UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

CLUSTERIZAÇÃO COMO TÉCNICA DE APOIO À DECISÃO PARA UM *MARKETPLACE* ELETRÔNICO LOGÍSTICO

Lucas Gomes Pereira

Itajubá

Agosto de 2023

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Lucas Gomes Pereira

CLUSTERIZAÇÃO COMO TÉCNICA DE APOIO À DECISÃO PARA UM *MARKETPLACE* ELETRÔNICO LOGÍSTICO

Dissertação submetida ao Programa de Pós-graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de Mestre em Ciências em Engenharia de Produção.

Área de concentração: Engenharia de Produção
Orientador: Prof. Dr. Renato da Silva Lima

Itajubá

Agosto de 2023

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Lucas Gomes Pereira

CLUSTERIZAÇÃO COMO TÉCNICA DE APOIO À DECISÃO PARA UM *MARKETPLACE* ELETRÔNICO LOGÍSTICO

Dissertação aprovada por banca examinadora em 21 de setembro de 2023, conferindo ao autor o título de Mestre em Ciências em Engenharia de Produção.

Banca examinadora:
Prof. Dr. Alexandre Ferreira de Pinho
Prof. Dr. André Luiz Barbosa Nunes da Cunha
Prof. Dr. Fabio Favaretto

Orientador:
Prof. Dr. Renato da Silva Lima

Itajubá

Agosto de 2023

DEDICATÓRIA

A Jesus, por ter me dado forças; à minha família, por ter me incentivado; aos meus amigos, por terem feito os meus dias mais alegres de se viver.

AGRADECIMENTOS

Esta dissertação é uma conquista muito grande, foram muitos desafios. Por isso, primeiramente, agradeço a Deus por ter me fortalecido e me mantido durante todo o processo. Eu não sou nada sem a Sua presença em mim. Quero registrar aqui que pertenço a Ele e que entreguei a minha vida ao Autor e Consumador da minha fé: Jesus Cristo. Meu único e suficiente salvador.

Agradeço aos meus pais, Nilma e Lúcio, que sempre me incentivaram a estudar, sempre acreditaram em meu potencial. Com muito trabalho e dedicação, eles me deram uma ótima criação. Jesus caprichou ao ter me dado pais tão maravilhosos.

Agradeço a minha irmã, Luísa, por ter sido o meu suporte em tantas vezes de desânimo e ansiedade. Não sei o que seria de mim sem seu apoio, "zimã"! Agradeço aos meus lindos sobrinhos, Ana, Pedro e Laura, por serem tão carinhosos com o tio.

Agradeço à minha avó, Geralda, e ao meu avô, Natanael, por sempre se alegrarem com as minhas conquistas. Agradeço também aos meus tios e tias, primos e primas, que são muito importantes para mim. Amo demais a minha família.

Agradeço aos meus colegas e amigos do LogTranS e do IEPG: Érica, Daniel, Rodrigo, Cecília, Andreza, Júlia, Juliana, Flávia Gontijo, Flávia Tuane, Larissa, Mirelli, Rafaela, Tiago, Rebecca, Mariana, Jayne. Vocês foram presentes de Deus para minha vida. Eu tenho certeza de que a nossa amizade vai ser para o resto da vida, independentemente de onde estivermos.

Agradeço ao meu orientador, Renato, por ter me conduzido com seus conhecimentos e experiências. Muito obrigado por ter se preocupado comigo no decorrer do mestrado e, principalmente, na minha estadia nos Estados Unidos. Obrigado por sempre ter me incentivado a correr atrás de aprimoramento.

Agradeço ao Prof. Dr. Hektor Monteiro, do Laboratório de Astrofísica Computacional (LAC), do Instituto de Física e Química da UNIFEI, pela contribuição neste trabalho, dando acesso aos computadores de alto desempenho do LAC.

Agradeço ao Professor José Holguín-Veras, que me recebeu tão bem no Rensselaer Polytechnic Institute, me dando o prazer de assistir a suas aulas. Obrigado aos meus queridos amigos que conheci nos Estados Unidos: Júlia, Milena, Andres, Adekunle, Oriana, Eli e Lina. I cannot wait to see you, guys, in a near future.

Agradeço a todos os professores do Instituto Federal de Minas Gerais e da Universidade Federal de Itajubá, por terem contribuído imensamente na minha formação e na minha carreira. Espelho-me muito em vocês.

Por fim, agradeço ao CNPq, à empresa parceira do projeto e à Universidade Federal de Itajubá pelo apoio financeiro na bolsa regular e na ajuda de custo do intercâmbio nos Estados Unidos.

RESUMO

O transporte rodoviário é uma das modalidades de maior impacto na matriz de transportes internacional. Apesar das suas vantagens de flexibilidade e disponibilidade, é um setor marcado pela alta fragmentação. Antigamente, o processo de intermediação dessa cadeia logística era operado por agentes de fretes, ineficiente em tempo e custo. Uma solução para atender às necessidades de agilidade e facilidade são os *marketplaces* logísticos eletrônicos, definidos por sistemas que permitem as transportadoras anunciarem suas cargas a caminhoneiros em busca de fretes. Todavia, como consequência da facilidade da automação de correspondência entre carga e capacidade, as provedoras de tecnologia desse modelo de negócio estão tendo que lidar com um volume de dados sem precedentes. Dessa variedade, pode-se extrair informações úteis sobre o comportamento dos usuários. Pode-se pontuar que, apesar da sua popularização, a literatura científica sobre os *marketplaces* logísticos não acompanhou o crescimento. Haja vista dessas oportunidades, este trabalho visa identificar padrões em uma base de dados de anúncio de cargas de um *marketplace* logístico por meio da clusterização, capaz de auxiliar na tomada de decisão. A pesquisa, seguindo o procedimento do CRISP-DM, obteve os dados de postagem de cargas na plataforma de 2019 a 2021. Empregando o *software* RStudio, a tendência de clusterização da base de dados foi confirmada e posteriormente implementado o algoritmo CLARA. A qualidade dos agrupamentos foi avaliada pelo índice Silhueta. Constatou-se que o grupo mais representativo, no âmbito nacional, pôde ser representado por fretes dentro do estado de São Paulo, que apresentavam carga completa, percorrendo distâncias de cerca de 500 km e demandando veículos de categoria pesada para o transporte. Já no contexto São Paulo, a partição mais expressiva foi a de fretes lotação, com viagens de pouco mais de cinco horas e que exigiam também veículos de categoria pesada. Associou-se a maior frequência desse tipo de frete aos seus benefícios, a saber, a eficiência em termos de utilização de espaço e recursos. As principais estratégias identificadas, no contexto nacional, consistem no oferecimento às transportadoras que realizam um alto volume de viagens descontos progressivos nos serviços adicionais e a divulgação do *marketplace* logístico entre os potenciais clientes com interesse específico em serviços de transporte de cargas pesadas e médias distâncias. Pôde-se identificar uma interessante oportunidade de negócio no Acre, em que a empresa poderia aumentar a sua atuação, dando o suporte nessa região do Brasil em que os fretes rodoviários não são tão recorrentes; bem como incentivar o uso da plataforma no estado de São Paulo para a postagens de cargas fracionadas, apresentando as vantagens também para os que detêm veículos de pequeno porte. Conclui-se que o CLARA trouxe resultados satisfatórios, diminuindo a complexidade computacional de uma base de mais de três milhões de entradas e revelando grupos de dados como uma oportunidade de crescimento da plataforma. Contudo, houve sobreposições de estruturas claramente avistadas como distintas nos gráficos de dispersão.

Palavras-chave: *Marketplace* logístico; transporte rodoviário; clusterização; Brasil.

ABSTRACT

Road transport is one of the most impactful modes in the global transportation matrix. Despite its advantages of flexibility and availability, the sector is characterized by high fragmentation. In the past, the intermediation process in this logistics chain was inefficiently handled by freight agents in terms of time and cost. An effective solution to address the need for agility and ease is through electronic logistics marketplaces, which are systems allowing carriers to advertise their loads to truck drivers searching for freight. However, the ease of automating load and capacity matching has resulted in technology providers dealing with an unprecedented volume of data. Valuable insights about user behavior can be derived from this diverse dataset. Despite the popularity of logistics marketplaces, scientific literature has not kept pace with their growth. Given these opportunities, this study aims to identify patterns in a cargo advertisement database of a logistics marketplace using clustering, which can assist in decision-making. Following the CRISP-DM procedure, data on load postings from 2019 to 2021 were collected, and the clustering trend of the database was confirmed using RStudio software. The CLARA algorithm was subsequently employed, and the quality of clusters was assessed using the Silhouette index. The most representative group at the national level consisted of freight within the state of São Paulo, featuring full loads, covering distances of around 500 km, and requiring heavy-duty vehicles for transportation. In the context of São Paulo, the most significant partition comprised full freight journeys of just over five hours, also requiring heavy-duty vehicles. The higher frequency of full freight was attributed to its benefits, such as efficiency in space and resource utilization. The main strategies identified for the national context involve offering progressive discounts on additional services to carriers conducting a high volume of trips and targeted promotion of the logistics marketplace to potential customers interested in transporting heavy loads over medium distances. An interesting business opportunity was identified in Acre, where the company could expand its operations and provide support in a region of Brazil where road freight is less common. Additionally, encouraging the use of the platform in São Paulo for posting fractional loads was suggested, highlighting the advantages for owners of small vehicles. In conclusion, CLARA produced satisfactory results by reducing the computational complexity of a database with over three million entries, and the study revealed data clusters as potential opportunities for platform growth. However, there were instances of overlaps of structures clearly seen as distinct in the scatterplots.

Keywords: Logistics e-marketplace; road transportation; clustering; Brazil.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	Controle Adaptativo de Cruzeiro
AGNES	*Agglomerative Nesting,* Encaixamento Aglomerativo
AI	*Artificial Intelligence,* Inteligência Artificial
B2B	*Business-to-business*
BI	*Business Intelligence,* Inteligência de Negócio
CIOT	Código Identificador de Operação de Transporte
CLARA	*Clustering Large Applications,* Clusterização de Grandes Aplicações
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CNT	Confederação Nacional do Transportes
COVID-19	Doença do coronavírus
CRAN	*Comprehensive R Archive Network*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DBSCAN	*Density Based Spatial Clustering and Applications with Noise,* Clusterização Espacial e Aplicações com Ruído Baseadas em Densidade
DIANA	*Divisive Analysis,* Análise Divisiva
EM	*Expectation-maximization*
ETL	*Extract, transform and load,* Extrair, transformar e carregar
FCM	*Fuzzy C-means*
FPN	*Frequent Pattern Network,* Rede de Padrões Frequentes
HKM	*Hierarchical K-means, K-means* Hierárquico
IEPG	Instituto de Engenharia de Produção e Gestão
IQR	*Interquartile Range,* Intervalo Interquartil
LogTranS	Grupo de Pesquisa em Logística, Transportes e Sustentabilidade
MAI	Mestrado Acadêmico para Inovação
PAM	*Partitioning Around Medoids*, Particionamento ao redor de Medoides
PCA	*Principal Component Analysis,* Análise do Componente Principal
PIB	Produto Interno Bruto
PRVCEL	Problema de Roteamento de Veículos de Coleta e Entrega com Lucro
qte	Quantidade
ROC	*Receiver Operating Characteristics*, característica de operação do receptor
SCM	*Supply Chain Managament,* Gestão da Cadeia de Suprimentos
WoS	Web of Science

SUMÁRIO

1 INTRODUÇÃO

O transporte rodoviário é uma das modalidades de maior impacto não só na matriz de transportes brasileira, mas também na de escala global. Por exemplo, na China, o mercado rodoviário registra saltos anuais no volume de fretes de 14,2%; chegou a totalizar cerca de 39 bilhões de toneladas métricas no ano de 2021 (CHINA, 2022). Nos Estados Unidos, outra grande economia mundial, segundo a American Trucking Associations (2020), pelas rodovias, movimentam-se cerca de 72% dos fretes nacionais. Já no Brasil, segundo a FreteBras (2022), o setor de fretes corresponde a 12,7% do Produto Interno Bruto (PIB) do país. Em 2021, conforme a organização, registrou-se o aumento de 37,6% no volume de cargas, mesmo em um cenário de pandemia.

Apesar das suas vantagens de flexibilidade e disponibilidade, o transporte rodoviário é marcado pela alta fragmentação, que causa competição no setor, elevando os custos envolvidos no processo de unir a demanda à oferta (COLLIGNON *et al.*, 2020; SAMPATH *et al.*, 2020; XIAO *et al.*, 2020). Uma grande parte das entidades de transporte é formada por pequenas e médias empresas. Nessas organizações, os processos de aquisição e de tomada de decisão em fretes são executados por apenas um responsável isolado e não por departamentos bem estabelecidos (COLLIGNON *et al.*, 2020).

Conforme Xiao, Xu, Liu e Liu (2020), outro fator apontado como causa para a intensa competição no setor é a sua segmentação. Há inúmeras possibilidades de combinação entre demanda, volume, tipo de carga e tamanho de veículos. Isso se deve à capacidade de uma variedade de caminhões comportarem o mesmo tipo de carga. Os autores batizam esse fenômeno de competição intersetorial.

Antigamente, o processo de intermediação de toda essa cadeia logística era operado por agentes de fretes. Através de um trabalho praticamente manual, realizavam a conexão entre quem precisava de capacidade para transportar suas cargas e quem tinha disponível os veículos, por meio de telefonemas ou mensagens de texto. Além das ineficiências inerentes às capacidades humanas e no pagamento de taxas pelo serviço prestado, o tempo até a execução efetiva do transporte nesse modelo era demasiadamente longo.

Uma solução para atender às necessidades de agilidade e facilidade são os *marketplaces* logísticos eletrônicos. Miller *et al*. (2020) os definem como sistemas eletrônicos que permitem as transportadoras anunciarem suas cargas publicamente para um vasto número de caminhoneiros em busca de fretes. Segundo esses autores e Jain *et al.* (2019), semelhantemente ao modelo de negócio de transporte urbano particular de passageiros, os *marketplaces* logísticos

conseguem combinar carga e capacidade em tempo real. Miller e Nie (2018) declaram que essa funcionalidade inovadora das plataformas proporcionou a mobilização, a consolidação e a gestão de grandes frotas por um considerável custo menor. Adicionalmente, houve a maximização de eficiência do sistema logístico por minimizar as ociosidades de capacidade e de tempo.

Investidores já vinham há anos percebendo que a indústria logística era promissora no processo de digitalização, mas só recentemente, com o advento das mídias sociais e dos *smartphones*, foi que várias *startups* iniciaram suas operações no ramo e estão recebendo grandes aportes (SHIPSTA, 2019). A Teleroute (www.teleroute.com), em operação na França e pioneira no ramo; a Truck Alliance, a maior da China (www.56qq.cn); a Coyote (www.coyote.com) e a Echo (www.echo.com), nos Estados Unidos, são exemplos de *marketplaces* logísticos de sucesso. Em países emergentes, como a África do Sul, pode-se identificar ações governamentais para a implementação de *marketplaces* logísticos como forma de aperfeiçoar a integração entre os modais de transporte (RANA, 2020). No Brasil, esse mercado também está em plena expansão. Um exemplo disso decorre do recém investimento de 100 milhões de reais recebido pela Sotran para a ampliação de sua base de usuários e o oferecimento de novos produtos digitais (FORBES, 2021).

1.1 Problema de pesquisa

Todavia, como consequência da facilidade da automação dos processos de correspondência entre carga e capacidade, as empresas provedoras de tecnologia dos *marketplaces* logísticos estão tendo que lidar com um volume de dados sem precedentes. Nessas plataformas, segundo Xiao; Gan; *et al.* (2020), Miller e Nie (2018) e Mallick *et al.* (2017), pode-se obter dados da mais rica variedade, que são gerados a todo momento: número de fretes por trecho/rota e data específica, o tipo de caminhão e a sua capacidade, o tipo de bem transportado, tamanho do veículo, hora do anúncio de carga, janela de tempo da entrega e até mesmo a localização do caminhão, idade e detalhes da licença de motorista.

Siqueira Junior (2021) conclui que, embora os volumes de informação estejam aumentando exponencialmente, apenas uma parcela deles é de fato entendida e analisada. Com essa grande quantidade de dados, enxergar padrões de comportamento dos usuários de um *marketplace* logístico é uma tarefa complexa demais para os métodos tradicionais; configurando-se como um problema típico de *Big Data*, que pode ser solucionado pelo *Machine Learning*, como dizem Ishwarappa e Anuradha (2015). Em uma era digital, extrair informações

valiosas de bases de dados robustas tem sido extremamente relevante para a sobrevivência das organizações.

Pode-se pontuar ainda que, apesar da sua popularização, existem poucos trabalhos que realizam análises sobre bases de dados de *marketplaces* logísticos. Os poucos exemplos consistem dos estudos de Miller e Nie (2018), Miller *et al.* (2020); Beraldi *et al.* (2021) e Mallick *et al.* (2017), que empregam técnicas da Heurística para construir modelos que otimizem a margem de lucro dos caminhoneiros; e Xiao, Xu, Liu, Yang, *et al.* (2020) que utilizam modelos de regressão para realizar previsões do preço proposto por fretes rodoviários. Somente Domashova e Zasypkina (2021) utilizam da técnica de agrupamento, a fim de detectar usuários atípicos em um *marketplace* logístico e, assim, evitar o impacto negativo deles na competição. Eles empregam dois algoritmos de clusterização avançada. Há, portanto, uma oportunidade de demonstrar se outros algoritmos são eficientes na tarefa de detectar padrões no comportamento de usuários de *marketplaces* logísticos. Outro fato importante é que, até o momento, não foi observado nenhum estudo sobre *marketplaces* logísticos no Brasil publicado em bases científicas relevantes, como Scopus e Web of Science.

No que se refere à pesquisa sobre *marketplaces* eletrônicos genéricos e que abordam a Mineração de Dados, pode-se verificar sua ênfase nos processos de negociação entre compradores e vendedores. Utilizando uma base de dados reais, Meida *et al.* (2019) identificam padrões de compra por meio do Mineração de Regras de Associação (*FP-Growth*). Já Oh *et al.* (2014) propõem a metodologia *Frequent Pattern Network* (FPN), comparando-a com o método *FP-Growth* e atesta sua melhor performance. Ainda no âmbito dos *marketplaces* eletrônicos genéricos, dos estudos que envolvem a clusterização, lista-se o de Ordanini *et al.* (2004) e Lorenzini (2014). Ambos especificam os fatores de sucesso e de falha desses modelos de negócio, empregando a técnica em dados oriundos de questionários. Contudo, implementando o método de clusterização sobre uma base de dados propriamente, destacam-se apenas Aravazhi Irissappane e Zhang (2017), que realizam o agrupamento a fim de identificar o perfil de recomendadores desonestos, enquanto avalia a reputação dos vendedores por vários critérios; e Pereira *et al.* (2009), que, por meio do algoritmo *K-Means*, consegue listar práticas e estratégias de venda. Tem-se, assim, a ocasião para aplicar a clusterização para reconhecer padrões de comportamento em um *marketplace* que atue em um ramo específico, tal como o mercado de fretes rodoviários.

1.2 Objetivo geral e específicos

Desta forma, tendo em vista as oportunidades listadas, este trabalho visa identificar padrões no processo de postagem de carga em um *marketplace* logístico por meio da clusterização. Tendo a premissa de que esses conjuntos de dados são clusterizáveis, será possível, então:

- Traçar estratégias para a empresa de como tratar os grupos resultantes do processo de clusterização, tendo por base suas características;
- Analisar quais localidades e tipologia de frete se destacam nas postagens de carga na plataforma eletrônica.

Será possível também:

- Apresentar o funcionamento e o comportamento do algoritmo de clusterização selecionado em uma base de dados concreta de grandes dimensões;
- Verificar os pontos positivos e negativos de empregar a clusterização frente a outras metodologias.
- Comparar os resultados da pesquisa com as colocações de autores da literatura em *marketplaces* logísticos.

1.3 Justificativa

De acordo com Agarwal e Dhar (2014) e Miller *et al*. (2020), as oportunidades que a disponibilidade de dados traz são grandes e pode-se obter *insights* tanto para um planejamento de poucos dias quanto de longos períodos. Waller e Fawcett (2013) afirmam que os dados podem revolucionar a dinâmica das cadeias de suprimento e abastecimento. Segundo Pereira *et al.* (2009), através dos *marketplaces* eletrônicos, é possível analisar o perfil dos anunciantes e da atividade econômica que praticam. E, para isso, Deng *et al.* (2016) apontam a metodologia de clusterização como fundamental para a segmentação de bases de dados, proveitosa para identificar as necessidades específicas de cada grupo. Oh *et al.* (2014) reiteram que descobrir padrões é essencial para consolidação de estratégias de *marketing* mais inteligentes e poderosas.

Este trabalho, portanto, pode ajudar às empresas de tecnologia em um *marketplace* logístico a efetuarem planejamento de ações de *marketing* tanto para aumentar a divulgação entre caminhoneiros e transportadoras usuárias da plataforma em locais onde a atividade de fretes é intensa, quanto buscar parcerias a longo prazo com os atuantes de determinada região para o oferecimento de novos serviços e especialidades. Compreendendo os preços praticados, as distâncias percorridas, as durações das viagens, o pesos das cargas transacionadas, os

provedores de tecnologia dos *marketplaces* logísticos podem oferecer serviços diferenciados, favorecendo o crescimento da base de usuários.

Domashova e Zasypkina (2021), em seu trabalho, conseguiram, por meio da clusterização, detectar usuários anormais (que operam na plataforma para adquirir dados comerciais de seus concorrentes) em um *marketplace* logístico e afirmam que sua ferramenta poderia ser usada pelos proprietários da empresa parceira, bem como por outras companhias que operam no mesmo ramo, portanto, este trabalho pode demonstrar, de forma prática, a aplicabilidade do método de agrupamento em outros contextos, como o de fretes rodoviários e do mercado brasileiro, reforçando os benefícios de suas características.

Este estudo também abrirá oportunidades para o desenvolvimento de mais pesquisas em *marketplaces* logísticos no Brasil, visto que nenhum trabalho sobre o tema foi identificado nas bases de documentos científicos. Acredita-se que esse fato de não haver trabalhos acadêmicos pode estar relacionado à natureza empresarial/técnica do modelo de negócio de um *marketplace* logístico eletrônico.

Por fim, este trabalho faz parte do Programa de Mestrado Acadêmico para Inovação (MAI) na Universidade Federal de Itajubá (UNIFEI), pelo Edital nº 01/2021, financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). O Programa contou com a parceria de uma empresa provedora de tecnologia de um *marketplace* logístico eletrônico, que se prontificou a ceder dados, bem como fornecer apoio tecnológico, técnico e financeiro a este projeto. Acredita-se que este projeto atuou de forma sólida na interface pesquisa e extensão, trazendo mais conhecimento sobre o mercado de fretes brasileiro e sobre as particularidades de um modelo de negócio ainda em ascensão na literatura científica, ao passo que auxilia a empresa parceira a estabelecer suas estratégias de atuação e de *marketing* no país.

1.4 Estrutura do trabalho

Esta dissertação está estruturada em seis capítulos. O primeiro é a introdução, que traz a contextualização do tema, problema de pesquisa, objetivos e justificativa. O Capítulo 2 dedica-se à apresentação da fundamentação teórica. Inicia-se com uma análise bibliométrica, conduzida para entender o estado da arte em *marketplaces* logísticos eletrônicos, bem como um diagnóstico espacial, temporal e de área das publicações. Os conceitos, funcionamento, classificações, vantagens e desvantagens dos *marketplaces* eletrônicos, dos *marketplaces* logísticos, da Mineração de Dados e, em específico, da clusterização são explanados em seguida.

O terceiro capítulo aborda a classificação da pesquisa, o funcionamento da empresa parceira do estudo, as características das bases de dados analisadas, bem como os passos seguidos para o desenvolvimento do trabalho. Já o Capítulo 4 apresenta os resultados obtidos na execução dos passos propostos na metodologia. O capítulo seguinte traz a discussão dos resultados e as respostas das questões centrais deste trabalho. Por fim, o sexto capítulo apresenta as conclusões do estudo, adicionalmente, discute as suas limitações e sugere temas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que ampara este trabalho. Os conceitos de *e-marketplaces,* Mineração de Dados e clusterização são apresentados de maneira a esclarecer seu funcionamento e suas técnicas. A fundamentação teórica compreende artigos em periódicos nacionais e internacionais, bem como em congressos, livros, teses e dissertações.

2.1 Análise bibliométrica

A análise bibliométrica permite analisar, de modo estruturado, o quanto de publicações envolve um determinado tema em fontes indexadas. Segundo Ellegaard e Wallin (2015), consiste em uma análise baseada na identificação do corpo da literatura dentro de uma determinada área de conhecimento. De acordo com os mesmos autores, originalmente, possibilita uma visão geral da produção científica e selecionar as publicações mais citadas e também as que envolvem um contexto geográfico ou institucional em específico.

A análise bibliométrica foi conduzida no mês de junho de 2022 e obteve sua última atualização em junho de 2023 com o objetivo de identificar os trabalhos concernentes aos *marketplaces* logísticos, entendendo seus conceitos, as lacunas de abordagem e as tendências de pesquisa. Primeiramente, definiu-se a estratégia de busca, como as plataformas, operadores lógicos e palavras-chave. As bases de dados consultadas foram a Scopus, da Elsevier e a Web of Science (WoS), da Thomson Reuters. Essas bases são apontadas como as principais bases de dados bibliográficas usadas para análises bibliométricas, fontes de citação e revisão de literatura, principalmente nas disciplinas da Engenharia (MONGEON; PAUL-HUS, 2016; PRANCKUTĖ, 2021; ZHU; LIU, 2020).

As palavras-chave foram definidas observando a indicação de Marasco (2004), Rios, (2018), Picão (2017) e Martins (2019). Portanto, a busca envolveu os termos "*logistics marketplace*", "*electronic logistics marketplace*", "*transportation electronic marketplace*", "*electronic transportation marketplace*", "*transport e-marketplace*", "*transportation e-marketplace*", "*online freight exchange*", "*logistics e-marketplace*", "*electronic freight exchange*", "*logistics service e-marketplace*" e "*e-logistics marketplace*". Foi usado o operador "OU", visto que esses termos são sinônimos. Os campos envolvidos foram títulos, resumos e palavras-chave.

Para análise dos artigos propriamente dita foram selecionados apenas trabalhos de acesso gratuito. Trabalhos presentes nas duas plataformas também foram desconsiderados. A pesquisa resultou em 41 publicações na Scopus e 21 na WoS. Todos os artigos presentes na

WoS também se encontravam disponíveis na base Scopus. Pôde-se ter o acesso completo a 24 trabalhos. O resultado retornou somente publicações no idioma inglês.

Os trabalhos foram organizados num planilha com as seguintes informações: título da publicação, autores, data de publicação, país de origem, palavras-chave, metodologia, técnica e periódico. A Figura 2.1 exibe o gráfico de distribuição temporal das publicações a respeito de *marketplace* logístico.

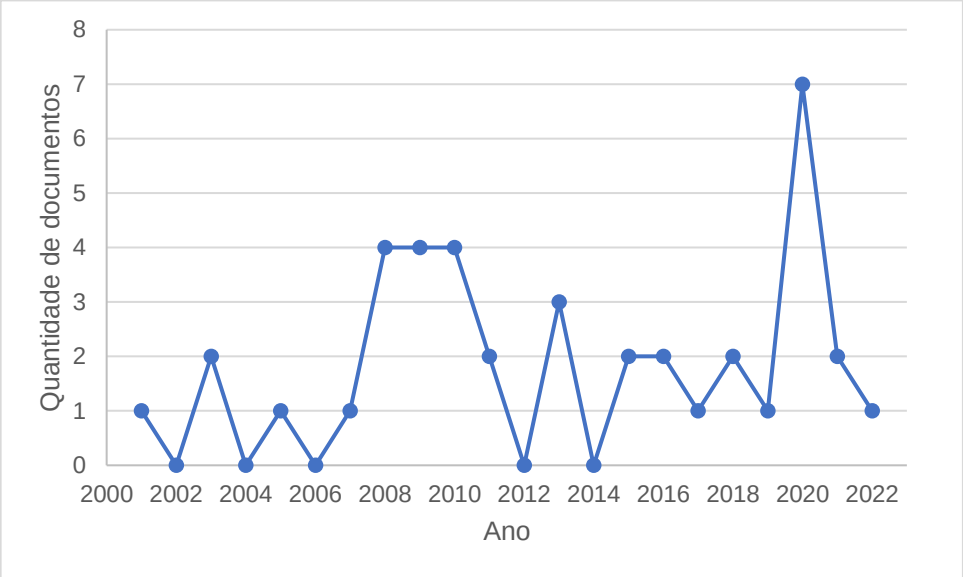


Figura 2.1 - Distribuição temporal de publicações em marketplace logístico

Fonte: Adaptado de Elsevier (2022)

Pode-se perceber que a quantidade de publicações a respeito dessa temática ainda é pouca. Por 21 anos de registro na plataforma, foram, em média, 1,9 trabalhos por ano. Em comparação à temática de *marketplace* eletrônico, são, em média, 30 trabalhos por ano. Os períodos de maior publicação sobre as plataformas de carga *online* são entre 2008 e 2010 e no ano de 2020. Foram esses, períodos de grande relevância para o mercado internacional. Em 2008, a crise econômica nos Estados Unidos e, em 2020, a pandemia do COVID-19 afetaram significativamente o panorama econômico de seus tempos. É plausível relacionar esses momentos à preocupação com a logística e o suprimento de recursos. É aceitável também relacionar, de certa forma, ao desemprego dessas épocas. Os *marketplaces* logísticos são tipicamente buscados para trabalhos eventuais, temporários e de contratação mais simples do que pelos modelos tradicionais. Esperava-se que, a partir de 2010, houvesse um aumento crescente de publicações no assunto, levando em consideração a popularização dos

smartphones e mídias sociais. Já a Tabela 2.1 exibe a distribuição geográfica dos trabalhos que abordam as plataformas de carga *online*.

Tabela 2.1 - Distribuição geográfica de publicações em marketplace logístico

País	Quantidade de publicações
Estados Unidos	8
Alemanha	7
China	6
Reino Unido	4
Grécia	3
Itália	3
Austrália	2
Hong Kong	2
Hungria	2
Índia	2
Malásia	2
Países Baixos	2

Fonte: Adaptado de Elsevier (2022)

Cabe ressaltar que Áustria, Canadá, Irã, Polônia, Eslováquia, Suíça e Ucrânia apresentam uma publicação cada. O país onde há mais publicações sobre *marketplaces* logísticos é os Estados Unidos, seguido por Alemanha e China. Pode-se traçar um paralelo com o nível de industrialização desses países, bem como a importância do modal rodoviário para estes. De acordo com ILOS (2020), nessas nações, o modo rodoviário é o de maior participação na matriz de transportes. Entretanto, nota-se que, em países emergentes como o Brasil, onde o transporte rodoviário é ainda mais expressivo, não houve nenhum publicação a respeito. Foi conduzida uma outra busca com os mesmos termos adicionando "*Brazil*" para identificar trabalhos sobre as plataformas de carga *online* no país, entretanto não houve o retorno de nenhum documento.

Com relação às áreas dos periódicos em que esses trabalhos foram publicados, destacam-se as áreas de Engenharia, com 29,3% do total, seguido pela Ciência da Computação (21,3%) e a Administração e Gestão (12%). Esperava-se que esses fossem os domínios mais comuns para publicação, haja vista das características inerentes aos *marketplaces* logísticos.

No que diz respeito à metodologia, foi possível identificar dois grupos majoritários de técnicas no estudo dos *marketplaces* logísticos: os que empregavam a modelagem e os que conduziram *surveys* ou construção de conceitos. Miller e Nie (2018), Miller *et al.* (2020) e Beraldi *et al.* (2021) estudaram o processo de roteamento dos veículos, visando encontrar um modelo que otimizasse os lucros dos caminhoneiros. Esses autores aplicam as técnicas de heurística, como o Processo de Decisão de Markov e o Problema de Roteamento de Veículos de Coleta e Entrega com Lucro (PRVCEL).

Mallick *et al.* (2017) empregam também a heurística (Processo de Decisão de Markov), só que para criar um algoritmo que recomenda o anúncio de cargas mais vantajoso para um caminhoneiro fazer seus lances em um modelo de *marketplace* baseado em leilões.

Os outros trabalhos desse grupo utilizam técnicas da estatística para realizar previsões. Xiao, Xu, Liu, Yang, *et al.* (2020) estabelecem um modelo de previsão para o preço oferecido por quilômetro por tonelada em uma plataforma de cargas na China. Já Xiao, Gan, *et al.* (2020) calculam o grau de variação do volume de frete no mercado chinês pelo tempo através da regressão. Por fim, Xiao, Xu, Liu e Liu (2020) verificam as variações do volume de frete entre as categorias de caminhões pesados. Todos os trabalhos descritos utilizaram dados reais de um *marketplace* de cargas e foram satisfatórios em seus resultados no que se refere a seus objetivos. Aponta-se a oportunidade de trabalhar as bases de dados com outras técnicas, por exemplo, a clusterização da Mineração de Dados.

Três trabalhos estudam os mecanismos de lances de *marketplaces* de cargas baseados em leilões. Motlagh *et al.* (2010), Xu e Huang (2013) e Mallick *et al*. (2017) analisam principalmente o recurso de leilões duplos, que é aquele que permite lances simultâneos tanto dos tomadores quanto dos ofertantes dos serviços logísticos. Motlagh *et al*. (2010) conseguem identificar que leilões combinatórios duplos são mais vantajosos que os leilões combinatórios simples. Já Xu e Huang (2013) conseguem concluir que os *marketplaces* logísticos são mais prováveis de obter maiores lucros com o uso dos leilões duplos com uma duração de leilão relativamente curta.

Dois trabalhos envolvem a proposta de trazer para os *marketplaces* logísticos a tecnologia de *blockchain*, definido por um sistema que possibilita o rastreamento do envio e do recebimento de informações pela Internet. Sampath *et al.* (2020) e Wester e Otto (2021) se preocupam com a questão da segurança do tráfego de informações nas interfaces. Ambos os artigos propõem um modelo de negócio sem a presença de um intermediador provedor de tecnologia. Ambos os trabalhos pontuam que a adoção de um sistema sem um provedor possibilitaria grande economia de custos.

De uma forma geral, pode-se pontuar que os trabalhos de modelagem em *marketplace* logístico focaram na experiência de operação do caminhoneiro. Indica-se, portanto, uma grande oportunidade de empregar a simulação para modelar o comportamento das transportadoras nessas plataformas digitais.

