UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ALEXSANDRA GOMES DE LIMA

Análise de Agrupamento Espacial para Dados Criminais

Recife

2023

ALEXSANDRA GOMES DE LIMA

Análise de Agrupamento Espacial para Dados Criminais

Tese apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito final para obtenção do título de Doutor em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Raydonal Ospina Martínez

Coorientador: Cristiano Ferraz

Recife

2023

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

L732a Lima, Alexsandra Gomes de
Análise de agrupamento espacial para dados criminais / Alexsandra Gomes de Lima. – 2023.
75 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2023.
Inclui referências.

1. Estatística aplicada. 2. Agrupamento. 3. Área integrada de segurança. I. Ospina Martínez, Raydonal (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2023 - 73

ALEXSANDRA GOMES DE LIMA

ANÁLISE DE AGRUPAMENTO ESPACIAL PARA DADOS CRIMINAIS

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutora em Estatística.

Área de concentração: Estatística Aplicada.

Aprovada em: 27 de fevereiro de 2023.

BANCA EXAMINADORA

_____
Prof. Dr. Raydonal Ospina Martínez
Presidente/Orientador, UFPE/DE

_____
Prof. Dr. Vinícius Quintas Souto Maior
Examinador Interno, UFPE/DE

_____
Prof. Dr. Marcel de Toledo Vieira
Examinador Externo, DE/UFJF

_____
Prof. Dr. José Luiz de Amorim Ratton Júnior
Examinador Interno, UFPE/DS

_____
Prof. Dr. Pedro Luis do Nascimento Silva
Examinador Externo, ENCE

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ser meu guia e dar sentido às minhas escolhas.

Ao meu pai Arnaldo e meus irmãos Juscelino e Vanda, que sempre investiram em meus estudos e a minha mãe, Lúcia (*in memorian*), por ser meu anjo da guarda.

A Renato, meu maior entusiasta, por seu amor, incentivo e paciência.

Ao meu orientador, o professor Raydonal, pela excelente orientação, pelas palavras de motivação e compreensão nos meus lapsos de desânimo e pela confiança no meu potencial.

Ao professor Cristiano, pela coorientação e sugestões ao longo de todo este trabalho, por sempre transmitir tranquilidade e bom humor.

Aos membros da banca avaliadora, os professores Pedro, Marcel, José Luis e Vinícius pela disponibilidade e contribuições tão significativas.

Aos meus colegas dessa jornada, Elisângela, Alison, Lucas Silva, Lucas David, César, Cris, Adenice, Jodavid, Joas, Thalyta, Diego, Codjo, André e Ivangillis, pelas boas risadas e troca de conhecimento durante as disciplinas.

Aos professores do Programa de Doutorado em Estatística, pela competência e dedicação e a Universidade Federal de Pernambuco que fez parte de toda a minha vida acadêmica, desde a graduação até o doutorado.

Por fim, agradeço à CAPES, pela bolsa de estudos, que possibilitou minha dedicação exclusiva ao doutorado.

RESUMO

Esta tese apresenta um estudo sob a perspectiva da análise de agrupamento envolvendo informação espacial e dados criminais. Foram considerados cinco métodos de agrupamento: K-Means, PAM (Partitioning Around Medoids), VNSKMED (Variable Neighborhood Search for K-Medoides), Ward-Like e SKATER (Spatial K'luster Analysis by Tree Edge Removal), além disso, foram propostas alterações nos algoritmos Ward-Like e SKATER modificando a estrutura de pesos e o processo de partição dos grupos usando a distância Gower, nomeados de Ward-Like.New e SKATER.New, respectivamente. Os métodos foram comparados, por meio de três índices de validação: índice Calinski-Harabasz, índice Dunn e índice Davies-Bouldin. Para a análise dos algoritmos, foram utilizados dados de 2007 a 2015 sobre a ocorrência de crimes nos bairros da cidade de Recife envolvendo as classificações das Áreas Integradas de Segurança. Os algoritmos permitiram explorar os padrões relacionados aos crimes, possibilitando mapeá-los em grupos de bairros da capital pernambucana. Os resultados apontam que as modificações Ward-Like.New e SKATER.New produziram os melhores resultados, sendo o SKATER.New o recomendado.

Palavras-chaves: agrupamento; área integrada de segurança; Recife; gower; SKATER.

ABSTRACT

This doctoral dissertation presents a study from the perspective of cluster analysis involving spatial information and criminal data. Five clustering methods were considered: K-Means, PAM (Partitioning Around Medoids), VNSKMED (Variable Neighborhood Search for K-Medoides), Ward-Like and SKATER (Spatial K'luster Analysis by Tree Edge Removal). proposed changes in the Ward-Like and SKATER algorithms by modifying the weight structure and the process of considering groups using the Gower distance, named Ward-Like.New and SKATER.New, respectively. The methods were compared using three validation indices: Calinski-Harabasz index, Dunn index and Davies-Bouldin index. For the analysis of algorithms, data from 2007 to 2015 on the occurrence of crimes in the neighborhoods of the city of Recife were used, influenced by the classifications of the Integrated Security Areas. The algorithms made it possible to explore patterns related to crimes, allowing them to be mapped in agglomerations of neighborhoods in the capital of Pernambuco. The results showed that the Ward-Like.New and SKATER.New modifications produced the best results, with SKATER.New being recommended, attesting to better performance in the use of criminal data and in the formation of new Integrated Security Areas.

Keywords: clustering; integrated security area; Recife; gower; SKATER.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Agrupar é um processo em que classificamos objetos a partir de propriedades que possuem em comum. A análise de agrupamento consiste em uma série de técnicas usadas para agrupar objetos, considerando as semelhanças ou discrepâncias entre eles. O termo objeto tem caráter geral e, dependendo da aplicação, um objeto pode assumir inúmeras formas como por exemplo, uma pessoa, uma localidade, um item de produção industrial, um animal, entre outros.

As áreas de análise de agrupamento e análise de classificação muitas vezes se confundem, porque ambas fazem uso de conjuntos de dados organizados em subconjuntos, mas elas possuem objetivos distintos. Por um lado, o principal objetivo de uma análise de agrupamento é formar grupos de objetos, a partir de critérios de semelhança aplicados às características observadas em cada um deles. Dessa forma, objetos pertencentes a um mesmo grupo tendem a ser mais semelhantes entre si do que objetos pertencentes a grupos distintos. A análise de classificação, por outro lado, parte da existência de várias classes (grupos de objetos) e seu objetivo está em descobrir a qual classe um determinado objeto de interesse, ainda não classificado, pertence.

A ideia de agrupamento está presente de muitas formas no nosso cotidiano e possui diversos exemplos de aplicação. Nas ciências sociais, por exemplo, tem-se a necessidade de agrupar as pessoas com respeito ao seu comportamento. Na biologia, agrupar os seres vivos de acordo com suas características físicas e fisiológicas comuns, bem como por suas relações de parentesco evolutivo. Em marketing, identificar segmentações de mercado, na geografia, indicar grupos de regiões, na medicina, identificar categorias de câncer, dentre outras tantas.

Essa tese se propõe a contribuir com uma proposta de método de agrupamento que seja útil para análise de dados de crimes. Os dados sobre criminalidade são disponibilizados por instituições oficiais cobrindo todo o território nacional, na cidade do Recife a Secretaria de Defesa Social (SDS) é responsável pelo armazenamento desse tipo de informação. Ressalta-se que é essencial o uso de métodos estatísticos para tratamento e análise dessas informações, propiciando transparência, conhecimento dos principais problemas por parte da sociedade e insumos para as instituições policiais realizarem as suas atividades.

Um crescente volume de recursos públicos é aplicado na realização de diversos programas focados na prevenção e na repressão de crimes. Em Pernambuco, o programa Pacto pela Vida (PPV) lançado em 2007 como o primeiro plano estadual de segurança, é uma política pública

de segurança integrada construída com a cooperação e articulação da sociedade, ministério público, poder judiciário, assembleia legislativa, municípios e a União. Desde sua elaboração, o PPV tem como principal objetivo, a prevenção e redução de homicídios, mas também cuida de um conjunto de crimes que despertam insegurança na população, como os crimes contra o patrimônio, estupro, agressão, tráfico de drogas e violência doméstica (RATTON; GALVÃO; FERNANDEZ, 2014; PERNAMBUCO, 2014; PERNAMBUCO, 2021).

Criado como estrutura de governança, o comitê gestor do PPV implantou um modelo de administração integrado, com o combate à violência estruturado em cima de suas Áreas Integradas de Segurança (AIS). Tais áreas são definidas como as menores unidades territoriais consideradas para fins de planejamento das ações policiais. O objetivo das AIS é, portanto, integrar as ações das polícias no combate à criminalidade. A divisão territorial que foi feita em Pernambuco para acompanhamento das ações e resultados do Pacto Pela Vida, dividiu o Estado em 26 AIS, enquanto a capital, Recife, em 5 Áreas.

Este modelo de gestão fundamentado em Áreas Integradas de Segurança também é aplicado em outros estados. Já adotam esse conceito, Minas Gerais, Rio de Janeiro, São Paulo, Espírito Santo, Distrito Federal, Ceará entre outros. Assim, uma alternativa de governança sobre a segurança pública se materializa no Brasil através da criação de áreas integradas, permitindo uma nova forma de atuação do trabalho policial, com articulação de dados e informações dessas regiões.

Olhando por outra perspectiva, também é possível fazer uso dos dados de violência acumulados ao longo do tempo para reavaliar espacialmente a distribuição dos crimes monitorados. Dessa forma, a análise de agrupamento é uma ferramenta com grande potencial estratégico de estudo das AIS. As técnicas de análise de agrupamentos seriam aplicadas com o intuito de identificar grupos de bairros que possuem características semelhantes com respeito aos tipos e frequências de crimes ocorridos em seus limites geográficos. Isso poderia servir, dentre outros, ao objetivo de reavaliar a definição geográfica das Áreas Integradas de Segurança.

Para realizar tal análise, no entanto, é preciso enfrentar alguns desafios metodológicos. O primeiro deles é a necessidade de ter algoritmos que lidem com variáveis de naturezas distintas. O segundo desafio está na necessidade de levar em consideração a falta de informação, reflexo de dados faltantes. Por último, a importância de utilizar medidas de distância adequadas ao tipo de dado utilizado e de impor contiguidade aos grupos que são formados pelo algoritmo. Esta última condição, de contiguidade, é frequente em procedimentos que lidam com objetos espaciais e geográficos, evitando que regiões que não possuam limites contíguos participem

de um mesmo grupo.

Dentre as diferentes técnicas existentes, os métodos tradicionais K-Means (JAIN; DUBES, 1988; ORTEGA et al., 2019) e Partitioning Around Medoids (PAM) (JAIN; DUBES, 1988; KAUFMAN; ROUSSEEUW, 1990) são os mais utilizados, sobretudo, servindo de base de comparação. Ambos possuem o foco em trabalhar na divisão de um conjunto de dados em grupos, minimizando a distância entre os objetos e o centro desses grupos. Tais técnicas são populares devido à sua simplicidade, sendo um bom ponto de partida. No entanto, é possível identificar limitações destes métodos, como ineficiência em conjunto de dados com altas dimensões e só poder ser usados em conjuntos de dados com variáveis numéricas.

Outro algoritmo como o Variable Neighborhood Search for K-Medoides (VNSKMED) (BRITO; SEMAAN; FADEL, 2022), propõe uma nova abordagem para problemas de agrupamentos através do método Variable Neighborhood Search (VNS) i.e., Busca em Vizinhança Variável, caracterizado por realizar buscas em vizinhanças distantes (MLADENOVIĆ; HANSEN, 1997). O VNSKMED procura boas soluções de agrupamento, mas suas desvantagens residem na dificuldade de escapar de ótimos locais e no maior tempo de processamento computacional. Já o agrupamento hierárquico Ward-Like (CHAVENT et al., 2018; AGUIAR; SÁNCHEZ; CAMÊLO, 2020), que usa duas matrizes de dissimilaridade, uma com informação dos dados e outra com informação geográfica, tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido a sua homogeneidade interna, mas é sensível à presença de outliers.

Do mesmo modo, usando restrições espaciais, a literatura apresenta uma abordagem baseada em grafos, que são estruturas que representam relações entre os objetos, o algoritmo Spatial K'luster Analysis by Tree Edge Removal (SKATER) (ASSUNÇÃO et al., 2006; TEIXEIRA; ASSUNÇÃO; LOSCHI, 2019), fácil de programar e que trata a regionalização. Os algoritmos que utilizam a regionalização observam as relações de vizinhança, identificando a localização espacial e a estrutura de vizinhança dos objetos. Contudo, pode ter desperdício de memória, caso o grafo for disperso.

Todos esses métodos são determinados, basicamente, pela medida de proximidade empregada. As medidas de proximidade são quantidades comparativas entre as observações e são escolhidas com base na natureza dos dados. Existem várias maneiras de se obter tais medidas. As métricas mais comuns são a distância Euclidiana e a distância Manhattan, utilizadas para dados numéricos, mas outras medidas, como a distância Gower, que abrange dados mistos, também são muito úteis.

Logo, são muitas técnicas de agrupamento existentes, cada uma com seus critérios, o

que pode ser confuso inicialmente escolher qual empregar. A dificuldade nessa escolha pode comprometer os resultados obtidos, devido ao fato dos agrupamentos finais serem fortemente dependentes do método e da distância usada. Assim, estudos que comparam métodos e distâncias contribuem na identificação dos mais satisfatórios para uma determinada situação. O desempenho dos algoritmos estudados, combinados com as medidas de proximidade adequadas, é um problema importante que precisa ser estudado.

Justificada pelo problema da violência e da possibilidade de usar técnicas estatísticas para estudo das AIS, esta tese utiliza dados de 2007 a 2015 sobre a ocorrência de crimes e informações espaciais dos bairros da cidade de Recife envolvendo classificações das Áreas Integradas de Segurança, para avaliar cinco métodos de agrupamento, K-Means, PAM, VNSKMED, Ward-Like e SKATER, além de propor alterações nos algoritmos Ward-Like e SKATER, modificando a estrutura de pesos e o processo de partição dos grupos usando a distância Gower, nomeados de Ward-Like.New e SKATER.New, respectivamente. Os algoritmos permitem explorar os padrões relacionados aos crimes, possibilitando mapeá-los em grupos de bairros da capital pernambucana.

O desenvolvimento deste trabalho é útil pois algoritmos de agrupamento são testados e novas extensões foram criadas, podendo ser utilizadas para os mais diversos agrupamentos e assim informações serão adquiridas e desenvolvidas que darão base ao assunto abordado e servirão para orientar e instigar novas pesquisas.

Por fim, esta tese está estruturada em seis capítulos, incluindo este primeiro. No segundo capítulo é feita uma descrição da análise de agrupamento, descrevendo os algoritmos K-Means, PAM, VNSKMED, Ward-Like e SKATER e os índices de validação Calinski-Harabasz, Dunn e Davies-Bouldin. O terceiro capítulo apresenta as modificações propostas, Ward-Like.New e SKATER.New. O quarto capítulo trata de elucidar o estudo das Áreas Integradas de Segurança e descrever os dados utilizados. No quinto capítulo são apresentados os resultados obtidos e por fim, o sexto capítulo apresenta as considerações finais e sugestões de trabalhos futuros.

2 ANÁLISE DE AGRUPAMENTO

Neste capítulo é feito um detalhamento do problema de agrupamento, apresentando os seus conceitos básicos. Ao longo do conteúdo são descritas as distâncias, as técnicas de agrupamento e os métodos de validação dos algoritmos.

2.1 CONCEITOS BÁSICOS

A característica de um problema de agrupamento é a necessidade de se agrupar objetos de um conjunto de dados de forma a manter os mais semelhantes no mesmo grupo. Apesar de intuitiva, essa semelhança é representada por alguma característica que seja comum entre os objetos. Dessa forma, agrupamentos podem ser definidos como grupos de objetos que possuem características semelhantes e são aplicados em uma ampla variedade de campos e circunstâncias (KIANI; MAHDAVI; KESHAVARZI, 2015). Além disso, a análise de agrupamento pode ser usada não apenas para identificar grupos, como também prescrever uma nova estrutura, mais ou menos homogênea para um conjunto de dados já dividido (JAIN; DUBES, 1988).

As técnicas de agrupamento são ferramentas importantes na análise exploratória de dados e vêm sendo desenvolvidas desde, principalmente, Sokal e Sneath (1963) e o fator principal que respalda o interesse no tema é a vasta utilização desses métodos que encontram aplicações em diversas áreas. Além disso, o avanço computacional e o desenvolvimento tecnológico proporcionam uma aplicabilidade da análise de agrupamentos cada vez maior com a construção de bancos de dados extensos, tratamento de qualquer tipo de variável, e sobretudo, uma crescente preocupação em fazer com que os métodos tratem de forma adequada as informações existentes e sejam também de menor complexidade.

O processo de agrupamento de dados, também conhecido como clusterização, procura encontrar conjuntos de objetos que se agrupam naturalmente por alguma similaridade, e esse processo compreende algumas etapas necessárias para realização da tarefa de agrupamento:

1 Representação dos dados;

2 Definição de uma medida de proximidade;

3 Escolha e aplicação de algum algoritmo de agrupamento de dados;

4 Validação dos algoritmos; e

5 Interpretação dos resultados.

Na etapa 1, têm-se a escolha dos atributos relevantes dos objetos para realizar o agrupamento. Nessa etapa, se faz a preparação dos dados que envolve, por exemplo, transformações nas variáveis originais, tais como normalizações e mudanças de escala. Na etapa 2, tem-se a definição de uma medida de proximidade, a qual se trata de uma métrica para quantificar a distância entre os objetos analisados. Na etapa 3, aplica-se um algoritmo de agrupamento. Em seguida, têm-se as etapas de validação (etapa 4) e interpretação dos grupos obtidos (etapa 5).

2.1.1 Representação dos Dados

Considere o conjunto de dados $X = \{x_1, x_2, \ldots, x_n\}$ no espaço $\mathbb{R}^n$ com n objetos a serem agrupados, e cada objeto x_i possui p variáveis ou atributos. Os algoritmos de agrupamento normalmente operam sobre uma das três estruturas de dados a seguir:

- Matriz de dados

A matriz de dados corresponde aos dados propriamente ditos. Assim, um conjunto de dados é representado por uma matriz $n \times p$, sendo n o número de objetos e p o número de atributos ou variáveis de cada objeto. Com isso, uma linha e uma coluna da matriz correspondem, respectivamente, a um objeto e um atributo desse objeto (JAIN; DUBES, 1988). A estrutura é a seguinte:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

onde cada observação x_{ij} representa o j-ésimo atributo do i-ésimo objeto, para todo $i = 1, \ldots, n$ e $j = 1, \ldots, p$.

Os conjuntos de dados podem apresentar dados inconsistentes ou incompletos, isso ocorre quando alguns de seus atributos apresentam dados ausentes. A matriz de dados utilizada aqui é uma matriz completa, i.e. não há valores ausentes em nenhum atributo.

- Matriz de proximidade

Já na matriz de proximidade, também chamada de matriz de similaridade ou dissimilaridade, cada elemento, s_{ij}, representa a medida de distância entre os objetos i e j. Esta matriz pode ser especificada por uma matriz diagonal inferior contendo $n(n-1)/2$ elementos:

$$\begin{bmatrix} 0 & & & & \\ s_{21} & 0 & & & \\ s_{31} & s_{32} & 0 & & \\ \vdots & \vdots & \vdots & & \\ s_{n1} & s_{n2} & \cdots & \cdots & 0 \end{bmatrix}$$

em que s_{ij} é o coeficiente de similaridade ou dissimilaridade entre os objetos i e j. Em geral s_{ij} é representado por um valor não negativo que é próximo de zero quando os objetos i e j são similares, e vai aumentando conforme mais diferentes os objetos se tornam (PAIVA, 2013; SERPA, 2011).

Para o cálculo do coeficiente s_{ij} usa-se as medidas de distância descritas na seção 2.1.2. Para ir de uma similaridade para uma distância, basta utilizar a fórmula: $d(i,j) = 1 - s_{ij}$, em que $d(i,j)$ é a distância entre i e j e s_{ij} o coeficiente de similaridade.

- Grafo de proximidade

A partir da matriz de distâncias é possível definir o grafo de proximidade (com mais detalhes na seção 2.2.4). Considere um grafo $\mathcal{G} = (V, L)$, em que V é o conjunto de nós e L o conjunto de arestas não direcionadas que ligam cada nó a todos os outros nós. Cada nó do grafo $\mathcal{G}$ representa um objeto do conjunto de dados e cada aresta recebe um valor de peso que corresponde ao valor da proximidade entre os objetos, ou seja, supondo v_i e v_j dois nós de $\mathcal{G}$ que representam os dois objetos x_i e x_j, então o peso da aresta l_{ij} que liga os nós v_i e v_j será a distância entre os objetos x_i e x_j (SERPA, 2011).

2.1.1.1 Tratamento dos dados

Um aspecto importante na etapa de representação dos dados, trata da homogeneidade entre as variáveis que venham a participar da tarefa de agrupamento. A padronização dos

dados é uma etapa essencial feita para evitar distorções na estrutura do agrupamento. As medidas de distância podem sofrer influência pela diferença de grandeza dos atributos, dessa forma, a padronização trata de transformar os dados numéricos para a mesma escala de valores, ou seja, é atribuído o mesmo peso para cada atributo (FAVEIRO et al., 2009; GUERREIRO et al., 2021).

