Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa

Sérgio Luiz da Silva Campos

JUIZ DE FORA

JULHO, 2019

Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa

Sérgio Luiz da Silva Campos

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Wagner Antonio Arbex

JUIZ DE FORA

JULHO, 2019

Campos, Sérgio
Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa / Sérgio Campos - 2019
149p. : il. color.

Orientador: Wagner Antonio Arbex
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2019.

1. Mineração de Dados 2. Análise de agrupamento. 3. Métodos não supervisionados 4. K-Means 5. Fuzzy C-Means. I. Arbex, Wagner, orient. II. Prof. Dr.

Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa

Sérgio Luiz da Silva Campos

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil (COPPE/UFRJ)

Heder Soares Bernardino
Doutor em Modelagem Computacional (LNCC/MCTI)

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação (UFRJ)

JUIZ DE FORA
08 DE JULHO, 2019

Dedico este trabalho a meus pais, por todo amor, carinho, apoio, sustento, encorajamento e educação que sempre me deram e, sobretudo, por serem tudo de mais valioso que tenho na vida.

Resumo

A análise de agrupamento é uma etapa no processo de descoberta de conhecimento em bases de dados (KDD), fornecendo mecanismos adequados à compreensão dos dados, além de possibilitar a formulação de hipóteses referentes à natureza dos mesmos. Quando utilizada como instrumento de pesquisa no âmbito da Zootecnia, como na bovinocultura de leite, amplia seu campo investigativo e contribui para a revelação de padrões que favorecem a predição e/ou tomada de decisões para a sustentabilidade dos sistemas produtivos. Este estudo tem por objetivo a identificação de conjuntos de animais com características fenotípicas distintas em relação à produção de leite. O fenótipo é o valor de uma dada característica, i.e., o que pode ser observado ou mensurado. Ele é dependente do valor genético do indivíduo, do ambiente no qual é produzido e da ação conjunta desses dois fatores, denominada interação genótipo-ambiente. Isso equivale a dizer que o fenótipo é influenciado individual e aditivamente pelo genótipo e pelo ambiente a que está sujeito. Logo, na análise de agrupamento, deve-se levar em conta esses três fatores na observância das possíveis variações fenotípicas dentro de um mesmo ambiente, provavelmente por questões genotípicas e/ou por diferentes respostas da interação genótipo-ambiente; como também das possíveis variações pela mudança de ambiente. O fator ambiental considerado foi a dieta utilizada na nutrição dos animais. Os resultados mostraram associações entre diferentes tipos de regimes alimentares em relação à performance produtiva dos animais, o que consequentemente pode gerar implicações práticas, como reduções de custo com dietas bovinas. Além disso, também permitiu a identificação de conjuntos bem definidos de animais com diferentes desempenhos produtivos em todas as dietas consideradas, possivelmente contendo relevância para a incorporação em avaliações genéticas, visando a formação de novilhas geneticamente superiores para a produção de leite. Os métodos de agrupamento utilizados foram o K-Means e Fuzzy C-Means.

Palavras-chave: análise de agrupamento, bovinocultura, fenótipo, fuzzy c-means, genótipo, kdd, k-means, produção de leite, zootecnia

Abstract

Clustering analysis is one of techniques of knowledge discovery in databases (KDD), providing an effective way to identify patterns and relationships in complex data and to form hypotheses about their structure. When used as a data analysis tool in the branch of Zootechny, as in dairy farming, broadens its field of research and allows the pattern discovery that supports prediction and/or decision making for the sustainability of production systems. This study aims to identify cattle clusters with different phenotypic characteristics in relation to milk production. The phenotype or observed value of a certain trait depends on genetic value of the individual, environment in which it is produced and their interactions. In another words, it is influenced by genotype and environment in an additive and independent way. Therefore, one must to take into account these three factors in the analysis of potential phenotypic variations within the same environment, most likely due to genotypic issues and/or different responses of the genotype-environment interaction; as well as the potential variations due to the environment changing. The environmental factor considered was the diet used in animal nutrition. The results showed associations between different types of diets regarding to the productive performance of the cows, which consequently can generate practical implications, such as cost reductions in dairy cattle feeding and nutrition. In addition, they also allowed identifying well-defined clusters of cows with different productive performances, which may be useful for genetic evaluations, aiming at generating of genetically superior heifers for milk production. The clustering methods used were K-Means and Fuzzy C-Means.

Keywords: clustering analysis, dairy farming, dairy cattle, data analysis, fuzzy c-means, genotype, kdd, k-means, milk production, phenotype, zootechny

Agradecimentos

A Deus, por toda a força para enfrentar e superar os obstáculos em minha vida e por todas as minhas realizações.

Aos meus pais e irmão, por serem o alicerce da minha vida e o motivo da minha felicidade.

Ao professor Wagner, por sua orientação e pela oportunidade na elaboração deste trabalho.

A Cristiano Borges, analista da Embrapa Gado de Leite, por suas valiosas considerações acerca deste estudo.

Ao professor Heder, por toda sua atenção, auxílio e disponibilidade.

A todos os professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Ao Centro Nacional de Pesquisa de Gado de Leite (Embrapa Gado de Leite) da Empresa Brasileira de Pesquisa Agropecuária pelo fornecimento da base de dados utilizada neste trabalho.

"O essencial é invisível aos olhos".

Antoine de Saint-Exupéry (O Pequeno Príncipe)

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

FCM	Fuzzy C-Means
KM	K-Means
KDD	*Knowledge Discovery in Database*
MSE	*Mean Squared Error*
SSE	*Sum of Squared Error*
TCF	Teoria dos conjuntos *fuzzy*

1 Introdução

A análise de agrupamento, também conhecida como análise de cluster (*cluster analysis*), é uma técnica de mineração de dados que tem por objetivo promover a segmentação de dados em categorias ou grupos baseando-se em suas características de forma que objetos pertencentes a um mesmo grupo possuam alto grau de similaridade, mas grande poder de distinção em relação a elementos de outros grupos. Ao contrário da classificação supervisionada, a análise de agrupamento utiliza amostras de dados não rotuladas nas quais a noção acerca dos grupos é ínfima ou inexistente (Tan; Steinbach; Kumar, 2009, p. 585; Linden, 2009, p. 18–19; Cassiano, 2014, p. 59).

Segundo Backer (1995 apud Yonamine et al., 2002, p. 1), houve um crescente aumento da necessidade de categorização de objetos decorrente do grande volume de dados gerados com a informatização, que torna o processo complexo e não intuitivo. Em meio a esse problema, há distintas abordagens que podem ser utilizadas para tratá-lo, dentre as quais se destacam a abordagem clássica (*crisp*), que atribui de forma única e exclusiva cada objeto a um grupo; e a abordagem *fuzzy*, na qual os objetos pertencem a todos os grupos em algum grau de pertinência compreendido no intervalo [0, 1] (Hoppner et al., 1999, p. 17), em que 0 é a exclusão total do objeto em relação a determinado grupo e 1 a sua inclusão integral e, portanto, diferentemente da abordagem *crisp*, um objeto pode estar associado a mais de um cluster.

Entre a vasta gama de algoritmos de agrupamento presentes na literatura, foram selecionados o K-Means (KM) e o Fuzzy C-Means (FCM) neste projeto justamente por trabalharem sob diferentes abordagens de modo a permitir uma comparação entre os resultados obtidos com a aplicação dos mesmos visando facilitar a identificação, análise e compreensão de padrões no conjunto e, logo, promover uma melhor percepção dos dados. Ambos os métodos são dependentes da escolha inicial de um parâmetro referente ao número de grupos fornecido pelo usuário que, em geral, representa um problema, visto o desconhecimento acerca do número de categorias existentes inicialmente (Linden, 2009, p. 24). Além disso, o KM é um dos algoritmos mais utilizados no processo de clusterização

de dados e o FCM é sua adaptação que incorpora a abordagem da teoria dos conjuntos *fuzzy* para lidar com incertezas nos dados que são comumente encontradas na prática.

O grande ganho proveniente do uso dos métodos de clusterização é a definição mais clara das propriedades individuais dos grupos, permitindo a formulação de hipóteses a respeito do comportamento dos dados (Cassiano, 2014, p. 60) e, por isso, a análise de agrupamento é um recurso aplicável em diversas áreas do conhecimento quando se tem por objetivo a categorização de uma massa de dados em grupos que, por definição, apresentam características gerais que os distinguem.

1.1 Definição do problema

Para entender as dificuldades associadas a problemas de agrupamento em geral, suponha que uma operadora de telefonia móvel deseja traçar o perfil de seus clientes com base em seus atributos individuais (idade, sexo, média de gasto com o plano, etc.), bem como os tipos serviços por eles contratados para um melhor direcionamento das ofertas. A identificação de padrões de comportamento provavelmente ampliaria o êxito da companhia na venda de seus produtos, já que indivíduos dentro de um mesmo perfil possuem maiores chances de aderir às mesmas ofertas em contraste a clientes de outros perfis.

Diante dessa situação, evidencia-se o grande desafio de separação dos clientes em grupos frente ao número de atributos. Além disso, outro questionamento pertinente ao problema é em relação a quantidade de perfis a se criar, uma vez que não se sabe essa informação a priori, o que exige esforço de análise (Meira Jr., 2008, on-line). De acordo com Hruschka e Ebecken (2003, p. 16), o problema de obtenção do agrupamento ótimo é NP-Completo, pois o número possível de divisões distintas de um conjunto com n objetos em k grupos cresce, aproximadamente, a uma taxa de $\frac{k^n}{k!}$.

Isto significa que se desconhece um algoritmo em tempo polinomial capaz de resolvê-lo. Apresenta complexidade de ordem exponencial e, portanto, torna-se inviável a enumeração de todas as possíveis soluções a fim de encontrá-lo. Logo, vê-se a necessidade de utilização de heurísticas eficientes que possam tratar o problema em tempo hábil, onde a qualidade da solução encontrada deve ser avaliada empiricamente. Ankerst et al. (1999, p. 49) apontam três importantes fatores associados que representam um problema para

a efetivação dos algoritmos de clusterização:

- a dependência aos parâmetros de entrada pela grande maioria dos métodos;
- a sensibilidade aos parâmetros iniciais que, por muitas vezes, produzem resultados distintos;
- alta dimensionalidade do conjunto de dados que o impede revelar suas características intrínsecas apenas com um ajuste de parâmetro global.

Na Zootecnia, a análise de agrupamento pode ser utilizada na bovinocultura de leite para a identificação de grupos de animais com variações na perfomance produtiva, visando a sua otimização. Entretanto, deve-se levar em conta o fato de que o fenótipo ou valor observado de uma dada característica depende do potencial genético do animal, do ambiente no qual ele é produzido e da ação aditiva desses dois fatores, denominada interação genótipo-ambiente.

Em outras palavras, o fenótipo sofre influência do genótipo e do ambiente de forma conjunta e individual. Logo, melhorias no ambiente só serão proficientes se houver resposta do genótipo às mesmas. Portanto, a interação genótipo-ambiente compõe outro importante fator a ser considerado no melhoramento genético de animais para consequente aumento da produção de leite (Brito et al., 2007, on-line; Zampar, 2009, on-line).

Do ponto de vista molecular, o genótipo é tido como a identidade genética de cada ser. Desse modo, vale a pena mencionar que não existem indivíduos com genótipos totalmente iguais para todas as características, exceto nos casos de reprodução assexuada. Contudo, na prática, essa terminologia se estende a grupos de indivíduos, referindo-se a um genótipo parcial, onde se estabelece um conjunto de genes ou características de interesse (Canhas, 2011?, on-line).

Nessa definição mais ampla, considera-se como genótipo as raças, por exemplo, embora, em sentindo estrito, pode haver diferenças genéticas entre indivíduos de uma mesma raça devido às particularidades em sua composição. O ambiente, por sua vez, engloba uma série de variáveis, tais como: tipo de alimentação, clima, região, sistema de manejo, etc (Brito et al., 2007, on-line). Diferentemente do genótipo, o fenótipo muda

constantemente ao longo do tempo para algumas características que possuem muitos genes envolvidos em sua expressão, como a produção de leite (Polycarpo, 2008, on-line).

1.2 Justificativa

A ciência de dados, utilizada como instrumento de pesquisa no campo da Zootecnia, amplia seu campo de investigação e favorece a descoberta de conhecimento em conjuntos de dados para a predição e/ou tomada de decisões. Torna-se necessário o emprego de diferentes abordagens além da clássica, visto que, isoladamente, os métodos tradicionais de agrupamento não são capazes de analisar dados em massa com a escalabilidade demandada pelas aplicações atuais.

A produção de leite média diária por vaca fica abaixo do potencial de produção. Uma das formas para seu aumento é através da seleção de animais geneticamente superiores para a produção de leite, como o Gir Leiteiro, que, além de alta produtividade, adaptaram-se aos aspectos socioeconômicos e ambientais do território brasileiro (Abcz, s.d, on-line). Adicionalmente, possuem maior tolerância a doenças, condições climáticas e parasitas tropicais (Panetto et al., 2019, p. 9, 11).

A rentabilidade na atividade leiteira é atrelada à eficiência reprodutiva do animal. Logo, a busca por uma alta eficiência deve ser um dos principais focos de atenção dos produtores para aquisição de produtividade e retorno econômico. Para isso, é necessário que se adote mecanismos adequados de criação das novilhas, já que posteriormente elas serão as reprodutoras do rebanho. Animais com defict no desenvolvimento, seja por má alimentação ou meios de criação inadequados, ou mesmo por questões ligadas a sua genética, não são aptos a expressar todo o seu potencial de produção ao longo da vida (Carvalho et al., 2002, on-line).

Eficiência reprodutiva é a capacidade de fazer com que a vaca entre no período gestacional o mais breve possível após o período natural de espera. A baixa eficiência gera a redução da lucratividade, pela queda na produção de leite e diminuição do número de novilhas para reposição, como também o aumento dos gastos com sêmen, fármacos e serviços veterinários (Santos; Vasconcelos, 2007, on-line).

A identificação de características fenotípicas que separaram os animais em grupos

distintos possivelmente viabiliza a sua incorporação em avaliações genéticas ou aumenta a informatividade das mesmas na pesquisa. Além disso, a inclusão de características adicionais à produção de leite, no processo de avaliação e seleção de animais em programas de melhoramento genético, deve permitir a obtenção de fenótipos mais adequados às condições ambientais, contribuindo para a eficiência e sustentabilidade dos sistemas de produção, com foco em aspectos sociais, ambientais e econômicos.

Conforme anteriormente visto, o fenótipo é o valor observado de uma dada característica que, por sua vez, é dependente da genética, do ambiente e da interação genótipo-ambiente. Portanto, avaliar e compreender o relacionamento entre esses três fatores para utilizá-los em programas de melhoramento genético, otimizando os lucros por meio dos diferentes sistemas produtivos, pode oferecer grandes retornos econômicos à bovinocultura de leite (Brito et al., 2007, on-line).

1.3 Objetivos

Considerando as dificuldades associadas à clusterização de dados e o problema biológico anteriormente descritos, o principal objetivo deste trabalho é identificar conjuntos de animais com características fenotípicas distintas em relação à produção de leite. A suposição é de que existem variações fenotípicas dentro de um mesmo ambiente, possívelmente por questões genéticas e/ou variações de resposta da interação genótipo-ambiente. Além disso, essas varições também são esperadas com a mudança de ambiente, ponderadas em termos do tipo alimentação ao qual os animais foram submetidos. Este objetivo geral pode ser subdividido em:

i. Avaliar o comportamento dos algoritmos KM e FCM;

ii. Comparar os resultados entre os dois métodos;

iii. Verificar a consistência dos padrões identificados e tendências de agrupamento dos dados.

1.4 Organização do texto

Este trabalho está estruturado em cinco capítulos. No Capítulo 2 são apresentados diversos conceitos referentes ao processo de descoberta de conhecimento em bases de dados (KDD), como a análise de agrupamento. No Capítulo 3 é apresentada a metodologia empregue, bem como a descrição da base de dados, a condução das análises e os parâmetros utilizados nos algoritmos. No Capítulo 4 são exibidos os resultados pela abordagem proposta e discussões acerca destes. Por fim, o Capítulo 5 mostra as conclusões obtidas a partir da interpretação dos resultados.

2 Revisão Bibliográfica

A análise de agrupamento ou clusterização de dados é uma técnica de mineração de dados que permite a divisão dos dados em grupos (clusters) baseando-se em suas próprias características ou atributos, de forma que os objetos contidos em um mesmo grupo sejam mutuamente semelhantes e, ao mesmo tempo, distintos de elementos de outros grupos (Cavalcante Júnior, 2006, p. 4; Tan; Steinbach; Kumar, 2009, p. 585; Linden, 2009, p. 18).

A similaridade entre os objetos é medida segundo um critério predefinido que, geralmente, é alguma métrica de proximidade (Linden, 2009, p. 18; Peres et al., 2012, p. 127) que diz o quão semelhante um dado é em relação aos demais a partir da distância entre eles. Quanto maior a similaridade intracluster (homogeneidade interna) e maior a diferença intercluster (heterogeneidade externa), melhor a qualidade do agrupamento (Tan; Steinbach; Kumar, 2009, p. 585).

Essa técnica é considerada um método de aprendizado não supervisionado (Yonamine et al., 2002, p. 2), pois, a priori, não existem informações de grupos associados aos objetos da base e, diferentemente da classificação de dados, não dispõe de exemplos de treinamento de grupos previamente rotulados. Logo, os grupos são aprendidos de forma natural a partir dos atributos dos dados (Cassiano, 2014, p. 59).

A análise de agrupamento pode ser feita consoante com a abordagem clássica (*crisp*), onde cada elemento do conjunto de dados é associado exclusivamente a um único cluster. Entretanto, uma partição *crisp* pode ser muito restritiva em diversas aplicações práticas por não levar em conta imprecisões e/ou incertezas que frequentemente são encontradas nos dados (Yonamine et al., 2002, p. 3; Oliveira, 2016, p. 13). Além disso, pode haver um alto custo associado à exatidão dos valores e dificuldade de decisão acerca da relevância dos atributos que melhor definem um objeto.

Nessas situações, é válido a busca por abordagens alternativas que sejam capazes de incorporá-las ao domínio do problema, oferecendo assim uma melhor representação a dados imprecisos. A abordagem *fuzzy* mostra-se viável a esse cenário devido a flexibi-

lização relacionada à atribuição dos elementos aos clusters, visto que ela permite que um dado seja associado a mais de um grupo com diferentes níveis de pertinências no intervalo [0, 1].

Cada objeto que compõe a base é formado por um vetor de atributos que podem ser numéricos ou categóricos e, dependendo da heurística de agrupamento a ser utilizada, pode haver necessidade de conversão dos dados para um tipo que atenda à mesma. Considera-se atributo numérico todo aquele que possui uma representação numérica e que permite efetuar operações aritméticas sobre os mesmos, desde que façam sentido, podendo ser discretos (e.g., número de carros que uma pessoa possui) e contínuos (e.g., peso em kg, altura em cm). Em contrapartida, os atributos categóricos assumem valores qualitativos, i.e., são definidos por categorias, podendo ser nominais (e.g., gênero, tipo sanguíneo); e ordinais, que expressam a noção de ordem (e.g., grau de instrução, meses do ano). Há, ainda, os atributos binários, i.e., caso especial em que um atributo pode assumir apenas dois estados: 0 ou 1.

Tendo em vista a gama de dados gerada pelas aplicações atuais e a crescente necessidade de avaliação destes na busca por informações que possam agregar novos conhecimentos às organizações para tomada de decisões, pode-se dizer que a análise de agrupamentos é um recurso útil em diversas áreas. Todavia, o processo KDD a partir de grandes volumes de dados representa um dos principais desafios enfrentados pelas instituições, haja vista a dificuldade de armazenagem, processamento, gerência e análise desses dados, além do cuidado exigido na manipulação dos mesmos para a obtenção de interpretações válidas (Cavalcante Júnior, 2006, p. 1).

Para Vale (2005, p. 20), existem basicamente três macroetapas que compõem o processo KDD utilizando a análise de agrupamento, detalhadas a seguir:

- Preparação de dados;
- Agrupamento dos dados;
- Avaliação e interpretação dos resultados.

2.1 Preparação de dados

Segundo Louzada Neto e Diniz (2002, p. 25–26), a fase de preparação envolve a seleção de atributos, pré-processamento e transformação de dados e pode exigir cerca de 60% a 80% do tempo gasto em todo o processo, sendo que boa parte é utilizada na limpeza de dados.

2.1.1 Seleção de atributos

Um dos grandes desafios relacionados a atividades que envolvem análise de dados é a alta dimensionalidade do conjunto, posto que a maioria das aplicações do mundo real são definidas por um vasto número de atributos (Kantardzic, 2011, p. 56), muitos dos quais podem ser desnecessários para a mineração (Han; Kamber; Pei, 2012, p. 103). Dados tendem a se tornar cada vez mais esparsos conforme o aumento da dimensão (i.e., do número de atributos). Como efeito, isso pode dificultar substancialmente o processo de análise e comprometer a qualidade dos resultados obtidos. Esse fenômeno é tido na literatura como maldição de dimensionalidade. No agrupamento de dados, a alta dimensionalidade reduz a significância das noções de densidade e distância entre objetos que, por suas vezes, são cruciais para a definição dos grupos (Tan; Steinbach; Kumar, 2009, p. 60–62).

Na prática, a redução de dimensionalidade traz diversos benefícios ao processo KDD em conjuntos de dados com alto número de dimensões. Seu uso pode conduzir a modelos mais concisos e compreensíveis. Além disso, também auxilia no processamento de dados por diminuir o espaço de atributos, otimizando assim a performance dos algoritmos de mineração (Kantardzic, 2011, p. 56–57; Goldschmidt; Passos; Bezerra, 2015, p. 41).

O termo redução de dimensionalidade é normalmente associado ao conjunto de técnicas que reduzem a dimensão da massa de dados pela criação de novos atributos a partir da combinação dos atributos originais, sendo, portanto, sinônimo de extração de atributos (Tan; Steinbach; Kumar, 2009, p. 61). Todavia, a redução do número de atributos também pode ser realizada através da seleção de atributos, que descarta determinadas variáveis do processo de mineração. Ao contrário do que possa parecer,

essa abordagem não gera perda de informações, já que apenas os atributos irrelevantes ou redundantes são removidos (Sorzano; Vargas; Montano, 2014, p. 1).

Atributos irrelevantes geralmente contêm pouca ou nenhuma informação útil à mineração de dados, enquanto que os atributos redundantes replicam a informação contida nos dados. Logo, tornam-se susceptíveis à remoção. Em determinados casos, esses tipos de atributos podem ser facilmente identificados pelo conhecimento de domínio. Entretanto, a tarefa de escolha de atributos exige o uso de alguma metodologia de seleção (Tan; Steinbach; Kumar, 2009, p. 62–63).

O problema de se obter o melhor subconjunto de atributos é NP-Completo (Goswami et al., 2015, p. 28), dado que o espaço de busca requer $2^f - 1$ combinações distintas de atributos até encontrá-lo, onde f é o número de dimensões ou atributos (Goldschmidt; Passos; Bezerra, 2015, p. 41; Teng; Dong; Zhou, 2017, p. 4). Desse modo, é trivial perceber a inviabilidade do processo de busca exaustiva em conjuntos de dados de alta dimensionalidade na aquisição do subconjunto ótimo, considerando o fato de que o número de combinações cresce exponencialmente com o aumento do número de atributos. Naturalmente, isso cria a necessidade de se utilizar aborgagens alternativas na atividade de seleção. Tais estratégias são tipicamente divididas em três categorias: *filter*, *wrapper* e *embedded* (Kantardzic, 2011, p. 57; Garcia; Nievola; Paraiso, 2016, p. 639).

- *Filter*

  Na abordagem *filter*, a seleção do subconjunto ocorre previamente à aplicação do algoritmo de mineração, ou seja, a seleção de atributos é independente do processo de data mining (Tan; Steinbach; Kumar, 2009, p. 63). A avaliação da relevância dos atributos, em geral, é feita com base em uma medida estatística (Arunadevi; Nithya, 2016, p. 13557).

  Portanto, os filtros são uma maneira simples e rápida de se obter um subconjunto de atributos úteis para a tarefa de mineração (Garcia; Nievola; Paraiso, 2016, p. 640). Adicionalmente, embora não haja garantia de otimalidade, os filtros exigem menos esforço computacional, quando comparados às estratégias de abordagem *wrapper* (Suman; Thirumagal, 2013, p. 1).

- *Wrapper*

  Consiste na utilização do algoritmo de mineração como ferramenta de auxílio à obtenção do melhor subconjunto de atributos dentre as combinações de teste (Tan; Steinbach; Kumar, 2009, p. 63). Logo, a abordagem *wrapper* considera a seleção de atributos como um problema de busca, onde se definem diferentes subconjuntos de teste, os quais são comparados segundo um critério de validação, a depender do método escolhido para a tarefa de mineração (Arunadevi; Nithya, 2016, p. 13557), fornecendo assim um subconjunto subótimo de atributos por meio da busca heurística (Chandrashekar; Sahin, 2014, p. 18).

  Ao contrário dos filtros, as estratégias *wrapper* permitem a descoberta de possíveis associações entre atributos (Garali et al., 2018, p. 2). No entanto, conforme já mencionado, o principal revés dessa abordagem é o seu alto custo computacional (Kantardzic, 2011, p. 57) como resultado da necessidade de chamada ao algoritmo para cada subconjunto de atributo.

- *Embedded*

  O processo de seleção de atributos por essa abordagem é embutido no método de mineração, daí o seu nome. Em outras palavras, o algoritmo possui livre-arbítrio na escolha dos atributos (Tan; Steinbach; Kumar, 2009, p. 63).

  A abordagem *embedded* é uma proposta de compensação das fragilidades de ambas as metodologias anteriores. Logo, os modelos *embedded*, diferentemente dos filtros, possuem interação com o processo de aprendizagem e são computacionalmente menos intensivos do que as estratégias de abordagem *wrapper*, dado que não requerem a realização de múltiplas chamadas ao algoritmo para avaliar a relevância dos atributos (Ananthi; Palanivel, 2017, p. 25–26). Porém, são específicos a dadas máquinas de aprendizagem (Dhote; Agrawal; Deen, 2015, p. 1377).

Geralmente, a abordagem *filter* é escolhida para a tarefa de seleção de atributos em conjuntos de dados de alta dimensionalidade e com grande número de objetos graças a sua eficiência e autonomia em relação ao processo de mineração (Kantardzic, 2011, p. 57). O coeficiente de correlação de Pearson, também conhecido como coeficiente de correlação produto-momento (Han; Kamber; Pei, 2012, p. 96), é uma medida de proximidade (Suman; Thirumagal, 2013, p. 1) que, embora bastante criticada pela suposição de linearidade entre variáveis, é a mais amplamente conhecida e utilizada como método de filtro (Goswami et al., 2015, p. 32).

Como já enunciado, o coeficiente de Pearson (ρ), descrito pela equação 2.1 (Han; Kamber; Pei, 2012, p. 96), avalia a relação linear entre pares de atributos numéricos. Uma correlação é linear quando o relacionamento entre duas variáveis pode ser ajustado por uma reta, ou seja, há uma variação proporcional de seus valores. O coeficiente ρ pode assumir qualquer valor no intervalo $[-1, 1]$, onde 1 indica a existência de uma correlação positiva perfeita; e -1, uma correlação negativa perfeita. Valores próximos de 0 indicam uma fraca relação de linearidade. Entretanto, pode haver uma dependência não linear entre o par de atributos avaliado. Assim, o caso $\rho = 0$ exige a utilização de outros meios estatísticos para análise de outros tipos de correlação (Tan; Steinbach; Kumar, 2009, p. 91–93).

$$\rho = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B} \tag{2.1}$$

Onde n é o número de objetos, a_i e b_i são os i-ésimos valores de seus respectivos atributos (A e B), $\bar{A}$ e $\bar{B}$ representam suas respectivas médias e σ_A e σ_B são os desvios padrões de A e B, nessa ordem. Quanto maior o valor de ρ, mais forte é a correlação entre A e B. Logo, isso pode ser um indicativo de que A ou B possa ser removido do processo de mineração como um atributo redundante. Vale destacar também que o fato de A e B estarem correlacionados não implica na existência de causalidade entre os mesmos, ou seja, não necessariamente que A causa B ou vice-versa (Han; Kamber; Pei, 2012, p. 96–97).

De acordo com Hinkle, Wiersma e Jurs (2003, p. 109), considera-se a correlação

como desprezível, se $0 < |p| < 0.30$; fraca, se $0.3 \leq |p| < 0.5$; moderada, se $0.5 \leq |p| < 0.7$; forte, se $0.7 \leq |p| < 0.9$; e muito forte, se $|p| \geq 0.9$. Assim, à medida que o valor de ρ se aproxima de 1 ou -1, mais intensa é a associação linear entre duas variáveis, seja ela positiva ou negativa, respectivamente.

Independentemente da abordagem escolhida, pode-se dizer que o intuito da seleção de atributos é a obtenção de um subconjunto de atributos que expressam relevância e poder de distinção na caracterização dos objetos. Conforme Vale (2005, p. 21), isso tende a auxiliar a etapa de agrupamento de dados, evitando prejuízos à formação dos clusters e reduzindo o tempo de extração do conhecimento, uma vez que se remove atributos desnecessários do processo.

2.1.2 Pré-processamento

A anomalia de dados é um problema recorrente em situações do mundo real. O processo de coleta de dados é, geralmente, pouco controlado e susceptível a falhas. Dados tendem a ser imprecisos, incompletos e inconsistentes por diversos fatores. Primeiro porque podem haver falhas nos equipamentos que realizam a coleta dos mesmos. Segundo, devido a erros na transmissão de dados. Terceiro, por limitações de infraestrutura para armazená-los e gerenciá-los. Quarto, em virtude de erros na entrada de dados que, por vezes, resultam em falsas ausências de dados. Usuários podem incorretamente definir valores padrões para atributos quando na verdade não o querem. Quinto, pela falta de convenções de nomenclaturas ou incoerências no formato dos campos de entrada. Por último, porque o mesmo dado pode ser capturado mais de uma vez, gerando duplicatas dentro da base (Han; Kamber; Pei, 2012, p. 84).

O pré-processamento de dados tem como objetivo a identificação e tratamento dessas anomalias com vistas à qualidade dos dados para a aquisição de resultados válidos. Doni (2004, p. 15) lista alguns dos principais problemas a serem abordados no tratamento ou limpeza de dados:

- Remoção de duplicatas e dados corrompidos

  Nesse passo, dados corrompidos e/ou duplicados, i.e., objetos que possuem os mesmos valores para todos atributos da base, são eliminados.

- Tratamento de *outliers*

  Ruídos ou *outliers* são pontos de dados que apresentam valores de atributos muito destoantes dos demais, os quais "prejudicam a análise de dados" (Oliveira, 2015, p. 17), afetando diretamente a qualidade dos resultados alcançados com as heurísticas de agrupamento e, portanto, devem ser tratados.

  A estatística descritiva dispõem de mecanismos úteis à identificação de *outliers*. Usualmente, o método empregado, presente na obra de Kumar et al. (2015, p. 671), é baseado na amplitude interquartil (IQR - *Interquartile Range*), que define um intervalo aceitável de valores para cada atributo. As equações 2.2 e 2.3 representam, respectivamente, o limitante inferior e superior para detecção de *outliers*:

$$L_{inf} = Q_1 - 1.5 \times IQR \tag{2.2}$$

$$L_{sup} = Q_3 + 1.5 \times IQR \tag{2.3}$$

  Q_1 e Q_3 são os 1º e 3º quartis, nessa ordem. IQR, conforme anteriormente citado, é a amplitude interquartil e, portanto, corresponde à diferença entre Q_3 e Q_1, ou seja, $IQR = Q_3 - Q_1$. Por ambas as equações, tem-se que *outliers* são todos os valores fora do intervalo $[L_{inf}, L_{sup}]$

  Há diferentes maneiras de lidar com esse tipo de anomalia. Uma delas é excluí-las das análises tal como é feito no caso de dados repetidos e corrompidos. Contudo, essa ação deve ser considerada de forma criteriosa, pois poderia-se remover dados que, possivelmente, contivessem valor interpretativo e agregassem valiosas informações ao estudo. Outras formas de tratamento incluem o descarte de períodos de observação e a correção de dados, que realiza ajustes sobre esses valores.

  De todo modo, a opinião de especialistas de domínio somada ao conhecimento prévio em relação à semântica dos atributos são fundamentais na determinação do método mais apropriado (Sassi, 2006, p. 32).

- Tratamento de valores faltantes (*missing values*) e inconsistências de dados

  Louzada Neto e Diniz (2002, p. 29–30) justificam as ausências de valores nos con-

juntos de dados como decorrências da indisponibilidade do dado no momento de sua coleta ou simplesmente devido a falhas humanas ou conflitos de informações obtidas de diferentes fontes, que podem gerar inconsistências de dados. Um procedimento exequível nesses casos é a utilização de medidas estatísticas, como a média de valores de objetos aparentemente semelhantes, por exemplo, na tentativa de preenchê-los.

Essas anomalias também podem ser corrigidas pela mesma técnica aplicável a dados duplicados e corrompidos, onde descarta-se as instâncias que as contenham. Apesar de mais simples, esse método é menos eficiente que a solução anterior, visto que há possibilidade de perda de informações caso haja um número expressivo desses valores dentro do conjunto de dados.

Outro tratamento bastante eficaz é a predição do valor de maior probabilidade por técnicas de regressão, indução por árvores de decisão ou com o uso de ferramentas de inferência, que, segundo Côrtes, Porcaro e Lifschitz (2002, p. 25), tende a manter a relação entre o atributo predito e os atributos independentes, posto que mais informações são consideradas no processo de estimativa.

Em razão da dificuldade de se manipular e processar grandes volumes de dados como consequência de limitações técnicas e/ou restrições temporais, é importante considerar a redução de numerosidade (*numerosity reduction*) como parte do pré-processamento. Seu objetivo é promover a redução da quantidade de dados para representações menores. Pode-se dizer que o foco dessa tarefa é o ganho de desempenho pela diminuição do volume de dados processados, preservando-se a qualidade dos resultados (Souza, 2017, p. 52).

As estratégias relacionadas à redução da quantidade de dados dividem-se em dois tipos: Métodos paramétricos e não paramétricos (Han; Kamber; Pei, 2012, p. 99–100, 105–110; Delaporte, 2015, p. 6–7). Métodos paramétricos mantêm apenas os parâmetros estimados do modelo em vez dos dados (e.g., modelos de regressão e log-lineares).

- Modelos de regressão

  Em uma regressão linear simples, os dados são ajustados a uma reta mediante uma

função linear do tipo $y = ax + b$, em que x e y são variáveis aleatórias, onde x corresponde à variável explicativa (independente) e y à variável resposta (dependente). a e b são os coeficientes da regressão e representam, respectivamente, a inclinação da reta e a interceptação com o eixo y.

Às vezes, uma única variável independente pode não ser suficiente para explicar a variável de interesse. Sendo assim, é necessário utilizar outros modelos de regressão que permitem considerar mais de uma variável explicativa no processo, tal como a regressão linear múltipla.

- Modelos log-lineares

  Aplicáveis na descrição de relações entre duas ou mais variáveis categóricas. Os modelos são construídos com base na distribuição de frequências e estimam a probabilidade de cada ponto no espaço multidimensional pela combinação linear dos parâmetros do modelo.

Comparativamente, ambos os modelos paramétricos lidam bem com distribuições assimétricas e podem ser utilizados em situações de esparsidade de dados. Entretanto, regressões podem ser computacionalmente caras em grandes conjuntos de dados, enquanto que modelos log-lineares são escaláveis para conjuntos de até dez atributos, embora restritos a dados categorizados (Han; Kamber; Pei, 2012, p. 105–106). Por outro lado, métodos não paramétricos dispensam o uso de modelos na redução e sumarizam os dados através de amostras ou de algum recurso de visualização de dados (e.g., amostragem, histogramas).

- Amostragem

  Consiste na redução do volume de dados a um subconjunto de amostras representativas para análise e compreensão do todo. Na amostragem de dados, é importante levar em conta a qualidade da amostra escolhida, dado que uma má seleção do subconjunto pode enviesar o estudo (Nogueira Júnior; Monico; Tachibana, 2004, p. 103). Sendo assim, é necessário a adoção de processos adequados de amostragem a fim de garantir a manutenção das características da população e,

consequentemente, a fidelização dos resultados.

A representatividade do subconjunto, por sua vez, é dependente do tamanho da amostra, pois tamanhos grandes aumentam a probabilidade de representação da massa original de dados, mas também reduzem suas vantagens de utilização e vice-versa, já que tamanhos muito reduzidos podem gerar perda de informações (Tan; Steinbach; Kumar, 2009, p. 58). Em vista disso, nota-se que a determinação do tamanho adequado da amostra pode não ser uma tarefa trivial.

Considerando esse problema, existem dois índices estatísticos que são usados como delimitadores de possíveis erros cometidos na generalização dos resultados: margem de erro e grau de confiança (Samohyl, 2009, p. 62). A margem de erro indica a precisão entre o resultado amostral e o resultado esperado sobre a população. Em outras palavras, esse parâmetro expressa o limite aceitável para o erro amostral. O grau de confiança, entretanto, refere-se ao nível de certeza de que os dados estejam dentro da margem de erro.

Os métodos de amostragem classificam-se em probabilísticos e não probabilísticos. Para Silva (2009, p. 88), em geral, as abordagens probabilísticas são mais utilizadas devido a maior representividade amostral proporcionada pela randomicidade presente nas mesmas, além de permitirem o cômputo do erro associado à inferência. O autor ainda descreve quatro importantes técnicas de amostragem probabilística: amostragem aleatória simples, amostragem estratificada, amostragem sistemática e amostragem por conglomerados.

A amostragem aleatória simples constitui o método mais básico dentre as técnicas probabilísticas (Tan; Steinbach; Kumar, 2009, p. 57). Nela, todos os objetos possuem a mesma probabilidade de seleção, onde um determinado número de elementos são escolhidos ao acaso dentro da população para compor a amostra. Pode ser realizada sem repetição, i.e., quando há remoção do objeto selecionado do conjunto original de dados; ou com repetição, caso contrário.

É natural supor que, na amostragem com repetição, quanto maior o tamanho populacional, menor será a chance de reelegibilidade de um elemento. Com base nisso,

é possível dizer que suas diferenças para a amostragem sem repetição tendem a se reduzir em grandes populações (Stewart, 2009, p. 699; Danish, 2017, p. 161; Rudas, 2018, p. 22). Cascarino (2017, p. 35) afirma que, na prática, o uso da amostragem sem repetição é mais comum, porém, segundo Osler II (2012, p. 418), é difícil contemplar todas as características da população no conjunto amostral quando o universo é muito grande.

Para isso, existem outras técnicas de amostragem que, de acordo com o mesmo, são mais eficientes por reduzirem o erro amostral, porém mais caras. Uma delas é a amostragem estratificada que, basicamente, é a aplicação de outro método de amostragem, como, por exemplo, a amostragem aleatória simples, sobre subconjuntos bem definidos da população, denominados estratos, assumindo que existam heterogeneidade entre os estratos e homogeneidade dentro dessas subunidades (Atangana, 2017, p. 41–42).

O processo de amostragem estratificada consiste na seleção aleatória de um número de elementos dentro de cada estrato previamente identificado e apresenta duas formas distintas de utilização. Na primeira versão, as amostras obtidas dos estratos contêm o mesmo número de objetos. Em uma outra abordagem, o tamanho das amostras é proporcional ao tamanho de cada estrato (Tan; Steinbach; Kumar, 2009, p. 57). Assim sendo, pode-se dizer que esse método é útil em situações de desbalanceamento de dados, i.e., quando o número de objetos de determinados grupos do conjunto é muito superior aos demais. Isso porque a estratificação promove um equilíbrio entre elementos no subconjunto amostral.

No entanto, é mais conveniente usar a amostragem sistemática nos casos em que existe uma relação de "ordem natural" dos dados (Jelihovschi, 2014, p. 13). O funcionamento dessa técnica se dá por meio da seleção aleatória de um objeto do conjunto. Após isso, toma-se os i-ésimos elementos acessíveis da população dentro de um padrão amostral predeterminado, de forma sistemática (Guimaraães, 2012, p. 21). De maneira mais formal, seja N o tamanho da população, a amostra será definida pelos elementos nas posições p, dadas por $p = x + \left\lfloor \frac{N}{n} \right\rfloor \cdot i$, onde x é um

número sorteado de forma aleatória entre 1 e $\left\lfloor \frac{N}{n} \right\rfloor$ (padrão amostral), n = tamanho desejado da amostra e i um inteiro que varia de 0 a n - 1.

Para entender melhor o seu funcionamento, suponha que se deseja obter uma amostra de tamanho 10 em um universo de 100 elementos ordenados. No passo inicial, sorteia-se um número aleatório entre 1 e 10. Considere agora que o número sorteado foi 5. Por fim, toma-se para a amostra todos os elementos cujas posições satisfazem $p = 5 + 10 \cdot i$, com i variando de 0 a 9. Portanto, a amostra obtida no final do processo é formada pelo subconjunto de posições {5, 15, 25, 35, 45, 55, 65, 75, 85, 95}.

Acharya et al. (2013, p. 331) destacam algumas das principais vantagens da amostragem sistemática. De acordo com eles, esse método possui um custo de aplicação moderado e altos níveis de validades interna e externa, que associam, respectivamente, à corretude do resultado com base no subconjunto amostral e à generalização dos mesmos a outras populações. Por outro lado, afirmam também que a representatividade de algumas parcelas da população pode ser comprometida, dado que a aleatoriedade é considerada apenas na seleção do primeiro objeto. Logo, alguns elementos têm chance nula de escolha nas seleções posteriores.

Finalmente, o método de amostragem por conglomerados que, segundo Kumar (2010, p. 204), baseia-se na existência de grupos naturais com características visíveis ou de fácil percepção dentro da população para definir a subunidade amostral. À primeira vista, esse método pode levar à falsa impressão de similaridade com a amostragem estratificada. Todavia, o processo de seleção da amostra é essencialmente divergente (Barata et al., 2005, p. 185; Khan, 2011, p. 87).