Tratando, agora, do outro grupo de pesquisas, que envolvem *surveys* e a construção de conceitos, dois trabalhos tratam de *marketplaces* logísticos com enfoque no modo ferroviário. São os de Endemann *et al.* (2012) e Föhring e Zelewski (2015). Já Jain *et al.* (2019) trazem uma

abordagem multimodal. Todos esses trabalhos se preocupam em identificar os fatores de sucesso para a implementação desses modelos na Europa, continente em que a integração dos modais rodoviário e ferroviário é bem desenvolvida. O método empregado foi o de entrevistas com representantes de empresas europeias, para que sejam apontadas as urgências. Esses trabalhos mostram um baixo nível de maturidade da intermodalidade no continente e um grande potencial para os modelos dessa natureza no mundo todo.

Wang *et al*. (2007, 2011) são os principais trabalhos com relação a *marketplaces* logísticos fechados. Em seu primeiro trabalho, foram analisados seis estudos de caso para identificar as vantagens e desvantagens dessa tipologia de plataforma de cargas. No seu segundo estudo, introduzem o conceito de *marketplaces* colaborativos. Utilizam-se de entrevistas semiestruturadas, mapeamento de processos e demonstrações do sistema. Registra-se que Chenyan *et al.* (2009) tenha trabalhado melhor o conceito de colaboração na logística, identificando as vantagens e desafios de promovê-la por meio dos *marketplaces.* Por fim, Nandiraju e Regan (2008) discutem os modelos operacionais que seguem os atuais *marketplaces* de cargas e, então, listam os principais problemas enfrentados pelos seus participantes, como a combinação das cargas, a contratação a longo tempo e as estratégias de colaboração. Vale ressaltar que os estudos de Wang *et al*. (2007, 2011) e Nandiraju e Regan (2008) são os únicos a envolverem a perspectiva de todos os participantes, isto é, transportadoras, caminhoneiros e provedores de tecnologia.

Os trabalhos de Kovács (2009, 2010) se encarregam de estudar os benefícios e as desvantagens de unir o intercâmbio de frete com o de estoque em uma única plataforma. Demonstra os objetivos, a estrutura e os serviços que um empreendimento como este teria idealmente. O autor conclui que as plataformas de carga *online* produzem efeitos de aumento da competição no mercado, diminuição dos custos de frete, melhor organização dos carregamentos, o aparecimento de uma extensiva base de dados de fretadores e o melhoramento dos padrões dos serviços de transporte.

Outros dois trabalhos buscam identificar a adoção dos *marketplaces* para serviços de transporte em um contexto industrial particular. Manthou *et al.* (2005) através de uma *survey* identificam as percepções dos benefícios e das limitações do uso das plataformas digitais no setor de comida enlatada. Igualmente Salleh *et al.* (2009) identificam a possibilidade de implementação do modelo de negócio para atender as necessidades das empresas do agronegócio. Ambos os trabalhos concluem a urgência da adoção dos *marketplaces* logísticos para aperfeiçoar a logística e as atividades de distribuição dos respectivos países.

Por fim, Golsby e Eckert (2003) trazem uma abordagem mais financeira para a temática. Esses autores aplicam a estrutura de Avaliação do Custo de Transação para reconhecer quando é mais proveitoso um transportador operar por meio de um *marketplace* logístico ou desempenhar esses serviços com recursos próprios. Eles apontaram que transportadoras que valorizam o relacionamento com caminhoneiros em específico irão evitar o uso de plataformas públicas e neutras como primeira opção para o transporte de suas cargas.

É cabível pontuar que os trabalhos que utilizaram de técnicas qualitativas tiveram um enfoque na experiência das transportadoras. Indica-se, por isto, uma oportunidade válida de se investigar mais a percepção das vantagens e limitações dos *marketplaces* logísticos por parte dos caminhoneiros, embora seja esta uma tarefa mais desafiadora do que conduzir pesquisas com as transportadoras, haja vista da grande população.

Acrescentando à lista de observações, foi possível constatar que se o estudo que se pretende conduzir objetiva abordar o aspecto mais operacional dos *marketplaces* logísticos, as técnicas mais empregadas são as de caráter quantitativo. Por sua vez, se a pesquisa delineia uma abordagem dos aspectos táticos e estratégicos das plataformas de carga *online,* os métodos mais aplicados são os de cunho qualitativo.

Por fim, a Figura 2.2 ilustra as palavras-chave mais indicadas pelos autores dos documentos analisados através de uma nuvem de palavras, que foi elaborada com o auxílio do Pro Word Cloud, um suplemento do Microsoft Office®. Verifica-se que o termo "*shipper*", ou seja, transportadora em inglês, precisa ser mais explorado. Não ocorre os termos de "*data mining*" e nem clusterização, apontando oportunidade dessas ferramentas. Os conceitos intermodal e multimodal não estão em destaque, possibilitando novos trabalhos nesta proposta tão necessária atualmente. Identifica-se, de igual modo, oportunidades de abordar as outras entidades que possam integrar o *marketplace* logístico, como as gerenciadoras de risco e as mediadoras de crédito. Por fim, os termos de maior destaque são "*collaboration*", colaboração em inglês, "*time*", tempo e "*commerce*", comércio; sugerindo serem estes aspectos importantes a serem tratados em qualquer estudos sobre a temática aqui apresentada.

Figura 2.2 - Nuvem de termos relacionados a marketplaces logísticos

2.2 *Marketplaces* eletrônicos

Os *marketplaces* eletrônicos são plataformas digitais que permitem a realização de negócios entre fornecedores e compradores. Bakos (1991), um dos pioneiros nos estudos científicos em comércio eletrônico, define o termo como um sistema interorganizacional que possibilita aos compradores e vendedores o intercâmbio de informações sobre ofertas de produtos e serviços. Já Janita e Miranda (2013) o interpretam como uma forma de fazer negócio, utilizando a tecnologia da Internet para reunir múltiplas empresas em uma única plataforma. Chang e Wong (2010) dão a sua definição como plataformas organizacionais que permitem os integrantes a troca de informações sobre preços e ofertas.

Picão (2017) e Wang *et al.* (2007) afirmam que os *marketplaces* eletrônicos também recebem as terminologias de *marketspace*, intercâmbio eletrônico, mercado eletrônico, *e-hub*, rede eletrônica e comércio eletrônico. Tipicamente, se referem aos modelos de negócio que atuam no *business-to-business* (B2B), isto é, nas transações realizadas de empresa para empresa.

Historicamente, os *marketplaces*, segundo Grieger (2003), eram locais de encontro físico entre vendedores e compradores. Eles anunciavam suas intenções e eventualmente tinham suas necessidades correspondidas. Só que, com a demanda por agilidade e facilidade, esse conceito foi remodelado, impulsionado pela evolução das redes sociais e da Internet (KUBOTA; MILANI, 2011).

Os *marketplaces* eletrônicos foram concebidos inicialmente como meio de ampliação de mercado e também como forma de recuperar o capital investido de produtos não vendidos

no mercado convencional. Mas, com o tempo, passou a fazer parte das principais operações das cadeias de suprimento. Muitas até deixaram de operar no mercado físico. Chang e Wong (2010) declaram que os principais fatores para isso foram a atração pela economia de tempo e de custo obtida nas plataformas; bem como a tendência de informatização dos concorrentes.

A ideia desses modelos de negócio é possuir a informação dos produtos e serviços em tempo real, na garantia de usufruir de economias de escala e liquidez. É também um mecanismo de mitigar a complexidade do processo e o custo de integração entre as empresas (PICÃO, 2017; WANG *et al.*, 2007).

Em suma, Turban *et al.* (2015) listam que os *e-hubs* possuem quatro funções principais, muito semelhantes ao mercado físico: possibilitar transações comerciais, permitir fluxos de informações, tornar disponível serviços de pagamento e dar a provisão de serviços auxiliares, a título de exemplo, seguros e documentos fiscais. Contudo, por serem virtuais, as atualizações e melhorias desses serviços ocorrem de uma forma muito mais rápida e eficiente.

Geralmente, as entidades participantes de um *marketplace* eletrônico são o comprador, o vendedor e o provedor de tecnologia. O provedor de tecnologia é o agente que se encarregará de manter a automatização das transações na plataforma, bem como cuidar da proteção dos dados e oferecer o suporte técnico. Outros agentes podem integrar o ambiente, mas isso depende da complexidade das operações. Os mais comuns são as gerenciadoras de riscos e as oferecedoras de crédito. Grieger (2003) faz um adendo, afirmando que deve haver múltiplos compradores e múltiplos fornecedores para a caracterização de um *marketplace* eletrônico.

Stockdale e Standing (2002) e Grieger (2003) provêm uma estruturação de caracterização dos *marketplaces* da seguinte forma:

- **Pela ênfase nas partes envolvidas**: a plataforma pode estar focada em atender os interesses do comprador ou do fornecedor; ou atuar de maneira neutra. Normalmente, os que são operados por um provedor de tecnologia são neutros;
- **Pelo foco do mercado**: podem ser verticais ou horizontais. Os primeiros atendem a um nicho específico, já os últimos não se restringem a um setor pré-determinado;
- **Pelo método de transação**: leilões, intercâmbio ou catálogo;
- **Pela acessibilidade**: podem ser abertos ou fechados. No primeiro caso, qualquer organização pode participar do sistema; já no segundo, é necessário um convite;
- **Pelo suporte à transação**: auxiliam até a fase de informação, ou a fase de negociação, a fase de execução e até a fase de pós-venda; por último,

- **Pelo mecanismo de mercado**: agregação ou correspondência. Na agregação, os próprios integrantes fazem a busca pelos seus melhores anúncios. Já na correspondência, os algoritmos automatizam esse processo.

Tratando do processo transacional dos *marketplaces* eletrônicos, segundo Schmid e Lindemann (1998) acontecem em três fases:

1) **Etapa da informação**: são providenciadas informações sobre as entidades participantes, sua atuação, seus documentos legais, registros e certificações para que possam operar. Além do mais, são providenciadas as informações do anúncio que farão;
2) **Etapa da negociação**: são executados os processos de correspondência entre oferta e demanda, bem como o fechamento da contrato;
3) **Etapa da execução**: os serviços e/ou produtos são realizados/entregues. As documentações fiscais são expedidas. Os pagamentos são oficializados.

Grieger (2003) acrescenta a fase de pós-venda, que se refere aos processos de avaliação da qualidade dos serviços e de satisfação dos produtos, como também de trocas e devoluções. A Figura 2.3 resume de uma forma gráfica o processo transacional.

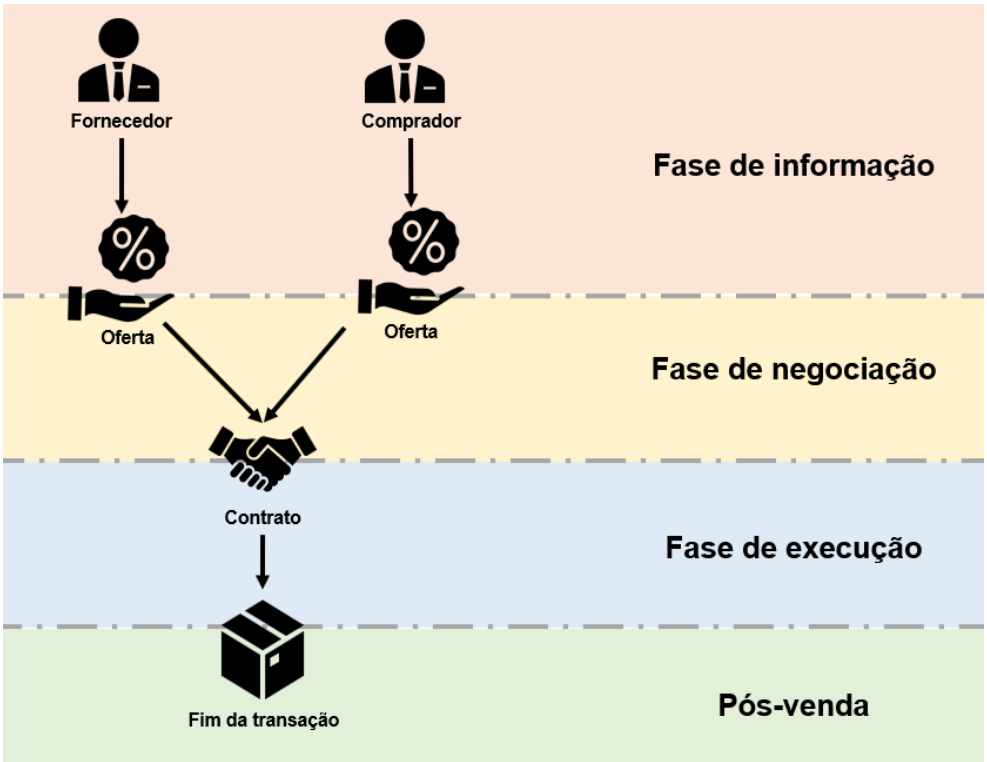


Figura 2.3 - Fases de transação num marketplace eletrônico

Fonte: Adaptado de Grieger (2003) e Schmid e Lindemann (1998)

Atualmente, está cada vez mais raro perceber *marketplaces* que só cobrem a fase informacional da transação. Embora seja uma questão de sobrevivência no mercado, muitas empresas encaram como oportunidade de lucro oferecer mais serviços de suporte. Ordanini *et al*. (2004) declaram que a plataforma mais completa é a mais procurada pelas partes interessadas, ao passo que oferece soluções inovativas para as necessidades empresariais.

A literatura é unânime quanto à listagem dos benefícios dos *e-marketplaces*. Conforme Wang *et al.* (2007), eles oferecem custo de procura mais baixo, custos de transação reduzidos; flexibilidade, automação do processos de negócio; mitigação dos custos com estoque; melhoramento da qualidade dos serviços pela ampliação da base de compradores e de vendedores, além disso ampliação geográfica. Picão (2017) menciona que tudo isso é possível pelo fato de que os integrantes não estão fisicamente presentes, tornando possível a transação de qualquer lugar do mundo com acesso à Internet.

Entretanto, é descrito na literatura também que o processo de ganhar novos usuários nesse modelo de negócio pode ser complicado e caro. Pela rapidez do processo, a confiança entre os participantes se torna um elemento crítico (JANITA; MIRANDA, 2013).

Stockdale e Standing (2002) reiteram que muitos empresários são relutantes em operar nos *e-marketplaces*, pois têm dificuldades de desenvolver uma estratégia de vendas no ambiente. Marasco (2004) acrescenta que a confiabilidade e a veracidade das certificações dos participantes e a segurança dos dados formam um desafio ainda maior do que encontrar o menor preço. Chang e Wong (2010) reforçam que a confiança é um pré-requisito para o uso dessas plataformas digitais.

2.3 *Marketplaces* logísticos eletrônicos

Marketplaces logísticos são um modelo de *marketplace* eletrônico especializado em serviços logísticos (WANG *et al.*, 2007; WESTER; OTTO, 2021). Atuam especificamente na esfera do *business-to-business* (B2B) (CHENYAN *et al.*, 2009). Schwind *et al.* (2011) e Collignon *et al.* (2020) concordam que são interfaces eletrônicas que permitem a união entre transportadoras e caminhoneiros, no intuito da troca de serviços logísticos. Segundo os mesmos autores, as transportadoras anunciam a necessidade de serviços de transporte e os caminhoneiros disponibilizam a capacidade de carga de seus veículos. O principal objetivo, de acordo com Kovács (2010), é a sincronização da demanda dos transportadores e a oferta dos caminhoneiros com o auxílio de um modelo de negócio eletrônico.

Os *marketplaces* logísticos eletrônicos também são denominados de plataformas de transporte de carga *online*, *marketplaces* eletrônicos de transporte, *marketplace* de cargas, plataformas de intercâmbio de cargas, como sugerem Picão (2017) e Marasco (2004). Em inglês, existem mais possibilidades de termos sinônimos, como consta na Seção de Análise Bibliométrica. Wang *et al.* (2007) sugerem que há três elementos básicos num *marketplace* logístico, conforme ilustra a Figura 2.4: um provedor de tecnologia, as transportadoras e os caminhoneiros. Pode haver a interação de mais entidades, mas isso depende das características das funções ofertadas, como, por exemplo, uma mediadora de crédito ou uma gerenciadora de riscos. Outros autores identificam a substituição da figura de vários caminhoneiros autônomos como prestadores do serviço logístico por transportadoras, que possuem a capacidade de transporte, porém, não têm cargas para movimentar (MARTINS, 2019).

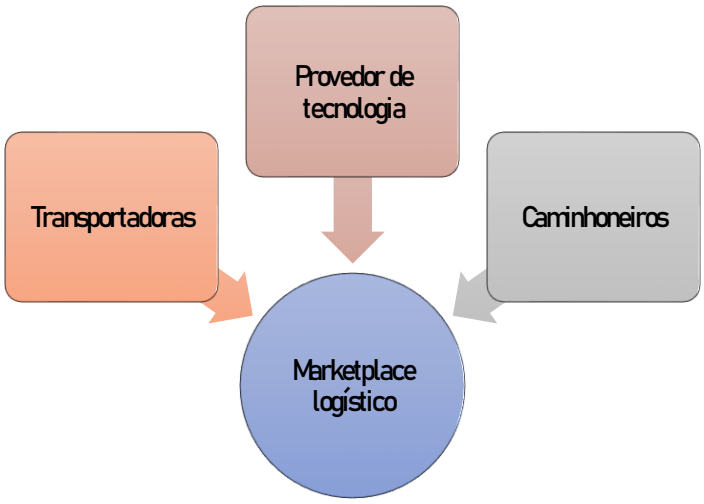


Figura 2.4 - Entidades integrantes de um marketplace logístico

Fonte: Adaptado de Wang *et al.* (2007)

Apesar de sua popularização atual, as plataformas de carga online não são recentes. Seu conceito básico nasceu nos anos 1970 quando, em conformidade com Alt e Klein (1998), uma organização alemã desenvolveu uma base de dados com informações de fretes e capacidade de carga que podia ser acessada pelos integrantes por meio de telefonemas. Miller e Nie (2018) afirmam que o primeiro modelo de negócio propriamente dito data do ano de 1985, pela Teleroute, com atuação por toda a Europa. A ideia com o seu surgimento era de reunir os anúncios de fretes em uma única interface, trazendo eficiência na busca.

Nandiraju e Regan (2008) declaram que, de início, a aceitação desse novo modelo comercial foi difícil. Eles falharam em obter a participação no mercado necessária. Marasco

(2004) afirma que os caminhoneiros foram relutantes na adoção do sistema devido à diminuição das margens de lucro, relacionada à disponibilidade de comparação dos preços, o que dificultava a estratégia de precificação desses integrantes. O mesmo autor pontua que isso fez com que muitos *marketplaces* logísticos encerrassem suas operações ou se fundissem a outras organizações. Outros fatores observados foram a preocupação com a segurança das informações deixadas na plataforma, dificuldades de navegabilidade e a precariedade do *design* da interface.

Com o crescimento do acesso à Internet, *smartphone* e mídias sociais, iniciado nos anos 2000, os *marketplaces* logísticos abriram novos canais de venda e alguns desses problemas foram de certa forma lidados. Föhring e Zelewski (2015) e Salleh *et al*. (2009) afirmam que esses sistemas evoluíram de uma simples plataforma de combinação de cargas para um recurso importante de gestão holística dos serviços logísticos. Esses comércios eletrônicos passaram a disponibilizar também outros serviços, como a integração com o sistema de controle de custos dos seus clientes, a gestão de documentos, planejamento das rotas, rastreamento de cargas, gerenciamento da janela de tempo, segurança eletrônica, mediação de créditos, processamento de pedidos e estatísticas (SALLEH *et al.*, 2009; SCHWIND *et al*., 2011). A adição desses serviços de valor agregado maior fez com houvesse fidelização dos usuários e a aumento da percepção de importância dos *marketplaces*.

Nandiraju e Regan (2008) indicam que historicamente os *marketplaces* logísticos têm sido recorridos para o transporte de demandas não previstas ou não cumpridas. Os mesmos autores afirmam que, no tocante às cargas esperadas, as transportadoras as delegam para seus caminhoneiros contratados. Entretanto, Wang *et al.* (2007) declaram que o uso das plataformas digitais de cargas não se limitam ao mercado à vista (*spot market*), mas também a firmação de contratos de longa duração.

As negociações entre as partes em um *marketplace* logístico são de um para um ou de um para muitos. Diferentemente do sistema tradicional de intermediação de fretes, a maioria dos provedores de tecnologia não participam do processo de contratação propriamente (SALLEH *et al.*, 2009). Geralmente, as transportadoras realizam o anúncio de suas cargas e também fazem a busca por motoristas compatíveis. Os caminhoneiros entram na plataforma em busca de fretes que haja correspondência com o seu veículo. As partes entram em contato mútuo. Visando a segurança das cargas, as transportadoras consultam o perfil do motorista em gerenciadoras de risco. A documentação do caminhoneiro é solicitada e, com o seu perfil aprovado, firmam o contrato de transporte. É comum a liberação de parte do pagamento ao

motorista para a cobertura de despesas, como pedágios e combustível. Ao findar do serviço, é realizado o pagamento integral.

Os *marketplaces* logísticos podem ser dimensionados de várias maneiras. Schwind *et al.* (2011) os organizam de acordo com oito atributos: pelo operador, foco do mercado, suporte à transação; mecanismo da transação, acessibilidade do mercado, estruturação dos contratos, serviços negociados e a fonte de lucro. Marasco (2004) acrescenta os atributos do viés e do modo de transporte.

Segundo o operador, Schwind *et al.* (2011) alegam que os *marketplaces* logísticos podem ser operados por um grupo de transportadoras ou um grupo de caminhoneiros, numa espécie de consórcio, sendo que estes participam do processo de negociação focando nos seus próprios interesses; ou operados por um provedor de tecnologia totalmente neutro. Grieger (2003) acrescenta que há também os que são desenvolvidos por um único operador autônomo. Essa classificação comuta semelhanças com a de Marasco (2004) quanto ao viés.

De acordo com a acessibilidade, podem ser abertos ou fechados, isto é, públicos ou privados (NANDIRAJU; REGAN, 2008). Os *marketplaces* abertos não possuem restrição de acesso. Já os de sistema fechado estão disponíveis a apenas um grupo específico de usuários, que podem se candidatar à membresia ou receber o convite à integração. Wang *et al.* (2011) ainda trazem um subgrupo dentro dessa tipologia, que são denominados de colaborativos. Nestes, seus integrantes se organizam num estrutura heterárquica. Os autores afirmam que cada participante pode entrar em contato com o outro e o conhecimento derivado da rede de contato é o que age como supervisor das operações. A principal motivação por trás da adoção de um *marketplace* logístico como este é de caráter relacional, de modo a eliminar as complexidades de comunicação.

No tocante ao foco no mercado, os *marketplaces* podem ser classificados em verticais ou horizontais. Os horizontais abrangem a necessidade de um setor de atividade econômica específico, como, por exemplo, o agronegócio. Já os verticais, também chamados de universais, atendem a todo tipo de demanda (SCHWIND *et al*., 2011). Marasco (2004) inclui nessa distinção a organização dos *marketplaces* logísticos de acordo com sua especialização em um modal de transporte. A autora declara que a maioria das plataformas de carga estão concentradas no transporte rodoviário, mas que existem as que se especializaram no transporte aéreo, ferroviário e marítimo. Com o advento da globalização, é uma tendência cada vez mais presente a integração entre os sistemas de transporte, surgindo assim os *marketplaces* logísticos multimodais.

As plataformas de carga *online* também podem se dividir pela extensão do seu suporte dado no processo transacional. Existem modelos de negócio que só abrangem o processo de submissão do anúncio, garantindo a sua visibilidade. Já outros oferecem apoio na fase de negociação, auxiliando na etapa de verificação de certificação dos caminhoneiros, por exemplo. Por último, existem aqueles que dão assistência também na fase de execução, garantindo a emissão dos documentos oficiais e meio de pagamento (SCHWIND *et al.*, 2011).

No que se refere à forma como provêm o serviço, isto é, o mecanismo da transação, podem ser divididos em quadro informativo de fretes, sistema baseado em leilões e intercâmbio de fretes (NANDIRAJU; REGAN, 2008). Segundo Collignon *et al.* (2020), o primeiro modelo é muito empregado em sistemas abertos. É caracterizado por apenas apresentar uma listagem dos anúncios. Os mesmos autores explicam que a preferência por essa modalidade se deve à flexibilidade na comunicação com o potencial parceiro. Já no modelo baseado em leilões, podem ser operados pelos tomadores ou pelos prestadores do serviço logístico ou na forma de leilões reversos, em que a transportadora anuncia uma carga e indica os critérios de seleção; e os caminhoneiros oferecem seus lances (MARASCO, 2004). A transportadora elege a menor dessas ofertas (MALLICK *et al.*, 2017). Já no intercâmbio de fretes, as transportadoras anunciam suas necessidades de serviços e os caminhoneiros postam sua capacidade de carga simultaneamente. O provedor de tecnologia, por sua vez, executa a combinação desses dois processos por um preço competitivo (NANDIRAJU; REGAN, 2008).

Acerca da estruturação dos contratos, os *marketplaces* logísticos podem focar em tanto contratos de curto prazo quanto os de longa duração. Sobre os serviços negociados, Schwind *et al.* (2011) pontuam que há modelos de negócio que negociam apenas as cargas postadas necessitadas de transporte, outros que só oferecem a capacidade ociosa de caminhões, e ainda os que disponibilizam ambos. Estes últimos são denominando pelos autores como intercâmbio de fretes clássicos.

Por fim, as fontes de lucro de um *marketplace* logístico podem ser percebidas pelo pagamento de taxas relacionadas à transação, que consiste na retenção de um percentual por cada frete realizado dentro da plataforma; pagamento pela membresia ou pelo licenciamento de usuários registrados de forma única, mensal ou anual; a venda de informações da indústria e do mercado; pelas taxas de prestação de serviços de valor agregado; e o lucro advindo de anúncios e propagandas (STOCKDALE; STANDING, 2002). É comum que os *marketplaces* considerem mais de uma fonte de lucro em sua plataforma. A Figura 2.5 resume o esquema de caracterização dos *marketplaces* logísticos.

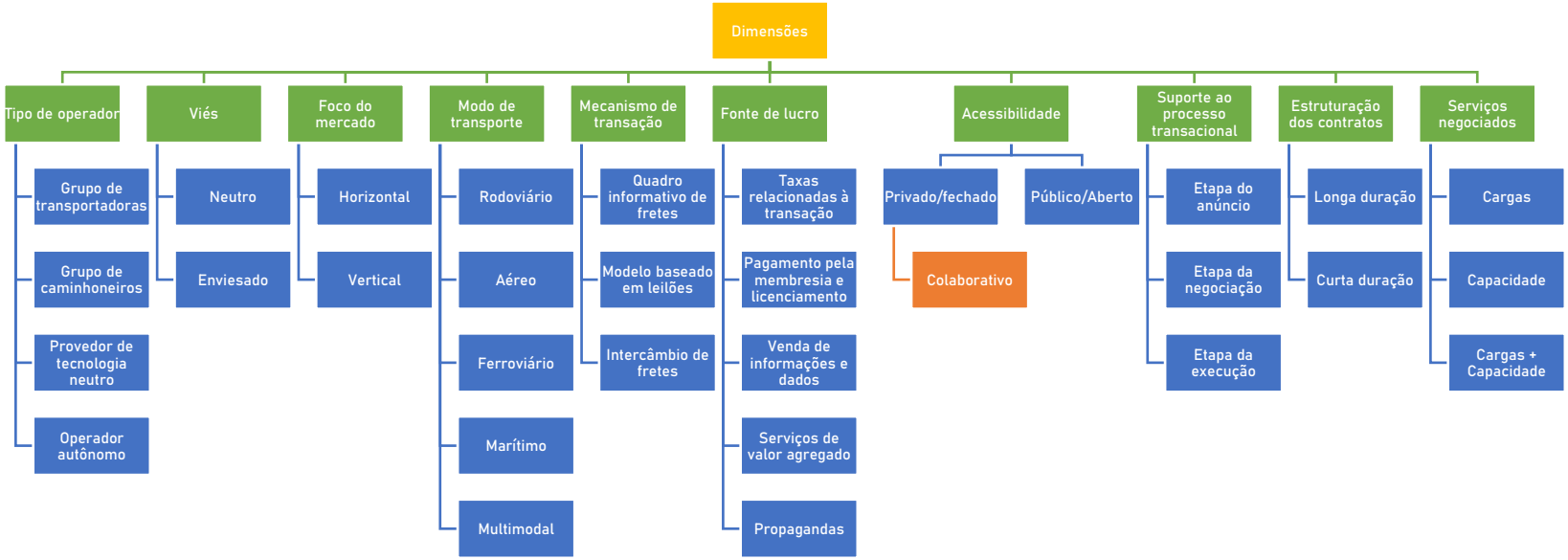


Figura 2.5 - Classificação dos marketplaces logísticos

Fonte: Adaptado de Marasco (2004) e Schwind *et al.* (2011)

A respeito da interação entre os agentes participantes de um *marketplace* logístico, Chenyan *et al*. (2009) tratam da colaboração como o fluxo de informação compartilhada. O nível de colaboração depende muito da forma como as plataformas de carga *online* operam e as características da atividade comercial das entidades integrantes. Há setores que são mais interdependentes do que outros. Os autores indicam três esferas de relacionamento: entre transportadoras, entre caminhoneiros e entre transportadoras e caminhoneiros. Para cada uma dessas esferas, existem três níveis: a menos intensa é a cooperação, passando pela coordenação e, por último, a sinergia. Na cooperação, as relações de colaboração são eventuais. Por exemplo, uma transportadora com um embarque menor que uma carga completa se associa a outra que tenha a mesma necessidade com o mesmo tipo de carga, destino e restrições de tempo. Um caminhoneiro com demanda acima da sua capacidade pode passar o excedente para outro. Ou ainda, dois caminhoneiros podem consolidar suas demandas a fim de aumentar a capacidade de um fretamento específico. Entre transportadoras e caminhoneiros, há simplesmente um monitoramento entre as partes para o cumprimento do frete, tanto em questão do serviço quanto do pagamento por ele.

Já no nível mais elevado de interação, as partes planejam a longo prazo. Na sinergia, as transportadoras elegem um representante para harmonizar as diferentes necessidades entre os integrantes e intermediam negociações com os caminhoneiros. Entre os caminhoneiros, há o compartilhamento de lucros, de informação da demanda de fretes e também os custos operacionais. Entre transportadoras e caminhoneiros, há grandes investimentos para a manutenção de um relacionamento próximo. São responsivos e compartilham informações gerenciais de alta confidencialidade (CHENYAN *et al.*, 2009). A tendência é que em *marketplaces* fechados, as partes colaboram entre si em um nível sinérgico. Já nas plataformas de carga *online* abertas, os integrantes mantêm apenas um vínculo cooperativo entre si.

Como vantagens do uso de um *marketplace* logístico, na visão de uma transportadora, pode-se pontuar o acesso rápido e fácil a uma vasta gama de parceiros e oportunidades. Estas oportunidades incluem a ampliação geográfica, isto é, a atuação em regiões que antes não seria possível o atendimento. Kovács (2010) afirma que há o desenvolvimento de cooperação entre os participantes e o uso de tecnologia da informação mais modernas e avançadas. Ele acrescenta que seu uso se torna significante, pois é possível encontrar fretes com a capacidade de utilização otimizada. A transportadora pode melhorar a qualidade de seus serviços logísticos, pela diversificação de sua frota. Além disso, a segurança das cargas é aumentada pelo rastreamento dos veículos.

Para os caminhoneiros, a organização de viagens de retorno com carregamento se torna mais fácil, diminuindo as viagens com carga vazia. Kovács (2009) descreve que, com a construção de uma base de dados robusta, é possível a comparação rápida dos preços de frete e com isso a eliminação dos valores absurdos, favorecendo a competitividade justa. O caminhoneiro consegue customizar a rota que deseja realizar e fazer o planejamento de suas viagens. Pode minimizar o seu tempo de ociosidade e contar com a segurança do pagamento de seus trabalhos. Para ambas as entidades, Golsby e Eckert (2003) e Marasco (2004) afirmam que, de uma perspectiva essencialista, o *marketplace* logístico suaviza a complexidade do processo de negociação e diminui o custo da transação, pela automatização que possibilita a eficiência do processo, a redução do corpo de colaboradores, economia de tempo e melhoria do fluxo de informações.

Entretanto, alguns desafios ainda precisam ser vencidos. Os *marketplaces* precisam aumentar o senso de confiança a seus usuários, já que as informações presentes na plataforma podem ser consideradas como competitivas, causando resistência na adesão. Föhring e Zelewski (2015) dizem que se um *marketplace* logístico falha em abranger o público-alvo crítico por falta de aderência e opera numa taxa de intermediação menor que 5%, a solução trazida acaba não compensando, e as transportadoras e caminhoneiros fecham contratos por si só, da forma tradicional.

Outro problema relatado por Wang *et al*. (2011) diz respeito ao fato de que muitos caminhoneiros se sentem julgados pelos contratantes apenas pela quantidade de viagens já operadas na plataforma e não pela qualidade dos serviços nomeadamente. Isso também torna complicada a obtenção de trabalhos por parte de novos usuários. Adicionalmente, Nandiraju e Regan (2008) apontam que transportadoras não aderem à solução dos *marketplaces* de cargas, pois essas organizações não tomam responsabilidade alguma pela movimentação dos fretes.

2.4 Ciência dos Dados – *Data Science*

Provost e Fawcett (2013) afirmam que a Ciência dos Dados envolve processos, técnicas e princípios que buscam entender um fenômeno por meio da análise automática de dados. Já Schutt e O'Neil (2014) trazem duas perspectivas sobre sua definição: uma acadêmica e outra profissional. Para a academia, seria a ciência que trabalha com uma grande quantidade de dados e precisa lidar com problemas computacionais relacionados à estruturação, tamanho, desorganização e complexidade dos dados, ao mesmo tempo que resolve problemas reais. Já de um ponto de vista profissional, os mesmos autores declaram que a Ciência dos Dados lida com

a maneira de extrair o significado dos dados, o que exige conhecimento das técnicas da Estatística e do *Machine Learning*. Acrescenta-se também o bom senso e criatividade do ser humano ao processo.

Waller e Fawcett (2013) também declaram a Ciência de Dados como a aplicação de métodos quantitativos e qualitativos para a resolução de problemas e predição de resultados. Ela estuda o ciclo de vida dos dados, isto é, desde a sua criação até o processo de descarte. Basso (2020) declara que sua finalidade está em buscar conhecimento acerca de soluções e problemas e principalmente auxiliar as tomadas de decisão.

A Ciência dos Dados é interdisciplinar/multidisciplinar, pois envolve o conhecimento derivado de uma variedade de áreas, tipicamente a Ciência da Computação, a Administração e Gestão e a Estatística (BASSO, 2020). Em diversos contextos, pode haver a integração com outros ramos. Waller e Fawcett (2013) demonstram, por exemplo, que a Ciência dos Dados no contexto da Gestão da Cadeia de Suprimentos (SCM, do inglês *Supply Chain Management*), envolve disciplinas como a Previsão, Otimização, Simulação a Eventos Discretos, Probabilidade, Finanças, Economia, *Marketing* e Contabilidade.

