde. O município *outlier* é destacado dos demais, sendo representado por uma linha com a cor vermelha e uma largura maior. Cada grupo prioritário (*feature*) é um eixo do PCP. O gráfico da figura é referente ao município de Barbacena (*outlier*) na semana que foi do dia 07/06/2021 ao dia 13/06/2021, que foi a semana em que o município se destacou de acordo com o LOF. A interface oferece algumas ferramentas para interação, exibidas no topo da figura.

Caso existam muitas *features*, a visualização pode ficar poluída. Dessa forma, a ferramenta oferece ao usuário a opção de selecionar apenas as N principais *features* do *outlier*, definidas de acordo com a técnica descrita na seção 4.2.2 usando OAM. O usuário pode aumentar e diminuir o valor de N de forma interativa. Na figura 12, por exemplo, estão sendo exibidos apenas os 11 grupos mais relevantes, através da opção "Nº grupos". Além disso, como proposto por Benghi (2020), optou-se por deixar os grupos mais relevantes mais próximos ao centro. O grupo mais relevante é apresentado bem no meio do gráfico; o segundo e o terceiro mais relevantes ficam imediatamente ao redor do

[1] <https://pandas.pydata.org/docs/reference/api/pandas.plotting.parallel_coordinates.html>

Figura 12 – Interface básica da visualização PCP



Fonte: Autoria própria

primeiro, e assim por diante.

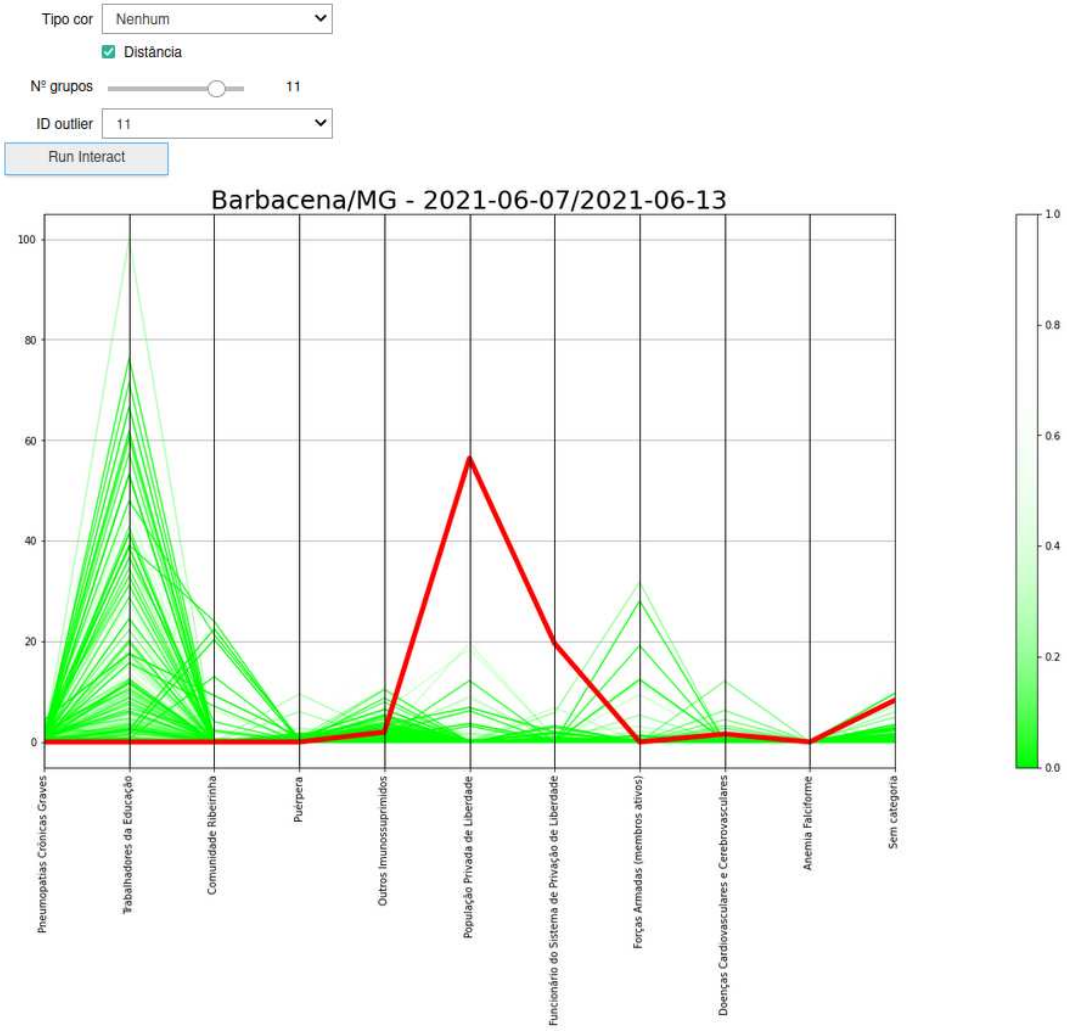
Essa visualização em PCP é bastante útil, pois permite a comparação de várias *features* simultaneamente, mesmo num contexto multidimensional, e muitas vezes torna evidente quais foram as *features* em que o *outlier* mais se destaca dos demais dados. Ainda na figura 12, pode-se observar que os grupos apresentados mais ao centro do gráfico realmente são grupos em que Barbacena apresentou um comportamento bem diferente dos demais municípios.