Enquanto que, na amostragem estratificada, os estratos são mutuamente diferentes em relação às características dos elementos que, por suas vezes, compartilham propriedades em comum dentro de cada subconjunto; na amostragem por conglomerados, espera-se que os grupos sejam internamente heterogêneos e externamente homogêneos, de modo que não existam grandes variações entre o estudo isolado de cada grupo. Portanto, pode-se notar que, nessa abordagem, os grupos devem ser

capazes de resumir toda a população individualmente devido a grande variabilidade de objetos no interior dos mesmos.

Além disso, a técnica de amostragem por conglomerados também difere da estratificada quanto à unidade de amostragem, visto que na composição amostral, a unidade considerada pelos métodos são, respectivamente, os grupos e os elementos da população. Desse modo, é trivial perceber que, no processo de formação da subunidade pela amostragem por conglomerados, todos os objetos contidos em um grupo são selecionados para a amostra.

Segundo Dhivyadeepa (2015, p. 85), o número de grupos a serem amostrados por essa técnica é igual à razão do tamanho da amostra desejado pelo número médio de elementos em cada grupo. A exemplo, considere que se queira amostrar 5000 alunos de uma universidade com 20000 estudantes. Naturalmente, uma característica lógica de divisão desses alunos é o tipo de curso. Suponha agora que, dos 100 cursos oferecidos pela instituição, o número médio de alunos por curso seja de 200. Portanto, o número de grupos (cursos) a serem selecionados para a amostra é igual a $\frac{5000}{200}$ ou 25. Na prática, isso quer dizer que dos 100 cursos disponíveis, 25 serão escolhidos de forma aleatória. Por conseguinte, todos os 5000 alunos estarão presentes na amostra ao final do processo, visto que cada um dos 25 cursos selecionados contém 200 indivíduos.

Dhivyadeepa (2015, p. 86–87) também define alguns prós e contras do uso dessa técnica. Para ele, essa abordagem apresenta menor custo em comparação com outros métodos de amostragem, uma vez que é mais fácil amostrar grupos a elementos, tornando o processo de seleção da amostra mais rápido e simplificado. Contudo, o autor ainda revela que, dentre as técnicas probabilísticas, é a que menos oferece representatividade populacional, pois, em geral, há tendência de similaridade entre os objetos de um mesmo grupo, prejudicando assim, a qualidade da amostra e, consequentemente, os resultados. Logo, é um método com mais susceptibilidade a apresentar maior erro amostral em decorrência da possibilidade de haver grupos limitados na amostra que impedem a inclusão de grande parcela da população no subconjunto, comprometendo a sua representatividade.

Face ao maior enfoque dado ao processo de amostragem, descrito anteriormente, é fácil perceber que é o principal método de mineração de dados empregado na redução populacional (Ricci; Rokach; Shapira, 2015, p. 231), haja vista a sua simplicidade e seus benefícios sobre grandes volumes de dados. Através dela, é possível otimizar a performance dos algoritmos de agrupamento pela diminuição da quantidade de objetos processados pelos mesmos e, consequentemente, viabilizar o processo KDD neste cenário.

- Histogramas

  Histograma é uma técnica de visualização de dados amplamente utilizada na representação de objetos para análise e compreensão do conteúdo de uma ou mais variáveis mediante à sumarização de sua distribuição (Han; Kamber; Pei, 2012, p. 54). Constitui-se, basicamente, de um gráfico em colunas que exibem a distribuição de frequências de um determinado atributo. Em essência, a principal diferença dessa representação para um gráfico de barras verticais é a ausência de espaçamento entre as mesmas (Jelihovschi, 2014, p. 30–31), ou seja, no histograma as barras são justapostas.

  Gráficos de barras são usados na representação de variáveis categóricas, onde cada coluna indica uma categoria distinta da mesma. A altura das barras é definida pela frequência de ocorrência da respectiva categoria no conjunto de dados, i.e., o número de vezes em que ela se repete. Por outro lado, utilizam-se histogramas para representar variáveis numéricas. Nesse caso, é realizada uma divisão do intervalo de valores em subconjuntos disjuntos e justapostos, denominados *buckets* ou *bins*, que, em geral, possuem o mesmo comprimento (Han; Kamber; Pei, 2012, p. 54). A altura de cada *bucket* é dada pelo número de objetos nele contidos.

  Histogramas são uma maneira simples e rápida de avaliar a centralidade da distribuição, i.e., em quais faixas de intervalos há maior concentração de dados. Além disso, mostram-se úteis tanto a distribuições uniformes quanto em situações de distorção de dados (Han; Kamber; Pei, 2012, p. 108). Sendo assim, pode-se dizer que esse método permite avaliar o tipo de distribuição que os dados tendem a seguir pelo seu formato. Para isso, é necessário conhecê-los. Kume (1993, p. 53–55)

caracteriza em sua obra os tipos mais comuns de histogramas, também inclusos em César (2011, p. 74–78) e Stapenhurst (2013, p. 339), descritos a seguir.

– Simétrico, normal ou em forma de sino (*Bell-shaped*)

  A configuração característica de um histograma em forma de sino é a centralização do valor médio dos dados (i.e., o pico do gráfico localiza-se no centro da distribuição) e há redução gradual e simétrica da frequência em direção aos extremos. Geralmente, verifica-se esse tipo de histograma quando o processo é estável e não existem grandes variações nos dados (Grupta; Valarmathi, 2009, p. 70).

– Pente ou multimodal (*Saw-toothed / comb*)

  O histograma multimodal é identificado pela alternância entre baixas e altas frequências, ou seja, a distribuição apresenta vários picos. Pode ocorrer quando há grande variabilidade do número de elementos entre classes vizinhas, de forma intercalada.

– Assimétrico (*Skewed*)

  A assimetria é justificada pelo deslocamento do valor médio para um dos lados do gráfico de modo que a frequência diminui gradualmente de um lado e mais bruscamente do outro. Tende a ser observado nos casos onde existe um limite de especificação da variável, não permitindo que ela assuma valores maiores ou menores que o mesmo.

– Abrupto ou despenhadeiro (*Precipice*)

  Tal como no caso anterior, em um histograma despenhadeiro, o valor médio também se encontra deslocado do centro do gráfico. Pode haver diminuição abrupta da frequência em um dos lados e de forma tênue do outro, ou ainda, de modo abrupto por ambos os lados. Ocorre, provavelmente, pela remoção de dados do processo, causando a impressão de incompletude do gráfico; ou

quando a assimetria se torna mais evidente.

– Achatado ou platô (*Plateau*)

As classes centrais de um histograma achatado possuem altas frequências de aproximadamente mesmo valor. Em contrapartida, as extremidades apresentam frequências menores. Isso pode ser causado pela miscelânia de distribuições com diferentes médias.

– Dois picos ou bimodal (*Double-peaked*)

Intuitivamente reconhecido pela existência de dois picos no gráfico dispostos não ipsilateralmente. No histograma bimodal, a frequência é menor em torno do centro e os picos localizam-se em lados opostos. Acontece quando há combinação de duas distribuições com médias distintas.

– Pico isolado ou ilhas isoladas (*Isolated-peaked*)

Nesse caso, observam-se duas regiões separadas uma da outra, ambas com formato semelhante ao de um histograma normal. Em geral, ocorre quando há falhas de medições e no registro de dados oriundos de um processo distinto.

2.1.3 Transformação de dados

É muito comum que, em aplicações reais, conjuntos de dados de alta dimensionalidade apresentem diversidade de atributos. Nesse contexto, devido a restrições impostas pelas técnicas de mineração quanto ao seu tipo, muitas vezes, é necessário modificar a natureza desses dados, já que grande parte delas estão limitadas à manipulação de determinados tipos de atributos. A transformação de dados é a etapa na qual se altera a representação inicial dos mesmos para um formato adequado, de modo a tornar o processo de mineração mais eficiente, facilitando a compreensão dos padrões identificados (Han; Kamber; Pei, 2012, p. 111). Algumas das estratégias de transformação de dados incluem a conversão de atributos e a normalização de dados, definidas a seguir.

- Conversão de atributos

  Conforme previamente mencionado, os atributos podem ser classificados como ordinais, nominais, binários e numéricos. No entanto, as formas de conversão variam dependendo do seu tipo. Para o caso ordinal, a conversão de seus valores para representações numéricas deve, de alguma forma, preservar a noção de ordem existente entre os mesmos. Por exemplo, considere que uma variável que expressa o nível de satisfação dos clientes referente a determinado serviço possa assumir os valores *péssimo*, *ruim*, *regular*, *bom*, *muito bom* e *ótimo*. Tais valores podem ser, respectivamente, mapeados nos elementos do conjunto {0, 2, 4, 6, 8, 10}, visto que a relação subjetiva de ordem é mantida após a conversão.

  A transformação de atributos nominais, por outro lado, é feita por binarização, i.e., técnica utilizada na conversão de valores nominais em atributos binários. Inicialmente, nesse processo, para cada categoria distinta da variável, cria-se um novo atributo. Em seguida, atribuem-se 1 às instâncias que as contenham e 0, quando não. Caso o atributo a ser transformado contenha apenas duas categorias (e.g., masculino, feminino), o processo de conversão é imediato. É óbvio que, pelo fato dessa abordagem aumentar a dimensionalidade do conjunto de dados, pode ser interessante agrupar alguns dos valores antes de transformá-los se o atributo em questão apresentar um número grande de categorias ou quando um valor ocorre de forma esporádica (Tan; Steinbach; Kumar, 2009, p. 69).

  Entretanto, há certos algoritmos de mineração de dados que, mesmo na atividade de agrupamento, exigem que os dados sejam nominais para que seja possível manipulá-los. Sendo assim, pode haver necessidade de transformar valores contínuos em valores nominais, agrupando-os dentro de uma faixa de valores. Tal técnica é conhecida como discretização (Tan; Steinbach; Kumar, 2009, p. 69). Para ilustrar, suponha que se deseja discretizar a estatura de indivíduos de uma população cujos valores variam entre 1.50 e 1.80 m. O conjunto de alturas h pode ser dividido em três subintervalos: baixo ($h \leq 1.60$ m), médio (1.60 m $< h <$ 1.70 m) e alto ($h \geq 1.70$ m). Essa abordagem, portanto, é constituída de duas etapas. Na primeira delas, definem-se o número de categorias, os limites e tamanho de seus intervalos.

Por fim, associa-se todos os valores de um intervalo à mesma categoria.

- Normalização

  A análise de dados, em muitos casos, pode ser afetada pela unidade de medida utilizada. A mudança de unidade dos atributos na mineração de dados pode produzir resultados bem distintos (Han; Kamber; Pei, 2012, p. 113). Além disso, a diferença de escala entre as variáveis também tende a ser um dos fatores que causam essa discrepância, dado que atributos com valores muito grandes exercem maior influência no processo (Tan; Steinbach; Kumar, 2009, p. 77). Vale (2005, p. 23) diz que, na clusterização, por exemplo, isso se torna evidente no cômputo da semelhança ou diferença entre objetos pelas métricas de proximidade.

  A normalização de dados é um artifício capaz de lidar com tal problema. Esse método visa reduzir a dependência da escolha de unidade existente no processo (Han; Kamber; Pei, 2012, p. 113). Desse modo, permite-se que os atributos possuam mesma influência, já que os dados são representados em uma mesma escala de valores após serem normalizados. Apesar disso, Faceli et al. (2011, p. 45) apontam que, dependendo da situação, é necessário considerar essas diferenças devido sua relevância ao contexto da aplicação.

  Geralmente, utilizam-se os intervalos [-1, 1] ou [0, 1] como faixas de valores na mudança de escala (Han; Kamber; Pei, 2012, p. 113). Contudo, existem diferentes métodos que podem ser empregados na etapa de normalização de dados. Alguns dos seus tipos são (Goldschmidt; Passos; Bezerra, 2015, p. 61–65):

  - Min-Max

    Nesse método, são redefinidos os limites mínimo (min_{novo}) e máximo (max_{novo}) dos atributos à nova escala de valores. Consiste, basicamente, na realização de uma transformação linear sobre os dados a serem normalizados. Segundo Faceli et al. (2011, p. 45), na primeira etapa do processo de normalização de um dado valor v_i, subtrai-se o mesmo pelo valor mínimo do atributo (min_{attr}). Em seguida, divide-se o resultado anterior pela diferença entre o maior (max_{attr}) e menor valores do atributo. O quociente da divisão é então multiplicado pela

diferença entre os limites da nova escala. Por fim, adiciona-se o menor valor do atributo da nova escala, i.e., min_{novo}, ao resultado do produto efetuado no passo anterior. Isso faz com que v_i seja mapeado a um valor v_{novo} dentro do intervalo $[min_{novo}, max_{novo}]$. A equação 2.4, definida a seguir, ilustra o processo de normalização por esse método (Han; Kamber; Pei, 2012, p. 114).

$$v_{novo} = \frac{v_i - min_{attr}}{max_{attr} - min_{attr}}.(max_{novo} - min_{novo}) + min_{novo} \tag{2.4}$$

Em particular, a equação de normalização de um dado valor v_i ao intervalo [0, 1] é dada por:

$$v_{novo} = \frac{v_i - min_{attr}}{max_{attr} - min_{attr}} \tag{2.5}$$

Como exemplo, considere a seguinte situação. Suponha que o menor e maior valores de determinado atributo sejam, respectivamente, 1000 e 5000. Deseja-se utilizar a normalização Min-Max para mapear seus valores ao intervalo [0, 1]. Como os limites da nova escala são 0 e 1, é mais lógico utilizar diretamente a equação 2.5 nesse processo. Seja 3800 um dos valores existentes no conjunto de dados para esse atributo a ser normalizado, então seu novo valor após o mapeamento corresponde a $\frac{3800-1000}{5000-1000} = 0.7$

A normalização Min-Max mantém a relação existente entre os dados antes de sua utilização. Entretanto, pode haver valores fora do intervalo da nova escala caso uma entrada adquirida após a aplicação da técnica esteja além dos limites da escala original do atributo (Han; Kamber; Pei, 2012, p. 114).

– Z-score

O cômputo do novo valor (v_{novo}) através dessa técnica é baseado na média (μ_{attr}) e desvio padrão (σ_{attr}) dos atributos. De acordo com Faceli et al. (2011, p. 45), após a transformação, essas medidas de localização e dispersão do conjunto de valores normalizados, i.e., média e desvio padrão, valem 0 e 1, nessa ordem. Portanto, ao contrário do que acontece na normalização Min-Max, onde se define uma nova escala de valores aos atributos; no método

Z-score, o que muda são os valores de média e desvio padrão (distintos para cada atributo) a valores comuns a todos eles.

A primeira etapa desse processo de normalização consiste em calcular a média e o desvio padrão para cada atributo. Após isso, para cada valor v_i a ser normalizado, subtrai-se o valor de média do respectivo atributo, i.e., $v_i - \mu_{attr}$. Por último, divide-se o resultado da operação anterior pelo valor de desvio padrão desse mesmo atributo. Logo, v_i é então normalizado a v_{novo} através da equação 2.6, definida como:

$$v_{novo} = \frac{v_i - \mu_{attr}}{\sigma_{attr}} \tag{2.6}$$

Considere o exemplo apresentado no método anterior para ilustrar esse processo. Suponha que os valores de média e desvio padrão do atributo sejam, respectivamente, 2200 e 400. Pela normalização Z-score, o valor 3800 equivale a $\frac{3800-2200}{400} = 4$.

É conveniente utilizar a normalização Z-score quando se desconhece os valores limites do atributo. Além disso, essa técnica é preferível à normalização Min-Max na presença de *outliers* (Han; Kamber; Pei, 2012, p. 114), pois mesmo após a sua aplicação, ainda é possível detectá-los.

– Escalonamento decimal

A normalização por essa técnica desloca a casa decimal dos valores dos atributos à esquerda. O pré-requisito para sua utilização é conhecer o maior valor absoluto do atributo a ser normalizado (Han; Kamber; Pei, 2012, p. 115), o qual serve de base para a determinação do expoente j do denominador da equação 2.7, que define o processo de normalização desse método, descrita a seguir.

$$v_{novo} = \frac{v_i}{10^j} \tag{2.7}$$

Onde j é o menor inteiro tal que $max(|V_{novo}|) < 1$, ou seja, j representa o menor valor de modo que o máximo valor absoluto do atributo normalizado

seja inferior a 1. Logo, para encontrá-lo, basta verificar o inteiro de menor valor que satisfaz a relação $max(|V|) < 10^j$, em que $V = \{v_1, ..., v_n\}$, i.e., V é o conjunto original de valores. Depois de determinado, calcula-se o valor do denominador (10^j) da referida equação. Finalmente, v_i é mapeado a v_{novo} no intervalo $[\frac{min_{attr}}{10^j}, \frac{max_{attr}}{10^j}]$ dividindo-se seu valor pelo resultado da exponenciação do passo anterior.

Tomando como exemplo o mesmo cenário ilustrativo dos métodos previamente descritos, tem-se que 5000 é o maior valor absoluto do atributo. Portanto, $j = 4$, pois é o menor inteiro que satisfaz $5000 < 10^j$. Daí, por escalonamento decimal, o valor 3800 é normalizado a $\frac{3800}{10^4} = 0.38$. Observe que, como 1000 e 5000 são, respectivamente, o menor e maior valores do atributo, o intervalo de abrangência do conjunto de valores normalizados corresponde, nesse caso, a [0.1, 0.5].

– Máximo absoluto

No critério de valor máximo absoluto, intuitivamente, utiliza-se o maior valor absoluto do atributo como referência no processo de normalização. Com isso, v_i é mapeado a um novo valor (v_{novo}) a partir de sua divisão pelo maior valor em módulo ($max(|V|)$ do atributo ao qual ele pertence, conforme a equação 2.8, definida a seguir. Nesse caso, os possíveis intervalos de mapeamento são [-1, 0], [0, 1] ou [-1, 1], a depender do sinal dos valores de cada atributo, ou ainda, alguma variação dentro de algum deles.

$$v_{novo} = \frac{v_i}{max(|V|)} \tag{2.8}$$

Com a utilização dessa técnica sobre o mesmo exemplo dos demais métodos, o valor 3800 é normalizado a $\frac{3800}{5000} = 0.76$. Adicionalmente, a faixa de valores [0.2, 1] corresponde ao intervalo do conjunto após a normalização dos mesmos, tendo em vista que os limites inferior e superior do atributo são, respectivamente, 1000 e 5000.

– Soma total

O método da soma total utiliza como base do processo de normalização a soma dos valores de cada atributo. Logo, o mapeamento de um valor v_i a v_{novo} é dado por meio da razão entre v_i e o somatório dos n valores do respectivo atributo. Nessa abordagem, é natural supor que, quanto maior a ordem de magnitude do atributo a ser normalizado e/ou maior o número de objetos, menor é o intervalo no qual seus valores são mapeados, dado o aumento do valor do somatório, presente na equação 2.9, definida a seguir, a qual define seu processo de normalização.

$$v_{novo} = \frac{v_i}{\sum_1^n v_k} \tag{2.9}$$

Considere a aplicação dessa técnica sobre a situação ilustrativa dos métodos de normalização anteriores. Suponha que o conjunto de dados é composto por pouco mais de 1000 objetos e que a soma de todos os seus valores é igual a 2500000. Logo, o valor 3800 é normalizado a $\frac{3800}{2500000} = 0.00152$ no intervalo [0.0004, 0.002], já que o os valores do conjunto original variam de 1000 a 5000. Note que, mesmo com um pequeno número de elementos, pode haver uma grande redução do intervalo do atributo normalizado.

2.2 Agrupamento dos dados

O agrupamento dos dados constitui a segunda macroetapa do processo KDD por clusterização. O objetivo dessa fase é agrupar os dados pré-processados com o auxílio de heurísticas de agrupamento. Acerca disso, a divisão mais comum dos algoritmos de clusterização leva em conta o formato do agrupamento (Goldschmidt; Passos; Bezerra, 2015, p. 97) e a questão da obrigatoriedade ou não de pré-especificação do número de grupos. Portanto, tradicionalmente são classificados como métodos hierárquicos e métodos particionais (Tan; Steinbach; Kumar, 2009, p. 586–587).

Métodos hierárquicos criam uma sequência aninhada de grupos sob a forma de

um dendograma ou estrutura de árvore (Kantardzic, 2011, p. 252). Em outras palavras, esses métodos fornecem uma estrutura de organização hierárquica dos dados a partir de sucessivas divisões dos mesmos (Tan; Steinbach; Kumar, 2009, p. 587), permitindo assim, visualizar o processo de composição dos clusters em cada nível da árvore e a relação entre eles. Além disso, não exigem a escolha inicial do número de grupos para uso (Kantardzic, 2011, p. 259).

Existem duas abordagens associadas ao esquema hierárquico de agrupamento: aglomerativa e divisiva (Linden, 2009, p. 25; Kantardzic, 2011, p. 260; Sidhu; Kaur, 2013, p. 711), também denominadas *bottom-up* e *top-down* (Han; Kamber; Pei, 2012, p. 449; Goldschmidt; Passos; Bezerra, 2015, p. 98–99), respectivamente. Na primeira, i.e., abordagem aglomerativa, cada objeto do conjunto de dados define um grupo individualmente no início do processo. Em seguida, os grupos ou objetos mais próximos são unidos uns aos outros de forma sucessiva até que todos compartilhem uma ligação em comum no nível mais alto da hierarquia.

Logicamente, pode-se deduzir que a semelhança entre os grupos tende a diminuir ao longo dos níveis à medida que a estrutura é formada, ou seja, a similaridade é mais alta nas folhas e decai em direção à raiz. Das duas versões de agrupamento hierárquico, a abordagem *bottom-up* é a mais amplamente usada (Grigoras; Cartina, 2011, p. 208), haja vista a maior simplicidade dos métodos aglomerativos sobre os métodos divisivos, com custos computacionais iguais a $\mathcal{O}(n^2)$ e $\mathcal{O}(2^n)$, respectivamente, (Zidek et al., 2016, p. 2), onde n é o número de objetos.

A formação da estrutura pela abordagem *top-down* é feita de maneira inversa à aglomerativa. Logo, nos métodos divisivos, há apenas um grande cluster no inicio da construção da árvore, formado por todos os objetos do conjunto de dados. A cada passo, os clusters são subdivididos em clusters menores até que, no final, cada objeto represente um cluster individualmente. Uma vez que a gênese do processo é a raiz, a similaridade entre os objetos aumenta no decorrer dos níveis em direção às folhas, conforme a estrutura é criada.

Embora a abordagem *top-down*, como já mencionado, seja mais complexa que a abordagem *bottom-up*, ainda assim, apresentam algumas vantagens sobre os métodos

aglomerativos. Uma delas é o acesso a todos os dados, facilitando a visão global da estrutura que, consequentemente, pode levar a uma melhor solução (Namratha; Prajwala, 2012, p. 28–29). Além disso, métodos divisivos podem ser mais eficientes em casos onde não há formação completa da árvore (Madhulatha, 2012, p. 720), já que, opcionalmente, pode-se especificar o número de clusters como critério de parada dos algoritmos hierárquicos (Han; Kamber; Pei, 2012, p. 459).

De modo geral, os pontos fortes dos métodos hierárquicos são a facilidade de percepção do agrupamento através do dendograma e sua versatilidade em atender qualquer tipo de similaridade. Além disso, o processo de formação dos grupos é bastante simples, já que basta calcular as distâncias entre os padrões em vez de computar os centroides de cada cluster. Entretanto, não são escaláveis com o número de objetos e possuem dificuldades em lidar com clusters de tamanhos variados (Islam; Ahmed, 2013, p. 178–179). Como desvantagem, vale também mencionar a incapacidade desses métodos na diferenciação de possíveis sobreposições de grupos (Kumar; Chhabra; Kumar, 2016, p. 296).

Grande parte das heurísticas de agrupamento hierárquico não se baseiam na ideia de otimização. Em vez disso, visam obter aproximações através das iterações, as quais norteiam as ramificações da árvore para a formação dos grupos (Kantardzic, 2011, p. 260). Por outro lado, os métodos particionais, usualmente baseados em distância (Han; Kamber; Pei, 2012, p. 448), utilizam algum critério de melhoria da solução na busca pela melhor separação dos dados.

A otimização pode ser feita segundo um critério local, i.e., considera apenas um subconjunto de objetos na criação dos clusters; ou global, que leva em conta todos os elementos da base na divisão do conjunto. O erro quadrático, considerado como um critério global, é a estratégia mais comumente usada no agrupamento particional (Kantardzic, 2011, p. 263–264), servindo de referência para encontrar a partição por meio de sua minimização.

Logo, conforme Faceli et al. (2011, p. 212), a otimização do critério é feita com base em um processo iterativo. Em um primeiro momento, define-se a partição inicial. Posteriormente, os objetos migram de forma sucessiva entre os cluster até que não haja mudanças significativas no valor do erro quadrático. Esse erro, também chamado de

variação intracluster, representa a soma da variação interna dos grupos, ilustrado pela equação 2.10 a seguir , considerando k clusters (Han; Kamber; Pei, 2012, p. 451–452).

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \tag{2.10}$$

Onde p é um objeto dentro do cluster C_i, c_i é o centroide do cluster C_i e $dist(p, c_i)$ é a distância euclidiana entre um objeto p e o centroide c_i. Através da minimização da SSE (*Sum of Squared Error*), as heurísticas particionais procuram obter uma partição de modo que os grupos contenham mínima coesão e máxima separação possíveis.

O erro quadrático médio (MSE - *Mean Squared Error*) também é uma função de otimização global bastante comum, utilizada como alternativa à SSE. Seu valor é calculado através da divisão da SSE pelo número total de objetos da base (n), ou seja, $MSE = \frac{SSE}{n}$. Potanto, a minimização da equação 2.11 surte os mesmos efeitos da SSE no que diz respeito ao particionamento do conjunto de dados (Malinen, 2015, p. 1, 3).

$$MSE = \frac{1}{n} \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \tag{2.11}$$

Ainda que os métodos particionais, foco deste trabalho, necessitem do fornecimento a priori do número de clusters pelo usuário (Han; Kamber; Pei, 2012, p. 451), estes são considerados computacionalmente mais eficientes do que as técnicas hierárquicas (Malki et al., 2016, p. 74). Além disso, levam maior vantagem ao lidarem com grandes massas de dados pelo fato da alta complexidade na construção do dendograma (Kantardzic, 2011, p. 263).

Os métodos de agrupamento em geral podem ainda ser divididos em exclusivos e difusos (*fuzzy*) (Han; Kamber; Pei, 2012, p. 448). A abordagem exclusiva ou *crisp* faz uso da teoria dos conjuntos clássica na designação dos objetos aos clusters. Sendo assim, cada objeto é atribuído unicamente a um grupo (Han; Kamber; Pei, 2012, p. 587). Em definição formal, seja X uma representação do universo de dados e {0,1} um conjunto de dois estados, onde 0 indica o não pertencimento de um elemento ou objeto de X a um dado conjunto ou grupo; e 1, caso contrário. A função característica que define

uma partição *crisp* é dada por (Peres et al., 2012, p. 125):

$$f_{crisp} = \mathrm{X} \longrightarrow \{0, 1\}$$

A abordagem difusa ou *fuzzy*, por sua vez, como o próprio nome induz, baseia-se no princípio da teoria dos conjuntos *fuzzy* (TCF), que nada mais é do que uma extensão da teoria dos conjuntos clássica. Por meio da TCF, permite-se que os elementos pertençam, ao mesmo tempo, a mais de um grupo, eliminando a noção de unicidade categórica, presente na teoria clássica. No agrupamento difuso, a associação objeto-cluster é feita com base no grau de pertinência do objeto a cada um dos grupos, o qual varia entre 0 e 1 (Tan; Steinbach; Kumar, 2009, p. 588, 688). A função característica, também chamada de função de pertinência no agrupamento difuso, que descreve um conjunto *fuzzy* é definida como (Peres et al., 2012, p. 125–126):

$$f_{fuzzy} = \mathrm{X} \longrightarrow [0, 1]$$

Onde X é o conjunto de dados e [0, 1] um intervalo de estados cujo limite inferior indica a não pertinência de um objeto a um dado cluster. Já o limite superior expressa sua pertinência total a um grupo. Em adição, uma pseudopartição *fuzzy* deve obedecer às seguintes restrições (Tan; Steinbach; Kumar, 2009, p. 689):

1. A soma de todos os graus de pertinência de um objeto x_i a cada um dos cluster C_j, dado por w_{ij}, é igual a 1.

$$\sum_{j=1}^{k} w_{ij} = 1 \tag{2.12}$$

2. Considerando um conjunto de dados de tamanho n, cada cluster C_j possui ao menos um objeto com grau de pertinência diferente de 0, mas não todos de valores iguais a 1.

$$0 < \sum_{i=1}^{n} w_{ij} < n \tag{2.13}$$

Na prática, o uso da abordagem *fuzzy* mostra-se um recurso alternativo válido devido a sua capacidade em lidar com a incerteza nos dados, presente em grande parte

dos casos. Às vezes, um objeto pode ser melhor representado por um conjunto de clusters, dado que o mesmo pode estar suficientemente próximo a todos eles e, logo, não seria justo atribuí-lo somente a um dos grupos, pois não haveria diferença numérica significativa no valor das distâncias entre o objeto e cada um dos clusters adjacentes. Nessas situações, portanto, há uma flexibilização relativa ao princípio de boa separação dos grupos (Tan; Steinbach; Kumar, 2009, p. 588, 688). Consequentemente, quando utilizadas em conjunto com as abordagens clássicas da análise de dados, podem permitir uma compreensão mais adequada acerca dos padrões.

Na seção 2.2.1 seguinte, serão apresentadas algumas das medidas de proximidade que podem ser usadas pelos algoritmos no cômputo da similaridade ou diferença entre os objetos. Posteriormente serão descritas duas heurísticas amplamente adotadas no processo de agrupamento particional de dados (Malki et al., 2016, p. 74).

2.2.1 Medidas de proximidade

No agrupamento de dados, é necessário uma forma de verificar a similaridade ou dissimilaridade entre os dados para avaliar e mensurar o quão semelhantes ou diferentes são os objetos em relação aos demais. Tal fato constitui um dos pontos-chave do processo de divisão do conjunto de dados, já que, por definição, um cluster é a reunião de objetos com características semelhantes. Adicionalmente, espera-se que os objetos de um mesmo cluster sejam distintos dos objetos de outros grupos (Linden, 2009, p. 18; Tan; Steinbach; Kumar, 2009, p. 585; Han; Kamber; Pei, 2012, p. 65–66).

Em terminologia mais ampla, utiliza-se proximidade como expressão para se referir tanto à similaridade quanto à dissimilaridade, em virtude da correlação existente entre ambas. Em essência, as duas possuem a mesma finalidade, apesar de que, conceitualmente, uma é o inverso da outra (Han; Kamber; Pei, 2012, p. 67). Na similaridade entre dois objetos, quanto mais alto for o seu valor, mais semelhante o são. Analogamente, quanto maior a dissimilaridade, maior a diferença entre eles.

Considerando que um conjunto de dados pode ser entendido como uma coleção de pontos em um espaço f-dimensional, onde f denota o número de atributos, a ideia de proximidade entre dois pontos x e y quaisquer pode ser definida como a distância entre os

mesmos (Goldschmidt; Passos; Bezerra, 2015, p. 71), embora haja outros meios de computar a proximidade entre elementos, a depender do tipo do atributo. A seguir, são apresentadas três funções de distância (Goldschmidt; Passos; Bezerra, 2015, p. 72) que podem ser utilizadas como medidas de proximidade voltadas a atributos numéricos.

- Distância Euclidiana

  É a estratégia de uso mais comum (Grabusts, 2011, p. 70) e mais bem conhecida (Kantardzic, 2011, p. 254; Han; Kamber; Pei, 2012, p. 72) dentre as funções de distância. A distância Euclidiana entre dois pontos $x = (x_1, ..., x_f)$ e $y = (y_1, ..., y_f)$, dada pela equação 2.14, corresponde ao comprimento do segmento linear que os une.

$$\sqrt{\sum_{i=1}^{f}(x_i - y_i)^2} \tag{2.14}$$

  Essa medida contribui para a formação de clusters de formato globular ou hiperesférico (Pereira; Mello, 2013, p. 34; Khan; Zomaya, 2015, p. 1114). Vale destacar também que, apesar de ser quase uma função padrão de distância, não quer dizer que seja a melhor (Kantardzic, 2011, p. 230). Logo, dependendo da natureza dos dados e do domínio no qual estão inseridos, outras funções podem ser mais adequadas.

- Distância Manhattan

  É também conhecida como *city block*. Isso se deve ao fato de que a distância Manhattan entre dois pontos x e y quaisquer é medida em termos de blocos (Han; Kamber; Pei, 2012, p. 72). Portanto, representa a distância de x a y medida em ângulos retos (Goldschmidt; Passos; Bezerra, 2015, p. 72), definida como:

$$\sum_{i=1}^{f}|x_i - y_i| \tag{2.15}$$

Baseado em suas características, normalmente a distância Manhattan tende à formação de grupos hiper-retangulares (Khan, 2011, p. 1114). Além disso, em comparação com a função Euclideana, nota-se que o cálculo da distância Manhattan é menos complexo devido ao menor número de operações matemáticas envolvidas.

- Distância Minkowski

  A distância Minkowski é uma generalização de ambas as funções anteriores. Através da equação 2.16, portanto, é fácil perceber que as distâncias Manhattan e Euclideana são obtidas ao fazer $p = 1$ e $p = 2$, respectivamente (Han; Kamber; Pei, 2012, p. 73).

$$\left(\sum_{i=1}^{f} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{2.16}$$

  Onde p é um número real tal que $p \geq 1$. Nessa função, há tendência de domino por parte dos atributos com maiores escalas sobre os demais. Uma alternativa a esse problema é a normalização do conjunto (Kiwanuka; Wilkinson, 2015, p. 132). Por sua grande capacidade de generalização ao abranger muitas das funções de distância utilizadas na literatura, uma de suas vantagens é permitir a adequação da função às necessidades da aplicação com um simples ajuste do parâmetro p. Consequentemente, isso permite diversificar as suposições a respeito do formato dos clusters (Groenen; Kaymak; Rosmalen, 2007, p. 54–55; Gueorguieva; Valova; Georgiev, 2017, p. 229).

Formalmente, dado um conjunto X, uma medida de distância entre dois pontos x e y é uma função: $\mathrm{X} \times \mathrm{X} \rightarrow \mathbb{R}^+$, denotada por $d(x, y)$, que atende às seguintes propriedades (Linden, 2009, p. 20; Tan; Steinbach; Kumar, 2009, p. 84–85; Han; Kamber; Pei, 2012, p. 72–73):

1. Não negatividade: $d(x, y) \geq 0$. A distância é positivamente definida para todo $x, y \in \mathrm{X}$

2. Identidade: $d(x, y) = 0 \iff x = y$. A distância de um ponto a ele mesmo é sempre nula.

3. Simetria: $d(x, y) = d(y, x)$. A distância entre dois pontos x e y é invariante, ou seja, tem o mesmo valor entre eles independentemente do ponto de origem.

4. Desigualdade triangular: $d(x, y) \leq d(x, z) + d(z, y)$. A distância linear entre dois pontos é sempre menor ou igual a qualquer outro caminho que ligue x a y. Em outras palavras, a menor distânca entre dois pontos é uma reta.

Medidas de proximidade que satisfazem essas quatro propriedades são denominadas métricas. A distância Minkowski, por exemplo, é uma métrica apenas para o caso em que $p \geq 1$ (Sladoje; Lindblad; Nyström, 2011, p. 128, 134). Quando $p < 1$, a distância entre o ponto $x = (0, 0)$ e $y = (1, 1)$ é $2^{1/p} > 2$. Considerando um terceiro ponto $z = (0, 1)$, a distância de x a y passando por z é igual a 2, já que ambos distam 1 unidade de z, violando o axioma da desigualdade triangular, pois $d(x, y) > d(x, z) + d(z, y)$ (Serra; Greco; Tagliaferri, 2015, p. 3). As distâncias Manhattan e Euclideana também são consideradas métricas (Han; Kamber; Pei, 2012, p. 72–73), uma vez que representam dois casos especiais da distância Minkowski, onde p vale 1 e 2, respectivamente.

2.2.2 Algoritmo K-Means

O KM, proposto por MacQueen (1967), também conhecido como *hard* C-Means, é um dos métodos particionais mais conhecidos e "amplamente usados" (Tan; Steinbach; Kumar, 2009, p. 593) no processo de clusterização por seu grande destaque na apresentação de bons resultados a dados numéricos (Bathla; Aggarwal; Rani, 2018, p. 1712), além de possuir um grande número de variações. O intento desse algoritmo é promover a divisão de um conjunto de dados de tamanho n em k grupos mutuamente exclusivos através da minimização da distância dos pontos a um conjunto de centroides denotado por $C = \{c_1, c_2, ..., c_k\}$.

"A distância entre um" objeto x_i "e um conjunto de clusters" P $= \{p_1, ..., p_k\}$, "dada por" $d(x_i, P)$, "é definida como sendo a distância do objeto ao centroide mais

próximo dele. A função a ser minimizada é dada por" (Linden, 2009, p. 24):

$$d(X, C) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, C)^2 \tag{2.17}$$

Linden (2008, p. 152) fornece uma descrição formal para o KM:

Algoritmo 1: K-Means

início
- Defina o número k de grupos a ser usado;
- Designe cada objeto para um grupo aleatoriamente;
- **repita**
  - Calcule o centroide de cada grupo;
  - Reorganize os objetos, designando-os para o grupo cujo centroide lhe for mais próximo;
- **até** *que nenhum elemento mude de grupo*;

fim

Esse algoritmo se destaca por sua simplicidade de implementação e rapidez, pois atinge a convergência com um baixo número de iterações (Daister, 2007, p. 20; Perim, 2008, p. 23; Linden, 2009, p. 24), e por sua eficiência, com complexidade temporal $\mathcal{O} = iknf$ (Tan; Steinbach; Kumar, 2009, p. 603), onde i = número de iterações, k = número de clusters, n = número de objetos e f = número de atributos.

Em contrapartida, Kainulainen (2002, p. 1) diz que a simplicidade do KM também representa uma desvantagem, visto que os resultados podem sofrer grandes variações dependendo da escolha dos parâmetros e centroides iniciais e da métrica de distância usada, havendo possibilidade de se criar grupos que não fazem sentido.

Portanto, uma má inicialização do parâmetro k e dos centroides iniciais compromete diretamente a qualidade de serapação dos grupos, pois um conjunto reduzido de grupos pode unir clusters que não compartilham das mesmas propriedades, ao passo que um alto valor de k pode promover a divisão de um grupo natural (Linden, 2009, p. 24). Logo, conforme Tesser et al. (2013, p. 26), pode-se dizer que quando se obtém resultados muito diferentes pela mudança das condições iniciais do algoritmo ou quando o mesmo converge lentamente são indícios de uma má escolha do parâmetro k ou que o conjunto de dados não possui uma estrutura adequada à aplicação do KM.

Outro ponto negativo a se mencionar é sua sensibilidade a *outliers* que po-

dem alterar substanciamente a "posição do centroide" e, consequentemente, prejudicar a formação do cluster (Pinheiro, 2006, p. 12).

2.2.3 Algoritmo Fuzzy C-Means

O algoritmo FCM, desenvolvido por Dunn (1973) e aprimorado por Bezdek (1981), consiste em uma adaptação *fuzzy* do KM para lidar com incertezas, devido às próprias características dos dados; e imprecisões, usualmente observadas na prática (Peres et al., 2012, p. 120), sejam por erros instrumentais ou eventos atípicos que ocasionam o surgimento de anomalias (Bezdek et al., 2005, xii). Esse método faz uso da abordagem *soft* para fragmentar o conjunto de dados em grupos *fuzzy*, onde um elemento do conjunto é associado a mais de um grupo através de um grau distinto de pertinência (Yonamine et al., 2002, p. 3–4), sendo que a escolha do grupo no qual um determinado objeto é atribuído é dada em função do maior grau de pertinência (Guiera et al., 2005, n.p.).

O particionamento do universo de amostras, tal como no KM, é feito através da minimização de uma função objetivo. Segundo Rocha (2003, p. 46), pontos de dados mais próximos a seus respectivos centroides são associados a maiores graus de pertinência e vice-versa. Essa função, chamada de índice de desempenho, é definida por ele em termos dos centroides dos clusters:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \, ||x_i - c_j||^2 \tag{2.18}$$

Onde m é um número real no intervalo (1,∞), u_{ij} é o grau de pertinência do objeto x_i em relação ao cluster j, x_i é o i-ésimo objeto do conjunto de dados, $c_1, ..., c_k$ são os centroides dos clusters e $||x_i - c_j||$ é qualquer métrica de distância usada no cálculo da similaridade entre os pontos e os centroides e, a exemplo, fora considerada a norma euclidiana nas equações 2.18 e 2.20.

O parâmetro m, denominado índice de fuzzificação (Yonamine et al., 2002, p. 7) ou índice de peso *fuzzy* por Rocha (2003, p. 46), ou ainda, coeficiente *fuzzy* segundo Vale (2005, p. 53), é um número real que controla a influência dos graus de pertinência e, de acordo com Rocha (2003, p. 46), expressa a "nebulosidade da solução", definindo a

distância permitida entre os pontos e os centroides de forma que, quando m se aproxima de 1, o algoritmo tem comportamento semelhante ao método KM. Entretanto, à medida que $m \to \infty$, os grupos tornam-se mais *fuzzy*, uma vez que mais elementos do conjunto passam a ser considerados nos grupos da pseudopartição.

Bezdek (1981) apontou, a partir de testes empíricos, que bons valores de m estariam no intervalo $[1.10, 5]$ e mostrou que, para $m = 2$, obtinha-se resultados satisfatórios com a aplicação do método. Mais tarde, Pal e Bezdek (1995) reduziram o intervalo de valores apropriados para m ao intervalo [1.5, 2.5] e que 2 poderia ser considerado como um valor padrão desse parâmetro. Nuevo, Catania e Palesi (2006, n.p.) ainda destacam que, geralmente, considera-se o intervalo [1.5, 2] para m, embora melhores resultados possam ser alcançados com valores maiores ou menores dependendo da aplicação.