Agarwal e Dhar (2014) afirmam que a Ciência dos Dados já existe há um considerável tempo, mas só com o advento da disponibilidade de um grande volume de dados, conhecido por *Big Data* e os avanços na Inteligência Artificial (AI), foi que novas questões e novas oportunidades foram surgindo. Os mesmos autores indicam que a captura de *insights* através de dados já é um campo antigo da Estatística e data do século XVIII. As mudanças para a atualidade dizem respeito à alta rapidez com que as transações sociais e econômicas ocorrem e migram para o ambiente virtual. O acesso aos *softwares* também tornou democrático o estudo dessa temática tanto por profissionais quanto por acadêmicos.

No contexto da globalização, Provost e Fawcett (2013) expõem que os dados em si juntamente com capacidade de extrair um conhecimento útil deles são o ativo estratégico crucial para as organizações contemporâneas. Os negócios estão sendo guiados pela análise de dados. A exploração tanto de novos dados quanto dos antigos e arquivados tem se tornado uma forma de adquirir vantagem estratégica dentro do mercado.

Schutt e O'Neil (2014) e Basso (2020) especificam o processo da Ciência dos Dados, representado pela Figura 2.6.

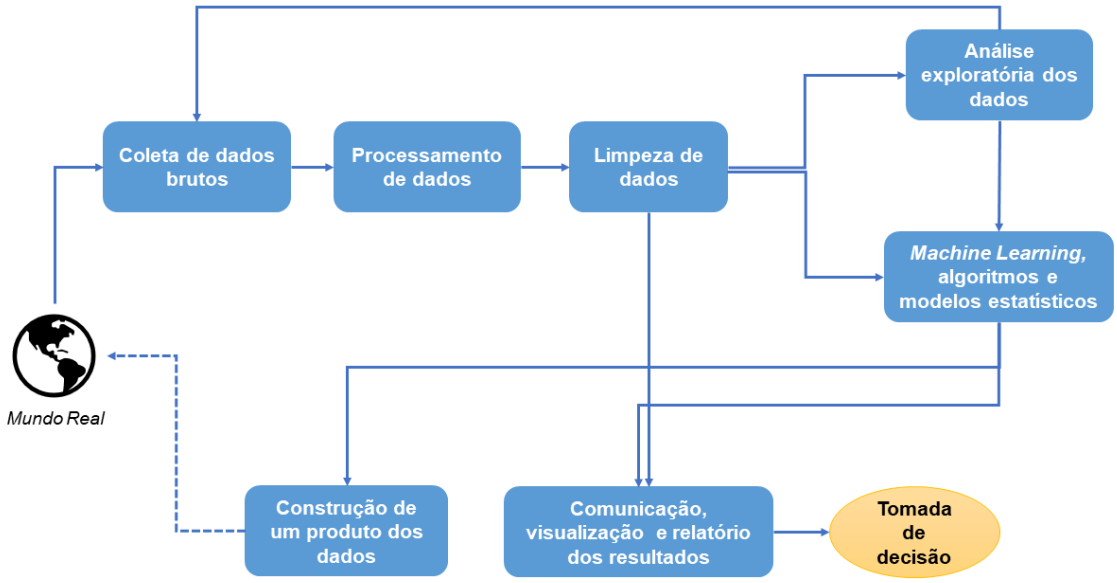


Figura 2.6 - Processo da Ciência de Dados

Fonte: Adaptado de Schutt e O'Neil (2014)

Observando a Figura 2.6, do mundo real se extraem os dados em sua forma bruta. Os dados digitais são gerados por um dispositivo, seja um sensor ou computador. Eles são armazenados em algum tipo de mídia. Pela etapa de processamento, eles são formatados, isto é, ganham o aspecto de tabelas. Na etapa de limpeza, valores duplicados, faltantes ou absurdos são excluídos. A partir daí, é possível aplicar os algoritmos sobre os dados para classificação ou predição. Executadas as análises, a etapa de reportar os achados, geralmente, faz uso de gráficos para que as ideias sejam captadas de forma fácil e rápida. Basso (2020) ainda apresenta a possibilidade desses dados serem deletados ou descartados, ao passo que ficam obsoletos.

Dos benefícios que a Ciência de Dados oferece, Schutt e O'Neil (2014) listam: análise exploratória de dados; *insights* do negócio; tomada de decisões orientada por dados; patentes; previsão de comportamentos ou desempenhos; publicações científicas; proficiência em linguagem de programação; aperfeiçoamento de algoritmos de otimização; correlação de dados; estabelecimento de causalidades; e, interação com especialidades diferentes.

Um princípio importante da Ciência de Dados é a Mineração dos Dados. Há quem trate esses termos como intercambiáveis, entretanto, Provost e Fawcett (2013) definem a Mineração dos Dados como o processo de obter informações relevantes de bases extensas por um processo de estágios bem definidos. Aggarwal (2015) afirma que é o estudo da coleta, da limpeza, do processamento, da análise e da obtenção de *insights* a partir de dados. As tarefas que envolvem a Mineração de Dados estão sumarizadas no Quadro 1.

Quadro 2.1 - Tarefas da mineração de dados

Tarefa da Mineração de Dados	Definição
Classificação	Visa prever, para cada indivíduo em uma população, a qual classe ele pertence. Normalmente, as classes são mutuamente exclusivas.
Regressão	Estima ou prediz o valor numérico de uma variável de um determinado indivíduo.
Correspondência por similaridade	Identifica indivíduos similares baseado nos dados conhecidos sobre eles.
Clusterização	Agrupa os indivíduos de uma população levando em consideração a sua similaridade, mas sem um propósito específico.
Agrupamento por co-ocorrência (associação)	Encontra associações entre as entidades baseado nas transações em que estão envolvidos.
Perfilamento (Descrição de comportamento)	Caracteriza o comportamento típico de um indivíduo, grupo ou população inteira.
Predição de elo	Busca prever as conexões entre os itens das bases de dados e o quão forte estão associados.
Redução de dados	Toma uma base de dados extensa e a substitui por uma menor que contenha as informações importantes da base maior, tornando-a mais fácil em questão de processamento.
Modelagem causal	Entende quais eventos ou ações de fato influenciam os outros.

Fonte: Adaptado de Provost e Fawcett (2013)

Aggarwal (2015) alega que, dentre essa lista, os principais problemas são a classificação, a associação e a clusterização. A Mineração de Dados é difícil de ser realizada sem a automação providenciada pelo *Machine Learning* (ALPAYDIN, 2014). Güemes-Peña *et al.* (2018) declaram que a ideia geral das técnicas do *Machine Learning* é que o próprio computador aprenda a implementar as tarefas, estudando um conjunto de exemplos de treinamento ou dados passados. Faz uso das teorias da estatística para a construção de modelos matemáticos, tendo em conta que o principal objetivo é fazer inferências por meio de uma amostra de dados. Logicamente, para que um computador resolva um problema, é necessário um algoritmo. A saber um algoritmo é o procedimento de passos finitos que busca perfazer uma determinada tarefa. É uma sequência computacional que transforma uma entrada numa saída

requisitada (MARAPPAN; BHASKARAN, 2022). A máquina, através do algoritmo, irá buscar a otimização dos parâmetros dos modelos matemáticos, ao mesmo tempo que armazena e processa a grande quantidade de entradas.

2.5 *Big Data*

Segundo Provost e Fawcett (2013), *Big Data* é definido como um conjunto de dados extensos demais para os sistemas tradicionais; demandando, portanto, novas tecnologias de processamento. Segundo Basso (2020), são ativos de alto volume, velocidade e variedade de informações. Já para Astill *et al*. (2018), o termo designa um grande conjunto de dados envolvendo informações relevantes para o processo de tomada de decisão. Informações essas que mudam rapidamente e necessitam, assim, de análises em tempo real.

O conceito se tornou popular ao passo da evolução da tecnologia nos últimos anos. A população passou também a ser usuária dos recursos tecnológicos antes apenas da indústria, como a Internet *wireless*, *smartphones* e *notebooks*; e passou a realizar as mais diversas transações por meio deles. As empresas, por sua vez, puderam, gradativamente, oferecer serviços e produtos que pudessem atender às necessidades reais de seus clientes, identificando-as em seu comportamento na rede (LIMA; CALAZANS, 2013).

Do ponto de vista de Hofmann e Rutschmann (2018), podem ser empregadas cinco técnicas no *Big Data*, são elas: a exploração de dados, que consiste na compreensão e exploração de uma ampla extensão de dados, com o objetivo da obtenção de *insights* sobre o negócio; a análise avançada, que abrange as questões de estudo mais complexas pela integração do *Machine Learning*, Mineração de Dados e Estatística; a análise e planejamento iterativo, configurando a análise de BI (*Business Intelligent*); a análise incorporada, que possibilita operações com base de dados por meio da automatização de processos; e, por fim, a análise de fluxo de dados, permitindo a programação automática de dados, para filtragem, transformação, detecção de correlação de dados, tendências e erros, além da previsão através de modelos.

Cavique (2014) e Basso (2020) descrevem as características do *Big Data* em grande volume, grande velocidade e grande variedade, popularmente conhecido com os 3 V ’s, como mostra a Figura 2.7. Existem outros que incluem a esses pilares os componentes de variabilidade, valor, e ainda veracidade, como em Comuzzi e Patel (2016) e Gandomi e Haider (2015).

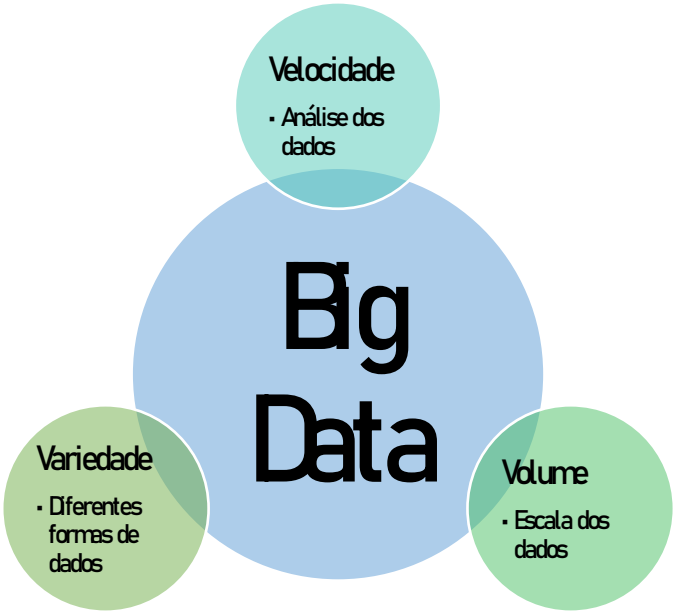


Figura 2.7 - Características do Big Data

Fonte: Adaptado de Cavique (2014)

Conforme Gandomi e Haider (2015), a característica de volume está associada à quantidade de dados gerados, sendo a principal atribuição do *Big Data.* A criação de dados alcança a escala de *terabytes*, se não *pentabytes*, por meio de várias fontes e dispositivos. Já a velocidade diz respeito ao tempo necessário para captura e análise dos dados. Os dados são produzidos a uma alta velocidade; como consequência, eles têm de ser sondados com a mesma rapidez.

A variedade concerne à heterogeneidade da estrutura dos dados acessíveis. Oliveira (2020) afirma que os dados têm origens diversas e podem conter o estado de uma vasta pluralidade de variáveis. Eles podem ser estruturados, isto é, cada informação está organizada em uma linha e as colunas recebem a identificação das variáveis; ou não estruturados, como em um texto ou uma imagem. Lima e Calazans (2013) declaram que o principal desafio está em dar sentido aos dados não estruturados, a partir de cruzamentos e interpretações, sendo um esforço útil em agregar valor estratégico para as diversas camadas do mercado.

A variabilidade, característica não inclusa na visão tradicional do *Big Data*, se refere à inconsistência dos dados, o que traz dificuldades para o seu gerenciamento (OLIVEIRA, 2020). Já a veracidade diz respeito à qualidade dos dados. Conforme Qader *et al.* (2020), se refere a certeza, ruído e abnormalidade dos dados. Ishwarappa e Anuradha (2015) afirmam que, quando se lida com alto volume, velocidade e variedade de dados; não é possível que todos os dados sejam 100% corretos; muito possivelmente, haverá dados “sujos”, isto é, dados repetidos, não

verificados ou que não servem para nenhum propósito. A qualidade do conjunto de dados a ser armazenado pode variar grandemente; e disso depende a acuracidade da análise a ser realizada.

No tocante a valor, Ishwarappa e Anuradha (2015) certificam ser o aspecto mais importante em *Big Data.* Dizem que, embora seja bom ter acesso a grandes volumes de dados, de nada adianta o processo se não houver a valoração das informações, levando em consideração que implementar sistemas de infraestrutura em tecnologia da informação para o armazenamento de dados é oneroso e, certamente, o negócio exigirá o retorno do investimento.

Apesar dos desafios, o *Big Data* desempenha um importante papel organizacional e influencia diretamente a estratégia de *marketing* de grandes negócios. Todas as tecnologias e ferramentas associadas ao *Big Data* podem ser empregadas para alavancar a eficiência de produção de uma empresa e ser útil na concepção de serviços e produtos orientados por dados (QADER *et al*., 2020). Como resultado, as aplicações do *Big Data* estão abrindo novos caminhos em vários campos de atuação como nas Telecomunicações, Serviços Financeiros, Mídias Sociais, Ciências da Saúde, Varejo e, como citam Ghalehkhondabi *et al.* (2020), têm contribuído com ainda mais relevância para a Manufatura, nos temas de Manufatura Ágil, Desenvolvimento Estratégico, Sustentabilidade; e também para a Gestão da Cadeia de Suprimentos e Logística, nos temas de Melhoria das Operações, Sustentabilidade, Cadeia de Suprimentos Alimentícios, Gestão de Riscos e Vendas.

2.6 Clusterização - *Clustering*

A clusterização, do inglês *clustering*, é também conhecida como “reconhecimento de padrões” ou “agrupamento”. De acordo com Saxena *et al.* (2017), é definida como um método pelo qual os objetos são agrupados tendo por base sua similaridade. Armano e Farmani (2016) a interpretam como o processo de particionar os dados em grupos com as propriedades almejadas e os dados em cada agrupamento são similares, ao passo que, em comparação aos outros, são diferentes. O termo amplamente empregado para os grupos é *cluster*.

Aggarwal (2015, p. 42) dá uma significação informal à clusterização por: “dada uma base de dados *D*, divida suas linhas (registros) em conjuntos C_1 ... C_k, de tal forma que as linhas em cada agrupamento sejam similares umas às outras”. O autor afirma que esta definição é informal, pois a clusterização pode assumir diversos outros significados, a depender do contexto. Saxena *et al.* (2017) trazem a estrutura formal e convencional da técnica: um conjunto *S* de subconjuntos S_1, S_2, ..., S_k de tal forma que:

$$S_1 \cap S_2 \cap S_3 \cap \ldots \cap S_k = \emptyset \quad (1)$$

Isto é, o elemento pertencente ao grupo S_1 obviamente não fará parte do grupo S_2 e assim por diante. Esses mesmos autores declaram que a clusterização é um problema mais complexo que a classificação, pois não há nenhum rótulo anexados aos dados previamente.

Outra questão importante na clusterização diz respeito à alta dimensionalidade da base de dados, visto que, quanto mais variáveis forem analisadas, maior o custo computacional em processá-las, afetando também negativamente a consistência do algoritmo. Por isso, devem ser avaliadas quais variáveis são mais relevantes de serem estudadas.

Santhanam e Velmurugan (2010) dizem que a principal vantagem da clusterização é que os padrões e a estrutura dos dados podem ser encontrados diretamente com pouco ou nenhum conhecimento prévio. As mais relevantes aplicações da metodologia são demonstradas a seguir:

- **Segmentação de clientes**: determina similaridades entre clientes para as estratégias de promoção de produtos específicos;
- **Resumo dos dados**: a base de dados pode ser explicada por meio dos *clusters* formados;
- **Aplicação em outros problemas de mineração de dados**: pode auxiliar no problema de análise de *outliers*, a título de exemplo;
- **Análise de redes sociais**: utilizada para a identificação de relacionamentos, grupos de amizade e comunidades. Muito útil no estudo dos comportamentos humanos (AGGARWAL, 2015; PROVOST; FAWCETT, 2013).

A clusterização é um processo de aprendizagem de máquina não supervisionada, pois o próprio computador aprende com a classificação de cada objeto. A cada iteração, esse conhecimento vai se tornando mais apurado (HALKIDI *et al.*, 2002).

Saxena *et al*. (2017) especificam que existem duas abordagens possíveis para a clusterização: a hierárquica e a particional. A hierárquica pode ser conduzida por uma metodologia de aglomeração ou divisão; que, por sua vez, podem adotar a correspondência simples, completa ou média. Já a particional pode ser baseada na distância, num modelo ou na densidade. Os algoritmos baseados na distância empregam o erro quadrado; os outros casos utilizam a probabilidade. A Figura 2.8 esquematiza as diferentes abordagens na clusterização.

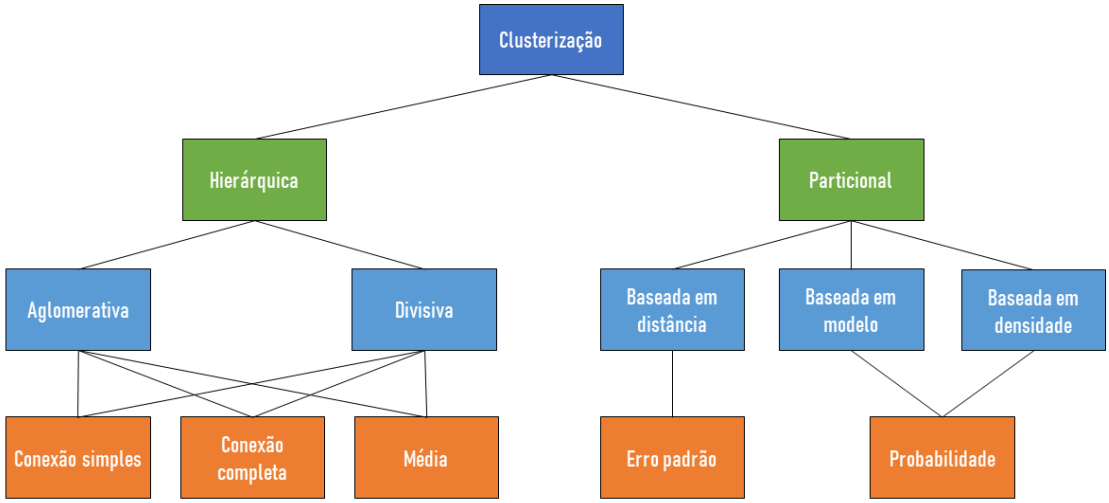


Figura 2.8 - Caracterização da clusterização
Fonte: Adaptado de Saxena *et al.* (2017)

Na abordagem hierárquica, os *clusters* são formados por um processo iterativo dividindo os padrões "de cima para baixo" ou "de baixo para cima". A metodologia não precisa de determinação prévia de uma quantidade de *clusters* a ser produzida. A técnica "de cima para baixo" é chamada de aglomerativa, pela qual cada dado é tido como um *cluster* por si só. A partir daí, os grupos mais similares são fundidos progressivamente, até que se resulte em um grande e único *cluster*. Já a estratégia de divisão é oposta, parte-se "de baixo para cima". O conjunto completo de observações é tomado dentro de um mesmo *cluster*; e então os grupos mais heterogêneos são separados, até que cada observação pertença a seu próprio *cluster* (KASSAMBARA, 2015; SAXENA *et al*., 2017).

A clusterização hierárquica é ainda caracterizada pela suas ligações. O método de clusterização hierárquica resulta em gráficos chamados de dendrogramas, que demonstram a relação entre elementos em um formato de árvore.

Já os algoritmos particionais, ou de particionamento, requerem do analista uma especificação preliminar da quantidade de *clusters* a ser construída. Segundo Chen e Hu (2006), os métodos de particionamento dividem *n* objetos em *k clusters*. A distância euclidiana é tida como a mais popular para medir a similaridade entre elementos: quanto maior, mais dissimilar serão as observações. Nas técnicas particionais, dois pressupostos são tomados: de que o elemento só pode pertencer a um único *cluster* e cada objeto deve ser agrupado ao *cluster* mais próximo. Os algoritmos mais populares dessa categoria são o *K-Means*, o *K-Medoids*, conhecido por *Partitioning Around Medoids* (PAM) e o *Clustering for Large Applications* (CLARA).

Dentro da categoria dos particionais, existem os algoritmos que empregam modelos probabilísticos, mencionados como *soft clustering* ou clusterização baseada em modelos, em que um elemento pode ter uma probabilidade de pertencer a dois grupos simultaneamente (AGGARWAL, 2015). Os dados são ponderados como provenientes de uma distribuição, que é a combinação de dois ou mais *clusters*. Tem-se também a clusterização baseada em densidade, denominada por DBSCAN (do inglês, *Density-Based Spatial Clustering and Application with Noise*), empregada para a identificação de grupos de qualquer forma em um conjunto de dados com ruídos e de pontos fora da curva (KASSAMBARA, 2015).

Existe a categoria de algoritmos descrita como avançada. Fazem parte dela a clusterização baseada em modelo, a baseada em densidade, os algoritmos de *Fuzzy Clustering* e o *K-Means* hierárquico.

A clusterização pode ser aplicada em vários domínios de conhecimento. Na literatura científica, podem ser encontrados trabalhos como o de Perotin *et al.* (2022), que implementou uma análise de *cluster* para identificar subgrupos de pacientes, baseado em seus dados de capacidade pulmonar, para avaliar as sequelas do COVID-19. Outros como o de Khajeh *et al.* (2022), que propôs a clusterização para organizar trajetórias similares de veículos que usam o controle adaptativo de cruzeiro (ACC). Já nas áreas sociais, Belcastro *et al*. (2022) realizaram um estudo sobre o comportamento humano nas redes sociais com relação às eleições presidenciais dos Estados Unidos no ano de 2020. Eles aplicaram as técnicas de clusterização para extrair os principais tópicos de discussão das redes sociais e avaliaram o seu impacto sobre as conversas nas plataformas digitais.

2.6.1 Algoritmos clássicos

São apontados como algoritmos clássicos de clusterização o *K-Means, K-Medoids,* também chamado de PAM, o CLARA, o AGNES e o DIANA. Um dos mais simples algoritmos de clusterização, o *K-Means* foi primeiramente publicado em 1955, e desde lá tem sido amplamente empregado (JAIN, 2010). O algoritmo encontra uma partição tal que o erro quadrado entre a média empírica do *cluster* e dos pontos no *cluster* seja mínimo. Os centros dos agrupamentos são tidos como referenciais (DENG *et al.*, 2016; JAIN, 2010).

O algoritmo pode ser definido pelos seguintes passos:

1) Determine a quantidade de *clusters*;
2) Escolha arbitrariamente *k* elementos do conjunto de dados para serem os centros ou a média dos *clusters*;

3) Designe cada elemento ao seu centroide mais próximo, considerando a distância euclidiana;
4) Corrija o centro do *cluster* de cada um dos *k* agrupamentos, determinando o novo valor de média de todos os pontos de dados do agrupamento. Vale ressaltar que o centroide do *k*-ésimo *cluster* define-se por um vetor de tamanho *p* (isto é, quantidade de variáveis), compreendendo as médias das variáveis para as observações do grupo;
5) Minimize a soma dos quadrados total pela repetição dos passos 3 e 4, até que as características do *cluster* cessem as mudanças ou o número máximo de iterações seja atingida (KASSAMBARA, 2015).

Como se percebe, o algoritmo do *K-Means* exige a indicação prévia de três parâmetros: a quantidade de *clusters k*, a inicialização do *cluster* e a métrica da distância. Jain (2010) declara ser a escolha do *k* a mais crítica.

Outra questão a ser considerada é a convergência da técnica. Nunes (2016) afirma que o *K-Means* converge para uma solução ótima parcial de um caso em uma quantidade finita de iterações. Capó *et al.* (2017), entretanto, aponta a eficiência da técnica, até mesmo para extensas bases de dados, a despeito da necessidade de inicialização e de tempo de convergência.

Já o algoritmo de particionamento ao redor de medoides, em inglês, *Partitioning Around Medoids* (PAM), é também chamado de *K-medoids.* Foi pela primeira vez introduzido por Kaufman e Rousseeuw (1990). Consiste em dois algoritmos: o *BUILD*, que objetiva uma clusterização inicial; o *SWAP*, que realiza os melhoramentos buscando um ótimo local. O algoritmo exige a construção de uma matriz de dissimilaridade, empregando a rotina *DAISY*, também proposta por Kaufman e Rousseeuw (1990) (SCHUBERT; ROUSSEEUW, 2019).

O algoritmo é análogo ao *K-Means,* que se propõe a dividir um conjunto de pontos em vários grupos que minimizam o somatório das distâncias entre um elemento e um centro. A principal diferença entre os métodos é que o PAM retorna os verdadeiros pontos dos dados como representantes do *cluster* (RUIZ *et al.*, 2020).

Na primeira parte, *BUILD* escolhe por *k* vezes o ponto que detém o menor somatório de distância de todos os demais pontos. A motivação para esse passo é encontrar um bom ponto de partida para que a etapa de refinamento demande menos iterações (SCHUBERT; ROUSSEEUW, 2019).

A segunda parte, *SWAP*, é baseada em trocar um único medoide por um elemento de dado único e aperfeiçoar a escolha de *k* medoides, tendo em mente todos os *k(n-k)* vizinhos do nó que conduz à otimização na função custo (NGUYEN; RAYWARD-SMITH, 2011). A saber, medoides são objetos mais centrais de um determinado *cluster*.

O procedimento para a execução do algoritmo é descrito por:

1) Tomar *k* objetos para serem os medoides;
2) Se não disponível previamente, computar a matriz de dissimilaridade;
3) Dispor cada objeto ao seu medoide mais próximo;
4) Avaliar se algum objeto diminui o coeficiente médio de dissimilaridade, observando cada *cluster*. Se houver mudanças, tomar o elemento que mais reduziu o parâmetro como novo medoide para o agrupamento;
5) Observar se houve substituição de, pelo menos, um medoide. Caso sim, volte à etapa 3. Caso contrário, conclua o algoritmo (KASSAMBARA, 2015).

A matriz de dissimilaridade é uma matriz *n* x *n* das distâncias entre os *n* objetos. Zerzucha e Walczak (2016) declaram que, para seu cálculo, é possível o emprego das distâncias euclidianas, definidas pela raiz quadrada da soma de quadrados das diferenças entre elas ou a distância *Manhattan*, o somatório das distâncias absolutas. O algoritmo possui limitação quanto a extensão da base de dados, consumindo bastante tempo e também sendo muito sensível a *outliers.*

O *Clustering for Large Applications* (CLARA), Clusterização para Grandes Aplicações, é uma combinação da abordagem de amostragem com o algoritmo PAM. O algoritmo foi idealizado por Kaufman e Rousseeuw (1990). Em vez de encontrar os medoides para toda uma base de dados, o CLARA toma uma amostra do conjunto inteiro e emprega o PAM para a seleção de um conjunto ótimo de medoides. O algoritmo é implementado repetidas vezes, juntamente com o processo de amostragem, para mitigar o viés de amostragem (WEI *et al*., 2003).

Aboubi *et al.* (2016) e Nguyen e Rayward-Smith (2011) afirmam que o CLARA é um melhoramento do PAM para lidar com base de dados muito grandes, visto que o PAM é computacionalmente caro, sendo este adequado para conjunto de dados pequenos (menores que mil observações). Os mesmos autores revelam que o algoritmo produz resultados satisfatórios em um tempo de computação razoável, mas é preciso ponderar a sua menor efetividade ao considerar amostras e não a base de dados completamente. Os seus idealizadores, Kaufman e Rousseeuw (1990) sugerem a tomada de cinco amostras independentes, cada uma de tamanho $(40 + 2k)$, onde *k* é o número de *clusters* pretendido.

O algoritmo CLARA pode ser detalhado nos seguintes passos:

1) Separar, de maneira aleatória, a base de dados em diversos subconjuntos com tamanho fixo;
2) Executar o algoritmo PAM em cada subconjunto;

3) Computar a média ou o somatório das dissimilaridades dos elementos para o medoide mais próximo;
4) Manter o subconjunto de dados no qual o parâmetro do passo anterior é mínimo. Os medoides representativos são aqueles resultantes da amostragem com menor custo de clusterização. Assim sendo, uma análise mais abrangente pode ser conduzida na partição definitiva (KASSAMBARA, 2015; NGUYEN; RAYWARD-SMITH, 2011).

Segundo Yan *et al.* (2022), da categoria de algoritmos hierárquicos, tem-se o *Agglomerative Nesting* (AGNES) e o *Divisive Analysis* (DIANA). A estratégia de processamento do AGNES segue-se "de baixo para cima". De acordo com os mesmos autores, cada observação é constituinte de um *cluster* propriamente. A partir daí, os que estão localizados a uma distância menor são fundidos até que a condição de parada seja satisfeita. O algoritmo é apontado como de baixa complexidade de implementação, sendo, por essa razão, muito empregado em análise de *cluster.*

De acordo com Wijuniamurti *et al*. (2022), a técnica pode ser detalhada da seguinte forma:

1) Inicie com *n clusters.* Cada *cluster* contém um objeto como membro;
2) Suponha uma matriz de proximidade e encontre uma matriz de inequalidade para o par mais similar;
3) Conecte os objetos utilizando uma função de ligação em um novo *cluster* e diminua a quantidade de *clusters* em um, removendo a linha e a coluna do objeto na matriz de proximidade. Calcule a inequalidade entre o *cluster* e os demais, adicionando linha e coluna à nova matriz de inequalidade;
4) Repita os passos 2 e 3 até (*n-1*) vezes para formar todos os objetos em um único agrupamento.

As funções de ligação mais comuns são a simples, completa e a média. Na simples, conhecido como método mínimo, individual ou do vizinho mais próximo, a distância entre dois *clusters* é determinada pela menor distância de qualquer membro de um para qualquer membro do outro. Na completa, tida como o método máximo ou do vizinho mais distante, a distância entre dois agrupamentos é concebida pela maior distância de qualquer membro de um *cluster* para qualquer outro membro de outro. Já na média, nomeada como método da mínima variância, é calculada a distância entre dois agrupamentos pela distância média entre qualquer membro de um a qualquer membro de outro (KASSAMBARA, 2015; SAXENA *et al*., 2017).

A distância interna dos *clusters* influencia consideravelmente o algoritmo AGNES, mas este se apresenta autêntico por ser independente da seleção preliminar de valores e desimpedido da curva das distribuição das amostragens (WANG *et al*., 2020).

O *Divisive Analysis* (DIANA), ou análise divisiva, considera previamente todos os dados como um único *cluster*, e, a partir deste, vai gerando partições cada vez mais menores. Consiste num algoritmo de lógica inversa ao do AGNES. A cada iteração, o conjunto maior de dados é dividido em dois *clusters.* O algoritmo cessa ao passo que cada observação seja um *cluster*. A hierarquia é conduzida, portanto, em (*n-1*) passos, sendo *n* a quantidade de observações da base de dados. Se no AGNES as possibilidades de combinação de cluster é dado por *n*(*n-1*)/2 combinações, no DIANA esse número é ainda maior: são 2^{n-1}-1 combinações (PATNAIK *et al*., 2016). Mesmo para conjuntos de dados de médio porte, esse número se torna computacionalmente inviável. Entretanto, é possível o método divisivo que não considere todas as divisões (KAUFMAN; ROUSSEEUW, 1990).

A abordagem do algoritmo DIANA é "de cima para baixo" e, segundo Gostkowski *et al.* (2021), pode ser mais eficiente do que o AGNES em certas situações. A medida de distância empregada com frequência no DIANA é a distância euclidiana. Dzikrullah e Ahmad (2021) afirmam que o método foca no cálculo da distância média de cada objeto fora do *cluster* em relação aos objetos dentro do grupo e também calcula a distância média entre os objetos dentro do próprio *cluster.* Esta medida é chamada de *Splinter Average Distance Method,* Método da Distância Média de Fragmentação, sendo definida por:

$$\bar{x}_k = \frac{1}{n-1}\sum_{k=1}^{n} x_k \qquad (2)$$

Onde $\bar{x}_k$ é a média dos *k*-ésimos objetos, enquanto x_k são os valores de *k* (*k = 1, 2, 3, ..., n*), *n* é a quantidade de objetos (DZIKRULLAH; AHMAD, 2021).

Detalhadamente, o algoritmo DIANA pode ser descrito pelas seguintes etapas:

1) Encontrar os objetos que têm a maior dissimilaridade de todos os outros objetos. Este será o grupo *splinter;*
2) Para cada objeto *i*, calcule a diferença da distância média entre os elementos não pertencentes ao grupo *splinter* e a distância média dos elementos pertencentes a esse grupo;
3) Ache um objeto *h* para o qual essa diferença de distância é a maior. Se for positiva, então o objeto *h* está próximo ao grupo *splinter*;

4) Repita o passo 2 até que as distâncias sejam negativas;
5) O *cluster* que apresenta o maior diâmetro será selecionado como a maior dissimilaridade entre qualquer dois objetos. Após, este *cluster* será dividido;
6) Todos os passos são executados outra vez, até que cada *cluster* seja composto por apenas um objeto (SHARF; RAZZAK, 2017).

2.6.2 Algoritmos avançados

Na categoria de algoritmos avançados, tem-se como mais populares a clusterização baseada em modelo, os algoritmos de *Fuzzy Clustering* com o *Fuzzy C-Means* e o *K-Means* hierárquico.

No *Fuzzy Clustering*, é possível que um elemento seja atribuído a vários *clusters* ao mesmo tempo. É um método com base nas partições *fuzzy* e cada elemento de dado é designado a um *cluster* com um grau de pertencimento, que depende do quão perto este objeto está do centro do agrupamento: quanto mais próximo, mais chances de ele estar vinculado ao grupo e vice-versa. O algoritmo mais popular é o *Fuzzy C-Means* (FCM), introduzido em 1974 (GHOSH; KUMAR, 2013).

O FCM difere do *K-Means* e do PAM, pois estas técnicas são determinísticas. No FCM, o grau de pertencimento de um objeto é dado por um valor de intervalo [0,1] e o centroide é computado como a média de todos os pontos (KASSAMBARA, 2015).

O FCM move, de maneira iterativa, os centros dos *clusters* para o local correto num base de dados (GHOSH; KUMAR, 2013). Suganya e Shanthi (2012) afirmam que o algoritmo apresenta vantagens de ser uma técnica não supervisionada e convergente, mas lida com os desafios de tempo computacional extenso e de sensibilidade à inicialização, a ruídos e *outliers*.