Assim, antes de implementar os algoritmos foi feita a padronização dos dados através da fórmula 2.1, por meio da subtração de sua média e divisão pelo seu desvio padrão.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{DP(x_{.j})} \tag{2.1}$$

onde, x_ij é o objeto i avaliado no atributo j $\bar{x}_j$ é a média a média dos objetos para o atributo j e $DP(x_{.j})$ é o desvio padrão do j-ésimo atributo. Agora, a dimensão transformada terá uma média zero e desvio padrão um.

2.1.2 Medida de Proximidade

Nesta etapa é definida a medida de proximidade apropriada ao agrupamento, neste trabalho definimos que a medida de proximidade é precisamente a distância escolhida, essa escolha deve levar em consideração os tipos e as escalas dos objetos das variáveis utilizadas. Em geral, os atributos podem ser de vários tipos, binários, que são aqueles definidos por apenas dois valores; discretos, definidos por um número finito de valores; ou contínuos, compostos por um número infinitos de valores. Já as escalas dos atributos definem se eles são qualitativos ou quantitativos (JAIN; DUBES, 1988).

Definidos os tipos e escalas das variáveis, é possível escolher a distância a ser utilizada. Segundo Jain e Dubes (1988) uma distância deve satisfazer algumas propriedades. Seja Ω o conjunto de objetos a serem agrupados, a distância entre os objetos i e j, $d(i,j)$, é tal que:

- $d(i,j) \geq 0$, $\forall\ i,j \in \Omega$, garantindo que uma distância seja não negativa;
- $d(i,i) = 0$, $\forall\ i \in \Omega$, a distância de um objeto para ele mesmo é zero;
- $d(i,j) = d(j,i)$, $\forall\ i,j \in \Omega$, a distância é uma função simétrica;
- $d(i,j) \leq d(i,h) + d(h,j)\ \forall\ i,j,h$, garantindo a desigualdade triangular.

Neste trabalho, serão utilizadas três medidas de distância: a distância Euclidiana e a distância Manhattan, adequadas quando os atributos são numéricos, não sendo aplicáveis a atributos nominais e a distância Gower, que abrange atributos de outros tipos.

2.1.2.1 Distância Euclidiana

A distância Euclidiana é a distância mais conhecida e comumente utilizada. Essa distância é a menor distância entre dois pontos em qualquer dimensão, que pode ser provada pela aplicação repetida do Teorema de Pitágoras (JAIN; DUBES, 1988).

As métricas de distância também são consideradas medidas de dissimilaridade. Seja x_{ik} o k-ésimo atributo do objeto i, a distância Euclidiana $\mathsf{d}_{Euc}(i,j)$ entre cada par de objetos i e j, é definida como a raiz quadrada da soma da diferença ao quadrado entre x_{ik} e x_{jk} para todo $k = 1, \ldots, p$ em suas respectivas dimensões:

$$\mathsf{d}_{Euc}(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2.2}$$

Que satisfaz as propriedades matemáticas de uma função de distância, citadas em 2.1.2.

2.1.2.2 Distância Manhattan

Outra distância bem conhecida é a Manhattan, representada por:

$$\mathsf{d}_{Manh}(i,j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}| \tag{2.3}$$

A distância Manhattan tem uma definição mais simples na qual é apenas a soma do módulo das diferenças entre x_{ik} e x_{jk} em cada dimensão. E também satisfaz as propriedades matemáticas de uma função de distância, citadas anteriormente.

2.1.2.3 Distância Gower

Por vezes, os conjuntos de dados apresentam variáveis do tipo misto, ou seja, com a presença de variáveis contínuas, discretas e até mesmo, categóricas. Para essas situações, a distância de Gower (GOWER, 1971) foi desenvolvida. Essa distância mede a semelhança entre dois objetos com base nas informações dos atributos, permitindo informações ausentes.

A similaridade s_{ij} entre os objetos i e j é expressa na fórmula a seguir:

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}\delta_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}} \tag{2.4}$$

em que δ_{ijk} é um peso dada a comparação ijk, atribuindo valor 1 para comparações válidas e valor 0 para comparações inválidas (quando o valor da variável está ausente em um ou ambos indivíduos); s_{ijk} é a contribuição da variável k na similaridade entre os os objetos i e j, possuindo valores entre 0 e 1.

Essa similaridade permite valores ausentes, assim, se uma unidade apresenta valores ausentes, então a similaridade com qualquer outra unidade seria indefinida, para isso, (GOWER, 1971) adicionou o elemento δ_{ijk}, que funciona como uma ponderação simples, informando quando há presença ou ausência de determinada informação, então $\delta_{ijk} = 0$ se não há nenhuma informação pertencente a variável k e $\delta_{ijk} = 1$ quando há informação.

O cálculo de s_{ijk} vai depender do tipo de variável. Se a k-ésima variável é qualitativa, então:

$$s_{ijk} = \begin{cases} 0, \text{se } x_{ik} = x_{jk} \\ 1, \text{se } x_{ik} \neq x_{jk} \end{cases} \tag{2.5}$$

Se for quantitativa, tem-se

$$s_{ijk} = 1 - \left| \frac{x_{ik} - x_{jk}}{\text{max}(x_{ik}) - \text{min}(x_{ik})} \right| \tag{2.6}$$

em que x_{ik} se refere ao valor da k-ésima variável para o objeto i, max(x_{ik}) é valor máximo da k-ésima variável e min(x_{ik}) é valor mínimo da k-ésima variável. Para o cálculo dessa distância no ambiente R (LÊ; JOSSE; HUSSON, 2008) foi utilizada a função *daisy* do pacote *cluster* (MAECHLER et al., 2021).

2.1.3 Agrupamento

Esta etapa compreende a escolha de um algoritmo de agrupamento adequado ao problema. De acordo com a literatura, os métodos de agrupamento podem ser classificados como particionais, hierárquicos e baseados em grafos (JOSÉ-GARCÍA; GÓMEZ-FLORES, 2016; PAIVA, 2013; DUQUE; RAMOS; SURIÑACH, 2007).

Nos algoritmos de clusterização particionais, os grupos (clusters) são mutuamente exclusivos, ou seja, cada objeto pertencente a exatamente um grupo (cluster) e é possível determinar o número k de grupos a priori, de forma que cada grupo deve conter pelo menos um objeto (LEAL, 2004).

Tais técnicas dividem um conjunto de dados em vários grupos com base em determinado critério. Depois a tarefa de particionamento é convertida em um problema de otimização, baseado na minimização de distância (GUERREIRO et al., 2021). Os grupos são formados com base em semelhanças e diferenças entre os objetos dos grupos e as medidas de similaridade diferem de aplicação para aplicação. Na Figura 1a tem-se um exemplo de um conjunto de 94 objetos formado por 4 grupos visivelmente separados por um método particional.

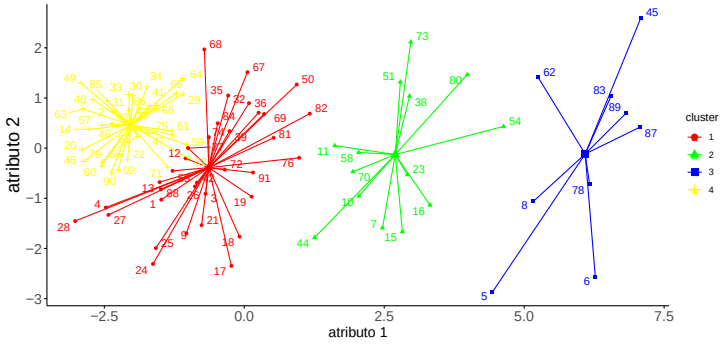
A técnica de partição mais popular é o algoritmo K-Means (MACQUEEN et al., 1967). Este algoritmo é extremamente veloz, geralmente convergindo em poucas iterações, e simples de implementar em linguagens computacionais, no entanto, sofre algumas desvantagens, como a dificuldade na determinação do número de grupos e alto enviesamento (KIANI; MAHDAVI; KESHAVARZI, 2015; NANDA; PANDA, 2014). Ao longo dos anos, muitas propostas para aumentar seu desempenho foram desenvolvidas, como o K-Medoids (KAUFMAN; ROUSSEEUW, 1990).

Já os métodos hierárquicos produzem uma hierarquia de agrupamento chamada dendrograma (ou estrutura em árvore), que consiste em uma estrutura em árvore, a qual representa a sequência hierárquica de partições do conjunto de dados, e onde pode-se inferir o número de grupos cortando o dendograma em diferentes níveis de acordo com o número de grupos desejados (JOSÉ-GARCÍA; GÓMEZ-FLORES, 2016). O procedimento constrói níveis sucessivos de agrupamento, nos quais o agrupamento atual é baseado na solução obtida no nível anterior. Portanto, o agrupamento hierárquico não requer que seja definido um número a priori de grupos, no entanto, os grupos obtidos são estáticos porque os objetos atribuídos a um determinado grupo não podem ser movidos para outro (MONDAL; CHOUDHURY, 2013).

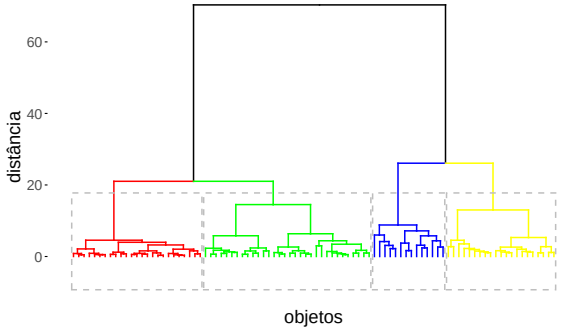
Na Figura 1b tem-se um exemplo de dendograma que ilustra o agrupamento hierárquico. Nos níveis superiores pode-se observar como os grupos tendem a fusão até formar um único agrupamento. Ainda na Figura 1b, pode-se ver o ponto de parada (19), na linha pontilhada cinza, indicando que nesse ponto há formação de exatamente $K = 4$ grupos.

Figura 1 – Exemplo de clusterização de um método particional formando por 4 grupos (Figura 1a) e exemplo de dendograma resultante de um algoritmo hierárquico (Figura 1b).

(a) Clusterização particional



(b) Dendograma



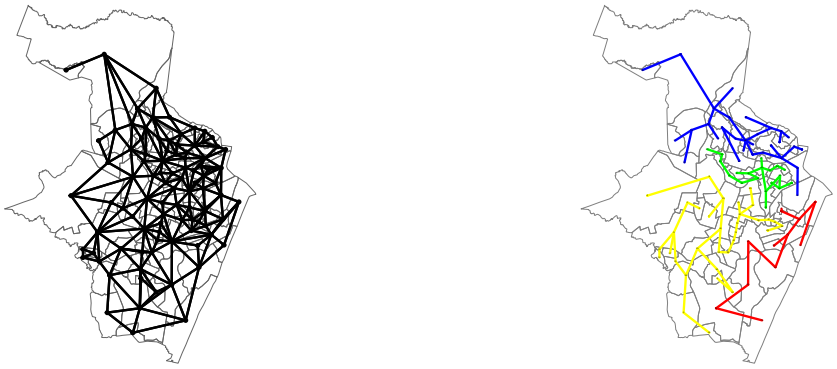
Fonte: Elaboração Própria.

No agrupamento baseado em grafos, um grafo é criado conectando os dados de acordo com alguma medida de similaridade. Um grafo é uma estrutura de representação de dados que consiste num diagrama composto de pontos, a partir dos quais os vizinhos são ligados entre si por linhas (TEIXEIRA; ASSUNÇÃO; LOSCHI, 2019).

No grafo de vizinhança, todas os dados que apresentam distâncias menores que um determinado limite estão conectados. Após o grafo totalmente conectado, todos os objetos semelhantes estão ligados entre si e são ponderados pelo valor da similaridade (POURBAHRAMI; KHANLI, 2018). Na Figura 2 é apresentado um exemplo de um modelo de agrupamento baseado em grafos, onde são destacadas as arestas e os grupos formados.

Figura 2 – Exemplo de mapa da cidade de Recife dividido em 94 bairros sobrepostos pelo grafo de vizinhança (Figura 2a) e exemplo de partição em quatro grupos espaciais (Figura 2b).

(a) Grafo (b) Partição



Fonte: Elaboração Própria.

No entanto, essas classificações listadas, particional, hierárquico e grafo, não são únicas e definitivas. Como já mencionado, o principal objetivo dessas técnicas é o de reunir objetos semelhantes e separar objetos diferentes. Existe um grande número de algoritmos descritos na literatura, cada um com seu propósito e restrições, nas próximas seções são apresentados alguns deles.

2.2 ALGORITMOS DE AGRUPAMENTO

Neste trabalho serão abordados diferentes algoritmos de agrupamento, haja vista a melhor resposta para o problema de agrupamento dos bairros de Recife. Os escolhidos foram os tradicionais, K-Means e PAM, o baseado em busca em vizinhança, VNSKMED, e os com restrição de contiguidade, Ward-Like e SKATER. Além destes, são propostas as modificações Ward-Like.New e SKATER.New, utilizando a distância Gower. Nas próximas seções serão discutidos de maneira mais detalhada cada um dos algoritmos citados.

2.2.1 Algoritmo K-Means

O algoritmo K-Means[1] é um dos métodos mais populares das técnicas particionais, proposto por MacQueen et al. (1967). Esse algoritmo busca particionar os dados em K agrupamentos mutuamente exclusivos. Ele define um protótipo em torno do ponto médio de cada agrupamento, comumente chamado de centroide (centro do grupo). Este é um ponto artificial gerado aleatoriamente e que possui a mesma dimensão dos dados a serem agrupados (ALAM; DOBBIE; REHMAN, 2015).

O algoritmo busca encontrar a partição ideal, particionando iterativamente os centroides e minimizando a função 2.7:

$$M = \sum_{i=1}^{K} \sum_{x \in C_i}^{n} d(x, b_i) \tag{2.7}$$

aqui b_i é o centro do grupo C_i (grupo dos centroides), enquanto $d(x, b_i)$ é a distância entre um ponto x e o centroide b_i.

A função objetivo M tenta minimizar a distância entre cada objeto do centro do grupo em um grupo. Inicialmente, atribuímos aleatoriamente K centroides, em seguida ele começa a atribuir cada registro do conjunto de dados ao grupo cujo centroide é o mais próximo usando alguma medida de distância e recalcula os centroides. Esse processo de alocar os objetos e atualizar os centroides se repete até que não haja mudanças nos centroides (MENDES, 2017; PAIVA, 2013).

Considerando um conjunto de dados X com n objetos, $X = \{x_1, x_2, \ldots, x_n\}$, as etapas do algoritmo são as seguintes:

Algoritmo 1 Algoritmo K-Means

Entrada: dados e quantidade de grupos

1: Gerar os K centroides aleatoriamente.
2: Agrupar a matriz de protótipo de grupo $d(x, b_i)$ (matriz de distância entre os objetos de dados e os centroides) de tamanho $K \times b$.
3: Atribua cada objeto do conjunto de dados ao grupo mais próximo.
4: Calcule a média dos elementos de cada grupo e troque os K centroides dos grupos por suas médias.
5: Calcule novamente a matriz de protótipo de grupo.
6: Repita os passos 4, 5 e 6 até que não haja alteração para cada grupo.

Saída: os grupos formados

[1] Implementado no software R e disponível no pacote base do R (R Core Team, 2023).

Este algoritmo é amplamente utilizado para lidar com bancos de dados de alta dimensão devido ao baixa custo computacional e convergência rápida, contudo, é sensível a outliers, uma vez que o centroide é definido pela média dos valores dos objetos de cada grupo.

2.2.2 Algoritmo PAM

Um outro método de agrupamento particional é o Partitioning Around Medoids (PAM)[2], também chamado de K-Medoides, proposto por Kaufman e Rousseeuw (1990). Sua estrutura e seu funcionamento são bem similares ao K-Means, a diferença é que enquanto no K-Means o centroide não precisa corresponder a um objeto pertencente ao conjunto de dados, no PAM o centroide é um dos pontos do conjunto, chamado de medoide.

Segundo Brito, Semaan e Fadel (2022) o termo medoide refere-se a um objeto dentro de um grupo para o qual a dissimilaridade média entre ele e todos os outros membros do grupo é mínima, ele corresponde ao ponto mais central do grupo. A forma mais comum é definir como medoides K objetos do conjunto de dados, isto é feito para minimizar os efeitos da inicialização aleatória do K-Means e com isso, conseguir eliminar ruídos e discrepâncias, o que torna o algoritmo PAM mais robusto (MONDAL; CHOUDHURY, 2013).

Basicamente, o algoritmo PAM inicia o processo de formação dos agrupamentos selecionando os K medoides arbitrariamente dentre os objetos da base de dados. Em cada etapa, uma troca entre um objeto selecionado O_i e um objeto não selecionado O_h é feita, desde que essa troca resulte em uma melhoria no agrupamento. Sobretudo, para calcular o efeito dessa troca entre O_i e O_h, um custo C_{ih} é calculado, que está relacionado com a qualidade de particionamento dos objetos não selecionados para K grupos representados pelos medoides. O algoritmo PAM consiste em:

Este algoritmo funciona satisfatoriamente para pequenos conjuntos de dados e poucos grupos k, em grandes volumes de dados e grupos, torna-se ineficiente.

2.2.3 Algoritmo VNSKMED

Outro algoritmo utilizado foi o VNSKMED proposto por Brito, Semaan e Fadel (2022) e baseado na meta-heurística Variable Neighborhood Search (VNS). As meta-heurísticas vêm sendo recentemente utilizadas para agrupamento de dados. As meta-heurísticas são algoritmos

[2] Implementado no software R e disponível no pacote cluster (MAECHLER et al., 2021).

Algoritmo 2 Algoritmo PAM

Entrada: dados e quantidade de grupos

1: Selecione aleatoriamente K objetos representativos. Marque esses objetos como "selecionados" e marque os demais como "não selecionados".

2: Repita até que não haja mais objetos a serem classificados.

a. Atribua cada objeto restante ao grupo do medoide mais próximo.

b. **Faça** para todos os objetos selecionados O_n.

i. **Faça** para todos os objetos não selecionados O_i.
Calcular C_{ih} (custo de troca)
Fim do Faça

Fim do Faça

c. Se o custo para todos os objetos selecionados e não selecionados $C_{imin,hmin} < 0$

d. Em seguida marque O_i como não selecionado e O_h como selecionado.

3: Vá para o passo 3 até que não haja mudança no medoide.

Saída: os grupos formados

sistemáticos capazes de resolver um problema de agrupamento como problema de otimização e encontrar uma solução ótima com um menor custo computacional e várias delas são adotadas por algoritmos de agrupamento particional (AL-SULTAN, 1995; SCHEUERER; WENDOLSKY, 2006; NASCIMENTO; TOLEDO; CARVALHO, 2010). A versão VNSKMED proposta com o objetivo de trabalhar com o problema de agrupamentos com medoides, tem o diferencial de ser mais eficaz na definição de estruturas de vizinhanças e procedimentos de busca local e perturbação, garantindo a produção de ótimos globais (BRITO; SEMAAN; FADEL, 2022).

Seja um conjunto de dados $X = \{x_1, x_2, \ldots, x_n\}$ com n objetos a serem agrupados, seleciona-se desse conjunto K objetos que definam os medoides utilizados para formação dos grupos. Os medoides são determinados de forma que seja mínima a soma das distâncias dos objetos restantes do conjunto X até o seu medoide mais próximo, equivalendo a minimizar a Equação 2.8.

$$V = \sum_{i=1}^{K} \sum_{x \in G_i}^{n} d(x, m_i) \tag{2.8}$$

em que m_i é o medoide do grupo G_i e $d(x, m_i)$ é a matriz de distâncias que contém o valor da distância entre um objeto qualquer de X e o medoide m_i do respectivo grupo. A função objetivo 2.8 minimiza a soma das distâncias dos objetos dos grupos aos seus respectivos medoides.

Neste algoritmo, cada solução (vizinhança) é representada por um vetor $s = (m_1, m_2, \ldots, m_k)$

com K posições correspondentes aos medoides. No processo de geração da solução inicial (s_0), são gerados g vetores s, que correspondem aos objetos associados aos medoides que são selecionados aleatoriamente de X. Dessa forma, são geradas várias soluções e somente a que obtiver o menor valor na função objetivo (2.8) é que será considerada solução inicial.

A estrutura de vizinhanças, $N_v^p(s)$, onde $p = \{1, \ldots, p_{max}\}$ e $N_v^p(s^*)$ é o conjunto de soluções da p-ésima vizinhança de s^*, é definida da seguinte maneira: dada uma solução s, toma-se, para cada um dos seus medoides, objetos de X mais próximos, de acordo com a distância utilizada. Os índices desses objetos são armazenados em uma matriz $M_{k\times p}$. Uma solução inicial $s' \in N_v^p(s)$ difere de s por, exatamente, v medoides (BRITO; SEMAAN; FADEL, 2022).