Para conjuntos de dados em que existam variáveis do tipo coordenada, essa ferramenta de visualização também oferece um *checkbox* com a opção de representar a distância de cada município com relação ao *outlier* através da transparência da linha:

quanto mais distante for o município, mais transparente é a linha. Com isso, o gráfico dá maior destaque aos vizinhos mais próximos do *outlier* sem desconsiderar o restante dos dados. Isso pode ser conveniente caso o usuário queira comparar o comportamento do *outlier* com seus vizinhos sem deixar de ter uma visão geral do comportamento dos dados.

Na figura 13, é exibido o mesmo gráfico da figura anterior, mas desta vez aplicando a técnica de transparência, através do uso da latitude e da longitude de cada município. Podemos observar que no grupo "População Privada de Liberdade", que foi o que o município de Barbacena mais se destacou, as amostras que apresentavam um valor próximo a 20% ficaram bem transparentes, então são municípios distantes de Barbacena. Isso demonstra que o comportamento de Barbacena nesse grupo é ainda mais destoante quando comparado apenas aos municípios próximos.

Figura 13 – PCP considerando as distâncias como transparência



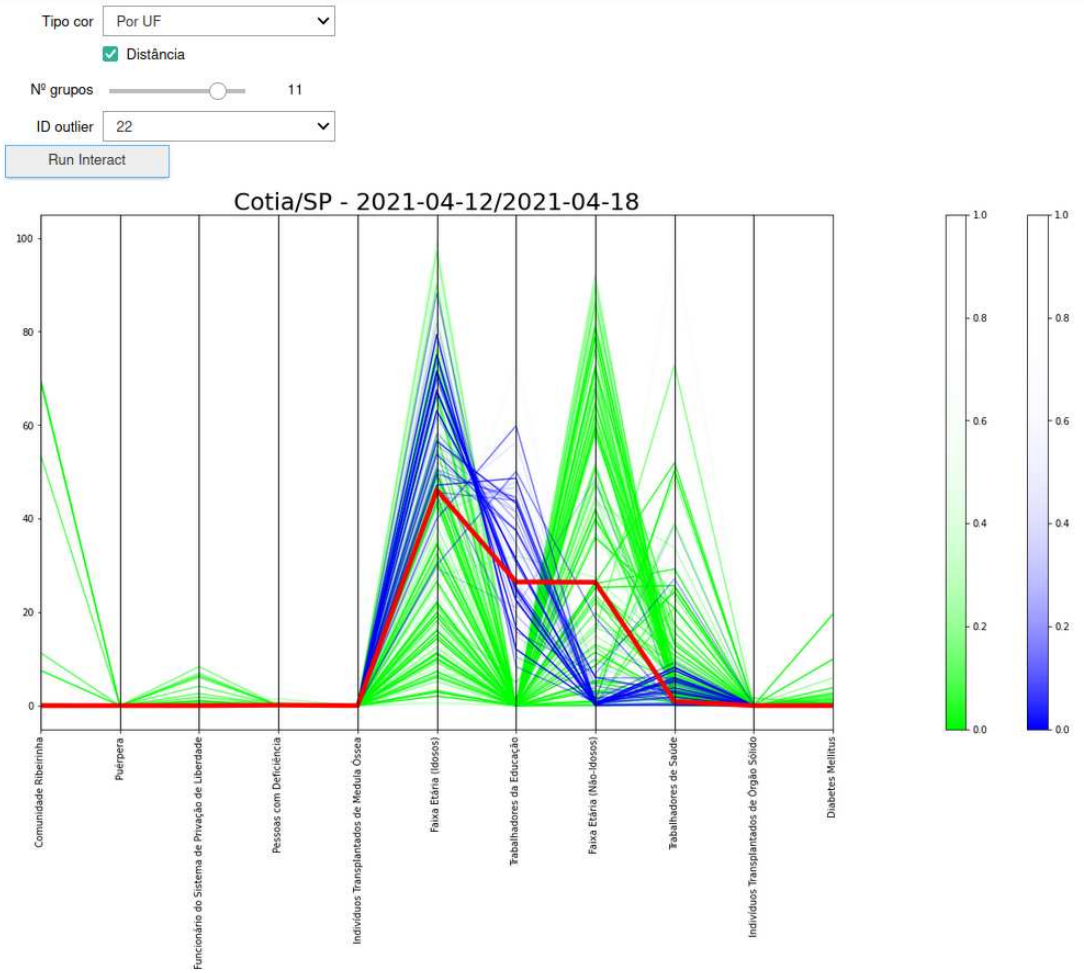
Fonte: Autoria própria

Uma outra opção disponível nesta visualização é usar uma cor diferente para destacar linhas de uma mesma região geográfica que o *outlier*. Através de um menu

dropdown, o usuário pode escolher se quer destacar registros do mesmo estado, região ou país que o *outlier*. A ideia dessa opção é possibilitar a contextualização do *outlier* com padrões regionais de forma separada dos padrões globais do conjunto.

Na figura 14 é demonstrado um exemplo de uso dessa opção, referente à semana do dia 12/04 ao dia 18/04. O *outlier* é o município de Cotia (SP), e todos os municípios do estado de SP foram desenhados com a cor azul. Pode-se perceber na figura que o município estava bem acima da média no grupo "Faixa Etária (Não-Idosos)" quando comparado ao restante do estado, mas não quando comparado ao restante do Brasil.