Já o *K-Means* hierárquico, em inglês *Hierarchical K-Means* (HKM), visa cobrir um problema do *K-Means* tradicional quanto à sua inicialização. Segundo Qi *et al*. (2017), é difícil não selecionar ruídos ou pontos muito aglomerados como sementes, visto que o *K-Means* as escolhe de forma arbitrária.

Nguyen *et al.* (2019) descrevem o algoritmo pelos seguintes passos:

1) Configure $C = \{c_i \mid i = 1, ..., n\}$como um atributo do vetor *n*-dimensional. Então, configure $Y = \{y_i \mid i = 1, \ldots, r\}$ como cada observação de C;
2) Determine a quantidade de *clusters* iniciais;
3) Estabeleça o número de iterações *b*;
4) O estado inicial do contador é definido em $i = 1$;

5) Aplique o algoritmo *K-Means;*
6) Registre os centroides como $A_i = \{a_{ij} \mid j = 1, ..., K\}$;
7) Incremente o contador em 1 unidade;
8) Execute novamente o passo 5 enquanto $i < b$;
9) Tome $A_i = \{a_i \mid i = 1, ..., b\}$ como um novo conjunto de dados, com *K* como a quantidade de *clusters* previamente definida;
10) Aplique um algoritmo hierárquico;
11) Registre os centroides resultantes como $D = \{d_i \mid i = 1, ..., K\}$;
12) D pode ser tomado como os centros de *clusters* iniciais para a execução do *K-Means*

O HKM melhora significantemente o processo de clusterização combinando as vantagens da clusterização hierárquica e do *K-Means*. É apontada também a eficiência do HKM nas análises de base de dados extensas, porém como desafios a melhoria da sua velocidade e precisão (LIAO *et al.*, 2013).

Já a clusterização baseada em modelo (do inglês, *Model-Based Clustering*), de acordo com Fop e Murphy (2018), os *clusters* são formados em uma estrutura de modelagem e o processo de geração de dados é representado por uma mistura finita de distribuições de probabilidade. Já Mcnicholas (2016) descreve esta técnica como o uso de finitos modelos combinados para a formação dos agrupamentos. Segundo McParland e Gormley (2016), a combinação de duas os mais funções densidade de probabilidade é usada quando não existem meios de identificar a qual população cada observação é pertencente.

A técnica de clusterização baseada em modelos exige uma etapa de seleção de variáveis, pela qual apenas uma parte delas acrescentará significado para o agrupamento; o que facilita as interpretações, a produção de estimativas mais condizentes com a realidade, partições efetivas e a seleção de um modelo mais simples (BOUVEYRON; BRUNET-SAUMARD, 2014; SCRUCCA; RAFTERY, 2018).

Os parâmetros do modelo (a saber, vetor das médias, probabilidade associada na combinação e matriz de covariância) são provenientes de uma distribuição normal e são estimados geralmente pelo algoritmo *Expectation-Maximization* (EM). Nele, cada grupo é localizado no centro do vetor das médias com densidade aumentada para os pontos de maior proximidade, sendo a geometria do *cluster*, isto é, seu volume, orientação e forma, definida pela matriz de covariância (KASSAMBARA, 2015).

2.6.3 Abordagem por amostragem

Segundo Hicks *et al.* (2021), apesar de ser simples de implementar, os algoritmos tradicionais de clusterização, como o *K-Means* e o PAM, requerem recursos computacionais mais complexos, como uma memória de acesso aleatório de maior capacidade para armazenar os dados e os cálculos intermediários. Isso ocorre porque é necessário carregar todo o conjunto na memória de uma vez para realizar a clusterização. Portanto, eles podem apresentar lentidão ou até mesmo falhar completamente.

Xu *et al.* (2022) pontuam que os algoritmos originais consomem muita memória ao realizar a clusterização em conjunto de dados em grande escala, devido à complexidade de tempo e espaço de ordem quadrática em relação ao número de dados. Descrevem ainda que eles podem produzir um número excessivo de pequenos *clusters*. E, por fim, podem ter dificuldade de convergir, o que resulta em maior tempo gasto no ajuste fino dos parâmetros.

Para lidar com esses problemas, ao realizar uma análise de *clusters* em conjuntos de dados extensos, existem duas soluções: a paralelização e a amostragem. Na paralelização, o trabalho é dividido em partes menores que podem ser executadas simultaneamente em múltiplos processadores ou núcleos de processamento. Já a abordagem por amostragem, conforme Hicks *et al.* (2021), trabalha com pequenas amostras aleatórias dos dados, chamadas de “mini-lotes”, em inglês *mini-batches*, que podem ser armazenados na memória de computadores convencionais. Os algoritmos que se popularizaram sob esse enfoque foram o *Mini-Batch K-Means* e o *Mini-Batch* DBSCAN.

Tang e Fong (2018) complementam que a amostragem é uma estratégia eficaz para lidar com grandes volumes de dados, visando encontrar um perfil reduzido, porém representativo desses dados. Eles indicam que cada bloco de dados precisa ser pequeno o suficiente para atender às restrições de tamanho dos algoritmos clássicos, mas, ao mesmo tempo, seus resultados devem ser confiáveis e aproximados o bastante para representar o sistema com um todo. Chavan *et al.* (2015) confirmam que, por essa abordagem, os algoritmos podem exibir uma ordem de convergência mais rápida que a dos métodos originais estocásticos.

É importante destacar que, embora os algoritmos desse tipo operem apenas em pequenas amostras dos dados por vez, eles ainda são capazes de minimizar a mesma função objetivo global avaliada em todas as amostras, assim como nas implementações tradicionais (YADAV; BARIA, 2014).

O agrupamento por lotes, como também pode ser chamada a clusterização por amostragem, é também adequado para o cenário de aprendizagem de máquina onde novos dados são constantemente obtidos. Ao utilizar pequenos conjuntos de dados, é possível realizar

a clusterização em tempo real à medida que novas informações são recebidas, tornando o processo de atualização do modelo mais dinâmico e adaptável (KANUNGO *et al.*, 2002).

Entretanto, vale ressaltar alguns pontos que são sacrificados em questão da escalabilidade e da eficiência computacional. A precisão pode ser reduzida se comparado ao modelo clássico. A escolha do tamanho do lote também pode afetar os resultados. Se for muito pequeno, a variabilidade entre os lotes pode ser alta, afetando a consistência das saídas; e se muito grande, pode-se perder a eficiência computacional. Os resultados do estudo de Hicks *et al.* (2021) comprovam que, desde que o tamanho da amostragem não seja excessivamente pequeno (ao menos, 500 a 1.000 dados por lote), os algoritmos sob esse enfoque são tão precisos quanto às versões tradicionais. Além do mais, Kanungo *et al.* (2002) apontam que a troca entre qualidade da clusterização e o tempo de execução do algoritmo pode ser suavemente ajustada ao considerar mais iterações.

O mais popular algoritmo, o *K-Means* que utiliza essa abordagem por amostragem, pode ser descrito da seguinte forma: como entrada, tem-se o conjunto de dados *X*, número de clusters *k*, tamanho do *mini-batch B* e o número máximo de iterações *T*. Como saídas do processo: os centroides dos *clusters* e atribuição de cada ponto de dados a um *cluster*.

1. Inicialize os centroides aleatoriamente;
2. Para $t = 1$ até T:
    a. Amostrar um *mini-batch* de tamanho *B* a partir do conjunto de dados *X*;
    b. Para cada ponto de dados *x* no *mini-batch*, calcule a distância entre *x* e todos os centroides;
    c. Atribua *x* ao cluster cujo centroide está mais próximo;
3. Atualize os centroides dos clusters:
    a. Para cada *cluster k*:
        i. Calcule o novo centroide como a média dos pontos de dados atribuídos ao *cluster;*
4. Verifique o critério de convergência (por exemplo, se o número máximo de iterações foi atingido):
    a. Se sim, saia do *loop;*
5. Retorne os centroides dos clusters e a atribuição de cada ponto de dados a um *cluster* (BÉJAR, 2013).

Apesar das grandes vantagens, a literatura científica ainda carece de trabalhos que abrangem a temática. Os poucos registros podem ser descritos pelos seguintes trabalhos: Bahmani *et al.* (2012) propõem um método escalável para inicialização dos centroides no

algoritmo *K-Means* chamado *K-Means++*. Em vez de selecionar os centroides iniciais aleatoriamente, *o K-Means++* utiliza um processo de seleção ponderado para escolher pontos de dados iniciais que sejam mais distantes uns dos outros, melhorando a qualidade do agrupamento e a convergência do algoritmo. Já Zhang *et al*. (2017) apoiam um algoritmo de clusterização em *mini-batch* que utiliza técnicas de mesclagem e amostragem para melhorar a eficiência e escalabilidade. O algoritmo realiza mesclagem de *clusters* próximos e amostragem de pontos de dados para reduzir a complexidade computacional e garantir uma convergência precisa. O método é adequado para fluxos contínuos de dados e demonstra um bom desempenho em grandes conjuntos de dados. Peng *et al.* (2018) conseguem combinar o *Mini-Batch K-Means* com a Análise de Componente Principal (PCA). Eles utilizaram a inicialização dos centros dos *clusters* para evitar a convergência em ótimos locais. Afirmam que, quando comparado com outros métodos, os resultados experimentais e a análise de complexidade temporal mostram que o método proposto é eficaz e eficiente.

2.6.4 Validação

Moulavi *et al*. (2014) afirmam que as decisões de clusterização não estão somente relacionadas à escolha de qual algoritmo executar para uma determinada tarefa, mas também de ajustar adequadamente seus parâmetros. Kassambara (2017) alerta que, antes de utilizar qualquer método de clusterização nos dados, é crucial avaliar se os conjuntos de dados contêm agrupamentos significativos, isto é, estruturas não-aleatórias. Caso existam, é importante determinar quantos agrupamentos estão presentes. Esse processo é conhecido como avaliação da tendência de clusterização ou a viabilidade da análise de clusterização. No entanto, um desafio importante na análise de *clusters* é que os métodos de clusterização podem gerar agrupamentos, mesmo quando os dados não possuem agrupamentos reais, ou seja, estão distribuídos uniformemente. Aplicando-se cegamente um método de clusterização em um conjunto de dados, ele tentará dividir os dados em grupos, pois essa é a sua finalidade.

Há dois métodos principais de avaliação da tendência de clusterização: estatístico e visual. A principal técnica no método visual são os *heatmaps*, mapas de calor, que consistem numa maneira gráfica de compreender a dinâmica de dados de alta dimensionalidade. É elaborado como uma matriz de células empregando-se um gradiente de cores. É comum o uso da técnica em conjunto com dendrogramas. Os *heatmaps* auxiliam na avaliação geral dos maiores e menores valores de um conjunto de dados. As linhas e colunas são frequentemente clusterizadas de forma a possibilitar a interpretação (KASSAMBARA, 2015; PAULSON *et al*., 2012).

Já Aggarwal (2015) afirma que a Estatística de Hopkins é frequentemente usada na análise estatística. Ela leva em consideração uma amostra de dados sintéticos gerados por um processo aleatório dentro do domínio do espaço dos dados. Segundo Semaan *et al.* (2019), a técnica é também chamada de Teste de Aleatoriedade Espacial e consiste num teste de hipótese em que a hipótese nula não é rejeitada quando o valor obtido para a estatística for igual ou menor que 0,5. Ou seja, se os pontos estiverem uniformemente distribuídos, não há proposta de agrupamento adequado para eles. Em contrapartida, para valores próximos de 1, a Estatística demonstra que os dados são fortemente "clusterizáveis", apresentam uma tendência de agrupamento.

Agora, o principal problema a ser lidado após a execução da clusterização é a validação dos seus resultados, que consiste em, por meio de índices, avaliar a eficiência do método quanto ao alcance de seus objetivos. De forma geral, a clusterização busca uma alta similaridade de dados, considerando o interior do *cluster*; e uma alta dissimilaridade entre os agrupamentos (HÄMÄLÄINEN *et al*., 2017).

As saídas de uma clusterização são as partições de um conjunto de dados, mas também podem ser observadas a estrutura espacial dessa saída, a título de exemplo, a sua compacidade (BRUN *et al*., 2007).

A complexidade do processo está em ajustar uma função de índice de validação, executar os algoritmos e concluir a investigação da quantidade otimizada de *clusters* ao passo que há a averiguação dos resultados (HALKIDI *et al.*, 2002).

Os índices de validação podem ser classificados de três formas:

- **Validação interna:** é baseada no cálculo das propriedades dos agrupamentos resultantes, como compacidade, separação e conectividade. Não precisa de informações extras da base de dados;
- **Validação relativa:** consiste em comparar as partições resultantes pelo mesmo algoritmo, mas com diferentes parâmetros ou outros subconjuntos de dados. Também não demanda informações adicionais dos dados;
- **Validação externa:** fundamenta-se da mesma forma na comparação de partições. Os *clusters* a serem comparados consistem no que é gerado pelo algoritmo e de uma partição já pré-estabelecida dos dados. Está associada à medição de erro (BRUN *et al*., 2007).

Os índices de validação interna são empregados para a escolha do algoritmo mais preciso na clusterização e também para a definição da quantidade ótima de grupos. A compacidade, considerada nessa tipologia de validação, é a medida que explica o quão bem

relacionados estão os elementos de um *cluster* (distância intra-*cluster*), pelo emprego da variância (uma variância de baixo valor aponta uma compacidade boa). Já a separação é uma medida que avalia o quão distinto um *cluster* é dos outros (distância inter-*cluster*); a distância mínima de pares é amplamente usada (LIU *et al.*, 2010). A Figura 2.9 exemplifica o que são essas medidas graficamente.

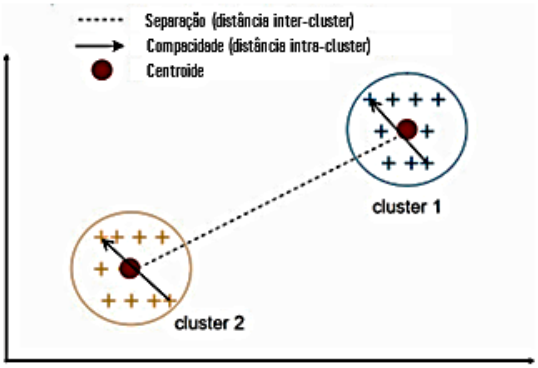


Figura 2.9 - Gráfico ilustrativo da compacidade e separação de clusters

Fonte: Adaptado de Ben Ncir *et al.* (2021).

A validação externa da clusterização é relativamente complexa, visto que os algoritmos são conduzidos de uma forma não supervisionada e não há um critério de validação externa disponível, isto é, não há um resultado pré-estabelecido como referência (AGGARWAL, 2015).

Seghier (2018) e Brun *et al.* (2007) apontam na literatura um vasto número de índices para a validação de clusterização, mais que 50 métodos, mas neste estudo será abordado dois dos principais.

2.6.4.1 Coeficiente Silhueta

O Coeficiente Silhueta, ou mais conhecido por *Silhouette*, segundo Siqueira Junior (2021), é um índice que avalia a boa clusterização de uma observação, calculando a distância média entre os *clusters.* Revela também a quão perto está cada ponto dos grupos vizinhos. É, portanto, uma medida de compacidade e de separação.

De acordo com Brun *et al.* (2007), se x é um ponto de um determinado *cluster* C_k e n_k é a quantidade de elementos desse grupo, então a largura do Coeficiente Silhueta de x é dada por:

$$S(x) = \frac{b(x) - a(x)}{\max[b(x), a(x)]} \quad (3)$$

Em que *a(x)* é a distância média entre o ponto e todos os outros do *cluster*. Já *b(x)* é a distância média mínima entre *x* e os pontos dos outros grupos a que não é pertencente. Por fim, o índice Silhueta global pode ser calculado por:

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} [\sum_{x \in C_k} S\,(x)] \quad (4)$$

De acordo com Kassambara (2015), a largura do Coeficiente Silhueta de determinado ponto abrange um intervalo de [-1,1]. De acordo com o mesmo autor, se o valor for próximo de -1, significa que este ponto está mais próximo de outro *cluster* do que aquele ao qual pertence. Caso próximo de 1, indica que a distância para o seu próprio *cluster* é significativamente menor que para outros grupos. Em geral, quanto maior o índice Silhueta (S), mais compacto e mais separado estão os *clusters.*

2.6.4.2 Índice de Validação de Dunn

O Índice Dunn (do inglês, *Dunn Index,* DI) consiste em outra forma de validação interna que determina a razão entre a distância mínima entre dois *clusters* e o tamanho do maior *cluster* (BRUN *et al.*, 2007).

Kassambara (2015) descreve as etapas para a aplicação da técnica:

1) Calcule a distância entre cada um dos objetos no *cluster* e os elementos dos demais;
2) Assuma a mínima distância de pares como a separação entre *clusters* (*min.separacao*);
3) Tomando cada *cluster,* calcule a distância entre os elementos internos desse grupo;
4) Como a compacidade interna do agrupamento, assuma a máxima distância intra-*cluster* (*max.diametro*);
5) Compute o Índice Dunn pela divisão de (*min.separacao*) por (*max.diametro*).

Ben Ncir *et al.* (2021) declaram que, quanto maior o valor do índice, mais evidências se tem dos bons resultados de clusterização.

3 METODOLOGIA

Este capítulo se dedica a explicar em detalhes os passos envolvidos na pesquisa, bem como os *softwares*, base de dados e características da empresa estudada.

3.1 Classificação da pesquisa

A Figura 3.1 mostra a classificação da pesquisa, pela ótica proposta por Turrioni e Mello (2012).

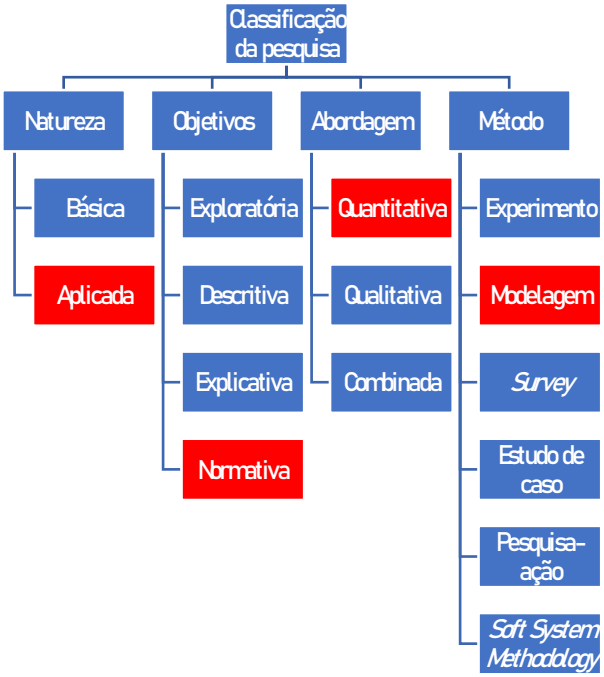


Figura 3.1 - Classificação da pesquisa

Fonte: Adaptado de Turrioni e Mello (2012).

Segundo Cauchick Miguel *et al.* (2010), as pesquisas em Engenharia de Produção são classificadas em quatro dimensões: natureza, objetivo, abordagem e método. Esta pesquisa pode ser classificada como aplicada, pois, de acordo com Santos e Parra Filho (2012), é conduzida por objetivos que almejam sua utilização prática.

Quanto à sua abordagem, este trabalho pode ser caracterizado como quantitativo. Segundo Appolinário (2013), uma pesquisa preponderantemente quantitativa é aquela em que a análise dos dados normalmente é realizada empregando-se a Estatística. Esse autor ainda afirma que o pesquisador se põe numa posição de neutralidade quanto ao objeto de pesquisa.

No tocante ao objetivo, pode ser classificado como normativo, pois está focada em definir estratégias após a identificação e análise dos *clusters*. Uma pesquisa normativa, levando em consideração o que afirmam Bertrand e Fransoo (2002), dá ênfase ao desenvolvimento de ações para a melhoria dos resultados disponíveis na literatura.

Por fim, quanto ao método, esta pesquisa se enquadra na modelagem, uma vez que utiliza "técnicas matemáticas para descrever o funcionamento de um sistema ou de parte de um sistema produtivo" (CAUCHICK MIGUEL *et al.*, 2010, p. 64) e também faz o uso de "técnicas computacionais para simular o funcionamento de sistemas produtivos a partir de modelos matemáticos" (CAUCHICK MIGUEL *et al.*, 2010, p. 64).

3.2 Materiais

O estudo foi desenvolvido em parceria com uma *startup*, que iniciou as suas atividades no ano de 2013 e promove um *marketplace* logístico no setor de transporte rodoviário de cargas. Ela é caracterizada por ser como provedora de tecnologia neutra, vertical, de mecanismo de transação quadro informativo, pública, que intermedia contratos de curta duração, oferece suporte até a fase de execução e transaciona cargas (ver *Marketplaces* Logísticos). A empresa disponibiliza uma plataforma *online* que realiza a correspondência de caminhoneiros autônomos a transportadoras de cargas. A Figura 3.2 apresenta o esquema como é realizada a conexão entre esses participantes da forma mais básica na plataforma.

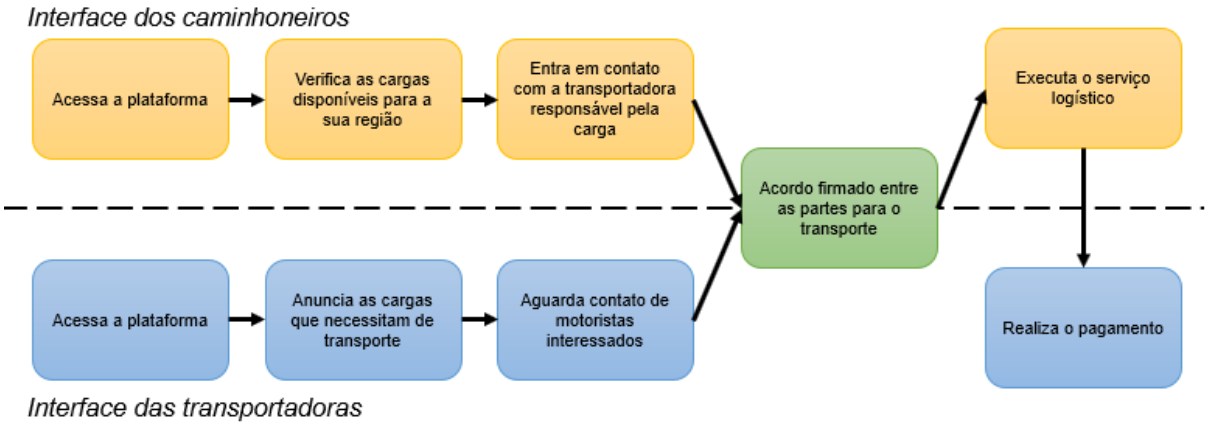


Figura 3.2 - Processo de transação básico entre transportadoras e caminhoneiros

Conforme a Figura 3.2, os caminhoneiros acessam à plataforma disponibilizada pelo *marketplace* pela interface apropriada. Através da geolocalização do motorista, o aplicativo

indica os fretes disponíveis nas proximidades. Em cada anúncio, os caminhoneiros podem acessar às informações de destino da carga, data de coleta e entrega, tipo de carga e valor do frete. O caminhoneiro escolhe o transporte que deseja realizar e entra em contato com a transportadora responsável pela carga. Do ponto de vista das transportadoras, estas também acessam a plataforma por uma interface idealizada para as suas necessidades e, então, disponibilizam as informações dos fretes que necessitam de transporte. Após o contato dos motoristas interessados, as empresas de transporte e os caminhoneiros estabelecem um acordo e o detalhamento da viagem. O caminhoneiro realiza o serviço logístico e recebe o pagamento da transportadora. A plataforma disponibiliza um sistema de pagamento próprio, parceria com empresas gerenciadoras de riscos e possibilidade de emissão de vale-pedágio e CIOT, isto é, Código Identificador de Operação de Transporte.

Os dados utilizados do presente trabalho são de caráter privado, de propriedade da empresa parceira. Os dados foram repassados em arquivos de extensão .csv (tipo de arquivo de texto para a transferência de informações, separadas por vírgulas) e são referentes ao território brasileiro e correspondem aos anúncios de carga na plataforma nos anos de 2019, 2020 e 2021, de janeiro a dezembro. Para se ter ideia do volume de suas transações, a empresa afirma que são mais de 18 mil empresas de frete e cerca de um milhão de motoristas cadastrados; totalizando cerca de 100 mil fretes mensalmente. As informações contidas nas bases de dados são explicadas na próxima seção, na etapa de compreensão dos dados.

3.3 Método

Como estrutura complementar para o processamento dos dados e aplicação dos métodos de clusterização, optou-se pelo desenvolvimento do *Cross Industry Standard Process for Data Mining* (CRISP-DM), de Chapman *et al*. (2000). Consiste numa metodologia para guiar os projetos de Mineração de Dados, desde a concepção de seus objetivos até a apresentação dos resultados. Providencia um esquema de todo o ciclo de vida dos dados. Provost e Fawcett (2013) a indicam como uma codificação do processo de Mineração de Dados. É iterativo, isto é, a cada ciclo de implementação, o conhecimento sobre os dados estudados é aperfeiçoado.

Schröer *et al.* (2021) e Palacios *et al.* (2017) a caracterizam por um modelo de processo independente da indústria, confiável e de grande aceitação na comunidade de Ciência de Dados. Apresenta a qualidade da padronização e de fácil aplicação, em comparação a outras metodologias. Engloba seis fases, que são demonstradas na Figura 3.3.

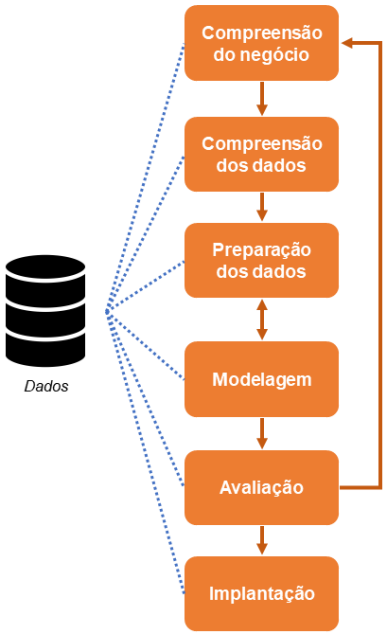


Figura 3.3 - Processo CRISP de mineração de dados

Fonte: Adaptado de Chapman *et al.* (2000) e Provost e Fawcett (2013)

Na etapa de compreensão do negócio, é entendido o problema que a organização almeja resolver e definido o que se pretende fazer com os dados. Essa etapa deve cuidadosamente abranger o cenário atual da entidade.

A segunda etapa consiste no reconhecimento de como um certo conjunto de dados poderia ajudar na resolução do problema. São listadas as limitações e pontos fortes dos dados, quais bases estão disponíveis e qual o preço para obtê-las. Se for em um ambiente organizacional, define-se também qual o responsável pela liberação. São realizadas aqui as tarefas de verificação e avaliação da qualidade dos dados.

Na terceira etapa, ocorre a preparação dos dados, onde são selecionadas as informações relevantes, definindo-se os critérios de inclusão e exclusão dos dados. Também entra nessa etapa a formatação das tabelas para uma extensão adequada para os algoritmos.

O passo seguinte, a modelagem, consiste em selecionar a técnica de modelagem, construindo o teste e o modelo. Todas as técnicas da mineração de dados podem ser usadas. Tudo dependerá do problema que está sendo abordado e das características dos dados. É relevante mostrar o motivo das escolhas. Nessa etapa também são configurados os parâmetros específicos do modelo. Os algoritmos são propriamente executados.

Na quinta fase, de avaliação, os resultados são verificados à luz dos objetivos do negócio. As saídas são interpretadas e também são pensadas as ações futuras. Acontece aqui também uma revisão do método, tendo em mente a performance dos algoritmos.

Por fim, a etapa de implantação consiste em dispor os resultados da mineração de dados para uso real de forma a retornar os investimentos dispendidos. Provost e Fawcett (2013) afirmam que esta etapa pode ser bem sutil, como apenas mudar os procedimentos de aquisição de dados, mudanças na estratégia de *marketing* ou novas operações resultantes dos *insights* da mineração de dados.

Na etapa de preparação dos dados também se descobre a tendência de clusterização pela Estatística de Hopkins e por *heatmaps*, já explicitada na Subseção Validação. Com essa tendência verificada, a modelagem dos métodos para a tarefa de clusterização poderá ser realizada, sendo, em seguida, conduzida a validação interna, já que a externa não é aplicável ao caso. Os parâmetros dos algoritmos poderão ser aperfeiçoados. Por fim, são apresentadas as propostas elaboradas a partir dos *insights* da *clusters*, configurando a etapa de implantação.

Especificamente, na etapa de preparação de dados é executado o processo ETL (do inglês, *extract, transform, and load*), que, de acordo com a IBM (2020), é um processo de integração de dados que combina dados de múltiplas fontes em uma só base consistente, que, posteriormente, é carregada em outro sistema.

Na etapa de extração, os dados são obtidos na forma estruturada, semiestruturada ou não estruturada. Já na etapa de transformação, considerada a mais crítica, os dados são limpos e propriamente formatados levando em consideração o sistema em que se objetiva conduzir as análises (CODELESS PLATFORMS, 2023). Essa etapa é crucial, pois nela se garante que os dados duplicados e anômalos (pontos de observação distantes das outras observações dentro de uma população normal) foram excluídos, bem como a padronização dos formatos, agregação dos dados, aplicação de cálculos específicos e a filtragem dos dados. No que diz respeito à etapa de carregamento, os dados são dispostos no sistema ou base de dados alvo do estudo (CODELESS PLATFORMS, 2023; IBM, 2020). Existe ainda a inclusão da etapa de análise de dados propriamente dita.

Khandavilli (2023) afirma que o processo ETL melhora a qualidade e a consistência dos dados, que, por sua vez, aperfeiçoa a acurácia e a confiabilidade da tomada de decisão. Acrescenta, ainda, que a integração de dados possibilita a visualização do modelo de negócio em uma só perspectiva.

No tocante a limpeza de *outliers*, é empregado o Método do Intervalo Interquartil, também conhecido em inglês por sua sigla IQR (*Interquartile Range*). De acordo com Perez e Tah (2020), a técnica consiste em medir a variabilidade dos dados e dividi-los em um conjunto de elementos ordenados em quartis. O IQR define um *outlier* como uma observação fora da

amplitude [$Q_1 – k(Q_3 – Q_1)$, $Q_3 + K(Q_3-Q_1)$], onde Q_1 é o primeiro quartil, Q_3 é o terceiro e k é uma constante. IQR seria a distância entre os primeiro e terceiro quartis.

Para o procedimento de clusterização, é crucial que os dados estejam na mesma escala para que eles sejam comparáveis. Mohamad e Usman (2013) declaram que este é o passo central da fase de pré-processamento dos dados. Eles afirmam que é necessário para a adequação às métricas de distância utilizadas nos algoritmos, tais como a Euclidiana, que é sensível a variações na magnitude ou escala dos atributos. Com a padronização, nenhuma variável se sobrepõe às outras em termos de relevância ou peso. Patro e Sahu (2015) indicam que há três métodos de padronização, o *Z-score* e o *Min-Max* e o Decimal. Para o trabalho, é proposta a adoção do Método *Min-Max* pela sua facilidade de implementação. Ele consiste na razão entre a diferença do valor pela observação mínima do conjunto de dados e a amplitude do conjunto de dados, dado pela diferença entre os valores máximo e mínimo.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \quad (3)$$

Sendo X, o conjunto de dados, x, uma das observações e x', o dado transformado.

Neste trabalho, são utilizados a linguagem R e o *software* RStudio®. A linguagem de programação R é tida como um recurso relevante para análise de dados, operações com matrizes e construções gráficas. Segundo Siqueira Junior (2021), recentemente, a linguagem se popularizou entre os profissionais como analistas de riscos, estatísticos e pesquisadores devido às suas configurações orientadas a objetos e à capacidade de desempenhar funções e pacotes idealizados por usuários, além de possuir uma sintaxe flexível e de fácil edição. O *software* é gratuito e de *open source*. Para Charalampopoulos (2020), essas características permitem que qualquer grupo de pesquisa o utilize sem limitações, independente dos sistemas operacionais e de fundos de apoio e, por ser *open source*, garante o desenvolvimento de novas ferramentas para o programa rapidamente.

Nas últimas três décadas, o programa tem exercido grande influência nos estudos do campo da Estatística, Bioinformática e Ciência de Dados. Conforme o que testificam Giorgi *et al.* (2022), a linguagem R é popularmente levantada como uma das 10 linguagens de programação mais difundidas mundialmente. De acordo com a Simplilearn (2023), o R possui mais de 10 mil pacotes armazenados no repositório CRAN (*Comprehensive R Archive Network*); e essa quantidade aumenta continuamente ao passo da produção de sua vasta comunidade. PanData (2023) aponta que o R é uma das linguagens mais populares empregadas

na análise de dados e no *Machine Learning.* De acordo com a mesma entidade, o R possui extensos pacotes e bibliotecas para a clusterização, tornando-o a escolha ideal para a execução da técnica. Dentre os recursos, estão os pacotes *factoextra,* que propicia uma visualização refinada dos resultados da clusterização; *cluster*, que contém os algoritmos e *hopkins*, que possui funções para avaliação da tendência de clusterização (DATANOVIA, 2018). Dentre os diversos trabalhos que utilizam a linguagem, pode-se mencionar o estudo de Choudhary e Saxena (2023) que aplicou a técnica de clusterização *Fuzzy C-Means* (FCM) em uma base de dados bancários a fim de verificar seu desempenho em relação ao algoritmo *K-Means.* Tem-se também o trabalho de Goksuluk *et al.* (2016) em que empregam a linguagem R para desenvolver uma ferramenta interativa para análise da curva de característica de operação do receptor (ROC, do inglês *Receiver Operating Characteristics*). Por fim, pode ser Schliep (2011) que, por meio da linguagem R, consegue construir um pacote para a reconstrução e análise filogenética através do método da máxima verossimilhança.

Já o RStudio® é um dos vários ambientes de desenvolvimento integrado que permitem o desenvolvimento do *script* de códigos em linguagem R de forma facilitada e desobstruída. O programa também é disponibilizado de forma grátis. De acordo com Paulson *et al.* (2012), o programa faz a combinação de vários componentes do R como *console,* edição, gráficos, histórico, suporte, entre outros, em uma plataforma de trabalho dinâmica e produtiva. De acordo com os mesmos autores, o *software* possibilita a definição clara de um fluxo de trabalho dentro de projetos e a utilização de poderosas ferramentas estatísticas.