Em cada iteração, o VNSKMED seleciona-se, uma solução aleatória s' contida na vizinhança $N_v^p(s)$ da solução corrente ($s' \in N_v^p(s^*)$. Em seguida, aplica-se uma busca local sobre o vizinho s'. Na busca local, considerando a mesma definição de estrutura de vizinhanças, é construída uma matriz $M_{k\times l}^2$, onde l são os vizinhos dos medoides na busca local. Logo depois, os subconjuntos de medoides de s', C_k^{k-1}, são combinados com os elementos de M^2, produzindo q soluções s_j ($q = k^2 l$). Posteriormente, é calculado o valor da função objetivo, determinada na Equação 2.8 para cada s_j e define-se:

$$\arg\min_{j=1,\ldots,g} f(s_j) \quad \text{e} \quad s'' = s_w \tag{2.9}$$

Caso o valor da função objetivo (Equação 2.8) da solução ótima local s'' ($f(s'')$) seja inferior ao valor de $f(s)$, é realizada a atualização da solução atual s com s'' e retorna-se à primeira vizinhança $N_v^1(s)$. Caso contrário, incrementa-se a ordem de estrutura de vizinhança, gera-se um novo vizinho em relação à solução s e aplica-se novamente a busca local. O critério de parada do VNSKMED tem como condição de parada o número de iterações.

As etapas do algoritmo VNSKMED são descritas em:

Algoritmo 3 Algoritmo VNSKMED

Entrada: v_{max}, $MAXITER$, p, l

1: Gere uma solução inicial s
   seja p_{max} o número de estruturas diferentes de vizinhança
   Inicialize a solução s^* para a melhor solução encontrada $s^* \leftarrow s$
2: **enquanto** critério de parada não satisfeito **faça**
   gere um vizinho s' da vizinhança $N_v^p(s^*)$
   aplique a busca local em s' obtendo ótimo local s''
   $(f(s'')) < (f(s^*))$ então
   $s^* \leftarrow s''$
   $p \leftarrow 1$
   Senão
   $p \leftarrow p + 1$
3: **fim do enquanto**

Saída: os grupos formados

A Tabela 1 traz a descrição dos parâmetros que precisam ser fornecidos como entrada para a execução do VNSKMED. De acordo com Brito, Semaan e Fadel (2022), tais parâmetros tem impacto direto no algoritmo e estes produzem as melhores soluções para o problema.

Tabela 1 – Parâmetros calibrados no algoritmo VNSKMED

Parâmetro	Descrição	Valor recomendado
v_{max}	máximo de vizinhanças	3
$MAXITER$	máximo de iterações sem melhoria	10
p	nº de vizinhos dos medoides na perturbação	10
l	nº de vizinhos dos medoides na busca local	30

Fonte: Elaboração própria.

A função no R que implementa o algoritmo VNSKMED está disponível pelos autores em <https://github.com/jambrito/VNSKMED>.

2.2.4 Algoritmo Ward-Like

O método de agrupamento Ward-Like[3], proposto por Chavent et al. (2018) é um algoritmo de agrupamento hierárquico que inclui restrições espaciais usando duas matrizes de dissimilaridade D_0 e D_1 e um parâmetro de mistura $\alpha \in [0; 1]$. O método consiste em combinar as duas matrizes, $D_0 = [d_{0,(ij)}]$ que representa a matriz de dissimilaridades obtida com base nos valores das variáveis não espaciais, que nesse trabalho são as variáveis criminais, e $D_1 = [d_{1,(ij)}]$ que

[3] Implementado no software R e disponível no pacote ClustGeo (CHAVENT et al., 2021).

leva em conta a vizinhança entre os n objetos, i.e., os pesos espaciais entre os bairros, usando um parâmetro α, que define o peso das restrições, ou seja, permite definir a importância de cada matriz no agrupamento.

Suponha que um conjunto de n objetos sejam particionados em K grupos, denotados por $\mathcal{C}_k^\alpha$, com $k = 1, \ldots, n$, que formam a partição $\mathcal{P}_k^\alpha = (\mathcal{C}_1^\alpha, \ldots, \mathcal{C}_n^\alpha)$. A abordagem Ward-Like, implica na minimização da inércia dentro do grupo da partição $\mathcal{P}_k^\alpha$ definida como:

$$W_\alpha(\mathcal{P}_k^\alpha) = \sum_{k=1}^{n} I_\alpha(\mathcal{C}_k^\alpha) \tag{2.10}$$

onde $I_\alpha(\mathcal{C}_k^\alpha)$ é a inércia de $\mathcal{C}_k^\alpha$ e assume a forma de:

$$\mathcal{I}_\alpha(\mathcal{C}_k^\alpha) = (1-\alpha) \sum_{i \in \mathcal{C}_k^\alpha} \sum_{j \in \mathcal{C}_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{0,(ij)}^2 + \alpha \sum_{i \in \mathcal{C}_k^\alpha} \sum_{j \in \mathcal{C}_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{1,(ij)}^2 \tag{2.11}$$

em que $\mu_k^\alpha = \sum_{i \in \mathcal{C}_k^\alpha} w_i$ é o peso de $\mathcal{C}_k^\alpha$ e $d_{0,ij}$ e $d_{1,ij}$ são a dissimilaridade normalizada entre as observações i e j em D_0 e D_1, respectivamente.

Basicamente, a proporção da inércia total mista explicada pela partição $\mathcal{P}_k^\alpha$ em K grupos é:

$$\mathcal{Q}_\beta(\mathcal{P}_k^\alpha) = 1 - \frac{W_\beta(\mathcal{P}_k^\alpha)}{W_\beta(\mathcal{P}_1)} \in [0, 1] \tag{2.12}$$

Quando $\beta = 0$, o denominador $W_0(\mathcal{P}_1)$ é a inércia total, e o numerador é a inércia dentro do grupo $W_0(\mathcal{P}_k^\alpha)$, ambos baseados na matriz D_0. Portanto, quanto maior o valor de $\mathcal{Q}_0(\mathcal{P}_k^\alpha)$, mais homogêneo é o grupo em termos de atributos não espaciais. Já quando $\beta = 1$, o denominador $W_1(\mathcal{P}_1)$ é a inércia total e o numerador é a inércia dentro do grupo $W_1(\mathcal{P}_k^\alpha)$, ambos baseados na matriz D_1. Portanto, quanto maior o valor do critério $\mathcal{Q}_1(\mathcal{P}_k^\alpha)$, maior a homogeneidade do grupo em termos de atributos espaciais (AGUIAR; SÁNCHEZ; CAMÊLO, 2020).

Para a construção de D_1 é utilizado o critério de contiguidade, dado por:

$$C = (c_{ij})_{n \times n} \tag{2.13}$$

onde, $c_{ij} = 1$, se o i-ésimo e o j-ésimo objeto são contíguos e 0 caso contrário. A matriz de adjacência A, que é criada pelo critério de contiguidade, é a base para a construção da matriz de vizinhança:

$$D_1 = 1 - A \tag{2.14}$$

onde $A = [a_{ij}]_{n \times n}$ será igual a 1 se os objetos i e j são vizinhos e 0 caso contrário, e a diagonal $a_{ii} = 1$.

Para esse tipo de matriz de dissimilaridades, a coesão geográfica, quando se tem poucos grupos, é pequena, ou seja, $W_1(P_k)$ pode ser muito pequeno e $Q_1(P_k)$ assumirá valores muito maiores que os obtidos por $Q_0(P_k)$, uma vez que pares de objetos próximos, mas que não possui vizinhança (bairros que não possuem vizinhos), recebem o mesmo valor que objetos (bairros) muito distantes. Para minimizar essa desigualdade nas duas matrizes usadas pelo método, $\mathcal{Q}_\beta$ precisa ser normalizado (Equação 2.15). Dessa forma, são obtidas escalas similares para Q_0 e Q_1.

$$\mathcal{Q}_{\beta norm}(\mathcal{P}_k^\alpha) = \mathcal{Q}_\beta(\mathcal{P}_k^\alpha)\mathcal{Q}_\beta(\mathcal{P}_k^\beta) \tag{2.15}$$

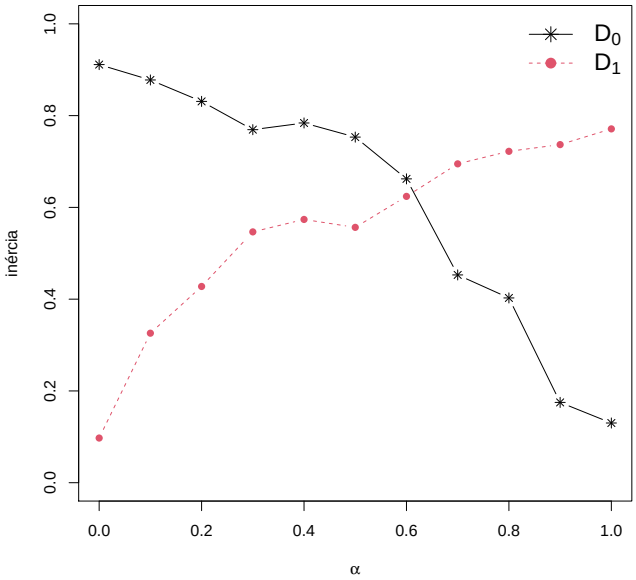
Por fim, é feita a análise e escolha do parâmetro de mistura α. Este parâmetro controla a parte da pseudo-inércia devida a D_0 e D_1. Quando $\alpha = 0$ a Equação 2.10 baseia-se apenas em D_0, nomeadamente atributos não espaciais, e quando $\alpha = 1$ a Equação 2.10 baseia-se apenas em D_1, nomeadamente pesos espaciais. O interesse aqui é em determinar um valor de α que aumente a homogeneidade espacial de uma partição de grupos sem afetar negativamente a homogeneidade não espacial. Para um determinado número K de grupos, a ideia é considerar uma dada grade de valores de J para $\alpha \in [0, 1]$:

$$\mathcal{G} = \{\alpha_1 = 0, \alpha_2, \ldots, \alpha_J = 1\} \tag{2.16}$$

Para cada valor $\alpha_j \in \mathcal{G}$, a partição correspondente $\mathcal{P}_k^{\alpha j}$ em K grupos é obtida usando o algoritmo Ward-Like. Para as J partições $\{\mathcal{P}_k^{\alpha j}, j = 1, \ldots, J\}$, o critério $\mathcal{Q}_0(\mathcal{P}_k^{\alpha j})$ é avaliado. O gráfico dos pontos $\{(\alpha_j, \mathcal{Q}_0(\mathcal{P}_k^{\alpha j}))j = 1, \ldots, J\}$ fornece uma maneira visual de observar a perda de homogeneidade não espacial da partição $\mathcal{P}_k^{\alpha j}$ (da partição criminal pura $\mathcal{P}_k^0$) à medida que α_j aumenta de 0 para 1. Da mesma forma, para as partições as J partições $\{\mathcal{P}_k^{\alpha j}, j = 1, \ldots, J\}$, o critério $\mathcal{Q}_1(\mathcal{P}_k^{\alpha j})$ é avaliado.

O gráfico dos pontos $\{(\alpha_j, \mathcal{Q}_1(\mathcal{P}_k^{\alpha j}))j = 1, \ldots, J\}$, exemplificado na Figura 3, fornece uma maneira visual de observar a perda de homogeneidade espacial da partição $\mathcal{P}_k^{\alpha j}$ (da partição espacial pura $\mathcal{P}_k^1$) à medida que α_j aumenta de 0 para 1. Esses dois gráficos, sobrepostos na mesma figura, permitem ao usuário escolher um valor adequado para $\alpha \in \mathcal{G}$ que é um trade-off entre a perda de homogeneidade não espacial e maior coesão espacial.

Figura 3 – Exemplo de visualização gráfica da proporção das inércias explicadas $Q_0(P_k^\alpha)$ versus α (na linha preta) e $Q_1(P_k^\alpha)$ versus α (na linha vermelha).



Fonte: Elaboração própria.

A Figura 3 apresenta o gráfico da proporção de inércia explicada calculada com D_0 que é igual a $0,91$ quando $\alpha = 0$ e diminui quando α aumenta (linha preta). Pelo contrário, a proporção de pseudo inércia explicada calculada com D_1 é igual a $0,77$ quando $\alpha = 1$ e diminui quando α diminui (linha vermelha). Aqui, o gráfico sugere a escolha de $\alpha = 0,6$, pois as linhas se cruzam.

A intenção do algoritmo Ward-Like é agregar os dois grupos $\mathcal{A}$ e $\mathcal{B}$ de uma determinada partição $\mathcal{P}_k^\alpha$ em $k+1$ grupos, para que a nova partição tenha inércia mínima mista dentro do grupo. O problema de otimização pode ser expresso da seguinte forma:

$$\underset{\mathcal{A},\mathcal{B}\in\mathcal{P}_{k+1}^\alpha}{\arg\min} \quad \mathcal{I}_\alpha(\mathcal{A}\cup\mathcal{B}) - \mathcal{I}_\alpha(\mathcal{A}) - \mathcal{I}_\alpha(\mathcal{B}) \tag{2.17}$$

O pseudocódigo do algoritmo Ward-Like (Algoritmo 4) é descrito como:

Algoritmo 4 Algoritmo Ward-Like

Entrada: matrizes de distância D_0 e D_1 e α

1: Obter a partição em K grupos da partição em $k+1$ grupos

A cada passo, o algoritmo agrega os dois grupos $\mathcal{A}$ e $\mathcal{B}$ de $\mathcal{P}_k^\alpha$ de acordo com o problema de otimização 2.17 tal que o aumento da inércia intra-grupo seja mínimo para a partição selecionada sobre as demais em K grupos

2: **repita**

3: Mesclar os dois grupos $\mathcal{A}$ e $\mathcal{B}$ de modo que a medida de agregação δ_α seja mínima:

$$\delta_\alpha(\mathcal{A},\mathcal{B}) := W_\alpha(\mathcal{P}_{k+1}^\alpha) - W_\alpha(\mathcal{P}_k^\alpha) = I_\alpha(\mathcal{A}\cup\mathcal{B}) - I_\alpha(\mathcal{A}) - I_\alpha(\mathcal{B})$$

4: **até que** $K=1$, a partição $\mathcal{P}_1^\alpha =: \mathcal{P}_1$ em um grupo seja obtida

5: **repita**

6: **até que** tenha K grupos desejados.

Saída: os grupos formados

Ao considerar restrições espaciais, o agrupamento hierárquico Ward-Like torna-se um algoritmo completo na detecção de grupos em conjuntos de dados de diferentes dimensões.

2.2.5 Algoritmo SKATER

O algoritmo Spatial K'luster Analysis by Tree Edge Removal (SKATER)[4], proposto por Assunção et al. (2006), utiliza-se de grafos para capturar as relações de contiguidade entre os objetos. No grafo, cada objeto está associado a um vértice e ligado por arestas aos seus vizinhos. Esse algoritmo permite criar grupos espaciais contíguos e homogêneos em relação aos atributos de interesse.

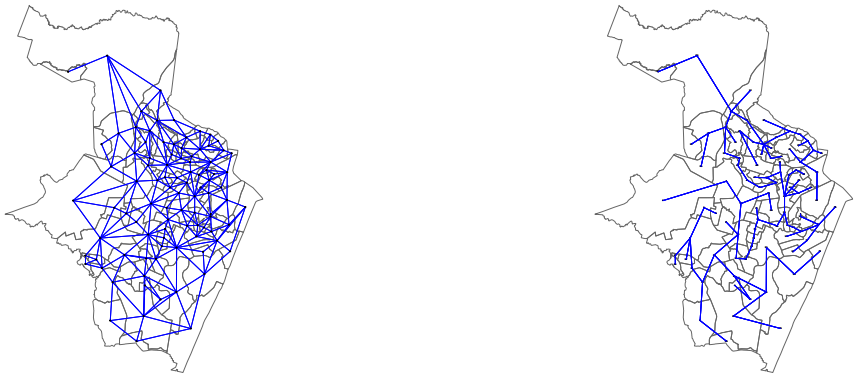
Considerando n regiões geográficas contíguas, usando como exemplo os bairros de Recife, o mapa é identificado com um grafo não direcionado $\mathcal{G} = (V, L)$, onde V é o conjunto de vértices ou nós que representam os bairros e L é o conjunto de arestas conectando pares de vértices e representando a relação de adjacência entre eles (veja a Figura 4a). Se existe uma aresta entre os vértices i e j dizemos que os bairros são vizinhos. Um caminho do nó v_1 para o nó v_k é uma sequência de nós $v_1, v_2, \ldots, v_k$. No caso de o primeiro e o último nó serem os mesmos, ou seja, $v_1 = v_k$, o caminho é designado como um circuito. Um grafo é dito conectado se, para qualquer par de nós v_i e v_j, existe pelo menos um caminho conectando-os (TEIXEIRA; ASSUNÇÃO; LOSCHI, 2019).

O processo de funcionamento do Skater é constituído por três etapas. Inicialmente, é gerado o grafo representativo do problema (Figura 4a), a partir dele, a Árvore Geradora Mínima (AGM), que representa um gráfico de vizinhança, é montada (Figura 4b), e em seguida é feita

[4] Implementado no software R e disponível no pacote spdep (BIVAND; PEBESMA; GOMEZ-RUBIO, 2013).

Figura 4 – Exemplo de mapa da cidade de Recife dividido em 94 bairros sobrepostos pelo grafo (Figura 4a) e exemplo de árvore geradora mínima (Figura 4b) construída pelo algoritmo Skater.

(a) Grafo (b) AGM



Fonte: Elaboração própria.

a partição dessa árvore para formar os grupos, de modo a conseguir a melhor similaridade entre as regiões.

Um árvore geradora $\mathcal{T}$ de um grafo $\mathcal{G}$ é um conceito importante em nossa análise. $\mathcal{T}$ é um subgrafo conectado sem circuitos contendo todos os nós de $\mathcal{G}$. Em uma árvore geradora, quaisquer dois nós de $\mathcal{G}$ são conectados por um único caminho e o número de arestas em $\mathcal{T}$ é $n-1$ (TEIXEIRA; ASSUNÇÃO; LOSCHI, 2019).

O algoritmo se inicia com uma árvore $\mathcal{T}_1$, contendo apenas um vértice. A cada iteração, uma nova aresta e um novo vértice são adicionados à árvore. Na iteração n, a árvore $\mathcal{T}_n$ contém todos os n vértices de V e um subconjunto, L_m de L, com $n-1$ arestas. O custo de cada aresta é proporcional à dissimilaridade do par de objetos (bairros) (ASSUNÇÃO et al., 2006).

Uma árvore geradora mínima é uma árvore geradora com custo mínimo, onde o custo é medido como a soma das dissimilaridades em todas as arestas da árvore. A AGM é única se os custos entre qualquer nó e todos os seus vizinhos forem distintos (TEIXEIRA; ASSUNÇÃO; LOSCHI, 2019).

O custo da distância $d(i,j)$ associado a aresta (v_i, v_j) mede a dissimilaridade entre os objetos i e j usando seus vetores de atributos x_i e x_j. Para criar a AGM adotam-se os procedimentos descritos em 5 (ASSUNÇÃO et al., 2006):

Após concluída a geração da AGM, a terceira etapa do processo consiste em particioná-la

Algoritmo 5 Construção da AGM

1: Tome qualquer vértice v_i e faça $\mathcal{T}_k = \mathcal{T}_1 = v_i$
2: Encontre a aresta de menor custo $(l') \in \mathcal{T}$ que conecta qualquer vértice de $\mathcal{T}_k$ a outro vértice, v_j, pertencente a V, mas não a $\mathcal{T}_k$
3: Acrescente o vértice v_j e a aresta l' à árvore $\mathcal{T}_k$, criando uma nova árvore $\mathcal{T}_{k+1}$
4: **repita**
os passos 2 e 3
5: **até que** todos os vértices tenham sido incluídos na árvore $(\mathcal{T}_n)$

para obter os grupos. O algoritmo usa uma função de custo para cada aresta l removida da árvore geradora mínima através de 2.18

$$f(S_l^{\mathcal{T}}) = SSD_{\mathcal{T}_i} - SSD_{\mathcal{A}} \tag{2.18}$$

onde $SSD_{\mathcal{T}_i}$ corresponde à soma dos quadrados dos desvios da árvore $\mathcal{T}_i$ e $SSD_{\mathcal{A}}$ à soma dos quadrados dos desvios das duas árvores $\mathcal{T}_{i1}$ e $\mathcal{T}_{i2}$ geradas a partir da eliminação da aresta l, $SSD_{\mathcal{A}} = SSD_{\mathcal{T}_{i1}} + SSD_{\mathcal{T}_{i2}}$.

O algoritmo de particionamento produz um grafo $\mathcal{G}^*$ que contém um conjunto de árvores $\mathcal{T}_1, \ldots, \mathcal{T}_n$ onde cada árvore é conectada, mas não possui arestas ou vértices em comum. Na primeira iteração, $\mathcal{G}^*$ tem apenas uma árvore, que é a AGM (árvore $\mathcal{T}$). A cada iteração, a AGM é dividida em duas sub-árvores ($\mathcal{T}_{i1}$ e $\mathcal{T}_{i2}$) removendo a primeira aresta e determinando seu custo. Novamente a árvore $\mathcal{T}$ é subdividida em duas, agora removendo a segunda aresta e calcula-se o seu custo. Repete-se o procedimento até que sejam determinados os custos de todas as arestas.