A atualização dos centroides c_j e da pertinência u_{ij} para cada iteração t são dadas similarmente a Guiera et al. (2005, n.p.) e Coelho et al. (2012, p. 101):

$$c_j^{(t)} = \frac{\sum_{i=1}^{n} (u_{ij}^{(t)})^m \cdot x_i}{\sum_{i=1}^{n} (u_{ij}^{(t)})^m} \tag{2.19}$$

$$u_{ij}^{(t+1)} = \frac{1}{\sum_{p=1}^{k} \left(\frac{||x_i - c_j^{(t)}||^2}{||x_i - c_p^{(t)}||^2} \right)^{\frac{2}{m-1}}} \tag{2.20}$$

O FCM assume como parâmetros de entrada: s = o conjunto de dados, k = número de clusters, m = coeficiente *fuzzy*, e ϵ = valor de erro máximo permitido. Guiera et al. (2005, p. 101) apresentam um passo a passo do funcionamento desse algoritmo, apresentado a seguir.

O critério de parada é satisfeito quando $\delta \leq \epsilon$ ou um número máximo de iterações é atingido. Uma das vantagens do FCM em relação ao KM é que ele consegue representar as incertezas nos dados, promovendo uma melhor descrição (Ferreira, 2012, p. 87). Além disso, esse método favorece um entendimento mais claro acerca dos padrões identificados e interações humanas e pode fornecer resultados aproximados mais rapidamente (Ghosh; Dubey, 2013, p. 38).

Todavia, semelhante ao problema do KM, esse método apresenta dificuldades ao

Algoritmo 2: Fuzzy C-Means

início

1. Defina um valor para k, m e ϵ;
2. Gere aleatoriamente a partição *fuzzy* U^0, obedecendo às seguintes restrições: $M_{fnk} = \{U \in U_{nk} : u_{ij} \in [0,1],\ \sum_{j=1}^{k} u_{ij} = 1,\ 0 < \sum_{i=1}^{n} u_{ij} < n\}$ onde U_{nk} representa um grupo de matrizes reais de dimensões $n \times k$ em que n = tamanho do conjunto de dados;
3. Atribua ao contador de iterações t o valor 0;
4. . Atribua $J_m^{(t)} = 0$;
5. Calcule os centroides $c_j^{(t)}$ segundo a equação 2.19;
6. Calcule a função objetivo $J_m^{(t+1)}$ segundo a equação 2.18;
7. Calcule os graus de pertinência $u_{ij}^{(t+1)}$ segundo a equação 2.20;
8. Calcule $\delta = J^{(t+1)} - J^{(t)}$;
9. Incremente o contador de iterações t;
10. Se critério de parada = falso, então retorne ao passo 5. Senão, finalize o algoritmo.

fim

lidar com *outliers* (Peres et al., 2012, p. 156) e na definição das partições iniciais. Outra desvantagem pertinente reside em sua complexidade computacional, com complexidade temporal na ordem de $\mathcal{O} = ik^2nf$ (Rao; Vidyavathi, 2010, p. 150), onde i = número de iterações, k = número de clusters, n = número de objetos e f = número de atributos. Em vista disso, observa-se que esse algoritmo é "computacionalmente mais caro que o" KM (Silva, 2015, p. 38) e "não é considerado escalável para grandes conjuntos de dados" (Cavalcante Júnior, 2006, p. 28).

Para Peres et al. (2012, p. 131), a escolha de k e m influencia diretamente o resultado a ser obtido ao final do processo e, devido a sua susceptibilidade a mínimos locais, recomenda-se realizar diversas execuções com diferentes valores para esses parâmetros a fim de se obter soluções mais consistentes.

2.3 Avaliação e interpretação dos resultados

Após a etapa de clusterização, é importante avaliar a qualidade dos resultados obtidos, dado que as heurísticas de agrupamento fornecem uma solução independentemente da existência de uma real estrutura de clustering. Para isso, é necessário utilizar metodologias

avaliativas que permitam discernir uma boa solução de um mau agrupamento, além de verificar se o resultado alcançado é mera casualidade ou consequência do algoritmo de agrupamento ou se de fato reflete a natureza dos objetos (Tan; Steinbach; Kumar, 2009, p. 634; Kantardzic, 2011, p. 275).

Portanto, a validação de agrupamento tem como foco a análise da viabilidade do processo de agrupamento aplicado sobre um conjunto de dados e da qualidade das soluções produzidas pelas heurísticas. As principais questões ligadas a essa tarefa incluem a avaliação da tendência de agrupamento, a determinação do número de clusters e a medição da qualidade do agrupamento (Han; Kamber; Pei, 2012, p. 484).

2.3.1 Avaliação da tendência de agrupamento

Nessa tarefa, é avaliada a possibilidade de uma estrutura aleatória estar presente nos dados, o que poderia gerar resultados enganosos e, consequentemente, levar a falsas conclusões a respeito de sua natureza (Han; Kamber; Pei, 2012, p.). Intuitivamente, poderia-se utilizar o próprio algoritmo de clusterização como forma de se determinar se um conjunto de dados contém grupos. Desse modo, a tendência de agrupamento seria verificada caso alguns dos grupos formados fossem de boa qualidade, por exemplo. No entanto, a geometria pesquisada pelo algoritmo poderia divergir do formato real dos possíveis grupos existentes nos dados. Uma forma de lidar com isso seria comparar os resultados provenientes da aplicação de diversas heurísticas. A uniformização da baixa qualidade dos grupos pode ser um indicativo da ausência de uma estrutura de agrupamento (Tan; Steinbach; Kumar, 2009, p. 651).

Todavia, dependendo da quantidade de dados e dos algoritmos de clusterização, tal metodologia pode exigir um alto custo computacional, como também demandar tempo até que se possa julgar a aleatoriedade da estrutura. Existem, portanto, outras estratégias mais diretas que dispensam o uso do processo de agrupamento.

A abordagem alternativa mais comumente usada na verificação da tendência de agrupamento consiste em realizar testes estatísticos de aleatoriedade espacial sobre o conjunto, como a estatística de Hopkins (Hopkins; Skellam, 1954). Nesse método, inicialmente são obtidas duas amostras, U e W, a partir do conjunto original de dados de

tamanho n, ambas contendo p elementos, de modo que $p \ll n$ (Jain; Dubes, 1988, p. 218). A diferença entre essas subunidades está no processo de formação, considerando que U é composta por p pontos aleatoriamente gerados no espaço, enquanto que W possui p pontos reais de dados. Em seguida, computa-se as distâncias entre os p elementos das amostras e seus respectivos vizinhos mais próximos no conjunto original D. A equação que define a estatística de Hopkins, denotada por H, é dada por (Tan; Steinbach; Kumar, 2009, p. 651–652; Han; Kamber; Pei, 2012, p. 485):

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i} \tag{2.21}$$

Onde u_i é a distância do i-ésimo ponto sinteticamente gerado em U a seu vizinho mais próximo em D. De modo análogo, w_i é a distância entre $i \in$ W e seu vizinho mais próximo em D. Em relação ao tamanho de p, se um número muito reduzido de objetos for escolhido, as distâncias entre os elementos de U e W e seus vizinhos mais próximos em D não são capazes de retratar toda a distribuição de distâncias, gerando assim, suposições inválidas (Lawson; Jurs, 1990, p. 37). Portanto, geralmente toma-se 10% dos dados para as subunidades ($p = 0.1n$) (Theodoridis; Koutroumbas, 2003, p. 629) com sucessivas amostragens visando a representatividade do conjunto (Lawson; Jurs, 1990, p. 41; Banerjee; Davé, 2004, p. 151).

Os limites inferior e superior de H são, respectivamente, 0 e 1 (Hopkins; Skellam, 1954, p. 227; Aggarwal, 2015, p. 158). A hipótese nula é de que os dados são uniformemente distribuídos, i.e., não há estrutura de agrupamento. Nesse caso, o valor de H fica em torno de 0.5 em decorrência da similaridade entre u_i e w_i. Por outro lado, na hipótese alternativa, u_i é consideravelmente menor que w_i. Logo, quando D contém grupos estatisticamente significativos, o valor de H tende a ser próximo de 0 (Han; Kamber; Pei, 2012, p. 485–486; Kassambara, 2017, p. 124; Krishna; Babu; Kumar, 2018, p. 304).

Adicionalmente, um valor próximo de 1 indica que os dados são regularmente distribuídos no espaço (Tan; Steinbach; Kumar, 2009, p. 652). Outra possível variação da fórmula para o cômputo dessa estatística, como vista em Aggarwal (2015, p.

157–158), consiste na substituição do termo do numerador pelo somatório das distâncias dos i-ésimos pontos sinteticamente gerados em U a seus respectivos vizinhos mais próximo em D, i.e., $\sum_{i=1}^{p} u_i$.

$$\overline{H} = 1 - H = \frac{\sum_{i=1}^{p} u_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i} \tag{2.22}$$

Nesse caso, a interpretação de $\overline{H}$ é feita de maneira inversa à definida previamente para H. Portanto, $\overline{H} \to 1$ quando D contém dados significativamente agrupáveis. Tendo como referência a equação 2.21, um estudo realizado por Banerjee; Davé (2004, p. 150) mostrou que, frequentemente, conjuntos aleatórios, agrupáveis e com dados regularmente distribuídos no espaço possuem, respectivamente, valores de H próximos de 0.5, entre 0.01 e 0.3 e entre 0.7 e 0.99. Por ser um método simples (Lavine; Nuguru; Nikhil, 2011, p. 119), a estatística de Hopkins é frequentemente usada como teste para avaliar a tendência de agrupamento (Aggarwal, 2015, p. 157).

2.3.2 Determinação do número ideal de clusters

Identificar o número ideal de grupos em um conjunto de dados é um dos problemas mais difíceis da análise de agrupamento (Kaufman; Rousseeuw, 1990, p. 87; Morozkov et al., 2012, p. 2001). Determiná-lo é fundamental não somente porque algumas heurísticas de agrupamento, tais como o KM e FCM, exigem sua especificação como parâmetro, mas principalmente por controlar o processo de divisão dos dados de modo adequado. A descoberta do número correto de grupos deve levar em conta o formato da distribuição e a escala do conjunto. Além disso, muitas vezes, a complexidade do problema é associada à subjetividade da tarefa, já que a escolha também depende da decisão do usuário (Han; Kamber; Pei, 2012, p. 486). Segundo Hruschka e Ebecken (2003, p. 18), grande parte das abordagens presentes na literatura estimam o número ideal de grupos conforme critérios baseados no agrupamento do conjunto estudado, i.e., que o avaliam em termos da coesão e/ou separação.

O *elbow method* (Thorndike, 1953) é uma estratégia comumente utilizada como critério de validação. Baseia-se na suposição de que o aumento do número de grupos (k)

provoca redução da SSE, descrita anteriormente pela equação 2.10. Logo, a ideia desse critério consiste basicamente no cômputo do valor da SSE para cada valor distinto de k. A abordagem sugere que o valor mais adequado de grupos é aquele em que, a partir de um dado k, o aumento do número de clusters não gera ganho significativo de informação. Isso é facilmente identificado na representação do número de grupos em relação à SSE, por um ponto de inflexão, proeminência ou "joelho"no gráfico (Madhulatha, 2012, p. 723). Em outras palavras, esse método indica a escolha do k no qual a SSE decai abruptamente, produzindo o chamado "efeito elbow"(Duvvada; Naidu; Sri, 2017, p. 1358).

O coeficiente de silhueta (Rousseeuw, 1987) é também um critério bastante conhecido que pode ser empregado nessa tarefa. Baseia-se nos conceitos de coesão (i.e., medida referente à semelhança interna dos objetos em um mesmo grupo) e separação (i.e., medida referente à diferença externa dos objetos em grupos distintos) (Tan; Steinbach; Kumar, 2009, p. 644). Como normalmente ambas as medidas são avaliadas em termos de distância entre objetos, nesse caso, é desejável se ter grupos com baixa coesão e alta separação, o que indica que os clusters são heterogêneos e que apresentam alta similaridade entre os elementos de um mesmo grupo.

Para cada elemento i, a silhueta (sil_i) é dada pela diferença entre a menor distância média intercluster e a distância média intracluster, dividida pela maior das duas distâncias (Norušis, 2012, p. 397), conforme a equação 2.23.

$$sil_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{2.23}$$

Na silhueta original, a_i corresponde à distância média de i a todos os elementos do mesmo grupo, e b_i é a distância média mínima de i a todos os pontos dos demais clusters, o que resulta em uma complexidade quadrática, $\mathcal{O}(mn^2)$, onde n e m são, respectivamente, o número de objetos e o número de atributos. Por outro lado, a silhueta simplificada, introduzida por Hruschka, Campello e Castro (2006, p. 1912), geralmente possui complexidade linear, $\mathcal{O}(mn)$, já que, diferentemente da abordagem original, baseia-se na distância entre objetos e os centroides dos grupos. No entanto, se $k \approx n$, i.e., se o número de grupos for próximo do número de objetos, esse método adquire a mesma

complexidade de sua versão original (Vendramin; Campello; Hruschka, 2010, p. 214–215). O coeficiente de silhueta para um agrupamento contendo k clusters, detonado por CS, é dado pela média do valor de sil_i sobre todos os n objetos do conjunto, ilustrado pela equação a seguir (Wang et al., 2017, p. 295).

$$CS = \frac{1}{n} \sum_{i=1}^{n} sil_i \tag{2.24}$$

Note que o valor da silhueta é definido no intervalo $[-1, 1]$, já que o denominador da equação 2.23 é apenas um termo de normalização. Quanto mais próximo de um 1, menos coeso e mais bem separado é o agrupamento. Entretanto, um valor negativo para a silhueta (i.e., $b_i < a_i$) representa um caso indesejável, pois significa que um dado objeto está mais perto de elementos de outros grupos do que de pontos em seu próprio cluster (Tan; Steinbach; Kumar, 2009, p. 644; Han; Kamber; Pei, 2012, p. 490; Aggarwal, 2015, p. 19; Wang et al., 2017, p. 294). Desse modo, ao se plotar o número de grupos versus o coeficiente de silhueta, o k ideal é indicado pelo ponto máximo do gráfico (Kaufman; Rousseeuw, 1990, p. 86–87).

Pela regra geral, um valor de CS entre 0.70 e 1 indica um forte agrupamento; entre 0.50 e 0.70 é considerado um resultado aceitável; entre 0.25 e 0.5 revela a obtenção de uma fraca estrutura de clustering, a qual pode ser artificial e, logo, nesse caso, é conveniente utilizar outras abordagens na determinação do número ideal de grupos. Por outro lado, quando o valor de CS é menor ou igual a 0.25, significa que nenhuma estrutura substancial foi identificada nos dados (Kaufman; Rousseeuw, 1990, p. 88; Deng, 2016, p. 4235).

Dentre as estratégias propostas para estimar o número de grupos, vale também destacar o método *gap statistic* (Tibshirani; Walther; Hastie, 2001). Essa técnica compara a variação intracluster (e.g., SSE) para diferentes valores de k com seus valores esperados sobre uma distribuição de referência para a hipótese nula (Kassambara, 2017, p. 130), i.e., uma distribuição sem estrutura de agrupamento evidente (Krishna; Babu; Kumar, 2018, p. 305). Seja SSE_i a variação intracluster para cada grupo C_i de uma partição P_k com k clusters, dada pela equacão 2.25.

$$SSE_i = \sum_{p \in C_i} dist(p, c_i)^2 \tag{2.25}$$

A SSE normalizada, detonatada por W_k, representa uma típica medida de coesão para P_k (Yan; Ye, 2007, p. 1031), definida a seguir.

$$W_k = \sum_{i=1}^{k} \frac{1}{2n_i} SSE_i \tag{2.26}$$

Onde n_i é o numero de objetos no cluster C_i. A partir da equação anterior, pode-se dizer que quanto menor o valor de W_k, maior a semelhança entre os objetos de um mesmo cluster, já que ele reflete a homogeneidade interna dos grupos. A ideia desse método é comparar a curva do $log(W_k)$ obtida do conjunto original de dados com a curva $E_n^*\{log(W_k)\}$, i.e., a curva esperada para uma distribuicão normal. Para cada P_k, o *gap statistic* é calculado da seguinte forma (Tibshirani; Walther; Hastie, 2001, p. 412–413; Yan; Ye, 2007, p. 1032):

$$Gap(k) = E_n^*\{log(W_k)\} - log(W_k) \tag{2.27}$$

O desvio padrão (sd_k) do $log(W_k)$ também é computado no intuito de determinar o erro padrão (s_k) da simulação, expresso como (Koch, 2014, p. 219; Vengadeswaran; Balasundaram, 2017, n.p.):

$$s_k = sd_k \sqrt{\frac{1}{1+B}} \tag{2.28}$$

B é a quantidade de conjuntos de referência gerados com uma distribuição uniforme. Fazendo $B = 500$, obtêm-se resultados bastante precisos, já que, para esse valor, o gráfico *gap*, obtido pela representacão do número de clusters pelo $Gap_n(k)$, permanece praticamente inalterado nas próximas chamadas ao método. Logo, o número ideal de grupos é dado pelo menor valor de k de modo que $Gap(k) \geq Gap(k+1) - s_{k+1}$. Em outras palavras, o k ideal é aquele no qual o *gap statitisc* está dentro de um desvio padrão em $k+1$ (Kassambara, 2017, p. 131).

As três técnicas previamente apresentadas para estimar o número ideal de gru-

pos são indicadas a partições *crisp*. Em agrupamentos *fuzzy*, por outro lado, existem estratégias mais adequadas as suas características por fazerem uso da matriz de pertinência dos objetos. Basicamentente, podem ser divididas em duas categorias (Hu; Meng; Shi, 2008, p. 193): estratégias que utilizam somente a matriz de pertinência (e.g., Partition Coefficient (PC) (Bezdek, 1973), Partition Entropy (PE) (Bezdek, 1981) e Modified Partition Coefficient (MPC) (DAVE, 1996)); e estratégias que utilizam tanto a matriz de pertinência quanto o conjunto de dados (e.g., índice de Fukuyama-Sugeno (FS) (FUKUYAMA; SUGENO, 1989 apud Pal e Bezdek, 1995, p. 374), índice de Xie-Beni (XB) (Xie; Beni, 1991) e Coeficiente de Silhueta Fuzzy (CSF) (Campello; Hruschka, 2006)).

Sejam X = $\{x_1, ..., x_n\}$ o conjunto de dados, P = $\{p_1, ..., p_k\}$ o conjunto de clusters de centroides C = $\{c_1, ..., c_k\}$, u_{ij} o grau de pertinência de x_i ao cluster p_j, e $m \in (1, \infty)$ o índice de fuzzificação. Sejam $||x_i - c_j||^2$ a distância euclidiana entre o objeto x_i e o centroide c_j, e $||c_j - \overline{c}||^2$ a distância euclidiana entre o centroide c_j e a média de seu conjunto. As expressões que definem PC, PE, MPC, FS, XB e CSF são dadas, respectivamente, a seguir.

$$PC = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^2 \tag{2.29}$$

O valor de PC está no intervalo $[\frac{1}{k}, 1]$. Assume-se que o número ideal de grupos é obtido pela maximização de PC sobre $k = \{2, ..., k_{max}\}$. Logo, um valor próximo de 1 pode ser uma boa aproximação para um agrupamento com grupos heterogêneos e de baixa coesão. Entretanto, quando PC $\rightarrow \frac{1}{k}$, a noção de grupos bem separados e com alta similaridade interna diminui (Bezdek, 1973, p. 61–65).

$$PE = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} \; log_a(u_{ij}) \tag{2.30}$$

Os valores de PE estão compreendidos no intervalo $[0, log_a(k)]$, onde $a \in (1, \infty)$ é uma base logarítmica. Ao contrário de PC, o número ideal de grupos pela função de PE é obtido pela sua minimização sobre $k = \{2, ..., k_{max}\}$. Desse modo, espera-se valores próximos de 0 em agrupamentos com clusters bem definidos (Bezdek, 1981, p.

110–112).

$$MPC = 1 - \frac{k}{k-1}(1 - PC) \tag{2.31}$$

O MPC é uma proposta de modificação de PC para eliminar a dependência em k através de uma transformação linear. Com isso, o alcance dos valores de MPC muda para o intervalo [0, 1]. Semelhante ao PC, a otimização de MPC se dá pela maximização da função 2.31 sobre $k = \{2, ..., k_{max}\}$ (DAVE, 1996, p. 617).

$$FS = \sum_{i=1}^{n}\sum_{j=1}^{k}(u_{ij})^m \ (||x_i - c_j||_A^2 - ||c_j - \overline{c}||_A^2) \tag{2.32}$$

Na expressão de FS, A representa uma matriz simétrica e positiva definida. Sejam p e q, respectivamente, o primeiro e segundo termos entre parênteses da equação 2.32. Nesse caso, p refere-se à coesão dos grupos, enquanto que q reflete a distância dos grupos. Sendo assim, quanto menor o valor da função de FS, i.e., quanto menor o valor de p e maior o valor de q, melhor a definição dos grupos. Logo, o número ideal de grupos é obtido pela minimização de FS sobre $k = \{2, ..., k_{max}\}$ (Vathy-Fogarassy; Abonyi, 2013, p. 21).

$$XB = \frac{1}{n}\ \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(u_{ij})^m \ ||c_j - x_i||^2}{\min_{i \neq j} ||c_j - x_i||^2} \tag{2.33}$$

Em termos gerais, o índice de XB é definido como a razão entre a coesão e a separação do agrupamento, i.e., $XB = \frac{coesão}{separação}$. Portanto, quanto maior o numerador e menor o denominador da equação 2.33, melhor o valor de XB, o que indica que a partição possui clusters bem definidos. Assim, o número ideal de grupos é obtido pela minimização de XB sobre $k = \{2, ..., k_{max}\}$ (Xie; Beni, 1991, p. 843).

$$CSF = \frac{\sum_{i=1}^{n}(u_i^p - u_i^q)^\alpha \ sil_i}{\sum_{i=1}^{n}(u_i^p - u_i^q)^\alpha} \tag{2.34}$$

O Coeficiente de Silhueta Fuzzy, como o próprio nome sugere, é uma adequação de CS ao domínio *fuzzy*, uma vez que a silhueta *crisp* não faz uso da matriz de pertinência

dos objetos e, portanto, CS pode não ser capaz de diferenciar clusters sobrepostos. Na equação 2.34, u_i^p e u_i^q são, respectivamente, o primeiro e segundo maiores valores da i-ésima linha da matriz de pertinência $\boldsymbol{\mu} = [u_{ij}]_{n \times k}$; e $\alpha \geq 0$ é um coeficiente de ponderacão, cujo valor padrão é 1. Nota-se que CS é um caso particular de CSF quando $\alpha = 0$. À medida que α cresce, a importância relativa dos objetos em regiões de sobreposição diminui. Consequentemente, CSF se afasta de CS. Apesar disso, o alcance dos valores de CSF também está no intervalo $[-1.1]$ e as interpretações são análogas. Logo, o número ideal de grupos é obtido pela maximização de CSF sobre $k = \{2, ..., k_{max}\}$ (Campello; Hruschka, 2006, p. 2863–2865).

2.3.3 Medição da qualidade do agrupamento

Quando a tendência de agrupamento é realmente confirmada em um conjunto, i.e., existe de fato uma estrutura não aleatória nos dados, faz sentido aplicar um ou mais métodos de clustering a fim de obtê-lo. No entanto, devido à grande variabilidade de soluções, é essencial avaliar a qualidade do agrupamento gerado para uma correta interpretação dos dados, posto que, comparativamente, certos resultados podem ser mais adequados que outros (Kantardzic, 2011, p. 275; Han; Kamber; Pei, 2012, p. 487).

Algumas medidas de qualidade determinam o quanto uma estrutura de agrupamento reflete a natureza do conjunto estudado. Outras, por suas vezes, comparam o resultado gerado com uma estrutura preestabelecida. Há também medidas que comparam diferentes partições obtidas a partir do mesmo conjunto (Han; Kamber; Pei, 2012, p. 484). Logo, as abordagens empregadas na validação de agrupamento são tradicionalmente divididas em três tipos: externa, interna e relativa (Jain; Dubes, 1988, p. 161; Tan; Steinbach; Kumar, 2009, p. 635–637; Kantardzic, 2011, p. 275).

- Validação externa

  Medidas de validação externa, também conhecidas como índices externos, utilizam o conhecimento prévio sobre os dados para avaliar a qualidade do agrupamento, ou seja, comparam-no com uma estrutura previamente definida por um especialista de domínio (Tan; Steinbach; Kumar, 2009, p. 637). Sendo assim, o objetivo

dessa abordagem, segundo Faceli et al. (2011, p. 238, 251), é verificar a validade da solução obtida em relação à estrutura esperada.

O Rand index (Rand, 1971) é um índice bem conhecido da área de validação externa, descrito formalmente a seguir. Sejam $R = \{r_1, ..., r_u\}$ e $Q = \{q_1, ..., q_v\}$ duas partições de um mesmo conjunto de dados, em que R corresponde ao agrupamento de referência utilizado na validação do agrupamento gerado (Q), comparados na Tabela 2.1. Além disso, considere as variáveis definidas na Tabela 2.2, na qual I_R e I_Q são, respectivamente, as funções indicadoras das partições R e Q (Jain; Dubes, 1988, p. 172–173).

cluster	q_1	q_2	...	q_v	
r_1	n_{11}	n_{12}	...	n_{1v}	$\sum_{j=1}^{v} n_{1j}$
r_2	n_{12}	n_{22}	...	n_{2v}	$\sum_{j=1}^{v} n_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r_u	n_{u1}	n_{u2}	...	n_{uv}	$\sum_{j=1}^{v} n_{uj}$
	$\sum_{i=1}^{u} n_{i1}$	$\sum_{i=1}^{u} n_{i2}$	...	$\sum_{i=1}^{u} n_{iv}$	$\sum_{i=1}^{u}\sum_{j=1}^{v} n_{ij}$

Tabela 2.1: Tabela de contingência para R e Q.

		I_Q	
		1	0
I_R	1	a	b
	0	c	d

Tabela 2.2: Tabela de contingência para I_R e I_Q.

- a = Número de pares de objetos que estão contidos nos mesmos grupos em ambas as partições;

- b = Número de pares de objetos nos mesmos grupos em R, mas em grupos distintos em Q;
- c = Número de pares de objetos em grupos distintos em R; mas nos mesmos grupos em Q;
- d = Número de pares de objetos pertencentes a grupos distintos tanto em R como em Q.

A formulação do Rand index (RI) é dada a seguir (Rand, 1971, p. 847; Jain; Dubes, 1988, p. 174):

$$RI = \frac{a+d}{M} \tag{2.35}$$

Onde M é o número total de pares de objetos de um conjunto de tamanho n (Jain; Dubes, 1988, p. 173). Logo:

$$M = a + b + c + d = \binom{n}{2} = \frac{n(n-1)}{2} \tag{2.36}$$

Assim como M, todos os termos da Tabela 2.2 podem ser obtidos por uma combinação simples de pares, conforme as equações 2.37, 2.38, 2.39 e 2.40 (Jain; Dubes, 1988, p. 173–174).

$$a = \sum_{i=1}^{u} \sum_{j=1}^{v} \binom{n_{i=j}}{2} \tag{2.37}$$

$$b = \sum_{i=1}^{u} \binom{\sum_{j=1}^{v} n_{ij}}{2} - \sum_{i=1}^{u} \sum_{j=1}^{v} \binom{n_{i=j}}{2} \tag{2.38}$$

$$c = \sum_{j=1}^{v} \binom{\sum_{i=1}^{u} n_{ij}}{2} - \sum_{i=1}^{u} \sum_{j=1}^{v} \binom{n_{i=j}}{2} \tag{2.39}$$

$$d = \binom{n}{2} - a - b - c \tag{2.40}$$

Onde $n_{i=j}$ é o número de objetos contidos nos mesmos grupos em ambas as partições, i.e., cada elemento da diagonal principal da matriz definida na Tabela 2.1, em que $i = j$.

O valor de RI varia de 0, indicando que R e Q não apresentam quaisquer similaridades; a 1, quando os dois agrupamentos comparados são idênticos. Sendo assim, quanto maior o seu valor, maior o nível de concordância entre as partições (Rand, 1971, p. 847; Jain; Dubes, 1988, p. 174).

Contudo, existe um histórico de severas críticas relacionadas a esse índice. Uma das principais está no fato de que o Rand index não é corrigido para o acaso, i.e., o valor esperado de RI não é 0 entre duas partições aleatórias (Hubert; Arabie, 1985, p. 193; Jain; Dubes, 1988, p. 174–175). Outro problema típico é que o RI atribui a mesma importância aos termos a e d, ou seja, o índice não faz qualquer distinção entre pares de objetos nos mesmos grupos e pares em grupos diferentes em ambas as partições, R e Q, o que é discutível em alguns casos (Saporta; Youness, 2002, p. 248).

Além disso, existe a possibilidade de um elevado número de clusters enviesar o valor do Rand index, visto que ele pode ser dominado pelo termo d, i.e., pelos objetos corretamente agrupados, mas que estão em grupos distintos, fazendo com que o RI seja incapaz de distinguir adequadamente entre uma boa partição e um mau agrupamento (Hu et al., 2013, p. 70). Felizmente, foi proposta uma variante do RI capaz de prover uma medida descritiva mais apropriada, conhecida como Adjusted Rand index (ARI) (Hubert; Arabie, 1985).

O ARI é uma versão corrigida do Rand index para lidar com o problema de aleatoriedade e que surge a partir da normalização do RI de modo que, ao se comparar duas partições selecionadas ao acaso, o valor esperado para o índice seja 0. É definido como (Hubert; Arabie, 1985, p. 197–198; Jain; Dubes, 1988, p. 175; Vendramin; Campello; Hruschka, 2010, p. 221):

$$ARI = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{2a+b+c}{2} - \frac{(a+c)(a+b)}{M}} \tag{2.41}$$

Semelhante ao Rand index, o ARI também é limitado superiormente por 1, indicando uma concordância plena entre os dois agrupamentos comparados; e assume o valor 0 quando R e Q são partições distintas (Hubert; Arabie, 1985, p. 198; Jain; Dubes, 1988, p. 175). Consequentemente, o Adjusted Rand index é considerado um dos índices de validação externa mais bem sucedidos (Fisset; Dhaenens; Laetitia, 2015, p. 301). O limite inferior, por outro lado, é -1, embora valores negativos não possuem uso substancial (Hubert; Arabie, 1985, p. 193, 198), podendo ser interpretado como um total desacordo entre R e Q (Kassambara, 2017, p. 142).

Para ilustrar o cálculo dos dois índices, considere a seguinte tabela de contingência:

cluster	q_1	q_2	q_3	
r_1	1	0	1	2
r_2	0	4	1	5
r_3	1	0	2	3
	2	4	4	10

Tabela 2.3: Exemplo de tabela de contingência.

Primeiramente, computa-se o valor de M. No exemplo, $M = \binom{10}{2} = 45$. Em seguida, deve-se encontrar o valor dos termos a, b, c e d, já que as equações 2.35 e 2.41 dependem dos mesmos. Para os dados da Tabela 2.3, $a = \binom{1}{2} + \binom{4}{2} + \binom{2}{2} = 0+6+1 = 7$; $b = \binom{2}{2}+\binom{5}{2}+\binom{3}{2}-a = 1+10+3-7 = 7$; $c = \binom{2}{2}+\binom{4}{2}+\binom{4}{2}-a = 1+6+6-7 = 6$ e $d = M-a-b-c = 45-7-7-6 = 25$. Logo, o RI entre as duas partições é de $\frac{7+25}{45} \approx 0.711$, enquanto o $ARI = \frac{7-\frac{(7+6)(7+7)}{45}}{\frac{2\times7+7+6}{2}-\frac{(7+6)(7+7)}{45}} \approx 0.313$. Como o Adjusted Rand Index é uma melhoria do Rand Index para garantir que o valor esperado no caso aleatório seja 0, é natural supor que $ARI < RI$.

- Validação interna

A validação interna torna-se útil na ausência de informações externas a respeito dos dados, o que é muito comum na prática (Aggarwal, 2015, p. 196, 198).

O propósito dessa abordagem é medir a qualidade da estrutura obtida com base apenas em informações intrínsecas ao agrupamento gerado. Em geral, os índices internos baseiam-se nas noções de coesão e separação. Logo, validam uma estrutura pela análise da boa separação dos grupos e do quão internamente relacionados são os objetos de um mesmo cluster (Tan; Steinbach; Kumar, 2009, p. 636; Han; Kamber; Pei, 2012, p. 489).

O grande problema desses critérios é que podem ser tendenciosos em relação aos algoritmos de agrupamento, dado que, muitas vezes, os índices empregados na validação interna são as próprias funções objetivo dos métodos de clustering. Sendo assim, um critério sempre favorece um algoritmo que adota uma função similar em sua otimização (Aggarwal, 2015, p. 196).

Segundo Faceli et al. (2011, p. 248–249), uma forma de lidar com isso é aplicá-los como índices relativos para estimar o número de clusters, por exemplo. As estratégias *elbow method*, coeficiente de silhueta e *gap statistic*, descritas na subseção 2.3.2, são alguns dos diversos índices internos presentes na literatura que podem ser empregados na validação de partições *crisp*, utilizando o KM, por exemplo. Os demais índices apresentados na mesma subseção, a saber, PC, PE, MPC, FS, XB e CSF, também são índices internos, porém utilizados na validação de partições *fuzzy*, com o uso do FCM, por exemplo.

- Validação relativa

  O intuito da validação relativa é comparar múltiplas estruturas geradas ou a partir da aplicação de diferentes métodos de agrupamento ou pela variação dos parâmetros de um mesmo algoritmo (Kantardzic, 2011, p. 275). Os índices relativos não são um tipo isolado de medidas de validação, mas sim um uso particular dos critérios internos ou externos (Tan; Steinbach; Kumar, 2009, p. 637).

  Faceli et al. (2011, p. 237–238, 242) destaca que tais índices podem ser utilizados para encontrar qual dentre os métodos de clustering aplicados é o mais adequado ao conjunto de dados ou para determinar o melhor valor para um ou mais parâmetros

em relação a um conjunto distinto de valores de entrada para uma mesmo algoritmo. Ainda segundo o autor, a maneira mais comum de uso desses critérios é na estimação do número ideal de grupos. Nesse caso, o melhor k é dado ou pelo valor mínimo ou máximo ou ponto de inflexão no gráfico, a depender do índice utilizado.

Por fim, posteriormente à aquisição de uma boa estrutura de agrupamento, i.e., cuja qualidade é verificada com o uso de índices de validação; é possível efetuar uma análise individual sobre seus grupos. Por meio da sumarização dos dados, para cada cluster, pode-se identificar propriedades comuns a seus objetos. A partir disso, podem ser definidos rótulos para representá-los (Goldschmidt; Passos; Bezerra, 2015, p. 96).

Conforme Faceli et al. (2011, p. 207), esse processo de descrição de grupos é conhecido como interpretação dos resultados. O autor destaca a grande importância do apoio do especialista de domínio nessa tarefa, visto que a análise crítica sob a ótica experiente é capaz de produzir acepção prática, o que, consequentemente, gera o conhecimento para a predição e tomada de decisões do mundo real.

3 Materiais e Métodos

3.1 Descrição da base de dados

A base de dados utilizada neste trabalho foi cedida pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e refere-se ao controle leiteiro de bovinos, efetuado entre 1960 e 2014, da produção total durante o período de lactação. O controle se inicia no momento do parto e se encerra quando naturalmente o animal deixa de produzir leite ou quando se aproxima do próximo parto.

Os registros nela contidos são proveniente de diversos estados brasileiros e reúnem informações de gados das raças Gir, Holandês e de sintéticos, denominados de Girolandos, formados por diferentes níveis de cruzameto entre as duas raças, expresso pelo grau de genótipo, comum e equivocadamente denominado de grau de sangue devido à falsa crença antiga de que as características genéticas eram herdadas através do sangue. A referida base contém 75597 objetos e 18 atributos, descritos a seguir.

- coda: código do animal cujo valor é único para cada registro.
- rebanho: Rebanho ao qual o animal pertence.
- uf: Estado de coleta do registro.
- municipio: Município de coleta do registro.
- nasc: Data de nascimento do animal.
- parto: Data de algum parto do animal.
- gs: Grau de genótipo cujos valores variam de 1 a 9.
- ra: Regime alimentar ao qual o animal foi submetido cujos valores estão definidos no conjunto {1, 2, 3, 4, 5, 9}.
- pl305: Produção de leite em até 305 dias de lactação, expressa em litros.
- pl305c: Produção de leite corrigida para 305 dias de lactação, expressa em litros, sendo necessária para comparações entre animais com diferentes períodos de lactação.

- porgor: Porcentagem de gordura no leite.
- porprot: Porcentagem de proteína no leite.
- porsol: Porcentagem de sólidos no leite.
- ccs: Contagem de células somáticas no leite.
- lccs: Logarítmo natural de ccs
- idade_p1: Idade ao primeiro parto do animal que se tem registro, podendo ou não ser o primeiro parto efetivo.
- idp2: Idade ao segundo parto do animal que se tem registro, podendo ou não ser o segundo parto efetivo.
- dlac: Duração da lactação, expressa em dias.

O grau de genótipo, referente ao nível de cruzamento genético entre as raças Gir (GIR) e Holandês (HOL), é interpretado da seguinte forma:

- gs=1 $\Leftrightarrow$ GIR puro
- gs=2 $\Leftrightarrow \frac{7}{8}$ GIR $\times \frac{1}{8}$ HOL
- gs=3 $\Leftrightarrow \frac{3}{4}$ GIR $\times \frac{1}{4}$ HOL
- gs=4 $\Leftrightarrow \frac{5}{8}$ GIR $\times \frac{3}{8}$ HOL
- gs=5 $\Leftrightarrow \frac{1}{2}$ GIR $\times \frac{1}{2}$ HOL
- gs=6 $\Leftrightarrow \frac{3}{8}$ GIR $\times \frac{5}{8}$ HOL
- gs=7 $\Leftrightarrow \frac{1}{4}$ GIR $\times \frac{3}{4}$ HOL
- gs=8 $\Leftrightarrow \frac{1}{8}$ GIR $\times \frac{7}{8}$ HOL
- gs=9 $\Leftrightarrow$ HOL puro

O atributo ra é uma codificação interna que diz respeito ao regime alimentar ao qual uma ou mais matrizes foram submetidas, podendo ser totalmente distinto para cada valor, bem como uma composição de regimes com algum tipo de suplementação, ou mesmo algum tipo especial de regime para competições. Em relação as ordens de parto, i.e., idade_p1 e idp2, vale mencionar que cada registro contém apenas o valor de um dos dois atributos.

3.2 Visão geral dos dados

A fim de se ter uma noção acerca da distribuição dos dados por grau de genótipo e por regime alimentar, bem como do número de ausências em cada coluna da base, visando uma melhor compreensão inicial dos mesmos, contabilizou-se esses valores, os quais foram reunidos em tabelas, apresentadas a seguir.

Atributo	**Qtd. de valores ausentes [NA]**	**% NA**	**Qtd. de 0's**	**% 0**
coda	-	-	-	-
rebanho	-	-	-	-
uf	16072	21.26%	-	-
municipio	16072	21.26%	-	-
nasc	-	-	-	-
parto	-	-	-	-
gs	-	-	-	-
ra	-	-	-	-
pl305	-	-	-	-
pl305c	-	-	-	-
porgor	-	-	46302	61.25%
porprot	-	-	65955	87.24%
porsol	-	-	69281	91.64%
ccs	-	-	62313	82.43%
lcss	-	-	70466	93.21%
idade_p1	-	-	45648	60.38%
idp2	-	-	41786	55.27%
dlac	-	-	-	-
total de objetos do dataset			75597	

Tabela 3.1: Quantidades iniciais de valores ausentes e zeros por atributo.

Pela Tabela 3.1, notou-se que há um grande número de atributos com valor 0, os quais correspondem a mais de 50% do total de registros. O alto percentual desses valores nos atributos porgor, porprot, porsol, ccs e lcc se justifica pela arbitrariedade de análise

e coleta dos mesmos. Portanto, na verdade, esses valores constituem-se de ausências, do mesmo modo que os zeros em idade_p1 e idp2. Entretanto, para esses dois últimos, o número de zeros se deve, muito provavelmente, a erros no momento da coleta, como no caso das ausências em uf e municipio.

gs	**nº de objetos**	**%**
1	57823	76.49%
2	434	0.57%
3	4425	5.85%
4	4864	6.43%
5	5152	6.82%
6	478	0.63%
7	1998	2.64%
8	359	0.47%
9	64	0.08%

Tabela 3.2: Distribuição dos dados por gs.

Conforme os dados da Tabela 3.2, foi possível observar que a grande maioria dos registros contidos na base de dados são de gados Gir. Além disso, viu-se que os gados da raça Holandês representam o menor percentual dos dados.

ra	**nº de objetos**	**%**
1	1911	2.53%
2	22614	29.91%
3	1164	1.54%
4	8018	10.61%
5	5990	7.92%
9	35900	47.49%

Tabela 3.3: Distribuição dos dados por ra.

De acordo com a distrição dos dados por regime alimentar, notou-se que, na maioria das matrizes, foram adotados os regimes 2 e 9 na alimentação.

3.3 Seleção da amostra

A amostra de dados utilizadas nas simulações, a serem descritas mais adiante, contém apenas animais da raça Gir Leiteiro. Com isso, a identifição da distinção fenotípica entre conjuntos de animais, proposta como objetivo deste trabalho, foi avaliada dentro de um mesmo grau de genótipo.

Segundo informações da **?**, on-line (Associação Brasileira de Criadores de Zebuínos), o Gir Leiteiro possui um alto valor genético tanto para a formação de outras raças, quanto por suas particularidades, apresentando grande adaptabilidade às condições ambientas e socioeconômicas do Brasil. Diante disso, analisou-se a amostra de animais contendo o seu genótipo a fim de se obter conjuntos de animais bem definidos com potencial para exploração genética e, consequentemente, permitir a seleção de animais de mais valia para aumento na produção de leite.

ra	**nº de objetos**	**%**
1	1337	2.54%
2	20084	38.17%
3	1015	1.93%
4	7271	13.82%
5	4166	7.92%
9	18747	35.63%

Tabela 3.4: Distribuição dos objetos com gs=1 por ra.

Comparando-se as Tabelas 3.3 e 3.4, notou-se que a proporção de animais por regime alimentar se manteve na amostra, exceto pelo fato de um melhor balanceamento entre os registros com regimes 2 e 9.