Com relação aos códigos dos algoritmos de clusterização, são empregadas adaptações encontradas no trabalho de Siqueira Junior (2021). Os pacotes utilizados no trabalho são o *cluster*, de Maechler *et al.* (2022), que contém os algoritmos de clusterização propriamente ditos como o CLARA, *K-Means*, PAM; o pacote *hopkins,* de autoria de Wright (2022), que contém os algoritmos necessários para o cálculo da Estatística de Hopkins; e a biblioteca *factoextra,* de Kassambara e Mundt (2022), essencial para a plotagem dos resultados da clusterização, bem como os gráficos *heatmap.* Acrescenta-se a biblioteca *scales*, de Wickham *et al.* (2022), para a normalização dos dados.

Para que seja possível trabalhar com a alta dimensionalidade dos dados nos algoritmos, é empregada a Análise de Componente Principal (PCA), em inglês, *Principal Component Analysis.* Ela é um método que objetiva a redução da dimensionalidade de um conjunto de dados, mas, na medida do possível, mantendo a similaridade e o relacionamento entre eles. Atualmente, os dados estão sendo tomados em multivariedade para dar suporte à compreensão

dos relacionamentos numa base de dados. Com isso, os dados podem ser visualizados em uma tela de computador (KARAMIZADEH *et al.*, 2013).

Em suma, o algoritmo PCA pode ser explicado em delinear dados de um espaço dimensional maior (R^M) em um subespaço de dimensão menor (R^k); transformando os dados $X = \{x_1, x_2, ..., x_N\}$, onde N representa o total de observações e x_i a ordem dessas observações. Vale ressaltar que todas as observações devem possuir a mesma dimensão, isto é, cada ponto é representado por M variáveis. A direção do espaço PCA representa a direção da máxima variância do conjunto de dados. Esse espaço é concebido de um número de componentes principais. Cada componente principal tem uma robustez diferente de acordo com o valor da variância em sua direção (THARWAT, 2018).

4 DESENVOLVIMENTO

Neste capítulo, são descritas as etapas percorridas seguindo o modelo CRISP-DM. Começa-se pela compreensão do negócio, no que diz respeito à provedora de tecnologia do *marketplace* logístico. As etapas seguintes são a compreensão e a preparação dos dados, que abrangem os processos de análise da dimensionalidade da base de dados, remoção de valores inconsistentes e a verificação da tendência de formação de agrupamentos. Prossegue-se, portanto, com a modelagem, consistindo na execução do algoritmo escolhido e, por fim, a validação dos resultados da clusterização, conferida pelos índices de qualidade. Aqui, também é executado uma modelagem a nível de estado.

4.1 Compreensão do negócio

Este estudo fez uso da base de dados de anúncios de carga em um *marketplace* logístico nos anos de 2019, 2020 e 2021. A empresa opera com mais de 18 mil empresas de frete e cerca de um milhão de motoristas cadastrados; chegando à marca de 100 mil fretes mensais. A primeira etapa do CRISP-DM se refere a entender qual o problema de pesquisa. Esse problema consiste em identificar padrões de postagem de carga pelas empresas de frete em sua plataforma *online* pelos anos registrados, visto que, em razão da competitividade do setor rodoviário, é importante que a organização sempre procure atender às necessidades de seus clientes, isto é, transportadoras e caminhoneiros, oferecendo soluções estratégicas de acordo com a localização e tipo de serviços que eles oferecem.

Entretanto, os conjuntos de dados são extensos (três anos de registro totalizam cerca de 4 milhões de dados), aumentando a complexidade da análise. Para isso, tem-se as técnicas de clusterização aliadas ao *Machine Learning*, já explanadas no Capítulo de Fundamentação Teórica, que possibilitam a análise de um grande conjunto de dados e faz o agrupamento deles sem uma prévia classificação, pois os dados não estão previamente agrupados. Adicionalmente, para a resolução do problema de clusterização, é preciso compreender os algoritmos com relação às situações e à forma de suas aplicações.

Por meio da linguagem de programação R, espera-se auxiliar o processo decisório da empresa parceira, ao passo que identifica os melhores algoritmos para esse ramo logístico.

4.2 Compreensão dos dados

Os dados se referem a anúncios de cargas na plataforma logística de 2019, 2020 e 2021.

Em conformidade com Eberendu (2016), os dados enviados pela empresa podem ser caracterizados por estruturados, visto que seguem um formato específico de nomenclatura de variáveis e tem um tamanho definido. A estrutura das informações é explicada como se segue:

- **Data de criação completa (*create_at*):** Apresenta a data e hora em que o anúncio de carga foi realizado;
- **Data de criação (*date_create_at*):** Fornece somente o dia, mês e ano da criação do anúncio;
- **Código de identificação (*uuid*):** Uma série de 32 caracteres que distingue os anúncios;
- **Tipo de frete (*freight_type*):** Se agregação, fracionado ou completo, também denominado de lotação. Na modalidade lotação, toda a capacidade de um veículo é requisitada. Já na fracionada, apenas uma parte dela. O caso de agregação se refere ao caminhoneiro que é autônomo e fecha o contrato de prestação de serviços de transporte com uma determinada empresa por um certo período;
- **Distância (*distance_km*):** Dada em quilômetros, é a distância a ser percorrida entre a cidade de origem da carga até o seu destino;
- **Cidade de origem (*origin_city*):** De onde a carga será despachada;
- **Estado de origem (*origin_state*):** Se refere à unidade federativa de onde a carga será despachada;
- **Cidade de destino (*destination_city*):** Onde a carga deverá ser entregue;
- **Estado de destino (*destination_state*):** De forma análoga, o estado para onde a carga deverá ser transportada;
- **Preço a combinar (*price_to_match*):** Variável booleana que diz se a transportadora prefere não disponibilizar informações sobre o valor do frete e deixar em aberto para negociação com o caminhoneiro. Esta coluna é preenchida com **"t"** para verdadeiro e **"f"** para falso;
- **Preço por tonelada (*price_per_ton*):** Variável também booleana que se refere à preferência de a transportadora disponibilizar o preço do frete em razão da tonelada a ser transportada ou não. Esta coluna é preenchida com **"t"** para verdadeiro e **"f"** para falso;
- **Peso (*weight_kg*):** Dado em quilogramas, se refere à massa de produtos a ser transportada;

- **Quantidade de veículos (*truck_amount*):** Quantidade de caminhões necessários para o transporte da carga total. A transportadora pode optar em fazer um único anúncio para casos de múltiplos fretes idênticos. No *marketplace* logístico, podem ser contratados até 5 motoristas distintos para a mesma carga;
- **Valor (*value*)**: Dado em reais (R$), é o valor que a transportadora determina para o frete. Se a variável *price per ton* for verdadeira, então esse é o valor por tonelada transportada;
- **Valor do frete (*truck_value*):** Dado também em reais (R$), é o montante a ser pago pelo serviço logístico, observando a marcação da variável *price per ton*. Se o valor for pago em razão da tonelada, então esta informação é o resultado da multiplicação da variável *value* pela *weight km* dividida por 1000, já que o peso é dado em quilogramas. Caso contrário, é apenas a variável *value* replicada.

A Tabela 4.1 exibe a quantidade de anúncios que se encontram em cada ano:

Tabela 4.1 - Quantidade de dados das bases

Base de dados	Quantidade de anúncios registrados
2019	490.616
2020	1.533.296
2021	1.965.105
Total	**3.989.017**

Observa-se que, no ano de 2019, houve uma menor quantidade de registros. Acredita-se que esteja relacionado à estratégia de operação da empresa. Como *startup*, o seu negócio precisa ser escalável. Nesse ano, a organização recebeu investimentos estrangeiros para dobrar a sua base de usuários no ano seguinte.

Fez-se também uma análise exploratória dos dados. No que se refere ao conjunto total dos dados, tem-se como média de peso, 36.746 kg, a distância média percorrida, 792,28 km, a quantidade de caminhões em 2,9 veículos e o valor do frete médio em R$ 4.653,33. A tipologia de frete mais comum é a lotação, tendo São Paulo como estado de origem e de destino mais comuns. A Figura 4.1 mostra um recorte das entradas da base de dados da postagem de cargas.

created_at	date_created_at	uuid	freight_type	distance_km	origin_city	origin_state	destination_city	destination_state	price_to_match	price_per_ton	weight_kg	truck_amount	value	truck_value
2019-01-01 10:21:38+00	01/01/2019	5d0e2f69-de0b-11e9-8d1c-02cfbfd65790	fractioned	1905,327622	Guarulhos	São Paulo	Camaçari	Bahia	f	f	1000	1	600	600
2019-01-01 17:05:48+00	01/01/2019	5d0e359d-de0b-11e9-8d1c-02cfbfd65790	full	256,828922	Inácio Martins	Paraná	Ortigueira	Paraná	f	t	37000	20	23	851
2019-01-01 17:09:07+00	01/01/2019	5d0e37ba-de0b-11e9-8d1c-02cfbfd65790	full	296,603069	Guarapuava	Paraná	Ortigueira	Paraná	f	t	37000	12	30	1110
2019-01-01 17:27:12+00	01/01/2019	5d0e3dfd-de0b-11e9-8d1c-02cfbfd65790	full	2910,914321	São Paulo	São Paulo	Fortaleza	Ceará	f	f	14400	1	7500	7500
2019-01-01 17:27:23+00	01/01/2019	5d0e44fe-de0b-11e9-8d1c-02cfbfd65790	full	2910,914321	São Paulo	São Paulo	Fortaleza	Ceará	f	f	14400	1	7500	7500
2019-01-02 05:56:16+00	02/01/2019	5d0e4cda-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	1500	5	500	500
2019-01-02 05:56:50+00	02/01/2019	5d0e54d2-de0b-11e9-8d1c-02cfbfd65790	aggregation	35,877063	Cajamar	São Paulo	São Paulo	São Paulo	f	f	1500	10	400	400
2019-01-02 06:53:39+00	02/01/2019	5d0e5c17-de0b-11e9-8d1c-02cfbfd65790	fractioned	632,455918	São Paulo	São Paulo	Andradina	São Paulo	f	f	1500	1	1200	1200
2019-01-02 06:59:17+00	02/01/2019	5d0e6396-de0b-11e9-8d1c-02cfbfd65790	full	44,954224	Limeira	São Paulo	Leme	São Paulo	f	f	10000	1	700	700
2019-01-02 07:02:52+00	02/01/2019	5d0e6a97-de0b-11e9-8d1c-02cfbfd65790	full	146,490664	Pindamonhangaba	São Paulo	São Paulo	São Paulo	t	f	12000	1	0	0
2019-01-02 07:03:23+00	02/01/2019	5d0e71a9-de0b-11e9-8d1c-02cfbfd65790	full	1182,056873	Guaratinguetá	São Paulo	Carazinho	Rio Grande do Sul	t	f	20000	1	0	0
2019-01-02 07:11:16+00	02/01/2019	5d0e78d0-de0b-11e9-8d1c-02cfbfd65790	full	613,855024	Sorriso	Mato Grosso	Rondonópolis	Mato Grosso	f	t	1	50	160	0,16
2019-01-02 07:16:19+00	02/01/2019	5d0e7f62-de0b-11e9-8d1c-02cfbfd65790	aggregation	41,201056	Cajamar	São Paulo	São Paulo	São Paulo	f	f	1600	2	220	220
2019-01-02 07:16:42+00	02/01/2019	5d0e88a0-de0b-11e9-8d1c-02cfbfd65790	aggregation	41,201056	Cajamar	São Paulo	São Paulo	São Paulo	f	f	2500	2	341	341
2019-01-02 07:18:23+00	02/01/2019	5d0e9019-de0b-11e9-8d1c-02cfbfd65790	aggregation	41,201056	Cajamar	São Paulo	São Paulo	São Paulo	f	f	6000	2	896	896
2019-01-02 07:18:55+00	02/01/2019	5d0e9c32-de0b-11e9-8d1c-02cfbfd65790	aggregation	41,201056	Cajamar	São Paulo	São Paulo	São Paulo	f	f	12500	2	533	533
2019-01-02 07:19:58+00	02/01/2019	5d0e9ed3-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	1600	10	5700	5700
2019-01-02 07:21:06+00	02/01/2019	5d0ea179-de0b-11e9-8d1c-02cfbfd65790	full	436,904555	Duque de Caxias	Rio de Janeiro	Barueri	São Paulo	f	f	10000	2	1800	1800
2019-01-02 07:23:08+00	02/01/2019	5d0ea47f-de0b-11e9-8d1c-02cfbfd65790	full	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	2500	1	7000	7000
2019-01-02 07:23:31+00	02/01/2019	5d0ea757-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	6000	5	558	558
2019-01-02 07:23:33+00	02/01/2019	5d0ea92d-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	6000	5	558	558
2019-01-02 07:24:24+00	02/01/2019	5d0eb007-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	12000	10	551	551
2019-01-02 07:24:25+00	02/01/2019	5d0eb214-de0b-11e9-8d1c-02cfbfd65790	aggregation	54,718628	Jundiaí	São Paulo	São Paulo	São Paulo	f	f	12000	10	551	551
2019-01-02 07:26:04+00	02/01/2019	5d0eb3e7-de0b-11e9-8d1c-02cfbfd65790	aggregation	21,412344	São Bernardo do Campo	São Paulo	São Paulo	São Paulo	f	f	2500	8	283	283
2019-01-02 07:26:51+00	02/01/2019	5d0ebb83-de0b-11e9-8d1c-02cfbfd65790	aggregation	21,412344	São Bernardo do Campo	São Paulo	São Paulo	São Paulo	f	f	4000	5	340	340
2019-01-02 07:27:42+00	02/01/2019	5d0ebd93-de0b-11e9-8d1c-02cfbfd65790	aggregation	21,412344	São Bernardo do Campo	São Paulo	São Paulo	São Paulo	f	f	12500	8	683	683
2019-01-02 07:44:41+00	02/01/2019	5d0ec003-de0b-11e9-8d1c-02cfbfd65790	fractioned	128,844289	Rio de Janeiro	Rio de Janeiro	São José do Vale do Rio Preto	Rio de Janeiro	f	f	90	1	300	300
2019-01-02 07:46:50+00	02/01/2019	5d0ec75c-de0b-11e9-8d1c-02cfbfd65790	full	407,534511	Boa Ventura de São Roque	Paraná	Paranaguá	Paraná	f	t	8000000	20	76	608000
2019-01-02 07:55:41+00	02/01/2019	5d0ec996-de0b-11e9-8d1c-02cfbfd65790	full	387,844734	CÃ¢ndido de Abreu	Paraná	Paranaguá	Paraná	f	t	700000	10	76	53200
2019-01-02 08:01:13+00	02/01/2019	5d0ed107-de0b-11e9-8d1c-02cfbfd65790	full	194,757599	Jaguariaíva	Paraná	Ortigueira	Paraná	f	t	37000	8	13	481
2019-01-02 08:02:07+00	02/01/2019	5d0ed31a-de0b-11e9-8d1c-02cfbfd65790	full	576,787475	Cotia	São Paulo	Guarani	Minas Gerais	f	f	12500	1	1831,68	1831,68
2019-01-02 08:02:10+00	02/01/2019	5d0eda38-de0b-11e9-8d1c-02cfbfd65790	full	2650,3557	São Bernardo do Campo	São Paulo	Parauapebas	Pará	f	f	6000	1	7300	7300
2019-01-02 08:02:21+00	02/01/2019	5d0edc49-de0b-11e9-8d1c-02cfbfd65790	full	416,904623	Candói	Paraná	Paranaguá	Paraná	f	t	1000000	10	76	76000
2019-01-02 08:03:36+00	02/01/2019	5d0ee31a-de0b-11e9-8d1c-02cfbfd65790	aggregation	7,449992	São Caetano do Sul	São Paulo	São Bernardo do Campo	São Paulo	f	f	1500	1	200	200
2019-01-02 08:05:39+00	02/01/2019	5d0ee55b-de0b-11e9-8d1c-02cfbfd65790	full	1766,467028	Primavera do Leste	Mato Grosso	Paranaguá	Paraná	f	t	1	50	180	0,18
2019-01-02 08:10:29+00	02/01/2019	5d0eec83-de0b-11e9-8d1c-02cfbfd65790	full	521,858932	Honório Serpa	Paraná	Paranaguá	Paraná	f	t	300	10	76	22,8
2019-01-02 08:20:52+00	02/01/2019	5d0eeed6-de0b-11e9-8d1c-02cfbfd65790	fractioned	576,427745	Contagem	Minas Gerais	Porteirinha	Minas Gerais	f	f	10000	1	1150	1150
2019-01-02 08:45:00+00	02/01/2019	5d0ef61b-de0b-11e9-8d1c-02cfbfd65790	full	298,547951	Marechal CÂndido Rondon	Paraná	Maringá	Paraná	f	t	1	50	50	0,05
2019-01-02 08:45:00+00	02/01/2019	5d0efcac-de0b-11e9-8d1c-02cfbfd65790	fractioned	24,150057	Rio de Janeiro	Rio de Janeiro	Rio de Janeiro	Rio de Janeiro	f	f	60	1	130	130
2019-01-02 08:45:15+00	02/01/2019	5d0f046d-de0b-11e9-8d1c-02cfbfd65790	aggregation	21,412344	São Bernardo do Campo	São Paulo	São Paulo	São Paulo	f	f	1600	8	241	241
2019-01-02 08:48:55+00	02/01/2019	5d0f0b9d-de0b-11e9-8d1c-02cfbfd65790	fractioned	111,518299	Guarulhos	São Paulo	Sorocaba	São Paulo	f	f	33	1	100	100
2019-01-02 08:49:27+00	02/01/2019	5d0f131f-de0b-11e9-8d1c-02cfbfd65790	full	191,723528	Ubiratã	Paraná	Sarandi	Paraná	f	t	1	50	38,07	0,03807

Figura 4.1 - Recorte da base de dados de postagem de anúncio de cargas

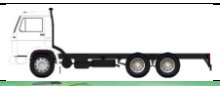
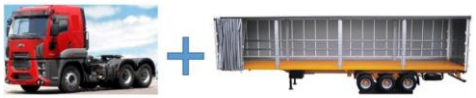
4.3 Preparação dos dados

Nessa etapa foi executado o processo de ETL, como fora explanado no Capítulo Metodologia. A etapa de extração dos dados foi realizada por um funcionário da empresa provedora a partir da base própria.

A transformação dos dados foi a etapa seguinte. Os arquivos foram enviados na forma de arquivo de texto, utilizando-se o RStudio. Utilizando a função nativa do RStudio *rbind*, as bases foram empilhadas em ordem cronológica.

Foi necessário também retirar as inconsistências no que se refere a valores de peso da carga (*weight_kg*), valor do frete (*truck_value*), distância (*distance_km*) e na variável quantidade de veículos (*truck_amount*). Se referindo às inconsistências, foram observados, por exemplo, valores de frete da ordem de R$ 10 bilhões de reais por um único veículo requisitado. Detectou-se também peso de carga da ordem de 250 milhões de quilos. Esse tipo de entrada claramente se configuravam como inconcebíveis para análise de dados. A plataforma não limita o usuário a digitar valores para essas variáveis. O tratamento dado a variável peso da carga foi feito com base na relação de veículos admitidos na plataforma, como são mostrados no Quadro 4.1.

Quadro 4.1 - Tipologia de veículos de transporte da plataforma

Veículo	Tipo	Categoria	Quantidade de eixos	Capacidade de carga
	Utilitário/Fiorino	Leve	2	500kg
	Van	Leve	2	1200kg
	HR	Leve	2	1.500kg
	VuC	Médio	2	3.000kg
	Toco	Médio	2	8.000kg
	Truck	Médio	3	14.000kg
	BiTruck	Médio	4	22.000kg
	Carreta simples	Pesado	5	25.000kg
	Carreta LS	Pesado	6	32.000kg
	Bitrem 7 eixos	Pesado	7	40.000kg

Quadro 4.1 - Tipologia de veículos de transporte da plataforma (cont.)

Veículo	Tipo	Categoria	Quantidade de eixos	Capacidade de carga
	Bitrem 9 eixos	Pesado	9	52.000kg
	Rodotrem	Pesado	9	50.000kg
	Tritrem	Pesado	9	65.000kg

Portanto, foi considerado como limite máximo de peso de caminhão a configuração do tritrem de 65.000 kg. Todos os valores acima foram considerados como inconsistentes, sendo eliminados da base. Para as outras variáveis, foi realizado a exclusão dos *outliers* pelo Método do Intervalo Interquartil, já que não se tem limitação para a entrada dessas variáveis e nem a indicação por parte da empresa dos valores máximos e mínimos admitidos. O valor de *k* foi selecionado em 1,5, ou seja, os limites inferior e superior distam do intervalo interquartil em 1,5 vezes. Como resultado da limpeza, a base integrada possui agora 3.642.836 entradas de fretes, ou seja, os dados removidos representam cerca de 8,67% da base original.

Foi realizado o processo de distinção da base nacional em duas: uma contendo apenas as variáveis quantitativas, isto é, peso da carga, distância percorrida, valor do frete e quantidade de veículos, chamada aqui de base **B1**; e outra contendo as variáveis quantitativas e as categóricas, denominada aqui de **B2**. As variáveis categóricas foram transformadas no tipo *dummy*, que, de acordo com Güneri e Durmuş (2020), são variáveis binárias (0 ou 1), concebidas para a representação de uma variáveis com duas ou mais categorias. Conforme o mesmo autor, elas devem ser utilizadas sempre que se deseja a inclusão de variáveis categóricas em modelos que admitem somente variáveis numéricas, que é o caso dos algoritmos de clusterização.

Essas variáveis foram adicionadas para demarcar o estado de origem e de destino da carga a partir das variáveis *origin_state* e *destination_state*. Escolheu-se fazer o processo por estado pelo seu menor número se comparado a quantidade de cidades, o que tornaria o trabalho demasiadamente demorado e complexo. Utilizou-se a seguinte nomenclatura: a variável *o_ac* é a variável *dummy* que indica se a carga saiu do estado do Acre. Já a variável *d_ac* indica se a carga tem como destino uma das cidades do Acre. O mesmo esquema é aplicado para as demais 26 unidades federativas da União. Aplicou-se esse tratamento também à variável *freight_type*, que passou a ser *full*, *fractioned* e *aggregation*.

A etapa seguinte consistiu em padronizar as bases para que os dados pudessem ser comparáveis entre si. Para esse processo, utilizou o método de normalização *Min-Max*. O pacote utilizado para este fim foi o *scales*, de Wickham *et al.* (2022). Os valores máximos e mínimos registrados após a limpeza são mostrados na Tabela 4.2.

Tabela 4.2 - Máximos e mínimos das variáveis quantitativas das bases

Variável	Máximo	Mínimo
Distância (km)	2.210	0
Quantidade de veículos	10	0
Peso da carga (kg)	65.000	0
Valor do frete (R$)	10.749	0

Para explicar os valores mínimos, alguns embarcadores colocam o peso de 0 kg e quantidade de veículos 0 não sabendo, na verdade, a totalidade a ser carregada, preferindo passar essas informações no contato com o motorista. É comum, no setor agropecuário, o carregamento pelo máximo da capacidade do veículo que o motorista contatado possui. Outra questão importante diz respeito à variável valor do frete com valores de zero, indicando que a transportadora deixou o preço a combinar. As distâncias zero demarcam fretes que acontecem em pontos dentro de uma mesma cidade, pois o nível de detalhamento da rota no *marketplace* logístico é municipal.

A etapa final do processo ETL foi consolidada com o carregamento da base tratada para o procedimento de Análise de Tendência de Clusterização. E para essa etapa, foi utilizado o método estatístico pela Estatística de Hopkins. Para o favorecimento da eficiência computacional, tomou-se 100 observações, utilizando-se a função *hopkins* da biblioteca *hopkins*, de Wright (2022). Também foi configurada uma semente de 123 no início do processo, para que os processos fossem reprodutíveis. Obteve-se o resultado apresentado na Tabela 4.3.

Tabela 4.3 - Resultados para a Estatística de Hopkins

Base	Estatística de Hopkins
B1	0,99
B2	1

A partir desses resultados, é possível identificar que ambas as bases de dados possuem uma tendência de formação de agrupamentos, já que os valores da Estatísticas de Hopkins são maiores que 0,5.

Fez-se também o Método Visual da Tendência de Clusterização por meio de *heatmaps*. Foram tomadas 1.000 observações aleatórias de cada uma das bases para a condução do experimento, utilizando as funções da biblioteca *factoextra*, de Kassambara e Mundt (2022). A Figura 4.2 apresenta o *heatmap* da base **B1**.

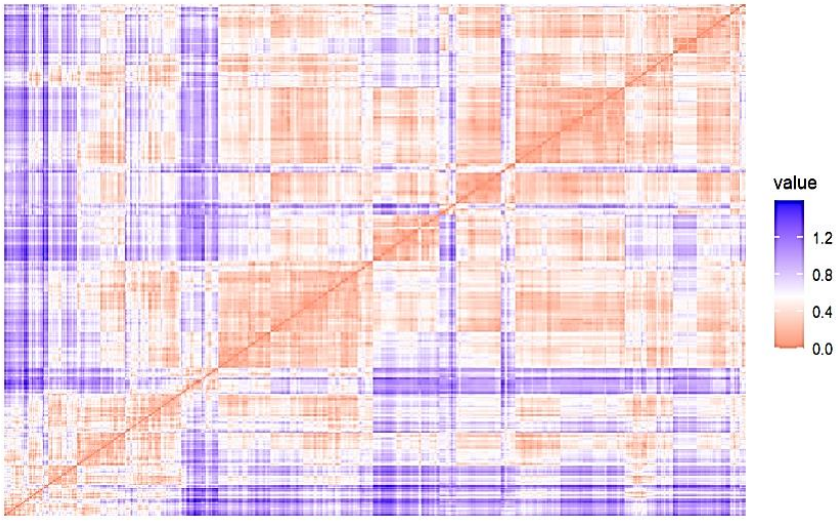


Figura 4.2 - *Heatmap* da base B1

Indivíduos com a cor vermelha indicam uma alta similaridade entre as observações e em azul uma alta dissimilaridade. Segundo Kassambara (2017), o Método Visual de Tendência de Clusterização detecta a tendência visualizando os blocos escuros de forma quadrada ao longo da diagonal do gráfico. Nota-se uma tendência de formação de cinco agrupamentos, mas não tão homogêneos, pois verifica-se a presença de muitos pontos azuis dentro dos blocos vermelhos. Já a Figura 4.3 mostra o resultado de *heatmap* para a base B2.

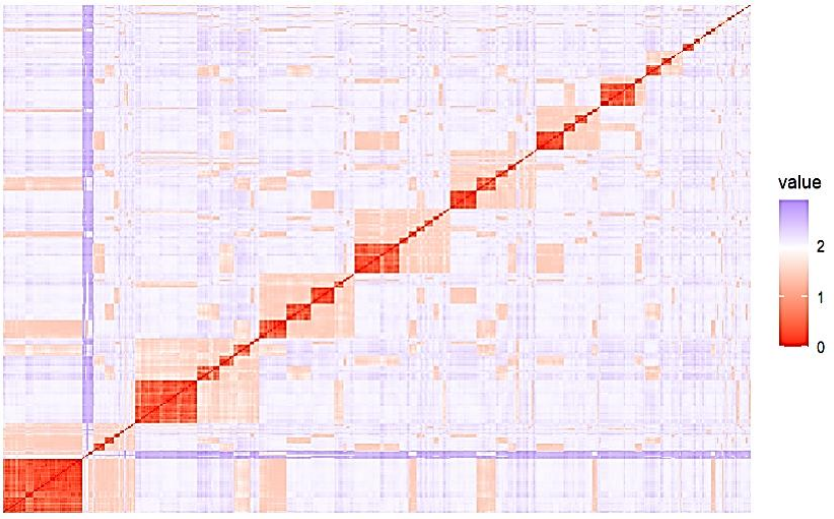


Figura 4.3 – *Heatmap* da base B2

Pode-se verificar uma tendência de clusterização, representada pelos oito blocos em vermelho pela diagonal. Pode-se observar que, dentro do *clusters*, ainda existem elementos de

maior similaridade do que o resto do agrupamento, representados pelas regiões de vermelho ainda mais escuro. Portanto, confirma-se a clusterabilidade dos dados.

4.4 Modelagem

Para realização da clusterização, escolheu-se a técnica de particionamento, mais precisamente o algoritmo CLARA, em razão das dimensões das bases de dados. O algoritmo consegue ser eficiente computacionalmente, visto que os dados não precisam ser todos carregados e analisados, mas apenas uma amostra deles. O algoritmo basicamente é uma versão do PAM, que é aplicado múltiplas vezes, e tem-se como resultado a melhor dentre as amostras. A sua interpretação também é facilitada, pois ao invés de retornar centroides como representantes do *clusters*, isto é, um ponto artificial contendo a média, ele retorna um dos integrantes do grupo. Para esta etapa, foi empregado o pacote *cluster*, de Maechler *et al.* (2022).

Kassambara (2017) afirma que determinar o número ótimo de *clusters* é um problema fundamental na clusterização de particionamento. O mesmo autor reitera que não há uma resposta definitiva a essa questão, sendo, de cera forma, subjetiva e que depende muito do método a ser usado para a mensuração das similaridades e dos outros parâmetros do algoritmo. Para este estudo, portanto, foi considerado uma amplitude para o número de *clusters* (k) de 2 a 10, já que maiores valores aumentariam consideravelmente o esforço computacional e, a princípio, não teriam utilização prática. A determinação do melhor particionamento é feita por medidas de validação.

Primeiramente, fez-se a clusterização utilizando a base **B1**, isto é, sem as variáveis de indicação do tipo de frete e da origem e do destino da carga. Utilizando a função *clara*, configurou-se os valores de k, mantendo-se a métrica de distância Euclideana e 100 amostras de 1.000 elementos cada. Para mais informações, ver Apêndice A. A quantidade padrão de amostras (*samples*) da função é cinco, devido às razões de compatibilidade na época em que foi concebido o algoritmo. Contudo, Maechler *et al.* (2022) recomendam a configuração desse parâmetro em uma proporção bem maior que a padrão. Para o tamanho da amostra (*sampsize*), os mesmos autores também recomendam um valor maior do que o padrão, que seria o mínimo entre o número de observações e $(40 + 2 * k)$. No caso de $k = 2$, o tamanho da amostra seria de 44 pontos. Todavia, eles recomendam cautela nessa escolha visto que o tempo computacional aumenta quadraticamente. Para a plotagem dos gráficos utilizando PCA, foi empregada a biblioteca *factoextra*, de Kassambara e Mundt (2022). Os resultados são mostrados na Figura 4.4.

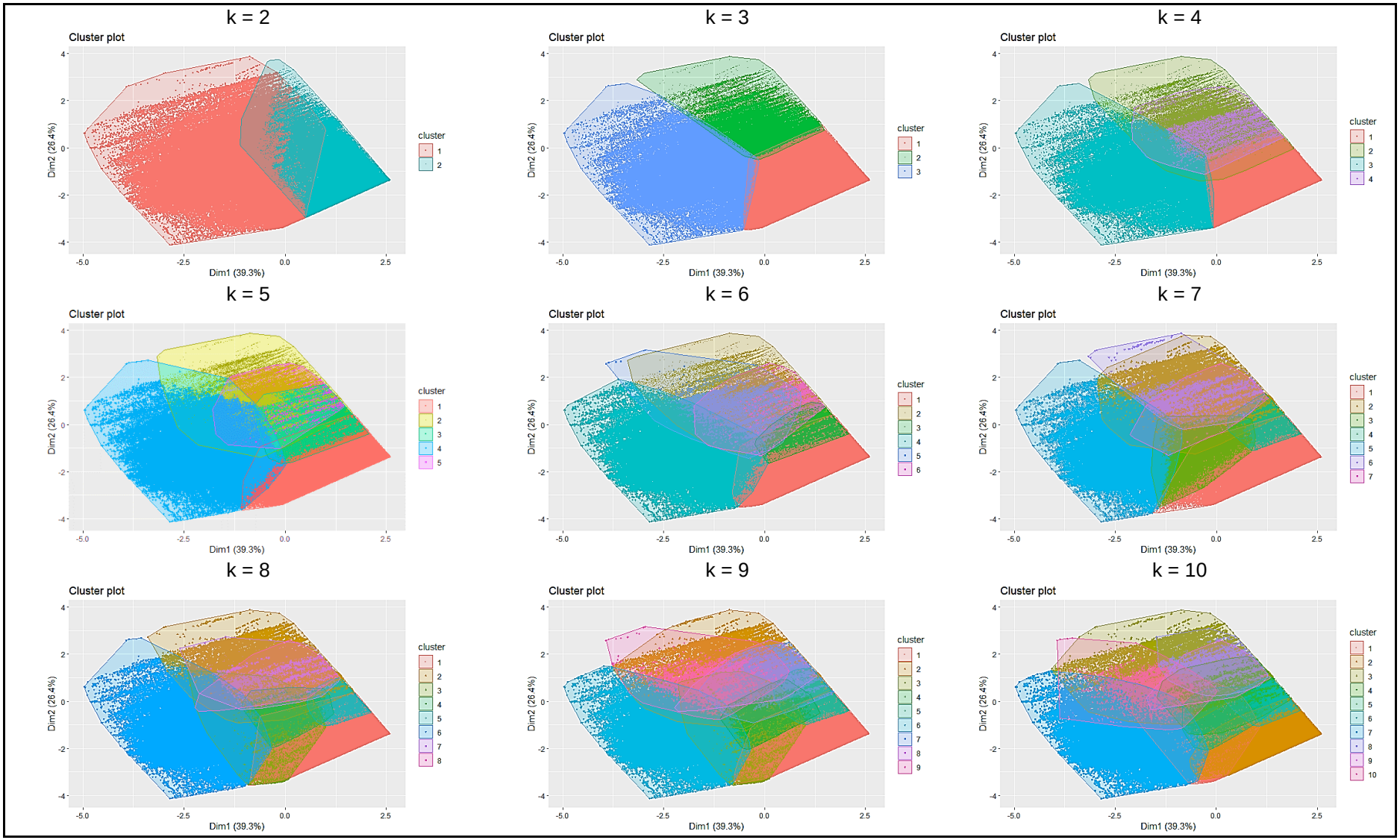


Figura 4.4 - Gráficos de dispersão com a indicação dos *clusters* para a base B1

De acordo com a Figura 4.4, percebe-se, no canto inferior esquerdo dos gráficos, uma grande porção de dados com baixa quantidade de veículos e baixo peso (é necessário visualizar o cruzamento dos eixos zero na dimensão 1 e na dimensão 2). Pontua-se também uma distribuição uniforme dos dados. Para todos os valores de *k*, houve sobreposição de agrupamentos, levando a crer que os elementos nessas regiões apresentam um certo grau de pertencimento a mais de um *cluster*. Já a Figura 4.5 apresenta os resultados da clusterização para a base **B2**, isto é, com a indicação das variáveis do tipo de frete e da origem e destino da carga.