Figura 14 – PCP. Os municípios de SP foram desenhados com a cor azul



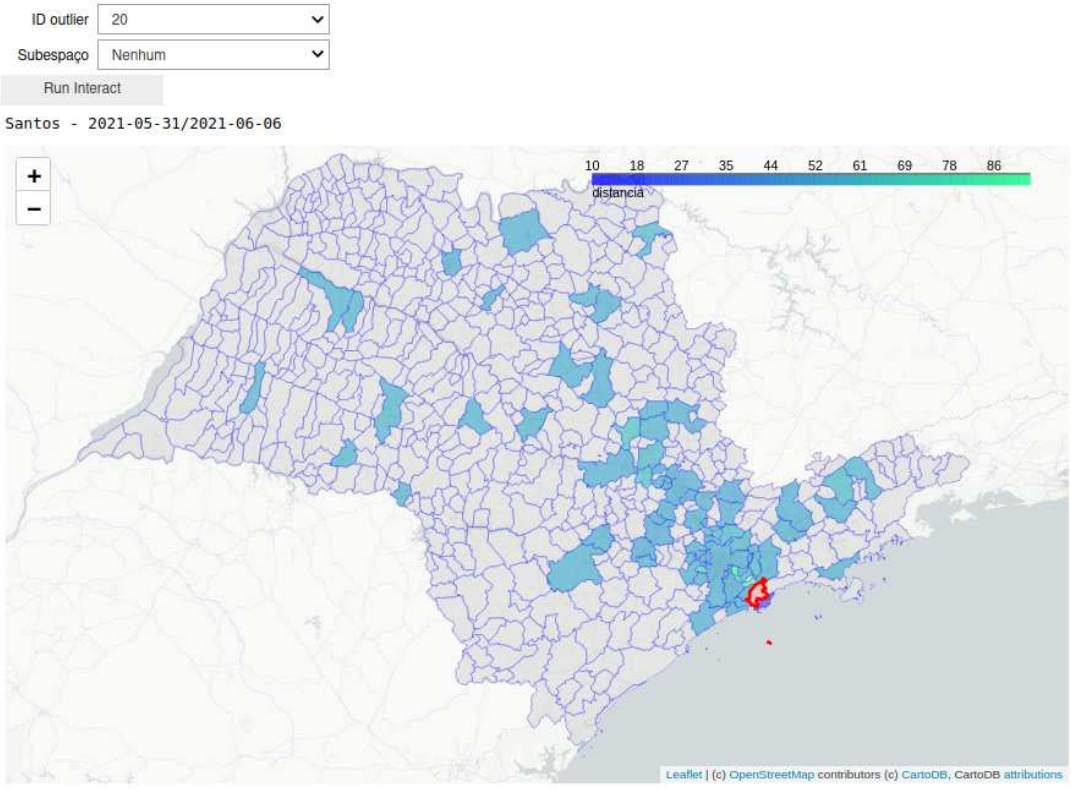
Fonte: Autoria própria

Vale notar que as opções podem ser usadas em conjunto. Ainda na figura 14, a técnica de transparência foi aplicada tanto nas linhas verdes quanto nas linhas azuis, como pode ser observado nas legendas no lado direito da figura. Essa mistura de técnicas pode ser interessante, pois podem existir linhas verdes mais próximas do que algumas linhas azuis, por exemplo, no caso de um município que esteja próximo da fronteira com outro estado.

4.3.3 Visualização em mapa

O segundo tipo de visualização oferecido pelas ferramentas propostas é uma visualização em mapas coropléticos interativos, como a apresentada na figura 15. Essa visualização foi desenvolvida com o uso da biblioteca *folium*[2], em python. O *outlier* novamente é destacado em vermelho. No caso da figura, referente à semana do dia 31/05 ao dia 06/06, o *outlier* é o município de Santos (SP). A cor de todos os demais municípios é calculada de acordo com a sua distância com relação ao *outlier* no espaço das features, ou seja, considerando as *features* como coordenadas para o cálculo da distância. Vale notar que, como mencionado na seção 3.2, municípios com menos de 100 mil habitantes foram filtrados do conjunto de dados e, portanto, são exibidos com a cor cinza.

Figura 15 – Mapa de distâncias no espaço das *features*



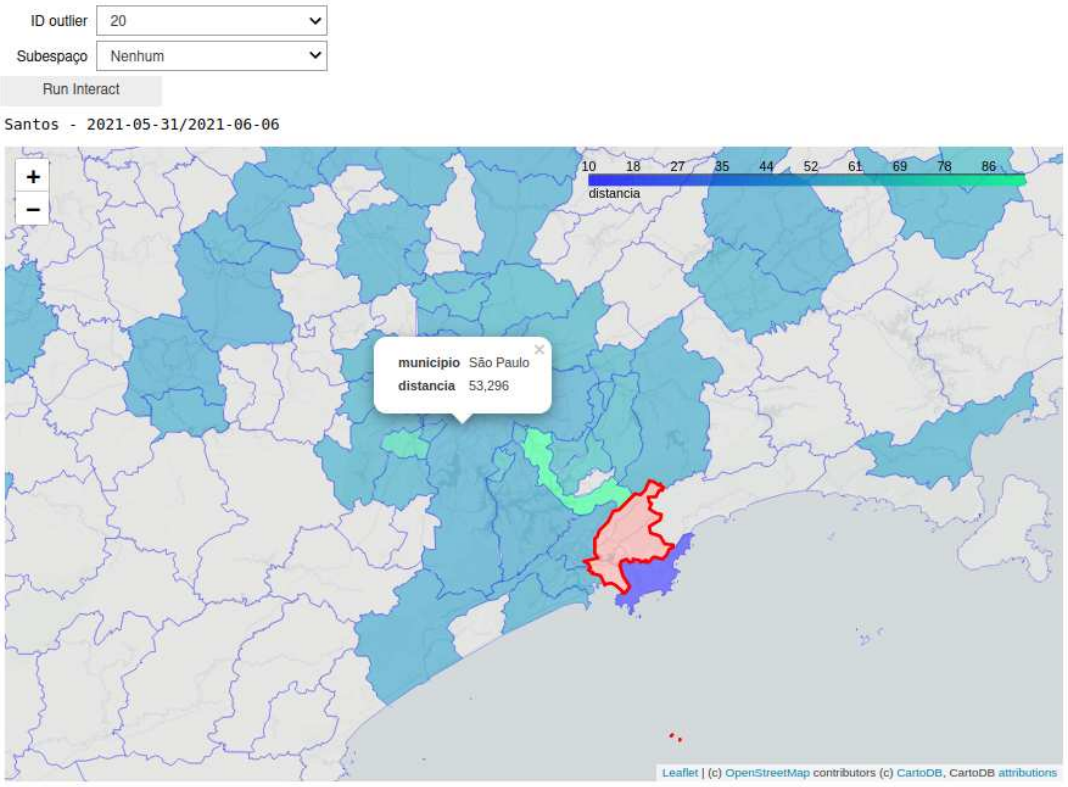
Fonte: Autoria própria

O usuário pode navegar livremente pelo mapa, dar zoom e colocar o cursor sobre os municípios para exibir mais informações, como exibido na figura 16.