3.4 Derivação de atributos da base

Com base no conhecimento de especialistas de domínio, o Gir Leiteiro possui idade ao primeiro parto efetiva entre 1.5 e 4 anos. Como visto previamente na Tabela 3.1, as ordens de parto possuem um alto percentual de valores ausentes. Sendo assim, utilizou-

se a informação dos especialistas para a criação de um atributo auxiliar, denominado idade_p1_novo, formado pela diferença entre os valores dos atributos parto e nasc, i.e., idade_p1_novo = parto - nasc. Tal atributo serviu de base para a geração de dois novos atributos, os quais foram adicionados à amostra de dados, pela sua subdivisão, como mostra a Tabela 3.5.

	intervalo (dias)	**nº de objetos**	**%**
≤ idp1_novo	idade_p1_novo < 547	11	0.02%
idp1_novo	547 ≤ idade_p1_novo ≤ 1460	16432	31.23%
idp2_novo	idade_p1_novo > 1460	36177	68.75%

Tabela 3.5: Distribuição dos objetos com gs=1 por ordem de parto.

A interpretação dos valores dos dois novos atributos é dada da seguinte forma:

- idp1_novo = Idade ao primeiro parto efetiva do animal, i.e., corresponde de fato ao primeiro parto, expressa em dias.
- idp2_novo = Idade a algum outro parto que se tem registro, expressa em dias.

Essa alteração permitiu que cada registro tivesse um valor para um dos dois novos atributos criados, reduzindo assim o número de ausências da base referentes às ordens de parto. Os 11 objetos cujos valores de idade_p1_novo são inferiores a 547 dias (i.e., 1.5 anos) foram removidos da amostra. Além disso, um terceiro atributo novo, denominado prod_diaria, foi derivado da razão entre pl305 e dlac, de modo que:

$$\text{prod\_diaria} = \begin{cases} \frac{\text{pl305}}{\text{dlac}}, & \text{se dlac} \leq 305 \\ \frac{\text{pl305}}{305}, & \text{se dlac} > 305 \end{cases}$$

O significado dessa nova variável, também adicionada ao conjunto de atributos da amostra, é tido como a média diária da produção do animal, exressa em litros.

3.5 Identificação e tratamento de *outliers*

Considerando a visão geral dos dados anteriormente exposta, a interpretação de cada variável e o potencial de relevância das mesmas no agrupamento dos dados, além do

fato de que o KM e o FCM aplicam-se apenas a variáveis numéricas; foram construídos boxplots de algumas delas, observados na Figura 3.1, para a identificação de possíveis *outliers* que pudessem prejudicar a formação dos grupos pelos métodos, já que ambos os algoritmos são sensíveis a esse tipo dado.

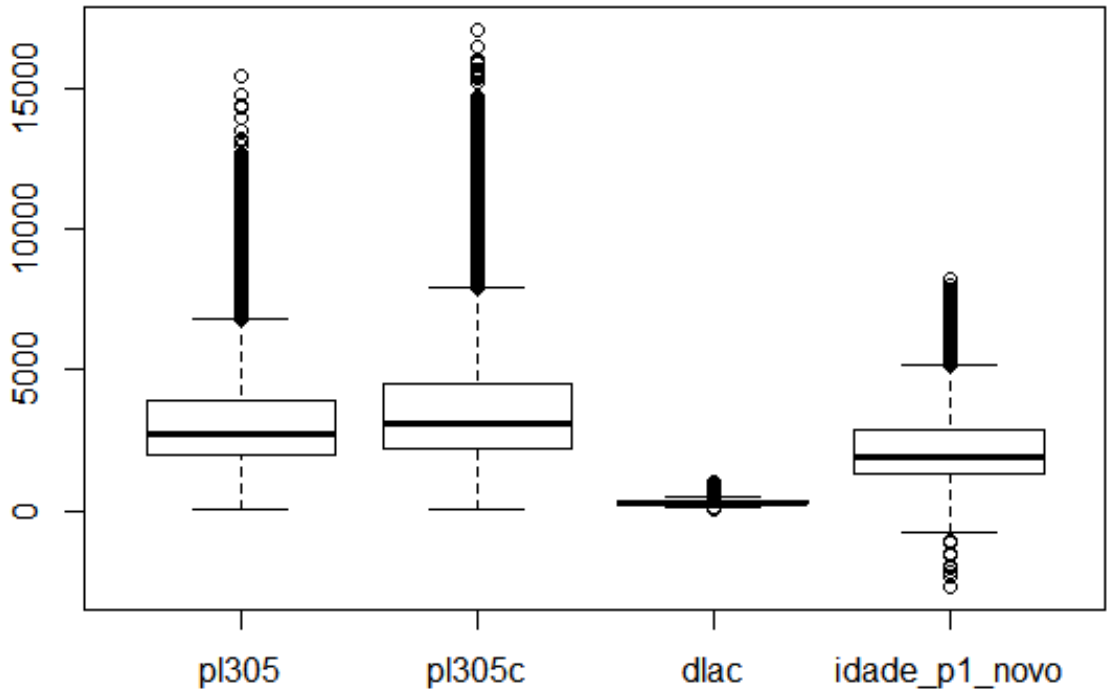

Figura 3.1: Boxplots dos atributos pl305, pl305c, dlac e idade_p1_novo

Devido às grandes diferenças de escala, a avaliação do boxplot relativo a duração da lactação foi prejudicada e, por isso, construi-se o boxplot individual para sua melhor visuzalição de acordo com a escala da variável, presente na Figura 3.2

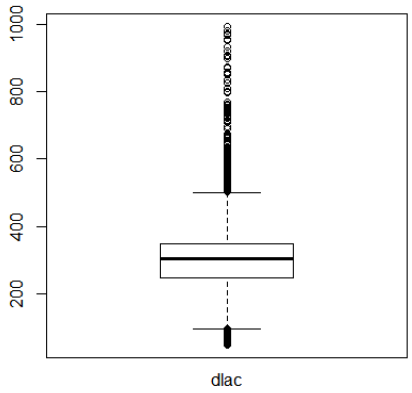

Figura 3.2: Boxplots do atributo dlac

Pela análise de todos os boxplots, foi possível identificar *outliers* acima do limite máximo de cada variável. Adicionalmente, notou-se alguns desses pontos abaixo do limite inferior de idade_p1_novo, que são provavelmente devido a erros de coleta na coluna parto, já que, como ela é derivada da diferença entre parto e nasc, a coluna parto, nesses casos, apresenta valores menores que nasc, o que não faz sentido.

Foi computada a correlação entre as variáveis pl305 e pl305c na amostra, cujo valor foi de 0.977 e, como esperado, observou-se uma alta correlação positiva entre elas. Logo, a remoção de todos os *outliers* foi realizada nas variáveis pl305c, dlac e idade_p1_novo. Viu-se também que há uma correlação parcial entre pl305c e dlac, com valor de 0.521.

3.6 Escolha dos atributos para utilização nos métodos de agrupamento

As variáveis numéricas selecionadas como referência para o particionamento dos dados pelos algoritmos nas simulações, tendo em vista o alto percentual de ausências em variáveis de análise do leite, foram definidas por combinações bidimensionais entre as variáveis pl305c, idp1_novo, idp2_novo e prod_diaria: (pl305c, idp1_novo), (pl305c, idp2_novo), (prod_diaria, idp1_novo) e (prod_diaria, idp2_novo). Tais combinações foram consideradas, pois, após o tratamento dos dados e a subdivisão da amostra, mostrada na seção a seguir, estes foram os atributos restantes para serem utilizados nos agrupamentos, incluindo também a duração da lactação.

Os valores dessas variáveis foram normalizados pelo método min-max antes da aplicação dos algoritmos, em decorrência das variações de escala entre os atributos mencionados, evitando que um atributo de maior grandeza exercesse maior influência sobre o outro no processo de formação dos grupos, já que a proximidade entre os objetos é calculada através de medidas de distância, como a euclidiana, utilizada nas simulações deste trabalho.

3.7 Subdivisão da amostra

Após a remoção de *outliers*, a amostra de dados foi subdividida de acordo com o regime alimentar. Para lidar com o problema dos valores ausentes ainda existentes em idp1_novo e idp2_novo, cada uma delas foi novamente subdividida em mais dois conjuntos, compondo 12 subconjuntos mutuamente exclusivos, conforme mostra a tabela a seguir, usados nas simulações.

	nº de objetos
gs1_ra1_idp1	367
gs1_ra2_idp1	5414
gs1_ra3_idp1	425
gs1_ra4_idp1	3685
gs1_ra5_idp1	2370
gs1_ra9_idp1	4171
gs1_ra1_idp2	970
gs1_ra2_idp2	14667
gs1_ra3_idp2	590
gs1_ra4_idp2	3585
gs1_ra5_idp2	1796
gs1_ra9_idp2	14569

Tabela 3.6: Número de objetos em cada subconjunto definido após a subdivisão dos casos com gs=1 de acordo com o regime alimentar e ordem de parto.

O nome de cada subconjunto diz respeito ao regime alimentar e a ordem de parto dos animais que o compõem. Por exemplo, o subconjunto denominado gs1_ra1_idp1 refere-se aos animais presentes na amostra (raça Gir, i.e., gs=1), cujo regime alimentar é dado por ra=1 e contém os gados com idade ao primeiro parto efetiva.

3.8 Parâmetros dos algoritmos

Como ambos os algoritmos, KM e FCM, dependem da prévia especificação do número k de grupos, utilizou-se os índices internos descritos neste trabalho para cada método a fim de determinar um *range* de valores adequados para efetuar as simulações, considerando um k_{max} inicial de 15 grupos. Para todos os agrupamentos com o FCM, definiu-se o valor 2 para o índice de fuzzificação m.

3.9 Condução da avaliação das simulações

Para cada algoritmo, avaliou-se individualmente os agrupamentos gerados pelos mesmos através de índices internos, em termos da baixa coesão e alta separação dos grupos. Posteriormente, comparou-se as soluções entre os dois algoritmos, verificando as semelhanças e diferenças existentes. Avaliou-se também a relação dos atributos considerados em cada simulação com a produção e o comportamento dos dados entre os diferentes tipos de regime alimentar.

Em relação à tendência de agrupamento dos dados, foram computadas as estatísticas de Hopkins para cada subconjunto, com diferentes combinações de atributos, observando seus efeitos nos valores de H para melhor julgamento da aleatoriedade dos subconjuntos. Realizou-se 6 execuções distintas com 10% de cada subconjunto para a composição das amostras U e V, presentes na definição dessa estatística (equação 2.21), contabilizando o valor mínimo, máximo e média dos valores de H.

3.10 Ferramentas

As simulações foram realizadas no RStudio, ambiente de desenvolvimento integrado para R, em uma máquina com um processador Intel® Core™ i3-3110M (2 núcleos, 2.40 GHz), 4GB de memória RAM e sistema operacional Windows 8.1 versão 6.3 (compilação 9600) x64. Foram utilizados diversos pacotes em R contendo funções de suporte à análise de dados. Foram também desenvolvidos diversos scripts na linguagem R para este fim.

4 Resultados e Discussão

4.1 Experimento 1: Determinação do *range* de valores ideais para k

Para se estimar o conjunto de valores ideais para o número de grupos, utilizados como parâmetros de inicialização do KM e FCM, foram computados os valores dos índices internos apresentados neste trabalho na subseção 2.3.2 para cada algoritmo, através das diferentes combinacões de atributos já mencionadas, a saber, (pl305c, idp1_novo), (pl305c, idp2_novo), (prod_diaria, idp1_novo) e (prod_diaria, idp2_novo). Além desses índices, utilizou-se a função *NbClust* do pacote 'NbClust' (v3.0) (Charrad et al., 2014) que provê 30 índices de validação interna para determinação de k com o método KM e métodos hierárquicos de agrupamento.

Dentre eles, foram calculados 26 dos 30 índices presentes no pacote. O k_{max} inicial considerado foi de 15 grupos. Os valores foram reunidos em tabelas, separadas por cada método e combinações de atributos utilizadas. Nas colunas do *gap statistic* das Tabelas 4.1 e 4.3, as células com mais de um valor se devem a diferentes abordagens utilizadas no cômputo dessa estatística, sendo que os valores acompanhados de "max" correspondem aos máximos globais da função, calculada para k entre 2 e 15, pela abordagem denominada globalmax, enquanto que, para o outro valor, utilizou-se a abordagem padrão, chamada de firstSEmax, em que o número ideal de grupos é dado como o menor valor de k tal que a função aplicada sobre ele não exceda um desvio padrão do máximo local.

Para a coluna dos índices do pacote 'NbClust', os valores entre colchetes indicam a quantidade deles que apontaram o valor anterior como sendo o número ideal de grupos. Em todas as quatro tabelas, mostradas a seguir, referentes a k, "*" indica a melhor escolha dentre os bons valores de cada célula. Além disso, as células contendo a expressão "Lack of Memory" é consequência da impossibilidade de cômputo do valor do índice em razão do alto consumo de memória exigido para a alocação da matriz de distância nos subconjuntos

de maior número de objetos.

	k ideal utilizando o KM			
	Índice			
	elbow (SSE)	**gap_statistic**	**CS**	**pacote NbClust (26 índices)**
gs1_ra1_idp1	4	1	2	2 [6] ou 3* [8]
gs1_ra2_idp1	5	2	2	2* [7] ou 3 [6]
gs1_ra3_idp1	4	1 ou 2_{max}	2	2* [6] ou 4 [6]
gs1_ra4_idp1	5	1	3	2 [6] ou 3* [7]
gs1_ra5_idp1	4	1	3	2 [6] ou 3* [8]
gs1_ra9_idp1	4	1	2	2* [7] ou 3 [6]
gs1_ra1_idp2	4	2	2	2 [6] ou 3* [8]
gs1_ra2_idp2	Lack of Memory	2	2	Lack of Memory
gs1_ra3_idp2	5	2 ou 3_{max}	2	2 [5] ou 3* [10]
gs1_ra4_idp2	6	1 ou 3_{max}	2	2 [5] ou 3* [11]
gs1_ra5_idp2	6	3	3	2 [4], 3* [9] ou 4 [3]
gs1_ra9_idp2	Lack of Memory	2	2	Lack of Memory

Tabela 4.1: k ideal utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c) pelo método KM.

	k ideal utilizando o FCM					
	Índice					
	CSF	**PC**	**MPC**	**PE**	**FS**	**XB**
gs1_ra1_idp1	3, 10 ou 13*	2	3	2	4, 9* ou 13	9 ou 13*
gs1_ra2_idp1	2	2	2	2	5	12
gs1_ra3_idp1	4 ou 11*	2	3	2	9 ou 12*	7, 11* ou 15
gs1_ra4_idp1	3	2	3	2	5	6 ou 9*
gs1_ra5_idp1	3	2	3	2	5	3, 9 ou 15*
gs1_ra9_idp1	2	2	2	2	3 ou 8*	14 ou 15*
gs1_ra1_idp2	2	2	2	2	7	2* ou 15
gs1_ra2_idp2	2	2	2	2	4	2, 12, 14*
gs1_ra3_idp2	3	2	2	2	4, 8 ou 14*	8 ou 9*
gs1_ra4_idp2	3	2	3	2	6	6* ou 13
gs1_ra5_idp2	3	2	3	2	6	3, 7* ou 10
gs1_ra9_idp2	2	2	2	2	4	2

Tabela 4.2: k ideal utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c) pelo método FCM.

	k ideal utilizando o KM			
	Índice			
	elbow (SSE)	**gap_statistic**	**CS**	**pacote NbClust (30 índices)**
gs1_ra1_idp1	5	3	2	2* [9] ou 3 [7]
gs1_ra2_idp1	4	3	2	2* [11] ou 3 [6]
gs1_ra3_idp1	4	3	2	2* [9] ou 3 [6]
gs1_ra4_idp1	4	3	2	2* [10], 3 [6] ou 4 [3]
gs1_ra5_idp1	4	3	2	2* [10] ou 3 [5]
gs1_ra9_idp1	4	1 ou 4_{max}	2	2* [8] ou 3 [7]
gs1_ra1_idp2	4	3	2	2* [8] ou 3 [8]
gs1_ra2_idp2	Lack of Memory	3	2	Lack of Memory
gs1_ra3_idp2	4	4	2	2 [7] ou 3* [9]
gs1_ra4_idp2	4	3	2	2* [9] ou 3 [8]
gs1_ra5_idp2	4	2	2	2 [9] ou 3* [10]
gs1_ra9_idp2	Lack of Memory	4	2	Lack of Memory

Tabela 4.3: k ideal utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria) pelo método KM.

	k ideal utilizando o FCM					
	Índice					
	CSF	**PC**	**MPC**	**PE**	**FS**	**XB**
gs1_ra1_idp1	2	2	2	2	3	2
gs1_ra2_idp1	2	2	2	2	3	2
gs1_ra3_idp1	2	2	2	2	3	2
gs1_ra4_idp1	2	2	2	2	3	2
gs1_ra5_idp1	2	2	2	2	3	2
gs1_ra9_idp1	2	2	2	2	4	2
gs1_ra1_idp2	2	2	2	2	3	2
gs1_ra2_idp2	2	2	2	2	3	2
gs1_ra3_idp2	2	2	2	2	3	2
gs1_ra4_idp2	2	2	2	2	3	2
gs1_ra5_idp2	2	2	2	2	3	2
gs1_ra9_idp2	2	2	2	2	3	2

Tabela 4.4: k ideal utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria) pelo método FCM.

Conforme os valores apresentados nas tabelas anteriores, notou-se de início que, nitidamente, os valores de maior ocorrência foram 2, 3 e 4. Além disso, os outros dois mais votados, contabilizados em termos das células de valor único e das melhores escolhas

das células com múltiplos valores, foram 5 e 6, que obtiveram 7 e 5 ocorrências, respectivamente. Desse modo, considerou-se $k = \{2, ..., 7\}$ como os diferentes valores ideais de grupo utilizados nos algoritmos.

4.2 Experimento 2: Agrupamentos dos subconjuntos com o KM e FCM

Cada subconjunto foi agrupado pelos algoritmos conforme o *range* ideal de k estabelecido no experimento anterior. Nos agrupamentos com o KM, os dois índices utilizados na validação interna das partições *crisp* em relação a boa definição dos grupos foram CS e SSE, mostrados, respectivamente, nas Tabelas 4.5 e 4.6; e Tabelas 4.10 e 4.11. No FCM, por sua vez, como os índices PC, MPC e PE fazem uso somente da matriz de pertinência no cômputo de seus valores, utilizou-se CSF, FS e XB na validação das partições *fuzzy*, como vistos nas Tabelas 4.7, 4.8 e 4.9; e Tabelas 4.12, 4.13 e 4.14. Para cada subconjunto, as linhas em amarelo nas tabelas representam o valor ótimo do índice dentre os distintos k utilizados, excetuando-se nas tabelas da SSE, onde a noção de melhor valor é dada como sendo o k em que a partir dele não se tem redução significativa do valor da variação intracluster. Nas tabelas relativas à mesma, a ausência de valores para alguns subconjuntos surgiu em decorrência do alto custo computacional envolvido na alocação da matriz de distância associado ao tamanho do conjunto no cômputo dessa medida.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**CS**	**k=2**	**0.397**	**gs1_ra1_idp2**	**CS**	**k=2**	**0.471**
		k=3	0.395			k=3	0.41
		k=4	0.362			k=4	0.383
		k=5	0.348			k=5	0.362
		k=6	0.345			k=6	0.349
		k=7	0.351			k=7	0.351
gs1_ra2_idp1	**CS**	**k=2**	**0.39**	**gs1_ra2_idp2**	**CS**	**k=2**	**0.465**
		k=3	0.369			k=3	0.388
		k=4	0.352			k=4	0.376
		k=5	0.339			k=5	0.367
		k=6	0.341			k=6	0.345
		k=7	0.336			k=7	0.342
gs1_ra3_idp1	**CS**	**k=2**	**0.369**	**gs1_ra3_idp2**	**CS**	**k=2**	**0.402**
		k=3	0.366			k=3	0.399
		k=4	0.363			k=4	0.38
		k=5	0.346			k=5	0.373
		k=6	0.353			k=6	0.341
		k=7	0.356			k=7	0.347
gs1_ra4_idp1	**CS**	k=2	0.35	**gs1_ra4_idp2**	**CS**	**k=2**	**0.413**
		k=3	**0.368**			k=3	0.405
		k=4	0.347			k=4	0.348
		k=5	0.349			k=5	0.349
		k=6	0.344			k=6	0.371
		k=7	0.34			k=7	0.368
gs1_ra5_idp1	**CS**	k=2	0.376	**gs1_ra5_idp2**	**CS**	k=2	0.368
		k=3	**0.393**			**k=3**	**0.421**
		k=4	0.371			k=4	0.403
		k=5	0.357			k=5	0.362
		k=6	0.334			k=6	0.38
		k=7	0.345			k=7	0.376
gs1_ra9_idp1	**CS**	**k=2**	**0.399**	**gs1_ra9_idp2**	**CS**	**k=2**	**0.486**
		k=3	0.365			k=3	0.419
		k=4	0.354			k=4	0.408
		k=5	0.335			k=5	0.357
		k=6	0.321			k=6	0.368
		k=7	0.322			k=7	0.348

Tabela 4.5: Valores de coeficiente de silhueta obtidos com o KM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**SSE**	k=2	12.701	**gs1_ra1_idp2**	**SSE**	k=2	40.084
		k=3	7.944			k=3	28.554
		k=4	**6.147**			**k=4**	**21.279**
		k=5	5.08			k=5	16.913
		k=6	4.332			k=6	14.772
		k=7	3.677			k=7	12.986
gs1_ra2_idp1	**SSE**	k=2	218.499	**gs1_ra2_idp2**	**SSE**	k=2	
		k=3	148.968			k=3	
		k=4	115.193			k=4	
		k=5	**94.367**			k=5	
		k=6	80.594			k=6	
		k=7	68.377			k=7	
gs1_ra3_idp1	**SSE**	k=2	20.531	**gs1_ra3_idp2**	**SSE**	k=2	34.344
		k=3	14.298			k=3	21.603
		k=4	**10.133**			k=4	18.336
		k=5	8.587			**k=5**	**13.849**
		k=6	7.029			k=6	11.736
		k=7	6.095			k=7	10.794
gs1_ra4_idp1	**SSE**	k=2	182.003	**gs1_ra4_idp2**	**SSE**	k=2	195.868
		k=3	121.313			k=3	116.036
		k=4	92.304			k=4	94.571
		k=5	**73.653**			k=5	76.292
		k=6	62.848			**k=6**	**61.461**
		k=7	53.524			k=7	54.528
gs1_ra5_idp1	**SSE**	k=2	110.93	**gs1_ra5_idp2**	**SSE**	k=2	105.093
		k=3	73.083			k=3	61.473
		k=4	**55.077**			k=4	50.137
		k=5	44.213			k=5	39.326
		k=6	37.863			**k=6**	**31.554**
		k=7	32.019			k=7	27.679
gs1_ra9_idp1	**SSE**	k=2	128.11	**gs1_ra9_idp2**	**SSE**	k=2	
		k=3	89.131			k=3	
		k=4	**71.377**			k=4	
		k=5	58.039			k=5	
		k=6	50.474			k=6	
		k=7	43.545			k=7	

Tabela 4.6: Valores de variação intracluster (SSE) obtidos com o KM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**CSF**	k=2	0.611	**gs1_ra1_idp2**	**CSF**	**k=2**	**0.724**
		k=3	**0.656**			k=3	0.605
		k=4	0.629			k=4	0.612
		k=5	0.602			k=5	0.527
		k=6	0.621			k=6	0.611
		k=7	0.643			k=7	0.595
gs1_ra2_idp1	**CSF**	**k=2**	**0.645**	**gs1_ra2_idp2**	**CSF**	**k=2**	**0.724**
		k=3	0.635			k=3	0.567
		k=4	0.629			k=4	0.644
		k=5	0.619			k=5	0.584
		k=6	0.623			k=6	0.593
		k=7	0.63			k=7	0.611
gs1_ra3_idp1	**CSF**	k=2	0.621	**gs1_ra3_idp2**	**CSF**	k=2	0.657
		k=3	0.647			**k=3**	**0.669**
		k=4	**0.65**			k=4	0.652
		k=5	0.631			k=5	0.652
		k=6	0.633			k=6	0.631
		k=7	0.644			k=7	0.629
gs1_ra4_idp1	**CSF**	k=2	0.606	**gs1_ra4_idp2**	**CSF**	k=2	0.643
		k=3	**0.65**			**k=3**	**0.684**
		k=4	0.622			k=4	0.605
		k=5	0.63			k=5	0.623
		k=6	0.622			k=6	0.665
		k=7	0.633			k=7	0.649
gs1_ra5_idp1	**CSF**	k=2	0.629	**gs1_ra5_idp2**	**CSF**	k=2	0.605
		k=3	**0.676**			**k=3**	**0.698**
		k=4	0.633			k=4	0.606
		k=5	0.638			k=5	0.641
		k=6	0.625			k=6	0.669
		k=7	0.638			k=7	0.665
gs1_ra9_idp1	**CSF**	**k=2**	**0.641**	**gs1_ra9_idp2**	**CSF**	**k=2**	**0.746**
		k=3	0.618			k=3	0.615
		k=4	0.608			k=4	0.679
		k=5	0.592			k=5	0.624
		k=6	0.581			k=6	0.55
		k=7	0.581			k=7	0.574

Tabela 4.7: Valores de coeficiente de silhueta *fuzzy* obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**FS**	k=2	-4.312	**gs1_ra1_idp2**	**FS**	k=2	-31.476
		k=3	-8.441			k=3	-29.141
		k=4	**-8.583**			k=4	-31.974
		k=5	-8.205			k=5	-30.324
		k=6	-8.229			k=6	-30.967
		k=7	-8.315			**k=7**	**-33.655**
gs1_ra2_idp1	**FS**	k=2	-106.886	**gs1_ra2_idp2**	**FS**	k=2	-553.24
		k=3	-131.837			k=3	-533.941
		k=4	-132.519			**k=4**	**-573.609**
		k=5	**-136.402**			k=5	-547.431
		k=6	-134.746			k=6	-535.998
		k=7	-132.487			k=7	-535.395
gs1_ra3_idp1	**FS**	k=2	-9.343	**gs1_ra3_idp2**	**FS**	k=2	-18.452
		k=3	-11.76			k=3	-24.399
		k=4	-13.544			**k=4**	**-25.075**
		k=5	-13.039			k=5	-23.448
		k=6	-13.255			k=6	-24.285
		k=7	**-13.625**			k=7	-24.3
gs1_ra4_idp1	**FS**	k=2	-72.599	**gs1_ra4_idp2**	**FS**	k=2	-82.065
		k=3	-99.8			k=3	-134.928
		k=4	-101.792			k=4	-126.799
		k=5	**-108.551**			k=5	-137.76
		k=6	-105.682			**k=6**	**-140.756**
		k=7	-106.49			k=7	-136.405
gs1_ra5_idp1	**FS**	k=2	-49.242	**gs1_ra5_idp2**	**FS**	k=2	-40.529
		k=3	-67.227			k=3	-68.445
		k=4	-70.453			k=4	-63.745
		k=5	**-73.712**			k=5	-68.957
		k=6	-71.794			**k=6**	**-75.859**
		k=7	-71.851			k=7	-74.301
gs1_ra9_idp1	**FS**	k=2	-55.831	**gs1_ra9_idp2**	**FS**	k=2	-585.033
		k=3	**-70.661**			k=3	-583.884
		k=4	-70.229			**k=4**	**-595.264**
		k=5	-69.277			k=5	-581.51
		k=6	-69.124			k=6	-542.238
		k=7	-69.816			k=7	-526.291

Tabela 4.8: Valores do índice de Fukuyama-Sugeno obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**XB**	k=2	0.362	**gs1_ra1_idp2**	**XB**	**k=2**	**0.181**
		k=3	0.244			k=3	0.412
		k=4	0.282			k=4	0.224
		k=5	0.247			k=5	0.341
		k=6	0.217			k=6	0.206
		k=7	**0.19**			k=7	0.288
gs1_ra2_idp1	**XB**	k=2	0.262	**gs1_ra2_idp2**	**XB**	**k=2**	**0.169**
		k=3	0.245			k=3	0.378
		k=4	0.231			k=4	0.188
		k=5	0.196			k=5	0.254
		k=6	0.201			k=6	0.265
		k=7	**0.174**			k=7	0.22
gs1_ra3_idp1	**XB**	k=2	0.279	**gs1_ra3_idp2**	**XB**	k=2	0.246
		k=3	0.215			k=3	0.205
		k=4	0.152			**k=4**	**0.158**
		k=5	0.169			k=5	0.167
		k=6	0.15			k=6	0.209
		k=7	**0.137**			k=7	0.159
gs1_ra4_idp1	**XB**	k=2	0.307	**gs1_ra4_idp2**	**XB**	k=2	0.305
		k=3	0.183			k=3	0.182
		k=4	0.221			k=4	0.262
		k=5	0.174			k=5	0.196
		k=6	**0.145**			**k=6**	**0.143**
		k=7	0.208			k=7	0.197
gs1_ra5_idp1	**XB**	k=2	0.283	**gs1_ra5_idp2**	**XB**	k=2	0.314
		k=3	**0.161**			k=3	0.145
		k=4	0.202			k=4	0.236
		k=5	0.22			k=5	0.187
		k=6	0.202			k=6	0.14
		k=7	0.168			**k=7**	**0.129**
gs1_ra9_idp1	**XB**	k=2	0.291	**gs1_ra9_idp2**	**XB**	**k=2**	**0.145**
		k=3	0.28			k=3	0.285
		k=4	0.244			k=4	0.18
		k=5	0.213			k=5	0.206
		k=6	0.223			k=6	0.335
		k=7	**0.207**			k=7	0.265

Tabela 4.9: Valores do índice de Xie-Beni obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

Pelos valores da Tabela 4.5, notou-se a existência de fracas estruturas de agrupamento utilizando as combinações (idp1_novo, pl305c) e (idp2_novo, pl305c), já que, em todos os subconjuntos, os valores de CS foram inferiores a 0.5. Tal fato também foi verificado através da Tabela 4.6 pela presença de valores distantes de 0, uma vez que grupos bem definidos possuem baixa coesão, i.e., baixa variação intracluster. Consequentemente, foram indícios de sobreposições entre os grupos das partições geradas com o KM usando essas combinações de atributos.

A Tabela 4.7 de CSF para as combinações (idp1_novo, pl305c) e (idp2_novo, pl305c) permitiu supor a existência tanto de objetos bem agrupados nos subconjuntos pelo FCM, quanto de objetos com baixos graus de pertinência aos grupos nos quais eles foram atribuídos, pois, para todos os subconjuntos, o valor ótimo de CSF oscilou entre 0.6 e 0.8, como também a maioria dos demais valores. Adicionalmente, viu-se que, pelas Tabelas 4.8 para FS e 4.9 para XB, as partições dos subconjuntos geradas pelo algoritmo FCM apresentaram separação de grupos superior a coesão interna de cada um deles, já que todos os valores ótimos de FS, computados por 2.32, tiveram valores negativos, além do que, todos os valores de XB, calculados através de 2.33, foram inferiores a 1.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**CS**	**k=2**	**0.509**	**gs1_ra1_idp2**	**CS**	**k=2**	**0.585**
		k=3	0.44			k=3	0.5
		k=4	0.456			k=4	0.44
		k=5	0.428			k=5	0.393
		k=6	0.377			k=6	0.414
		k=7	0.353			k=7	0.392
gs1_ra2_idp1	**CS**	**k=2**	**0.513**	**gs1_ra2_idp2**	**CS**	**k=2**	**0.581**
		k=3	0.431			k=3	0.501
		k=4	0.386			k=4	0.443
		k=5	0.39			k=5	0.403
		k=6	0.365			k=6	0.403
		k=7	0.373			k=7	0.379
gs1_ra3_idp1	**CS**	**k=2**	**0.5**	**gs1_ra3_idp2**	**CS**	**k=2**	**0.546**
		k=3	0.419			k=3	0.464
		k=4	0.373			k=4	0.406
		k=5	0.356			k=5	0.372
		k=6	0.362			k=6	0.366
		k=7	0.368			k=7	0.359
gs1_ra4_idp1	**CS**	**k=2**	**0.527**	**gs1_ra4_idp2**	**CS**	**k=2**	**0.583**
		k=3	0.431			k=3	0.468
		k=4	0.376			k=4	0.393
		k=5	0.333			k=5	0.374
		k=6	0.341			k=6	0.37
		k=7	0.351			k=7	0.37
gs1_ra5_idp1	**CS**	**k=2**	**0.52**	**gs1_ra5_idp2**	**CS**	**k=2**	**0.542**
		k=3	0.421			k=3	0.456
		k=4	0.363			k=4	0.382
		k=5	0.348			k=5	0.359
		k=6	0.339			k=6	0.37
		k=7	0.345			k=7	0.363
gs1_ra9_idp1	**CS**	**k=2**	**0.505**	**gs1_ra9_idp2**	**CS**	**k=2**	**0.505**
		k=3	0.441			k=3	0.441
		k=4	0.4			k=4	0.4
		k=5	0.351			k=5	0.351
		k=6	0.371			k=6	0.371
		k=7	0.377			k=7	0.377

Tabela 4.10: Valores de coeficiente de silhueta obtidos com o KM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**SSE**	k=2	3.933	**gs1_ra1_idp2**	**SSE**	k=2	19.26
		k=3	2.46			k=3	11.251
		k=4	1.967			**k=4**	**8.45**
		k=5	**1.408**			k=5	6.786
		k=6	1.19			k=6	5.583
		k=7	1.064			k=7	4.8
gs1_ra2_idp1	**SSE**	k=2	70.736	**gs1_ra2_idp2**	**SSE**	k=2	
		k=3	45.101			k=3	
		k=4	**35.017**			k=4	
		k=5	30.114			k=5	
		k=6	25.769			k=6	
		k=7	21.681			k=7	
gs1_ra3_idp1	**SSE**	k=2	6.268	**gs1_ra3_idp2**	**SSE**	k=2	14.161
		k=3	4.031			k=3	8.553
		k=4	**3.084**			**k=4**	**6.405**
		k=5	2.64			k=5	5.321
		k=6	2.18			k=6	4.427
		k=7	1.875			k=7	3.671
gs1_ra4_idp1	**SSE**	k=2	54.194	**gs1_ra4_idp2**	**SSE**	k=2	68.804
		k=3	33.782			k=3	44.116
		k=4	**26.247**			**k=4**	**34.425**
		k=5	22.451			k=5	27.243
		k=6	19.169			k=6	23.055
		k=7	16.541			k=7	19.891
gs1_ra5_idp1	**SSE**	k=2	35.257	**gs1_ra5_idp2**	**SSE**	k=2	38.371
		k=3	22.353			k=3	25.201
		k=4	**17.497**			**k=4**	**20.021**
		k=5	15.086			k=5	15.885
		k=6	12.953			k=6	13.085
		k=7	11.231			k=7	11.358
gs1_ra9_idp1	**SSE**	k=2	44.185	**gs1_ra9_idp2**	**SSE**	k=2	
		k=3	27.547			k=3	
		k=4	**21.201**			k=4	
		k=5	18.139			k=5	
		k=6	15.544			k=6	
		k=7	13.339			k=7	

Tabela 4.11: Valores de variação intracluster (SSE) obtidos com o KM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**CSF**	**k=2**	**0.774**	**gs1_ra1_idp2**	**CSF**	**k=2**	**0.834**
		k=3	0.706			k=3	0.761
		k=4	0.65			k=4	0.682
		k=5	0.712			k=5	0.643
		k=6	0.643			k=6	0.606
		k=7	0.607			k=7	0.552
gs1_ra2_idp1	**CSF**	**k=2**	**0.778**	**gs1_ra2_idp2**	**CSF**	**k=2**	**0.832**
		k=3	0.709			k=3	0.774
		k=4	0.654			k=4	0.717
		k=5	0.593			k=5	0.665
		k=6	0.531			k=6	0.614
		k=7	0.564			k=7	0.645
gs1_ra3_idp1	**CSF**	**k=2**	**0.769**	**gs1_ra3_idp2**	**CSF**	**k=2**	**0.803**
		k=3	0.705			k=3	0.74
		k=4	0.641			k=4	0.676
		k=5	0.613			k=5	0.648
		k=6	0.64			k=6	0.628
		k=7	0.641			k=7	0.651
gs1_ra4_idp1	**CSF**	**k=2**	**0.789**	**gs1_ra4_idp2**	**CSF**	**k=2**	**0.833**
		k=3	0.714			k=3	0.727
		k=4	0.657			k=4	0.644
		k=5	0.6			k=5	0.651
		k=6	0.599			k=6	0.641
		k=7	0.6			k=7	0.642
gs1_ra5_idp1	**CSF**	**k=2**	**0.785**	**gs1_ra5_idp2**	**CSF**	**k=2**	**0.798**
		k=3	0.698			k=3	0.714
		k=4	0.633			k=4	0.618
		k=5	0.573			k=5	0.632
		k=6	0.602			k=6	0.653
		k=7	0.607			k=7	0.645
gs1_ra9_idp1	**CSF**	**k=2**	**0.766**	**gs1_ra9_idp2**	**CSF**	**k=2**	**0.834**
		k=3	0.71			k=3	0.784
		k=4	0.66			k=4	0.74
		k=5	0.6			k=5	0.693
		k=6	0.547			k=6	0.658
		k=7	0.505			k=7	0.617

Tabela 4.12: Valores de coeficiente de silhueta *fuzzy* obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**FS**	k=2	-5.319	**gs1_ra1_idp2**	**FS**	k=2	-38.967
		k=3	**-5.925**			**k=3**	**-43.704**
		k=4	-5.835			k=4	-40.751
		k=5	-5.317			k=5	-40.727
		k=6	-5.356			k=6	-39.566
		k=7	-5.102			k=7	-36.756
gs1_ra2_idp1	**FS**	k=2	-102.212	**gs1_ra2_idp2**	**FS**	k=2	-671.238
		k=3	**-107.611**			**k=3**	**-738.704**
		k=4	-105.587			k=4	-711.409
		k=5	-99.493			k=5	-672.697
		k=6	-92.64			k=6	-618.192
		k=7	-89.11			k=7	-569.926
gs1_ra3_idp1	**FS**	k=2	-8.771	**gs1_ra3_idp2**	**FS**	k=2	-23.991
		k=3	**-9.237**			**k=3**	**-27.784**
		k=4	-8.693			k=4	-26.345
		k=5	-7.974			k=5	-22.909
		k=6	-8.28			k=6	-23.334
		k=7	-7.934			k=7	-22.632
gs1_ra4_idp1	**FS**	k=2	-86.073	**gs1_ra4_idp2**	**FS**	k=2	-137.107
		k=3	**-87.532**			**k=3**	**-141.013**
		k=4	-83.317			k=4	-133.382
		k=5	-77.492			k=5	-119.506
		k=6	-72.76			k=6	-117.928
		k=7	-69.28			k=7	-111.596
gs1_ra5_idp1	**FS**	k=2	-52.298	**gs1_ra5_idp2**	**FS**	k=2	-55.759
		k=3	**-52.885**			**k=3**	**-67.274**
		k=4	-49.277			k=4	-63.437
		k=5	-45.618			k=5	-58.827
		k=6	-42.599			k=6	-56.46
		k=7	-41.166			k=7	-53.633
gs1_ra9_idp1	**FS**	k=2	-56.949	**gs1_ra9_idp2**	**FS**	k=2	-677.828
		k=3	-62.906			**k=3**	**-748.887**
		k=4	**-64.13**			k=4	-736.657
		k=5	-59.95			k=5	-690.008
		k=6	-57.121			k=6	-676.974
		k=7	-54.31			k=7	-620.192

Tabela 4.13: Valores do índice de Fukuyama-Sugeno obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

	índice	nº de grupos	valor do índice		índice	nº de grupos	valor do índice
gs1_ra1_idp1	**XB**	**k=2**	**0.115**	**gs1_ra1_idp2**	**XB**	**k=2**	**0.083**
		k=3	0.154			k=3	0.136
		k=4	0.175			k=4	0.208
		k=5	0.128			k=5	0.255
		k=6	0.214			k=6	0.258
		k=7	0.245			k=7	0.307
gs1_ra2_idp1	**XB**	**k=2**	**0.11**	**gs1_ra2_idp2**	**XB**	**k=2**	**0.079**
		k=3	0.14			k=3	0.114
		k=4	0.157			k=4	0.16
		k=5	0.216			k=5	0.214
		k=6	0.254			k=6	0.28
		k=7	0.22			k=7	0.235
gs1_ra3_idp1	**XB**	**k=2**	**0.114**	**gs1_ra3_idp2**	**XB**	**k=2**	**0.097**
		k=3	0.134			k=3	0.137
		k=4	0.181			k=4	0.202
		k=5	0.239			k=5	0.295
		k=6	0.214			k=6	0.233
		k=7	0.184			k=7	0.197
gs1_ra4_idp1	**XB**	**k=2**	**0.103**	**gs1_ra4_idp2**	**XB**	**k=2**	**0.085**
		k=3	0.158			k=3	0.178
		k=4	0.166			k=4	0.298
		k=5	0.246			k=5	0.294
		k=6	0.236			k=6	0.219
		k=7	0.203			k=7	0.204
gs1_ra5_idp1	**XB**	**k=2**	**0.109**	**gs1_ra5_idp2**	**XB**	**k=2**	**0.11**
		k=3	0.179			k=3	0.181
		k=4	0.212			k=4	0.322
		k=5	0.259			k=5	0.298
		k=6	0.247			k=6	0.23
		k=7	0.26			k=7	0.229
gs1_ra9_idp1	**XB**	**k=2**	**0.122**	**gs1_ra9_idp2**	**XB**	**k=2**	**0.077**
		k=3	0.146			k=3	0.106
		k=4	0.17			k=4	0.134
		k=5	0.22			k=5	0.181
		k=6	0.237			k=6	0.194
		k=7	0.288			k=7	0.251

Tabela 4.14: Valores do índice de Xie-Beni obtidos com o FCM no agrupamento de cada subconjunto de dados utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

Nos agrupamentos com o KM utilizando as combinações (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), observou-se que os valores do coeficiente de silhueta, mostrados na Tabela 4.10, decaem para menos de 0.5 (ou ficam iguais a este) com o aumento de k, para todos os subconjuntos de dados. Além disso, mesmo o valor ótimo do índice fica em torno deste valor, sendo que o máximo entre eles foi de 0.585. Isso permitiu ver que as partições geradas com o KM, usando essas combições, também apresentam grupos não muito bem definidos a partir dos subconjuntos. Os valores de SSE distantes de 0 também contribuíram na verificação da baixa definição dos grupos.