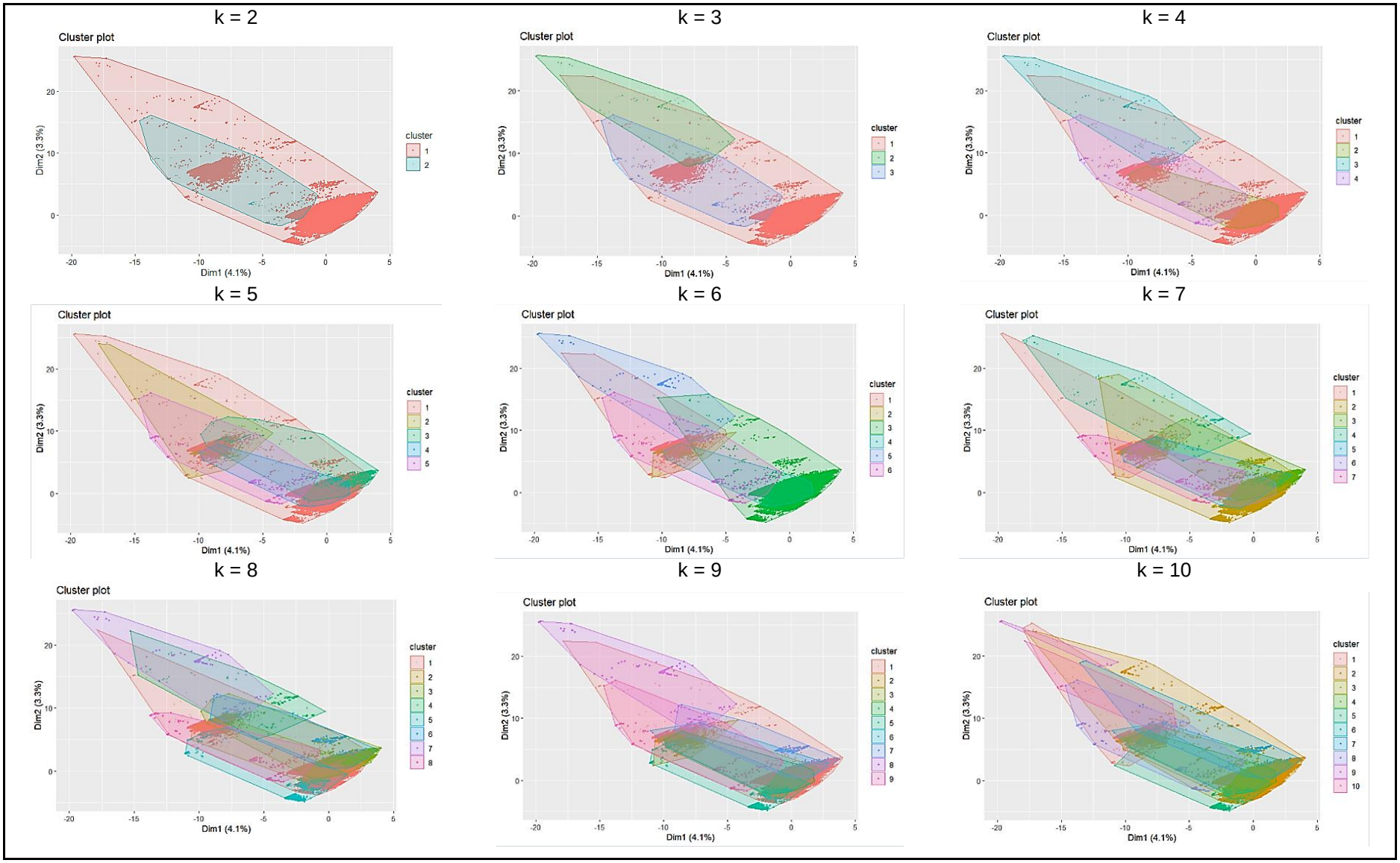


Figura 4.5 - Gráficos de dispersão com a indicação dos *clusters* para a base B2

A presença de muitos pontos extremos faz com a forma dos *clusters* se disperse. Um grande agrupamento de dados que apresentam valores de frete, distância elevados e tipologia de frete lotação, concebidos pelos pontos no canto inferior direito dos gráficos. Identifica-se também uma concentração de dados no canto superior esquerdo concebendo os fretes do tipo fracionado e agregado. Percebe-se uma alta sobreposição dos agrupamentos, mesmo com as concentrações identificadas anteriormente. A tendência é quanto maior o número de *clusters*, mais os agrupamentos se sobrepõem.

4.5 Avaliação

Para a validação interna do resultados, foi conduzida a análise do Índice Silhueta, importante para conferir a qualidade da clusterização. Para isso, foi utilizado a biblioteca *factoextra*, de Kassambara e Mundt (2022), executando-se a função *fviz_silhouette*. A Figura 4.6 exibe os resultados do Índice Silhueta de cada observação da amostra ótima, para cada *k*, para a base **B1**.

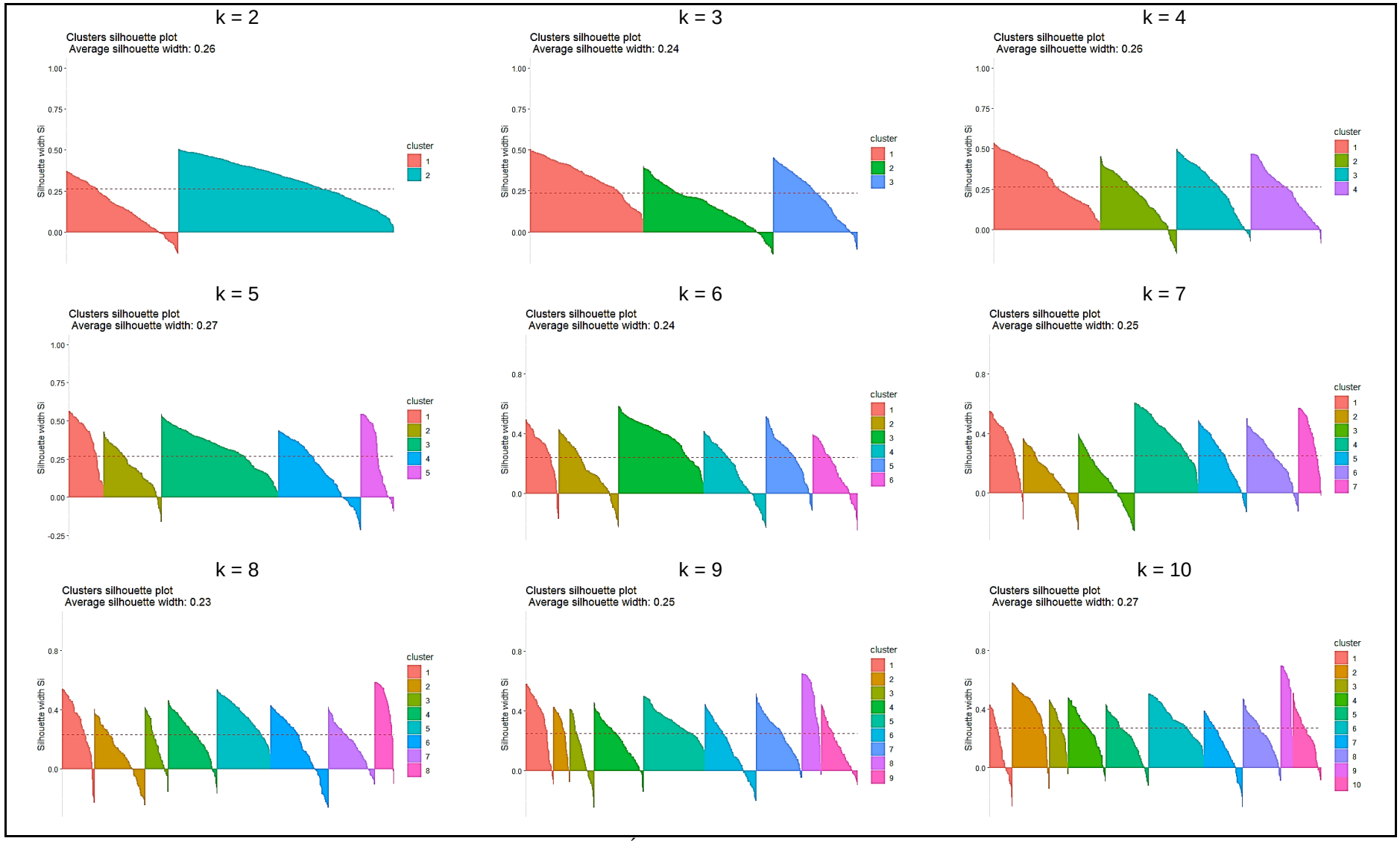


Figura 4.6 - Índice Silhueta para a base B1

A Figura 4.7 apresenta o Índice Silhueta médio das clusterizações implementadas para cada valor de *k* na base **B1**.

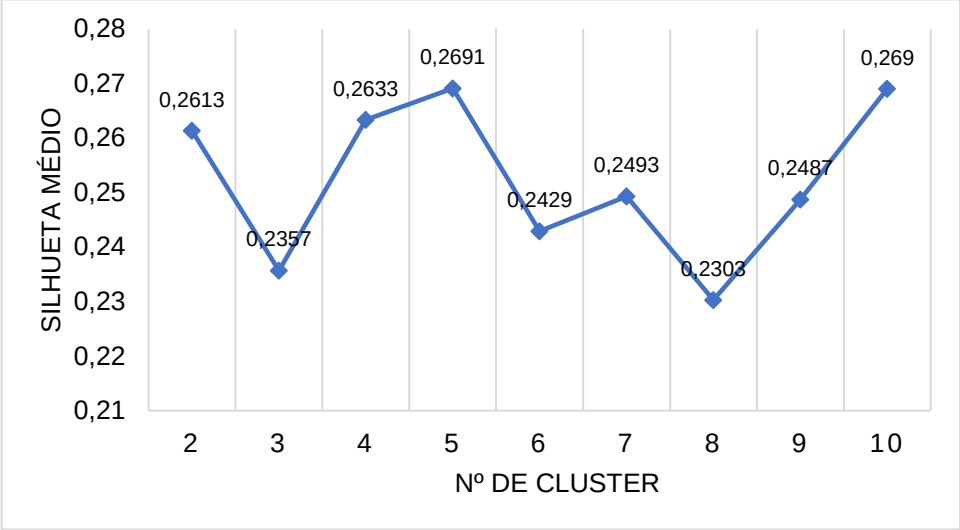


Figura 4.7 - Índice Silhueta médio para cada valor de k executado na base B1

Nota-se que os índices médios não apresentam ganhos significativos. O cenário que produziu o melhor resultado está para *k* = 5. O índice é uma estimativa da distância média entre os agrupamentos, e está compreendido entre -1 e 1. O índice médio mais próximo de 1 (um) revela boa compacidade e separação dos agrupamentos. Portanto, pode-se concluir, pelas Figuras 4.6 e 4.7, uma qualidade razoável de clusterização da base **B1**.

Se a empresa optar por esse processo de clusterização, a Tabela 4.4 revela, portanto, a definição dos medoides para cada *cluster*.

Tabela 4.4 - Características dos medoides para *k* = 5 da base B1

Medoide	Distância (km)	Peso (kg)	Qte. veículos	Valor do frete (R$)	Tamanho do *cluster*
1	398	10.000	1	450	391.053
2	493	32.000	6	1.920	786.757
3	380	32.000	2	1.600	1.265.863
4	1.225	37.000	3	5.550	849.821
5	591	51.340	1	1	349.342

Verifica-se que o grupo representado pelo medoide 3 é o mais representativo, possuindo a maioria das postagens de carga. Esse *cluster* apresenta viagens de média duração (cerca de 6 horas, considerando uma velocidade de tráfego dos veículos de 60 km/h); são executados por caminhões de categoria pesada (ver Quadro 4.1), portanto, veículos de 6 eixos.

O Grupo 4 representa os fretes de longas distâncias (mais de 20 horas de viagem) e os de maiores valores de frete. É o segundo grupo mais expressivo na plataforma. Também

configura a categoria pesada de veículos (7 eixos). Para esses, são necessários três veículos, consequentemente três motoristas para executar o frete.

O Grupo 5 caracteriza os fretes com carga maiores cargas, 51 toneladas. Cabe dizer que o valor de frete pago é expresso em R$ 1,00; sendo possivelmente como uma estratégia de preço a combinar. Pela sua menor quantidade de elementos, não é tão frequente no *marketplace*. Por fim, cabe ressaltar aqui as viagens representadas pelo medoide 1, são as que possuem menor peso, podendo simbolizar entregas de última milha.

É importante ressaltar que, ao tratar os grupos com interpretação extraída do medoide, deve-se atentar para a dispersão da clusterização. Neste caso, a variabilidade é expressiva.

Já para a base **B2**, tem-se os seguintes resultados do Índices Silhueta médio e individuais, exibidos pelas Figuras 4.8 e 4.9.

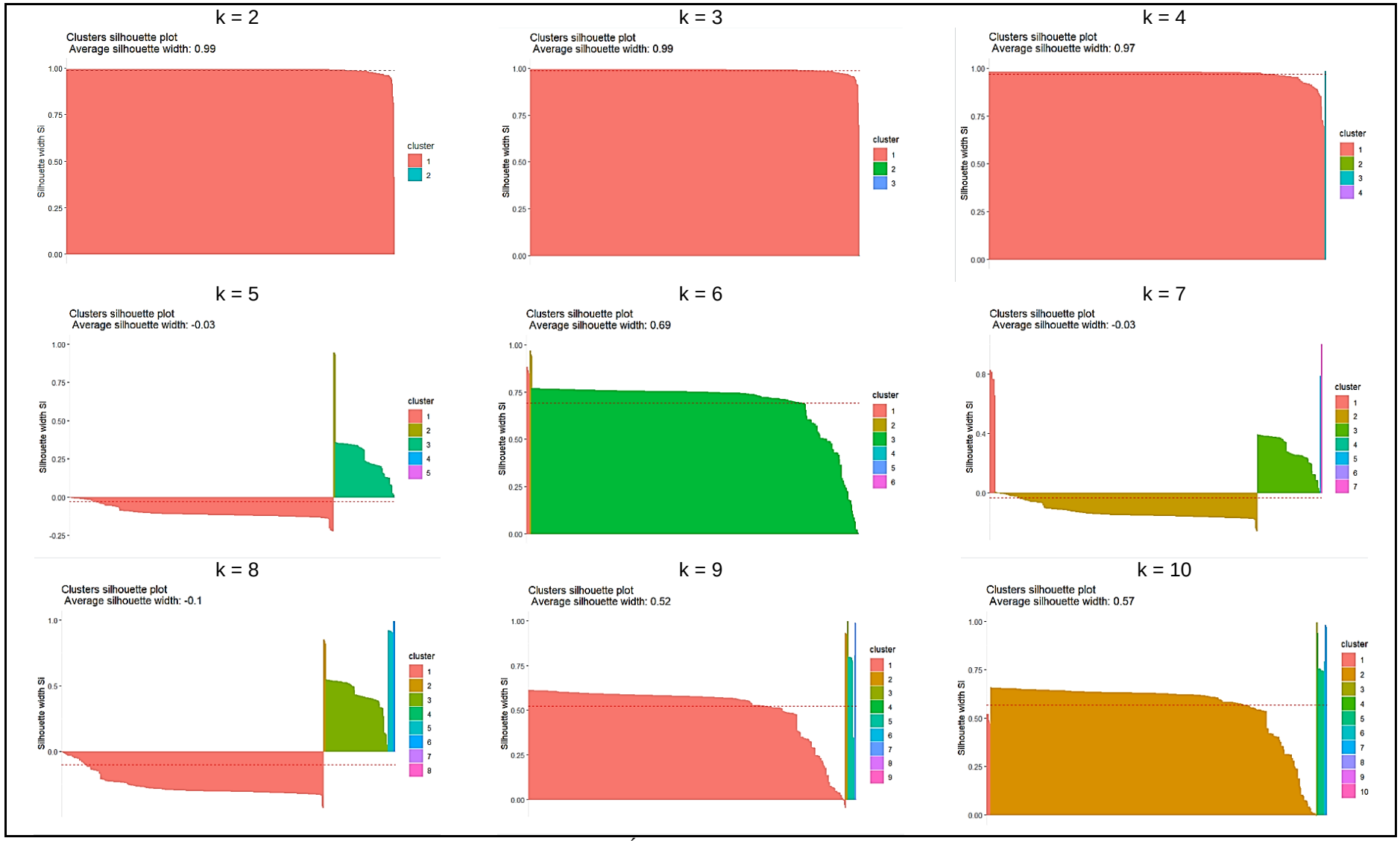


Figura 4.8 - Índice Silhueta para a base B2

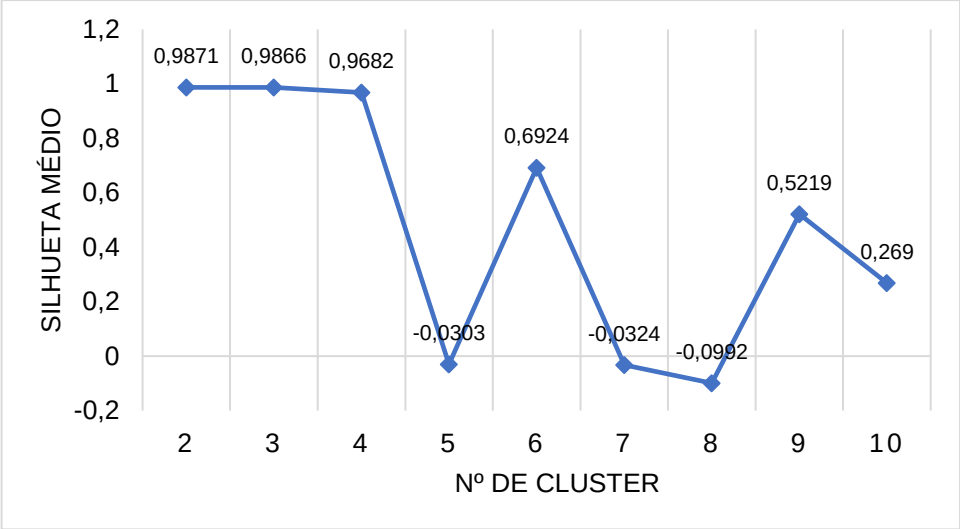


Figura 4.9 - Índice Silhueta médio para cada valor de *k* executado na base B2

O cenário de melhor resultado foi para *k* = 2, indicando uma ótima clusterização. Se a empresa optar por esse processo, tem-se os seguintes medoides, apresentados na Tabela 4.5.

Tabela 4.5 - Características dos medoides para *k* = 2 da base B2

Medoide	Tipo de frete	Origem	Destino	Distância (km)	Peso (kg)	Veículos	Valor do frete (R$)	Tamanho do *cluster*
1	Lotação	São Paulo	São Paulo	555	32.000	3	2.304	3.642.147
2	Lotação	Acre	Acre	0	8.000	1	300	689

O medoide 1 representa o grupo mais significativo, com mais 3 milhões de entradas. Todavia, dentro desse *cluster*, a partir da Figura 4.4, é possível perceber dois grupos bem distintos, que o algoritmo identificou como integrantes do mesmo agrupamento. De acordo com esse processo de clusterização, os fretes do cenário nacional seguem as características das viagens saindo do estado de São Paulo e chegando no estado de São Paulo. São fretes tipo lotação, percorrendo médias distâncias (viagens de 9 horas, considerando velocidade de tráfego de 60 km/h). Os veículos para o transporte são considerados de carga pesada, possuindo 6 eixos. O valor do frete é alto, podendo associar ao peso da carga elevada, demandando mais combustível e veículos mais potentes, o que aumenta o custo.

Já o medoide 2 simboliza um grupo mais singelo, com poucos representantes e com características peculiares. São fretes que ocorrem internamente em alguma das cidades do Acre

(indicado pela distância zero), com carga média, precisando de veículos de apenas dois eixos. O pagamento por eles fica em torno dos R$ 300,00.

4.6 Modelagem São Paulo

Graficamente, o *cluster* mais significativo da clusterização da base **B2** releva dois agrupamentos dentro de um só. Deste modo, buscou-se entender mais a fundo a estruturação por trás dos fretes internos do estado de São Paulo. A base que contém as variáveis categóricas (tipo de frete) e as quantitativas (peso da carga, distância, valor do frete, quantidade de veículos) é denominada **B3**. Já a base que considera apenas as variáveis quantitativas é denominada **B4**. Ambas as bases possuem 379.006 postagens de carga realizadas entre os anos 2019 e 2021.

Os dados foram extraídos do conjunto sem tratamento, desta forma, foi necessária a normalização novamente. A Tabela 4.6 exibe os valores mínimos e máximos dos novos conjuntos.

Tabela 4.6 - Valores máximos e mínimos das bases do estado de São Paulo

Variável	Máximo	Mínimo
Distância (km)	866	0
Peso (kg)	65.000	0
Quantidade de veículos	8	0
Valor do frete (R$)	6.446	0

Também foi verificado se os dados contidos nesse subgrupos não se distribuíam de forma uniforme pelos métodos estatístico e visual, apresentados pela Tabela 4.7 e as Figuras 4.9 e 4.10. Para a Estatística de Hopkins, foram tomadas 100 amostras. Já para o método visual, foram adotadas 1.000 observações.

Tabela 4.7 - Estatística de Hopkins para as bases de São Paulo

Base	Estatística de Hopkins
B3	0,99
B4	0,99

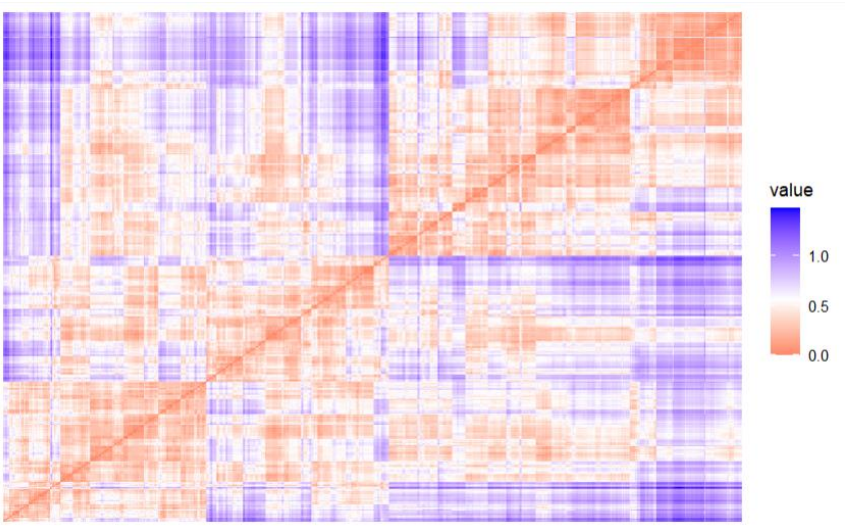


Figura 4.10 - *Heatmap* da base B3

A Figura 4.10 se apresenta com um nível de homogeneidade considerável, contudo contando-se os blocos em vermelhos pela linha diagonal do gráfico, pode-se dizer que a tendência é a formação de três *clusters*. Já para a base **B4**, tem-se a Figura 4.11.

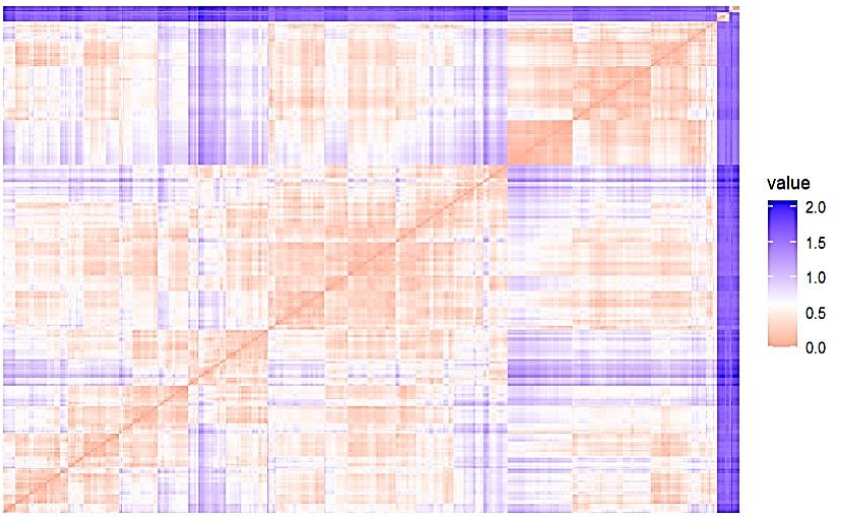


Figura 4.11 - *Heatmap* da base B4

A formação de agrupamentos pode ser deduzida de dois a três *clusters*. Tanto as Estatísticas de Hopkins quanto os *heatmaps* indicam a clusterabilidade das bases, embora a **B3** apresente melhor definição.

O algoritmo CLARA foi executado novamente, utilizando-se o pacote *cluster*, de Maechler *et al.* (2022), com os mesmos parâmetros da análise nacional: 100 amostras de 1.000 observações. A Figura 4.12 exibe os gráficos de dispersão.

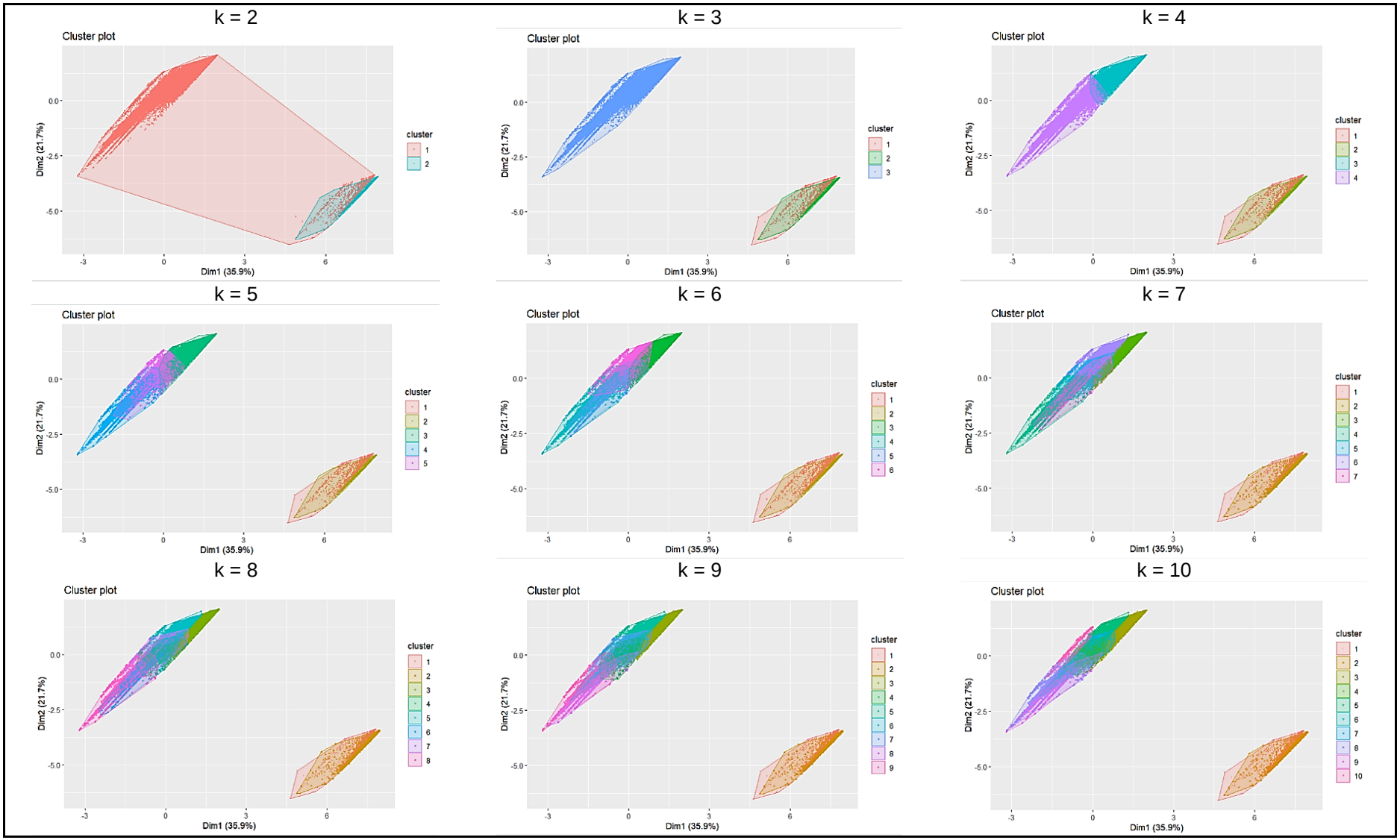


Figura 4.12 - Gráficos de dispersão com a indicação dos *clusters* para a base B3

Pode-se perceber graficamente a separação de duas estruturas: um agrupamento no canto esquerdo superior, representa fretes de carga completa, valores de frete e peso de carga altos; e outro, no canto inferior direito, caracterizando fretes de menores quantidades de veículos, peso de carga e valores de frete. Já as medidas de Silhueta para a base B3 são mostradas na Figura 4.13.

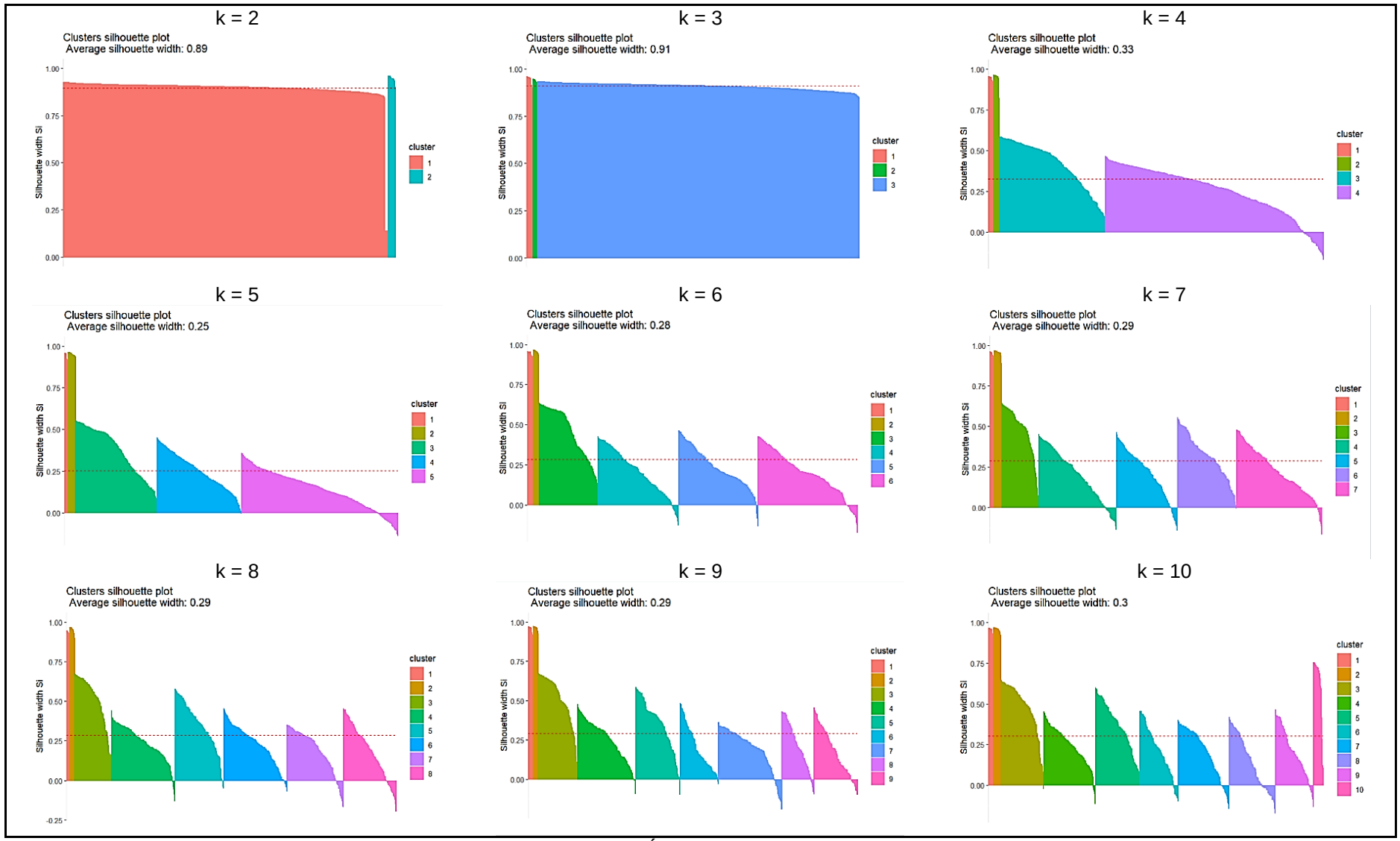


Figura 4.13 - Índice Silhueta da base B3

Já a Figura 4.14 apresenta os índices Silhueta médios para cada valor de *k* para a base **B3**.

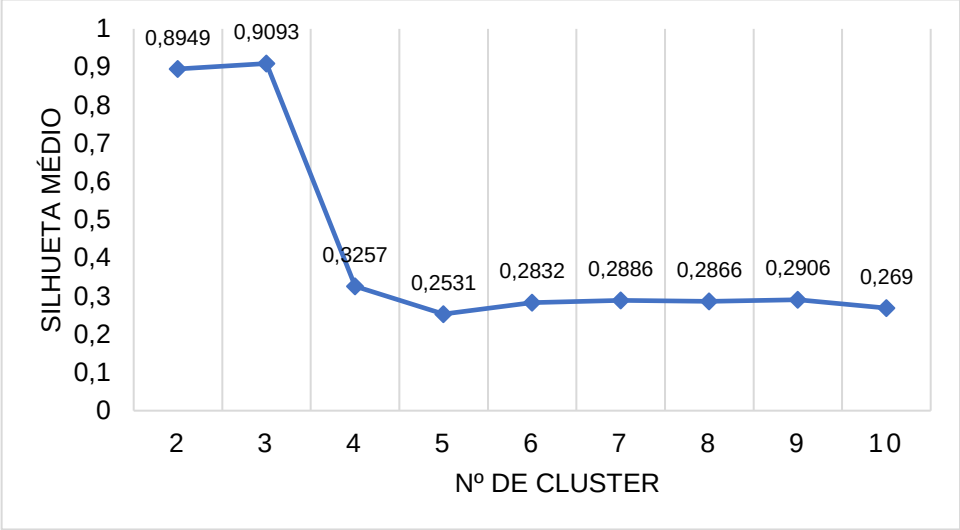


Figura 4.14 - Índices Silhueta médios para cada valor de *k* para a base B3

O algoritmo produziu melhor resultado para o valor de *k* = 3, observando que o índice é próximo de 1, indicando boa separação e compacidade. Se a empresa optar por esse processo de clusterização, obtém, portanto, os seguintes medoides, indicados na Tabela 4.8.