Feita todas essas divisões, as melhores soluções para cada uma das árvores $\mathcal{T}_1, \ldots, \mathcal{T}_n$ são comparadas através de uma função custo 2.18. A solução que maximiza essa função é a que melhor subdivide uma árvore $\mathcal{T}$ em duas novas árvores, também chamada de solução ideal S_*^R da função objetivo. Este processo é repetido até atingir o número de grupos desejados (ASSUNÇÃO et al., 2006).

A construção dos grupos é descrita em 6 (ASSUNÇÃO et al., 2006).

Finalmente, o algoritmo gera K grupos espaciais contíguos. A qualidade da divisão desses grupos pode ser medida pela soma dos quadrados dos desvios de cada grupo. O cálculo de $Q(\pi)$ para a partição dos dados em K grupos e da soma dos quadros dos desvios SSD_k para um grupo qualquer é realizado a partir das equações abaixo

$$Q(\pi) = \sum_{i=0}^{k} SSD_i, \tag{2.19}$$

Algoritmo 6 Algoritmo SKATER

Entrada: grafo $\mathcal{G}^* = \mathcal{T}_0$, dados e quantidade de cortes na árvore

1: Identifica-se a aresta que possui o custo $f(S_*^{\mathcal{T}^0})$ mais alto
2: **enquanto** a quantidade de sub-árvores de $\mathcal{G}$ for menor do que k **faça**
para todas as árvores em $\mathcal{G}^*$, selecione a $\mathcal{T}_i$ responsável por maximizar $f(S_*^{\mathcal{T}^i})$
divida $\mathcal{T}_i$ em duas novas subárvores e atualize $\mathcal{G}^*$
3: **fim do enquanto**

Saída: os grupos formados

$$SSD_k = \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ij} - \overline{x}_j)^2 \tag{2.20}$$

em que p é o número de atributos considerados na análise, n_k é o número de objetos espaciais no grupo K, x_{ij} é valor do j-ésimo atributo do objeto espacial i, $\overline{x}_j$ é a média do j-ésimo atributo para todos os objetos na árvore e π é uma partição dos objetos espaciais em k grupos e SSD_i é a soma dos desvios quadrados na região i. Assim, quanto menor o $Q(\pi)$, melhor a partição, pois regiões homogêneas produzem menores valores de SSD.

2.2.6 Validação dos Agrupamentos

Após as etapas de construção e aplicação dos algoritmos escolhidos, um momento importante desse processo é a avaliação dos grupos obtidos. Diferentes métricas de desempenho são usadas para avaliar diferentes algoritmos e verificar qual é o melhor em termos de separação e coesão. Os índices de validação de agrupamentos considerados foram Calinski-Harabasz, Dunn e Davies-Bouldin.

2.2.6.1 Índice Calinski-Harabasz

O índice Calinski-Harabasz (CH)[5], também conhecido como Critério de Razão de Variâncias, introduzido por Caliński e Harabasz (1974) pode ser usado para avaliar os agrupamentos, i.e. a validação de quão bem o agrupamento foi feito. O índice CH é uma medida de quão semelhante um objeto é ao seu próprio grupo em comparação com outros grupos.

Para um conjunto de dados X de tamanho n que foi agrupado em K grupos, o índice de Calinski-Harabasz é definido como a razão entre a dispersão média entre grupos e a dispersão

[5] Implementado no software R e disponível no pacote fpc (HENNIG, 2020).

dentro do grupo. O cálculo de CH é definido através da Expressão 2.21:

$$\mathrm{CH} = \frac{B(k)/(K-1)}{W(k)/(n-K)} \tag{2.21}$$

onde, n é o número de objetos, K é a quantidade de grupos, $W(k) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} ||x_i - c_j||^2$ é a dispersão interna dos grupos e $B(k) = \sum_{j=1}^{k} n_j ||c_j - c||^2$ é a dispersão entre os grupos, c_j o centro do j-ésimo grupo e c o centroide do conjunto de dados.

A partir da análise da Equação 2.21, como n e K são constantes se $B(k)$ for maior do que $W(k)$, o que indica grupos compactos e bem separados, o resultado dará alto. Desta forma, as soluções de grupo com valores maiores do índice correspondem as melhores soluções.

2.2.6.2 Índice Dunn

O índice de Dunn (DN)[6] Dunn (1974) também é uma métrica para avaliar o bom desempenho de algoritmos de agrupamento. Seu objetivo é identificar um conjunto de grupos compactos, com pequena variação entre os membros do grupo e bem separados dos membros de outros grupos.

A Equação 2.22 representa o cálculo do índice:

$$\mathrm{DN} = \min_{1 \le i \le K} \left\{ \min_{1 \le j \le K, i \ne j} \left\{ \frac{d(i,j)}{\max\limits_{1 \le h \le K} \{\Delta d(h)\}} \right\} \right\} \tag{2.22}$$

onde, K é o número de grupos, $d(i,j)$ representa a distância entre os grupos i e j e $\Delta d(h)$ mede a distância entre dois objetos do grupo h, o que pode ser considerado como uma medida de dispersão de agrupamento.

O índice compara as distâncias intergrupos com o tamanho do grupo mais disperso. Tem um valor que está no intervalo $[0, \infty)$, e quanto maior o seu valor, mais separados e compactos são os grupos.

2.2.6.3 Índice Davies-Bouldin

O índice de Davies-Bouldin (DB)[7] proposto por Davies e Bouldin (1979) é um esquema de avaliação interna, onde a validação de quão bem o agrupamento foi feito é dada usando

[6] Implementado no software R e disponível no pacote clValid (BROCK et al., 2008).
[7] Implementado no software R e disponível no pacote clusterSim (WALESIAK; DUDEK, 2020).

quantidades e recursos inerentes ao conjunto de dados.

O índice DB é definido como 2.23:

$$\mathrm{DB} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left\{ \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right\} \tag{2.23}$$

onde, K é o número de grupos, c_i é o centroide do grupo i, σ_i é a distância média de todos os elementos no grupo i ao centroide c_i e $d(c_i, c_j)$ é a distância entre os centroides c_i e c_j.

Nesse caso, pequenos valores de DB correspondem a grupos mais compactos e melhor será o resultado, significando, baixas medidas de dispersão intragrupo e grandes distâncias intergrupo.

3 MODIFICAÇÕES PROPOSTAS: ALGORITMOS DE AGRUPAMENTO ESPACIAIS PARA DADOS CRIMINAIS

Este trabalho apresentou inicialmente um estudo sobre o problema de agrupamento, onde foram descritos alguns algoritmos, cada um com suas particularidades. Durante o estudo, percebeu-se a dificuldade de determinadas técnicas em utilizar outros tipos de dados em seus atributos representados, por exemplo, como dados de contagem, categóricos e até mesmo dados com informações ausentes, os quais, por não terem uma métrica implícita, dificultam o trabalho dos algoritmos em termos de atribuição de pesos e distâncias para formação dos grupos.

Os dados de crimes possuem essas características. São dados que passam por uma rigorosa contagem dos crimes, todas as ocorrências são analisadas e conferidas. Porém, por vezes, possuem dados faltantes, ou pelo tipo de crime ocorrido, que não tem um desfecho concreto e por isso não é possível ter todas as informações do crime ou da vítima ou por problemas de registro nas bases oficiais.

Para realizar uma análise de agrupamento com dados desse tipo, a medida de distância sugerida é a distância Gower, que abrange as características do tipo de dado criminal citadas. Assim, propomos introduzir a distância Gower no funcionamento dos algoritmos Ward-Like e SKATER, que originalmente, usam distância Euclidiana.

A escolha desses algoritmos se dá pelo fato de serem algoritmos com restrições de contiguidade espacial, que no nosso objeto de estudo é fundamental termos grupos com objetos conexos. As extensões desses algoritmos foram construídas e chamadas de **Ward-Like.New** e **SKATER.New**, respectivamente.

Foram desenvolvidos algoritmos aptos para utilizar dados com informação criminal mas, bem como os outros métodos de agrupamento, também têm potencial de aplicação em qualquer outra situação, cujas características sejam similares aquelas de dados de crime.

3.1 WARD-LIKE.NEW

A modificação **Ward-Like.New** ocorreu em uma das etapas do algoritmo Ward-Like. Resumidamente, o algoritmo Ward-Like segue as etapas relatadas em 2.2.4:

1 Cálculo da matriz de dissimilaridade D_0;

2 Cálculo da matriz de dissimilaridade D_1; e

3 Definição do parâmetro de mistura α.

A modificação **Ward-Like.New** do algoritmo Ward-Like ocorreu na etapa 1, no cálculo da matriz de dissimilaridade D_0. Como visto em 2.2.4, essa matriz fornece as distâncias no "espaço de características" i.e. nas variáveis sem informação espacial, que no estudo de caso são determinadas pelas variáveis criminais. Não houve mudança no cálculo da matriz D_1 e do parâmetro α (etapas 2 e 3, respectivamente).

Para construir D_0 usando a distância Gower foi utilizada a função gower.dist do pacote StatMatch (D'ORAZIO, 2020) do software R. Esta função irá calcular a distância Gower entre os bairros no conjunto de dados determinado pelas informações criminais. Após o cálculo da distância, ela foi transformada em uma matriz pela função as.dist do pacote base do R que é usada para conversão entre objetos de classe "dist" e matrizes de distância e vice-versa.

Calculada a matriz D_0 com a distância Gower, o próximo passo é determinar a matriz D_1, que leva em consideração a restrição de vizinhança, dada pela transformação da lista de vizinhos em uma matriz de pesos espaciais, e determinar também o valor de α.

Após esses procedimentos, o algoritmo **Ward-Like.New** está pronto para ser utilizado.

Algoritmo 7 Algoritmo Ward-Like.New

1: Gerar D_0 pela distância Gower.
2: Transformar D_0 numa matriz.
Entrada: matriz de distância D_0, matriz de vizinhança D_1 e parâmetro α
3: Obter a partição em K grupos da partição em $k+1$ grupos.
A cada passo, o algoritmo agrega os dois grupos $\mathcal{A}$ e $\mathcal{B}$ de $\mathcal{P}_k^\alpha$ de acordo com o problema de otimização 2.17 tal que o aumento da inércia intra-grupo seja mínimo para a partição selecionada sobre as demais em K grupos.
4: **repita**
5: Mesclar os dois grupos $\mathcal{A}$ e $\mathcal{B}$ de modo que a medida de agregação δ_α seja mínima:
$$\delta_\alpha(\mathcal{A},\mathcal{B}) := W_\alpha(\mathcal{P}_{k+1}^\alpha) - W_\alpha(\mathcal{P}_k^\alpha) = I_\alpha(\mathcal{A}\cup\mathcal{B}) - I_\alpha(\mathcal{A}) - I_\alpha(\mathcal{B})$$
6: **até que** $K=1$, a partição $\mathcal{P}_1^\alpha =: \mathcal{P}_1$ em um grupo seja obtida.
7: **repita**
8: **até que** tenha K grupos desejados.
Saída: os grupos formados

Adiante, seguem trechos de código R usados na manipulação dos dados e na construção dos grupos espacias, tomando-se como exemplo a cidade de Recife. Os textos imediatamente após o símbolo # correspondem a comentários explicativos do código R, conforme apresentado a seguir:

```
#Primeiramente, carregamos os pacotes necessários
library(ClustGeo)
library(maptools)
library(spdep)
library(rgdal)
library(StatMatch)
library(readr)
#Ler do shapefile da cidade Recife
recife <- readOGR(...)
#Ler dos dados utilizados
dados <- read_csv(...)
set.seed(1938) #semente

#Calcular D0 usando o método Gower
D0gow = gower.dist(dados)
#Conversão de D0gow em uma matriz
D0gower = as.dist(D0gow)
#Criar a matriz de vizinhança
recife.nb <- poly2nb(recife)
A <- nb2mat(recife.nb, style="B")
#A matriz de dissimilaridade D1 é então 1 menos A
D1 <- as.dist(1-A)

#Escolher o parâmetro de mistura alpha
range.alpha <- seq(0,1,0.1)
K <- 5 #5 grupos
cr <- choicealpha(D0gower, D1, range.alpha, K, graph=F)

#Código para geração da Figura 5a
plot(cr)

cr$Q # proporção da inércia explicada
                 Q0         Q1
```

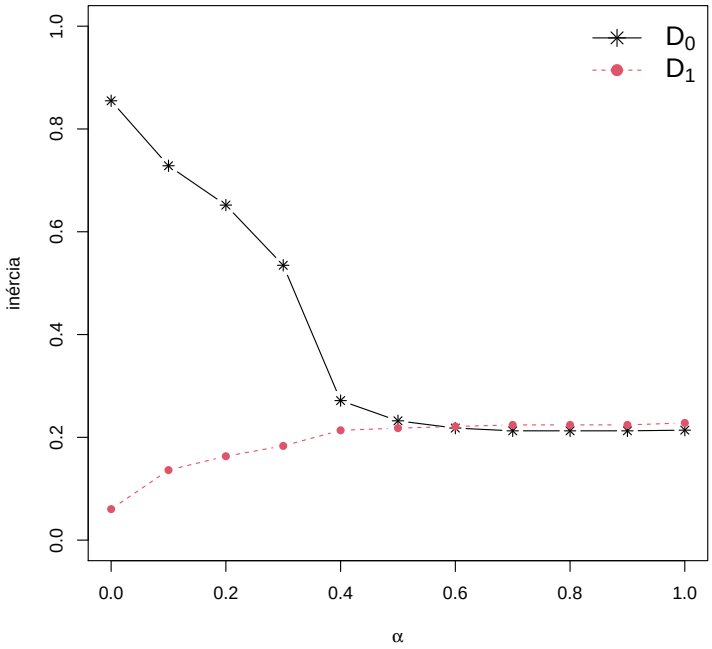
```
alpha=0   0.8548534 0.06031638
alpha=0.1 0.7284001 0.13621313
alpha=0.2 0.6518804 0.16320343
alpha=0.3 0.5347497 0.18328775
alpha=0.4 0.2714724 0.21364028
alpha=0.5 0.2321179 0.21790488
alpha=0.6 0.2180124 0.22109838
alpha=0.7 0.2126500 0.22417959
alpha=0.8 0.2126500 0.22417959
alpha=0.9 0.2126500 0.22417959
alpha=1   0.2139432 0.22802925
```

A inércia explicada calculada aqui com D_1 (curva vermelha) é menor do que a inércia explicada calculada com D_0 (curva preta). Para superar esse problema, a proporção normalizada da inércia explicada (Qnorm) é plotada.

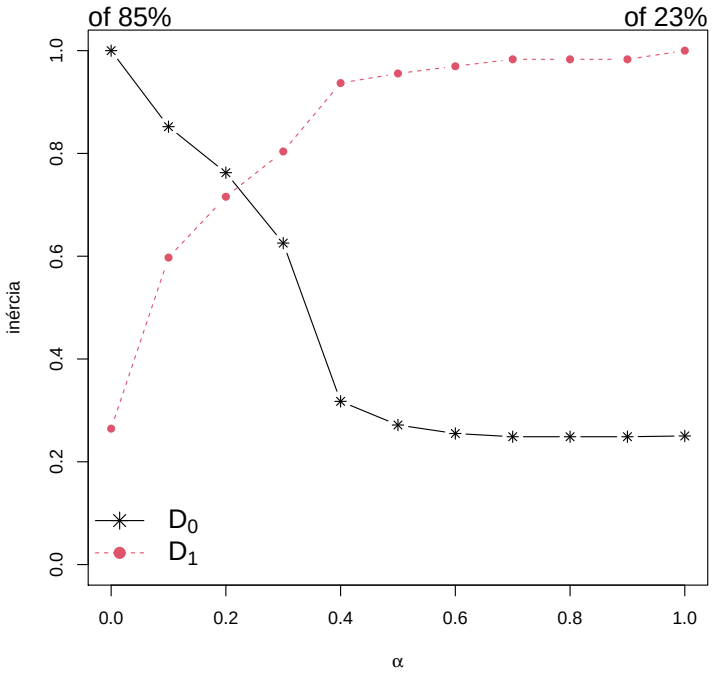
```
cr$Qnorm #proporção normalizada da inércia explicada
             Q0norm    Q1norm
alpha=0   1.0000000 0.2645116
alpha=0.1 0.8520761 0.5973494
alpha=0.2 0.7625640 0.7157127
alpha=0.3 0.6255455 0.8037905
alpha=0.4 0.3175660 0.9368986
alpha=0.5 0.2715294 0.9556006
alpha=0.6 0.2550290 0.9696054
alpha=0.7 0.2487561 0.9831177
alpha=0.8 0.2487561 0.9831177
alpha=0.9 0.2487561 0.9831177
alpha=1   0.2502689 1.0000000

#Código para geração da Figura 5b
plot(cr, norm = T)
```

Figura 5 – Inércias e α com distância Gower

(a) Proporção da inércia explicada versus α



(b) Proporção normalizada da inércia explicada versus α



Fonte: Elaboração própria.

Pela proporção da inércia explicada e Figura 5a vemos que a proporção de inércia explicada calculada com D_0 (as distâncias criminais) é igual a $0,85$ quando $\alpha = 0$ e diminui quando

α aumenta (linha preta). Pelo contrário, a proporção da inércia explicada calculada com D_1 (as distâncias espaciais) é igual a $0,22$ quando $\alpha = 1$ e diminui quando α diminui (linha vermelha). O traçado das curvas de Q_0 e Q_1 é uma ferramenta para escolher um valor de α que haja um equilíbrio entre a perda de homogeneidade criminal e o ganho de coesão espacial.

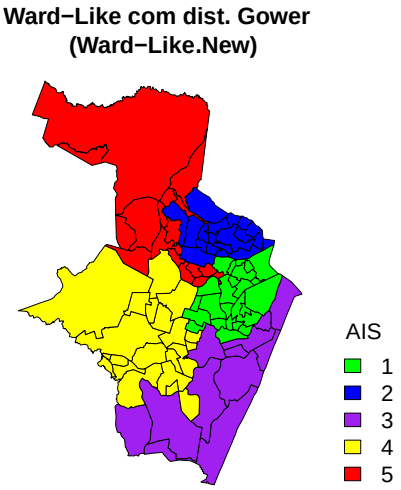
Em 5b com D_0, a curva começa em 100% e diminui à medida que α aumenta de 0 a 1. Com D_1, a curva começa em 100% (à direita) e diminui à medida que α diminui de 0 a 1. Este gráfico sugere a escolha de $\alpha = 0,2$, o que corresponde a uma perda de $(1 - 0,7625640 = 23,74\%)$ da homogeneidade criminal e um aumento de $(1 - 0,7157127 = 28,42\%)$ na homogeneidade geográfica.

Entendido o processo de escolha de α, o próximo passo é a execução do algoritmo:

```
#Implementar o Ward-Like.New
WL.New <- hclustgeo(D0,D1,alpha=0.2)
#Cortar o dendrograma obtido com a função hclustgeo e alpha=0,2
#para obter a nova partição em 5 clusters
WL.New5 <- cutree(WL.New, 5)
#Visualizar o número de bairros por grupo
table(WL.New5)
 1  2  3  4  5
21 24 12 18 19

#Código para geração da Figura 6
#Criando os grupos
grupoWLNewgow5 <- WL.New5
#Definir as cores para os grupos
cores = c("green", "blue", "purple","yellow", "red", "lightblue", "pink")
plot(recife, col = cores[grupoWLNewgow5])
```

Na Figura 6 temos o produto final da partição em 5 grupos do algoritmo **Ward-Like.New**. Assim, de acordo com a combinação dos pesos dados às diferenças geográficas, a solução terá grupos mais ou menos contíguos espacialmente.

Figura 6 – Grupos formados com o Ward-Like.New



Fonte: Elaboração própria.

Essa partição é espacialmente compacta, uma vez que as dissimilaridades são construídas a partir da matriz de adjacência que dá mais importância à contiguidade dos bairros.

3.2 SKATER.NEW

Já a modificação **SKATER.New** ocorrida no algoritmo SKATER ocorreu em três partes. Resumidamente, o algoritmo SKATER segue as etapas descritas em 2.2.5 e novamente repetidas:

1 Gerar o grafo representativo do agrupamento;

2 Calcular a Árvore Geradora Mínima (AGM); e

3 Fazer a partição na AGM para formar os cluters.

Foram três modificações do método, e elas foram realizadas no software R no código fonte do pacote spdep (BIVAND; PEBESMA; GOMEZ-RUBIO, 2013). Não houve mudança no processo de geração do grafo (etapa 1), as alterações ocorreram nas etapas 2 e 3. Na etapa 2, a modificação na construção da AGM, ocorreu na função nbcosts agora chamada de **nbcosts.new**. E na etapa 3, ocorreram duas modificações: uma na "poda"[1] da AGM, modificando a função prunecost

[1] No sentido de fazer cortes na AGM.

alterada para **prunecost.new** e outra no processo de particionamento intra grupo, na função ssw agora chamada de **ssw.new**. As alterações são descritas a seguir:

- **nbcosts.new**: A primeira modificação ocorreu na construção da Árvore Gerado Mínima. A construção da AGM é baseada em medidas de similaridade entre os objetos, analisando os custos das arestas do grafo gerado na etapa 1. Inicialmente os custos são calculados através de uma métrica que avalia a semelhança entre dois objetos. Essa métrica é medida pelo método de distância declarado na função. No algoritmo SKATER original os métodos de distância disponíveis são: Euclidiano, Manhattan, Máximo, Canberra, Binário ou Minkowski, e todos eles exigem que os atributos sejam variáveis aleatórias com variação contínua, caso contrário, uma mensagem de erro é apresentada e a AGM não é produzida. Foi incluída então, a distância Gower acrescentada no código original.