Pelo FCM, notou-se que, de um modo geral, os valores de CSF, mostrados na Tabela 4.12, com essas combinações de atributos, de modo semelhante aos valores de CSF das combinações anteriores, oscilaram entre 0.6 e 0.8, apesar da elevação do valor ótimo do índice nos subconjuntos. Viu-se também que as partições obtidas tiveram superioridade da separação em relação à coesão dos grupos, pelos valores de FS e XB das Tabelas 4.13 e 4.14, respectivamente.

A fim de se avaliar os limites do grau de certeza com que os objetos foram atribuídos aos clusters com o FCM, determinou-se o maior e o menor graus de pertinência máxima de um dado objeto a um dado cluster em cada subconjunto utilizando $k = \{2, ..., 7\}$ grupos, vistos nas tabelas seguintes. O conjunto de pertinências máximas, denotado por Ω nas Tabelas 4.15, 4.16, 4.17 e 4.18, reúne os maiores valores de cada linha da matriz de pertinência, i.e., as pertinências aos grupos que cada objeto foi atribuído.

	$\max(\Omega_{k=2})$	$\max(\Omega_{k=3})$	$\max(\Omega_{k=4})$	$\max(\Omega_{k=5})$	$\max(\Omega_{k=6})$	$\max(\Omega_{k=7})$
gs1_ra1_idp1	0.998	0.999	0.999	0.999	0.998	0.999
gs1_ra2_idp1	1	1	1	0.999	1	1
gs1_ra3_idp1	1	0.998	0.999	0.994	0.998	0.998
gs1_ra4_idp1	1	1	1	1	0.999	1
gs1_ra5_idp1	1	1	0.998	0.999	0.998	1
gs1_ra9_idp1	1	1	1	1	1	1
gs1_ra1_idp2	1	0.998	0.998	0.998	0.998	0.999
gs1_ra2_idp2	1	1	1	1	0.998	1
gs1_ra3_idp2	1	1	0.999	0.995	0.998	1
gs1_ra4_idp2	1	1	1	1	1	1
gs1_ra5_idp2	1	1	1	1	0.999	1
gs1_ra9_idp2	1	1	1	1	0.999	0.999

Tabela 4.15: Valores máximos de pertinência máxima alcançados no agrupamento de cada subconjunto com o FCM utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c), variando k de 2 a 7.

	$\max(\Omega_{k=2})$	$\max(\Omega_{k=3})$	$\max(\Omega_{k=4})$	$\max(\Omega_{k=5})$	$\max(\Omega_{k=6})$	$\max(\Omega_{k=7})$
gs1_ra1_idp1	1	0.999	1	1	0.999	0.995
gs1_ra2_idp1	1	0.999	1	1	0.999	1
gs1_ra3_idp1	1	1	0.995	0.997	0.996	0.996
gs1_ra4_idp1	1	1	0.999	0.999	1	1
gs1_ra5_idp1	1	0.999	1	0.998	0.999	1
gs1_ra9_idp1	1	1	0.999	1	0.997	1
gs1_ra1_idp2	1	1	1	0.999	0.999	0.999
gs1_ra2_idp2	1	1	1	1	1	0.999
gs1_ra3_idp2	1	0.998	0.999	0.998	0.999	0.999
gs1_ra4_idp2	1	1	1	1	1	0.999
gs1_ra5_idp2	1	1	1	1	0.997	0.999
gs1_ra9_idp2	1	1	1	1	0.999	0.999

Tabela 4.16: Valores máximos de pertinência máxima alcançados no agrupamento de cada subconjunto com o FCM utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

	min($\Omega_{k=2}$)	**min**($\Omega_{k=3}$)	**min**($\Omega_{k=4}$)	**min**($\Omega_{k=5}$)	**min**($\Omega_{k=6}$)	**min**($\Omega_{k=7}$)
gs1_ra1_idp1	0.504	0.357	0.318	0.278	0.276	0.245
gs1_ra2_idp1	0.5	0.341	0.307	0.282	0.251	0.24
gs1_ra3_idp1	0.502	0.377	0.31	0.285	0.28	0.259
gs1_ra4_idp1	0.5	0.355	0.312	0.265	0.237	0.239
gs1_ra5_idp1	0.5	0.34	0.309	0.267	0.226	0.262
gs1_ra9_idp1	0.5	0.347	0.312	0.287	0.251	0.228
gs1_ra1_idp2	0.502	0.361	0.293	0.23	0.228	0.213
gs1_ra2_idp2	0.5	0.357	0.292	0.246	0.228	0.212
gs1_ra3_idp2	0.5	0.364	0.311	0.285	0.246	0.194
gs1_ra4_idp2	0.5	0.344	0.299	0.243	0.244	0.255
gs1_ra5_idp2	0.5	0.346	0.31	0.27	0.268	0.239
gs1_ra9_idp2	0.5	0.366	0.294	0.25	0.209	0.186

Tabela 4.17: Valores mínimos de pertinência máxima alcançados no agrupamento de cada subconjunto com o FCM utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, prod_pl305c), variando k de 2 a 7.

	min($\Omega_{k=2}$)	**min**($\Omega_{k=3}$)	**min**($\Omega_{k=4}$)	**min**($\Omega_{k=5}$)	**min**($\Omega_{k=6}$)	**min**($\Omega_{k=7}$)
gs1_ra1_idp1	0.505	0.388	0.319	0.31	0.29	0.253
gs1_ra2_idp1	0.501	0.358	0.266	0.215	0.183	0.165
gs1_ra3_idp1	0.518	0.463	0.386	0.296	0.311	0.249
gs1_ra4_idp1	0.5	0.438	0.341	0.286	0.238	0.237
gs1_ra5_idp1	0.501	0.411	0.339	0.267	0.243	0.233
gs1_ra9_idp1	0.5	0.407	0.311	0.247	0.204	0.181
gs1_ra1_idp2	0.511	0.418	0.343	0.28	0.231	0.204
gs1_ra2_idp2	0.505	0.376	0.287	0.227	0.191	0.179
gs1_ra3_idp2	0.501	0.46	0.386	0.311	0.289	0.264
gs1_ra4_idp2	0.5	0.411	0.34	0.272	0.253	0.213
gs1_ra5_idp2	0.5	0.395	0.292	0.241	0.239	0.205
gs1_ra9_idp2	0.508	0.417	0.334	0.279	0.231	0.197

Tabela 4.18: Valores mínimos de pertinência máxima alcançados no agrupamento de cada subconjunto com o FCM utilizando as combinações de atributos (idp1_novo, prod_diaria) e (idp2_novo, prod_diaria), variando k de 2 a 7.

De acordo com as Tabelas 4.15 e 4.16, pôde-se confirmar a existência de objetos bem definidos nos grupos aos quais eles foram atribuídos no *range* de valores de k considerado em todos os subconjuntos, uma vez que os valores de Ω nessas tabelas são essencialmente 1. As Tabelas 4.17 e 4.18 permitiram observar um aumento gradual do grau de certeza com que os objetos são atribuídos aos grupos, dado que, se $\min(\Omega)$ decaisse proporcionalmente com o aumento do número de grupos, entao seus valores permaneceriam muito próximos a $\frac{1}{k}$, ou seja, próximos a $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}\}$, para k = {2,...,7}, respectivamente. Entretanto, notou-se que a razão entre $\min(\Omega)$, obtidos experimentalmente, e esses valores tende a aumentar gradualmente. Portanto, constatou-se um aumento do grau de certeza na atribuição dos objetos aos grupos conforme aumento de k. Além disso, percebeu-se que, com a elevação do número de grupos, é possível identificar objetos com pertinências máximas ainda menores, e, logo, aumentando a precisão de análise.

Com base no argumento de Zadeh (1973, p. 28) de que existe um limiar entre precisão e significância que, quando ultrapassado, elas tornam-se quase que "mutuamente exclusivas", definiu-se valores intermediarios dentre os valores de $k = \{2, ..., 7\}$, a saber, 3 e 4, como sendo os números ideais de grupos. Os gráficos dos agrupamentos por ambos os algoritmos com as diferentes combinações de atributos para cada subconjunto de dados foram plotados usando $k = 3$ e $k = 4$, exibidos nas figuras a seguir.

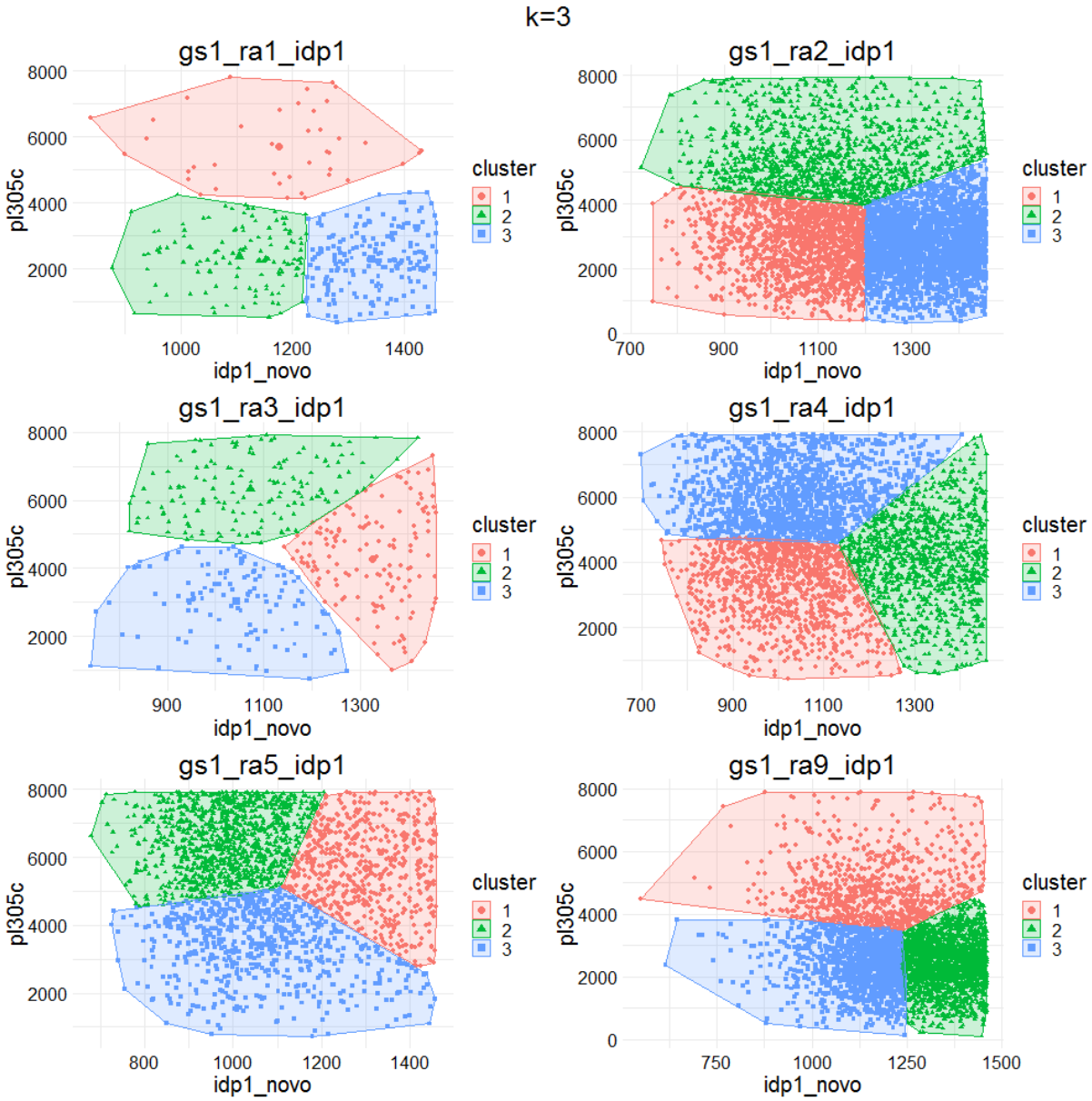

Figura 4.1: Agrupamento dos subconjuntos pelo KM com $k = 3$ utilizando a combinação de atributos (idp1_novo, pl305c)

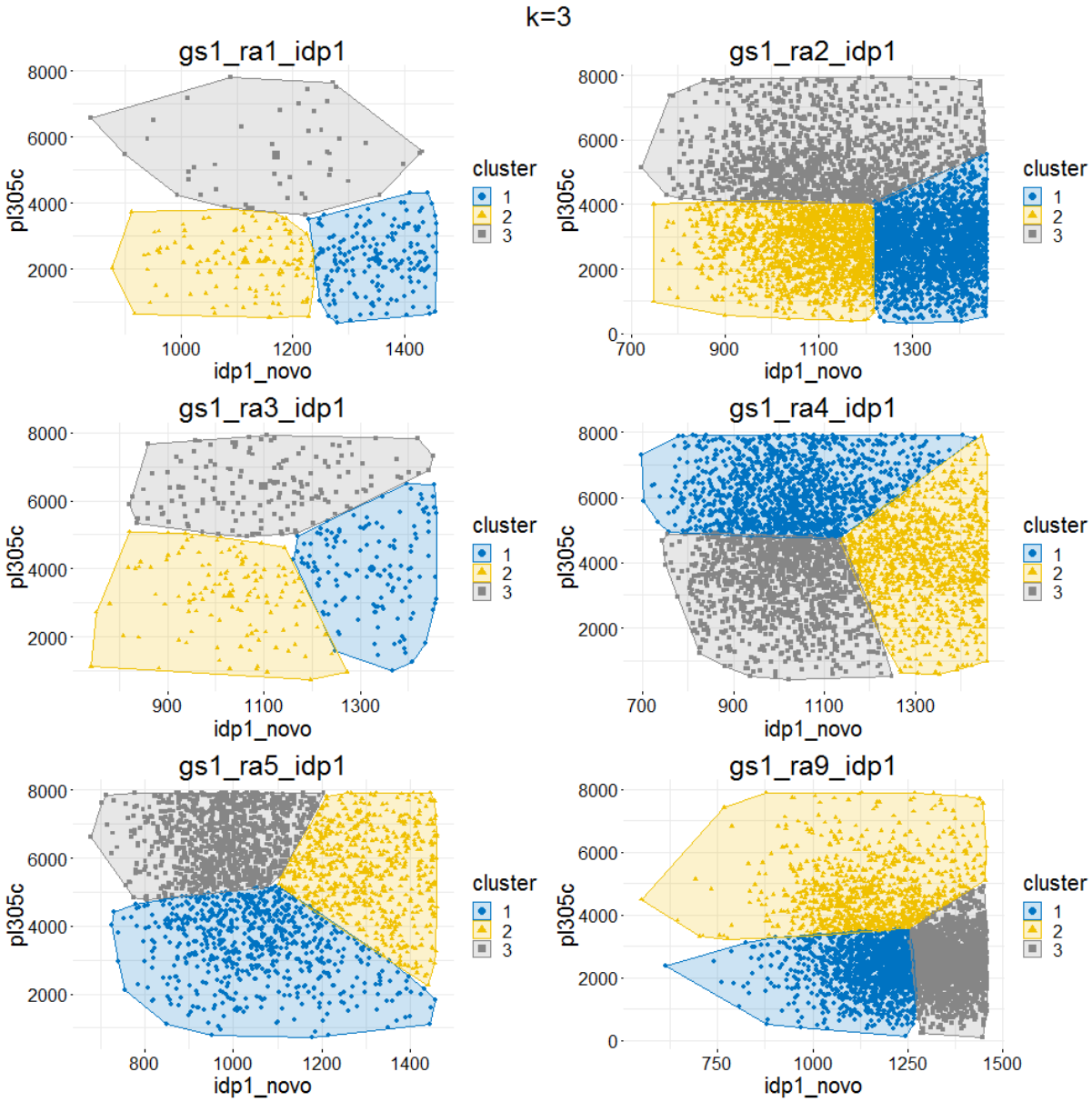

Figura 4.2: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp1_novo, pl305c)

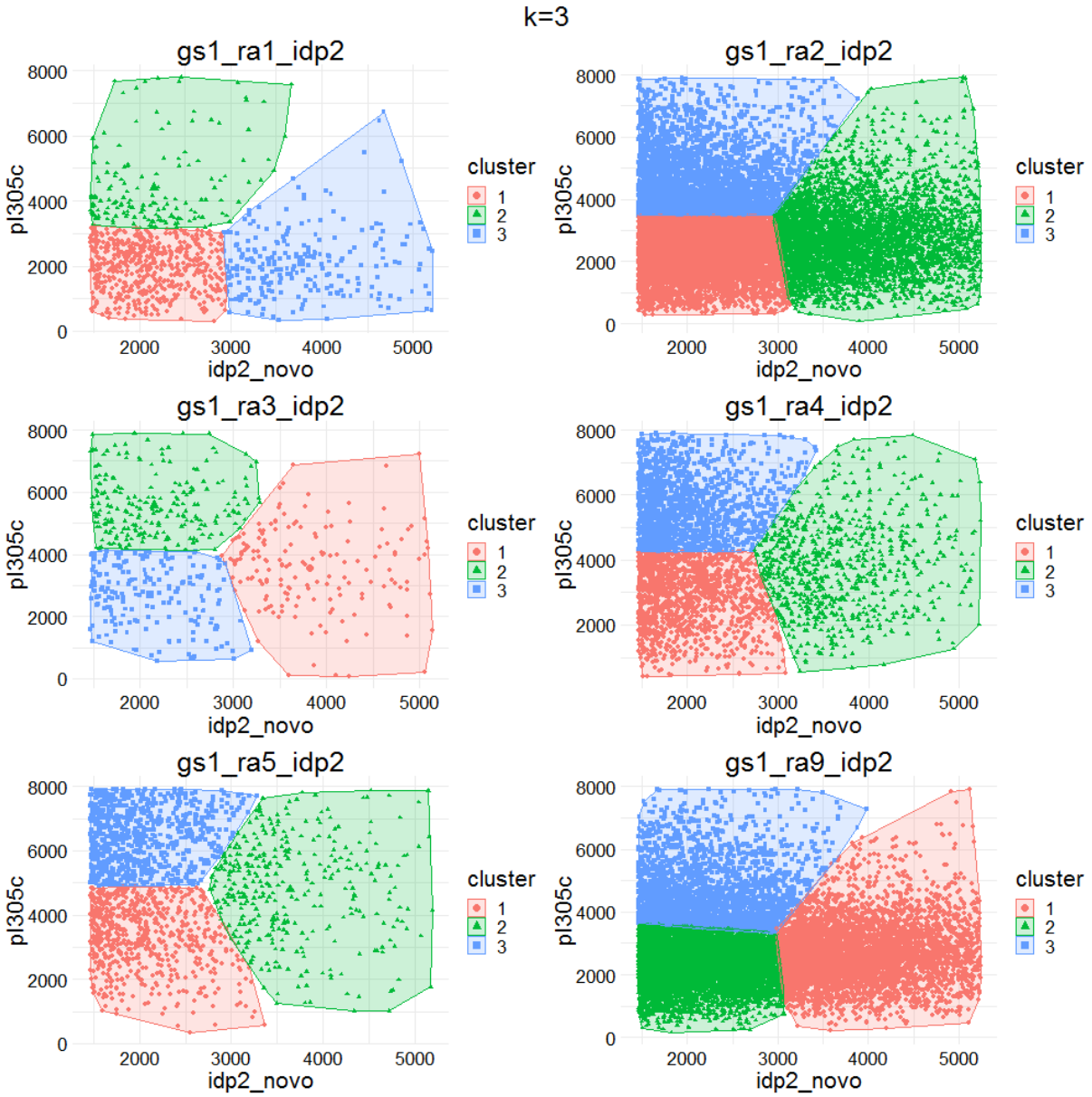

Figura 4.3: Agrupamento dos subconjuntos pelo KM com $k = 3$ utilizando a combinação de atributos (idp2_novo, pl305c)

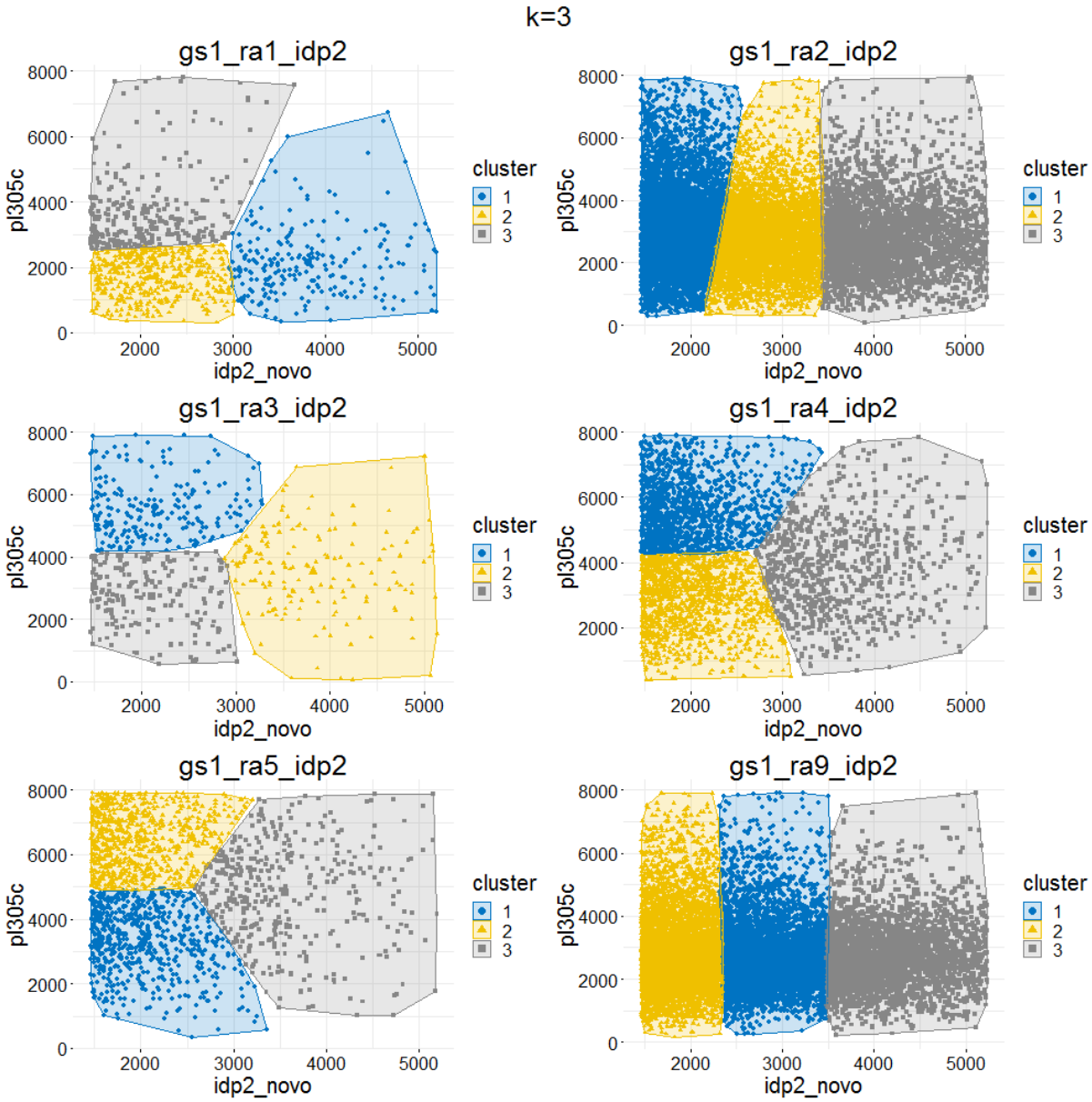

Figura 4.4: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp2_novo, pl305c)

Pela comparação das Figuras 4.1 e 4.2, notou-se semelhanças entre os algoritmos na divisão dos subconjuntos em relação às fronteiras entre cada grupo, bem como das disposições dos mesmos, e, consequentemente, houve poucas variações no deslocamento dos centroides. As diferenças mais significativas entre os métodos ocorreram nos agrupamentos com as demais ordens de parto, i.e., idp2_novo (Figuras 4.3 e 4.4), nos subconjunto de objetos com regimes alimentares 2 e 9. Nesses casos, o FCM particionou esses dois subconjuntos apenas com base em idp2_novo, observando-se perda de influência do atributo pl305c.

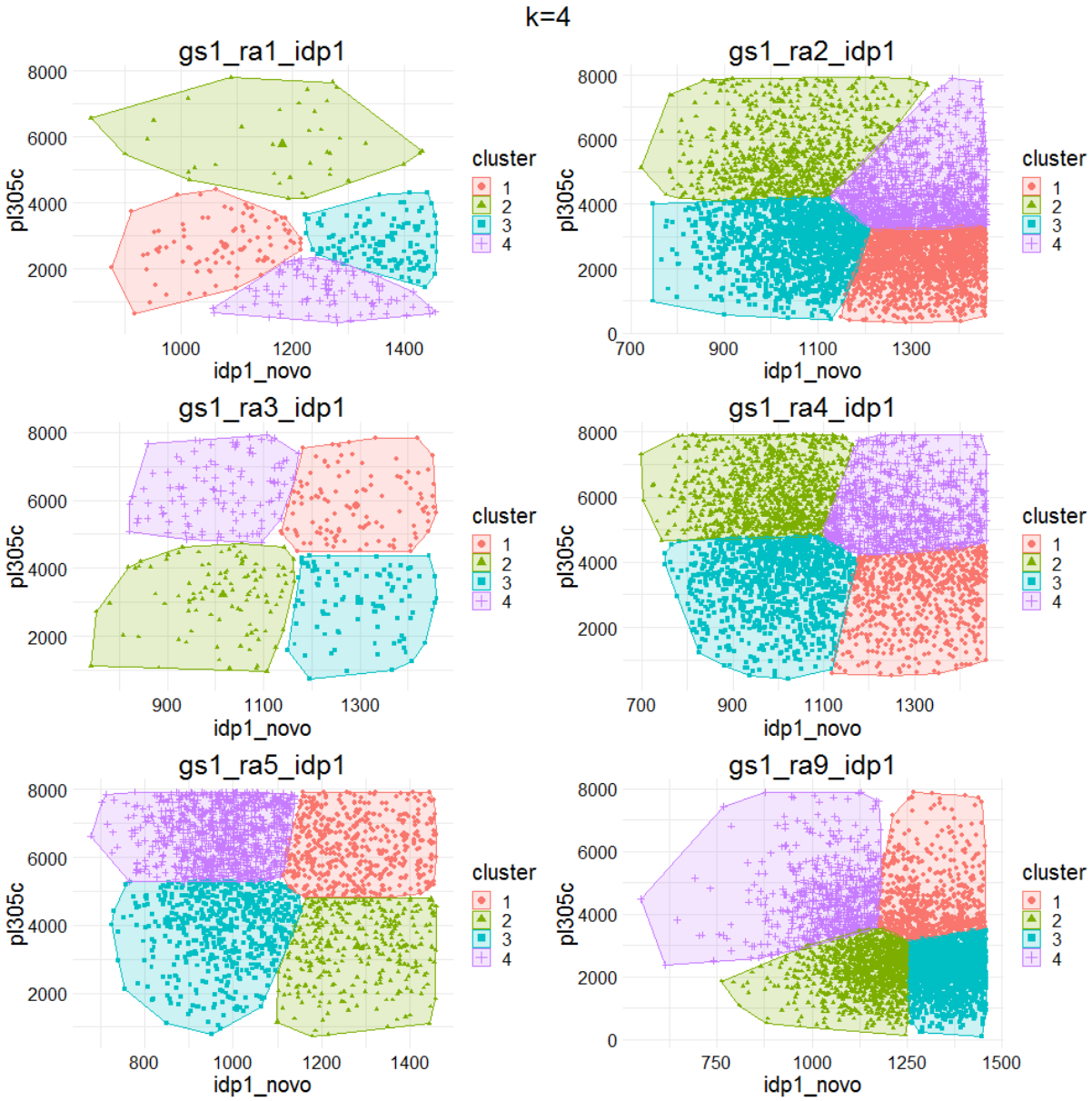

Figura 4.5: Agrupamento dos subconjuntos pelo KM com $k = 4$ utilizando a combinação de atributos (idp1_novo, pl305c)

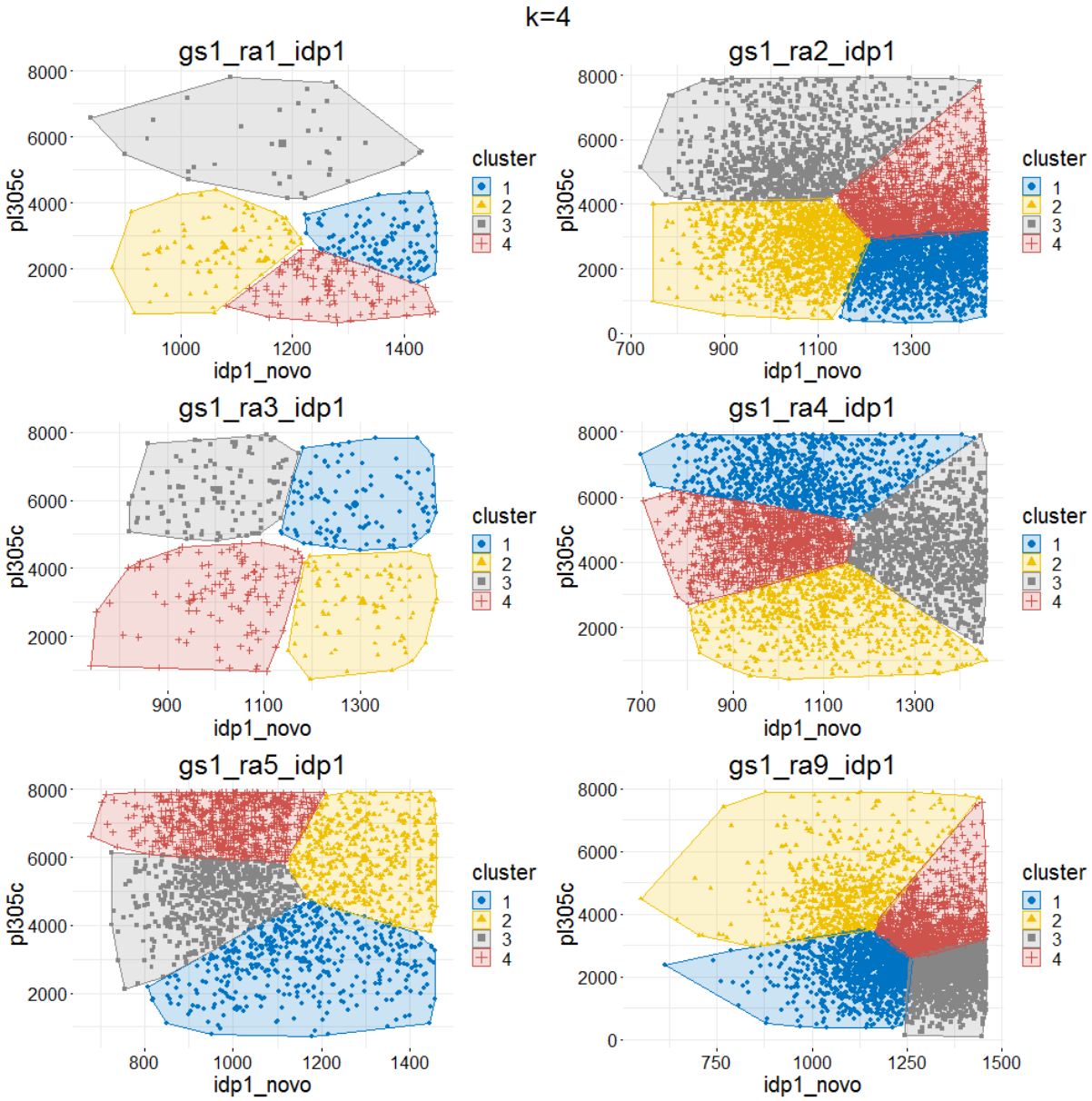

Figura 4.6: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp1_novo, pl305c)

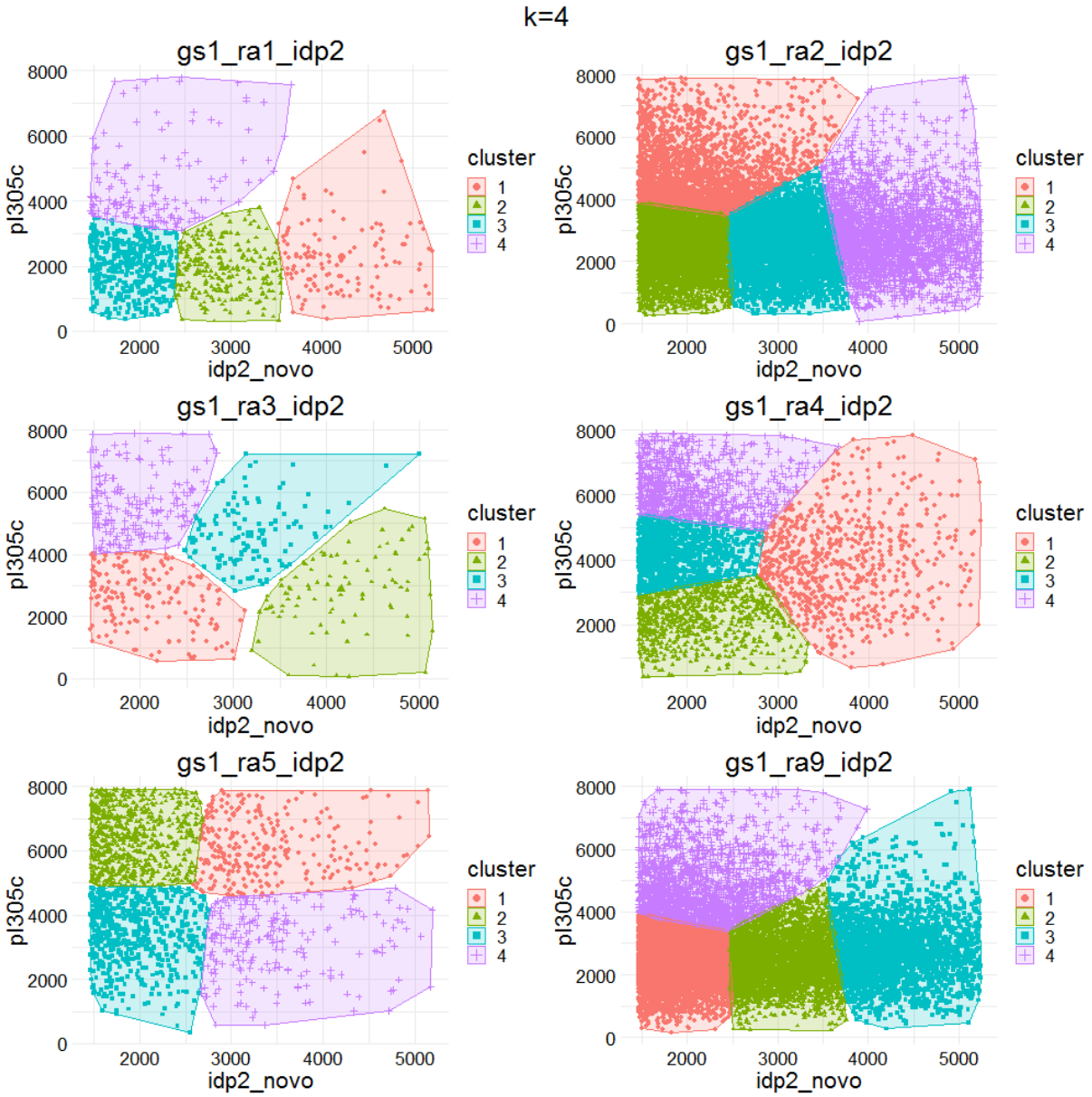

Figura 4.7: Agrupamento dos subconjuntos pelo KM com $k = 4$ utilizando a combinação de atributos (idp2_novo, pl305c)

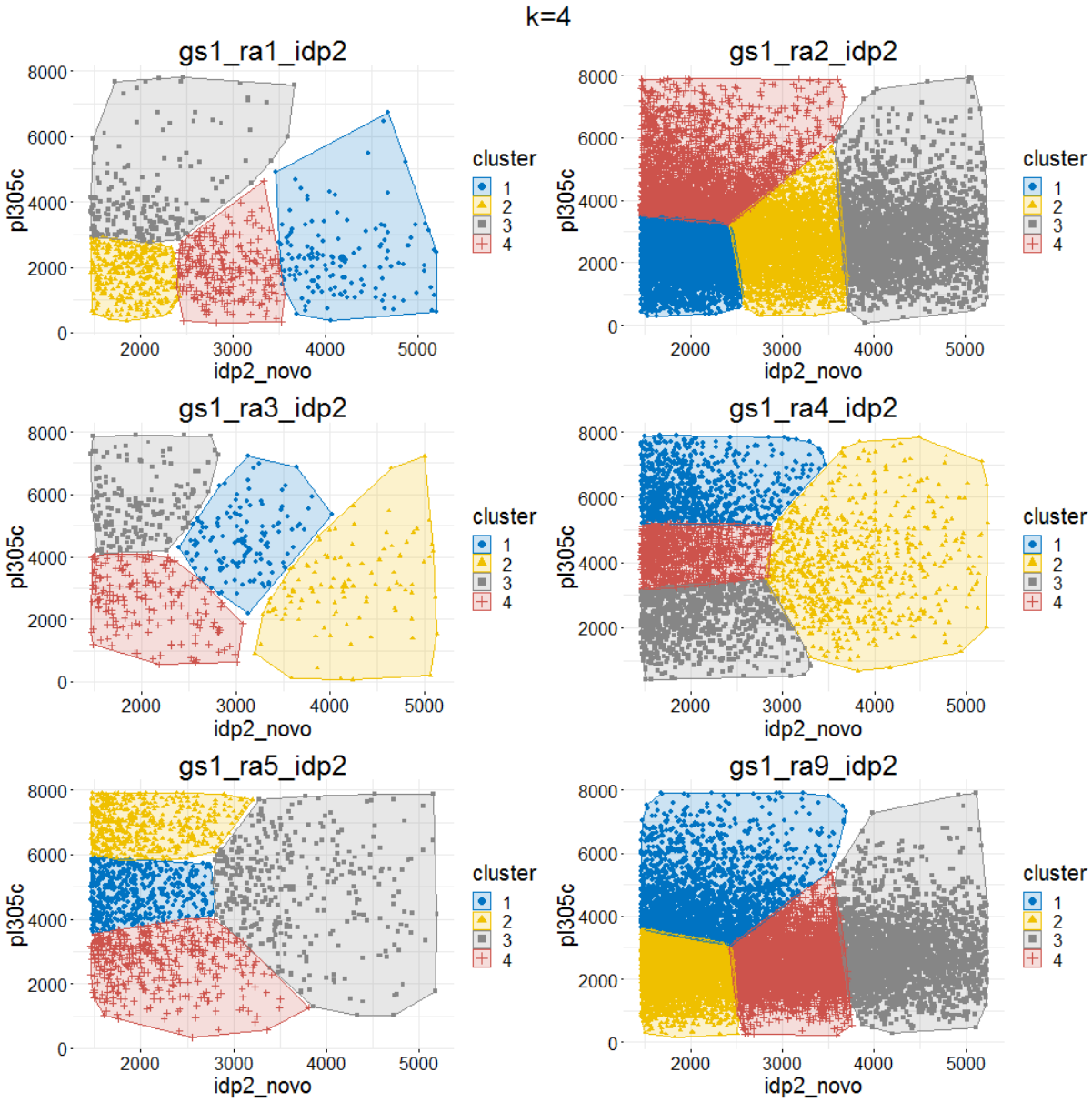

Figura 4.8: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp2_novo, pl305c)

Nos agrupamentos dos subconjuntos para a primeira ordem de parto com 4 grupos (Figura 4.5), viu-se que, no KM, ambas as variáveis, idp1_novo e pl305c, utilizadas como referência para o particionamento dos mesmos, apresentaram praticamente a mesma influência, dada a formação de regiões ortogonais, menos evidente no regime alimentar 1. Comparando esses resultados com os agrupamentos da Figura 4.6, observou-se maiores divergências entre os métodos na constituição dos grupos, sendo que as mudanças mais notáveis ocorreram nos subconjuntos com regimes alimentares 4 e 5. Além disso, os agrupamentos com 4 grupos através do FCM permitiram observar uma semelhança

da influência das variáveis na divisão dos subconjuntos com regimes alimentares 2 e 9, e parcialmente entre aqueles com regimes 4 e 5.