Tabela 4.8 - Características dos medoides para k = 3 da base B3

Medoide	Tipo de frete	Distância (km)	Peso (kg)	Qte. de veículos	Valor do frete (R$)	Tamanho do *cluster*
1	Agregação	37	3.500	1	400	4.968
2	Fracionado	169	700	1	300	6.374
3	Completo	334	32.000	3	1.504	367.664

O grupo mais significativo é representado pelo medoide 3, com pouco mais de 350 mil entradas. Possui características de frete lotação, caracteriza viagens de pouco mais de cinco horas, demanda veículos de categoria pesada. É bem similar ao medoide 1, encontrado na modelagem nacional (Tabela 4.5). Já os Grupos 1 e 2 são bem semelhantes, levando em consideração que, na Figura 4.11, eles se encontrem sobrepostos. Eles tipificam as viagens de carga fracionada e agregação, apresentando peso leve, percorrendo curtas distâncias. O valor do frete também é pequeno, traduzindo possivelmente em viagens de última milha.

Vale ressaltar que a variabilidade que se constata na Figura 4.11 reforça a cautela que se deve ter ao generalizar as características do medoide ao *cluster*. Para a base **B4**, a Figura 4.15 expõe os seguintes gráficos de dispersão.

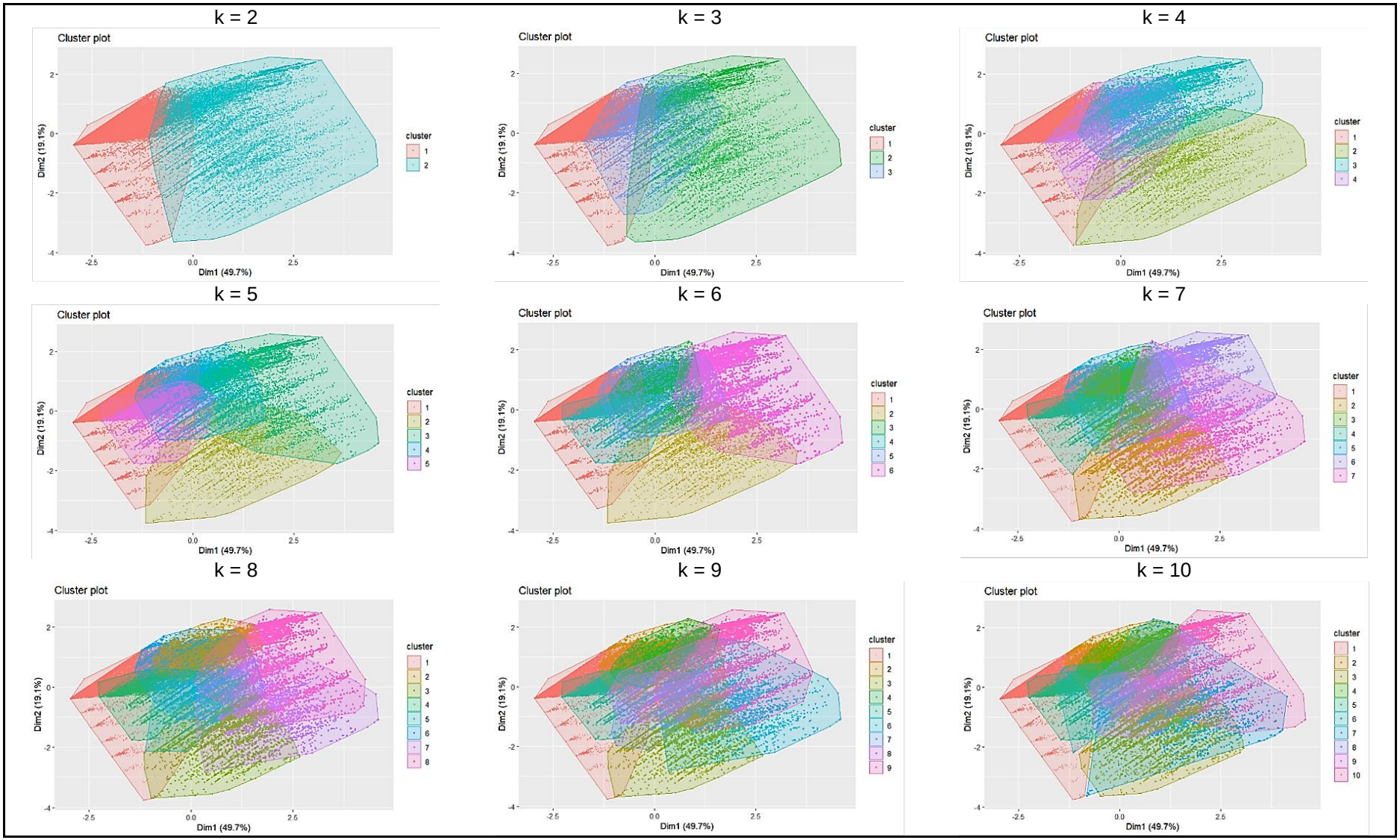


Figura 4.15 - Gráficos de dispersão com a indicação dos *clusters* para a base B4

Percebe-se que os dados estão bem esparsos, produzindo *clusters* sobrepostos. Levando a conclusão de que os dados, considerando as variáveis quantitativas, se distribuem uniformemente. Analisando o Índice Silhueta desse processo de clusterização, tem-se os resultados expostos na Figura 4.16.

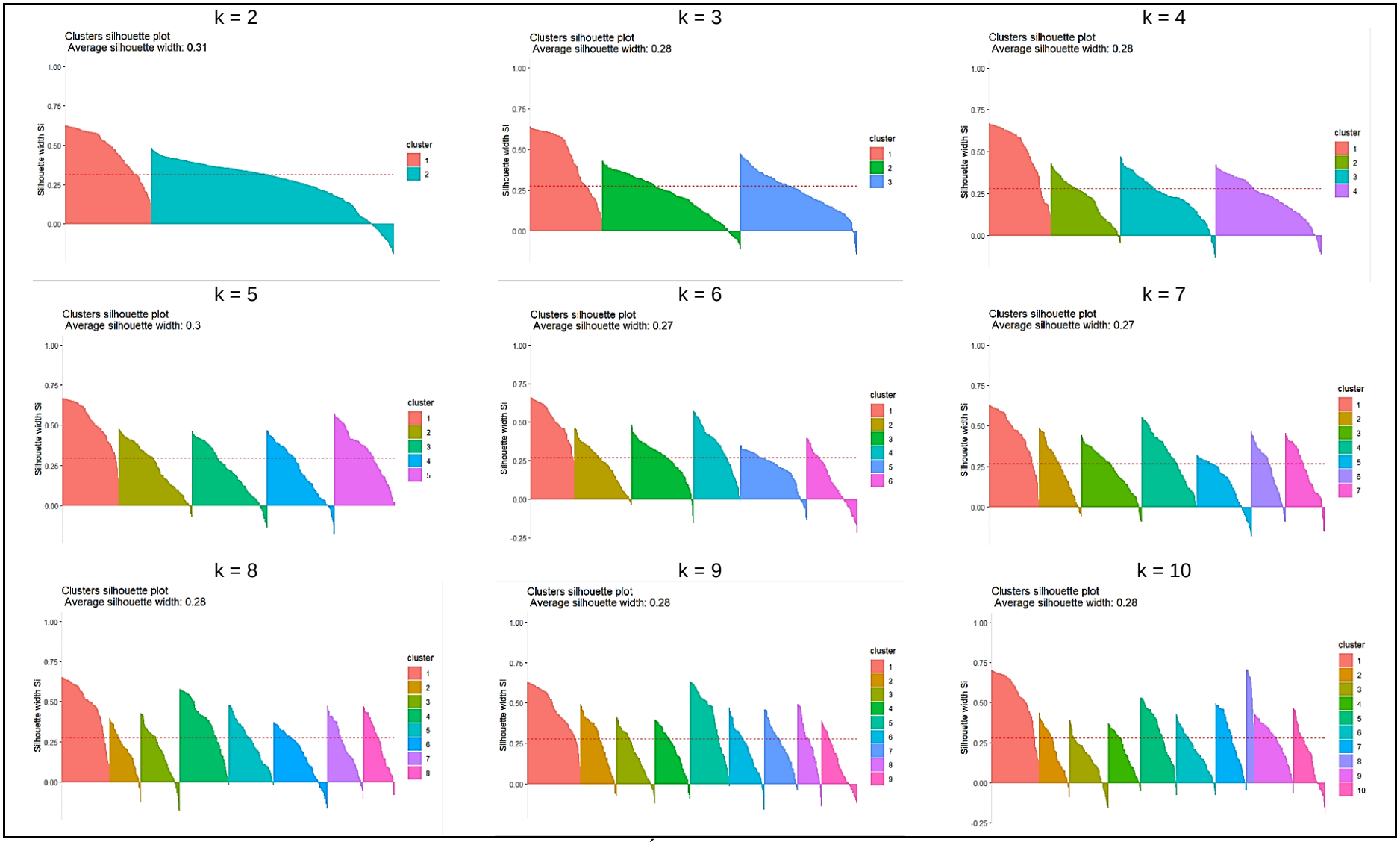


Figura 4.16 - Índice Silhueta da base B4

Já a Figura 4.17 apresenta os índices Silhueta médios para cada valor de *k* para a base **B4**.

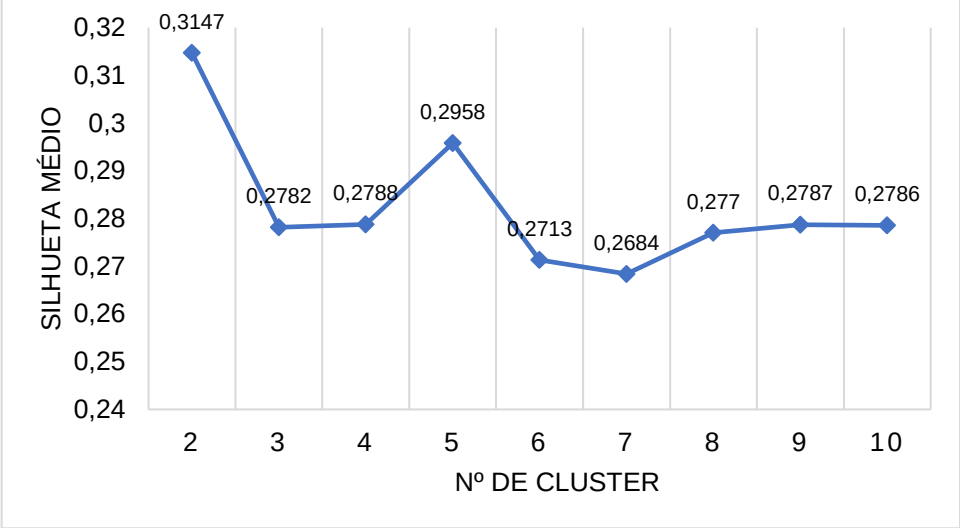


Figura 4.17 - Índices Silhueta médios para cada valor de k da base B4

Melhor resultado de clusterização é apresentado com o valor de *k* = 2, pois é o que mais se aproxima de 1. Se a empresa seguir os resultados desse processo, terá os seguintes medoides, descritos na Tabela 4.9.

Tabela 4.9 - Características dos medoides para k = 2 da base B4

Medoide	Distância (km)	Peso (kg)	Qte. de veículos	Valor do frete (R$)	Tamanho do *cluster*
1	114	15.000	1	470	107.191
2	384	32.000	3	1.824	271.815

O algoritmo divide a base em dois agrupamentos, em que um dos *cluster* contém mais que a metade dos dados. O agrupamento 1 estão inseridas as entradas de fretes que demandam caminhões de carga pesada (6 eixos); que percorrem médias distâncias, necessitando de três caminhoneiros para executar os trajetos. Já o grupo de menor quantidade, percorre distâncias menores, requerem veículos de médio porte e tem o valor de frete menor.

5 DISCUSSÃO DOS RESULTADOS

Este capítulo tem como objetivo trazer as respostas às questões levantadas nesta dissertação e estabelecer um paralelo com os resultados encontrados na literatura científica e profissional. Esta pesquisa buscou identificar padrões na postagem de carga em um *marketplace* logístico eletrônico, por meio da clusterização, a fim de desenvolver estratégias específicas para cada agrupamento. É importante que a empresa considere quais modelagens utilizar, sendo que a recomendação é pelas quais produziram melhores resultados, no caso, a modelagem da base **B2** para o nível nacional e da base **B3** para o estado de São Paulo, conforme descrito no capítulo anterior.

Tendo em vista os atributos dos medoides da clusterização da base **B2**, o algoritmo CLARA descreve que a maioria das postagens de carga compartilham características dos fretes que saem de São Paulo e chegam neste mesmo estado, percorrendo médias distâncias, isto é, viagens que duram cerca de nove horas; demandam veículos de carga pesada (seis eixos), e apresentam valores de frete altos, podendo-se associar ao peso do material a ser transportado, pois cargas densas exigem mais combustíveis e veículos maiores e mais potentes, que elevam as despesas. Esse resultado confirma o que a FreteBras (2022) afirma em seu 6º Relatório. A entidade reforça a representatividade do estado de São Paulo na postagem de cargas no contexto nacional. Em sua plataforma, no ano de 2021, o estado validou cerca de 23% dos 8 milhões de entradas. Portanto, pode-se concluir que a forte presença do estado de São Paulo nas postagens de carga não está associada a uma mera questão de atuação da empresa objeto de estudo desta pesquisa.

Os resultados da clusterização identifica que a lotação é a tipologia mais frequente no âmbito nacional. A Confederação Nacional do Transporte (2022) (CNT) constata o mesmo em sua Pesquisa Anual de Perfil Empresarial do Transporte de Cargas Rodoviário. A organização afirma que a maioria dos fretes no Brasil é de carga completa (57%). É plausível associar a maior frequência dessa tipologia aos benefícios que ela apresenta, como declaram a Confederação Nacional do Transporte (2022) e Golsby e Eckert (2003). São eles: a eficiência em termos de utilização de espaço e recursos, reduzindo os custos operacionais, bem como de combustível, pedágios e outros gastos, que são diluídos por unidade de carga transportada. Eles apontam também como benefícios o menor tempo de trânsito, já que não são necessárias paradas ao longo do percurso, facilitando o seu rastreamento. O processo de contratação do frete é mais simplificado, visto que as cargas são mais homogêneas, simplificando a contratação de seguros de carga, por exemplo. Essa característica de tornar os procedimentos menos

complexos é o que torna o modelo de negócio dos *marketplaces* logísticos atraente e é o que ele se propõe a fazer. Outra associação pode ser feita com o tipo de carga transportada no Brasil. O país apresenta um forte atividade agrícola e industrial, e esses produtos demandam veículos de grandes proporções por sua alta densidade.

Com relação à clusterização na base **B3**, a expectativa era que São Paulo apresentasse mais fretes do tipo fracionada em razão da intensa atividade do seu mercado consumidor de produtos e serviços. Contudo, os resultados apontam a maioria de fretes de carga completa e é verossímil relacionar com a grande concentração de indústrias e empresas em seu território, gerando a necessidade de transporte de cargas em larga escala. Assim, crê-se que as empresas que operam em São Paulo, especialmente aquelas com grandes volumes de carga, buscam constantemente formas de aumentar a eficiência e reduzir os custos logísticos, pela estratégia de fretes de carga completa. Também é importante vincular esse expressivo volume de cargas à atuação do estado no agronegócio. O Relatório da FreteBras (2022) confirma que 15% de fretes do agro e quase 30% de fretes da indústria vêm desse estado.

No contexto nacional, levando em consideração as características do *cluster* mais significativo, a provedora de tecnologia do *marketplace* logístico pode oferecer às transportadoras que realizam um alto volume de viagens descontos progressivos nos serviços adicionais por peso da cargas. Isso pode incentivar empresas com demanda regular a utilizar o *marketplace* de forma mais frequente. Seria interessante a possibilidade de contratação de serviços prioritários, como a divulgação diferenciada de anúncios patrocinados. Deve haver o compartilhamento de depoimento de transportadoras e motoristas satisfeitos que utilizaram o *marketplace* com êxito, pois, segundo Collignon *et al.* (2020), os *marketplaces* logísticos podem promover a confiança na plataforma se estabelecerem um conjunto de padrões profissionais mínimos e se oferecerem avaliações e análises do desempenho de seus participantes, auxiliando os usuários a determinar o nível de competência.

Outra estratégia pode ser o anúncio segmentado por parte da provedora de tecnologia do *marketplace* logístico, isto é, a divulgação da plataforma entre potenciais clientes com interesse específico em serviços de transporte de cargas pesadas e em distâncias de 500 km. Outro ponto é que para o tipo de frete identificado pelo medoide 1 são necessários caminhões de carga pesada, que exigem investimentos mais altos por parte dos caminhoneiros, portanto, seria relevante oferecer benefícios que incentivem a compra de veículos dessa capacidade, como parceria com empresas de crédito e serviços de manutenção especializada. Cabe dizer que as provedoras de tecnologia de *marketplaces* logísticos devem desenvolver um bom relacionamento com seus usuários, pois, como Nandiraju e Regan (2008) declaram, muitos

modelos desse negócio fracassaram por deixar se atentar a uma das partes (transportadoras e caminhoneiros). Por fim, é vantajoso que a empresa promova e participe de eventos e feiras sobre logística e transporte de carga pesada. Essas atividades podem aprimorar a visibilidade do negócio e gerar *networking* com potenciais clientes e parceiros.

O outro agrupamento do processo de clusterização da base **B2** representa os fretes no estado do Acre. A provedora de tecnologia do *marketplace* logístico pode simplesmente ignorar esse perfil ou encontrar nele uma oportunidade interessante de negócio. A empresa pode aumentar a sua atuação no estado, identificar as transportadoras que postam as suas cargas com mais frequência e dar a elas o suporte nessa região do Brasil em que os fretes rodoviários não são tão recorrentes. Da mesma forma, o *marketplace* logísticos pode utilizar o depoimento dessas transportadoras que utilizaram os serviços de intermediação com sucesso, identificando as razões que levaram a escolha pela plataforma como forma de atração de outros clientes. A empresa também pode criar conteúdo com informações sobre os desafios logísticos específicos do estado e soluções sob medida para superá-los. Todavia, pode ser também que os elementos desse *cluster* sejam entradas anormais como equívocos na usabilidade do aplicativo/interface por parte das transportadoras, cabendo mais investigação.

No que tange à clusterização das postagens de carga do estado de São Paulo, seria proveitoso o estabelecimento também de parcerias com empresas locais em São Paulo, como indústrias, distribuidoras e comércios, para promover o uso do *marketplace* logístico entre os seus parceiros comerciais. A Confederação Nacional do Transporte (2022) afirma que a maioria das empresas de serviço de transporte sempre contratam caminhoneiros autônomos para o atendimento de demandas excepcionais, portanto, também deve-se utilizar as mídias sociais para campanhas que possam alcançar potenciais caminhoneiros parceiros dentro do estado de São Paulo.

Percebe-se que as tipologias agregação e fracionada não são tão comuns no estado (medoide 1 e 2). A empresa pode também identificar esse fenômeno como uma oportunidade de negócio. Para esse perfil, conforme a Tabela 4.8, as necessidades de transporte são menores e as distâncias também. A empresa provedora pode destacar ao público a diversidade de cargas disponíveis, que podem não só atender aos detentores de veículos de grande porte, como também os de pequeno porte.

A quantificação do impacto dessas recomendações se torna um processo complexo, visto que os resultados dependem de uma série de fatores econômicos e gerenciais da empresa em questão, que se levantados fugiriam do escopo do trabalho. Além disso, alguns dos pontos levantados podem levar um longo tempo para surtirem efeito, bem como serem influenciados

por fatores externos, que estão fora do controle da organização. Haja vista dos desafios citados, uma abordagem que pode fazer sentido é categorizar as recomendações por sua facilidade de implementação, por seu público-alvo e pela sua abrangência geográfica. Dessa forma seria possível priorizá-las, mesmo que cabendo um aprofundamento posterior de cada medida.

Quadro 5.1 - Características das recomendações para o marketplace logístico

Recomendação	**Facilidade de implementação**	**Público-alvo**	**Abrangência geográfica**
1) Descontos progressivos nos serviços adicionais	Média	Transportadoras	Nacional
2) Anúncio pago de cargas	Fácil	Caminhoneiros	Nacional
3) Compartilhamento de depoimentos do uso da plataforma	Fácil	Transportadoras e caminhoneiros	Nacional
4) Eventos e feiras sobre logística de cargas pesadas	Fácil	Transportadoras e caminhoneiros	Nacional
5) Divulgação da plataforma para um perfil de potenciais usuários	Fácil	Transportadoras e caminhoneiros	Nacional
6) Parceria com concessionárias de caminhões, ofertadoras de créditos e empresas especializadas em manutenção de veículos de carga pesada	Difícil	Caminhoneiros	Nacional
7) Suporte às transportadoras e caminhoneiros do Acre	Fácil	Transportadoras e caminhoneiros	Estadual
8) Divulgação da plataforma para potenciais usuários interessados em cargas fracionadas e agregação em São Paulo	Fácil	Caminhoneiros	Estadual

O critério usado para definir se uma medida é fácil ou difícil de ser implementada envolveu a questão de a empresa já possuir ou não as informações para a identificação do público-alvo; se acarretaria a exigência de uma grande quantidade de recursos financeiros, tecnológicos e humanos, bem como a aceitação por parte dos parceiros. Levando em consideração que as medidas que exigem menos esforços e que abrangem um número máximo de usuários são as mais desejáveis, a empresa poderia começar a implementação das medidas pelas recomendações 3, 4 e 5.

Verificou-se que a organização já tem posto em prática parte dessas recomendações, como o estabelecimento de parcerias com entidades que realizam a integração para a emissão do CIOT e do Vale-pedágio, ou ainda com gerenciadoras de riscos e *softwares* de gestão. Marasco (2004) reitera que essa estratégia de oferta de serviços de valor agregado é crucial para que os *marketplaces* logísticos possam aumentar a fidelidade dos participantes e moldar sua proposta de valor. Além disso, ao depender também do lucro proveniente dos serviços de valor

agregado, é mais provável que esses modelos de negócio alcancem rentabilidade suficiente para a sua sobrevivência no mercado.

Agora com relação ao processo de clusterização, os *clusters* obtidos da base **B2** ficaram sobrepostos. A princípio, pensa-se que, com o aumento do número de partições, os resultados seriam melhores. Entretanto, tanto Kanungo *et al.* (2002) quanto Alpaydin (2014) sugerem que definir muitos *clusters* pode piorar o desempenho do algoritmo, aumentando a complexidade computacional e aumento o risco de *overfitting*, que seria o ajuste excessivo dos algoritmos de *Machine Learning*.

Acredita-se também que as demais clusterizações (base **B1** e **B4**) não produziram resultados satisfatórios por seus pontos se distribuírem de forma uniforme. Apesar da Estatística de Hopkins confirmar a clusterabilidade, mas esse fenômeno se torna mais visível pelo *heatmap*. Esse problema pode ser uma questão de a quantidade de amostras não compreender a dimensão dos dados. Wright (2022) revela que alguns autores sugerem que o número de amostras na função *hopkins* seja maior que 10, para evitar problemas de amostragem insuficiente.

É plausível também levar em consideração a análise da correlação entre as variáveis antes da clusterização como um método para diminuir a dimensionalidade do conjunto de dados. A presença de variáveis altamente correlacionadas pode caracterizar informações semelhantes ou redundantes, isto é, as variáveis não irão contribuir muito para a formação de *clusters* distintos, levando a uma divisão artificial dos dados. Domashova e Zasypkina (2021) utilizam o *heatmap* para produzir uma matriz de correlação e assim eliminar 23 características/atributos que estavam fortemente correlacionados. Entretanto, Kassambara (2017) defende que a técnica PCA já executa a tarefa de redução de dimensionalidade, encontrando um conjunto menor de variáveis que capturem a maior parte da variação nos dados.

Domashova e Zasypkina (2021) também não removeram os *outliers* de sua base de dados para que eles pudessem ser identificados pelo algoritmo como usuários com o propósito malicioso de coletar informações comerciais sensíveis de outros. Percebe-se que a estratégia de manter com os *outliers* nas bases de dados poderia render outras interpretações úteis, uma vez que Saxena *et al.* (2017) apontam pouca sensibilidade do CLARA a eles.

Pode-se ponderar também o dimensionamento insuficiente da amplitude do número de *clusters.* Halkidi *et al.* (2002) apresentam, em seu trabalho, um exemplo em que a base de dados claramente estava dividida em três agrupamentos, mas que o algoritmo em questão concluiu que o melhor particionamento era em quatro. Eles afirmam que configurar um valor impróprio para os parâmetros da clusterização pode afetar a qualidade das decisões tomadas a partir dela.

Possivelmente, valores maiores de *k* na bases **B1**, **B2** e **B3** poderiam produzir *clusters* mais definidos.

Identifica-se as sobreposições das duas estruturas claramente avistadas como distintas na base **B2**. Saxena *et al.* (2017) indicam que, para os casos em que o algoritmo encontra dificuldades na identificação dos agrupamentos, é necessário a aplicação de métodos que reconheçam *clusters* de formas não elípticas, como o DBSCAN ou até mesmo ponderar a aplicação de abordagens que considerem uma probabilidade de os pontos pertencerem a grupos distintos, como o FCM. Xu e Tian (2015), na verdade, indicam que os métodos tradicionais, como o PAM, o *K-Means* e o CLARA, não são ideais para a clusterização de dados de alta dimensionalidade. Acredita-se que seja por isso que o CLARA tenha implementado tão bem a base B3, com 7 variáveis e não a base B2, com 61. Esses mesmos autores recomendam o emprego dos algoritmos hierárquicos, melhores em termos de complexidade de tempo, escalabilidade e conjuntos de dados com muitas entradas e muitas variáveis.

Os melhores processos de clusterização conseguiram captar em partes as características das bases de dados apresentados na análise exploratória, isto é, média de peso da carga e quantidade de veículos. Ainda assim a clusterização apresentou mais vantagens frente a esse tipo de análise, visto que conseguiu revelar grupos de dados que podem ser uma oportunidade de crescimento da plataforma e também se mostrou útil como uma iniciação para técnicas mais complexas. Siqueira Junior (2021), em seu trabalho, consegue, de forma semelhante, por meio da clusterização, encontrar, em uma base de fornecedores, os que apresentam alto valor de inconformidades e conclui que, do ponto de vista prático, a ferramenta pode auxiliar gestores a decidir qual grupo investir recursos e melhorar os relacionamentos empresariais.

6 CONCLUSÕES

Esta dissertação teve por objetivo reconhecer padrões no procedimento de postagem de carga em um *marketplace* logístico, através da técnica de *Machine Learning* conhecida por clusterização. Foram avaliados os dados de frete da plataforma de 2019 a 2021.

Após a análise de tendência de clusterização do conjunto de dados, que confirmou a sua capacidade de formar agrupamentos, foi possível examinar essas bases, associando diferentes variáveis categóricas e quantitativas e contextos geográficos, através do algoritmo CLARA, concebida para clusterização de largas aplicações. Por meio do índice de validação Silhueta, verificou-se que a clusterização na base nacional, com a indicação do tipo de frete e da rota, para o número de *clusters* 2; e a clusterização das postagens de carga do estado de São Paulo, com a indicação do tipo de frete, para o número de *clusters* 3 resultaram nos melhores índices.

O grupo mais representativo, no âmbito nacional, pôde ser representado por fretes dentro do estado de São Paulo, que apresentavam carga completa, percorrendo distâncias de cerca de 500 km e demandando veículos de categoria pesada para o transporte. Já no contexto São Paulo, a partição mais expressiva foi a de fretes lotação, com viagens de pouco mais de cinco horas e que exigiam também veículos de carga pesada. Assim, o objetivo de analisar quais localidades e tipologia de frete se destacam nas postagens de carga na plataforma eletrônica foi atingido. Cabe pontuar que o trabalho levantou a associação da maior frequência do frete tipo lotação no *marketplace* logístico pelos benefícios que esta modalidade apresenta, a saber, a eficiência em termos de utilização de espaço e recursos e a diluição das despesas por unidade de carga transportada.

O objetivo de traçar estratégias para a empresa provedora de tecnologia de como tratar os grupos resultantes do processo de clusterização, tomando as características apontadas pelos medoides, também foi finalizado. As principais estratégias identificadas para a base nacional consistem no oferecimento às transportadoras que realizam um alto volume de viagens descontos progressivos nos serviços adicionais e a divulgação segmentada do *marketplace* logístico entre os potenciais clientes com interesse específico em serviços de transporte de cargas pesadas e em distâncias de 500 km. Pôde-se identificar uma interessante oportunidade de negócio no Acre, em que a empresa poderia aumentar a sua atuação no estado, identificar as transportadoras que postam as suas cargas com mais frequência e dar a elas o suporte nessa região do Brasil em que os fretes rodoviários não são tão recorrentes; bem como incentivar o uso da plataforma no estado de São Paulo para a postagens de cargas fracionadas, apresentando as vantagens também para os que detêm veículos de pequeno porte. Considerando que as

medidas de menor complexidade e que abrangem um maior número de usuários são as mais desejáveis, a provedora de tecnologia do *marketplace* logístico pode inicialmente promover o compartilhamento de depoimentos do uso da plataforma, bem como realizar eventos e feiras sobre logística de cargas pesadas e divulgar a plataforma para potenciais usuários, a saber os que se interessam por frete de carga pesada e médias distâncias.

Quanto ao objetivo de apresentar o funcionamento e o comportamento do algoritmo de clusterização selecionado, conclui-se que o CLARA trouxe resultados satisfatórios, diminuindo a complexidade computacional de uma base de mais de três milhões de entradas. Entretanto, houve sobreposições de estruturas claramente avistadas como distintas nos gráficos de dispersão. Verificou-se que alguns autores não indicam o algoritmo em questão para conjuntos de dados de alta dimensionalidade. Acredita-se que essa foi a razão de o CLARA ter executado uma divisão mais clara na base com sete variáveis do que a com 61.

Pontuou-se também os pontos positivos e negativos do emprego da clusterização ante a simples análise exploratória dos dados. Concluiu-se que o procedimento de clusterização apesar de captar em partes as características das bases de dados apresentados na análise exploratória, ainda apresenta mais vantagens frente a esse tipo de análise, visto que conseguiu revelar grupos de dados que podem ser uma oportunidade de crescimento da plataforma e também se mostrou útil como os primeiros passos para a implementação de técnicas mais complexas.

Por fim, o objetivo de comparar os resultados com as colocações dos autores em *marketplaces* logísticos também foi satisfeito, tendo em vista que foi possível associar a maior frequência de fretes tipo lotação na plataforma à simplificação dos processos de contratação que o modelo de negócio em questão se propõem a fazer. Da mesma forma, foi viável traçar algumas das recomendações tendo por base as boas práticas em um *marketplace* logístico que a literatura aconselha.

6.1 Limitações e sugestões para trabalhos futuros

Apesar dos objetivos terem sido alcançados, é importante salientar as dificuldades computacionais enfrentadas. Na tentativa de determinar o número ótimo de *clusters*, pela função *fviz_nbclust* do pacote *factoextra*, e também da construção do *heatmap* com a totalidade dos dados, o console do RStudio indicou que não era possível a alocação de 60 mil *gigabytes* na memória. Isso se deve ao fato de a função calcular uma matriz de dissimilaridade de 1.000.000 x 1.000.000.

Portanto, é interessante estudar mais a fundo os algoritmos de abordagem por amostragem, que tem se mostrado tendência, visto que aumentam a eficiência computacional quando as análises são executadas em computadores convencionais.

Sugere-se a condução de análises do conjunto de dados do *marketplace* logístico com outros algoritmos de clusterização clássicos, como o *K-Means*, AGNES, DIANA, e com aqueles apontados pela literatura como mais adequados para reconhecimento de agrupamentos de formas não elípticas, como o DBSCAN. Também, aconselha-se testar outros índices para a validação dos resultados, como o Índice Dunn.

Propõe-se a integração dos dados com a base do ano de 2022, bem como reduzir o nível de avaliação das rotas ao municipal, tendo, todavia, o cuidado de tomar somente as cidades de maior relevância, para não aumentar a complexidade computacional.

Por fim, recomenda-se a verificação da estabilidade do algoritmo CLARA, configurando outras sementes e a análise da base de dados de outras empresas de *marketplace* logístico no Brasil, como forma de comparação dos resultados aqui apresentados.

REFERÊNCIAS

ABOUBI, Y.; DRIAS, H.; KAMEL, N. BAT-CLARA: BAT-inspired algorithm for Clustering LARge Applications. **IFAC-PapersOnLine**, v. 49, n. 12, p. 243–248, 2016. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.ifacol.2016.07.607>. .

AGARWAL, R.; DHAR, V. Big data, data science, and analytics: The opportunity and challenge for IS research. **Information Systems Research**, v. 25, n. 3, p. 443–448, 2014.

AGGARWAL, C. C. **Data Mining**. Cham: Springer International Publishing, 2015.

ALPAYDIN, E. **Introduction to Machine Learning Second Edition**. 2014.

ALT, R.; KLEIN, S. Learning from failure: The myths and magic of electronic transportation markets. **Proceedings of the Hawaii International Conference on System Sciences**, v. 4, n. c, p. 102–110, 1998.

AMERICAN TRUCKING ASSOCIATIONS. Economics and industry data. Disponível em: <https://www.trucking.org/economics-and-industry-data>. Acesso em: 7/6/2022.

APPOLINÁRIO, F. **Metodologia da ciência - filosofia e prática da pesquisa**. 2º ed. 2013.

ARAVAZHI IRISSAPPANE, A.; ZHANG, J. Filtering unfair ratings from dishonest advisors in multi-criteria e-markets: a biclustering-based approach. **Autonomous Agents and Multi-Agent Systems**, v. 31, n. 1, p. 36–65, 2017. Springer US.

ARMANO, G.; FARMANI, M. R. Multiobjective clustering analysis using particle swarm optimization. , v. 55, p. 184–193, 2016. Elsevier Ltd.

ASTILL, J.; FRASER, E.; DARA, R.; SHARIF, S. Detecting and predicting emerging disease in poultry with the implementation of new technologies and big data: A focus on avian influenza virus. **Frontiers in Veterinary Science**, v. 5, n. OCT, p. 1–12, 2018.

BAHMANI, B.; MOSELEY, B.; VATTANI, A.; KUMAR, R.; VASSILVITSKII, S. Scalable κ-means++. **Proceedings of the VLDB Endowment**, v. 5, n. 7, p. 622–633, 2012.

BAKOS, J. Y. A strategic analysis of electronic marketplaces. **MIS Quarterly: Management Information Systems**, v. 15, n. 3, p. 295–310, 1991.

BASSO, D. E. **Big data**. Curitiba, 2020.