- **prunecost.new**: A segunda modificação ocorreu no corte da AGM para a formação dos grupos. Na nova função **prunecost.new**, também foi acrescentado o método de distância Gower como uma das opções a ser escolhida pelo usuário, dentre as outras já existentes (Euclidiano, Manhattan, Máximo, Canberra, Binário ou Minkowski) e modificada a função de custo **ssw.new** que está embutida em **prunecost.new**.

- **ssw.new**: A terceira e última modificação ocorreu na função de custo. A função custo é quem vai definir quais arestas serão removidas e como serão formados os grupos. Nesta etapa do procedimento a forma de atribuir custos às arestas é modificada, de modo a obter melhores resultados. É removida as arestas de menores custos e essa avaliação dos custos de cada partição é dada agora através da distância Gower.

Após as adaptações, o algoritmo **SKATER.New** está finalizado e pronto para ser utilizado.

Algoritmo 8 Algoritmo SKATER.New

1: Construir a AGM pela distância Gower, obtendo a nova função **nbcosts.new**.
2: Incluir a distância Gower nas funções de custo e corte da AGM, obtendo as novas funções **prunecost.new** e **ssw.new**, respectivamente.

Entrada: grafo $\mathcal{G}^* = \mathcal{T}_0$, dados, quantidade de cortes na árvore e novas funções.

3: Identifica-se a aresta que possui o custo $f(S_*^{\mathcal{T}^0})$ mais alto
4: **enquanto** a quantidade de sub-árvores de $\mathcal{G}$ for menor do que k **faça**
    para todas as árvores em $\mathcal{G}^*$, selecione a $\mathcal{T}_i$ responsável por maximizar $f(S_*^{\mathcal{T}^i})$
    divida $\mathcal{T}_i$ em duas novas subárvores e atualize $\mathcal{G}^*$
5: **fim do enquanto**

Saída: os grupos formados

Adiante, seguem trechos de código R usados para a manipulação dos dados e geração dos grupos espaciais, tomando-se como exemplo a cidade de Recife. Os textos imediatamente após o símbolo # correspondem a comentários explicativos do código R, conforme apresentado a seguir:

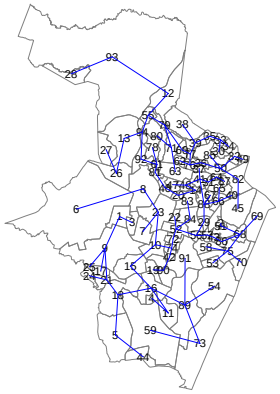
```
#Primeiramente, carregamos os pacotes necessários
library(maptools)
library(spdep)
library(rgdal)
library(cluster)
library(readr)
#Ler do shapefile da cidade Recife
recife <- readOGR(...)
#Ler dos dados utilizados
dados <- read_csv(...)
#Transformar dos dados para a mesma escala
sdat <-  scale(dados)
set.seed(1938) #semente

#Criar a matriz de vizinhança
recife.nb <- poly2nb(recife)
#Gerar a Árvore Geradora Mínima
# 1.Cálcular do custo das arestas incluindo o método Gower
lcosts <- nbcosts.new(recife.nb, sdat, method = "gower")
# 2.Transformar os custos de borda em pesos espaciais
recife.w <- nb2listw(recife.nb, lcosts, style="B")
# 3.Árvore Geradora Mínima
recife.mst <- mstree(recife.w)

#Código para geração da Figura 7a
plot((recife), border=gray(.5))
plot(recife.mst, coordinates(recife), col="blue",
     cex.lab=.8, cex.circles=0.5, add=TRUE)
```

```
#Partição da AGM em 5 grupos
#As funções prunecost.new e ssw.new estão embutidas na função skater.new
SK.New5 <- skater.new(recife.mst[,1:2], sdat, method = "gower", 4, crit = 10)
#Visualizar do número de bairros por grupo
table(SK.New5$groups)
 1  2  3  4  5
21 16 16 24 17

#Geração da Figura 7c
#Criando os grupos
grupoSKNewgow5 <- SK.New5$groups
#Definir as cores para os grupos
cores = c("green", "blue", "purple","yellow", "red")
plot(recife, col = cores[grupoSKNewgow5])

#Código para geração da Figura 7b
plot(recife, border=gray(.5))
plot(SK.New5, coordinates(recife), groups.colors=c("green", "blue", "purple","yellow","
cex.lab=.6, cex.circles=.5, add=T, bty="n", lwd = 2)
```

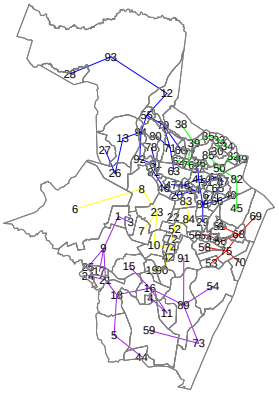
Na Figura 7a, apresenta-se a AGM para a região em estudo considerando as variáveis selecionadas, já na Figura 7b, tem-se a AGM particionada em cinco grupos e na Figura 7c, o produto final da regionalização do algoritmo SKATER.New (i.e. o algoritmo SKATER modificado para uso da distância Gower). Já a Figura 8c temos o algoritmo SKATER tradicional usando a distância Euclidiana, comparando, é possível verificar que o método SKATER.New produz grafos e, consequentemente, Árvores Geradoras Mínimas diferentes.

Figura 7 – Particionamento da Árvore Geradora Mínima em cinco grupos de bairros distintos para as Áreas Integradas de Segurança projetadas com o SKATER.New.

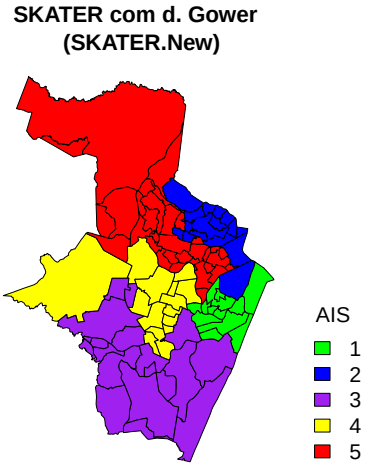
(a) AGM com a distância Gower



(b) AGM particionada com a distância Gower



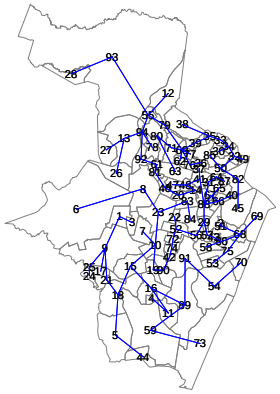
(c) Grupos formados com o SKATER.New



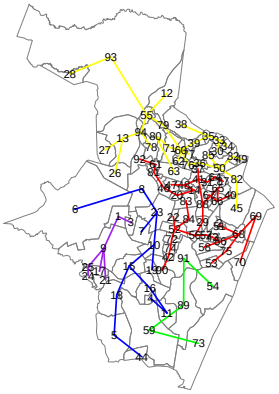
Fonte: Elaboração própria.

Figura 8 – Particionamento da Árvore Geradora Mínima em cinco grupos de bairros distintos para as Áreas Integradas de Segurança projetadas com o SKATER com distância Euclidiana.

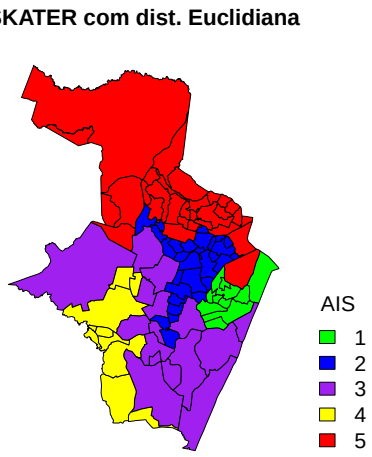
(a) AGM com a distância Euclidiana



(b) AGM particionada com a distância Euclidiana



(c) Grupos formados com o SKATER com a distância Euclidiana



Fonte: Elaboração própria.

A implementação e execução de pesquisas sobre a modificação do SKATER ainda são poucas, portanto, essa nova proposta implicará numa nova metodologia do algoritmo que privilegia o uso de dados mistos e também possibilidade de informações ausentes na base de dados. Comparamos este algoritmo com o original, ainda que os dois apresentem desempenhos semelhantes em termos de formação espacial dos grupos, o novo algoritmo possui desempenho médio superior em termos de homogeneidade interna dos grupos, confirmados pelos índices de validação e pelo indicador de qualidade de partição $Q(\pi)$ demonstrados no Capítulo 5.

4 ANÁLISE DE ÁREAS INTEGRADAS DE SEGURANÇA DE RECIFE - DADOS CRIMINAIS

Neste capítulo são apresentadas as atuais Áreas Integradas de Segurança do município de Recife e a base de dados utilizada na aplicação dos algoritmos de agrupamento estudados. A seção 4.1 apresenta as Áreas Integradas de Segurança e a seção 4.2 descreve os dados.

4.1 ÁREAS INTEGRADAS DE SEGURANÇA ATUAIS

A análise dos registros de crimes em uma determinada localidade é essencial para o provimento de informação sobre a criminalidade e suas tendências, objeto de interesse por parte da sociedade, particularmente dos governantes, formuladores de políticas públicas e agentes da segurança pública, principalmente no âmbito de planejamento e distribuição de recursos do Estado para prevenção e supressão das atividades criminais.

O uso de de uma metodologia adequada para analisar esses dados com base na frequência de ocorrência dos crimes cometidos em determinada área é um aspecto importante que deve ser abordado. Deste modo, a disponibilidade de dados de violência permite o uso de métodos estatísticos que busquem revelar eventuais padrões de regularidade de crimes.

A implantação do Plano Estadual de Segurança Pública do Estado de Pernambuco, ocorrida em 2007 no Pacto pela Vida, teve como foco principal integrar a atuação das polícias, além de fazer coincidir as áreas geo-técnicas de atuação policial, mediante um planejamento comum de ações e operações para definir conjuntamente os objetivos, estratégias e metas de enfrentamento à criminalidade. As Áreas Integradas de Segurança, seguindo a metodologia delineada por esse novo modelo de gestão, ficaram responsáveis em repassar informações e identificar as demandas e problemas numa mesma área de responsabilidade territorial. Consequentemente, com base nesses elementos, os órgãos operativos podem deliberar soluções e ter uma maior efetividade das ações.

Atualmente, as Áreas Integradas de Segurança dividem os 94 bairros do Recife geograficamente em cinco áreas, nomeadas por AIS 1, AIS 2, AIS 3, AIS 4 e AIS 5, que são supervisionadas, cada uma, por dois responsáveis, um da delegacia seccional e outro do batalhão de polícia militar (PERNAMBUCO, 2010). Essas áreas, como já mencionado, surgiram em 2007, com a implantação do Pacto Pela Vida (PPV), com isso, a informação criminal passou a ser gerada por área, sendo possível entender as diferentes realidades do crime em Pernambuco.

As AIS foram formadas de acordo com a divisão territorial já existente e com estudos técnicos específicos utilizando critérios cartográficos. Tal recorte, tem em vista uma correspondência geográfica que relaciona a área de cobertura de um Batalhão de Polícia Militar e uma ou mais circunscrições de Polícia Civil, seu objetivo é a articulação territorial regional, no nível tático, da Polícia Civil com a Polícia Militar (PERNAMBUCO, 2010; PERNAMBUCO, 2014).

Os bairros que compõem cada AIS da capital Recife são apresentados na Tabela 2:

Tabela 2 – Bairros que compõem cada Área Integrada de Segurança da cidade de Recife

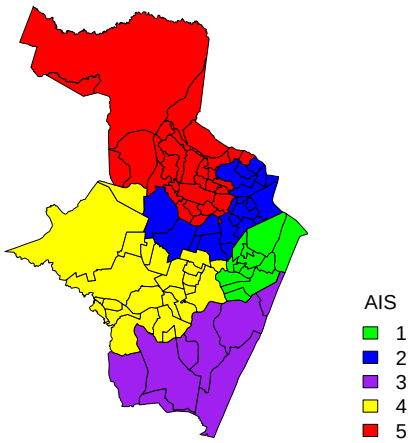
AIS	**Circunscrições**	**Bairros**
1 - Rio Branco	1ª – Rio Branco	Do Recife, Santo Antônio, São José e Cabanga
	2ª – Boa Vista	Boa Vista, Soledade e Santo Amaro
	3ª – Joana Bezerra	Ilha Joana Bezerra, Coelhos, Ilha do Leite e Paissandu
2 - Espinheiro	4ª – Espinheiro	Derby, Graças, Espinheiro, Aflitos, Rosarinho, Encruzilhada, Torreão, Ponto de Parada, Hipódromo e Campo Grande
	6ª – Cordeiro	Madalena, Torre, Zumbi, Cordeiro e Iputinga
	16ª – Água Fria	Água Fria, Arruda, Campina do Barreto, Cajueiro, Fundão e Peixinhos
3 - Boa Viagem	7ª – Boa Viagem	Brasília Teimosa, Pina e Boa Viagem
	8ª – Jordão	Jordão e Ibura
	9ª – Ipsep	Imbiribeira e Ipsep
	10ª – Cohab	Cohab
4 - Várzea	11ª – Afogados	Afogados, Jiquiá, Areias, Caçote e Estância
	12ª – Tejipió	Jardim São Paulo, Barro, Tejipió, Sancho, Totó e Coqueiral
	13ª – Mustardinha	Ilha do Retiro, Prado, Bongi, Mustardinha, Mangueira e San Martin
	14ª – Várzea	Torrões, Curado, Engenho do Meio, Cidade Universitária, Várzea e Caxangá
5 - Apipucos	5ª – Casa Amarela	Jaqueira, Santana, Poço da Panela, Parnamirim, Casa forte, Tamarineira e Casa Amarela
	15ª – Alto do Pascoal	Bomba do Hemetério, Alto Santa Terezinha, Alto José Bonifácio, Linha do Tiro, Dois Unidos, Passarinho, Beberibe e Porto da Madeira
	17ª – Vasco da Gama	Alto José do Pinho, Mangabeira, Morro da Conceição e Vasco da Gama
	18ª – Macaxeira	Macaxeira, Apipucos, Brejo de Beberibe, Brejo da Guabiraba, Córrego do Jenipapo, Dois Irmãos, Guabiraba, Monteiro, Alto do Mandu, Sítio dos Pintos, Nova Descoberta e Pau Ferro

Fonte: Elaboração própria a partir da Portaria GAB/SDS N° 1197, de 11 de junho de 2010.

O mapeamento dessas áreas leva em consideração, principalmente, a contiguidade dos bairros, como mostra o mapa de Recife e sua divisão das AIS (Figura 9), logo abaixo desta figura, é descrita a quantidade de bairros em cada AIS e sua respectiva porcentagem. A AIS 1, chamada de Rio Branco é composta por 11 bairros, o que corresponde a 21% do total de bairros da cidade, a AIS 2, Espinheiro, é formada por 21 bairros (22%), a menor AIS é a 3,

chamada de Boa Viagem, que possui 8 bairros (9%) localizados na parte sul do mapa, já a AIS 4, Várzea, é formada por 23 bairros (24%) e por fim, a AIS 5, Apipucos, a maior de todas, é formada por 31 bairros (33%) e localizada no norte de Recife.

Figura 9 – AIS atuais de Recife



(a) AIS - nº de bairros (%):
AIS 1 - 11 (21%)
AIS 2 - 21 (22%)
AIS 3 - 08 (09%)
AIS 4 - 23 (24%)
AIS 5 - 31 (33%)

Fonte: Elaboração própria.

Pretendeu-se nesta parte da pesquisa construir novas Áreas por meio das técnicas de agrupamento explanadas anteriormente. O propósito é analisar e discutir novas especificações para as Áreas Integradas de Segurança, levando em consideração não apenas a divisão territorial, como também, os dados de crimes ocorridos nos próprios bairros. Espera-se com isso oferecer um ferramenta estratégica inteligente de combate a crimes, fortalecendo a capacidade do Estado em gerir a violência e aumentando o potencial de criação de políticas públicas construídas com base em evidência de dados.

Esses novos mapeamentos consistem em novos cenários para as AIS e almeja contribuir para a sistematização e produção de uma metodologia de classificação que possa ser atualizada e corrigida a partir do momento em que futuros dados sejam disponibilizados. Isto porque

entende-se que as Áreas Integradas de Segurança podem ser modificadas ao longo do tempo, desde que surjam novos elementos que comprometam a real representação das regiões.

É verificado em outras cidades do Brasil, que novas demarcações das áreas foram realizadas ao longo do tempo, visando uma maior eficiência nos resultados de combate a crimes nessas regiões. Por exemplo, a capital cearense, Fortaleza, criou mais quatro Áreas Integradas de Segurança desde a sua elaboração, segundo a Portaria Normativa nº 436/2017 a criação de mais áreas leva em conta a importância da atualização dessas regiões (CEARÁ, 2017). No Rio de Janeiro, ponderando a importância da atualização da compatibilização e integração territorial das regiões, a capital por meio da Resolução nº 607 também fez alterações nas Áreas Integradas de Segurança Pública (AISP) da cidade (RIO DE JANEIRO, 2003).

Este tipo de estudo pode ser útil para futuras pesquisas, como também, pode ser ampliado para outras regiões. Diferentes algoritmos foram usados para chegar a um bom resultado neste objetivo. Cada um deles foi experimentado e testado e, finalmente, avaliamos quais deles funcionam melhor.

4.2 BASE DE DADOS

Os dados utilizados para formar os agrupamentos pertencem a uma base de domínio privado da Secretaria de Defesa Social de Pernambuco que foi obtida por meio de ofício e contém, para o período de Janeiro de 2007 a Dezembro de 2015, ocorrências de Crimes Violentos Letais Intencionais (CVLI's), lesão corporal e estupro na cidade de Recife. O indicador CVLI que engloba os homicídios, feminicídios, latrocínios, as lesões corporais seguidas de morte e as mortes decorrentes de confrontos policiais, nesse estudo foi chamado de crime letal.

Não foi possível atualizar a base para anos mais recentes devido ao Termo de Classificação de Informação (TCI) nº 01/2015 (PERNAMBUCO, 2012c), vinculado em 2015, que classifica as informações relativas ao bairro onde foi cometido o crime violento como sigilosas. O termo está fundamentado no Decreto nº 38.787 de 30/10/2012 (PERNAMBUCO, 2012a), o qual regulamenta a lei nº 14.804 de 29/10/2012 (PERNAMBUCO, 2012b) que regula o acesso a informação no âmbito do Poder Executivo Estadual.

O banco de dados é composto por 11 atributos/variáveis com as informações sobre a ocorrência de crimes, o sexo da vítima e também informações espaciais. As variáveis relacionadas ao sexo e o tipo de crime referem-se a frequência absoluta das mesmas, já as variáveis espaciais referem-se a latitude e a longitude do bairro. Feitas as modificações necessárias, seleção,

tabulação e padronização, o conjunto de dados passou a ter os bairros como as unidades de análise, e não os registros de crimes individuais.

Na Tabela 3, encontram-se o nome da variável, sua descrição, a média, o desvio padrão (DP), o valor mínimo e máximo entre os bairros, total e a porcentagem que cada variável assume. A amostra desse estudo é composta por 66.682 crimes, desse total 351 são os valores chamados "NaN" (do inglês "Not a Number"), que são os valores faltantes. Os valores faltantes foram substituídos por zero, em alguns casos este tipo de tratamento seria inadequado, pois se a informação observada e faltante tiverem caraterísticas diferentes a análise realizada é inadequada, porém, como estamos trabalhando com dados de contagem, onde o interesse da base de dados é informar quantos crimes ocorreram em determinado bairro, não há perda de informação, já que são poucos dados faltantes.

A frequência de vítimas mulheres (FEM) foi de $56,2\%$ $(n = 37.530)$, de homens (MAS) $43,5\%$ $(n = 29.020)$ e de vítimas que não tiveram seu sexo identificado (DES) apenas $0,3\%$ $(n = 214)$. Em relação aos tipos de crime, a lesão corporal (LCO) é o crime com maior incidência $(68,1\%)$, seguido de lesão corporal doméstica (LCD) com a porcentagem de $11,7\%$, crimes letais (LET) com $9,1\%$, os crimes de lesão corporal ocorridas no trânsito (LCT) correspondentes a $5,5\%$ dos crimes, estupro (EST) $4,9\%$ e outros tipos de lesão (OUT) $(0,6\%)$.