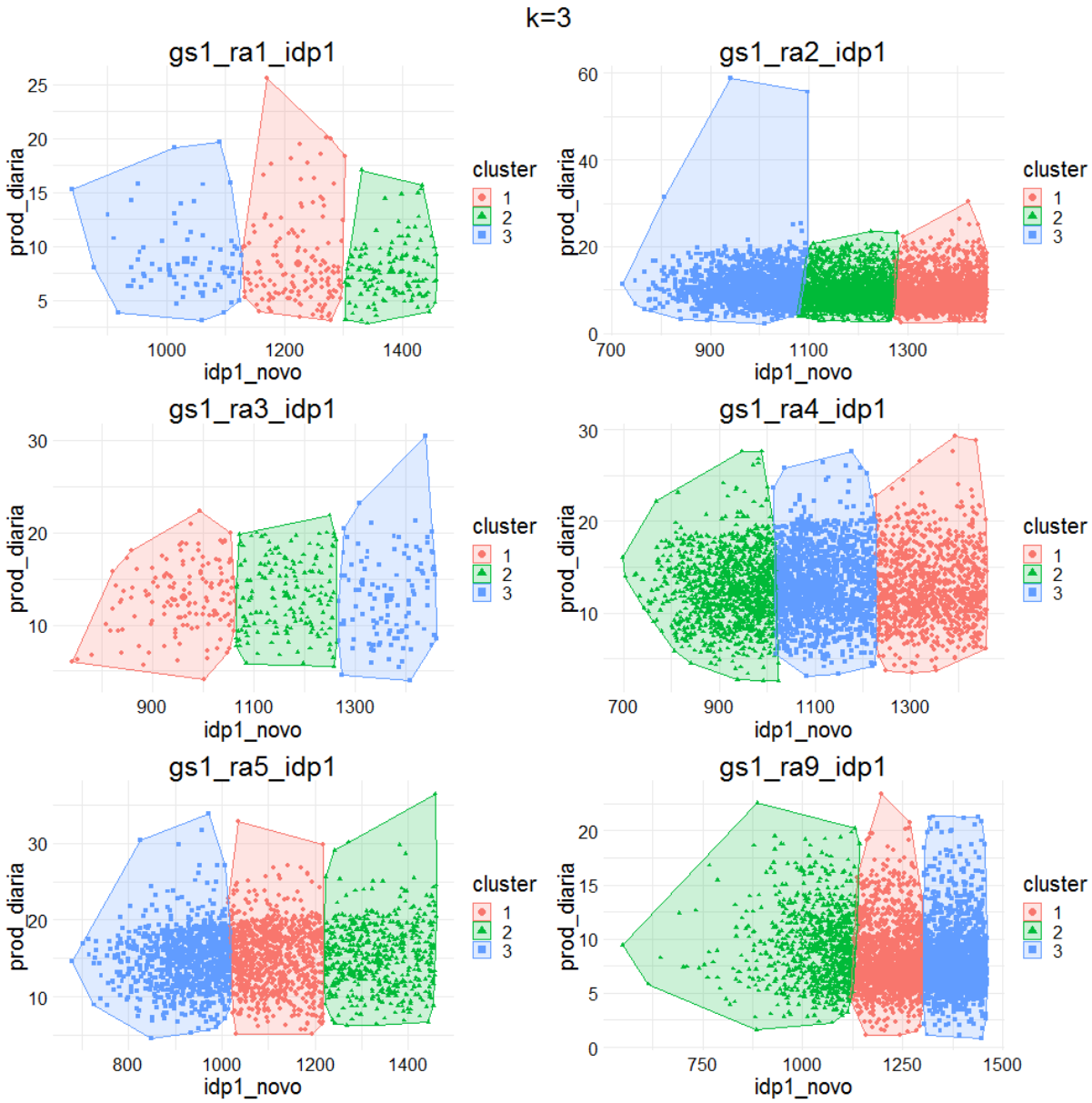

Figura 4.9: Agrupamento dos subconjuntos pelo KM com $k = 3$ utilizando a combinação de atributos (idp1_novo, prod_diaria)

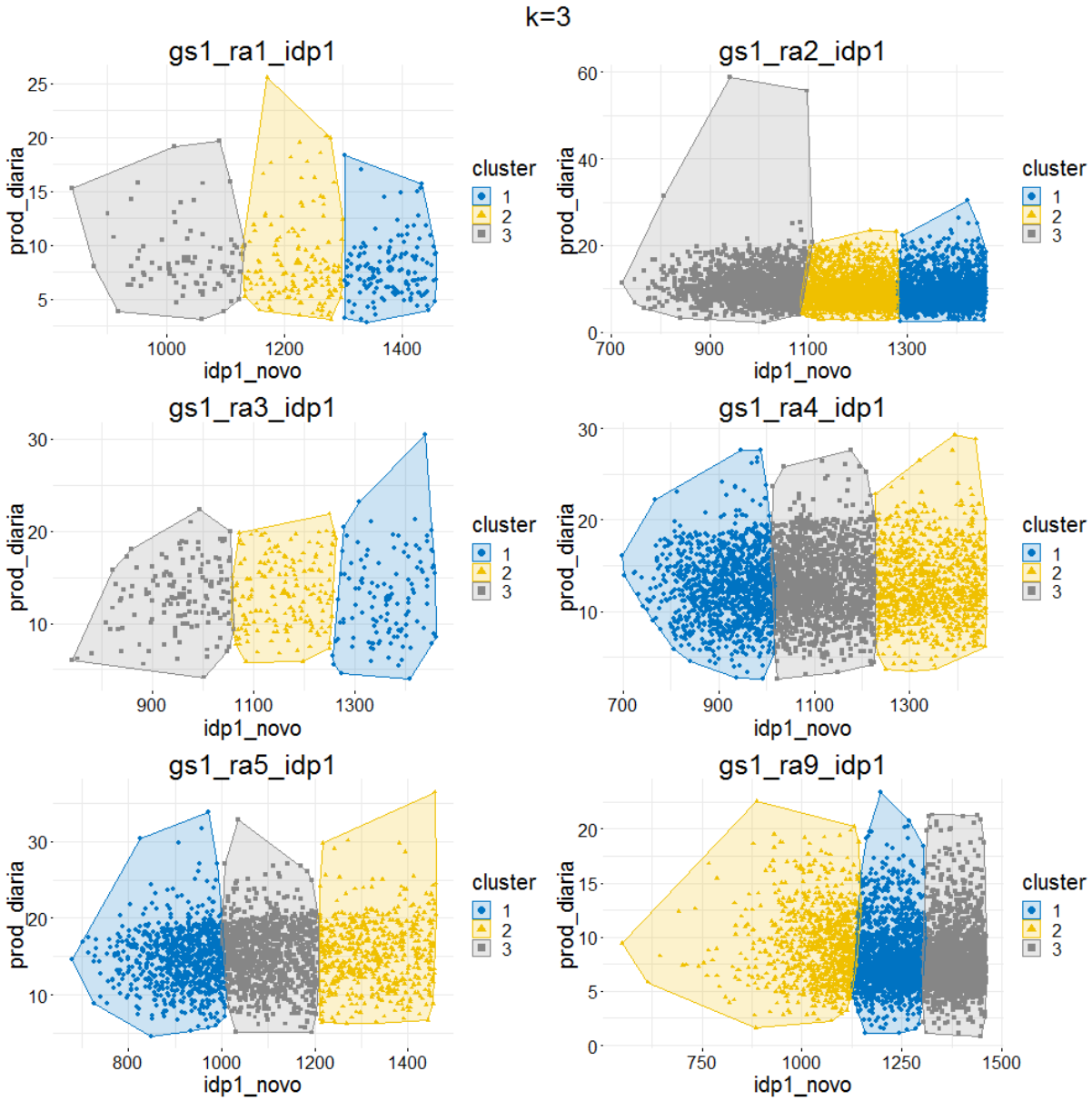

Figura 4.10: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp1_novo, prod_diaria)

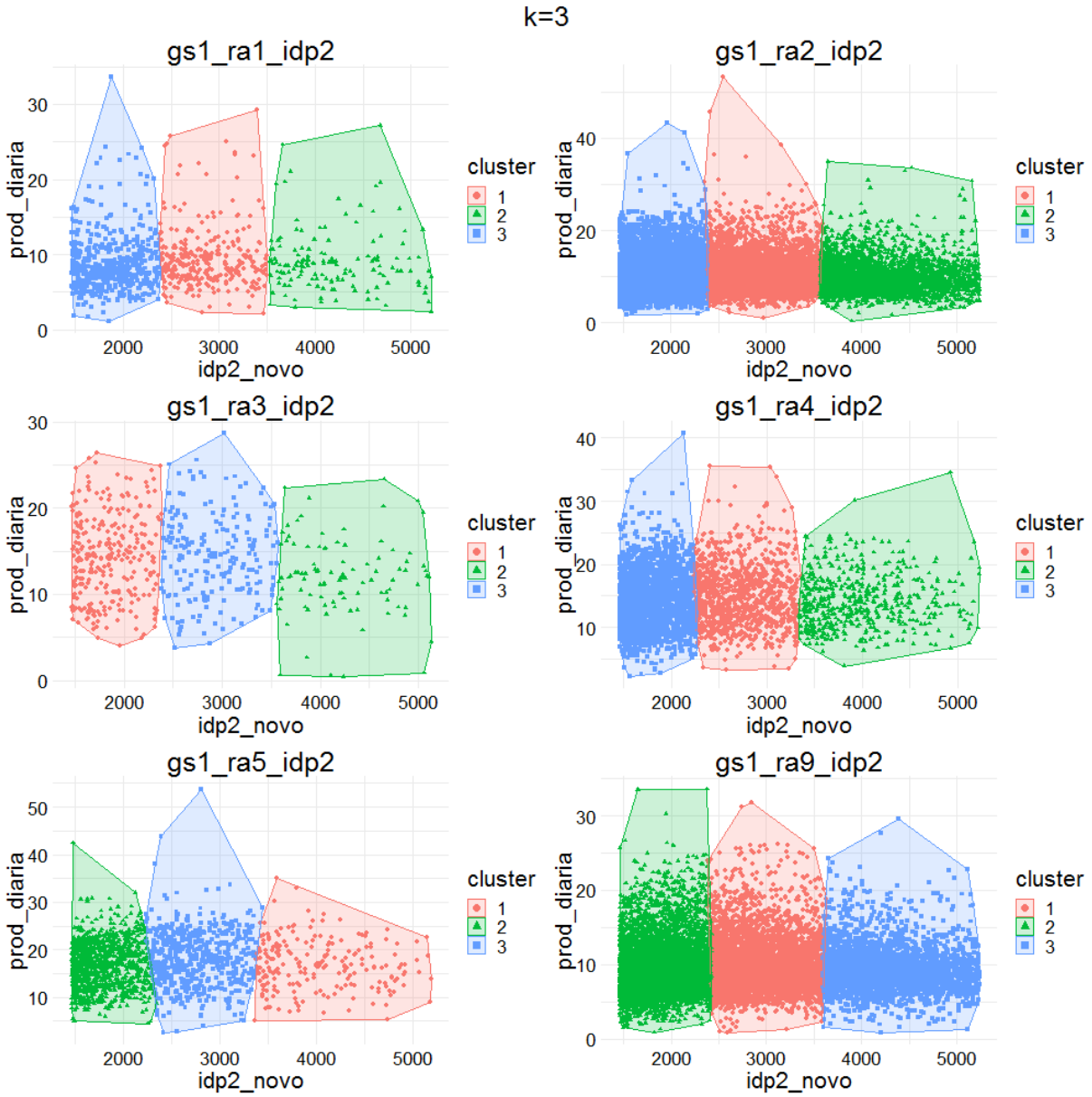

Figura 4.11: Agrupamento dos subconjuntos pelo KM com $k = 3$ utilizando a combinação de atributos (idp2_novo, prod_diaria)

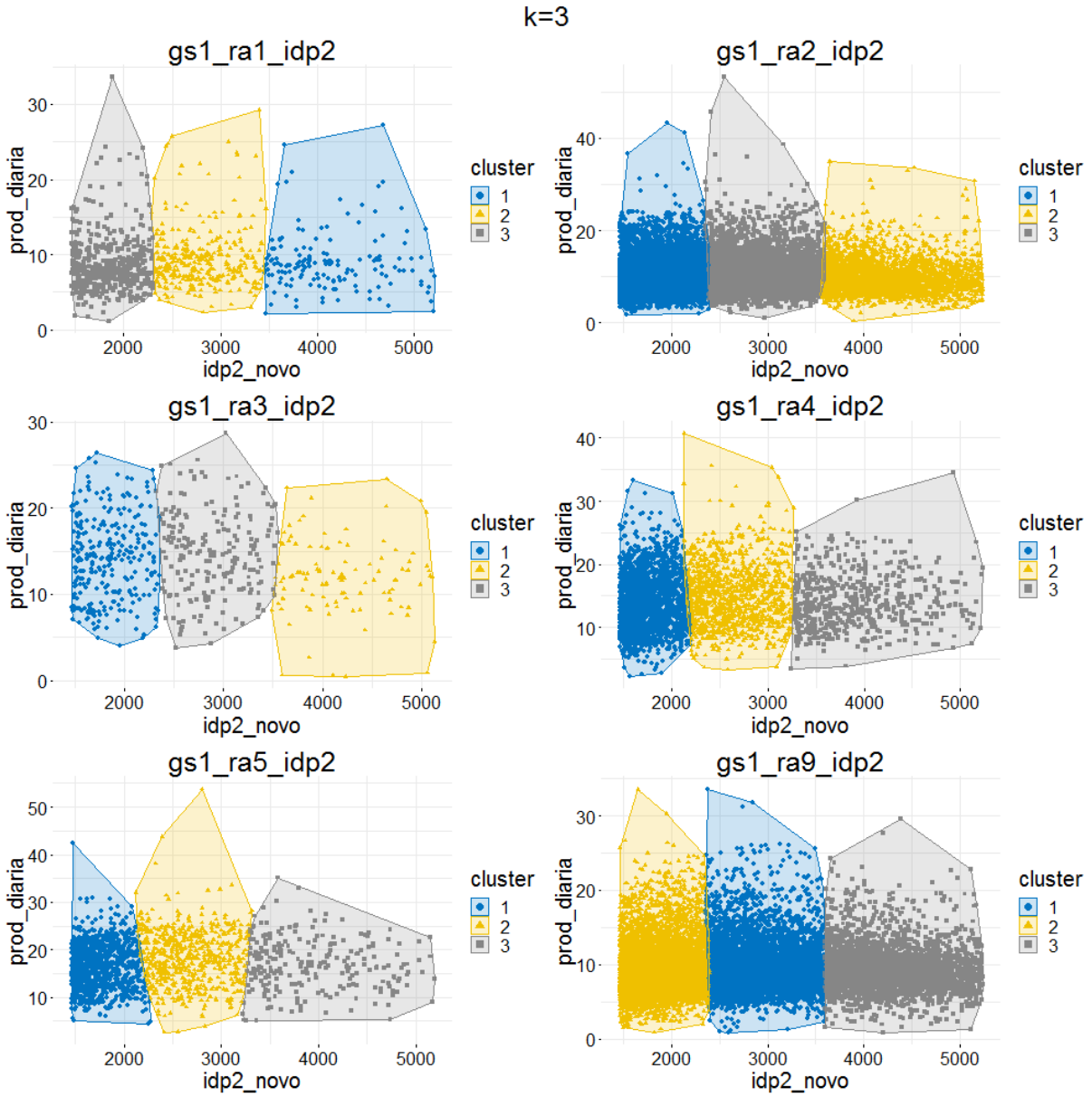

Figura 4.12: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp2_novo, prod_diaria)

Com a substituição do atributo pl305c por prod_diaria nos agrupamentos dos subconjuntos com 3 grupos utilizando-a juntamente com as ordens de parto, foi possível notar que, em ambos os métodos, essa nova váriavel, derivada da média entre pl305 e dlac, não exerceu qualquer influência na divisão dos subconjuntos, sendo esta realizada somente relativas às ordens de parto. Adicionalmente, as diferenças entre as fronteiras dos grupos, como também entre o deslocamento dos centroides, pelos métodos foram ínfimas, quase inexistentes.

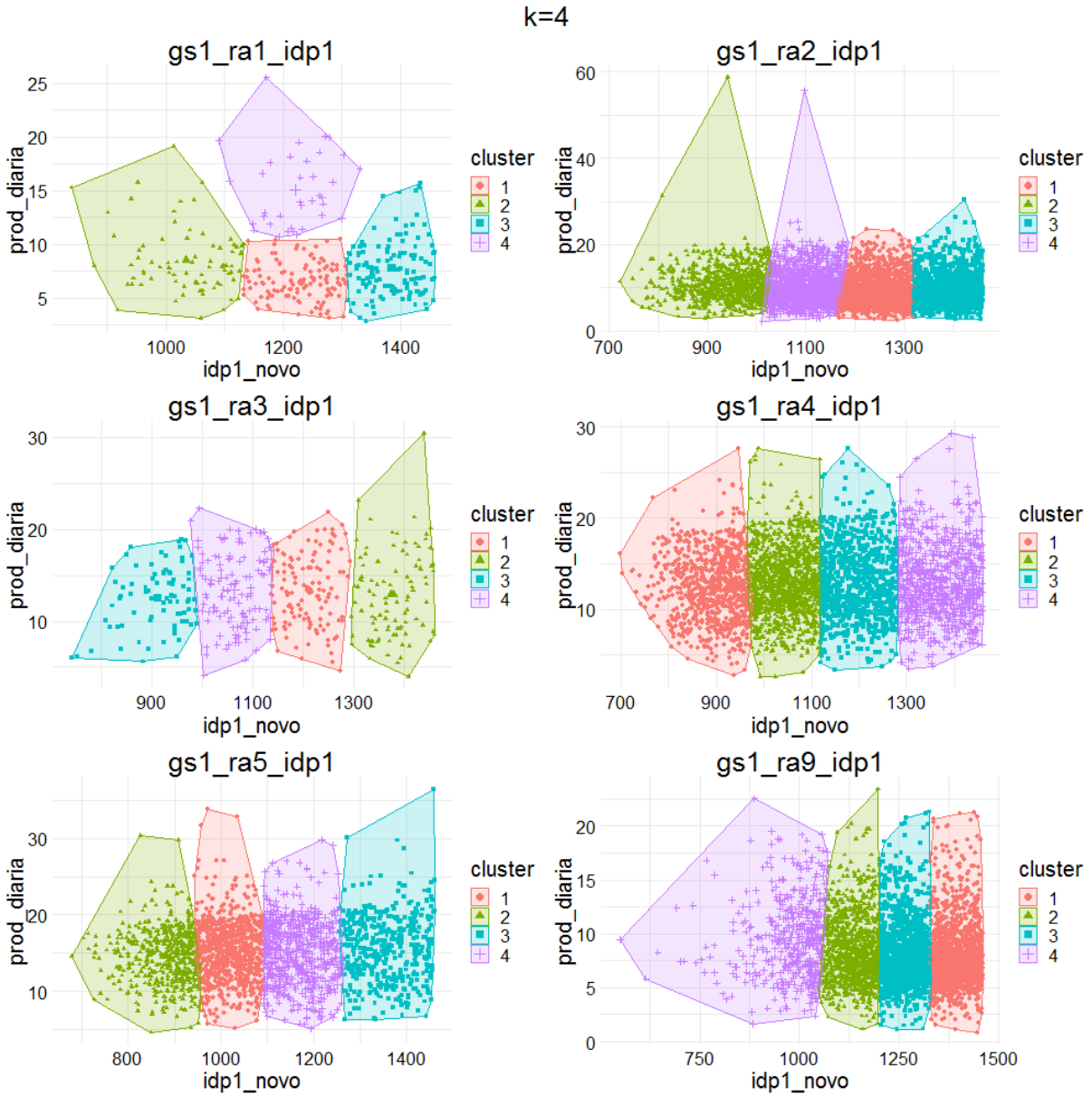

Figura 4.13: Agrupamento dos subconjuntos pelo KM com $k = 4$ utilizando a combinação de atributos (idp1_novo, prod_diaria)

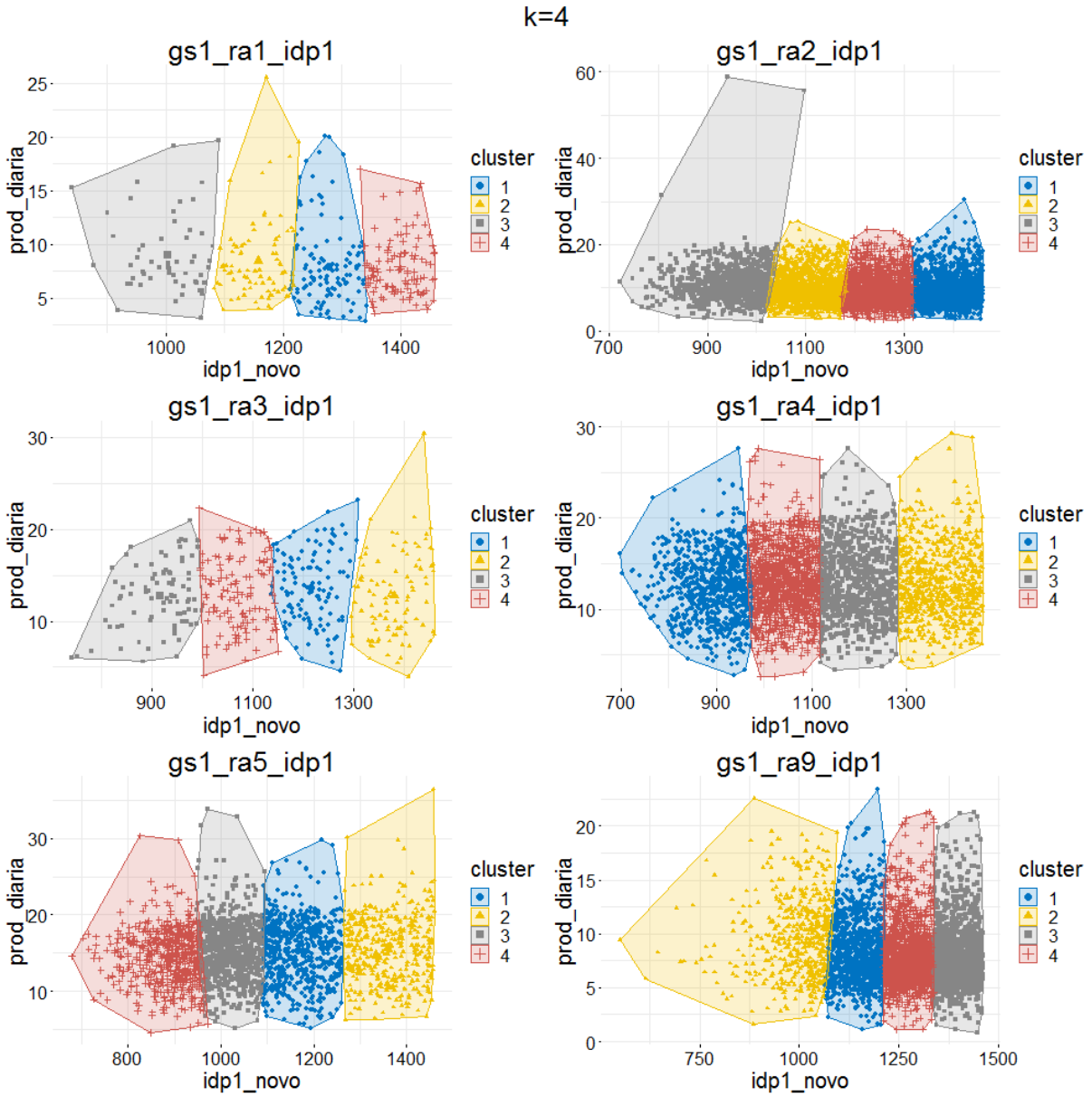

Figura 4.14: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp1_novo, prod_diaria)

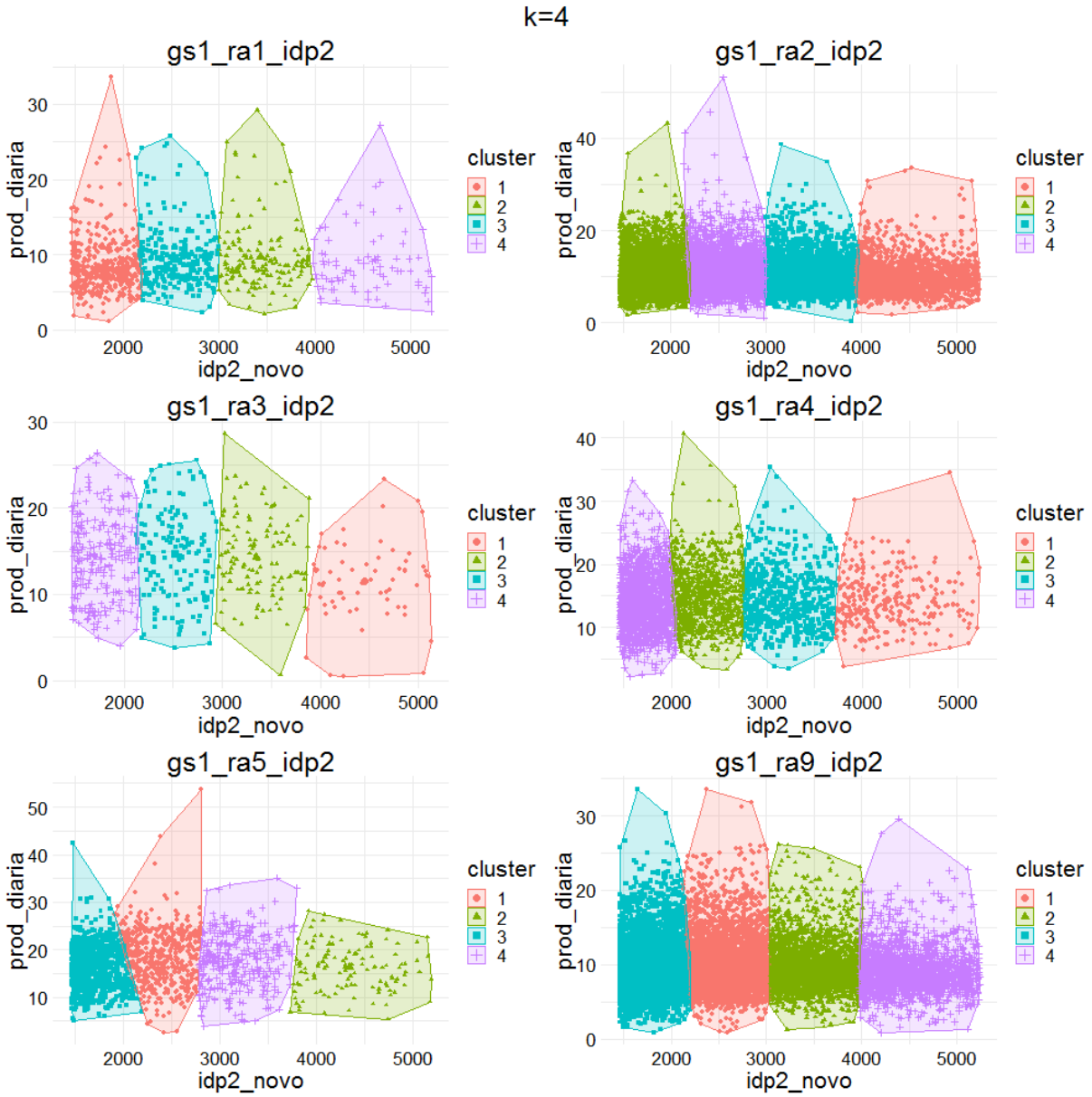

Figura 4.15: Agrupamento dos subconjuntos pelo KM com $k = 4$ utilizando a combinação de atributos (idp2_novo, prod_diaria)

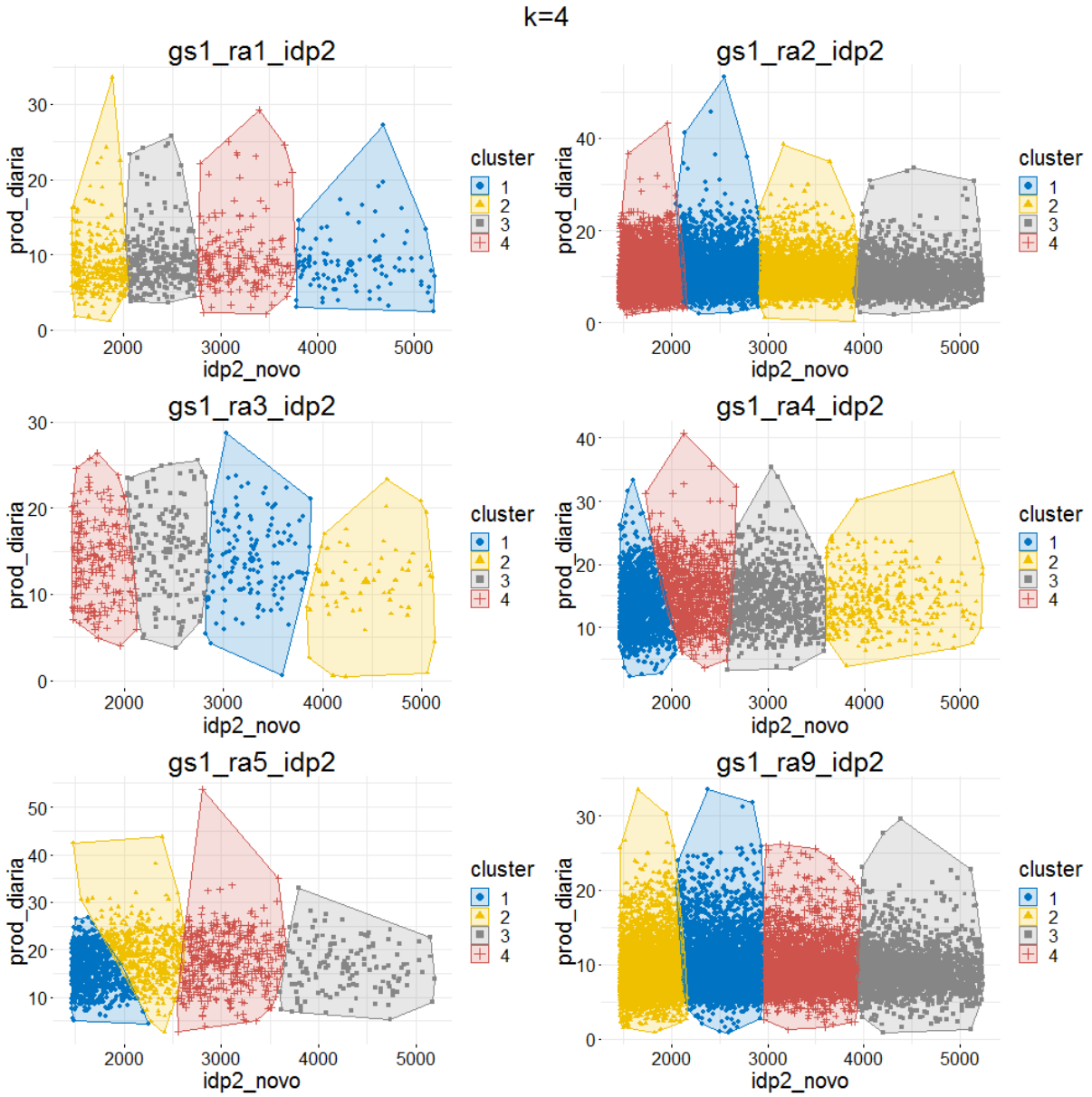

Figura 4.16: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp2_novo, prod_diaria)

De igual forma, o efeito da troca da variável pl305c por prod_diaria no agrupamento do subconjuntos com 4 grupos foi o mesmo, de modo que prod_diaria não fez distinção entre os grupos, exceto para o agrupamento com o KM do subconjunto da primeira ordem de parto e regime alimentar 1 (Figura 4.13). Além disso, houve também pequenas diferenças entre os métodos. Como esse atributo não produziu resultados relevantes com o uso dos algoritmos, decidiu-se eliminá-lo.

Foram também realizados os agrupamentos com o valor limite do *range* de k, i.e. $k = 7$, com as combinações (idp1_novo, pl305c) e (idp2_novo, pl305c), cujos grafícos são

mostrados nas Figuras 4.17 e 4.18; e em 4.19 e 4.20, respectivamente.

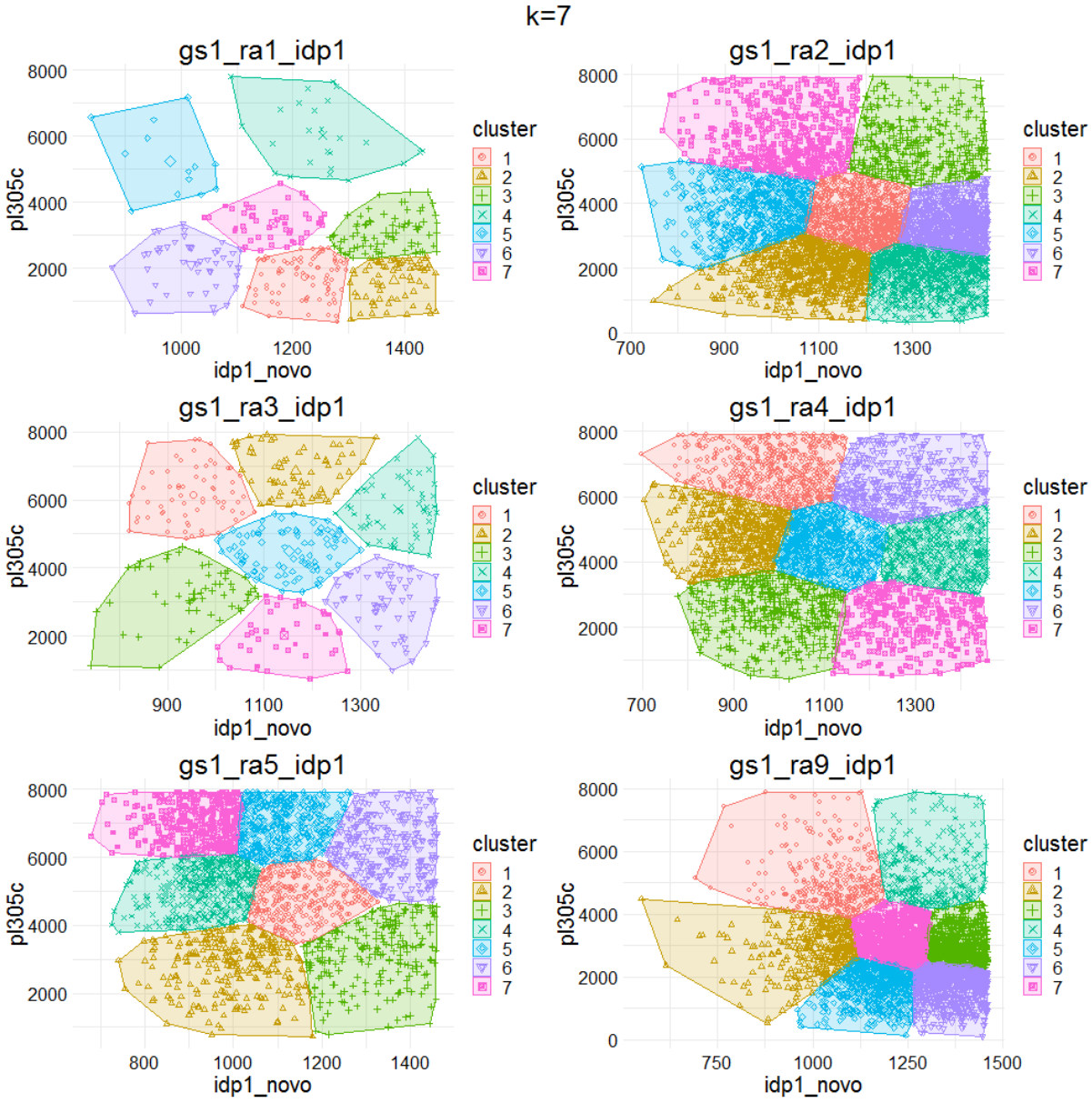

Figura 4.17: Agrupamento dos subconjuntos pelo KM com $k = 7$ utilizando a combinação de atributos (idp1_novo, pl305c)

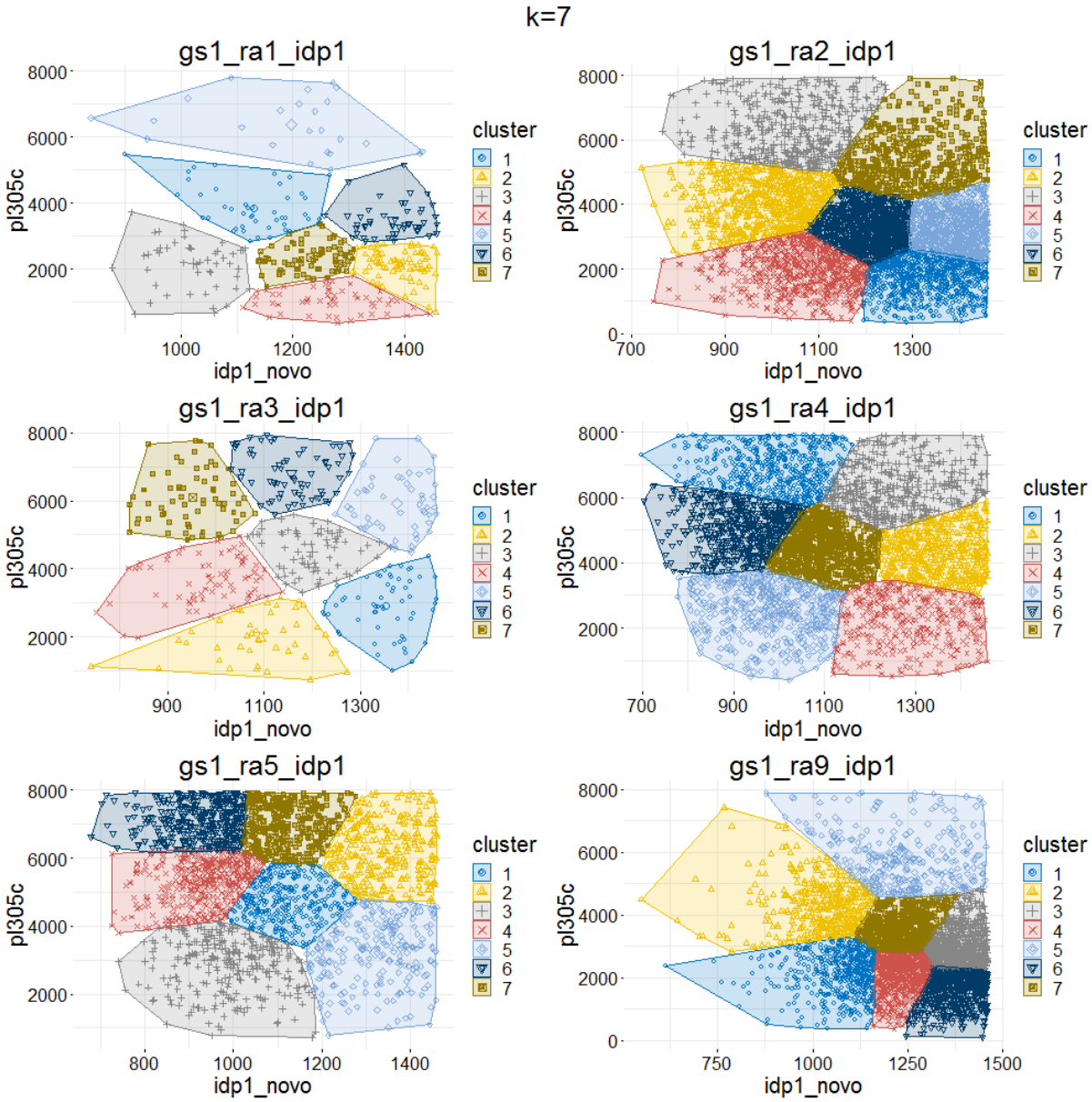

Figura 4.18: Agrupamento dos subconjuntos pelo FCM com $k = 7$ utilizando a combinação de atributos (idp1_novo, pl305c)

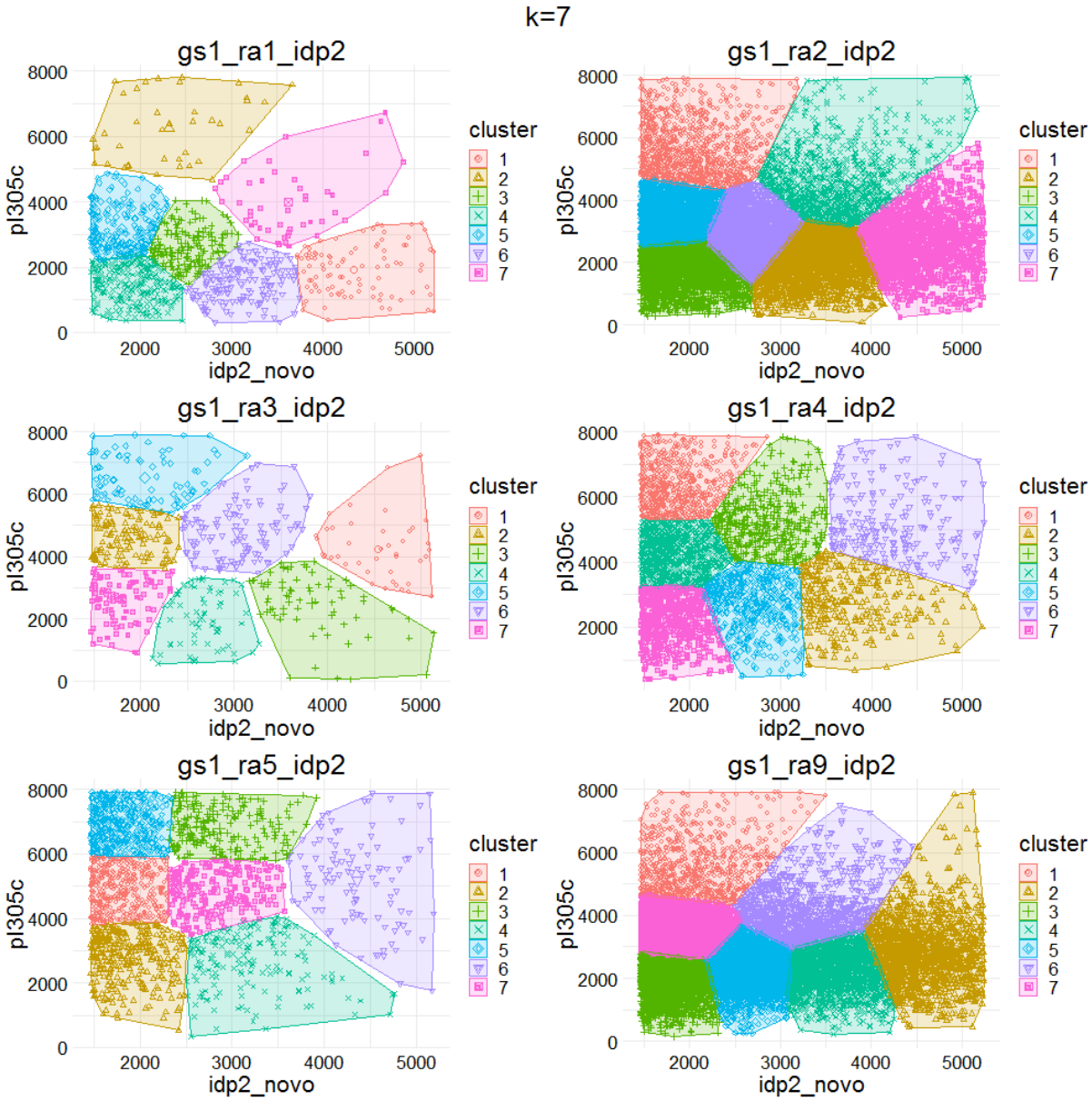

Figura 4.19: Agrupamento dos subconjuntos pelo KM com $k = 7$ utilizando a combinação de atributos (idp2_novo, pl305c)

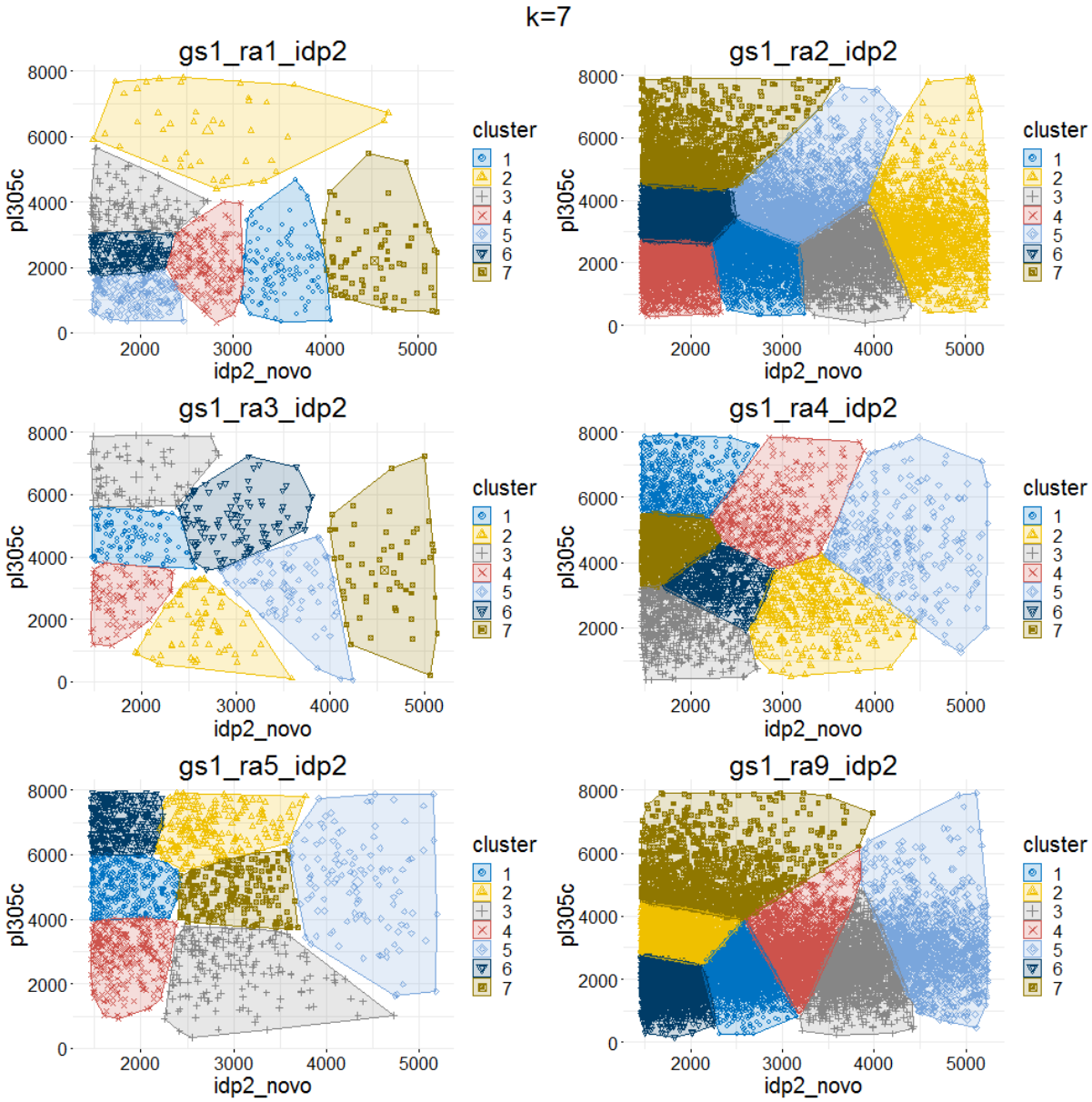

Figura 4.20: Agrupamento dos subconjuntos pelo FCM com $k = 7$ utilizando a combinação de atributos (idp2_novo, pl305c)

Com 7 grupos, pode-se verificar o surgimento de uma região central em todos os subconjuntos da primeira ordem de parto. Em quaisquer ordens de parto, as variáveis continuaram tendo influência no processo de formação dos grupos. Além disso, notou-se diferenças mais aparentes em todos os subconjuntos.