BÉJAR ALONSO, J. K-means vs Mini Batch K-means: a comparison. **Departament de Llenguatges i Sistemes Informàtics**, p. 1–12, 2013. Disponível em: <http://hdl.handle.net/2117/23414>. .

BELCASTRO, L.; BRANDA, F.; CANTINI, R.; et al. Analyzing voter behavior on social media during the 2020 US presidential election campaign. **Social Network Analysis and Mining**, 2022. Springer Vienna. Disponível em: <https://doi.org/10.1007/s13278-022-00913-9>. .

BERALDI, P.; DE MAIO, A.; LAGANÀ, D.; VIOLI, A. A pick-up and delivery problem for logistics e-marketplace services. **Optimization Letters**, v. 15, n. 5, p. 1565–1577, 2021. Springer Berlin Heidelberg. Disponível em: <https://doi.org/10.1007/s11590-019-01472-3>. .

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations and Production Management**, v. 22, n. 2, p. 241–264, 2002.

BOUVEYRON, C.; BRUNET-SAUMARD, C. Model-based clustering of high-dimensional data: A review. **Computational Statistics and Data Analysis**, v. 71, p. 52–78, 2014. Elsevier

B.V. Disponível em: <http://dx.doi.org/10.1016/j.csda.2012.12.008>. .

BRUN, M.; SIMA, C.; HUA, J.; et al. Model-based evaluation of clustering validation measures. **Pattern Recognition**, v. 40, n. 3, p. 807–824, 2007.

CAPÓ, M.; PÉREZ, A.; LOZANO, J. A. An efficient approximation to the K-means clustering for massive data. **Knowledge-Based Systems**, v. 117, p. 56–69, 2017.

CAUCHICK MIGUEL, P. A.; FLEURY, A.; MELLO, C. H. P.; et al. **Metodologia de pesquisa em Engenharia de Produção e Gestão de Operações**. Rio de Janeiro, 2010.

CAVIQUE, L. Big Data e Data Science. **Boletim APDIO**, v. 51, p. 11–14, 2014. Disponível em: <https://repositorioaberto.uab.pt/handle/10400.2/3918>. .

CHANG, H. H.; WONG, K. H. Adoption of e-procurement and participation of e-marketplace on firm performance: Trust as a moderator. **Information and Management**, v. 47, n. 5–6, p. 262–270, 2010. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.im.2010.05.002>. .

CHAPMAN, P.; CLINTON, J.; KERBER, R.; et al. CRISP-DM 1.0: Step-by-step Data Mining Guide. **SPSS inc**, v. 78, p. 1–78, 2000. Disponível em: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72%0Ahttp://www.crisp-dm.org/CRISPWP-0800.pdf>. .

CHARALAMPOPOULOS, I. The R language as a tool for biometeorological research. **Atmosphere**, v. 11, n. 7, 2020.

CHEN, Y. L.; HU, H. L. An overlapping cluster algorithm to provide non-exhaustive clustering. **European Journal of Operational Research**, v. 173, n. 3, p. 762–780, 2006.

CHENYAN, Z.; HONGYAN, Y.; ZHONGYING, L. Logistics collaboration supported by electronic logistics marketplaces. **2008 IEEE Symposium on Advanced Management of Information for Globalized Enterprises, AMIGE 2008 - Proceedings**, p. 46–50, 2009.

CHINA. China's road freight volume up 14.2% in 2021. Disponível em: <http://english.www.gov.cn/archive/statistics/202201/22/content_WS61eba411c6d09c94e48a41ac.html>. Acesso em: 13/7/2022.

CHOUDHARY, B.; SAXENA, P. V. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Fuzzy C-Mean Technique for Accessing Large Database of Banking Sector. , v. 11, n. 4, p. 263–271, 2023.

CODELESS PLATFORMS. What is ETL? – A Comprehensive Guide to Extract, Transform, and Load. Disponível em: <https://www.codelessplatforms.com/blog/what-is-etl/#:~:text=using stored procedure.-,ETL Process Example%3A Extracting%2C Transforming%2C and Loading Data from,data from various store locations.>. .

COLLIGNON, S.; COOK, D. F.; LI, Y. Motor carrier spot market: trust-building in public e-marketplaces. **International Journal of Physical Distribution and Logistics Management**, v. 50, n. 2, p. 191–214, 2020.

COMUZZI, M.; PATEL, A. How organisations leverage Big Data: A maturity model. **Industrial Management and Data Systems**, v. 116, p. 1468–1492, 2016.

CONFEDERAÇÃO NACIONAL DO TRANSPORTE. **Pesquisa CNT Perfil Empresarial 2021: transporte rodoviário de cargas**. 2022.

DATANOVIA. Data Preparation and R Packages for Cluster Analysis. Disponível em:

<https://www.datanovia.com/en/lessons/data-preparation-and-r-packages-for-cluster-analysis/>. .

DENG, Z.; ZHU, X.; CHENG, D.; ZONG, M.; ZHANG, S. Efficient kNN classification algorithm for big data. **Neurocomputing**, v. 195, p. 143–148, 2016.

DOMASHOVA, J.; ZASYPKINA, A. Detection of non-typical users of the electronic marketplace "freight transportation" to prevent the competitive intelligence. **Procedia Computer Science**, v. 190, n. 2020, p. 210–216, 2021. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.procs.2021.06.026>. .

DZIKRULLAH AHMAD, A. D. **Bunga Rampai Skenario Covid-19 Prodi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia**. 2021.

EBERENDU, A. C. Unstructured Data: an overview of the data of Big Data. **International Journal of Computer Trends and Technology**, v. 38, n. 1, p. 46–50, 2016.

ELLEGAARD, O.; WALLIN, J. A. The bibliometric analysis of scholarly production: How great is the impact? **Scientometrics**, v. 105, n. 3, p. 1809–1831, 2015. Springer Netherlands.

ELSEVIER. Analyze search results. Disponível em: <https://www.scopus.com/term/analyzer.uri?sid=2da64b0f6140d8db458098cac1d3e678&origin=resultslist&src=s&s=%28TITLE-ABS-KEY%28%22logistics+marketplace%22%29+OR+TITLE-ABS-KEY%28%22electronic+logistics+marketplace%22%29+OR+TITLE-ABS-KEY%28%22transportation+el>. Acesso em: 13/7/2022.

ENDEMANN, P.; FRANKFURTRHEINMAIN, R.; TRACKSDORF, K.; et al. CODE 24 Online Rail Freight Exchange – Needs of Potential Users. **European Transport Conference**, v. 2010, n. Wittenbrink 2011, p. 1–14, 2012.

FÖHRING, R.; ZELEWSKI, S. AFEX: An autonomous freight exchange concept. **Transportation Research Procedia**, v. 10, n. July, p. 644–651, 2015. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.trpro.2015.09.018>. .

FOP, M.; MURPHY, T. B. Variable selection methods for model-based clustering. , v. 12, p. 18–65, 2018.

FORBES. Logtech Sotran recebe aporte de R$100 milhões do Arlon Group. Disponível em: <https://forbes.com.br/forbes-money/2021/02/logtech-sotran-recebe-aporte-de-r-100-mi-do-arlon-group/>. Acesso em: 21/5/2021.

FRETEBRAS. **6ª Edição Relatório FreteBras: uma visão de 2021**. 2022.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137–144, 2015. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>. .

GHALEHKHONDABI, I.; AHMADI, E.; MAIHAMI, R. An overview of big data analytics application in supply chain management published in 2010-2019. **Production**, v. 30, 2020.

GHOSH, S.; KUMAR, S. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. **International Journal of Advanced Computer Science and Applications**, v. 4, n. 4, p. 35–39, 2013.

GIORGI, F. M.; CERAOLO, C.; MERCATELLI, D. The R Language: An Engine for Bioinformatics and Data Science. **Life**, v. 12, n. 5, p. 1–25, 2022.

GOKSULUK, D.; KORKMAZ, S.; ZARARSIZ, G.; KARAAGAOGLU, A. E. EasyROC: An

interactive web-tool for roc curve analysis using r language environment. **R Journal**, v. 8, n. 2, p. 213–230, 2016.

GOLSBY, T. J.; ECKERT, J. A. Electronic transportation marketplaces: A transaction cost perspective. **Industrial Marketing Management**, v. 32, n. 3, p. 187–198, 2003.

GOSTKOWSKI, M.; ROKICKI, T.; OCHNIO, L.; et al. Clustering analysis of energy consumption in the countries of the visegrad group. **Energies**, v. 14, n. 18, p. 1–24, 2021.

GRIEGER, M. Electronic marketplaces: A literature review and a call for supply chain management research. **European Journal of Operational Research**, v. 144, n. 2, p. 280–294, 2003.

GÜEMES-PEÑA, D.; LÓPEZ-NOZAL, C.; MARTICORENA-SÁNCHEZ, R.; MAUDES-RAEDO, J. Emerging topics in mining software repositories: Machine learning in software repositories and datasets. **Progress in Artificial Intelligence**, v. 7, n. 3, p. 237–247, 2018. Springer Berlin Heidelberg. Disponível em: <https://doi.org/10.1007/s13748-018-0147-7>. .

GÜNERI, Ö. İ.; DURMUŞ, B. Dependent Dummy Variable Models: An Application of Logit, Probit and Tobit Models on Survey Data. **International Journal of Computational and Experimental Science and ENgineering (IJCESEN)**, v. 6, n. 1, p. 63–74, 2020. Disponível em: <http://dergipark.org.tr/en/pub/ijcesen>. .

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part I. **SIGMOD Record**, v. 31, n. 2, p. 40–45, 2002.

HÄMÄLÄINEN, J.; JAUHIAINEN, S.; KÄRKKÄINEN, T. Comparison of internal clustering validation indices for prototype-based clustering. **Algorithms**, v. 10, n. 3, 2017.

HICKS, S. C.; LIU, R.; NI, Y.; PURDOM, E.; RISSO, D. Mbkmeans: Fast clustering for single cell data using mini-batch k-means. **PLoS Computational Biology**, v. 17, n. 1, 2021. Public Library of Science.

HOFMANN, E.; RUTSCHMANN, E. Big data analytics and demand forecasting in supply chains: a conceptual analysis. **International Journal of Logistics Management**, v. 29, n. 2, p. 739–766, 2018.

IBM. What is ETL (Extract, Transform, Load)? Disponível em: <https://www.ibm.com/cloud/learn/etl>. .

ILOS. Matriz de transportes do Brasil à espera dos investimentos. Disponível em: <https://www.ilos.com.br/web/tag/matriz-de-transportes/>. Acesso em: 13/7/2022.

ISHWARAPPA; ANURADHA, J. A brief introduction on big data 5Vs characteristics and hadoop technology. **Procedia Computer Science**, v. 48, n. C, p. 319–324, 2015.

JAIN, A.; BRUCKMANN, D.; VAN DER HEIJDEN, R. E. C. M.; MARCHAU, V. A. W. J. Towards rail-related multimodal freight exchange platforms: Exploring regulatory topics at EU level. **Competition and Regulation in Network Industries**, v. 20, n. 2, p. 138–163, 2019.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2009.09.011>. .

JANITA, M. S.; MIRANDA, F. J. The antecedents of client loyalty in business-to-business (B2B) electronic marketplaces. **Industrial Marketing Management**, v. 42, n. 5, p. 814–823, 2013.

KANUNGO, T.; MOUNT, D. M.; NETANYAHU, N. S.; et al. An efficient k-means clustering

algorithm: Analysis and implementation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 881–892, 2002.

KARAMIZADEH, S.; ABDULLAH, S. M.; MANAF, A. A.; ZAMANI, M.; HOOMAN, A. An Overview of Principal Component Analysis. , v. 2013, n. August, p. 173–175, 2013.

KASSAMBARA, A. Alboukadel Kassambara Practical Guide To Cluster Analysis in R. , 2015.

KASSAMBARA, A. Multivariate Analysis I: Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning. **Taylor & Francis Group**, p. 188, 2017.

KASSAMBARA, A.; MUNDT, F. Package "factoextra". , 2022. Disponível em: <https://github.com/kassambara/factoextra/issues>. .

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: An introduction to cluster analysis**. 1990.

KHAJEH, M.; TALEBPOUR, A.; DEVUNURI, S. An unsupervised learning framework for detecting adaptive cruise control operated vehicles in a vehicle trajectory data. **Expert Systems With Applications**, v. 208, n. July, p. 118060, 2022. Elsevier Ltd. Disponível em: <https://doi.org/10.1016/j.eswa.2022.118060>. .

KHANDAVILLI, P. Importance of ETL: 3 Critical Benefits and Top ETL Tools Table of Contents. Disponível em: <https://hevodata.com/learn/importance-of-etl/>. .

KOVÁCS, G. The structure, modules, services, and operational process of modern electronic freight and warehouse exchanges. **Periodica Polytechnica Transportation Engineering**, v. 37, n. 1–2, p. 33–38, 2009.

KOVÁCS, G. Possible methods of application of electronic freight and warehouse exchanges in solving the city logistics problems. **Periodica Polytechnica Transportation Engineering**, v. 38, n. 1, p. 25–28, 2010.

KUBOTA, L. C.; MILANI, D. N. No Title. , 2011.

LIAO, K.; LIU, G.; XIAO, L.; LIU, C. Knowledge-Based Systems A sample-based hierarchical adaptive K -means clustering method for large-scale video retrieval. **Knowledge-Based Systems**, v. 49, p. 123–133, 2013. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2013.05.003>. .

LIMA, C.; CALAZANS, J. Pegadas digitais: Big Data e informação estratégica sobre consumidor. Performances Interacionais e Mediações Sociotécnicas. **Anais...** , 2013.

LIU, Y.; LI, Z.; XIONG, H.; GAO, X.; WU, J. Understanding of internal clustering validation measures. **Proceedings - IEEE International Conference on Data Mining, ICDM**, p. 911–916, 2010.

LORENZINI, E. Innovation and e-commerce in clusters of small firms: The case of a regional e-marketplace. **Local Economy**, v. 29, n. 8, p. 771–794, 2014.

MAECHLER, M.; ROUSSEEUW, P. J.; STRUYF, A.; et al. Package "cluster". , 2022. Disponível em: <https://cran.r-project.org/web/packages/cluster/index.html>. .

MALLICK, P.; SARKAR, S.; MITRA, P. Decision recommendation system for transporters in an online freight exchange platform. **2017 9th International Conference on Communication Systems and Networks, COMSNETS 2017**, p. 448–453, 2017. IEEE.

MANTHOU, V.; MATOPOULOS, A.; VLACHOPOULOU, M. Internet-based applications in the agri-food supply chain: A survey on the Greek canning sector. **Journal of Food Engineering**, v. 70, n. 3, p. 447–454, 2005.

MARAPPAN, R.; BHASKARAN, S. A Linear-Time Algorithm to Find the Second Smallest Number. **Trends Journal of Sciences Research**, v. 1, n. 1, p. 8–11, 2022.

MARASCO, A. Business models of transportation electronic marketplaces: an empirical survey. **Pomorski Zbornik**, v. 42, n. 1, p. 77–92, 2004.

MARTINS, R. J. M. **A relevância das plataformas de carga online num contexto nacional rodoviário**, 2019. Instituto Politécnico de Setúbal.

MCNICHOLAS, P. D. Model-Based Clustering. **Journal of Classification**, v. 33, n. 3, p. 331–373, 2016.

MCPARLAND, D.; GORMLEY, I. C. Model based clustering for mixed data: clustMD. **Advances in Data Analysis and Classification**, v. 10, n. 2, p. 155–169, 2016.

MEIDA, A.; RINI, D. P.; SUKEMI. Pattern of e-marketplace customer shopping behavior using improved tabu search and FP-growth algorithm. **Indonesian Journal of Electrical Engineering and Informatics**, v. 7, n. 4, p. 772–778, 2019.

MILLER, J.; NIE, Y. Dynamic trucking equilibrium through a freight exchange. **Transportation Research Procedia**, v. 38, p. 320–340, 2018. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.trpro.2019.05.018>. .

MILLER, J.; NIE, Y. M.; LIU, X. Hyperpath Truck Routing in an Online Freight Exchange Platform. **Transportation Science**, v. 54, n. 6, p. 1676–1696, 2020.

MOHAMAD, I. BIN; USMAN, D. Standardization and its effects on K-means clustering algorithm. **Research Journal of Applied Sciences, Engineering and Technology**, v. 6, n. 17, p. 3299–3303, 2013.

MONGEON, P.; PAUL-HUS, A. The journal coverage of Web of Science and Scopus: a comparative analysis. **Scientometrics**, v. 106, n. 1, p. 213–228, 2016.

MOTLAGH, S. M. H.; SEPEHRI, M. M.; IGNATIUS, J.; MUSTAFA, A. Optimizing trade in transportation procurement: Is combinatorial double auction approach truly better? **International Journal of Innovative Computing, Information and Control**, v. 6, n. 6, p. 2537–2550, 2010.

MOULAVI, D.; JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J. Density-based clustering validation. **SIAM International Conference on Data Mining 2014, SDM 2014**, v. 2, n. i, p. 839–847, 2014.

NANDIRAJU, S.; REGAN, A. Freight Transportation Electronic Marketplaces: A Survey of the Industry and Exploration of Important Research Issues. **Transportation Research Record**, v. 15, n. 4, p. 250–260, 2008. Disponível em: <https://escholarship.org/uc/item/9fj2c4jw>. .

BEN NCIR, C. E.; HAMZA, A.; BOUAGUEL, W. Parallel and scalable Dunn Index for the validation of big data clusters. **Parallel Computing**, v. 102, n. February 2020, p. 102751, 2021. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.parco.2021.102751>. .

NGUYEN, H.; BUI, X.; TRAN, Q.; MAI, N. A new soft computing model for estimating and controlling blast-produced ground vibration based on Hierarchical K-means clustering and Cubist algorithms. **Applied Soft Computing Journal**, v. 77, p. 376–386, 2019. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.asoc.2019.01.042>. .

NGUYEN, Q.; RAYWARD-SMITH, V. J. CLAM: Clustering Large Applications Using Metaheuristics. **Journal of Mathematical Modelling and Algorithms**, v. 10, n. 1, p. 57–78, 2011.

NUNES, D. H. F. Um Estudo Sobre O Algoritmo K-Means. , p. 60, 2016.

OH, K. J.; JUNG, J. G.; JO, G. S. Discovering Frequent Patterns by Constructing Frequent Pattern Network over Data Streams in E-Marketplaces. **Wireless Personal Communications**, v. 79, n. 4, p. 2655–2670, 2014.

OLIVEIRA, C. C. P. Benchmarking de técnicas de Business Analytics em Big Data. , 2020.

ORDANINI, A.; MICELLI, S.; DI MARIA, E. Failure and success of B-to-B exchange business models: A contingent analysis of their performance. **European Management Journal**, v. 22, n. 3, p. 281–289, 2004.

PALACIOS, H. J. G.; TOLEDO, R. A. J.; PANTOJA, G. A. H.; NAVARRO, Á. A. M. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. **Advances in Science, Technology and Engineering Systems**, v. 2, n. 3, p. 598–604, 2017.

PANDATA. The complete guide to clustering in Python and R. Disponível em: <https://levelup.gitconnected.com/the-complete-guide-to-clustering-in-python-and-r-2584f80b6fc6>. .

PATNAIK, A. K.; BHUYAN, P. K.; KRISHNA RAO, K. V. Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets. **Alexandria Engineering Journal**, v. 55, n. 1, p. 407–418, 2016. Faculty of Engineering, Alexandria University. Disponível em: <http://dx.doi.org/10.1016/j.aej.2015.11.003>. .

PATRO, S. G. K.; SAHU, K. K. Normalization: A Preprocessing Stage. **Iarjset**, p. 20–22, 2015.

PAULSON, J.; ALLAIRE, J. J.; CHENG, J. Getting the Most Out of RStudio ™. , p. 2012, 2012.

PENG, K.; LEUNG, V. C. M.; HUANG, Q. Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data. **IEEE Access**, v. 6, n. February, p. 11897–11906, 2018.

PEREIRA, A.; DUARTE, D.; MEIRA, W.; GÓES, P. Selling practices in online fixed-price marketplaces. **Proceedings - 2009 9th Annual International Symposium on Applications and the Internet, SAINT 2009**, p. 71–77, 2009.

PEREZ, H.; TAH, J. H. M. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. **Mathematics**, v. 8, n. 5, 2020.

PEROTIN, J. M.; GIERSKI, F.; BOLKO, L.; et al. Cluster analysis unveils a severe persistent respiratory impairment phenotype 3 - months after severe COVID - 19. **Respiratory Research**, p. 1–10, 2022. BioMed Central. Disponível em: <https://doi.org/10.1186/s12931-022-02111-9>. .

PICÃO, B. J. DE S. **Digital business transformation in transport and logistics companies: a global freight forwarder case study**, 2017. Universidade do Porto. Disponível em: <https://repositorio-aberto.up.pt/handle/10216/106181%0Ahttps://repositorio-aberto.up.pt/bitstream/10216/106181/2/203281.pdf>. .

PRANCKUTĖ, R. Web of science (Wos) and scopus: The titans of bibliographic information in today's academic world. **Publications**, v. 9, n. 1, 2021.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What you need to know about Data Mining and Data-Analytic Thinking**. 1º ed. Sebastopol, 2013.

QADER, W. A.; AMEEN, M. M.; AHMED, B. I. Big data characteristics, architecture,

technologies and applications. **Journal of Computer Science**, v. 16, n. 6, p. 817–824, 2020.

QI, J.; YU, Y.; WANG, L.; LIU, J.; WANG, Y. An effective and efficient hierarchical K-means clustering algorithm. **International Journal of Distributed Sensor Networks**, v. 13, n. 8, p. 1–17, 2017.

RANA, K. TCS partners with TPT South Africa to develop logistics e-marketplace platform. Disponível em: <https://www.logisticsinsider.in/tcs-partners-with-tpt-south-africa-to-develop-logistics-e-marketplace-platform/>. Acesso em: 26/6/2022.

RIOS, A. **Exploring the use of freight exchange e-marketplaces in Sweden: the perspective of the transport service provider**, 2018. Lund University. Disponível em: <http://lup.lub.lu.se/student-papers/record/8951501>. .

RUIZ, L. G. B.; PEGALAJAR, M. C.; ARCUCCI, R.; MOLINA-SOLANA, M. A time-series clustering methodology for knowledge extraction in energy consumption data. **Expert Systems with Applications**, v. 160, p. 113731, 2020. Elsevier Ltd. Disponível em: <https://doi.org/10.1016/j.eswa.2020.113731>. .

SALLEH, N. A. M.; MUKHTAR, M.; ASHAARI, N. S. Logistic E-Marketplace For Agro-Based Industries In Malaysia. 2009 International Conference on Electrical Engineering and Informatics. **Anais...** . p.338–342, 2009. IEEE.

SAMPATH, K.; KOTI REDDY DANDA, S.; KUMAR, K.; et al. Spot collaborative shipping sans orchestrator using Blockchain. **Proceedings - 2020 IEEE International Conference on Blockchain, Blockchain 2020**, p. 371–378, 2020.

SANTHANAM, T.; VELMURUGAN, T. Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. **Journal of Computer Science**, v. 6, n. 3, p. 363–368, 2010.

SANTOS, J. A.; PARRA FILHO, D. **Metodologia científica**. Cengage Learning Brasil, 2012.

SAXENA, A.; PRASAD, M.; GUPTA, A.; et al. A Review of Clustering Techniques and Developments. , 2017.

SCHLIEP, K. P. phangorn: Phylogenetic analysis in R. **Bioinformatics**, v. 27, n. 4, p. 592–593, 2011.

SCHMID, B. F.; LINDEMANN, M. A. Elements of a reference model for electronic markets. **Proceedings of the Hawaii International Conference on System Sciences**, v. 4, n. c, p. 193–201, 1998.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying CRISP-DM process model. **Procedia Computer Science**, v. 181, n. 2019, p. 526–534, 2021. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.procs.2021.01.199>. .

SCHUBERT, E.; ROUSSEEUW, P. J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 11807 LNCS, p. 171–187, 2019.

SCHUTT, R.; O'NEIL, C. **Doing data science**. 2014.

SCHWIND, M.; STENGER, A.; APONTE, S. Electronic Transportation Marketplaces : How C an Green-IS H elp to P romote S ustainable L ogistics ? 44th Hawaii International Conference on System Science. **Anais...** . p.1–8, 2011. IEEE.

SCRUCCA, L.; RAFTERY, A. E. Clustvarsel: A package implementing variable selection for

Gaussian model-based clustering in R. **Journal of Statistical Software**, v. 84, n. 1, 2018.

SEGHIER, M. L. Clustering of fMRI data: The elusive optimal number of clusters. **PeerJ**, v. 2018, n. 10, 2018.

SEMAAN, G. S.; FADEL, A. C.; BRITO, J. A. D. M.; OCHI, L. S. A hybrid heuristic with hopkins statistic for the automatic clustering problem. **IEEE Latin America Transactions**, v. 17, n. 1, p. 1–17, 2019.

SHARF, Z.; RAZZAK, M. The Informative Vector Selection in Active Learning using Divisive Analysis. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 10, 2017.

SHIPSTA. Digital marketplaces are conquering the logistics industry. Disponível em: <https://blog.shipsta.com/en/blog/logistics-marketplaces>. Acesso em: 26/6/2022.

SIMPLILEARN. What is R: Overview, its Applications and what is R used for? Disponível em: <https://www.simplilearn.com/what-is-r-article#:~:text=R offers a wide variety,for data importing and cleaning.>. .

SIQUEIRA JUNIOR, L. A. DE. **Uso de machine learning para classificação de fornecedores no contexto da Data Science**, 2021. Universidade Federal de Itajubá.

STOCKDALE, R.; STANDING, C. A framework for the selection of electronic marketplaces: A content analysis approach. **Internet Research**, v. 12, n. 3, p. 221–234, 2002.

SUGANYA, R.; SHANTHI, R. Fuzzy C- Means Algorithm- A Review. , v. 2, n. 11, p. 440–442, 2012.

TANG, R.; FONG, S. Clustering big IoT data by metaheuristic optimized mini-batch and parallel partition-based DGC in Hadoop. **Future Generation Computer Systems**, v. 86, p. 1395–1412, 2018. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.future.2018.03.006>. .

THARWAT, A. Principal Component Analysis (PCA) : An Overview Principal Component Analysis : An Overview Alaa Tharwat. , , n. March 2016, 2018.

TURBAN, E.; KING, D.; LEE, J. K.; LIANG, T.-P.; TURBAN, D. C. **Electronic commerce: A managerial and social networks perspective**. 2015.

TURRIONI, J. B.; MELLO, C. H. P. **Metodologia de pesquisa em Engenharia de Produção: estratégias, métodos e técnicas para condução de pesquisas quantitativas e qualitativas**. Itajubá: Universidade Federal de Itajubá, 2012.

WALLER, M. A.; FAWCETT, S. E. Data Science , Predictive Analytics , and Big Data : A Revolution That Will Transform Supply Chain Design and Management. , v. 34, n. 2, p. 77–84, 2013.

WANG, X.; YANG, Y.; CHEN, M.; et al. AGNES-SMOTE : An Oversampling Algorithm Based on Hierarchical Clustering and Improved SMOTE. , v. 2020, 2020. Hindawi.

WANG, Y.; POTTER, A.; NAIM, M. Electronic marketplaces for tailored logistics. **Industrial Management and Data Systems**, v. 107, n. 8, p. 1170–1187, 2007.

WANG, Y.; POTTER, A.; NAIM, M.; BEEVOR, D. A case study exploring drivers and implications of collaborative electronic logistics marketplaces. **Industrial Marketing Management**, v. 40, n. 4, p. 612–623, 2011. Elsevier Inc. Disponível em: <http://dx.doi.org/10.1016/j.indmarman.2010.12.015>. .

WEI, C. P.; LEE, Y. H.; HSU, C. M. Empirical comparison of fast partitioning-based clustering

algorithms for large data sets. **Expert Systems with Applications**, v. 24, n. 4, p. 351–363, 2003.

WESTER, M.; OTTO, B. Blockelm - A public blockchain freight exchange protocol. **Proceedings of the Annual Hawaii International Conference on System Sciences**, v. 2020-Janua, p. 5587–5596, 2021.

WICKHAM, H.; SEIDEL, D.; RSTUDIO. Package ‘ scales ’. , 2022. Disponível em: <https://cran.r-project.org/web/packages/scales/index.html>. .

WIJUNIAMURTI, S.; NUGROHO, S.; RACHMAWATI, R. JSDS : JOURNAL OF STATISTICS AND DATA SCIENCE Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method For Hierarchical Clustering On Some Distance Measurement Concepts. , v. 1, n. 1, p. 1–5, 2022.

WRIGHT, K. Package ‘ hopkins ’. , 2022. Disponível em: <https://cran.r-project.org/web/packages/hopkins/index.html>. .

XIAO, W.; GAN, M.; LIU, H.; LIU, X. Modeling and Prediction of the Volatility of the Freight Rate in the Roadway Freight Market of China. **Mathematical Problems in Engineering**, v. 2020, 2020.

XIAO, W.; XU, C.; LIU, H.; LIU, X. Volatility Transmission in Chinese Trucking Markets: An Application Using BEKK, CCC and DCC-MGARCH Models. **Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020**, , n. i, p. 1179–1188, 2020.

XIAO, W.; XU, C.; LIU, H.; YANG, H.; LIU, X. Short-term truckload spot rates’ prediction in consideration of temporal and between-route correlations. **IEEE Access**, v. 8, p. 81173–81189, 2020.

XU, D.; TIAN, Y. A Comprehensive Survey of Clustering Algorithms. **Annals of Data Science**, v. 2, n. 2, p. 165–193, 2015. Springer Berlin Heidelberg.

XU, S. X.; HUANG, G. Q. Transportation service procurement in periodic sealed double auctions with stochastic demand and supply. **Transportation Research Part B: Methodological**, v. 56, p. 136–160, 2013. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.trb.2013.07.015>. .

XU, Z.; LU, Y.; JIANG, Y. Research on mini-batch affinity propagation clustering algorithm. Proceedings - 2022 IEEE 9th International Conference on Data Science and Advanced Analytics. **Anais...** , 2022.

YADAV, K.; BARIA, J. Mini-Batch K-Means Clustering Using Map-Reduce in Hadoop. **International Journal of Computer Science and Information Technology Research**, v. 2, n. 2, p. 336–342, 2014.

YAN, J.; CHEN, J.; ZHAN, J.; et al. Automatic identification of rock discontinuity sets using modified agglomerative nesting algorithm. **Bulletin of Engineering Geology and the Environment**, p. 1–15, 2022. Springer Berlin Heidelberg. Disponível em: <https://doi.org/10.1007/s10064-022-02724-w>. .

ZERZUCHA, P.; WALCZAK, B. Concept of (dis) similarity in data analysis. , , n. January, 2016.

ZHANG, Y.; TANGWONGSAN, K.; TIRTHAPURA, S. Streaming k-means clustering with fast queries. **Proceedings - International Conference on Data Engineering**, p. 449–460, 2017.

ZHU, J.; LIU, W. A tale of two databases : the use of Web of Science and Scopus in academic papers A tale of two databases : the use of Web of Science and Scopus in academic papers. **Scientometrics**, v. 123, n. 1, p. 321–335, 2020. Springer International Publishing. Disponível em: <https://doi.org/10.1007/s11192-020-03387-8>. .

APÊNDICE A – EXEMPLO DE CÓDIGO R USADO NA CLUSTERIZAÇÃO

```
#Carrega as bibliotecas que serão utilizadas na sessão.
library(cluster)
library(factoextra)
library(FactoMineR)
library(scales)
library(hopkins)

#Configura uma semente para que os procedimentos sejam reprodutíveis.
set.seed(123)

#Empilha os dados.
data <- rbind(base.2019, base.2020.a, base.2020.b, base.2020.c, base.2021.a, base.2021.b, base.2021.c)

#Remove as entradas com peso maior que 65000 kg.
all.data <- subset(data, weight_kg < 65000)

#Aplica o Método do Intervalo Interquartil para remoção de outliers.
q1.distance <- quantile(all.data$distance_km, 0.25)
q3.distance <- quantile(all.data$distance_km, 0.75)
iqr.distance <- IQR(all.data$distance_km)
lim.inf.distance <- q1.distance - (1.5*iqr.distance)
lim.sup.distance <- q3.distance + (1.5*iqr.distance)
q1.amount <- quantile(all.data$truck_amount, 0.25)
q3.amount <- quantile(all.data$truck_amount, 0.75)
iqr.amount <- IQR(all.data$truck_amount)
lim.inf.amount <- q1.amount - (1.5*iqr.amount)
lim.sup.amount <- q3.amount + (1.5*iqr.amount)
q1.value <- quantile(all.data$truck_value, 0.25)
q3.value <- quantile(all.data$truck_value, 0.75)
```

```
iqr.value <- IQR(all.data$truck_value)
lim.inf.value <- q1.value - (1.5*iqr.value)
lim.sup.value <- q3.value + (1.5*iqr.value)
all_data <- subset(all.data, distance_km > lim.inf.distance & distance_km < lim.sup.distance & truck_amount > lim.inf.amount & truck_amount < lim.sup.amount & truck_value > lim.inf.value & truck_value < lim.sup.value)

#Grava os valores máximos e mínimos
min.distance <- min(all.data$distance_km)
max.distance <- max(all.data$distance_km)
min.weight <- min(all.data$weight_kg)
max.weight <- max(all.data$weight_kg)
min.amount <- min(all.data$truck_amount)
max.amount <- max(all.data$truck_amount)
min.value <- min(all.data$truck_value)
max.value <- max(all.data$truck_value)

#Normaliza os dados pelo Método Min-Máx
all.data$distance_km <- rescale(all.data$distance_km, to = c(0,1))
all.data$weight_kg <- rescale(all.data$weight_kg, to = c(0,1))
all.data$truck_amount <- rescale(all.data$truck_amount, to = c(0,1))
all.data$truck_value <- rescale(all.data$truck_value, to = c(0,1))

#Calcula a estatística de Hopkins
hopkins.results <- hopkins(all.data, m = 100)
print(hopkins_results)

#Método visual heatmap
datasample <- all.data[sample(nrow(all.data), 1000, replace = FALSE), ]
fviz_dist(dist(datasample), show_labels = FALSE)

#Implementa o algoritmo CLARA
clara_k2 <- clara(x = all.data, k = 2, metric = “euclidean”, stand = TRUE, cluster.only = FALSE, samples = 100, sampsize = 1000, rngR = TRUE)
```

```
#Plota a clusterização, utilizando a técnica PCA
fviz_cluster(clara_k2, geom = “point”, pointsize = 0.25, ellipse.type = “convex”)

#Plota os resultados do índice silhueta para clusterização
fviz_silhouette(clara_k2)

#Visualiza as informações da clusterização
print(clara_k2$medoids)
print(clara_k2$clusinfo)
print(clara_k2$silinfo)
```