A ideia desta visualização é que municípios com cor mais escura tenham um comportamento parecido com o *outlier* sob análise. No caso da figura 16, podemos ver que o município de Guarujá, imediatamente ao sul de Santos, está bem próximo de Santos no espaço das features, o que indica que esses dois municípios tiveram uma distribuição semelhante de grupos prioritários na semana.

[2] <https://python-visualization.github.io/folium/>

Figura 16 – Exemplo de interatividade na visualização em mapa coroplético



Fonte: Autoria própria

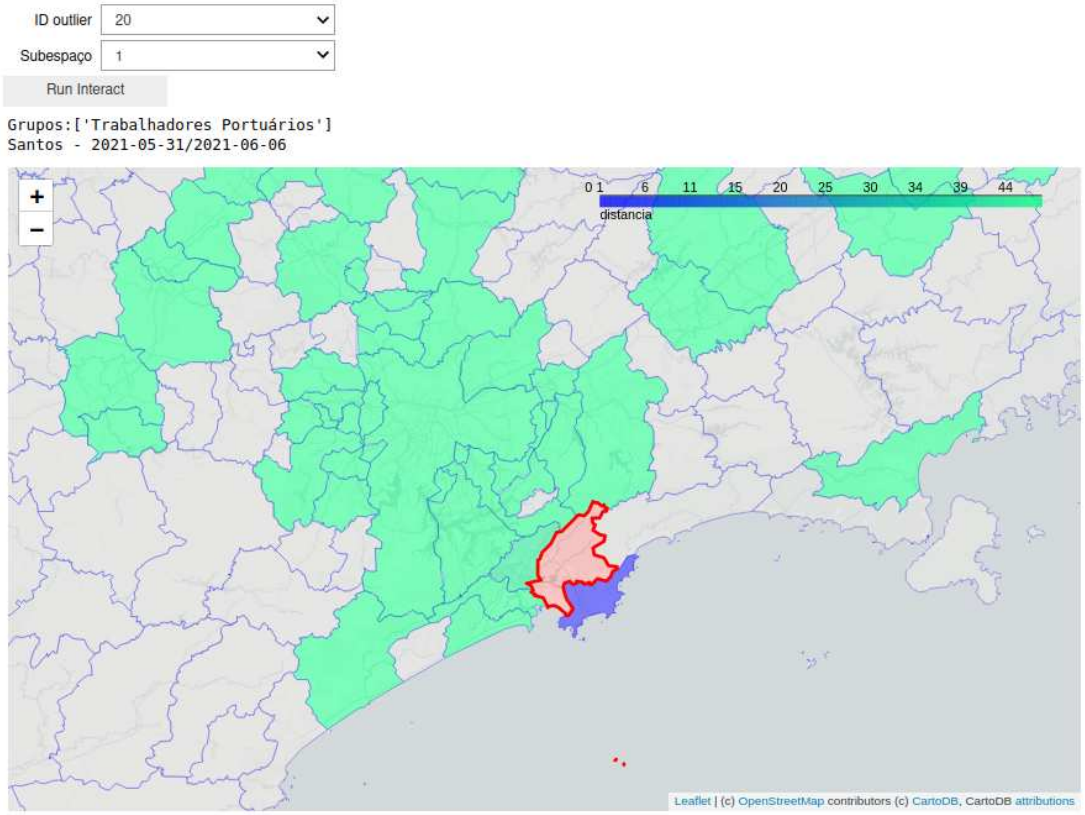
A visualização em mapas coropléticos é interessante por ser bastante intuitiva e de simples entendimento. Porém, uma de suas limitações é que só pode ser exibida uma cor por unidade geográfica, e portanto a análise simultânea de várias *features* não é possível. Para contornar essa necessidade, a ferramenta também oferece um menu *dropdown* para o usuário, listando os principais subespaços identificados pelo iPath. Caso o usuário escolha alguma das opções desse menu, as distâncias serão recalculadas considerando apenas os grupos que formam o subespaço selecionado.