Tabela 3 – Variáveis utilizadas para os agrupamentos

Nome	Descrição	Média; DP; [Min, Max]	Total	%
MAS	Quantidade de vítimas homens	$309; 315,12; [1, 1940]$	29.020	$43,5$
FEM	Quantidade de vítimas mulheres	$399; 403,79; [0, 1968]$	37.530	$56,2$
DES	Quantidade de vítimas de sexo desconhecido	$2,28; 2,93; [0, 15]$	214	$0,3$
EST	Quantidade de vítimas de estupro	$34,9; 37,17; [0, 239]$	3.282	$4,9$
LET	Quantidade de vítimas de crimes letais	$64,7; 72,55; [0, 348]$	6.080	$9,1$
LCO	Quantidade de vítimas de lesão corporal	$483; 498,47; [0, 2884]$	45.433	$68,1$
OUT	Quantidade de vítimas de outros tipos de lesão	$4,5; 5,2; [0, 22]$	423	$0,6$
LCD	Quantidade de vítimas de lesão corporal doméstica	$82,9; 84,47; [0, 439]$	7.797	$11,7$
LCT	Quantidade de vítimas de lesão corporal de trânsito	$39,0; 53,59; [0, 342]$	3.667	$5,5$
LAT	Latitude do bairro	$-34,9; 0,04; [-35,0, -34,9]$	-	-
LON	Longitude do bairro	$-8,05; 0,03; [-8,14, -7,96]$	-	-

Fonte: Elaboração própria.

Para a implementação dos algoritmos e demais análises foi utilizado o software R (R Core Team, 2023) e seus pacotes disponíveis.

5 RESULTADOS

Nesta seção os resultados provenientes dos cinco métodos de agrupamento, K-Means, PAM, VNSKMED, Ward-Like e SKATER, com as três distâncias utilizadas, Euclidiana, Manhattan e Gower, incluindo as novas propostas de implementação Ward-Like-New e SKATER.New, são apresentados e avaliados. A base de dados utilizada para o estudo de caso é composta por crimes ocorridos no município de Recife e por variáveis georreferenciais dos bairros, tais como latitude e longitude. As análises são baseadas em mapas, na apuração dos índices de validação e na análise descritiva e exploratória dos novos grupos obtidos. Os grupos são representados por cores distintas e abaixo de cada mapa, estão detalhadas a quantidade de bairros e entre parênteses a porcentagem correspondente a cada grupo.

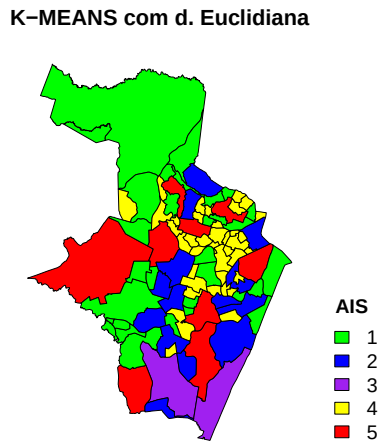
Com as técnicas de agrupamento apresentadas anteriormente, foram criados agrupamentos com cinco grupos ($K = 5$). A escolha do número de grupos deu-se avaliando a classificação atual das AIS, que é formada por cinco grupos.

O primeiro método avaliado foi o K-Means. Ele é conhecido por ser de rápida convergência para o resultado final. De fato, a rapidez foi comprovada com os experimentos realizados, como mostra a Tabela 4, em menos de um segundo, o algoritmo foi executado. Porém, apesar de ser um método que pode gerar grupos contíguos, essa característica não é garantida, já que ele não se baseia na conexidade dos objetos do grupo, e sim na semelhança dos objetos mais próximos do "centro" do grupo, e como mostram os mapas na Figura 10, o algoritmo não produziu grupos contíguos, independente da distância aplicada.

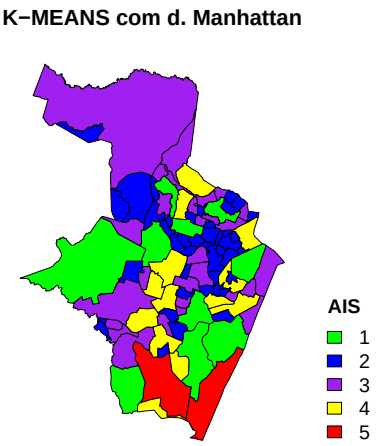
Além disso, no K-Means, os grupos possuem tamanhos mais discrepantes. Algumas AIS agruparam mais de 40% dos bairros, são elas: AIS 4 composta por 38 bairros (40%) na distância Euclidiana, AIS 2 com 41 bairros (44%) na distância Manhanttan e AIS 4 com 41 bairros (44%) na distância Gower, esses são os maiores grupos entre os cinco métodos analisados. O menor grupo, também foi gerado pelo K-Means, com 2 bairros, AIS 3, AIS 5 e AIS 3, nas distâncias Euclidiana, Manhattan e Gower, respectivamente, representando 2% do total de bairros de Recife. Na observação dos índices de validação, descritos na Tabela 2.2.6, o K-Means apresentou, dentre os demais algoritmos, um bom resultado do índice CH, aliás, o segundo melhor resultado nas três distâncias, CH $= 47,69$ na distância Euclidiana e CH $= 47,49$ nas distâncias Manhattan e Gower. Logo, por produzir grupos com muito mais bairros que outros, o método K-Means pode se tornar menos eficaz do que os outros métodos

em cenários com maior variação não espacial.

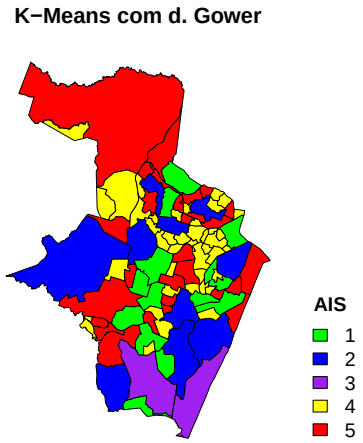
Figura 10 – Resultados do algoritmo K-Means



(a) AIS - nº de bairros (%):
AIS 1 - 31 bairros (33%)
AIS 2 - 14 bairros (15%)
AIS 3 - 02 bairros (02%)
AIS 4 - 38 bairros (40%)
AIS 5 - 09 bairros (10%)



(b) AIS - nº de bairros (%):
AIS 1 - 10 (11%)
AIS 2 - 41 (44%)
AIS 3 - 28 (30%)
AIS 4 - 13 (14%)
AIS 5 - 02 (02%)



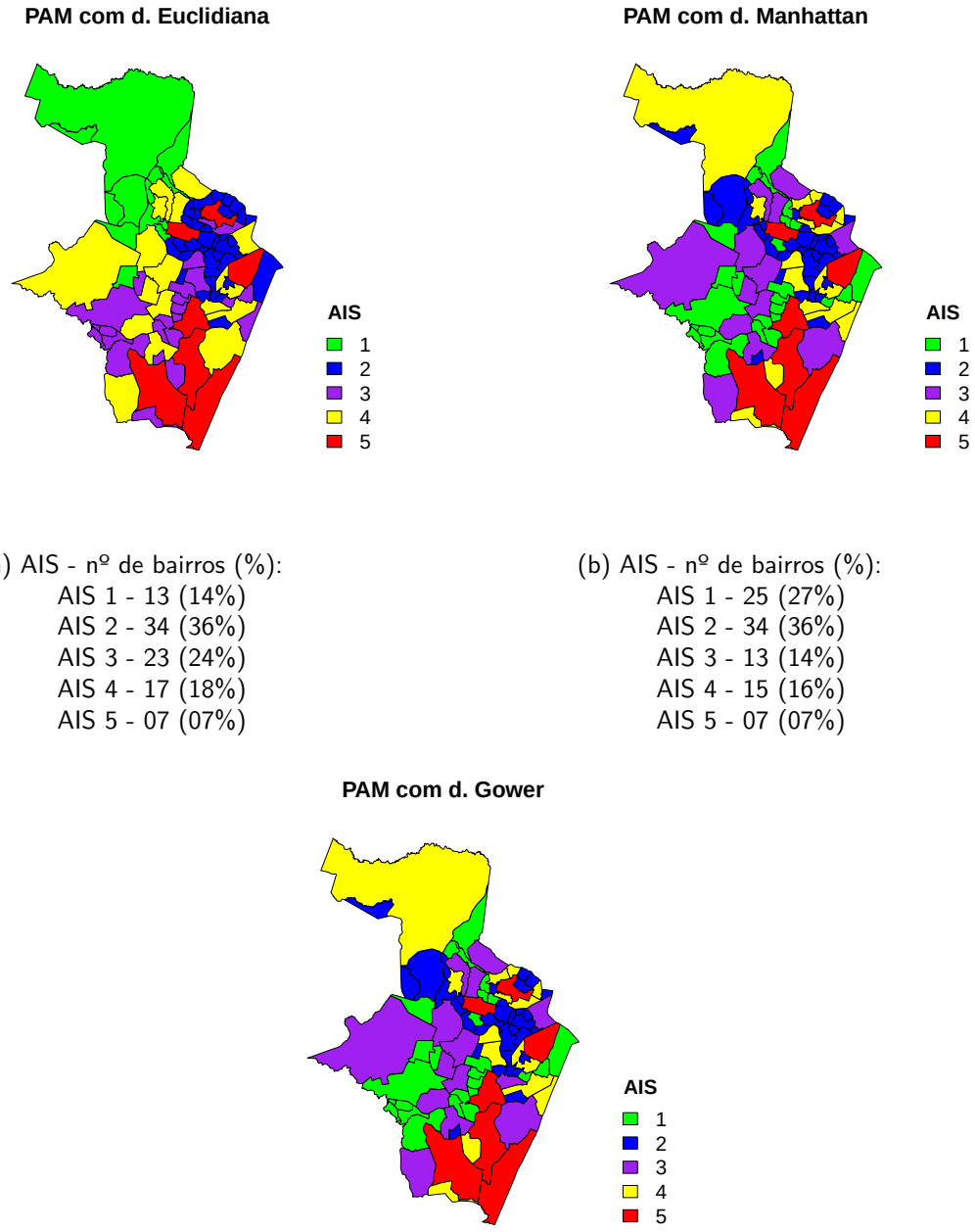
(c) AIS - nº de bairros (%):
AIS 1 - 13 (14%)
AIS 2 - 10 (11%)
AIS 3 - 02 (02%)
AIS 4 - 41 (44%)
AIS 5 - 28 (30%)

Fonte: Elaboração própria.

O método PAM, por sua vez, retorna o segundo maior grupo, a AIS 2 nas três distâncias

utilizadas, contém 34 bairros, que representa 36% dos bairros da cidade de Recife.

Figura 11 – Resultados do algoritmo PAM



(a) AIS - nº de bairros (%):
AIS 1 - 13 (14%)
AIS 2 - 34 (36%)
AIS 3 - 23 (24%)
AIS 4 - 17 (18%)
AIS 5 - 07 (07%)

(b) AIS - nº de bairros (%):
AIS 1 - 25 (27%)
AIS 2 - 34 (36%)
AIS 3 - 13 (14%)
AIS 4 - 15 (16%)
AIS 5 - 07 (07%)

(c) AIS - nº de bairros (%):
AIS 1 - 25 (27%)
AIS 2 - 34 (36%)
AIS 3 - 14 (15%)
AIS 4 - 14 (15%)
AIS 5 - 07 (07%)

Fonte: Elaboração própria.

Já o menor grupo desse algoritmo, a AIS 5, foi composto por 7 bairros, o que corresponde a 7% dos bairros da cidade. Os grupos construídos com as distâncias Manhattan e Gower foram bem parecidos, no quesito nº de bairros por grupo, houve diferença apenas de um bairro nas AIS's 3 e 4. Também não formou grupos contíguos, mas demonstrou um bom resultado no índice CH, utilizando a distância euclidiana $(\text{CH} = 43,56)$ e na rapidez em formar os grupos (menos de 1 segundo).

Já o VNSKMED, além de não haver contiguidade, o método apresentou uma grande desvantagem em relação aos demais métodos: o tempo de execução. Enquanto os demais métodos consumiram menos de 1 segundo, o VNSKEMD precisou de mais de 6 segundos para produzir a solução, como se verifica na Tabela 4.

Tabela 4 – Tempo de processamento por algoritmo (em segundos)

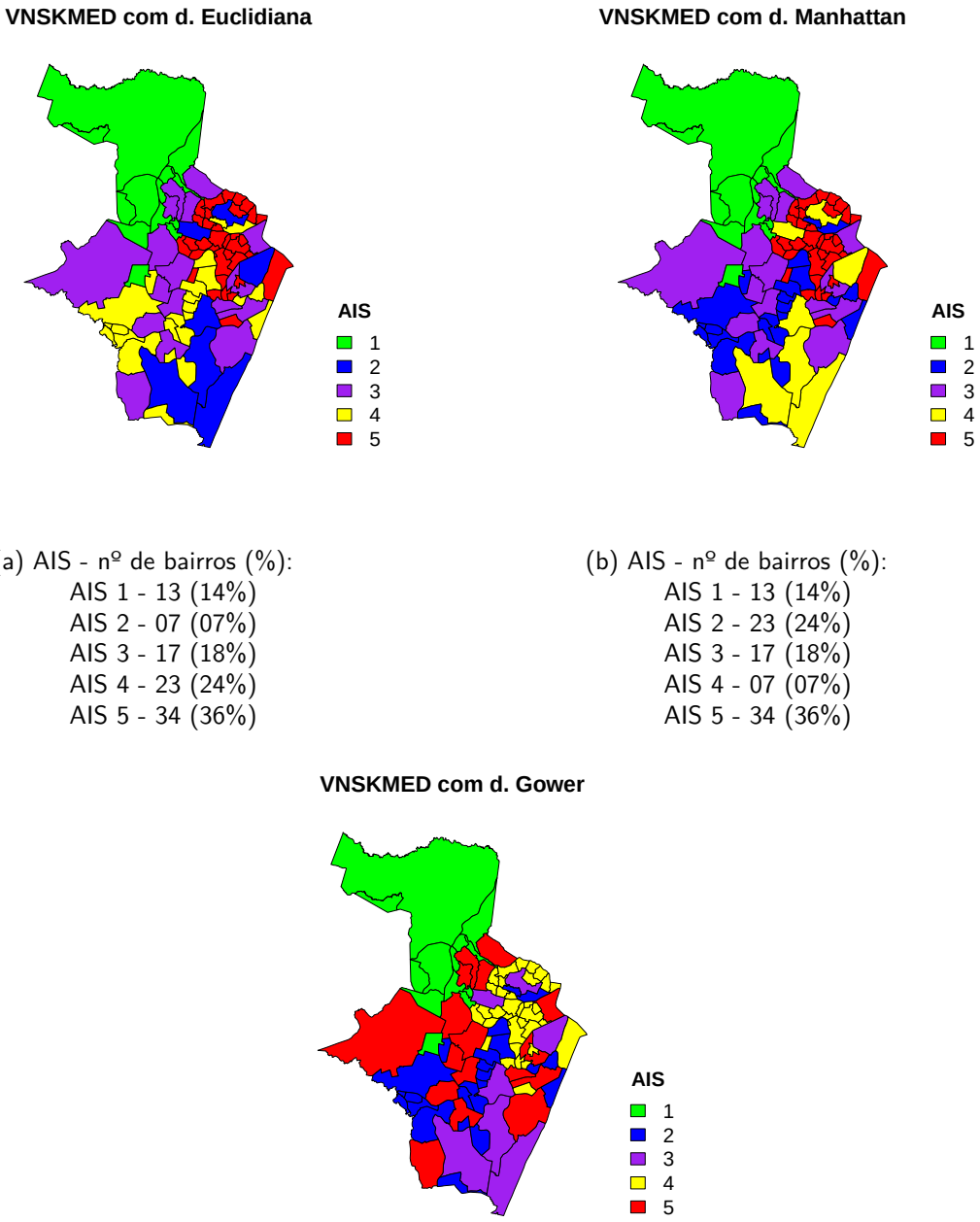
	Distância		
Método	Euclidiana	Manhattan	Gower
K-Means	< 1	< 1	< 1
PAM	< 1	< 1	< 1
VNSKMED	$6,2$	$5,8$	$6,0$
Ward-Like	< 1	< 1	< 1
SKATER	< 1	< 1	$1,5$

Fonte: Elaboração própria

Todavia, o método não oscilou no número de bairros por grupo, uma característica do algoritmo. Em qualquer distância utilizada os grupos foram formados por 7, 13, 17, 23 e 34 bairros, a mudança ocorrida foi na migração de bairros de uma AIS para outra.

Com relação aos índices de validação, o VNSKMED apresentou bons valores em todas as três distâncias. Inclusive, apresentou o melhor resultado do índice CH, entre todos os outros algoritmos ($\text{CH} = 71,35$ nas distâncias Euclidiana e Manhattan e $\text{CH} = 71,34$ na distância Gower).

Figura 12 – Resultados do algoritmo VNSKMED



(a) AIS - nº de bairros (%):
AIS 1 - 13 (14%)
AIS 2 - 07 (07%)
AIS 3 - 17 (18%)
AIS 4 - 23 (24%)
AIS 5 - 34 (36%)

(b) AIS - nº de bairros (%):
AIS 1 - 13 (14%)
AIS 2 - 23 (24%)
AIS 3 - 17 (18%)
AIS 4 - 07 (07%)
AIS 5 - 34 (36%)

(c) AIS - nº de bairros (%):
AIS 1 - 23 (24%)
AIS 2 - 17 (18%)
AIS 3 - 07 (07%)
AIS 4 - 13 (14%)
AIS 5 - 34 (36%)

Fonte: Elaboração própria.

Como visto, os algoritmos K-means, PAM e VNSKKMED não serviram para obter novas

Áreas Integradas de Segurança, pois não classificaram os bairros de forma contígua. Todavia, em alguns problemas de agrupamento, é relevante impor restrições de contiguidade no espaço em estudo. Tais restrições ocorrem quando os objetos em um grupo precisam não apenas ser semelhantes entre si, mas também compreender um conjunto de objetos vizinhos. Aqui, estamos definindo grupos contíguos como aqueles que possuem bairros vizinhos.

Já os algoritmos Ward-like e SKATER foram então desenvolvidos, incluindo restrições de contiguidade ou restrições espaciais. Dessa forma, esses algoritmos produzem grupos espacialmente coesos, compostos por bairros vizinhos, em comparação aos outros métodos de agrupamentos.

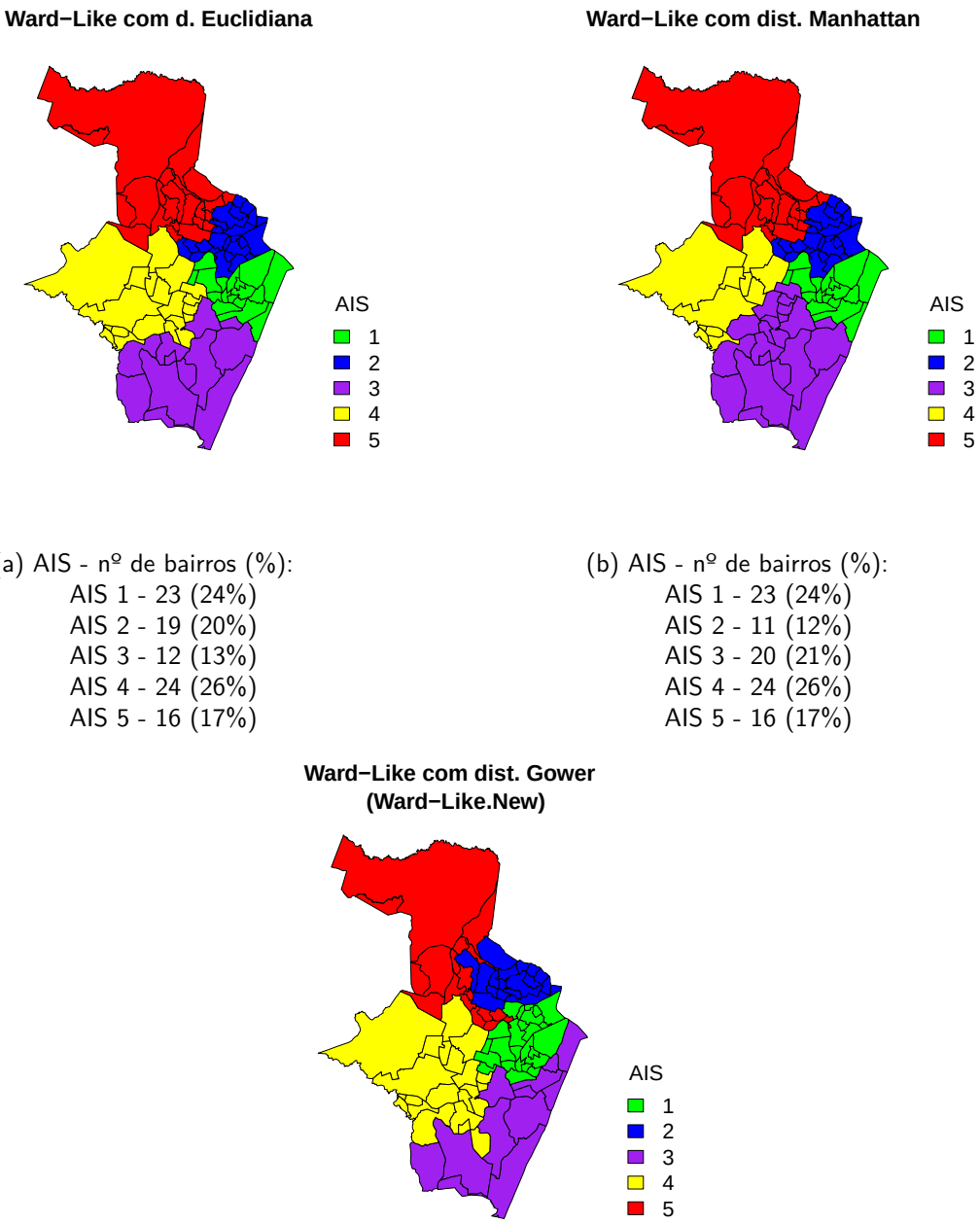
Tabela 5 – Comparação dos índices de validação

		Distância		
Índice	Método	Euclidiana	Manhattan	Gower
CH	K-Means	$\mathbf{47,69}$	$\mathbf{47,49}$	$\mathbf{47,49}$
	PAM	$\mathbf{43,56}$	$40,50$	$40,44$
	VNSKMED	$\mathbf{71,35}$	$\mathbf{71,35}$	$\mathbf{71,34}$
	Ward-Like	$5,71$	$3,59$	$\mathbf{7,57}$
	SKATER	$\mathbf{9,63}$	$\mathbf{10,29}$	$\mathbf{9,42}$
DN	K-Means	$1,66$	$1,68$	$1,68$
	PAM	$1,40$	$1,82$	$1,83$
	VNSKMED	$5,71$	$5,71$	$5,71$
	Ward-Like	$6,46$	$6,29$	$\mathbf{12,05}$
	SKATER	$2,42$	$2,15$	$2,46$
DB	K-Means	$0,11$	$0,10$	$0,10$
	PAM	$0,06$	$0,08$	$0,08$
	VNSKMED	$0,03$	$0,03$	$0,03$
	Ward-Like	$0,02$	$0,02$	$0,03$
	SKATER	$\mathbf{0,04}$	$\mathbf{0,02}$	$\mathbf{0,04}$

Fonte: Elaboração própria.