Consoante com o que fora verificado para todos os valores de k utilizados, as fronteiras entre os grupos estão próximas das de suas vizinhanças. Logo, independente de k, constatou-se que há regiões de sobreposição dos grupos, devido às proximidades entre as fronteiras dos mesmos. Consequentemente, os objetos em seu entorno não estão bem

definidos nos grupos aos quais eles foram atribuídos.

Pela análise da Tabela 4.17, definiu-se intervalos para o conjunto Ω de modo a permitir a observação de diferentes regiões dentro de cada grupo para separação dos objetos em relação ao valor de pertinência aos próprios grupos nos quais eles foram atribuídos. Sendo assim, foi possível identificar graficamente tanto as áreas em que os objetos estão bem definidos nos grupos, quanto as regiões que contém objetos com baixa pertinência, i.e., os objetos não muito bem agrupados. Além destas, também se identifou as áreas que delimitam os objetos parcialmente definidos nos clusters. Para isso, considerou-se as seguintes faixas de valores para Ω:

- ◆ $0.75 \leq \Omega_i \leq 1 \Leftrightarrow$ objeto bem definido no grupo
- ◆ $0.5 \leq \Omega_i < 0.75 \Leftrightarrow$ objeto parcialmente definido no grupo
- ◆ $0.25 \leq \Omega_i < 0.5 \Leftrightarrow$ objeto fracamente definido no grupo
- ◆ $\Omega_i < 0.25 \Leftrightarrow$ objeto mal definido no grupo

Onde Ω_i = Valor máximo da i-ésima linha da matriz de pertinência $\mu = [u_{ij}]_{n \times k}$, i.e., é o maior grau de pertinência de um objeto $x_i \in \mathrm{X} = \{x_1, ..., x_n\}$ a um dado cluster $p_j \in \mathrm{P} = \{p_1, ..., p_k\}$ que, consequentemente, representa a pertinência ao cluster no qual ele foi inserido pelo FCM. Com base nesses valores, plotou-se novamente os gráficos dos agrupamentos com $k = 3$, $k = 4$ e também para $k = 7$, utilizando as combinações de atributos (idp1_novo, pl305c) e (idp2_novo, pl305c). As variações de cores dentro de cada grupo correspondem aos intervalos para Ω, anteriormente definidos.

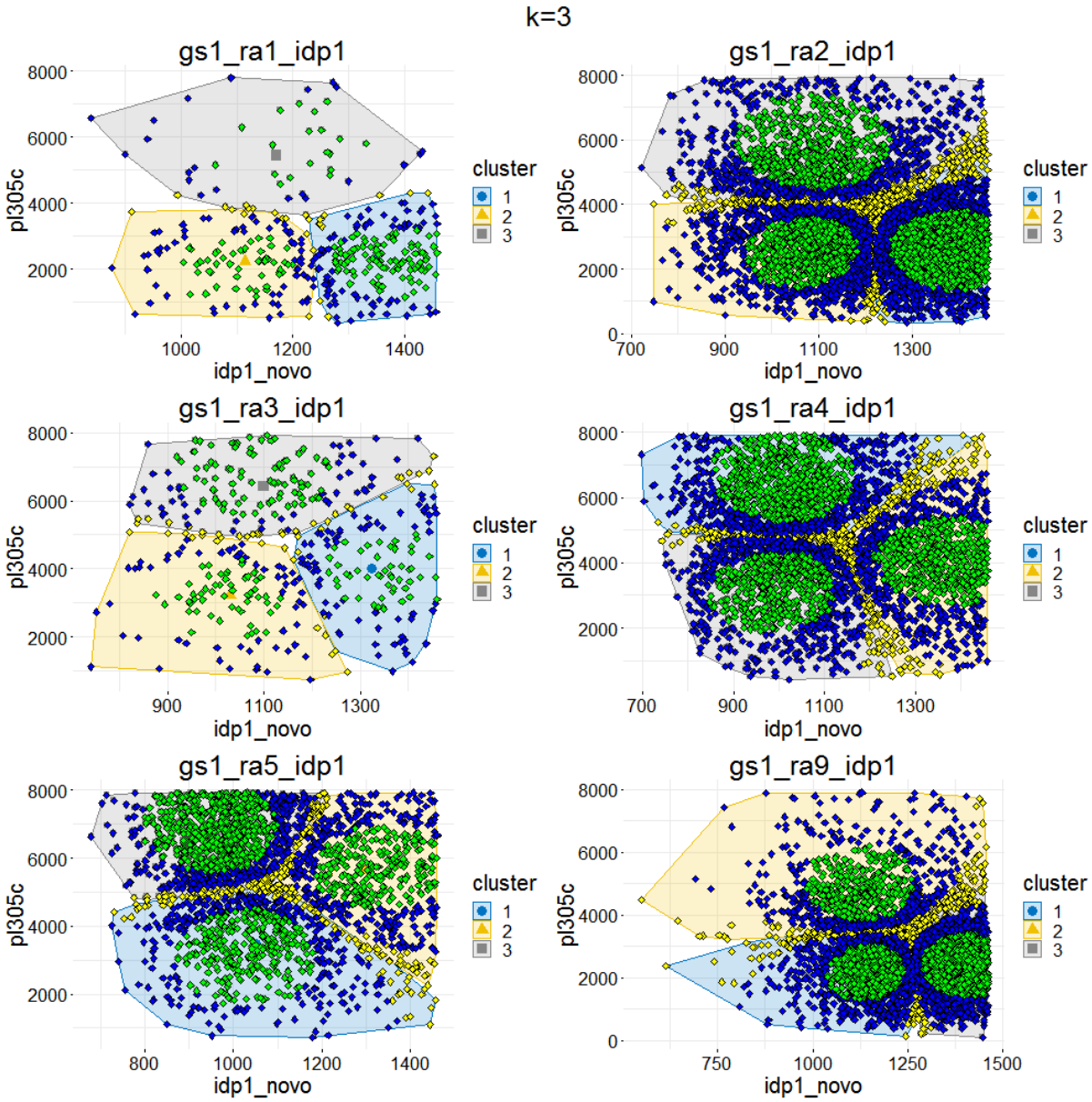

Figura 4.21: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp1_novo, pl305c), com cada objeto colorido por Ω_i

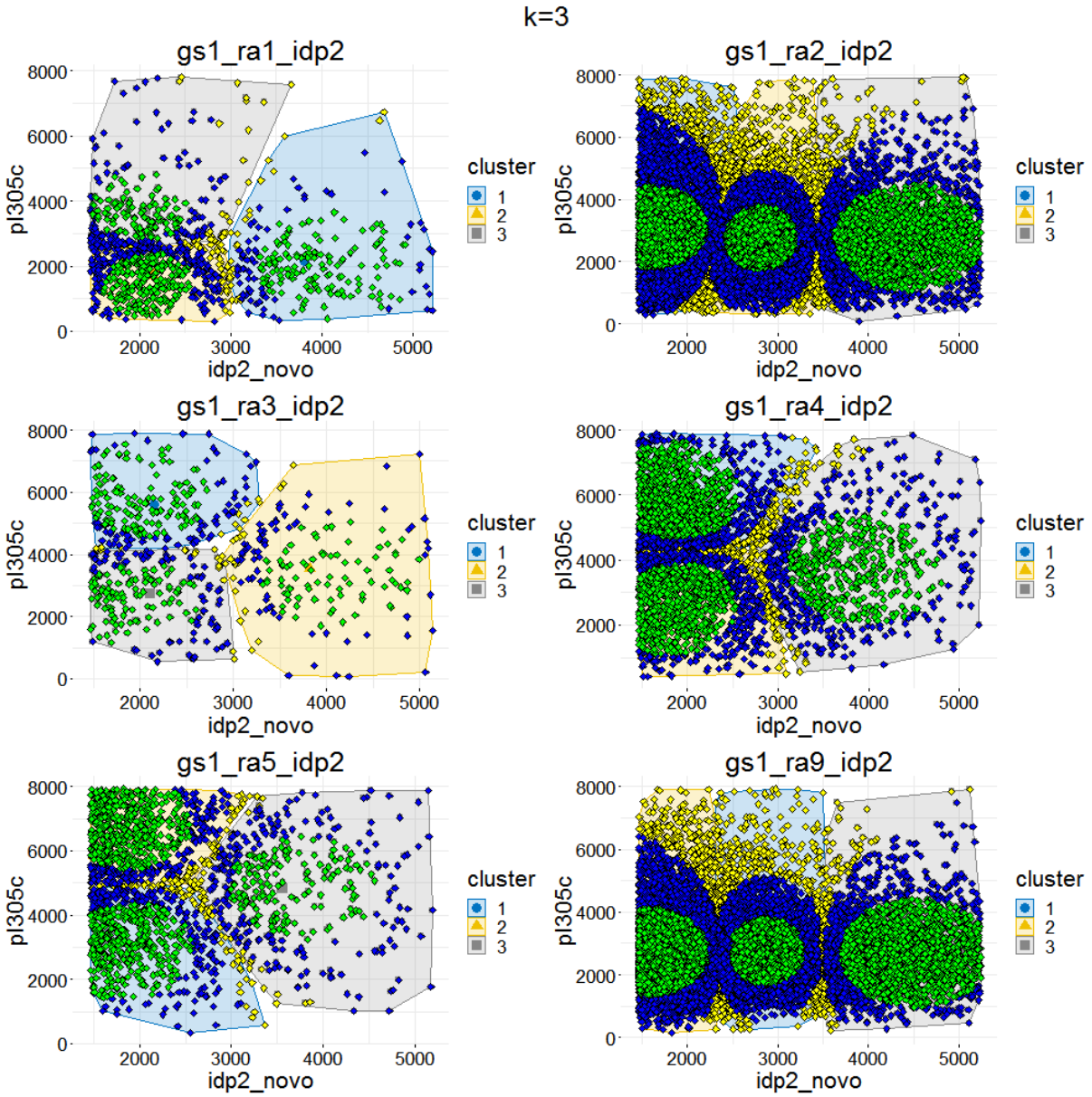

Figura 4.22: Agrupamento dos subconjuntos pelo FCM com $k = 3$ utilizando a combinação de atributos (idp2_novo, pl305c), com cada objeto colorido por Ω_i

A partir das Figuras 4.21 e 4.22 notou-se uma semelhança da disposição das áreas em verde dos grupos de subconjuntos com objetos de regimes alimentares 2 e 9. Observou-se também uma similaridade parcial entre as regiões bem definidas dos grupos dos subconjuntos com ra $= 4$ e ra $= 5$ contendo idp2_novo. A não existência de pontos na cor vermelha nos gráficos se deu por conta dos valores mínimos de Ω para $k = 3$, como constados na Tabela 4.17, dado que eles foram maiores que 0.25.

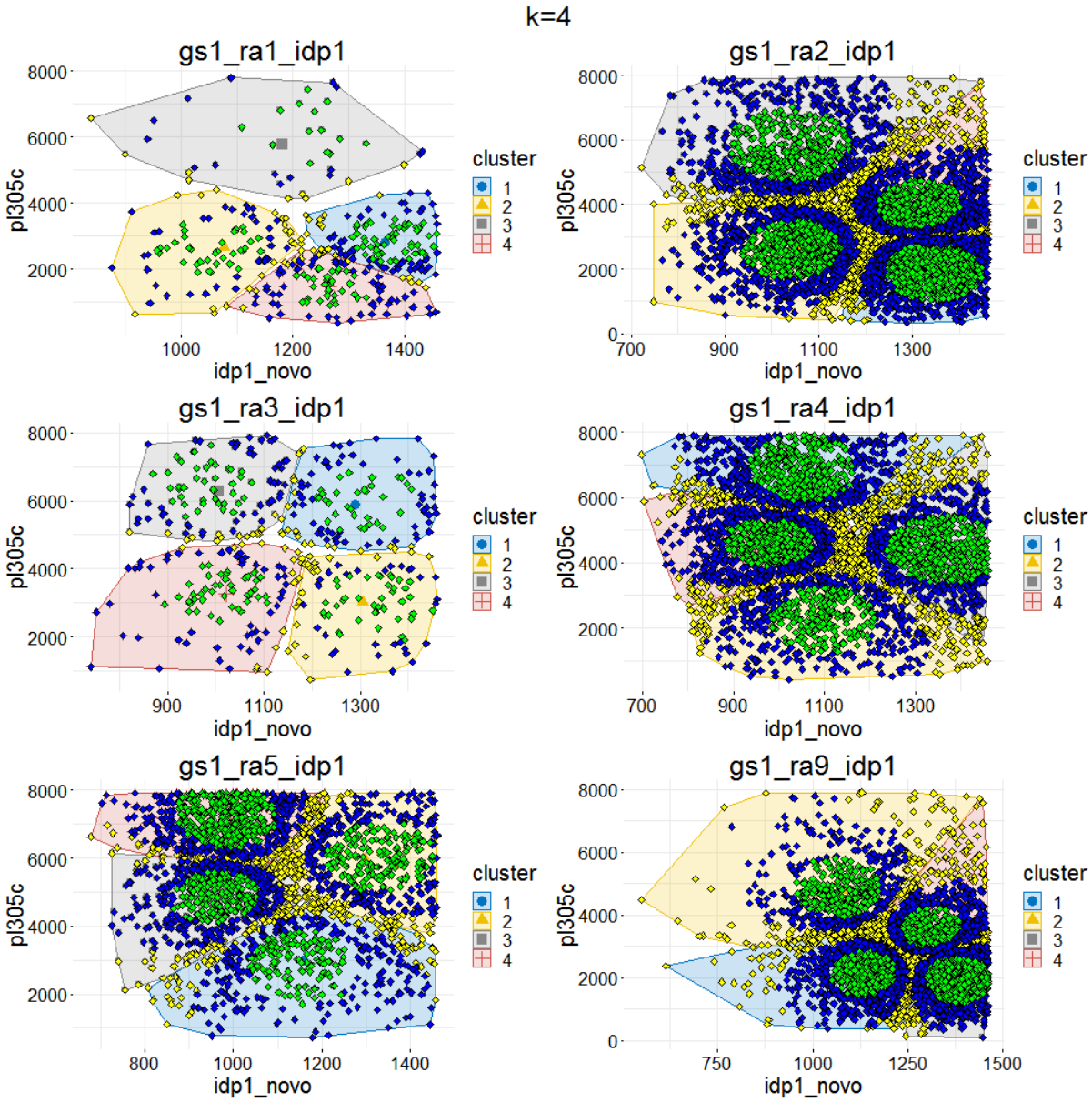

Figura 4.23: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp1_novo, pl305c), com cada objeto colorido por Ω_i

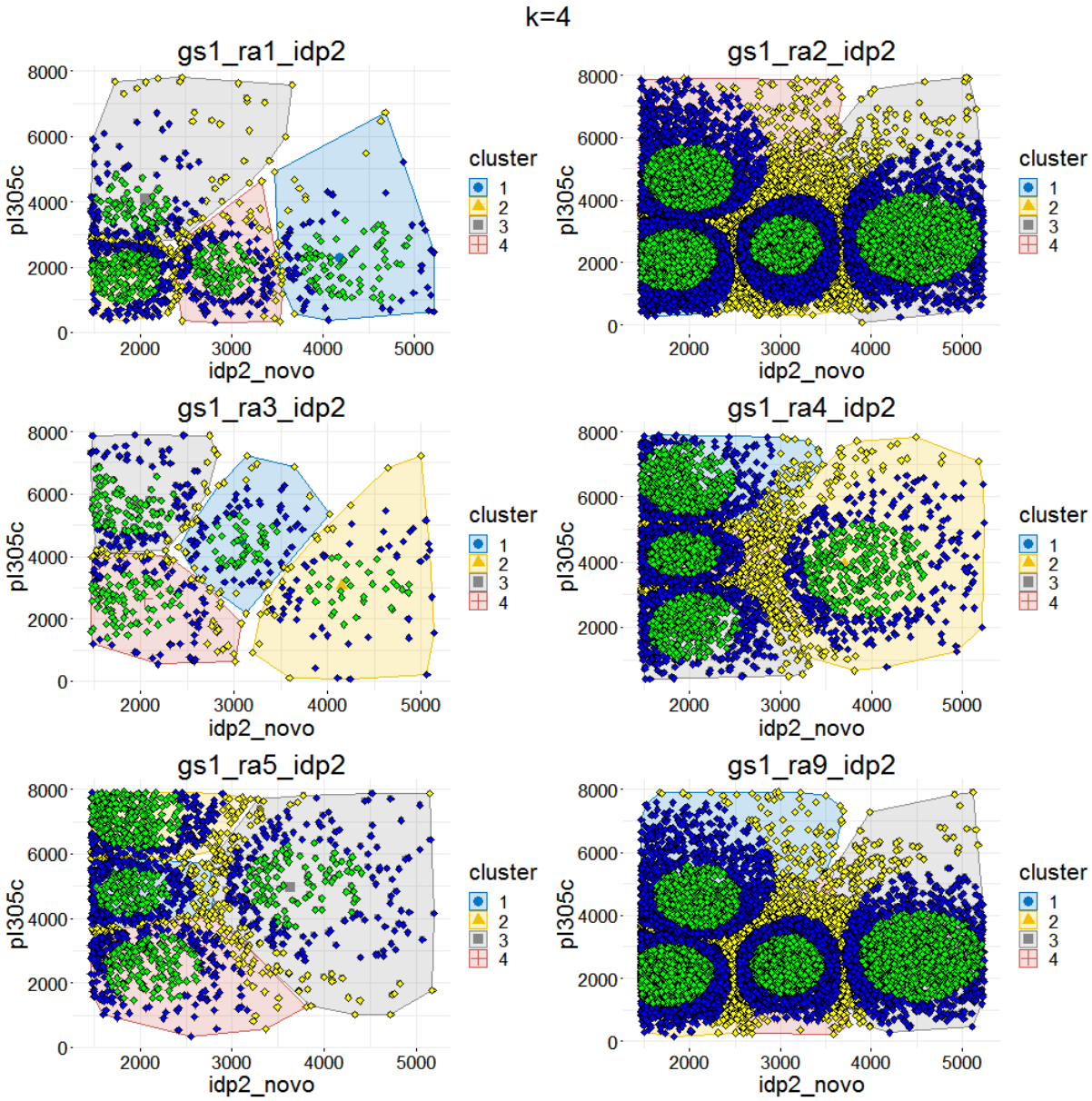

Figura 4.24: Agrupamento dos subconjuntos pelo FCM com $k = 4$ utilizando a combinação de atributos (idp2_novo, pl305c), com cada objeto colorido por Ω_i

As semelhanças entre os suconjuntos com regimes 2 e 9 foram preservadas nos agrupamentos com 4 grupos. Entretanto, houve uma diferenciação dos objetos bem definidos em um dos grupos com ra = 4 e ra = 5 na primeira ordem de parto (idp1_novo) em relação a produção. Todavia, notou-se um padrão de comportamento tanto entre os subconjuntos com regimes 2 e 9 quanto entre os de regime 4 e 5 em relação à formação dos grupos pela variação de k de 3 para 4 grupos. Da mesma forma como nos agrupamentos com 3 grupos, não se observou objetos na cor vermelha nas Figuras 4.23 e 4.24, dado que, para todo subconjunto, $\min(\Omega_{k=4}) > 0.25$, conforme visto na Tabela 4.17. Para ambos

os valores de k, pôde-se notar que, como esperado, os pontos na cor amarela se localizam próximos às fronteiras dos grupos ou estão mais distantes dos centroides.

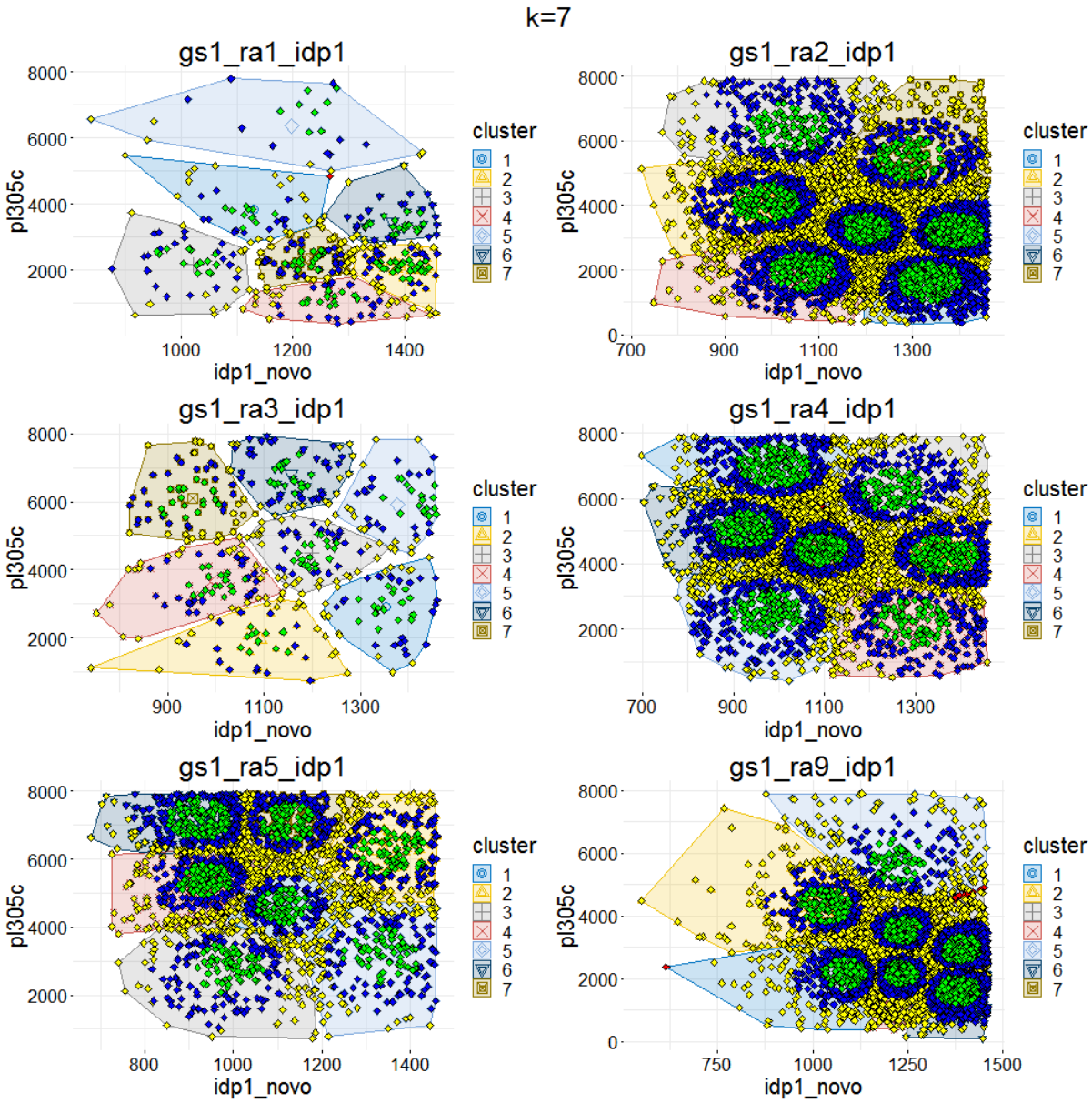

Figura 4.25: Agrupamento dos subconjuntos pelo FCM com $k = 7$ utilizando a combinação de atributos (idp1_novo, pl305c), com cada objeto colorido por Ω_i

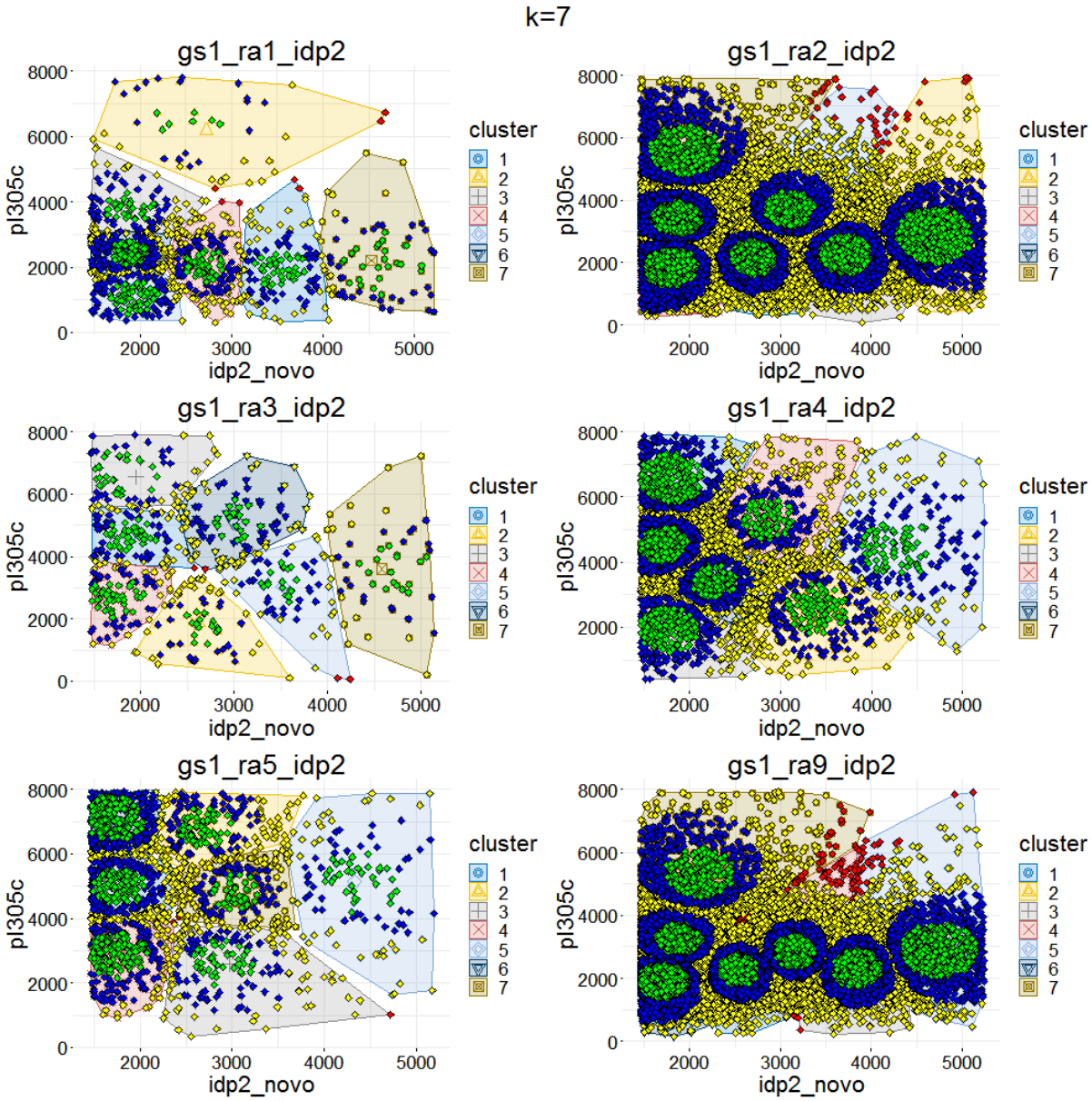

Figura 4.26: Agrupamento dos subconjuntos pelo FCM com $k = 7$ utilizando a combinação de atributos (idp2_novo, pl305c), com cada objeto colorido por Ω_i

Com 7 grupos, as relações ainda mantidas no agrupamento com $k = 4$ entre os subconjuntos de regimes alimentares 2 e 9 contendo idp2_novo tornaram-se menos evidente em um dos grupos. O aumento do número de grupos favoreceu o surgimento de pontos na cor vermelha, pois, conforme a Tabela 4.17, alguns dos subconjuntos apresentaram $\min(\Omega_{k=4}) \leq 0.25$, com maior incidência nos subconjuntos contendo idp2_novo.

A partir disso, constatou-se o aumento da precisão da avaliação, mas que, possivelmente, contém menos significância. Vale destacar que os pontos em vermelho identificados em alguns dos subconjuntos das Figuras 4.25 e 4.26 permitiram ver que eles se localizam

em regiões de fronteira ou estão a uma distância muito próxima a mais de um centroide, o que gera maior dúvida na atribuição do objeto ao grupo. Para todos os valores de k, notou-se que as áreas em verde constituem os núcleos de cada grupo nas quais os objetos por elas delimitados estão bem definidos nos mesmos.

No intuito de se observar se a correlação existente entre pl305c e dlac na amostra, cujo valor foi de 0.521, se manteve nos subconjuntos, computou-se esses valores para cada um deles, mostrados na Tabela 4.19.

	Correlação entre pl305c e dlac
gs1_ra1_idp1	0.624
gs1_ra2_idp1	0.622
gs1_ra3_idp1	0.633
gs1_ra4_idp1	0.637
gs1_ra5_idp1	0.618
gs1_ra9_idp1	0.552
gs1_ra1_idp2	0.614
gs1_ra2_idp2	0.613
gs1_ra3_idp2	0.579
gs1_ra4_idp2	0.646
gs1_ra5_idp2	0.609
gs1_ra9_idp2	0.509

Tabela 4.19: Valores de correlação entre os atributos pl305c e dlac para cada subconjunto.

Verificou-se que, nitidamente, houve uma elevação do valor da correlação na maioria dos subconjuntos, excetuando-se para gs1_ra9_idp2. Além disso, viu-se que, dentre os subconjuntos restantes, gs1_ra9_idp2 e gs1_ra3_idp2 foram os que apresentaram aumento menos significativo em relação aos demais.

4.3 Experimento 3: Análise da tendência de agrupamento dos dados

Os dados em cada subconjunto também foram avaliados quanto à tendência de agrupamento, evitando assim suposições equivocadas acerca dos resultados obtidos. Realizou-se

o cômputo da estatística de Hopkins separadamente para cada um deles utilizando diferentes combinações de atributos como forma de verificar os seus efeitos no valor da estatística. As tabelas a seguir exibem o valor de H ao longo de 6 execuções distintas para os subconjuntos. Como o valor da estatística foi computado conforme a equação 2.21, quanto mais próximo de 0, provavelmente, maior a tendência de agrupamento.

	n execução	**H (idp1_novo e pl305c)**	**H (idp1_novo, pl305c e dlac)**	**H (idp1_novo, pl305 e dlac)**
gs1_ra1_idp1	**1 (seed=25)**	0.324	0.294	0.294
	2 (seed=26)	0.286	0.239	0.242
	3 (seed=27)	0.282	0.268	0.266
	4 (seed=28)	0.338	0.267	0.267
	5 (seed=29)	0.351	0.254	0.256
	6 (seed=30)	0.364	0.256	0.262
	media	**0.324**	**0.263**	**0.264**
	min	**0.282**	**0.239**	**0.242**
	max	**0.364**	**0.294**	**0.294**

Tabela 4.20: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra1_idp1 com diferentes combinações de atributos.

	n execução	H (idp1_novo e pl305c)	H (idp1_novo, pl305c e dlac)	H (idp1_novo, pl305 e dlac)
gs1_ra2_idp1	**1 (seed=25)**	0.311	0.227	0.219
	2 (seed=26)	0.314	0.226	0.217
	3 (seed=27)	0.309	0.218	0.212
	4 (seed=28)	0.294	0.24	0.228
	5 (seed=29)	0.273	0.231	0.223
	6 (seed=30)	0.295	0.243	0.236
	media	**0.299**	**0.231**	**0.223**
	min	**0.273**	**0.218**	**0.212**
	max	**0.314**	**0.243**	**0.236**

Tabela 4.21: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra2_idp1 com diferentes combinações de atributos.

	n execução	H (idp1_novo e pl305c)	H (idp1_novo, pl305c e dlac)	H (idp1_novo, pl305 e dlac)
gs1_ra3_idp1	**1 (seed=25)**	0.41	0.345	0.336
	2 (seed=26)	0.389	0.332	0.326
	3 (seed=27)	0.404	0.356	0.352
	4 (seed=28)	0.448	0.346	0.339
	5 (seed=29)	0.426	0.371	0.356
	6 (seed=30)	0.423	0.33	0.327
	media	**0.416**	**0.347**	**0.339**
	min	**0.389**	**0.33**	**0.326**
	max	**0.448**	**0.371**	**0.356**

Tabela 4.22: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra3_idp1 com diferentes combinações de atributos.

	n execução	H (idp1_novo e pl305c)	H (idp1_novo, pl305c e dlac)	H (idp1_novo, pl305 e dlac)
gs1_ra4_idp1	**1 (seed=25)**	0.309	0.254	0.238
	2 (seed=26)	0.326	0.244	0.236
	3 (seed=27)	0.327	0.247	0.229
	4 (seed=28)	0.357	0.245	0.229
	5 (seed=29)	0.289	0.241	0.23
	6 (seed=30)	0.323	0.241	0.23
	media	**0.322**	**0.245**	**0.232**
	min	**0.289**	**0.241**	**0.229**
	max	**0.357**	**0.254**	**0.238**

Tabela 4.23: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra4_idp1 com diferentes combinações de atributos.

	n execução	H (idp1_novo e pl305c)	H (idp1_novo, pl305c e dlac)	H (idp1_novo, pl305 e dlac)
gs1_ra5_idp1	**1 (seed=25)**	0.34	0.255	0.256
	2 (seed=26)	0.351	0.246	0.237
	3 (seed=27)	0.317	0.24	0.231
	4 (seed=28)	0.336	0.254	0.248
	5 (seed=29)	0.273	0.252	0.25
	6 (seed=30)	0.27	0.246	0.24
	media	**0.315**	**0.249**	**0.244**
	min	**0.27**	**0.24**	**0.231**
	max	**0.351**	**0.255**	**0.256**

Tabela 4.24: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra5_idp1 com diferentes combinações de atributos.

	n execução	H (idp1_novo e pl305c)	H (idp1_novo, pl305c e dlac)	H (idp1_novo, pl305 e dlac)
gs1_ra9_idp1	**1 (seed=25)**	0.145	0.153	0.147
	2 (seed=26)	0.123	0.141	0.135
	3 (seed=27)	0.133	0.151	0.144
	4 (seed=28)	0.122	0.148	0.14
	5 (seed=29)	0.133	0.15	0.144
	6 (seed=30)	0.131	0.16	0.153
	media	**0.131**	**0.151**	**0.144**
	min	**0.122**	**0.141**	**0.135**
	max	**0.145**	**0.16**	**0.153**

Tabela 4.25: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra9_idp1 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra1_idp2	**1 (seed=25)**	0.295	0.265	0.258
	2 (seed=26)	0.299	0.241	0.235
	3 (seed=27)	0.248	0.236	0.227
	4 (seed=28)	0.246	0.26	0.252
	5 (seed=29)	0.277	0.244	0.236
	6 (seed=30)	0.274	0.229	0.218
	media	**0.273**	**0.246**	**0.238**
	min	**0.246**	**0.229**	**0.218**
	max	**0.299**	**0.265**	**0.258**

Tabela 4.26: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra1_idp2 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra2_idp2	**1 (seed=25)**	0.246	0.204	0.192
	2 (seed=26)	0.271	0.209	0.197
	3 (seed=27)	0.27	0.201	0.191
	4 (seed=28)	0.285	0.213	0.202
	5 (seed=29)	0.259	0.198	0.185
	6 (seed=30)	0.277	0.204	0.195
	media	**0.268**	**0.205**	**0.194**
	min	**0.246**	**0.198**	**0.185**
	max	**0.285**	**0.213**	**0.202**

Tabela 4.27: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra2_idp2 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra3_idp2	**1 (seed=25)**	0.361	0.32	0.312
	2 (seed=26)	0.351	0.275	0.266
	3 (seed=27)	0.349	0.291	0.288
	4 (seed=28)	0.326	0.303	0.299
	5 (seed=29)	0.388	0.284	0.271
	6 (seed=30)	0.369	0.261	0.252
	media	**0.357**	**0.289**	**0.281**
	min	**0.326**	**0.261**	**0.252**
	max	**0.388**	**0.32**	**0.312**

Tabela 4.28: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra3_idp2 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra4_idp2	**1 (seed=25)**	0.304	0.238	0.227
	2 (seed=26)	0.288	0.233	0.224
	3 (seed=27)	0.313	0.237	0.225
	4 (seed=28)	0.326	0.246	0.235
	5 (seed=29)	0.322	0.213	0.204
	6 (seed=30)	0.301	0.221	0.212
	media	**0.309**	**0.231**	**0.221**
	min	**0.288**	**0.213**	**0.204**
	max	**0.326**	**0.246**	**0.235**

Tabela 4.29: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra4_idp2 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra5_idp2	**1 (seed=25)**	0.277	0.251	0.248
	2 (seed=26)	0.296	0.243	0.238
	3 (seed=27)	0.299	0.251	0.244
	4 (seed=28)	0.309	0.258	0.252
	5 (seed=29)	0.309	0.261	0.26
	6 (seed=30)	0.303	0.26	0.251
	media	**0.299**	**0.254**	**0.249**
	min	**0.277**	**0.243**	**0.238**
	max	**0.309**	**0.261**	**0.26**

Tabela 4.30: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra5_idp2 com diferentes combinações de atributos.

	n execução	H (idp2_novo e pl305c)	H (idp2_novo, pl305c e dlac)	H (idp2_novo, pl305 e dlac)
gs1_ra9_idp2	**1 (seed=25)**	0.212	0.155	0.143
	2 (seed=26)	0.217	0.158	0.147
	3 (seed=27)	0.213	0.16	0.149
	4 (seed=28)	0.235	0.165	0.152
	5 (seed=29)	0.207	0.155	0.143
	6 (seed=30)	0.226	0.164	0.152
	media	**0.218**	**0.159**	**0.147**
	min	**0.207**	**0.155**	**0.143**
	max	**0.235**	**0.165**	**0.152**

Tabela 4.31: Valores da estatística de Hopkins calculados para o subconjunto de dados gs1_ra9_idp2 com diferentes combinações de atributos.