Na figura 17 é exibido um exemplo de uso dessa opção: foi selecionado o principal subespaço identificado pelo iPath, que consiste apenas no grupo prioritário "Trabalhadores Portuários". Então apenas essa *feature* foi usada como coordenada para o cálculo das distâncias. Podemos ver que a similaridade entre Santos e Guarujá fica ainda mais destacada, uma vez que apenas estes dois municípios apresentaram alto valor no grupo "Trabalhadores Portuários" em São Paulo na semana em questão. Todos os demais municípios estão com cores claras, o que indica que esse subespaço de fato isola bem o *outlier* sob análise.

4.4 Considerações

Baseando-se nas imagens apresentadas ao longo deste capítulo, acredita-se que as ferramentas propostas e as visualizações apresentadas podem facilitar consideravelmente

Figura 17 – Mapa de distâncias no espaço das *features*, considerando apenas *features* do subespaço com melhor score de acordo com o algoritmo iPath



Fonte: Autoria própria

a tarefa de contextualização geográfica de *outliers* em conjuntos de dados complexos. Gráficos de coordenadas paralelas, como os apresentados nas figuras 12, 13 e 14 são uma das principais maneiras de se visualizar dados multidimensionais, e as técnicas propostas por este trabalho apresentam potencial de tornar a visualização mais intuitiva. No entanto, seria interessante realizar testes com usuários reais.

Além disso, acredita-se que existe bastante potencial na utilização de técnicas de OAM para gerar visualizações em mapas coropléticos, como o apresentado na figura 17. Esse tipo de visualização é bastante popular, por ser naturalmente intuitivo, mas em conjuntos de dados muito complexos, a sua utilidade é limitada. Acredita-se que o uso dos subespaços de OAM cumpriu seu objetivo, pois aliviou a restrição desta visualização de só apresentar um único valor por município.

Como a arquitetura proposta é aplicável a qualquer conjunto de dados com variáveis geográficas, pretende-se eventualmente incorporar uma implementação das ferramentas apresentadas à biblioteca de OAM de Faria e Colli (2021).

5 CONCLUSÕES

5.1 Conclusões gerais

Com relação à contextualização geográfica de *outliers*, as ferramentas propostas demonstraram alto potencial para auxiliar a análise de um especialista, através de interações e visualizações intuitivas e de fácil compreensão. As figuras apresentadas ao longo deste trabalho são capazes de sintetizar uma grande quantidade de informações de uma forma que pode ser compreendida sem grandes esforços.

Como mencionado anteriormente, a aplicação de OAM a dados reais é um ramo pouco explorado. Com este trabalho, foi possível contribuir para esta área, destacando que o resultado do algoritmo iPath foi de grande utilidade para se gerar visualizações significativas e concisas nos dados de vacinação contra a Covid-19 no Brasil.

Espera-se que, de um modo geral, este trabalho sirva como contribuição no âmbito de análises de dados multidimensionais, e que motive trabalhos futuros na área, para que cada vez mais ferramentas sejam desenvolvidas para este tipo de problema.

5.2 Trabalhos futuros

Seria interessante, em um trabalho futuro, testar as ferramentas propostas em outros conjuntos de dados com características geográficas. Além disso, seria desejável realizar testes com usuários reais, para averiguar o verdadeiro potencial das ferramentas.

Além disso, considera-se acrescentar mais interatividade nas visualizações. Através do uso da biblioteca *plotly*, é possível gerar PCPs altamente interativos, nos quais o próprio usuário pode ordenar os eixos ou destacar linhas para analisar. Nas visualizações em mapa, apesar de já existirem elementos para interação, estes poderiam ser mais aprofundados. Por exemplo, clicar num município para analisá-lo.

Como mencionado anteriormente, pretende-se eventualmente incorporar as visualizações propostas neste trabalho à biblioteca de OAM desenvolvida por Faria e Colli (2021). Com isso, acredita-se que seria mais fácil que outros pesquisadores dessem continuidade ao trabalho atual, ou desenvolvessem trabalhos semelhantes.

Por fim, seria interessante incorporar às ferramentas algoritmos de clusterização geográfica e de detecção de mudanças significativas em padrões geográficos usando Local Moran[1].