Na análise da Figura 13, o método Ward-Like forma grupos contíguos e uniformes. Seu tempo de processamento foi de menos de 1 segundo. Além disso, em comparação com os demais métodos, o Ward-Like apresentou os melhores valores dos índices DN e DB, sobretudo na distância Gower (DN $= 12,05$).

Figura 13 – Resultados do algoritmo Ward-Like



(a) AIS - nº de bairros (%):
AIS 1 - 23 (24%)
AIS 2 - 19 (20%)
AIS 3 - 12 (13%)
AIS 4 - 24 (26%)
AIS 5 - 16 (17%)

(b) AIS - nº de bairros (%):
AIS 1 - 23 (24%)
AIS 2 - 11 (12%)
AIS 3 - 20 (21%)
AIS 4 - 24 (26%)
AIS 5 - 16 (17%)

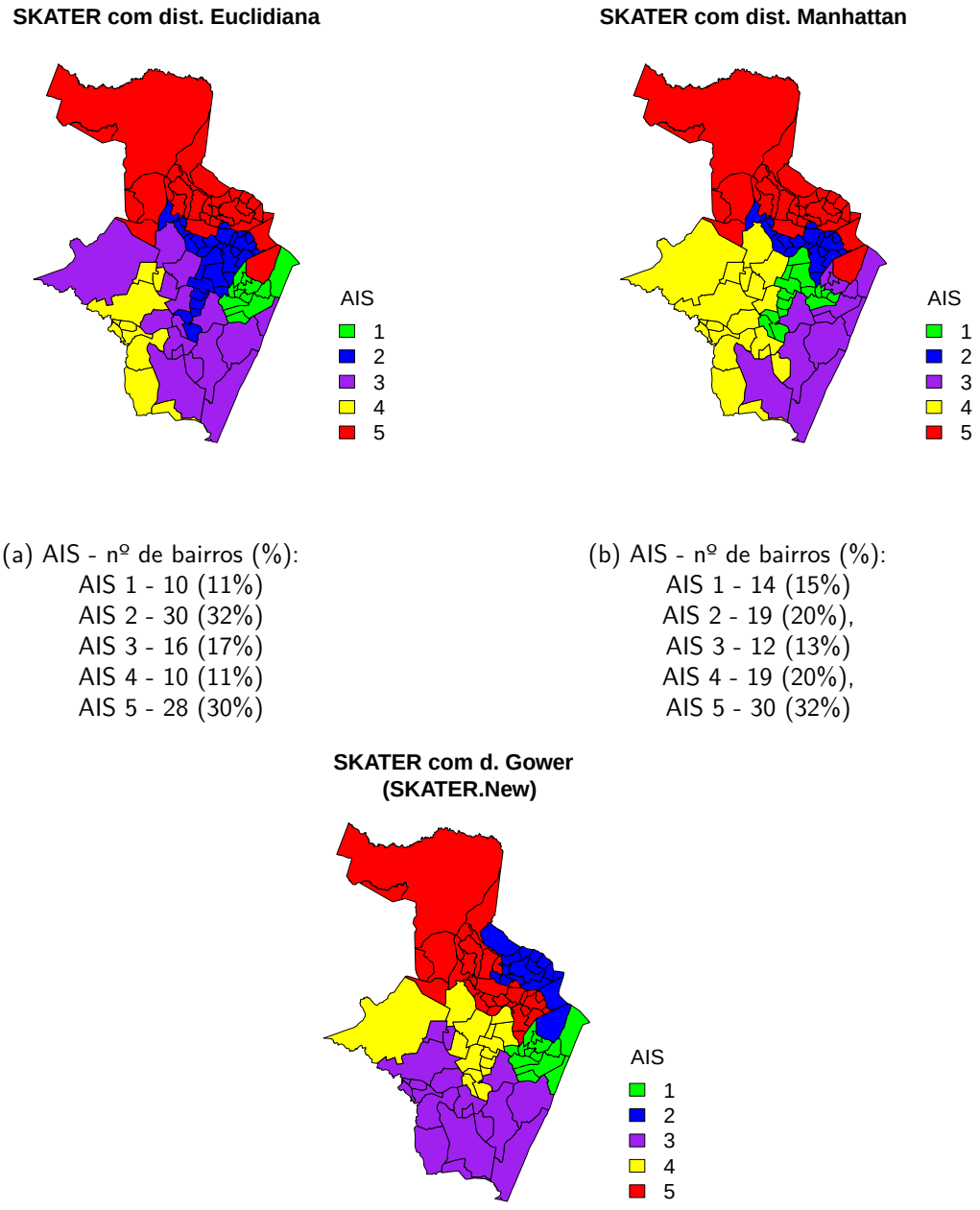
(c) AIS - nº de bairros (%):
AIS 1 - 24 (26%)
AIS 2 - 19 (22%)
AIS 3 - 12 (13%)
AIS 4 - 21 (22%)
AIS 5 - 18 (19%)

Fonte: Elaboração própria.

Observou-se também que no índice CH, o algoritmo Ward-Like teve melhor performance

também na distância Gower (CH $= 7,57$), i.e. na modificação Ward-Like.New, o que indica que a utilização da distância Gower mostrou-se mais efetiva na separação dos grupos.

Figura 14 – Resultados do algoritmo SKATER



(a) AIS - nº de bairros (%):
AIS 1 - 10 (11%)
AIS 2 - 30 (32%)
AIS 3 - 16 (17%)
AIS 4 - 10 (11%)
AIS 5 - 28 (30%)

(b) AIS - nº de bairros (%):
AIS 1 - 14 (15%)
AIS 2 - 19 (20%),
AIS 3 - 12 (13%)
AIS 4 - 19 (20%),
AIS 5 - 30 (32%)

(c) AIS - nº de bairros (%):
AIS 1 - 12 (13%)
AIS 2 - 16 (17%)
AIS 3 - 19 (20%)
AIS 4 - 14 (15%)
AIS 5 - 33 (35%)

Fonte: Elaboração própria.

Por fim, o último método utilizado foi o SKATER. Assim como no método Ward-Like, o SKATER também foi capaz de produzir classificações de boa qualidade em um tempo bom, apesar de um pouco maior com a distância Gower, pouco mais de 1 segundo.

Podemos observar na Figura 14a, algoritmo SKATER com distância Euclidiana, que todos os grupos são contíguos, embora a forma do grupo 3 (em roxo) pareça não assegurar contiguidade. A melhor forma de contiguidade para todas as AIS's, só foi conseguida nas distâncias Manhattan e Gower. A melhoria do agrupamento, com a proposta de utilização da distância Gower, i.e. com a extensão SKATER.New, pode ser vista, além na contiguidade, mas também na separação dos bairros por grupo (Figura 14c). O SKATER.New mostrou um desempenho satisfatório e demandou um pouco mais de tempo, 1,5 segundos. Na comparação dos índices de validação, o índice CH foi igual a $9,63$, $10,29$ e $9,42$ e o índice DB igual a $0,04$, $0,02$ e $0,04$, respectivamente com as distâncias Euclidiana, Manhattan e Gower, comprovando que foi o método contíguo com melhores resultados nesses dois índices.

Ainda analisando o algoritmo SKATER e sua extensão proposta SKATER.New, é possível verificar a qualidade do particionamento dos grupos pelo resultado da Equação 2.19, onde um valor menor indica a melhor a partição.

Tabela 6 – Comparação da qualidade da partição do algoritmo SKATER para diferentes distâncias

	Distância		
	Euclidiana	Manhattan	Gower
$Q(\pi)$	$262,03$	$763,37$	$\mathbf{13,41}$

Fonte: Elaboração própria.

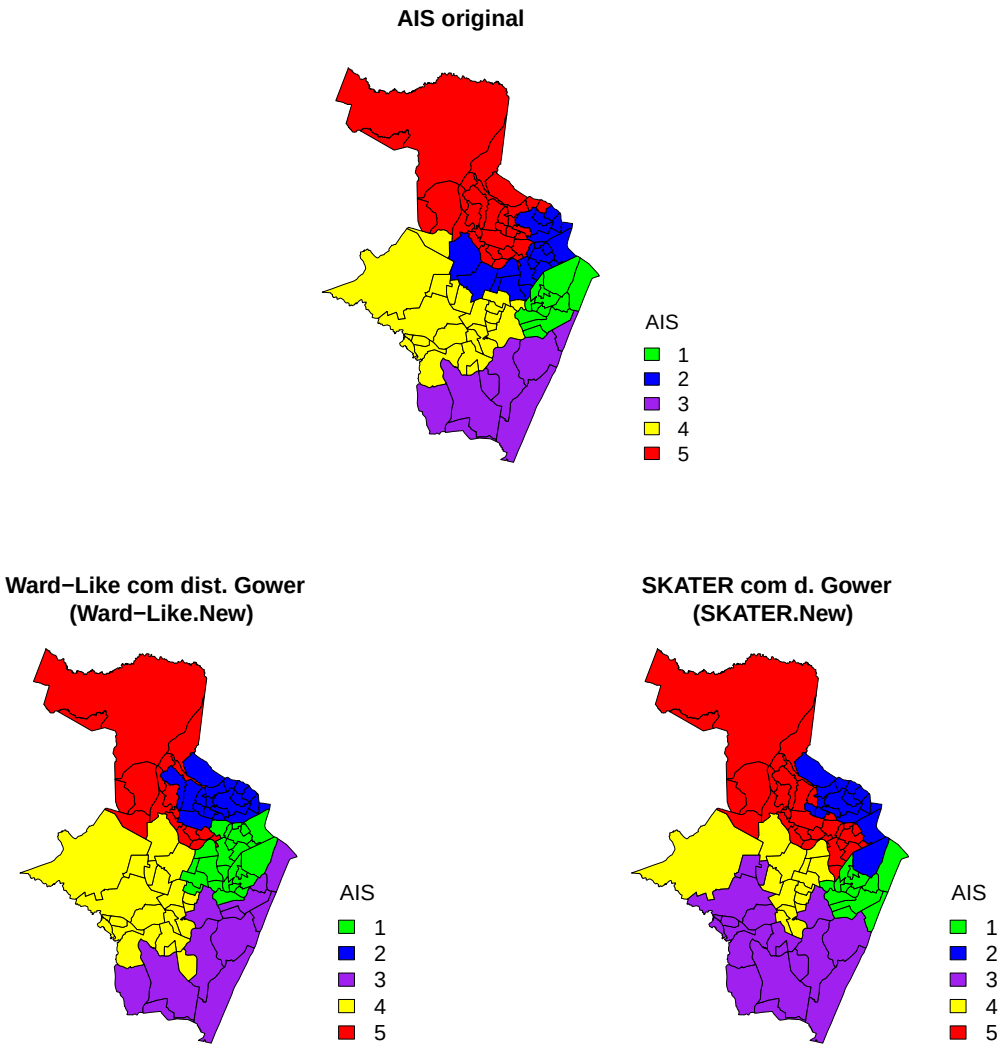
Com a Tabela 6 é possível visualizar que o índice de qualidade do agrupamento com a distância Gower ficou em $13,41$. Este valor representa um índice de qualidade muito bom, apontando mais uma vez que a extensão SKATER.New produziu um agrupamento com alta uniformidade dos grupos e de alta qualidade.

A priori os métodos K-Means, PAM, VNSKMED e Ward-Like podem ser executados sem necessidade de definições acerca do número de objetos em cada grupo. No entanto, uma especificação adicional relativa ao método SKATER precisa ser feita: pelas suas características, o algoritmo, tende, muitas vezes, a compor um grande grupo, englobando muitos bairros, em detrimento a grupos pequenos, eventualmente unitários, como ocorrido no K-Means e PAM. Neste sentido, é possível aplicar uma restrição de fixar um mínimo de bairros por grupo. Foi

fixado, então um mínimo de 10 bairros por grupo, o que corresponde a um mínimo de 10% de bairros em cada grupo.

Por fim, após a análise dos agrupamentos formados pelos algoritmos estudados, verificados os índices, tempo de processamento, tamanho e contiguidade dos grupos, temos dois algoritmos competitivos ao melhor método, são as propostas Ward-Like.New e SKATER.New. Com isso, iremos aprofundar a análise dos resultados e verificar de que forma os resultados apresentados por esses dois métodos se assemelham com a configuração atual das Àreas Integradas de Segurança.

Figura 15 – Comparação entre a formação das AIS's originais e as propostas



Fonte: Elaboração própria.

A Figura 15 contém o mapa de Recife divido pelas atuais AIS, servindo de "gabarito" para a comparação dos mapas de Recife divididos pelas novas AIS's formadas pelos algoritmos propostos. Geograficamente, algumas Áreas Integradas de Segurança se coincidem.

Os bairros que compõem cada nova área são apresentados nas Tabelas 7 e 8, os que estão marcados em negrito são os bairros que correspondem a mesma área da formação original, ou seja, correspondem a interseção entre a AIS original e a nova AIS do método proposto. Para não confundir, chamamos de "AIS" os grupos originais e demos um outro nome para os grupos dos algoritmos propostos, "WL" para os grupos do Ward-Like.New e "SK" para os grupos do SKATER.New.

No Ward-like.New, a WL 1, i.e. a Área Integrada de Segurança 1 do algoritmo Ward-Like.New, dada pela cor verde, é composta por 24 bairros, são eles: Soledade, Zumbi, Derby, Rosarinho, Torreão, Tamarineira, Ilha do Leite, Santo Amaro, Boa Vista, Prado, Ilha do Retiro, Paissandu, Ilha Joana Bezerra, Ponto de Parada, Encruzilhada, Espinheiro, Aflitos, Bongi, Campo Grande, Torre, Madalena, Coelhos, Hipódromo e Graças. Cinco desses bairros estão na AIS 1 original, apontados em negrito, são eles: Soledade, Ilha do Leite, Santo Amaro, Paissandu e Ilha Joana Bezerra.

Tabela 7 – Bairros que compõem cada Área Integrada de Segurança do algoritmo Ward-Like.New

AIS	**Bairros**
WL 1	**Soledade**, Zumbi, Derby, Rosarinho, Torreão, Tamarineira, **Ilha do Leite**, **Santo Amaro**, Boa Vista, Prado, Ilha do Retiro, **Paissandu**, **Ilha Joana Bezerra**, Ponto de Parada, Encruzilhada, Espinheiro, Aflitos, Bongi, Campo Grande, Torre, Madalena, Coelhos, Hipódromo, Graças.
WL 2	**Fundão**, **Campina do Barreto**, Porto da Madeira, **Cajueiro**, Beberibe, Bomba do Hemetério, Mangabeira, Dois Unidos, Linha do Tiro, **Peixinhos**, **Arruda**, Alto José Bonifácio, Morro da Conceição, Casa Amarela, Vasco da Gama, Alto José do Pinho, Alto Santa Terezinha, Nova Descoberta, **Água Fria**.
WL 3	**Cohab**, **Jordão**, Cabanga, **Pina**, **Ibura**, Santo Antônio, Do Recife, **Brasília Teimosa**, **Boa Viagem**, São José, **Imbiribeira**, Afogados.
WL 4	**Cidade Universitária**, **Engenho do Meio**, **Caçote**, **Várzea**, **Torrões**, Iputinga, **Curado**, **San Martin**, IPSEP, **Jardim São Paulo**, **Areias**, **Sancho**, **Barro**, **Estância**, **Tejipió**, Cordeiro, **Coqueiral**, **Totó**, **Mangueira**, **Mustardinha**, **Jiquiá**.
WL 5	**Passarinho**, **Dois Irmãos**, **Jaqueira**, **Santana**, Caxangá, **Sítio dos Pintos**, **Pau Ferro**, **Poço da Panela**, **Casa Forte**, **Parnamirim**, **Brejo da Guabiraba**, **Alto do Mandu**, **Macaxeira**, **Brejo de Beberibe**, **Monteiro**, **Apipucos**, **Guabiraba**, **Córrego do Jenipapo**.

Fonte: Elaboração própria.

Na WL 2, na cor azul, os bairros Fundão, Campina do Barreto, Cajueiro, Peixinhos, Arruda e Água Fria estão na mesma AIS 2 da formação original. Agora na nova formação, WL 2,

foram incluídos os bairros Porto da Madeira, Beberibe, Bomba do Hemetério, Mangabeira, Dois Unidos, Linha do Tiro, Alto José Bonifácio, Morro da Conceição, Casa Amarela, Vasco da Gama, Alto José do Pinho, Alto Santa Terezinha e Nova Descoberta.

A WL 3, na posição sul do mapa (em roxo) é composta agora por doze bairros, Cohab, Jordão, Cabanga, Pina, Ibura, Santo Antônio, Do Recife, Brasília Teimosa, Boa Viagem, São José, Imbiribeira, Afogados, onde sete bairros coincidiram exatamente nesse mesmo grupo da formação original, AIS 3, são eles: Cohab, Jordão, Pina, Ibura, Brasília Teimosa, Boa Viagem e Imbiribeira.

A WL 4, em amarelo, foi a área do Ward-Like.New que mais se aproximou geograficamente da área AIS 4 original. Saíram os bairros Afogados, Ilha do Retiro, Prado, Bongi, Torrões e Caxangá, acrescentaram os bairros Iputinga, IPSEP e Cordeiro e permaneceram Cidade Universitária, Engenho do Meio, Caçote, Várzea, Torrões, Curado, San Martin, Jardim São Paulo, Areias, Sancho, Barro, Estância, Tejipió, Coqueiral, Totó, Mangueira, Mustardinha e Jiquiá, tendo assim, dezoito bairros exatamente na mesma posição oeste do mapa da formação original.

Por último, na WL 5 (em vermelho), dezessete bairros estão na mesma AIS 5 original, Passarinho, Dois Irmãos, Jaqueira, Santana, Sítio dos Pintos, Pau Ferro, Poço da Panela, Casa Forte, Parnamirim, Alto do Mandu, Macaxeira, Brejo de Beberibe, Monteiro, Apipucos, Guabiraba e Córrego do Jenipapo. E foi incluído o bairro Caxangá, que na formação original faz parte da AIS 4.

No algoritmo SKATER.New (Tabela 8), os bairros Soledade, Ilha do Leite, Boa Vista, Cabanga, Paissandu, Ilha Joana Bezerra, Santo Antônio, Do Recife, São José e Coelhos estão na interseção das áreas SK 1 e AIS 1. Ilha do Retiro e Brasília Teimosa não estão na formação original, mas agora fazem parte desse grupo 1. Já Boa Vista, Santo Amaro e Ilha do Leite pertencem à formação original AIS 1 e não estão na SK 1.

A AIS 2 original é formada pelos bairros Derby, Graças, Espinheiro, Aflitos, Rosarinho, Encruzilhada, Torreão, Ponto de Parada, Hipódromo, Campo Grande, Madalena, Torre, Zumbi, Cordeiro, Iputinga, Água Fria Água Fria, Arruda, Campina do Barreto, Cajueiro, Fundão e Peixinhos. Agora, o novo grupo SK 2 é formado por Fundão, Campina do Barreto, Porto da Madeira, Cajueiro, Beberibe, Bomba do Hemetério, Dois Unidos, Linha do Tiro, Santo Amaro, Peixinhos, Arruda, Morro da Conceição, Alto José do Pinho, Alto Santa Terezinha, Campo Grande e Água Fria. Sete bairros coincidem-se, são eles: Fundão, Campina do Barreto, Cajueiro, Peixinhos, Arruda, Campo Grande e Água Fria.