Primeiramente, viu-se que a inclusão do atributo dlac provocou redução dos valores de H em quase todos os subconjuntos, excetuando-se para gs1_ra9_idp1, que teve um pequeno aumento no valor (Tabela 4.25). Percebeu-se também que os maiores valores foram verificados no subconjunto gs1_ra3_idp1 com base nos valores da primeira coluna da Tabela 4.22 que foram, em sua maioria, acima de 0.4 e, mesmo com a introdução do atributo dlac, eles ainda permaneceram acima de 0.3.

Desse modo, dentre os subconjuntos, ele provavelmente contém uma estrutura aleatória ou com muito pouca significância. Os segundos maiores valores foram identificados em gs1_ra3_idp2. Portanto, esse subconjunto também tende a apresentar uma fraca ou nenhuma estrutura de grupos relevante. De posse dos maiores valores de H da segunda coluna, foi possível notar pouca significância de gs1_ra1_idp1, cujo valor máximo foi de 0.294.

Os subconjuntos mais bem pontuados, i.e., com menores valor de H, foram gs1_ra9_idp1 e gs1_ra9_idp2, com médias 0.131 e 0.218, respectivamente, antes da inclusão de dlac. Portanto, são os grupos que, possivelmente, apresentam maior significância.

Dentre os subconjuntos restantes, tomou-se as médias daqueles que apresentaram valores menor que 0.3 antes da introdução de dlac, observando se havia notáveis reduções no valor de H posteriormente a sua inclusão. Com isso, viu-se que gs1_ra2_idp1 e gs1_ra2_idp2, depois dos mais bem avaliados, também podem conter grupos com maior significância.

A partir dos valores máximos de H inferiores ou muito próximos a 0.3 da primeira coluna também se identificou que os subconjuntos gs1_ra1_idp2 e gs1_ra5_idp2 tem chances de apresentar estruturas de grupos, dado que esses valores, correspondentes à 0.299 e 0.309, respectivamente, sofreram perceptíveis reduções com a inclusão de dlac. No caso dos subconjuntos gs1_ra4_idp1, gs1_ra5_idp1 e gs1_ra4_idp2 constatou-se também significativas diminuições dos valores de H após utilizar o atributo dlac, com médias inferiores a 0.25 na segunda coluna e, portanto, podendo conter alguma estrutura de grupos de relevância.

Em resumo, viu-se que, dos 12 subconjuntos utilizados, 3 deles, a saber, gs1_ra1_idp1, gs1_ra3_idp1, gs1_ra3_idp2, são os mais propensos a apresentarem pouca ou nenhuma informação relevante no que diz respeito aos grupos, com destaque para os dois últimos, onde possivelmente existe uma estrutura aleatória. As proximidades entre os valores das segunda e terceira colunas se deram pela alta correlação existente entre pl305 e pl305c.

Adicionalmente, notou-se significâcia dos grupos encontrados nos demais subconjuntos, mostrados no experimento anterior, dada às maiores reduções no valor de H após a utilização do atributo dlac no cálculo do teste para valores abaixo de 0.3. Logo, concluiu-se as associações anteriormente observadas entre os regimes 2 e 9, e entre os regimes 4 e 5, são válidas, podendo conter significado prático.

4.4 Consolidação dos resultados

O experimento 1 foi útil para se estabelecer valores ideais para o número de grupos com o auxílio dos índices de validação interna apresentados neste trabalho para o KM e o FCM. No KM, também se utilizou 26 índices adicionais de um pacote do R contendo 30 índices internos. O k_{max} inicial considerado foi de 15 grupos. Eles foram aplicados sobre cada subconjunto com todas as quatro combinações de atributos empregadas nos agrupamentos, a saber, (pl305c, idp1_novo), (pl305c, idp2_novo), (prod_diaria, idp1_novo), (prod_diaria, idp2_novo). Após isso, extraiu-se as maiores ocorrências para compor o *range* de valores

ideais para $k = \{2, 3, 4, 5, 6, 7\}$.

No experimento 2, os subconjuntos foram agrupados com os valores ideais de k do experimento anterior. As partições obtidas com o KM foram validadas internamente com o coeficiente de silhueta e a variação intragrupo. No FCM, por sua vez, utilizou-se o coeficiente de silhueta *fuzzy*, o índice de Fukuyama-Sugeno e o índice de Xie-Beni na validação interna. Pelo CS, viu-se que as partições geradas com o KM apresentaram sobreposições de grupos com base nos seus valores, inferiores ou muito próximos a 0.5. A análise dos valores de CSF permitiu supor que os grupos gerados pelo FCM contêm tanto objetos bem definidos quanto mal definidos dentro deles.

Em seguida, tomou-se os menores e maiores valores de pertinência máxima para cada subconjunto com todos os valores de k utilizados nos agrupamentos com o FCM. Pelos máximos, pôde-se confirmar a existência de objetos muito bem definidos nos grupos aos quais eles foram atribuídos, já que alguns dos maiores valores de pertinência foram praticamente 1. Os valores mínimos, entretanto, mostraram que, conforme aumento do número de grupos, aumenta-se gradualmente o grau de certeza com que os objetos são atribuídos aos grupos. Desse modo, constatou-se que a elevação do número de grupos também causa aumento da precisão de análise.

Posteriormente, selecionou-se valores intermediários no *range* ideal de k, a saber, 3 e 4, como o número ideal de grupos, considerando o contraste entre a significância e a precisão de análise, mencionado por Zadeh (1973, p. 28). Os gráficos gerados por ambos os métodos com 3, 4 e 7 grupos utilizando as combinações entre as ordens de parto e a produção diária não acrescentaram informações ao estudo, dado que a variável prod_diaria não exerceu qualquer influência no processo de agrupamento e, logo, elas foram desconsideradas das análises.

Todavia, pela análise gráfica dos agrupamentos dos subconjuntos com 3 e 4 grupos com as ordens de parto e a produção de leite corrigida para 305 dias de lactação, foi possível detectar semelhanças no comportamento dos dados presentes nos subconjuntos de regimes alimentares 2 e 9, e nos subconjuntos de regimes 4 e 5. Ao se utilizar $k = 7$, notou-se que essas similaridades entre esses regimes, anteriormente identificadas para 3 e 4 grupos, se desfizeram, provavelmente por consequência do aumento da precisão de

análise. Independentemente do valor de k, viu-se que, em todas partições obtidas com ambos os algoritmos, há proximidade entre fronteiras de grupos e, logo, os objetos no seu entorno não estão bem definidos dentro dos mesmos.

Tendo em vista que o FCM consegue lidar com a incerteza dos dados, com o auxílio do conjunto de valores de pertinência máxima dos objetos, definiu-se faixas de valores para caracterizá-los em relação à definição nos grupos. Sendo assim, foi possível identificar diferentes regiões de pertinência dentro dos grupos. As áreas de mais alta pertinência concentraram-se nos núcleos dos grupos, i.e., em torno dos centroides.

Como esperado, notou-se graficamente que, à medida que a distância entre os objetos e os centroides aumenta, os valores de pertinência decaem, compondo áreas de menores pertinência. Nos agrupamentos com $k = 7$, notou-se que os pontos em vermelho nos gráficos localizam-se muito próximos às fronteiras dos grupos. Logo, nesses casos, há menor variação entre a distância desses pontos a dois ou mais centroides. A geometria circular das regiões se deu por conta do uso da distância euclidiana como medida de proximidade nos agrupamentos. Pela Tabela 4.19, viu-se que a correlação entre a produção de leite corrigida e a duração da lactação para cada subconjunto é positiva, com valores acima de 0.5, indicando uma tendência de maiores lactações por parte dos animais mais produtivos.

A divisão dos objetos nos grupos por regiões de pertinência permitiu evidenciar maiores associações entre os regimes 2 e 9, como também entre os regimes 4 e 5, a partir da observação do comportamento dos núcleos na variação de k de 3 para 4 grupos. Nessa transição, viu-se a formação de dois ou três núcleos em diferentes faixas de produção de leite para idades ao parto abaixo de 8 anos (2920 dias). Como observado anteriormente, pela análise da disposição dos núcleos gerados com 7 grupos, notou-se que, nitidamente, as associações entre os regimes foram desfeitas como provável consequência do aumento da precisão.

O experimento 3 serviu para validar as associações de regimes através do teste de aleatoriedade espacial por estatística de Hopkins. Ela mostrou que, os subconjuntos de animais submetidos ao regime alimentar 3 possivelmente contêm uma estrutura aleatória nos dados, já que, mesmo com a introdução da duração da lactação, os valores máximos

do teste permaneceram acima de 0.3, sobretudo na primeira ordem de parto.

Além disso, o valor máximo de H da primeira coluna da Tabela 4.20, equivalente a 0.364, juntamente com o valor máximo da segunda coluna, i.e., 0.294, revelaram baixa significância dos grupos no subconjunto de regime alimentar 1 com a primeira ordem de parto. Observou-se significância estatística nos grupos dos demais subconjuntos, considerando as reduções nos valores de H após a utilização de dlac no seu cômputo, ficando inferiores a 0.3. Logo, notou-se que as associações previamente identificadas entre os regimes 2 e 9, e entre os regimes 4 e 5, são relevantes para o estudo.

5 Conclusão

A metodologia empregada neste trabalho mostrou-se útil à descoberta de conhecimento na base de dados considerada, visto que, por meio dela, foi possível adquirí-lo e validá-lo quanto às relações entre os atributos utilizados e a producão de leite, às associações entre subconjuntos, ao comportamento dos dados e aos grupos. Além disso, a utilização da abordagem *fuzzy* foi crucial para o alcance do objetivo proposto neste estudo, dado que ela acrescentou valiosas informações ao mesmo, permitindo compreender de melhor forma as incertezas observadas nos dados.

Como ambos os métodos particionais de agrupamento, KM e FCM, dependem da especificação a priori do número de grupos e, devido ao desconhecimento prévio deste, o uso de diferentes índices de validação interna foi importante para o estabelecimento inicial de um *range* de valores ideais para k, que permitiu comparar os diferentes agrupamentos conforme sua variação em termos da coesão e da separação dos grupos. A escolha de valores intermediários, 3 e 4 grupos, auxiliou a identificação de associações entre subconjuntos, bem como, provavelmente, a definição de grupos com maior relevância, dado o contraste entre a precisão e a significância destes de acordo com k.

No que diz respeito aos resultados, verificou-se que, com base nas análises feitas com 3 e 4 grupos, existe um padrão de comportamento similar tanto entre os animais submetidos aos regimes alimentares 2 e 9, quanto entre os de regimes 4 e 5. Logo, existe uma correspondência entre os elementos em cada par de regime, i.e., atuam de forma similar na produção. Observou-se também que há uma correlação moderada entre a duração da lactação e a produção, menos evidente para os animais sob o regime alimentar 9. Portanto, há maiores chances dos animais mais produtivos serem aqueles a terem maiores períodos de lactação.

Em relação à atuação dos algoritmos na formação dos grupos observou-se que:

- Com 3 grupos:
  - Na primeira ordem de parto, os métodos apresentaram comportamento relati-

vamente semelhantes, porém com algumas diferenças mais nítidas nos subconjuntos de animais com regimes 2 e 9.

  - Nas demais ordens, as mudanças mais drásticas ocorreram nos subconjuntos com esses regimes (2 e 9).

- Com 4 grupos:

  - Quanto a primeira ordem de parto, houve mudanças mais significativas em mais subconjuntos, com destaque para aqueles com animais sob os regimes 4 e 5.

  - A variação mais expressiva nas demais ordens de parto ocorreu no subconjunto de animais com o regime 5.

Apesar dos indícios de sobreposições de grupos pelo valor dos índices internos utilizados para a validação das partiçoes com o KM, somente com o FCM foi possível representá-las e compreendê-las. Os graus máximos de pertinência de cada objeto serviram de norte para a criação de faixas de valores entre os mesmos, conseguindo-se separar graficamente os objetos por regiões de pertinência dentro de cada grupo. Naturalmente, isso permitiu a identificação de *cores* ou núcleos de grupos, onde os objetos estão bem definidos nos clusters, que, consequentemente, possuem características individuais distintas de outros núcleos, e, portanto, com potencial valor para avaliação genetíca dos animais no interior destes. O formato arredondado dos *cores* se deu por conta da utilização da distância euclidiana no cômputo da similaridade entre os objetos no método de agrupamento.

Nos *cores* identificados com o uso do FCM com 3 e 4 grupos, notou-se que, de modo geral, a diferenciação da produção de leite em relação à idade do Gir Leiteiro dentro dessas regiões de alta pertinência ocorre nos animais com até pouco mais de 6 anos (2190 dias), aproximadamente, não ultrapassando 8 anos (2920 dias), notada a partir da mudança de $k = 3$ para $k = 4$ com o surgimento de mais núcleos de grupos abaixo desse limite em diferentes faixas de valores de produção.

O estudo acerca da aleatoriedade dos subconjuntos revelou que, provavelmente, aqueles com animais sob o regime alimentar 3 não contêm estrutura de grupos e, logo, os clusters das partições encontrados nesse regime pelos algoritmos muito provavelmente não

apresentar significância estatística, o que evitou equívocos na interpretação dos resultados. Ademais, também apontou baixa significância dos grupos no subconjunto de animais submetidos ao regime 1 da primeira ordem de parto. Entretanto, como os subconjuntos restantes apresentaram reduções significativas no valor da estatística de Hopkins com a introdução da duração da lactação no cálculo desse teste para valores abaixo de 0.3, provavelmente os grupos nesses subconjuntos contêm significância estatística. Desse modo, concluiu-se que, de fato, ambas as associações de regimes identificadas, a saber, entre os regimes 2 e 9 e entre os regimes 4 e 5, fazem sentido.

Isso significa que, na prática, pode-se ter reduções de gastos na alimentação dos animais a partir da utilização do regime alimentar de menor custo para os subconjuntos de animais onde há associação de regimes, dado que eles tendem a impactar de modo semelhante na produção de leite. Entretanto, deve-se verificar a existência de uma mesma ou até melhor interação genótipo-ambiente pela troca do regime para manutenibilidade ou elevação da performance produtiva, sem deixar de considerar outras variáveis do fator ambiente.

Pode-se também ter ganho de produtividade pela mudança da alimentação dos animais menos produtivos para um outro regime não associado na busca por melhores respostas da interação genótipo-ambiente e da adaptação do animal ao ambiente. Além disso, o aumento na produtividade pode se dar pela seleção de animais com maior potencial genético para a produção de leite a partir de avaliações genéticas dos núcleos dos grupos encontrados, visando a formação de novilhas geneticamente superiores para essa característica.

Os agrupamentos com 7 grupos permitiram constatar aumento da precisão com a elevação do número de grupos, onde foi possível identificar regiões de pertinências ainda menores com o FCM. Nesse caso, em ambos os métodos, as associações anteriormente identificadas entre os subconjuntos de animais sob os regimes 2 e 9, e entre aqueles sob regimes 4 e 5, se desfizeram. Pelo FCM, verificou-se que os objetos com maiores dúvidas na atribuição, i.e., que apresentaram menores graus de pertinência aos grupos nos quais foram inseridos, localizam-se nas adjacências de fronteiras de grupos próximas umas das outras, onde existe menor diferença numérica entre a distância desses objetos a dois ou

mais centroides.

Entretanto, como já dito, os grupos encontrados com $k = 7$, provavelmente, são mais susceptíveis a conter menos significância do que os para $k = 3$ e $k = 4$, dada a discordância entre a precisão e a mesma conforme aumento de k. Por fim, vale ressaltar que, embora computacionamente mais caro que o KM, o FCM conseguiu agregar valiosas informações ao estudo no que concerne ao entendimento dos dados, as quais não seriam possíveis somente com o uso da abordagem clássica de agrupamento.

Como possibilidades de trabalhos futuros, quanto ao uso da base dados, sugere-se efetuar análises a outras raças bovinas, em especial para os sintéticos de maior expressividade, visando avaliar e comparar suas performances produtivas sobre as mesmas dietas. Em relação ao processo de análise, há diversas sugestões de trabalhos futuros. Uma delas é realizar testes com variações no parâmetro m do FCM para agrupamento dos mesmos subconjuntos amostrais utilizados neste estudo e/ou de subdivisões de outras raças. Adicionalmente, sugere-se empregar outras métricas de distância no cômputo da similaridade entre os objetos na busca por diferentes geometrias.

Ainda nesse sentido, pode ser interessante utilizar outras medidas de proximidade, como as medidas baseadas em correlação, por exemplo. Indica-se também considerar a utilização de outros métodos de agrupamento, sobretudo os que não dependem da escolha de k, como os métodos baseados em densidade; ou mesmo o uso de redes neurais, e.g., mapas auto-organizáveis (*self-organizing maps* - SOM). Tais sugestões podem somar novos conhecimentos com potencial econômico e científico. Por fim, outra ideia é analisar dados provenientes de outros controles para avaliação do progresso da performance de produção. Além disso, sugere-se a utilização de dados mais recentes que contenham informações referentes às características do leite.

Bibliografia

ABCZ. **Raças zenuínas**. Disponível em: <http://www.abcz.org.br/Home/Conteudo/23985-Racas-Zebuinas>. Acesso em: 8 jul. 2019.

ACHARYA, A. S.; PRAKASH, A.; SAXENA, P. ; NIGAM, A. Sampling: Why and how of it. **Indian Journal of Medical Specialities**, v.4, n.2, p. 330–333, 2013.

AGGARWAL, C. C. **Data mining: the textbook**. Springer, 2015, 734p.

ANANTHI, M.; PALANIVEL, K. Survey on classification feature selection strategies. **International Journal of Computer Science and Mobile Applications**, v.5, n.9, p. 23–31, 2017.

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H. P. ; SANDER, J. Optics: Ordering points to identify the clustering structure. **SIGMOD Rec.**, v.28, n.2, p. 49–60, 1999.

ARUNADEVI, J.; NITHYA, M. J. Comparison of feature selection strategies for classification using rapid miner. **International Journal of Innovative Research in Computer and Communication Engineering**, p. 13556–13563, 2016.

ATANGANA, A. **Fractional Operators with Constant and Variable Order with Application to Geo-hydrology**. Elsevier Science, 2017, 414p.

BANERJEE, A.; DAVÉ, R. N. **Validating clusters using the hopkins statistic**. In: International Conference on Fuzzy Systems, volume 1, p. 149–153. IEEE, 2004.

BARATA, R. B.; MORAES, J. C.; ANTONIO, P. R. ; DOMINGUEZ, M. Inquérito de cobertura vacinal: avaliação empírica da técnica de amostragem por conglomerados proposta pela organização mundial da saúde. **Revista Panamericana de Salud Pública**, v.17, p. 184–190, 2005.

BATHLA, G.; AGGARWAL, H. ; RANI, R. A novel approach for clustering big data based on mapreduce. v.8, n.3, p. 1711–1719, 06 2018.

BEZDEK, J. C. Cluster validity with fuzzy sets. **Cybernetics and Systems**, v.3, n.3, p. 58–73, 1973.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York USA: Plenum Press, 1981, 256p.

BEZDEK, J. C.; KELLER, J.; KRISNAPURAM, R. ; PAL, N. R. **Fuzzy Models and Algorithms for Pattern Recognition and Image Processing (The Handbooks of Fuzzy Sets)**. Berlin, Heidelberg: Springer-Verlag, 2005, 776p.

BRITO, F. V.; CARDOSO, V.; CARVALHEIRO, R.; FRIES, L. A.; PEÑA, C. D. O.; PICCOLI, M. L.; ROSO, V. M.; SCKENKEL, F. ; SEVERO, J. L. P. S. **Interação genótipo-ambiente em bovinos de corte: aspectos técnicos e aplicabilidade**. Disponível em: <https://www.beefpoint.com.br/interacao-genotipo-ambiente-em-bovinos-de-corte-aspectos-tecnicos-e-aplicabilidade-38003/>, Porto Alegre: Gensys, 2007. Acesso em: 8 jul. 2019.

CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. A fuzzy extension of the silhouette width criterion for cluster analysis. **Fuzzy Sets and Systems**, v.157, n.21, p. 2858–2875, 2006.

CANHAS, I. **Genótipo**. Disponível em: <https://www.infoescola.com/genetica/genotipo/>, InfoEscola, 2011?. Acesso em: 8 jul. 2019.

CARVALHO, L. M.; NOVAES, L. P.; MARTINS, C. E.; ZOCCAL, R.; MOREIRA, P.; RIBEIRO, A. C. C. L. ; LIMA, V. M. B. **Reprodução**. Disponível em: <https://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Leite/LeiteCerrado/reproducao.html>, Embrapa Gado de Leite, 2002. Acesso em: 8 jul. 2019.

CASSIANO, K. M. **Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade**. Rio de Janeiro, RJ, 2014. 96p. Tese de Doutorado - Pontifícia Universidade Católica.

CASCARINO, R. E. **Data Analytics for Internal Auditors**. Auerbach Publications, 2017, 418p.

CAVALCANTE JÚNIOR, N. L. **Clusterização baseada em algoritmos fuzzy**. Refice, PE, 2006. 97p. Dissertação de Mestrado - Universidade Federal de Pernambuco.

CÉSAR, F. I. G. **Ferramentas Básicas da Qualidade**. 1. ed., Biblioteca 24 horas, 2011, 132p.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, v.40, n.1, p. 16–28, 2014.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V. ; NIKNAFS, A. Nbclust package: finding the relevant number of clusters in a dataset. **Journal of Statistical Software**, v.61, n.1, p. 36, 2014.

COELHO, T. S.; FERNANDES, M. A. R.; MIOT, H. A. ; YORIYAZ, H. Uso do método fuzzy c-means para segmentação de imagens dermatoscópicas de lesões de pele. **Revista Brasileira de Física Médica**, p. 99–102, 2012.

CÔRTES, S. D. C.; PORCARO, R. M. ; LIFSCHITZ, S. **Mineração de dados - Funcionalidades, Técnicas e Abordagens**. Rio de Janeiro, RJ: Pontifícia Universidade Católica, 2002, 34p.

DAVE, R. N. Validating fuzzy partitions obtained through c-shells clustering. **Pattern Recognition Letters**, v.17, n.6, p. 613–623, 1996.

DAISTER, L. P. **Estratégias para desenvolvimento de sistemas de múltiplos classificadores em aprendizado supervisionado**. Rio de Janeiro, RJ, 2007. 98p. Dissertação de Mestrado - COPPE - UFRJ.

DANISH, F. **Fundamentals of Statistics: A Brief Insight**. Onlinegatha, 2017, 165p.

DELAPORTE, D. Preparação de dados de plano de saúde suplementar para algoritmos de mineração de dados. **Universidade Tecnológica Federal do Paraná**. Monografia (Tecnólogo em Tecnologia em Sistemas para Internet). Campo Mourão, PR, 2015. 31p.

DENG, C. Automated construction of multiple regional libraries for neighborhoodwise local multiple endmember unmixing. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v.9, n.9, p. 4232–4246, 2016.

DHIVYADEEPA, E. **Sampling Techniques in Educational Research**. Raleigh, NC, EUA: Lulu Publication, 2015, 148p.

DHOTE, Y.; AGRAWAL, S. ; DEEN, A. J. **A survey on feature selection techniques for internet traffic classification**. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), p. 1375–1380. IEEE, 2015.

DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. **Universidade Presbiteriana Mackenzie**. Monografia (Bacharel em Sistemas de Informação). São Paulo, SP, 2004. 92p.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. **Cybernetics and Systems**, , n.3: 3, p. 32–57, 1973.

DUVVADA, H. P.; NAIDU, G. D. R. ; SRI, V. D. K-means cluster analysis of cities based on their inter-distances. **International Journal of Engineering Development and Research**, v.5, p. 1356–1363, 2017.

FACELI, K.; LORENA, A. C.; GAMA, J. A. ; CARVALHO, A. C. P. L. F. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro, RJ: LTC, 2011, 396p.

FERREIRA, C. C. M. Aplicação da lógica nebulosa (fuzzy cluster) na definição de unidades climáticas: Estudo de caso na bacia do rio paraibuna-mg/rj. **Revista Brasileira de Climatologia**, v.10, n.1, 2012.

FISSET, B.; DHAENENS, C. ; JOURDAN, L. **Mo–Mine$_{clust}$: A framework for multi-objective clustering**. In: International Conference on Learning and Intelligent Optimization, p. 293–305. Springer, 2015.

GARCIA, R.; NIEVOLA, J. C. ; PARAISO, E. C. **Estudo comparativo de métodos de seleção de atributos na predição de matrizes de conectividades em tdah obtidas pela técnica de resting-state fmri**. In: XIII Encontro Nacional de Inteligência Artificial e Computacional, p. 637–647. SBC, 2016.

GARALI, I.; ADEL, M.; BOURENNANE, S. ; GUEDJ, E. Histogram-based features selection and volume of interest ranking for brain pet image classification. **IEEE journal of translational engineering in health and medicine**, v.6, p. 1–12, 2018.

GHOSH, S.; DUBEY, S. K. Comparative analysis of k-means and fuzzy c-means algorithms. **International Journal of Advanced Computer Science and Applications**, v.4, n.4, 2013.

GOLDSCHMIDT, R.; PASSOS, E. ; BEZERRA, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. Elsevier Editora Ltda., 2015, 296p.

GOSWAMI, S.; SAHA, S.; CHAKRAVORTY, S.; CHAKRABARTI, A. ; CHAKRABORTY, B. A new evaluation measure for feature subset selection with genetic algorithm. **International Journal of Intelligent Systems and Applications**, v.7, n.10, p. 28–36, 2015.

GRABUSTS, P. **The choice of metrics for clustering algorithms**. In: Proceedings of the 8th International Scientific and Practical Conference, volume 2, p. 70–76, 2011.

GRIGORAS, G.; CARTINA, G. Improved fuzzy load models by clustering techniques in distribution network control. **International Journal on Electrical Engineering and Informatics**, v.3, n.2, p. 207–2016, 2011.

GROENEN, P. J. F.; KAYMAK, U. ; ROSMALEN, J. Fuzzy clustering with minkowski distance functions. **Advances in Fuzzy Clustering and its Applications**, v.10, p. 53–68, 2007.

GRUPTA, N. S.; VALARMATHI, B. **Total Quality Management**. 2. ed., McGraw-Hill Education, 2009, 212p.

GUEORGUIEVA, N.; VALOVA, I. ; GEORGIEV, G. M&MFCM: Fuzzy c-means clustering with mahalanobis and minkowski distance metrics. **Procedia Computer Science**, v.114, p. 224–233, 2017.

GUIERA, A. J. A.; CENTENO, T. M.; DELGADO, M. R. ; MÜLLER, M. Segmentação por agrupamentos fuzzy c-means em imagens lidar aplicados na identificação de linhas de transmissão de energia elétrica. **Espaço Energia**, v.3, p. 24–31, 2005.

GUIMARÃES, P. R. B. **Métodos quantitativos estatísticos**. 1 ed. rev., Curitiba, PR: IESDE Brasil S.A., 2012, 252p.

HAN, J.; KAMBER, M. ; PEI, J. **Data Mining: Concepts and Techniques third edition**. 3. ed., Waltham, MA, USA: Morgan Kaufmann publishers, 2012, 703p.

HINKLE, D. E.; WIERSMA, W. ; JURS, S. G. **Applied Statistics for the Behavioral Sciences**, volume 663 de **Applied Statistics for the Behavioral Sciences**. 5. ed., Houghton Mifflin, 2003.

HOPKINS, B.; SKELLAM, J. G. A new method for determining the type of distribution of plant individuals. **Annals of Botany**, v.18, n.2, p. 213–227, 1954.

HOPPNER, F.; KLAWONN, F.; KRUSE, R. ; RUNKLER, T. **Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition**. John Wiley Sons Ltd, 2009.

HRUSCHKA, E. R.; EBECKEN, N. F. F. A genetic algorithm for cluster analysis. **Intelligent Data Analysis**, v.7, n.1, p. 15–25, 2003.

HRUSCHKA, E. R.; CAMPELLO, R. J. G. B. ; CASTRO, L. N. Evolving clusters in gene-expression data. **Information Sciences**, v.176, n.13, p. 1898–1927, 2006.

HU, C.; MENG, L. ; SHI, W. Fuzzy clustering validity for spatial data. **Geo-spatial information science**, v.11, n.3, p. 191–196, 2008.

HU, P.; VENS, C.; VERSTRYNGE, B. ; BLOCKEEL, H. **Generalizing from example clusters**. In: International Conference on Discovery Science, p. 64–78. Springer, 2013.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, v.2, n.1, p. 193–218, 1985.

ISLAM, S.; AHMED, M. Implementation of image segmentation for natural images using clustering methods. **International Journal of Emerging Technology and Advanced Engineering**, v.3, n.3, p. 175–180, 2013.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Prentice-Hall Inc, 1988, 334p.

JELIHOVSCHI, E. **Análise exploratória de dados usando o R**. Ilhéus, BA: EDITUS, 2014, 85p.

KAINULAINEN, J. J. Clustering algorithms: basics and visualization. **Helsinki University of Technology, Laboratory of Computer and Information Science**, 2002.

KANTARDZIC, M. **Data mining: concepts, models, methods, and algorithms**. John Wiley & Sons, 2011, 552p.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning**, volume 1. STHDA, 2017, 187p.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis**. A Wiley-Interscience publication. Wiley, 1990, 342p.

KHAN, J. A. **Research Methodology**. APH Publishing Corporation, 2011, 334p.

KHAN, S. U.; ZOMAYA, A. Y. **Handbook on data centers**. Springer, 2015, 1134p.

KIWANUKA, F. N.; WILKINSON, M. H. F. **Cluster based vector attribute filtering**. In: International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing, p. 116–135. Springer, 2015.

KOCH, I. **Analysis of Multivariate and High-Dimensional Data**. Analysis of Multivariate and High-dimensional Data. Cambridge University Press, 2014.

KRISHNA, T. V. S.; BABU, A. Y. ; KUMAR, R. K. **Determination of optimal clusters for a non-hierarchical clustering paradigm k-means algorithm**. In: Proceedings of International Conference on Computational Intelligence and Data Engineering, p. 301–316. Springer, 2018.

KUME, H. **Métodos estatísticos para melhoria da qualidade**. Gente, 1993, 245p.

KUMAR, R. **Research Methodology: A Step-by-Step Guide for Beginners**. SAGE Publications, 2010, 440p.

KUMAR, U.; AHMADI, A.; VERMA, A. K. ; VARDE, P. **Current Trends in Reliability, Availability, Maintainability and Safety: An Industry Perspective**. Springer, 2015, 738p.

KUMAR, V.; CHHABRA, J. K. ; KUMAR, D. Data clustering using differential search algorithm. **Pertanika Journal of Science & Technology**, v.24, n.2, p. 295–306, 2016.

LAVINE, B. K.; NUGURU, K. ; MIRJANKAR, N. One stop shopping: feature selection, classification and prediction in a single step. **Journal of Chemometrics**, v.25, n.3, p. 116–129, 2011.

LAWSON, R. G.; JURS, P. C. New index for clustering tendency and its application to chemical problems. **Journal of chemical information and computer sciences**, v.30, n.1, p. 36–41, 1990.

LINDEN, R. **Algoritmos Genéticos**. 2. ed., Rio de Janeiro: Brasport, 2008, 428p.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistema de Informação da FSMA**, v.1, n.4, p. 18–36, 2009.

LOUZADA NETO, F.; DINIZ, C. A. R. **Técnicas estatísticas em data mining**, volume 31. São Carlos, SP: IMCA, 2002, 102p.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, p. 281–297. Oakland, CA, USA, 1967.

MADHULATHA, T. S. An overview on clustering methods. **IOSR Journal of Engineering**, v.2, n.4, p. 719–725, 2012.

MALINEN, M. **New alternatives for k-means clustering**. Joensuu, 2015. 82p. Dissertação de Mestrado - University of Eastern Finland.

MALKI, A. A.; RIZK, M. M.; EL-SHORBAGY, M. A. ; MOUSA, A. A. Hybrid genetic algorithm with k-means for clustering problems. **Open Journal of Optimization**, v.5, n.02, p. 71–83, 2016.

MEIRA JR., W. **Mineração de dados**. Disponível em: <`http://homepages.dcc.ufmg.br/~meira/DokuWiki/wiki/algoritmos?do=recent#agrupamento`>. Acesso em: 8 jul. 2019.

MOROZKOV, M.; GRANICHIN, O.; VOLKOVICH, Z. ; ZHANG, X. **Fast algorithm for finding true number of clusters. applications to control systems**. In: 2012 24th Chinese Control and Decision Conference (CCDC), p. 2001–2006. IEEE, 2012.

NAMRATHA, M.; PRAJWALA, T. R. A comprehensive overview of clustering algorithms in pattern recognition. **IOR Journal of Computer Engineering**, v.4, n.6, p. 23–30, 2012.

NOGUEIRA JÚNIOR, J. B.; MONICO, J. F. G. ; TACHIBANA, V. M. Tamanho da amostra no controle de qualidade posicional de dados cartográficos. **Boletim de Ciências Geodésicas**, v.10, n.1, p. 101–112, 2004.

NORUŠIS, M. J. **IBM SPSS statistics 19 statistical procedures companion**. Prentice Hall, 2012, 646p.

NUEVO, A. G. D.; CATANIA, V. ; PALESI, M. **The hybrid genetic fuzzy c-means: a reasoned implementation**. In: Proceedings of the 7th WSEAS International Conference on Fuzzy Systems, p. 33–38. ACM, 2006.

OLIVEIRA, D. S. Fuzzy clustering multi-critério para dados relacionados à genética. **Universidade Federal de Uberlândia**. Monografia (Bacharel em Estatística). Uberlândia, MG, 2016. 47p.

OLIVEIRA, A. G. **MiMi: Plataforma computacional para mineração de dados micrometeorológicos**. Cuiabá, MT, 2015. 84p. Tese de Doutorado - Universidade Federal do Mato Grosso.

OSLER II, J. E. **Interactive Statistics Methods ©**. Osler Studios Incorporated, 2012, 472p.

PAL, N. R.; BEZDEK, J. C. On cluster validity for the fuzzy c-means model. **IEEE Transactions on Fuzzy systems**, v.3, n.3, p. 370–379, 1995.

PANETTO, J. C. C.; SILVA, M. V. G. B.; VERNEQUE, R. S.; MACHADO, M. A.; FERNANDES, A. R.; MARTINS, M. F.; FAZA, D. R. L. R.; ARBEX, W. A.; OLIVEIRA, J. C.; VENTURA, H. T. ; PEREIRA, M. A. Programa nacional de melhoramento do gir leiteiro: Sumário brasileiro de touros - 2ª avaliação genômica de touros - Resultado do teste de progênie - Abril 2019. **Embrapa Gado de Leite - Documentos - 235 (INFOTECA-E)**, 98p., 2019.

PERIM, G. T. **Uso de métodos de inicialização combinados ao simulated annealing para resolver o problema de agrupamento de dados**. Vitória, ES, 2008. 76p. Dissertação de Mestrado - Universidade Federal do Espírito Santo.

PERES, S. M.; ROCHA, T.; BISCARO, H. H.; MADEO, R. C. ; BOSCARIOLI, C. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: Abordagens híbridas para tarefas de agrupamento e classificação. **Revista de Informática Teórica e Aplicada**, v.19, n.1, p. 120–163, 2012.

PEREIRA, C. M. M.; MELLO, R. F. Common dissimilarity measures are inappropriate for time series clustering. **Revista de Informática Teórica e Aplicada**, v.20, n.1, p. 25–48, 2013.

PINHEIRO, L. C. **Método de representação espacial de clustering**. Curitiba, PR, 2006. 123p. Dissertação de Mestrado - Universidade Federal do Paraná.

POLYCARPO, R. C. **Conceitos genéticos**. Disponível em: <`https://www.milkpoint.com.br/artigos/producao/conceitos-geneticos-44625n.aspx`>, MilkPoint, 2008. Acesso em: 8 jul. 2019.

RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical association**, v.66, n.336, p. 846–850, 1971.

RAO, V. S.; VIDYAVATHI, S. Comparative investigations and performance analysis of fcm and mfpcm algorithms on iris data. **Indian Journal of Computer Science and Engineering**, v.1, n.2, p. 145–151, 2010.

RICCI, F.; ROKACH, L. ; SHAPIRA, B. **Recommender Systems Handbook**. Springer US, 2015, 1003p.

ROCHA, J. E. N. **Sistemas inteligentes no estudo de pedras comerciais do setor de energia elétrica**. Rio de Janeiro, RJ, 2003. 198p. Dissertação de Mestrado - Pontifícia Universidade Católica.

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

RUDAS, T. **Lectures on Categorical Data Analysis**. Springer, 2018.

SAMOHYL, R. W. **Controle estatístico de qualidade**. Elsevier, 2009.

SANTOS, R. M.; VASCONCELOS, J. L. M. **Interpretação dos Índices da eficiência reprodutiva**. Disponível em: <https://www.milkpoint.com.br/colunas/jose-luiz-moraes-vasconcelos-ricarda-santos/interpretacao-dos-indices-da-eficiencia-reprodutiva-41269n.aspx>, MilkPoint, 2007. Acesso em: 8 jul. 2019.

SAPORTA, G.; YOUNESS, G. **Comparing two partitions: Some proposals and experiments**. In: Compstat, p. 243–248. Springer, 2002.

SASSI, R. J. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados: teoria dos rough sets e redes neurais artificiais mapas auto-organizáveis.** São Paulo, SP, 2006. 169p. Tese de Doutorado - Universidade de São Paulo.

SERRA, A.; GRECO, D. ; TAGLIAFERRI, R. **Impact of different metrics on multi-view clustering**. In: 2015 International Joint Conference on Neural Networks (IJCNN), p. 1–8. IEEE, 2015.

SIDHU, N. K.; KAUR, R. Clustering in data mining. **International Journal of Computer Trends and Technology (IJCTT)**, v.4, n.4, p. 710–714, 2013.

SILVA, A. L. C. Introdução à análise de dados. **Rio de Janeiro: E-papers**, 2009.

SILVA, L. R. **Uma plataforma intervalar para agrupamentos de dados**. Natal, RN, 2015. 109p. Tese de Doutorado - Universidade Federal do Rio Grande do Norte.

SLADOJE, N.; LINDBLAD, J. ; NYSTRÖM, I. Defuzzification of spatial fuzzy sets by feature distance minimization. **Image and Vision Computing**, v.29, n.2-3, p. 127–141, 2011.

SORZANO, C. O. S.; VARGAS, J. ; MONTANO, A. P. A survey of dimensionality reduction techniques. **arXiv preprint arXiv:1403.2877**, p. 1–35, 2014.

SOUZA, A. H. D. **Seleção de atributos relevantes: aplicando técnicas na base de dados do herbário virtual da flora e dos fungos**. 2017. 81p. Dissertação de Mestrado - Universidade Federal do Amazonas.

STAPENHURST, T. **Mastering Statistical Process Control**. Taylor & Francis, 2013, 456p.

STEWART, W. J. **Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling**. Princeton University Press, 2009, 776p.

SUMAN, K.; THIRUMAGAL, S. Feature subset selection with fast algorithm implementation. **International Journal of Computer Trends and Technology**, v.6, n.1, p. 1–5, 2013.

TAN, P. N.; STEINBACH, M. ; KUMAR, V. **Introdução ao Data mining: Mineração de Dados**. Rio de Janeiro: Ciência Moderna Ltda., 2009.

TENG, X.; DONG, H. ; ZHOU, X. Adaptive feature selection using v-shaped binary particle swarm optimization. **PloS one**, v.12, n.3, p. 1–22, 2017.

TESSER, D. P.; SANTOS, S. N.; REIS, D. R. ; OLIVEIRA JUNIOR, L. Reclamações de clientes como fonte de inovações a partir de uma base de help desk utilizando data mining-um exemplo de aplicação. **Revista Latino-Americana de Inovação e Engenharia de Produção**, v.1, n.1, p. 20–37, 2013.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. Elsevier Science, 2003, 689p.

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, v.18, n.4, p. 267–276, 1953.

TIBSHIRANI, R.; WALTHER, G. ; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v.63, n.2, p. 411–423, 2001.

VALE, M. N. **Agrupamentos de dados: Avaliação de métodos e desenvolvimento de aplicativo para análise de grupos**. Rio de Janeiro, RJ, 2005. 120p. Dissertação de Mestrado - Pontifícia Universidade Católica.

VATHY-FOGARASSY, A.; ABONYI, J. **Graph-Based Clustering and Data Visualization Algorithms**. SpringerBriefs in Computer Science. Springer London, 2013.

VENDRAMIN, L.; CAMPELLO, R. J. G. B. ; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical analysis and data mining: the ASA data science journal**, v.3, n.4, p. 209–235, 2010.

VENGADESWARAN, S.; BALASUNDARAM, S. R. **Significance of hierarchical and partitioning based clustering in grouping aware data placement for data intensive applications**. In: 2017 National Conference on Parallel Computing Technologies (PARCOMPTECH). IEEE, 2017.

WANG, F.; FRANCO-PENYA, H.-H.; KELLEHER, J. D.; PUGH, J. ; ROSS, R. **An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity**. In: International Conference on Machine Learning and Data Mining in Pattern Recognition, p. 291–305. Springer, 2017.

XIE, L. X.; BENI, G. A validity measure for fuzzy clustering. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, v.13, n.8, p. 841–847, 1991.

YAN, M.; YE, K. Determining the number of clusters using the weighted gap statistic. **Biometrics**, v.63, n.4, p. 1031–1037, 2007.

YONAMINE, F. S.; SPECIA, L.; CARVALHO, V. O. ; NICOLETTI, M. C. **Aprendizado não supervisionado em domínios fuzzy – algoritmo fuzzy c-means**. São Carlos: UFSCAR, 2002. 18p.

ZADEH, L. A. Outline of a new approach to the analysis of complex systems and decision processes. **IEEE Transactions on systems, Man, and Cybernetics**, v.smc-3, n.1, p. 28–44, 1973.

ZAMPAR, A. **A importância da interação genótipo-ambiente na bovinocultura leiteira - parte i**. Disponível em: <www.milkpoint.com.br/artigos/producao/a-importancia-da-interacao-genotipoambiente-na-bovinocultura-leiteira-parte-i-58465n.aspx>, MilkPoint, 2009. Acesso em: 8 jul. 2019.

ZIDEK, K.; MAXIM, V.; PITEL, J. ; HOSOVSKY, A. Embedded vision equipment of industrial robot for inline detection of product errors by clustering–classification algorithms. **International Journal of Advanced Robotic Systems**, v.13, n.5, p. 1–10, 2016.