[1] <https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html>

Referências

BENGHI, F. M. **Visual analytics e outlying aspect mining: contextualização de anomalias considerando questões temporais e multidimensionais**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2020. Citado 3 vezes nas páginas 18, 20 e 32.

BOUKELA, L. et al. A modified lof-based approach for outlier characterization in iot. **Annals of Telecommunications**, Springer, v. 76, n. 3, p. 145–153, 2021. Citado na página 20.

BRASIL. Ministério da saúde. **Plano nacional de operacionalização da vacinação contra a covid-19**, 2021. Citado na página 22.

BREUNIG, M. M. et al. Lof: identifying density-based local outliers. In: **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2000. p. 93–104. Citado 2 vezes nas páginas 13 e 14.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 41, n. 3, p. 1–58, 2009. Citado na página 13.

DATA, O. W. I. **Human Development Index (HDI)**. 2017. <https://ourworldindata.org/human-development-index>. Acesso em 2021-11-14. Citado na página 19.

DATASUS. **População Residente - Estudo de Estimativas Populacionais por Município, Idade e Sexo 2000-2020 - Brasil**. 2020. <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?popsvs/cnv/popbr.def>. Acesso em 2021-09-19. Citado na página 21.

DATASUS. **Campanha Nacional de Vacinação contra Covid-19 - Conjuntos de Dados**. 2021. <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao>. Acesso em 2021-09-19. Citado na página 21.

EDSALL, R. M. The parallel coordinate plot in action: design and use for geographic visualization. **Computational Statistics & Data Analysis**, Elsevier, v. 43, n. 4, p. 605–619, 2003. Citado 2 vezes nas páginas 10 e 20.

FARIA, R.; COLLI, T. **Biblioteca em python para explicabilidade de anomalias**. Monografia (Trabalho de Conclusão de Curso - Em andamento) — Universidade Tecnológica Federal do Paraná, 2021. Citado 3 vezes nas páginas 29, 38 e 39.

HAWKINS, D. M. **Identification of outliers**. 1st. ed. [S.l.]: Springer, 1980. v. 11. ISBN 978-94-015-3994-4. Citado na página 13.

HEINRICH, J.; WEISKOPF, D. State of the art of parallel coordinates. In: **Eurographics (State of the Art Reports)**. [S.l.: s.n.], 2013. p. 95–116. Citado na página 18.

INSELBERG, A. The plane with parallel coordinates. **The visual computer**, Springer, v. 1, n. 2, p. 69–91, 1985. Citado 2 vezes nas páginas 10 e 18.

KEIM, D. et al. Mastering the information age: solving problems with visual analytics. Goslar: Eurographics Association, 2010. Citado na página 10.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. **2008 eighth ieee international conference on data mining**. [S.l.], 2008. p. 413–422. Citado 2 vezes nas páginas 15 e 16.

MICENKOVÁ, B. et al. Explaining outliers by subspace separability. In: **2013 IEEE 13th International Conference on Data Mining**. [S.l.: s.n.], 2013. p. 518–527. Citado na página 17.

PECK, R.; OLSEN, C.; DEVORE, J. L. **Introduction to statistics and data analysis**. 5th. ed. [S.l.]: Cengage Learning, 2015. v. 1. ISBN 978-130526581-3. Citado na página 16.

PLOTLY. **Parallel Coordinates Plot**. 2021. <https://plotly.com/python/parallel-coordinates-plot/>. Acesso em 2021-11-13. Citado na página 19.

PRADO, K. S. do. **Dados relacionados aos municípios brasileiros: código IBGE, nome, capital, código UF, UF, estado, latitude, longitude, código SIAFI, DDD e Fuso Horário.** 2021. <https://github.com/kelvins/Municipios-Brasileiros>. Acesso em 2021-09-19. Citado na página 21.

SAMARIYA, D.; MA, J.; ARYAL, S. **A Comprehensive Survey on Outlying Aspect Mining Methods**. 2020. Citado na página 16.

VINH, N. X. et al. Discovering outlying aspects in large datasets. **Data mining and knowledge discovery**, Springer, v. 30, n. 6, p. 1520–1555, 2016. Citado 3 vezes nas páginas 10, 17 e 29.

WANG, H.; BAH, M. J.; HAMMAD, M. Progress in outlier detection techniques: A survey. **IEEE Access**, v. 7, p. 107964–108000, 2019. Citado na página 13.