Tabela 8 – Bairros que compõem cada Área Integrada de Segurança do algoritmo SKATER.New

AIS	Bairros
SK 1	**Soledade**, **Ilha do Leite**, **Boa Vista**, **Cabanga**, Ilha do Retiro, **Paissandu**, **Ilha Joana Bezerra**, **Santo Antônio**, **Do Recife**, Brasília Teimosa, **São José**, **Coelhos**.
SK 2	**Fundão**, **Campina do Barreto**, Porto da Madeira, **Cajueiro**, Beberibe, Bomba do Hemetério, Dois Unidos, Linha do Tiro, Santo Amaro, **Peixinhos**, **Arruda**, Morro da Conceição, Alto José do Pinho, Alto Santa Terezinha, **Campo Grande**, **Água Fria**.
SK 3	Cidade Universitária, Engenho do Meio, Caçote, **Cohab**, Curado, **Ipsep**, Jardim São Paulo, Areias, Sancho, Barro, Tejipió, Coqueiral, Totó, **Jordão**, **Pina**, **Ibura**, **Boa Viagem**, **Imbiribeira**, Afogados.
SK 4	Várzea, **Torrões**, Iputinga, **San Martin**, **Estância**, Zumbi, Cordeiro, **Mangueira**, **Prado**, **Bongi**, **Mustardinha**, Torre, Madalena, **Jiquiá**.
SK 5	**Passarinho**, **Dois Irmãos**, **Jaqueira**, **Santana**, Caxangá, **Sítio dos Pintos**, **Pau Ferro**, Derby, Rosarinho, **Mangabeira**, Torreão, **Tamarineira**, **Poço da Panela**, **Casa Forte**, **Parnamirim**, Brejo da Guabiraba, **Alto José Bonifácio**,**Alto do Mandu**, **Casa Amarela**, Ponto de Parada, Encruzilhada, Espinheiro, Aflitos, **Vasco da Gama**, **Macaxeira**, **Brejo de Beberibe**, **Nova Descoberta**, **Monteiro**, Hipódromo, Graças, **Apipucos**, **Guabiraba**, **Córrego do Jenipapo**.

Fonte: Elaboração própria

Na formação AIS 3, os bairros que compõem esse grupo são Brasília Teimosa, Pina, Boa Viagem, Jordão, Ibura, Imbiribeira, Ipsep e Cohab. Agora na formação SK 3, retirou-se o bairro Brasília Teimosa e acrescentou-se Cidade Universitária, Engenho do Meio, Caçote, Curado, Jardim São Paulo, Areias, Sancho, Barro, Tejipió, Coqueiral, Totó e Afogados.

A SK 4 é composta pelos bairros Várzea, Torrões, Iputinga, San Martin, Estância, Zumbi, Cordeiro, Mangueira, Prado, Bongi, Mustardinha, Torre, Madalena e Jiquiá, sendo Torrões, San Martin, Estância, Mangueira, Prado, Bongi, Mustardinha, e Jiquiá bairros que coincidem com a formação AIS 4.

Os bairros que compõem a AIS 5 original são bairros da zona norte de Recife e no mapa estão dispostos de vermelho. Exatamente vinte e dois bairros dessa área também estão na SK 5, são eles: Passarinho, Dois Irmãos, Jaqueira, Santana, Sítio dos Pintos, Pau Ferro, Mangabeira, Tamarineira, Poço da Panela, Casa Forte, Parnamirim, Alto José Bonifácio, Alto do Mandu, Casa Amarela, Vasco da Gama, Macaxeira, Brejo de Beberibe, Nova Descoberta, Monteiro, Apipucos, Guabiraba e Córrego do Jenipapo. Esse é o grupo do SKATER.New que mais se aproximou geograficamente da área AIS 5 original.

Para entender melhor essas interseções, a Tabela 9 foi montada. Ela mostra a quantidade de bairros que cada grupo possui e a quantidade de bairros que cada grupo tem em comum exatamente ao mesmo grupo da formação original. No total, o Ward-Like.New apresentou 53

bairros que coincidem com a formação original e o SKATER.New 54 bairros.

Tabela 9 – Comparação do número de bairros e interseções visualizadas

	nº de bairros					comum a mesma AIS					
Formação	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	Total
Original	11	21	08	23	31	-	-	-	-	-	
Ward-Like.New	24	19	12	21	18	5	6	7	18	17	53
SKATER.New	12	16	19	14	33	10	7	7	8	22	54

Fonte: Elaboração própria

Outra verificação importante é como esses novos grupos se diferenciam em relação aos tipos de crime ocorridos neles e qual crime está mais predominante em determinada grupo. A Tabela 10 apresenta para cada novo grupo a quantidade de cada tipo de crime ocorrido.

Verifica-se que lesão corporal (LCO) é o crime com maior ocorrência em todas as áreas, devido ao fato de ser o tipo de crime com maior incidência na base de dados ($68,1\%$) como visto na Tabela 3), porém, algumas áreas concentraram mais crimes desse tipo que outras, por exemplo, há uma discrepância de LCO da área WL 3 ($n = 12.541$) com a área SK 3 ($n = 2.375$) e da WL 4 ($n = 3.377$) com a SK 4 ($n = 19.550$).

Tabela 10 – Número de crimes que compõem cada nova área formada pelo algoritmos Ward-Like.New e SKATER.New

Tipo de Crime	WL 1	SK 1	WL 2	SK 2	WL 3	SK 3	WL 4	SK 4	WL 5	SK 5
EST	900	484	480	559	887	106	283	1.563	732	570
LET	1.890	854	808	1.060	1.686	129	484	3.192	1.212	845
LCO	10.738	8.419	9.227	8.991	**12.541**	**2.375**	**3.377**	**19.550**	9.550	6.098
OUT	85	84	90	68	134	38	34	173	80	60
LCT	690	535	809	789	1.179	333	271	1.621	718	389
LCD	1.992	1.198	1.380	1.695	1.882	281	587	3.355	1.956	1.268

Fonte: Elaboração própria.

Posto isso, verifica-se que a área WL 1 concentrou os crimes de lesão corporal, lesão corporal doméstica e crime letal. A WL 2 concentrou crimes de lesão corporal, lesão corporal doméstica e lesão corporal de trânsito. A WL 3 concentrou, dentre todas as áreas do Ward-Like.New, o maior número de ocorrências de crime de lesão corporal ($n = 12.541$). Na WL 4, além da lesão corporal, predominaram os crimes LCD, LET e EST. E a WL 5, segue o perfil da área 4.

Os crimes da área 1 do SKATER.New, a SK 1, tem como principais características crimes de natureza de lesão corporal, de lesão corporal doméstica e letal. Na SK 2, também predominam os crimes de natureza de LCO, LCD e LET. A SK 3 concentrou crimes de lesão corporal de trânsito e lesão corporal doméstica. Já a SK 4 é a área mais violenta do SKATER.New, todos os tipos de crimes têm a maior ocorrência nessa área, sobretudo, os crimes de LCO ($n = 19.550$). E na SK 5, predominou os crimes de lesão corporal, lesão corporal doméstica, letal e estupro.

Se fossemos utilizar essa mesma base de ocorrência de crimes, para analisar as Áreas Integradas de Segurança originais, a formação que mais assemelhou-se foi a formação do SKATER.New seja pela semelhança dos bairros, seja pela ocorrência dos crimes nas áreas.

6 CONSIDERAÇÕES FINAIS E SUGESTÕES FUTURAS

Neste trabalho foi realizado um estudo sobre cinco métodos de agrupamento já existentes e propostas duas extensões para uso com dados de crimes. Os experimentos realizados tiveram como propósito analisar os algoritmos no conjunto de dados criminais, de modo a verificar os resultados dos agrupamentos entre os bairros da cidade de Recife e suas Áreas Integradas de Segurança.

Foi possível notar que os diferentes métodos criaram agrupamentos muito variados. O método K-Means demonstrou ser muito rápido e apresentou bons resultados nos índices de validação. Porém, apresentou dois problemas na construção de grupos com a base de dados usada: discrepâncias no tamanho dos grupos e falta de contiguidade. Do mesmo modo, os métodos PAM e VNSKMED também tiveram bons rendimentos nos índices de validação, mas os objetos também não possuíram contiguidade, apesar de demonstrarem uma menor disparidade no número de bairros em cada grupo, além disso, o VNSKMED foi o mais dispendioso em tempo de processamento.

Já nos métodos que consideram restrições espaciais, como o Ward-Like e o SKATER, os agrupamentos mostraram-se mais completos na detecção dos grupos. De acordo com os pesos dados às diferenças geográficas, as soluções tiveram grupos mais ou menos contíguos espacialmente. Ademais, o ganho de homogeneidade nos grupos formados pelas propostas Ward-Like.New e SKATER.New foi confirmado pelos índices de validação e pelo indicador de qualidade de partição $Q(\pi)$ no SKATER.New.

Com base na análise dos resultados, concluímos que os métodos espaciais modificados apresentam melhores indicativos de reproduzir com mais eficácia as informações contidas nos dados criminais, baseados nos resultados dos índices de validação e, também, por produzirem soluções mais regionalizadas e contíguas apontadas nos mapas, podendo servir de suporte na definição de novas Áreas Integradas de Segurança. E obter resultados mais regionalizados pode facilitar a aplicação de iniciativas públicas no tocante às medidas de segurança.

Ademais, acredita-se que o SKATER.New seja o mais recomendado. Para sua aplicação, o usuário não precisa definir centroides ou parâmetros adicionais para definir os grupos, pois ele se organiza formando seus próprios grupos sem necessidade dessas informações iniciais. Como também, pode-se aplicar restrições sobre a quantidade de objetos por grupo, a fim de tornar os grupos mais homogêneos. Foi visto também, que o uso desse algoritmo obteve bons

resultados acerca da definição de novas Áreas Integradas de Segurança, chegando a ter mais bairros em comum com a estrutura de AIS original.

Novos estudos, análises e propostas para melhorar o desempenho e estabilidade do SKATER.New podem ser realizados, com o intuito de tonar esse algoritmo uma boa ferramenta para o agrupamento de dados criminais. Pelas aplicações, verifica-se um desempenho menos rápido em sua execução. Em trabalhos futuros, poderá haver mais ênfase nesse aspecto, como também automatizar o algoritmo, para que ele detecte automaticamente o melhor número de grupos, sem a necessidade de impor essa informação na entrada do algoritmo.

Além disso, a proposta de um estudo a nível desagregado de variáveis e, também estender o campo de estudo, para outros municípios do estado de Pernambuco e também outros estados, a análise de outros índices e a comparação com outras técnicas de agrupamento não consideradas nesse trabalho, além do uso do algoritmo em outras aplicações, podem ser explorados em trabalhos futuros.

REFERÊNCIAS

AGUIAR, D. C.; SÁNCHEZ, R. G.; CAMÊLO, E. L. S. Hierarchical Clustering with Spatial Constraints and Standardized Incidence Ratio in Tuberculosis Data. *Mathematics*, Mdpi, v. 8, n. 9, p. 1478, 2020. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/8/9/1478>.

AL-SULTAN, K. S. A tabu search approach to the clustering problem. *Pattern recognition*, Elsevier, v. 28, n. 9, p. 1443–1451, 1995. ISSN 0031-3203. Disponível em: <https://doi.org/10.1016/0031-3203(95)00022-R>.

ALAM, S.; DOBBIE, G.; REHMAN, S. U. Analysis of particle swarm optimization based hierarchical data clustering approaches. *Swarm and Evolutionary Computation*, Elsevier, v. 25, p. 36–51, 2015. ISSN 2210-6502. SI: RAMONA. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2210650215000784>.

ASSUNÇÃO, R. M.; NEVES, M. C.; CÂMARA, G.; FREITAS, C. da C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, Taylor & Francis, v. 20, n. 7, p. 797–811, 2006. Disponível em: <https://doi.org/10.1080/13658810600665111>.

BIVAND, R. S.; PEBESMA, E.; GOMEZ-RUBIO, V. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. Disponível em: <https://asdar-book.org/>.

BRITO, J. A. de M.; SEMAAN, G. S.; FADEL, A. C. An Effective VNS Algorithm for k-Medoids Clustering Problem. *IEEE Latin America Transactions*, v. 20, n. 5, p. 710–717, 2022. Disponível em: <https://latamt.ieeer9.org/index.php/transactions/article/view/5480>.

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. clValid: An R package for cluster validation. *Journal of Statistical Software*, v. 25, n. 4, p. 1–22, 2008. Disponível em: <https://www.jstatsoft.org/v25/i04/>.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/03610927408827101>.

CEARÁ (ESTADO). Portaria gs/sspds n° 436/2017, de 17 de abril de 2017. *Diário Oficial [do] Estado do Ceará*, Fortaleza, CE, 2017.

CHAVENT, M.; KUENTZ-SIMONET, V.; LABENNE, A.; SARACCO, J. ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, Springer, v. 33, n. 4, p. 1799–1822, 2018. ISSN 1613-9658. Disponível em: <https://doi.org/10.1007/s00180-018-0791-1>.

CHAVENT, M.; KUENTZ, V.; LABENNE, A.; SARACCO, J. *ClustGeo: Hierarchical Clustering with Spatial Constraints*. [S.l.], 2021. R package version 2.1. Disponível em: <https://CRAN.R-project.org/package=ClustGeo>.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

D'ORAZIO, M. *StatMatch: Statistical Matching or Data Fusion*. [S.l.], 2020. R package version 1.4.0. Disponível em: <https://CRAN.R-project.org/package=StatMatch>.

DUNN, J. C. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, Taylor & Francis, v. 4, n. 1, p. 95–104, 1974. Disponível em: <https://doi.org/10.1080/01969727408546059>.

DUQUE, J. C.; RAMOS, R.; SURIÑACH, J. Supervised regionalization methods: A survey. *International Regional Science Review*, Sage Publications Sage CA: Los Angeles, CA, v. 30, n. 3, p. 195–220, 2007. Disponível em: <https://doi.org/10.1177/0160017607301605>.

FAVEIRO, L.; BELFIORE, P.; SILVA, F.; CHAM, B. *Análise de dados: modelagem multivariada para tomada de decisão*. São Paulo: Campus, 2009.

GOWER, J. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, v. 27, n. 4, p. 857–871, 1971. Disponível em: <https://doi.org/10.2307/2528823>.

GUERREIRO, M. T. et al. *Análise de métodos de agrupamento de dados para detecção de anomalias na precificação e categorização de peças da indústria automotiva*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2021.

HENNIG, C. *fpc: Flexible Procedures for Clustering*. [S.l.], 2020. R package version 2.2-9. Disponível em: <https://CRAN.R-project.org/package=fpc>.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Englewood Cliffs: Prentice-Hall, 1988. ISBN 013022278X.

JOSÉ-GARCÍA, A.; GÓMEZ-FLORES, W. Automatic clustering using nature-inspired metaheuristics: A survey. *Applied Soft Computing*, Elsevier, v. 41, p. 192–213, 2016. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494615007772>.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 1990. ISBN 0-471-87876-6.

KIANI, R.; MAHDAVI, S.; KESHAVARZI, A. Analysis and Prediction of Crimes by Clustering and Classification. *International Journal of Advanced Research in Artificial Intelligence*, Citeseer, v. 4, n. 8, p. 11–17, 2015. Disponível em: <http://dx.doi.org/10.14569/IJARAI.2015.040802>.

LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: a package for multivariate analysis. *Journal of Statistical Software*, v. 25, n. 1, p. 1–18, 2008. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v025i01>.

LEAL, P. B. *Estudo e Aplicações do Método de Agrupamento Baseado no Modelo*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, PE, 2004.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967. v. 1, n. 14, p. 281–297. Disponível em: <https://cir.nii.ac.jp/crid/1572261550390329472>.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. *cluster: Cluster Analysis Basics and Extensions*. [S.l.], 2021. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source). Disponível em: <https://CRAN.R-project.org/package=cluster>.

MENDES, J. C. *Agrupamento de Dados e Suas Aplicações*. Dissertação (Mestrado) — Universidade Federal do Maranhão, São Luís, 2017.

MLADENOVIĆ, N.; HANSEN, P. Variable neighborhood search. *Computers & operations research*, Elsevier, v. 24, n. 11, p. 1097–1100, 1997. ISSN 0305-0548. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0305054897000312>.

MONDAL, B.; CHOUDHURY, J. P. A Comparative Study on K-means and Pam Algorithm Using Physical Characters of Different Varieties of Mango in Índia. *International Journal of Computer Applications*, Citeseer, v. 78, n. 5, p. 21–24, 2013. Disponível em: <https://www.ijcaonline.org/archives/volume78/number5/13485-1189>.

NANDA, S. J.; PANDA, G. A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation*, Elsevier, v. 16, p. 1–18, 2014. ISSN 2210-6502. Disponível em: <https://www.sciencedirect.com/science/article/pii/S221065021300076X>.

NASCIMENTO, M. C.; TOLEDO, F. M.; CARVALHO, A. C. de. Investigation of a New GRASP-based Clustering Algorithm Applied to Biological Data. *Computers & Operations Research*, Elsevier, v. 37, n. 8, p. 1381–1388, 2010. ISSN 0305-0548. Operations Research and Data Mining in Biological Systems. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0305054809000574>.

ORTEGA, J.; ALMANZA-ORTEGA, N.; VEGA-VILLALOBOS, A.; PAZOS-RANGEL, R.; ZAVALA-DIAZ, J. C.; MARTíNEZ-REBOLLAR, A. The K-Means Algorithm Evolution. In: _____. [S.l.: s.n.], 2019. ISBN 978-1-83880-333-9.

PAIVA, R. de O. *Análise Comparativa de Métodos de Agrupamentos*. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Juiz de Fora, Juiz de Fora, 2013.

PERNAMBUCO (ESTADO). Portaria gab/sds n° 1197, de 11 de junho de 2010. *Diário Oficial [do] Estado de Pernambuco*, Recife, PE, 2010.

PERNAMBUCO (ESTADO). Decreto nº 38.787, de 30 de outubro de 2012. *Diário Oficial [do] Estado de Pernambuco*, Pernambuco, PE, 2012. Disponível em: <http://www.portais.pe.gov.br/c/document_library/get_file?uuid=53f783a0-36e5-4811-a39f-f775fb522f33&groupId=17459>. Acesso em: 20 de dezembro de 2021.

PERNAMBUCO (ESTADO). Lei nº 14.804, de 29 de outubro de 2012. *Diário Oficial [do] Estado de Pernambuco*, Pernambuco, PE, 2012. Disponível em: <http://www.portais.pe.gov.br/c/document_library/get_file?uuid=d6b104f3-7aec-49f3-8120-6e899742a428&groupId=17459>. Acesso em: 20 de dezembro de 2021.

PERNAMBUCO (ESTADO). Termo de classificação de informação nº 01/2015. *Diário Oficial [do] Estado de Pernambuco*, Pernambuco, PE, 2012. Disponível em: <https://www.lai.pe.gov.br/sds/wp-content/uploads/sites/118/2019/04/TCI-001_2015.pdf>. Acesso em: 20 de dezembro de 2021.

PERNAMBUCO, S. de Planejamento e Gestão de. *Pacto pela Vida. (Coleção Cadernos de Boas Práticas de Gestão)*. Brasília, DF: Instituto Publix, 2014. v. 5.

PERNAMBUCO, S. de Planejamento e Gestão de. *Cartilha Pacto pela Vida: democratização e controle social da política de segurança dos municípios*. Recife,PE: SEGPR, 2021.

POURBAHRAMI, S.; KHANLI, L. M. A survey of neighbourhood construction models for categorizing data points. *ArXiv*, abs/1810.03083, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>.

RATTON, J. L.; GALVÃO, C.; FERNANDEZ, M. O Pacto pela Vida e a Redução de Homicídios em Pernambuco. *Instituto Igarapé*, Rio de Janeiro, RJ, 2014.

RIO DE JANEIRO (ESTADO). Resolução ssp n° 607, de 24 de março de 2003. *Diário Oficial [do] Estado do Rio de Janeiro*, Rio de Janeiro, RJ, 2003.

SCHEUERER, S.; WENDOLSKY, R. A scatter search heuristic for the capacitated clustering problem. *European Journal of Operational Research*, Elsevier, v. 169, n. 2, p. 533–547, 2006. ISSN 0377-2217. Feature Cluster on Scatter Search Methods for Optimization. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221704005533>.

SERPA, D. R. *Abordagens Heurísticas para Problemas de Agrupamentos*. Dissertação (Mestrado em Computação Aplicada) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2011.

SOKAL, R. R.; SNEATH, P. H. A. *Principle of Numerical Taxonomy*. San Francisco: Freeman, 1963. ISSN 1063-5157.

TEIXEIRA, L. V.; ASSUNÇÃO, R. M.; LOSCHI, R. H. Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research*, v. 20, p. 1–35, 2019. Disponível em: <http://jmlr.org/papers/v20/16-615.html>.

WALESIAK, M.; DUDEK, A. The choice of variable normalization method in cluster analysis. In: SOLIMAN, K. S. (Ed.). [S.l.]: International Business Information Management Association (IBIMA), 2020. p. 325–340. ISBN 978-0-9998551-4-1.